Abstract

As the economic opportunities fostered by large cities become more diverse, the travel patterns of public transport users become more heterogeneous. From personalized customer information, to improved travel demand models, understanding these heterogeneous travel patterns is useful for a number of applications relevant to public transport agencies. This thesis explores how smart card data can be used to analyze and compare the structure of individual travel patterns observed over several weeks. Specifically, the way in which multiple journeys and activities are ordered and combined into repeated patterns, both by the same individual over time and across individuals is evaluated from the journey sequence of each user.

The research is structured around three objectives. First, we introduce a representation of individual travel patterns and develop a measure of travel sequence regularity. The mobility of each individual is modeled as a stochastic process with memory, of which each new realization represents an activity or journey. Entropy rate, a measure of randomness in the stochastic process, is used to quantify repetition in the order of journeys and activities. This analysis reveals that the order of events is an important component of regularity not explicitly captured in previous literature. Second, we develop an approach to identify clusters of travel patterns with similar structure considered with respect to public transport usage and activity patterns. Finally, we present an exploratory evaluation of the associations between the identified clusters and socio-demographic characteristics by linking smart card data to an annual travel diary survey. These three objectives are considered in the context of a practical application using the transactions of a sample of approximately 100,000 users collected between February 10th and March 10th 2015 in London.
Acknowledgments

Exploring new ideas and probing unanswered questions is a very fulfilling experience. Over the last two years, my most valuable moments were those spent thinking about these kinds of questions out loud with my peers. I am deeply grateful to Haris Koutsopoulos and Jinhua Zhao for sharing many such moments with me and for helping me develop ideas with rigor. Thank you Haris for your dedication and enthusiasm for our work. Thank you Jinhua for always offering a fresh perspective on the research and its context.

This thesis would not have been possible without the precious help and contribution of the Transport for London staff. I cannot thank Lauren Sager Weinstein, Dale Campbell, James McNair, Duncan Horne, Neil Bichard, Graeme Fairnie, and Lucy Fish enough for their insight, feedback, and support over the last two years. I am also thankful to Guy Allen and his team for the technical support and after-work camaraderie.

I feel truly lucky for my far away loved ones. From Montreal, to Torino, to London, to Victoria, my friends and family always provided much needed love and support. Thank you maman and Laurence for your many visits and for pretending you weren’t bored of Boston yet.

Finally, thank you to all my MST colleagues for making my experience at MIT such a good one. From the free-food alerts to the nerdy jokes and puns, I am grateful for the many moments of joy shared with you. Thank you Gabriel Sanchez-Martinez for your sincere curiosity and generosity, and for your infinite resourcefulness. Thank you to my office neighbors for coping with the late night songs and spontaneous questions. Thank you Anne Halvorsen, Michel Babany, and Raphael Dumas for the adventures, discussions, and debates.
## Contents

List of Figures ................................. 11
List of Tables ................................ 13

1 Introduction .................................. 15
   1.1 Motivation ................................ 16
       1.1.1 Adapting Transit Systems ............. 17
       1.1.2 Influencing Passenger Behavior ........ 17
       1.1.3 Developing Commercial Opportunities .... 18
   1.2 Objectives and Approach ................. 18
   1.3 Thesis Organization ........................ 20

2 Background & Framework of Analysis ....... 21
   2.1 Introduction ............................... 21
   2.2 Concepts .................................. 22
       2.2.1 Activity Based Travel Theory .......... 22
       2.2.2 Lifestyle ................................ 23
   2.3 Activities, Lifestyle & Travel Patterns ... 25
   2.4 Observing Travel Patterns ................. 26
   2.5 Framework of Analysis ...................... 29

3 Constructing Travel Sequences ............. 31
   3.1 Introduction ................................ 31
   3.2 Geographical User-areas ................. 33
       3.2.1 Geographical Clustering Approaches .... 35
   3.3 Activity Sequence Inference ............... 38
   3.4 Application ................................. 40
       3.4.1 Data .................................. 40
       3.4.2 Geographical User-areas ............... 42
       3.4.3 Activity Sequence Inference ........... 47
   3.5 Conclusion .................................. 51

4 Quantifying Regularity in Travel Sequences 53
   4.1 Introduction ............................... 53
   4.2 Sequence Representation .................... 59
7.1.1 Activity Sequence Inference ........................................ 151
7.1.2 Travel Sequence Regularity ......................................... 152
7.1.3 Travel Sequence Clusters ........................................... 153
7.2 Limitations & Future Work ........................................... 155
  7.2.1 Inferring Activity Patterns ....................................... 155
  7.2.2 Travel Behavior Regularity ....................................... 157
  7.2.3 Travel Pattern Classification ..................................... 158

Bibliography ................................................................. 161
## List of Figures

2-1 Activity-Based Travel Theory ........................................... 23  
2-2 Partial Observation of Travel Behavior Through Smart Card Data  27  
2-3 Analysis Framework ...................................................... 29  

3-1 Example Sequential Representations of Travel Patterns .......... 32  
3-2 Travel Sequence Representation .................................... 33  
3-3 Complete Distance ...................................................... 36  
3-4 Distribution of Number of Journeys per User .................... 41  
3-5 Distribution of Number of Distinct Stations Visited by User ... 42  
3-6 Distribution of User-Area Diameter ................................ 43  
3-7 Journeys with Origin and Destination in the Same User-area ... 44  
3-8 Users with a Single Area .............................................. 45  
3-9 Geographical Areas of Example User ............................... 45  
3-10 Distribution of Number of Areas per User....................... 46  
3-11 Distribution of Number of Locations per User-Area .......... 46  
3-12 Activity Sequence Inference of Example User .................. 47  
3-13 Distribution of Activity Duration by Status .................... 49  
3-14 Distribution of Activity Duration by Activity Status ........ 50  

4-1 Event Order - Activity Location Sequence ....................... 56  
4-2 Event Order - First Journey Time Sequence ..................... 57  
4-3 Travel Sequence Representation ................................... 59  
4-4 Location-Duration Sequence ....................................... 61  
4-5 Activity Duration Distribution .................................... 62  
4-6 Calendar Events Sequence ......................................... 63  
4-7 Input Location Sequence - Example Representation ............ 70  
4-8 Durationless Location Sequence - Example Representation ... 70  
4-9 Sequence Length Distribution ..................................... 71  
4-10 Entropy Distributions .............................................. 72  
4-11 Location of User 1 .................................................... 73  
4-12 Location of User 2 .................................................... 74  
4-13 Comparison of users in the lower and upper 10\textsuperscript{th} percentile of irregularity 75  

5-1 Six Example Individual Travel Patterns .......................... 78  
5-2 Examples of Day-to-Day Transit Usage Patterns ................. 80
# List of Tables

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Table Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Activity Status Summary</td>
<td>39</td>
</tr>
<tr>
<td>3.2</td>
<td>Number of Stages per Journey</td>
<td>40</td>
</tr>
<tr>
<td>3.3</td>
<td>Missing Journey Information</td>
<td>41</td>
</tr>
<tr>
<td>3.4</td>
<td>Distribution of Time Inferred Across Activity Status</td>
<td>48</td>
</tr>
<tr>
<td>3.5</td>
<td>Inference by Interval Categories</td>
<td>50</td>
</tr>
<tr>
<td>4.1</td>
<td>Activity Status Summary</td>
<td>61</td>
</tr>
<tr>
<td>4.2</td>
<td>BWT Illustration (adapted from (Adjeroh et al., 2008))</td>
<td>68</td>
</tr>
<tr>
<td>4.3</td>
<td>Activity Sequence - Output Structure</td>
<td>69</td>
</tr>
<tr>
<td>5.1</td>
<td>Metrics of Transit Usage Frequency and Distribution</td>
<td>83</td>
</tr>
<tr>
<td>5.2</td>
<td>Cluster Description: 3 Clusters</td>
<td>89</td>
</tr>
<tr>
<td>5.3</td>
<td>Aggregated Activity Statuses</td>
<td>93</td>
</tr>
<tr>
<td>5.4</td>
<td>Clustering Results (k=4) - Descriptive Statistics</td>
<td>100</td>
</tr>
<tr>
<td>5.5</td>
<td>Cross Tabulation of $G_{ij}^{11}$ and $G_{ij}^{4}$</td>
<td>103</td>
</tr>
<tr>
<td>5.6</td>
<td>11 Clusters - Descriptive Statistics</td>
<td>104</td>
</tr>
<tr>
<td>5.7</td>
<td>Cross Tabulation of 1-PC and 2-PC Clusters (k = 11)</td>
<td>114</td>
</tr>
<tr>
<td>5.8</td>
<td>Reordered Cross Tabulation of 1-PC and 2-PC Clusters</td>
<td>114</td>
</tr>
<tr>
<td>5.9</td>
<td>Transit Usage Clusters - Period Comparison</td>
<td>116</td>
</tr>
<tr>
<td>5.10</td>
<td>February-October Clustering Overlap</td>
<td>119</td>
</tr>
<tr>
<td>5.11</td>
<td>Relative Cluster Size Comparison - 11 Clusters</td>
<td>119</td>
</tr>
<tr>
<td>6.1</td>
<td>Distribution of Card Types by Transit Usage Cluster</td>
<td>123</td>
</tr>
<tr>
<td>6.2</td>
<td>Odds Ratio - Cluster and Characteristic</td>
<td>125</td>
</tr>
<tr>
<td>6.3</td>
<td>Odds-Ratio Card Type - Transit Usage Clusters</td>
<td>126</td>
</tr>
<tr>
<td>6.4</td>
<td>Odds-Ratio Card Type - 11 Clusters</td>
<td>126</td>
</tr>
<tr>
<td>6.5</td>
<td>LTDS Transit Usage Classification</td>
<td>131</td>
</tr>
<tr>
<td>6.6</td>
<td>LTDS Travel Sequence Classification (4 Clusters)</td>
<td>132</td>
</tr>
<tr>
<td>6.7</td>
<td>LTDS Travel Sequence Classification (11 Clusters)</td>
<td>132</td>
</tr>
<tr>
<td>6.8</td>
<td>Descriptive Overview of Cluster Demographics</td>
<td>133</td>
</tr>
<tr>
<td>6.9</td>
<td>Distribution of Respondents by Occupation Status</td>
<td>134</td>
</tr>
<tr>
<td>6.10</td>
<td>Odds-Ratio Occupation Status - 11 Clusters</td>
<td>135</td>
</tr>
<tr>
<td>6.11</td>
<td>Odds-Ratio Age Categories - 11 Clusters</td>
<td>136</td>
</tr>
<tr>
<td>6.12</td>
<td>Odds-Ratio Income Categories - 11 Clusters</td>
<td>138</td>
</tr>
<tr>
<td>6.13</td>
<td>Odds-Ratio Household Composition - 11 Clusters</td>
<td>139</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>6.14 Household Access to Vehicle and Transit Usage</td>
<td>139</td>
<td></td>
</tr>
<tr>
<td>6.15 Odds-Ratio Household Access to Vehicle - Transit Usage Clusters</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>6.16 Odds-Ratio Household Access to Vehicle - 11 Clusters</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>6.17 Cluster Membership Model - Coefficient Estimates</td>
<td>147</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

As cities grow and attract ever more diverse populations, their social, cultural, and economic organization become more complex. Complex economies and the varied opportunities they foster are reflected in the increasingly heterogeneous lives of the inhabitants of large cities. Whereas 9-to-5 commuters once accounted for an overwhelming proportion of the urban population, nonconventional routines, for example driven by shift work, multi-employment, flexible work, or self-employment, now guide a significant proportion of urban and suburban lifestyles. With the growing heterogeneity of activity schedules comes a wide variety of travel needs and habits. Journeys, characterized by different locations, times, routes and modes, are combined within days and across days in line with increasingly diverse activity patterns. In parallel, public transport agencies operating in large metropolitan areas are faced with growing ridership, changing demand patterns, and changing customer expectations. Understanding these patterns is crucial to agencies’ ability to respond to developments in urban travel demand.

In order to understand and accommodate diverse travel patterns, assumptions of homogeneity in the population must carefully be revisited. What was once considered a single population of transit users must now be recognized as a composition of multiple types of users, each with different travel needs and behavior. This change begs for a shift in the conventional approach of agencies from the aggregate to the user-centric. Whereas agencies have typically focused on understanding travel demand at the aggregate level (e.g. aggregate ridership along a route or a network, or average travel time between origin-destination pairs), a user-centric view of the system focuses on understanding travel patterns and behavior on an individual basis.

To illustrate the contrast between the two approaches, consider 500 bus journeys between origin A and destination B observed between 13:00 and 14:00 on a given day. At the aggregate level, the 500 journeys performed during the same hour of the day, on the same mode, and between the same locations are identical. At the disaggregate level, however, each one of the 500 journey is completed in the context of a unique activity-travel pattern. For each individual, the journey from A to B
precedes and follows different journeys; some individuals may have traveled from C to A, or from D to A before arriving at A, and some may travel from B to D or from B to A after visiting B. As the length of the pattern within which the A-to-B journey is considered increases, each user’s journey combination becomes increasingly distinct. A few individuals may share the same journey combination on a given day, but very few users are likely to have visited the same locations in the same order over a week or a month. As routines and lifestyles become more diverse, so does the structure driving how users combine multiple journeys. Individual-focused analysis allows for heterogeneity in this structure to be explicitly captured and accounted for in transit operations and planning.

As illustrated in the example above, understanding the structure driving activity-travel patterns requires the journeys of a given individual not be considered independently, but rather, in relationship to the other journeys with which they are combined. Distilling such relationships at the individual level can only be done from multiple observations of the same user’s travel over time. Hence, longitudinal data provided by modern automatic fare collection systems offers unprecedented opportunities for user-centric analysis. Smart card fare systems, such as the Oyster card in London, the Octopus card in Hong Kong, or the Clipper card in the San Francisco Bay area, provide a continuous stream of disaggregate data about the travel of each user. Recent advances in the processing of these transactions (Chu and Chapleau, 2010, Gordon, 2012, and Ma et al., 2013) have demonstrated how detailed itineraries spanning multiple public transport modes can be reconstructed from smart card data. While studies have used smart card data to investigate a wide variety of topics, from transit service quality and reliability (Trépanier et al., 2009, Uniman et al., 2010, Wood, 2015, and Tarte, 2015), to route choice and demand modeling (Chu and Chapleau, 2008, Jang, 2010, and Viggiano et al., 2014), few studies have leveraged the longitudinal nature of this data to analyze the structure of activity-travel patterns captured over multiple days.

The thesis explores how smart card data can be used to analyze and compare the structure of individual travel patterns observed over several weeks. Specifically, the way in which multiple journeys and activities are ordered and combined into repeated patterns, both by the same individual over time and across individuals, constitutes the primary focus of this research.

1.1 Motivation

Understanding and accounting for user heterogeneity is beneficial to transit agencies in two ways. First, it allows for the interaction between agencies and riders to be tailored on an individual by individual basis. For example, from information provision about disruptions and crowding levels to customized journey planning, disaggregate information about users can be used to improve the way agencies communicate with individual passengers. Such customized agency-passenger interaction is
relevant across multiple applications, be they related to user communication, flexible fare structures (Brakewood and Kocur, 2013), or innovative incentive schemes for demand management (Pluntke and Prabhakar, 2013, Halvorsen, 2015). Second, understanding user heterogeneity is useful to identify similarities between users. While each user is unique to some level, some passengers share similar activity-travel patterns. Analyzing these similarities reveals how different types of users can be grouped according to the way they travel and use public transport. Given knowledge about such segments, it is possible to adapt transit operations and planning on a segment by segment basis. The benefits of user-centric analysis, related to both user by user customization and user segmentation, are relevant across a wide array of applications. We group these applications into three categories summarized below.

1.1.1 Adapting Transit Systems

Detailed understanding of user behavior can be leveraged to adapt multiple aspects of transit systems to the diverse needs of its riders. User heterogeneity can inform the functioning of agencies across most of their key roles, from customer information provision, to fare policy design, to service planning and network design, to performance measurement, to real-time service control (Wilson et al., 2009). In real-time, observed journeys can be used to tailor the information disseminated to each user. For example, knowing that a given individual traveled from station A to station B in the morning and that she usually returns to A after visiting B, a customized alert could be sent to the user if a disruption alters the service between A and B. In the medium-term, the types of services available on different portions of the network can be modified according to the needs of different user groups. For instance, stations used by many passengers unfamiliar with the transit system, such as tourists, may require more in station personnel than stations used almost exclusively by frequent commuters. In the long-term, retrospective analysis of common activity patterns can be used to improve demand models and to track the evolution of demand patterns over time. Such analysis ensures that long-term network planning and development is aligned with heterogeneities in user needs. In general, these three examples illustrate how the transit system can be better adapted to its users if a better understanding of individual travel patterns is available.

1.1.2 Influencing Passenger Behavior

In addition to adapting the transit system to its users, disaggregate analysis of smart card data can inform how user behavior can be influenced to achieve desirable aggregate demand changes. As transit networks in rapidly growing cities are faced with increased crowding issues, managing travel demand in peak periods has proved a useful complement to infrastructure expansions. Examples such as the Singapore (Pluntke and Prabhakar, 2013) and Hong-Kong (Halvorsen, 2015) schemes illustrate how incentives and information provision can contribute to diverting demand away
from highly crowded routes and time periods. In order to maximize the effectiveness of the information and incentives provided to passengers, it is necessary to identify and target users who are most likely to respond to the scheme. Recent attempts at demand management in London illustrate that indiscriminately diffused messages can induce frustration among members of the public who perceive they are unrealistically asked to change their travel route or time (Adams, 2013). However, analyzing the extend of regularity in travel patterns at the individual level may reveal users with more flexible schedules who should be targeted by customized travel demand management campaigns.

1.1.3 Developing Commercial Opportunities

A third motivation, less directly related to the core role of a public transport agency, relates to marketing opportunities provided by smart card data. The type of user-centric analysis previously described provides a detailed picture of the user population in specific stations and along specific lines or corridors. This information can be used to optimize revenue generated by in-station and in-vehicle advertising and commercial space. For example, certain types of retailers, such as prepared meal outlets, dry-cleaners, or business apparel stores may be willing to pay a premium for commercial spaces located in stations primarily used by specific user groups. User data can be used to provide tangible evidence of the distribution of different types of passengers across the network. Through this approach, agencies can strengthen the value of the advertising and retail spaces on their network, and hence increase revenue collected from non-fare sources.

1.2 Objectives and Approach

In line with the motivation described above, this research focuses on the individual by individual and segment by segment analysis of travel behavior. The overarching objective of the thesis is to develop methods to evaluate and quantify how activity travel patterns repeat over time for a given individual and across the user population. This is achieved by addressing the following objectives.

1. Develop a measure of user regularity capturing the extent of repetition in the travel behavior of an individual observed over time.

2. Identify clusters of travel patterns observed across individuals from smart card fare transactions.

3. Evaluate the association between these clusters of travel patterns and long-term elements of lifestyle as captured from socio-demographic characteristics revealed in a travel survey.
These objectives are addressed through conceptual and methodological contributions illustrated in the context of concrete applications. Throughout the thesis, a sample of approximately 100,000 smart card users from Transport for London’s network is used. The sample includes all transactions recorded for these users between February 10th and March 10th 2015. All bus, light-rail and heavy-rail journeys completed in London over the 4 weeks covered by the analysis period are captured.

The first step to achieving the objectives is to develop a representation of each individual which captures the sequence of journeys and activities performed over the analysis period. This travel sequence is reconstructed from the smart card transactions of each user. First, all stops and stations visited by a given user are clustered into geographical areas aligned with the activities of the individual. Second, the location of the individual is inferred for all intervals preceding and following journeys using the previously defined user-specific areas. The resulting output is a sequence of intervals, each characterized by a duration and by the area in which the user was inferred to be, which represents the activity pattern of the user over the 4-week period analyzed.

In line with the first objective, the regularity of each user’s travel sequence reconstructed over time is evaluated. An index of regularity sensitive to the relative frequency and order of activities is developed based on information theory concepts. The mobility of each individual is modeled as a stochastic process with memory, of which each new realization represents an activity. The regularity of each sequence is evaluated based on the average amount of randomness associated with each activity given the sequence of activities preceding it. The process entropy rate, estimated using the block-sorting algorithm described by Cai et al. (2004), is proposed as a measure of regularity.

In line with the second objective, users are compared based on the structure of their 4-week travel sequence. At a coarse level, users are first classified with respect to the distribution of their public transport usage over time. Focusing on the most frequent users, the comparison of activity patterns is then implemented by extracting recurrent elements of sequence structure through principal component analysis. The projection of each individual sequence onto the most important principal components is then used as input to the k-means cluster analysis. The robustness of the resulting clusters is evaluated through bootstrapping and cluster temporal stability is measured through cross-validation with a sample of transactions extracted at a later time period.

The association between the resulting clusters and different socio-demographic variables expected to impact activity-travel patterns is investigated using a yearly travel diary survey distributed among residents of greater London. Approximately 2,000 survey respondents for whom Oyster card number is known are classified according to the clusters previously identified. The distribution of respondents across clusters is analyzed to evaluate how certain travel patterns correlate with socio-demographic attributes.
1.3 Thesis Organization

The thesis is organized as follows. Chapter 2 summarizes the conceptual background used throughout this thesis, and introduces the research framework. Chapter 3 describes the methodology developed for the travel sequence and illustrates the resulting sequence for the user sample. Chapter 4 introduces the regularity metric used to characterize user sequences and applies the metric to the sample of travel sequences reconstructed in 3. Chapter 5 compares the similarities of user-sequences through clustering analysis and evaluates cluster stability and robustness. Chapter 6 presents an exploratory analysis of the socio-demographic characteristics of users allocated to clusters identified in Chapter 5. Finally, Chapter 7 presents the concluding remarks of the thesis.
Chapter 2

Background & Framework of Analysis

2.1 Introduction

smart card data contains partial traces of how people move through space and time within cities. This movement is driven by the activities people choose to engage in on a daily basis. In turn, individuals choose to engage in activities in line with the type of life they lead and with their long-term life choices, for example employment and housing situation. Hence, there is a connection between the way a person travels in the city and the type of life she leads. For example, whereas career oriented individuals who choose to live downtown, near their employment location in order to facilitate working long hours may walk to work, family oriented individuals who choose to live in suburban areas and work shorter hours in order to engage in more family related activities likely commute longer distances.

This connection implies that observing certain types of travel may provide information about the type of life driving it. Smart card data contains detailed information on the characteristics of each user’s travel. Therefore, it may be useful to infer information about certain aspects of users’ lives. Specifically, life choices related to housing, employment and car ownership have been observed to cluster around a limited number of discrete lifestyles. Smart card may provide insight about the long-term life choices of the people using public transit.

This chapter will develop this argument and propose a research framework for testing it. First, the concept of travel is contextualized by reviewing activity-based travel theory (ABTT), and the related concept of lifestyle. Second, practical considerations associated with the observation of travel patterns from smart card data are discussed. Finally, conceptual and practical considerations are integrated in the research framework presented in Section 2.5.
2.2 Concepts

In order to develop a framework for the analysis of individual travel patterns, it is first essential to contextualize the concept of travel and travel behavior. For this purpose, we rely on activity-based travel theory to discuss the underlying process driving travel. ABTT offers a rich framework to describe the relationship between travel, activity patterns, and long-term life choices. While this framework is traditionally used for travel demand forecasting, it constitutes a general basis to discuss travel behavior.

2.2.1 Activity Based Travel Theory

The basic idea behind activity-based travel theory is that individuals engage in activities taking place at different locations and must hence move through space to reach the location of each activity. Travel refers to this movement. Activities could include for example working, shopping, or visiting a friend. They are events with a purpose and a duration driving the individual to visit a location. Individuals have a finite amount of time to take part in activities and must choose which activities to engage in and in what order. Hence, as first conceptualized by Hägerstråand (1970), activities are organized in patterns which are structured according to the constraints facing each individual. Constraints related to the schedule, socio-demographics, and home and work location of a person govern her time-use decisions (Bhat and Koppelman, 1999). For example, employment, car ownership, and residential location all constrain what activity patterns can feasibly be take place in a finite 24 hour period.

The description above implies two important characteristics about travel. First, travel demand is derived from demand for activities distributed over space (Jones et al., 1990; Bhat and Koppelman, 1999; Bowman and Ben-Akiva, 2001; McNally and Rindt, 2008). In other words, travel is only a means to an end. Second, because travel is derived from activities organized in comprehensive schedules, individual journeys cannot be analyzed independently. Rather, they should be considered with a focus on the sequences or patterns within which they are realized (Jones et al., 1990).

Travel is derived from activities, which are organized in patterns according to the constrained time-use decisions of a person. While this idea represents the core of ABTT, significant developments in the theory over the last two decades have resulted in a more comprehensive framework (Ben-Akiva et al., 1996). This extended framework incorporates more explicitly the influence of longer-term factors, such as residential and employment choices, on short-term activity patterns. Figure 2-1 summarizes the relationship between short-term and long-term elements.

The diagram consists of two parts: the realized travel, consisting of observable journeys and their characteristics, and the decision process consisting of the underlying process generating the observed travel. The decision process is further broken-down according to the time-scale at which elements are relevant. The most fundamental
component is the life orientation of each individual. As will be exposed in the next section, life orientations drive the life choices and the preferences of each individual. In turn, these dictate the short-term constraints which structure activity patterns. The relationship between life orientation, life choices and preferences is best explained using the construct of lifestyle first formally introduced to ABTT by (Ben-Akiva et al., 1996). These long-term factors will be discussed in the next section.

2.2.2 Lifestyle

*Lifestyle* is commonly used in transportation and human geography research (Krizek and Waddell, 2002; Walker and Li, 2007). Though it has often been loosely defined, some key contributors have helped operationalize the concept in the context of travel theory (Walker and Li, 2007; Salomon, 1980; Krizek and Waddell, 2002; Ben-Akiva et al., 1996). The operational definitions presented by these authors all use a combination of four core concepts: *life orientations, life choices, lifestyle, and lifestyle groups*. *Life orientations*, also referred to as aspirations or lifestyle preferences, refer to future and long-term preferences for a way of living (Walker and Li, 2007). These describe individuals’ broad objectives for the type of life they target. For example, Salomon (1980) illustrates the concept by contrasting family and career oriented individuals.

In turn, life orientations drive *life choices*. Life choices, or life decisions, refer to long-term decisions which affect the constraints and preferences driving daily activity patterns. They are typically described as related to employment, housing, and vehicle ownership (Ben-Akiva et al., 1996; Waddell, 2001). Housing decisions include the choice of neighborhood and dwelling unit type. Employment decisions include the
choice of job location, job type, and working hours/work load. Vehicle ownership
decisions include choices on whether to own a vehicle, on the number and type of
vehicles to own, and on subscription to vehicle sharing systems. Of course, long-term
decisions are intimately related to each other, and life orientations do not guide them
independently.

As described by Krizek and Waddell (2002) life choices reinforce and mutually inform
each other, to form a strategic combination of decisions driven by life aspirations.
They describe the complex interactions between long-term choices and short-term
activity behavior as an integrated process. Similarly, Salomon (1980) and Salomon
and Ben-Akiva (1983) refer to a ‘pattern of behavior’ to describe the set of choices
made by individuals to implement their life orientations. Ben-Akiva et al. (1996);
Waddell (2001); Krizek and Waddell (2002); Salomon (1980) refer to this strategic
combination of life choices as lifestyle. In other words, a lifestyle is a specific pattern
of long-term life choices which dictate the short-term activity pattern of each individual.
As such, life decisions form an equilibrium aligned with specific life orientations.

While there are many possible combinations of life choices, only a small subset of these
constitute logical combinations for a given life orientation. For example, owning a car
and living in a low-density suburban neighborhood, or not owning a car and living in
a dense neighborhood where parking is scarce are two logical choice combinations in
line with two different life orientations. As choices are strategically combined with one
another, individuals with similar life orientations tend to have similar lifestyles.

For the reasons discussed above, there are not as many choice patterns, or lifestyles,
as there are individuals. Rather, certain choice patterns tend to repeat across the
population and cluster into lifestyle groups. A lifestyle group is a segment of individu-
als sharing with lifestyles, or life choice pattern. The relevance of lifestyle groups has
been recognized in urban geography and transportation research over the last decade.
The work of Walker and Li (2007); Krizek and Waddell (2002) demonstrates the ex-
istence of discrete lifestyle groups, and their impact on how life choices are made.
Krizek and Waddell (2002) used attributes related to travel and activity characteris-
tics, automobile ownership, and neighborhood urban form as clustering variables to
identify common life choice patterns. They identify 9 lifestyle groups: retirees, sin-
gle busy urbanists, elderly homebodies, urbanists with higher income, transit users,
suburban errand runners, family and activity-oriented participants, suburban worka-
holics, and exurban family commuters. Walker and Li (2007) studied the concept
of lifestyle in the context of residential choices. By modeling life orientation as a latent
variable in a discrete choice model, they identified three lifestyle groups: suburban
auto-oriented, urban oriented, and suburban transit-oriented. In contrast with Krizek
and Waddell (2002) who define lifestyle groups based on choice combination itself,
Walker and Li (2007) define lifestyle segments based on the latent orientations driv-
ing the choices. As pointed by Walker and Li (2007), this distinction is important
as the former approach assumes homogeneity in choices, whereas the later assumes
homogeneity in orientations. In both cases, however, the research demonstrates that
long-term choices tend to cluster into patterns because they are driven by underlying
life-orientations.

In summary, Life orientations drive long-term life decisions related to housing, employment, and car ownership. These decisions form a strategic combination, referred to as lifestyle, which tends to be repeated across the population. Repeated choice combinations are described as lifestyle segments or lifestyle groups. As described by ABTT, these long-term elements influence the constraints governing the travel and activity patterns implemented by individuals in the short-term.

2.3 Activities, Lifestyle & Travel Patterns

As introduced in Chapter 1 the overarching aim of this research is composed of the following three objectives:

1. Develop a measure of user regularity capturing the extent of repetition in the travel behavior of an individual observed over time.

2. Identify clusters of travel patterns observed across individuals from smart card fare transactions.

3. Evaluate the association between these clusters of travel patterns and long-term elements of lifestyle as captured from socio-demographic characteristics revealed in a travel survey.

The activity and lifestyle concepts presented in the previous section inform these objectives in two ways. First, they reveal how similarities in travel patterns might emerge in a population. Parallel to Krizek and Waddell’s (2002) and Waddell’s (2001) argument on strategic combinations of long-term choices, our second objective is to investigate the existence of strategic combinations of short-term travel behavior. The same way long-term life choices have been demonstrated to cluster into discrete lifestyle groups, we aim to identify whether short-term travel behavior also clusters into distinct travel patterns. Life decisions tend to organize in discrete lifestyles and these decisions dictate the constraints under which individuals operate in the short-term. In turn the constraints drive the way in which activities are organized, and hence the characteristics of travel (e.g. mode, time, route, location). This suggest that similarities in lifestyle may be associated with similarities in short-term travel patterns. This motivates the second and third objectives described above.

To the extent to they can be identified, clusters of travel patterns may be the result of different processes. On one hand, activity based travel theory suggests that similarities in lifestyle and socio-demographic characteristics may result in similar travel patterns. For example, full-time employed suburban dwellers in their forties who own a car may be more likely to travel regularly by car on working days. This kind of process could imply that certain travel patterns are associated with relatively homogeneous lifestyles. On the other hand, the same travel pattern could also be shared by users with distinct lifestyles and demographic attributes. For example, individuals
with different occupation statuses, such as self-employed individuals and university students, could have similar travel patterns. Hence, the third objective of this research is to evaluate the level of association between short-term travel patterns and long-term elements of lifestyle as described by Salomon and Ben-Akiva (1983); Krizek and Waddell (2002); Walker and Li (2007) captured through socio-demographic characteristics.

Second, activity based travel theory highlights the importance of analyzing journeys within the context of the sequence they are situated in. As described by Hägerstråand (1970) certain activity patterns include activities and journeys arranged in ‘non-permutable’ sequences. Hence, activity patterns are defined not only by the attributes of the activities and journeys they are composed of, but also by the order in which these activities are organized. For example, dropping-off the kids at the daycare, going to work, and returning to the daycare after work involves a logical sequence of activities. The combination of these three activities is constrained by the order in which they must be arranged. Therefore, the representation and analysis of this pattern would be incomplete without consideration for order. This insight is relevant to the first and second objectives. First, it reveals that a measure of individual regularity should capture not only repetition in the characteristics of activities and journeys, but also repetition in the order of combinations of activities and journeys. Repeated travel patterns are defined by the characteristics of travel events, but also by the sequence of these events, within a day and across multiple days. This observation extends to the comparison of travel patterns across individuals. Similarities in the travel of multiple users are defined with respect to activity and journey sequences. Identifying recurrent travel patterns across the user population should therefore be done with consideration for the order of journeys and activities.

### 2.4 Observing Travel Patterns

While the conceptual discussion introduced previously provides useful grounds for understanding the process driving travel behavior, it does not address the practical limitations involved in the observation of realized travel and activities. Indeed, different types of observations may only partially capture specific components of the process illustrated in Figure 2-1. Surveys have historically been the most common way of capturing individual and household travel patterns. Typical travel diary surveys commonly include reported information about trip and activity pattern, and about elements of lifestyle captured through socio-demographic variables. While they capture multiple elements of the process described in Figure 2-1, recent research has systematically questioned the accuracy of reported information on trip and activity patterns (Stopher et al., 2007; Trepanier et al., 2009; Riegel, 2013). In addition, partly due to their cost, travel dairy surveys are most commonly collected for short time periods, most typically for a single day, and cannot provide a full picture of the variation in activity and travel patterns over time.
More recently, technological developments have allowed for new sources of information to provide more objective observations of travel. Whether collected specifically for the purpose of studying travel, or passively for other purposes, digital traces of location from GPS enabled devices, cellphone records and smart card data, provide opportunities to infer travel patterns over longer time periods. Specifically, smart card data is widely available to a number of transit agencies and provides a continuous stream of information about the travel patterns of each card user.

**Figure 2-2:** Partial Observation of Travel Behavior Through Smart Card Data

While the information these new automated sources provide is more objective and covers longer time periods than travel diary surveys, smart card data only captures a portion of the total travel pattern of each individual. Figure 2-2 provides an analogy relating smart card data to the components of the process illustrated in Figure 2-1. Transactions recorded from Public transit journeys provide a window through which different portions of a person’s overall travel patterns can be observed. In turn, travel patterns may provide partial information about the lifestyle and attitudes by which they are driven. Smart card data is the result of the interaction between these three layers and it only reveals a specific portion of the travel behavior of each user. For example, some individuals use transit for all their trips, some use it occasionally for various purposes, some use it for specific purposes or to reach specific locations.
The size and disposition of the window varies across individuals, allowing for different portions of a person’s travel and hence lifestyle to be inferred. The individual’s choice to use public transportation to realize specific portions of her travel is represented by the size and disposition of this window

- Layer 1: The most directly observable layer consists of the usage of public transport. Public transport usage is considered at the journey level (e.g. mode, start time, origin, destination) and at the longitudinal level, accounting for the arrangement of journeys over time. Smart card transactions contain explicit information about each journey stage and can be used to infer unrecorded journey characteristics. For example, route choice models, destination inference, and passenger-to-train assignment models have all been successfully implemented in the literature.

- Layer 2: The second layer consists of each individual’s complete travel pattern. This includes all travel in the city, at different times of the day, on different modes, and for different activity purposes. This layer is only partially observed through the previous layer and the unobserved portions of the mobility are not randomly arranged. Rather, they correspond to travel for which the individual chooses not to use public transport. Hence, the spatial scale and time of day at which travel occurs and the type of activity driving it will influence the likelihood of capturing this travel through public transit.

- Layer 3: The third layer relates to the long-term elements of lifestyle and preferences described in section 2.2.2. As previously described, these elements influence the short-term constraints governing travel and activity patterns. Hence, the realized travel of an individual is, to some degree, a reflection of these lifestyle elements.

Research questions facing public transit agencies can focus on any of these levels. Identifying the nature of the question may instruct how different types of data should be used. Some questions might focus specifically on public transport usage (layer 1, e.g. bus user vs rail user, occasional user vs frequent user), while others might focus on preferences and attitudes (e.g. crowding averseness, habituality, cost sensitivity). As we move from the outer to inner layers, less information becomes available directly from smart card data, and other types of data, collected from surveys for instance, become necessary complements. For example, if we are trying to understand a specific users’ attitude towards unreliability, or crowding, objective journey characteristics can be obtained from smart card records allowing for surveys to focus on attitudinal question requiring the subjective input of users. Additionally, large transport agencies overseeing multiple modes across a metropolitan area may be able to combine data sources related to various transport modes to provide a more complete picture of each user’s mobility. For example, in the case of TfL, congestion pricing data, bike sharing data, and public transport data could be combined to reconstruct the multi-modal travel patterns of individuals.
2.5 Framework of Analysis

In the short-term, the realized travel of each individual, characterized by journey time, order, mode, and route, is driven by activity patterns organized according to time constraints, location constraints, and socio-demographic constraints. In the long-term, these constraints are established by elements of lifestyle, such as employment situation, housing, and vehicle ownership. Hence, to some extent, each individual’s realized travel pattern is connected to her socio-demographic situation. Smart card data provides partial traces of each user’s realized travel pattern. These traces may hence provide some information about aspects of the individual’s lifestyle.

Figure 2-3: Analysis Framework

Figure 2-3 illustrates the framework integrating the different pieces of analysis presented in this thesis. In the fashion of a reverse engineering approach, the framework parallels the activity-based travel theory process described in Figure 2-1. The traces of each user’s realized travel captured from smart card fare transactions are the initial input to the process. First, transactions are pieced together into a coherent representation of each individual’s activity pattern. In line with the importance of the links tying the multiple journeys and activities of a given person, this representation relies on the travel sequence formulation implemented in Chapter 3. Second, from the sequence representation of each user, similarities in individual travel patterns are extracted in order to identify clusters of travel sequences. Finally, the resulting clusters are related to the long-term elements of lifestyle described in section 2.2.2 in Chapter 6.
Chapter 3

Constructing Travel Sequences

3.1 Introduction

The fundamental departure of activity based travel theory (ABTT) from trip based travel theory was the recognition that trips cannot be analyzed in isolation. Rather, trips and activities are coherently organized over time (typically a day) and must be analyzed as a comprehensive combination of travel-related events. The way in which a person organizes multiple journeys and activities is equally as important as the characteristics of each journey or activity. While this principle has traditionally been considered in the context of demand forecasting, it is also highly relevant to our objectives. Comparing the travel behavior of multiple individuals with respect to the attributes of disconnected journeys alone would fail to capture important characteristics of each person’s travel. Hence, accounting for similarities and differences in both the attributes of journeys and activities and in the way they are organized over time is essential.

The ability to compare individuals’ travel patterns is inherently tied to the way in which these patterns are represented. In order to capture the characteristics of journeys and activities, as well as their organization, a representation of travel behavior which preserves the relationships between such journeys and activities is required. This chapter introduces a conceptual representation of travel behavior focused on capturing the relationship between journeys and activities, and presents an operational implementation of this representation using smart card data.

Prior to introducing the abstract definition of this representation, we illustrate the concepts involved with a concrete example. Figure 3-1 shows three different approaches to representing the travel patterns of an individual. The first consists of a sequence of activity purposes. The order in which activities are performed is preserved by the sequence, but the duration, time and location of each activity are abstracted. The second sequence summarizes a trajectory, including all geographical locations visited by an individual during a given period. The third sequence represents an
individual’s first journey departure time on different days of a 14-day period. Days on which the individual did not travel are represented by a value of 0. The fundamental structure of all three representations is similar; each representation consists of an ordered sequence of events. The feature distinguishing each sequence is the way in which events are defined and characterized. In the first, events are aligned with activities characterized by their purpose. In the second, events correspond to visits characterized by geographical locations. In the third, events correspond to days characterized by the time of the first journey on the day. The exact event definition and the attributes used to characterized events may vary with the research question and application considered.

Figure 3-2 illustrates a generalization of the three example sequences described above. In this general representation, an individual’s travel pattern is modeled by an ordered sequence of flexibly defined events which occur over time. Each travel sequence is defined by the attributes used to characterize its constituting events. Most intuitively, events can be aligned to the concept of activity as used by Ben-Akiva et al. (1996); Bhat and Koppelman (1999); Bowman and Ben-Akiva (2001); McNally and Rindt (2008). In this context, events are characterized by the attributes of activities. Each activity instance has a duration, a location, and a purpose. Journeys can also be considered as activities characterized by a route and a mode (rather than a static location). Each event occurs in time with respect to a background calendar of interest (e.g. time of day, day of week, month etc.). The relation of each event to the background calendar can also be integrated as an event characteristic (e.g activity time). The resulting representation captures the order in which an individual chooses to perform activities as well as the characteristic of each activity or journey. Depending on the research focus, certain event attributes can be abstracted. For instance, as illustrate above, activity location and time may be irrelevant to applications focused

Figure 3-1: Example Sequential Representations of Travel Patterns
While the most intuitive sequence representation aligns with activities, travel sequences can also be defined at a coarser level of aggregation. In this case, events represent some aggregation of multiple activities rather than aligning with a single activity. For example, events could be defined with respect to the aggregation of all activities performed on the same day. Then, each day becomes a distinct event characterized by some attributes of interest, such as first journey departure time (as illustrated in Figure 3-1).

The remainder of this chapter presents a methodology to implement this representation using smart card transactions. In order to do so, section 3.2 discusses approaches to group the individual public transport stops and stations visited by users into geographical areas aligned with different activities. Section 3.3 presents a methodology to infer the location of users between consecutive transit journeys. Finally, section 3.4 presents an application of these two methodologies using the transactions of a sample of smart card users from London.

### 3.2 Geographical User-areas

A high percentage of origin-destination pairs in dense metropolitan transit networks are connected by more than one route. Transport for London (TfL) passengers, for example, can often access the same geographical area of London through multiple stops and stations. Hence, what a user perceives as an origin or destination is not necessarily limited to a specific stop or station, but rather extends to an area of the city. We denote the individual stops and stations visited by a user by *locations* and clusters of locations used interchangeably by the same user by *user-areas*. The choice to use a specific stop within a given area may depend on a number of factors including the origin of the journey, the time of travel, the first available alternative, and crowding levels on different routes.

For the purpose of travel behavior analysis, it is useful to group locations into geographical clusters which reflect this user-centric view of the system. Segmenting the
city into discrete areas for each user is useful because travel is driven by discrete underlying activities. For each individual, a specific geographical area is often associated with a specific activity. The aim is to identify, for each user, groups of stops and stations which align with the same activity.

The geographical areas defined for each user should satisfy two important requirements in order to align with underlying activities. First, stops and stations in the same area should all lie within walking distance of the area center. Assuming this center aligns with the actual activity location, this ensures that the activity can be accessed by foot from any location in the user-area. Second, areas should be defined so that the number of full journeys with both origin and destination contained in the same area is minimized.

Different users might consider varying distances to be walkable, and may choose to use public transit for journeys of different length. This choice may depend, for example, on the user’s fare type, age, physical condition, and value of time. For example, Pay-as-you-go customers are likely to avoid making very short trips, while users holding seasonal tickets can complete any trip at no marginal cost. Similarly, elderly users may prefer using PT for shorter trips due to limited physical mobility, while young riders may choose to walk over longer distances to save money. The appropriate area size may hence vary from user to user.

Additionally, the minimum distance for which the same user chooses to use PT may vary from journey to journey. Examples of factors influencing this choice include trip purpose, trip location, time of day, and weather. The same user may be willing to walk longer distances to avoid waiting time in an ex-centric area with lower service frequency than in the urban core were service is more frequent. During the night time users may feel safer completing short journeys on public transport. Equally, a user carrying heavy items, for example after a shopping trip, may choose to use public transit for a short journey. Hence, what a user considers to be a walkable distance may vary across different OD pairs.

Given that the definition of walking distance may vary across users and across OD pairs for the same user, it is useful to consider origin-destination information to examine the relationship between stops and stations. Observing a public transit journey between two locations is indicative that distinct trip generating activities are associated with the origin and the destination location. Two locations between which a user completes journeys are likely to be associated with different activities given that the user perceives the two locations to be far enough apart to justify the use of public transportation. This property is especially relevant for individuals who travel locally and make short journeys. For example, pupils often use London buses to travel short distances between home and school, typically under 1000 m. Hence, user-areas should be defined based on both the geographical relationship and the observed journeys between locations.
3.2.1 Geographical Clustering Approaches

For each user \( u \), let \( X_u = \{x_1, x_2, \ldots, x_{n_u}\} \) be the set of all locations (stops or stations) visited by the user as the origin or destination of a journey over the period of analysis, where \( n_u \) represents the number of distinct locations visited by the user. Let \((x_l, x_m)\) denote a pair of locations, and \( P_u = \{(x_l, x_m)\} \) the set of all an undirected OD pairs visited by user \( u \), such that \( P_u \subseteq X_u \times X_u \). For each pair \((x_l, x_m)\in P_u\), \( t_{l,m} \) represents the number of journeys observed between locations \( l \) and \( m \), and \( T_u \) the total number of journeys completed by user \( u \). The set of all geographical user-areas defined for user \( u \) is \( A_u = \{A_1, A_2, \ldots, A_i, \ldots, A_{k_u}\} \) with \( k_u \) representing the number of areas defined for user \( u \), such that \( A_i \subseteq X_u \forall A_i \in A_u \). Recall that a separate set of geographical areas is defined for each user and that the same physical location may be associated with different user-areas from user to user. Finally, the distance for the maximum user-area size for user \( u \) is denoted by \( D_u \).

The area definition is flexible and can be adapted to accommodate observed travel between OD pairs for each user. For a given individual \( u \), the main approach to identify clusters follows the standard agglomerative hierarchical clustering procedure (Day and Edelsbrunner, 1984, for a detailed discussion of the algorithm). As described in Algorithm 1, areas consisting of sets of locations, are defined by iteratively merging the two closest areas until the two closest areas are separated by a distance above the predefined threshold \( D_u \). The different approaches introduced in the following sections vary with respect to the distance function \( \delta \) used to evaluate the distance between sets of points, and the definition of threshold \( D_u \). All approaches are implemented using MATLAB’s Statistics and Machine Learning Toolbox.

### Algorithm 1 Agglomerative Hierarchical Clustering

**Input:** All locations visited by user \( u \) \( X_u = \{x_1, x_2, \ldots, x_{n_u}\} \), maximum distance threshold \( D_u \), \( \delta \) distance function between two sets of locations

**Output:** Set of areas \( A_u \) for user \( u \) composed of clustered locations

1. Initialize each location to singleton area such that \( A_u = \{A_1, \ldots, A_{n_u}\} \)
2. \( \textbf{while } \min_{\{A_i, A_j \in A_u \times A_u, i \neq j\}} \delta(A_i, A_j) \leq D_u \textbf{ do} \)
3. \( (A_i, A_j) \leftarrow \arg \min_{\{A_i, A_j \in A_u \times A_u, i \neq j\}} \delta(A_i, A_j) \)
4. Merge \( A_i \) and \( A_j \) as one cluster
5. \( \textbf{end while} \)

**Simple Complete Distance** The simplest approach is presented as a basis to explain the more complex methods. In this approach the distance (or dissimilarity) between clusters is a function of geographical distance only. The distance \( \delta(A_i, A_j) \) between areas \( A_i \) and \( A_j \) is equal to the distance separating the two locations furthest away from each other in the two locations (see figure 3-3). This measure of distance between two sets of points is referred to as complete distance. The two least distant areas are merged at each iteration until \( \delta(A_i, A_j) > D_u \forall A_i \in A_u \). The size of the
merged area is equal to the distance between the two component areas $d(A_i, A_k)$. Hence, the largest area cannot be larger than the specified threshold $D_u$. Under this approach, the same threshold $D$ is specified across all users.

$$\delta(A_i, A_j) = \max_{x_l \in A_i, x_m \in A_j} (d(x_l, x_m))$$ (3.1)

where $d(x_l, x_m)$ denotes the euclidean distance between locations $x_l$ and $x_m$.

\[ \text{Figure 3-3: Complete Distance} \]

**Median Journey Threshold**  In order to incorporate variations in the perception of walkability across users, it is possible to specify a user specific threshold for the maximum area size. Let $m_u$ be the median journey length for user $u$, where journey length is the euclidean distance between the origin and destination location of the journey. Define a user specific threshold for user $u$ as $D_u = \min(D, \gamma \cdot m_u)$, where $D$ represents the population wide threshold and $0 < \gamma < 1$. For $\gamma = 0.5$, for example, the threshold specified for each user can be no larger than half the length of their median journey length, such that no travelled OD pair separated by this distance will be grouped in the same cluster. This approach is useful to account for variations across users, but fails when the same user completes both long and short journeys. For example, an individual commuting over long distances to work might also use public transit for shorter journeys around her home location. The median journey length of this individual would fail to distinguish the OD pairs separated by shorter distances.

**OD Flow Constraint**  An alternative to using an aggregate metric summarizing all journeys made by an individual, such as median journey length, is to consider each origin-destination pair individually. Under this approach, the area definition is constrained such that no area contains an origin destination pair accounting for over a given percentage $\tau$ of the journeys performed by the user. For this purpose,
the distance between locations is modified based on OD flow and then the complete distance approach described above is applied based on the modified distances. Specifically, the constraint is implemented by setting the distance between locations with high OD flows equal to the global area diameter threshold $D$. As a result, the two locations cannot be grouped in the same cluster using the complete distance clustering approach described previously. Equation 3.2 summarizes how the OD flow constraint is implemented.

$$d'(x_l, x_m) = \begin{cases} d(x_l, x_m) & \text{if } t_{l,m}/T_u < \tau \\ D & \text{if } t_{l,m}/T_u \geq \tau \end{cases}$$  \hfill (3.2)$$

where $d'(x_l, x_m)$ denotes the modified distance between two locations $l$ and $m$. This approach involves two parameters. $D$ is the global maximum area diameter across all users. Additionally, parameter $\tau$ specifies the maximum percentage of a user’s journeys for a given origin destination pair in the same area. In contrast with the median journey threshold approach, the OD flow constraint approach can account for users who complete both long and short public transit journeys. Regardless of a user’s median journey length, OD pairs which are frequently used by the user, indicating distinct activities, will not be merged in the same geographical area.

**Agglomeration by Number of Visits** Isaacman et al. (2011) propose a different type of approach which does not rely on hierarchical clustering with complete distance. Their approach, based on Hartigan’s leader algorithm (Hartigan, 1975), utilizes the number of visits made to each location. First, locations are ordered with respect to the number of times they are visited over a given period of analysis. Then, the top location is assigned to its own cluster. If the second location is within a specified distance of the first location, it is assigned to the first cluster and the centroid of the resulting cluster is computed. Otherwise, if the second location lies outside the specific threshold, a new cluster is formed. This process is repeated for each subsequent location, by considering the distance to the closest existing cluster centroid. The algorithm terminates once all locations have been allocated to a cluster. The major advantage of this approach is its computational efficiency. All locations can be clustered in a single pass. The approach was used by Isaacman et al. to cluster cell phone towers based on a large number of cellphone call records. While its simplicity may be useful for very large scale applications, the agglomeration by number of visits approach is not designed to leverage the origin-destination information available from smart card transactions.
3.3 Activity Sequence Inference

Once the stops and stations visited by a user are grouped into geographical areas aligned with activities, the user’s activity pattern can be reconstructed based on the information provided by consecutive pairs of journeys. Unlike other forms of digital traces such as cellphone records or GPS records, smart card transactions provide location information explicitly tied to the travel of an individual. Hence, the origins and destinations of a user’s journeys provide information about the user’s location beyond the exact time at which the user was observed at a stop or station. Relying on this information, this section introduces a methodology for inferring the activity location of a user over time given observations about her public transport journeys. Activity location is represented using the previously defined user-areas.

Each smart card transaction represents either the start or the end of a journey. The activity sequence inference focuses on assigning a discrete activity status to intervals delimited by all transactions. To do so, the journeys of each individual are ordered by time and each journey is considered sequentially to determine the activity status assigned to the intervals it delimits. For each journey $i$, the destination and origin of neighboring journeys $i - 1$ and $i + 1$, respectively, is used to infer the activity status as described below.

1. If the current journey $i$ started on the same day as journey $i - 1$ directly preceding it, or on the day directly following the day of $i - 1$, establish the user’s activity status from the end time of the preceding journey to the start of the current journey by comparing the destination cluster of journey $i - 1$ and the origin cluster of journey $i$. If the destination of $i - 1$ and the origin of $i$ are the same, infer it to be the activity status. If they are not the same, the user used another mode of transport during the interval considered and the location cannot be inferred.

2. If the current journey $i$ started on a day later than the day directly following journey $i - 1$, or if journey $i$ is the first journey made by the user, infer the location of the user from the start of the day on which the current journey was made to the start of the journey based on the origin of the journey. If the origin of the current journey is unknown, location cannot be inferred.

3. If the current journey ended on the same day as journey $i + 1$, or on the day directly preceding journey $i + 1$, the location from the end of journey $i$ to the start of journey $i + 1$ can be inferred as explained in 1 in the next step, when journey $i + 1$ is considered as the current journey.

4. If the current journey ended on a day earlier than the day directly preceding journey $i + 1$, or if journey $i$ is the last journey made by the user, the location from the end of journey $i$ to the end of the day can be inferred based on the destination of journey $i$. If the destination of $i$ is unknown, the location of the individual cannot be inferred for the rest of the day.
The approach is implemented by framing the current journey $i$ in a window including the previous and following journeys $i - 1$ and $i + 1$. At each step, this window rolls over all journeys, until all journeys have been considered. The algorithm terminates with a single pass through all user transactions.

The value of each possible activity status is summarized in the Table 3.1. The first three statuses distinguish between three different situations in which a user’s area cannot be inferred. Status -2 indicates that the user was on a public transport journey. Status -1 indicates that the origin of a journey $i$ did not correspond to the destination of the previous journey $i - 1$. In other words, status -1 denotes all intervals during which the user traveled between areas through a non-public transport mode are. Status 0 denotes intervals for which the absence of journeys, or missing origin or destination information prevents the user location from being inferred.

<table>
<thead>
<tr>
<th>Activity Status</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>User is in public transit journey</td>
</tr>
<tr>
<td>-1</td>
<td>User Location cannot be inferred because a non-public transit trip was completed during the interval</td>
</tr>
<tr>
<td>0</td>
<td>User Location cannot be inferred because origin or destination are not known</td>
</tr>
<tr>
<td>1</td>
<td>User is located at Area 1</td>
</tr>
<tr>
<td>2</td>
<td>User is located at Area 2</td>
</tr>
<tr>
<td>...</td>
<td>User is located at Area ...</td>
</tr>
</tbody>
</table>

Table 3.1: Activity Status Summary

Statuses above 0 indicate intervals for which the user-area was successfully inferred. Once the user’s activity status is inferred over the period of analysis, geographical areas are ordered with respect to the amount of time spent in each area. Hence, area 1, always aligns with the area in which the user was inferred to spend most time. Alternatively, areas could be ordered with respect to other criteria, such as the number of journeys started or ended in the cluster.

The inference process assumes that the actual location of users activity is in proximity of the stops or stations used to access the activity. This assumption may be inaccurate for users who use a combination of modes to access activities. For instance, park-and-ride users accessing public transit stations by car would inaccurately be inferred to be located near the station they use to commute. Activities in low transit-accessibility areas are most likely to be affected by such bias. Another limitation to acknowledge is that non-public transit journeys can only be inferred if they result in a discontinuity in the user’s public transit journey destination-origin sequence. Hence, return journeys to and from a given area would not be identified. For example, a journey to lunch and back completed by bike would not be captured.
3.4 Application

In this section, we apply the geographical location clustering and the activity sequence inference described in sections 3.2 and 3.3 to a sample of smart card transactions from London, UK. The process is used to construct the travel sequence of each user from smart card data. The resulting travel sequences constitute the input to the analysis presented in the remainder of this thesis.

3.4.1 Data

For the purpose of this research, the smart card records of a random sample of 99,925 TfL passengers is used. The sample contains records of all 3,160,276 bus, rail, and tram journey stages completed by these individuals between February 10th and March 10th 2014. The sample of cards is randomly selected from all non-staff cards with at least one travel transaction during this period. Rail records contain the origin and destination station and time of each rail stage, while bus and tram stages only include boarding time and location. Hence, the data is processed to infer the alighting time and location of bus and tram stages according to the methodology developed by Gordon (2012). His approach uses information about vehicle location and subsequent stage boarding time and location to infer alighting time and location. Rail, bus and tram stages are then linked together into inter-modal journeys to identify the origin and destination of users’ full journeys. Also in line with Gordon’s approach, stages are linked into journeys by comparing the alighting and boarding time and location of subsequent stages, as well as the direction of each stage. Table 3.2 summarizes the distribution of the number of stages inferred per journey. Over 97% of all inferred journeys are composed of 2 stages or less, and fewer than 0.5% are composed of 4 or more stages.

<table>
<thead>
<tr>
<th>Nbr of Stages</th>
<th>Nbr of Jnys</th>
<th>Prop. Jnys (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2112611</td>
<td>81.37</td>
</tr>
<tr>
<td>2</td>
<td>413783</td>
<td>15.94</td>
</tr>
<tr>
<td>3</td>
<td>60740</td>
<td>2.34</td>
</tr>
<tr>
<td>4</td>
<td>7751</td>
<td>0.30</td>
</tr>
<tr>
<td>5</td>
<td>1104</td>
<td>0.04</td>
</tr>
<tr>
<td>6</td>
<td>172</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>37</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 3.3 shows the proportion of observed journeys missing different elements of information. As shown, the boarding and alighting time and location are inferred for 83% of all journeys. The alighting location and time of 12.2% of journeys could not
Table 3.3: Missing Journey Information

<table>
<thead>
<tr>
<th>Origin</th>
<th>Destination</th>
<th>Boarding Time</th>
<th>Alighting Time</th>
<th>Nbr of Jnys</th>
<th>Prop. Jnys (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2157415</td>
<td>83.1</td>
</tr>
<tr>
<td>✓</td>
<td>?</td>
<td>✓</td>
<td>✓</td>
<td>85053</td>
<td>3.3</td>
</tr>
<tr>
<td>✓</td>
<td>?</td>
<td>✓</td>
<td>?</td>
<td>316058</td>
<td>12.2</td>
</tr>
<tr>
<td>?</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>8587</td>
<td>0.3</td>
</tr>
<tr>
<td>?</td>
<td>✓</td>
<td>?</td>
<td>✓</td>
<td>14300</td>
<td>0.6</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
<td>✓</td>
<td>✓</td>
<td>2543</td>
<td>0.1</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
<td>✓</td>
<td>?</td>
<td>12250</td>
<td>0.5</td>
</tr>
</tbody>
</table>

be inferred. Approximately 5% of journeys are lacking a combination of the other information components. Note that this table includes multi-stage journeys, and that intermediary origin and destinations are not included.

Figure 3-4 summarizes the distribution of the number of linked journeys observed per individual. As shown in the figure, a large proportion of users sampled performed 10 or less journeys during the analysis period. The distribution has a heavy right tail, with over 15% of users having completed 50 or more journeys. The mode of the distribution is 2 journeys per user. Overall, the 99,925 users sampled completed 2.6 million linked journeys between February and March 10th 2014.

Figure 3-4: Distribution of Number of Journeys per User
3.4.2 Geographical User-areas

Having identified linked journeys, the origins and destinations visited by a user can be analyzed to identify areas of London which users are assumed to visits for the same activity. As described in section 3.2, every stop and station, referred to as location, visited by a user as origin or destination, can be clustered based on their geographical distribution and on the journeys connecting them. Figure 3-5 summarizes the distribution of number of distinct locations visited per user. On average, users visited 15.4 distinct locations over the period of analysis, with the mode number of distinct stations visited equal to 2. The user having visited the most number of distinct locations was observed at 213 different stops or stations. 2888 users were observed to have visited only 1 location. These users performed few journeys, of which only one origin or destination was observed. Additionally, 59 users for whom no origin or destination was observed have 0 locations. These individuals are excluded from the remainder of the user location analysis.

![Figure 3-5: Distribution of Number of Distinct Stations Visited by User](image)

The geographical clustering approach implemented for the application relies on two parameters: $D$ the maximum distance between the farthest locations contained in the same user-area, and $\tau$ the maximum percentage of a user’s journeys observed between two given locations grouped in the same area. In order to determine appropriate value of these parameters, their effect is evaluated through sensitivity analysis. For each user, a different set of areas is defined for all 12 combinations of $\tau = \{1\%, 10\%, 50\%, 100\%\}$ and $D = \{500\text{m}, 1000\text{m}, 1500\text{m}\}$. Note that for $\tau = 100\%$, the OD flow constraint approach is equivalent to the simple complete distance approach.
Figure 3-6: Distribution of User-Area Diameter

Figure 3-6 shows the distribution of the user-area diameter, computed as the distance between the two farthest points contained in the same area. Areas including a single location are not included in the distribution. The color of each curve indicates the diameter threshold $D$, while values of $\tau$ are symbolized as indicated in the legend. As expected, the effect of $D$ on area diameter is larger than the effect of $\tau$. As $D$ decreases, the number of clusters with short diameters increases, and the distribution is shifted upwards. The effect of $\tau$ increases with increasing values of diameter threshold $D$ and is largest for areas of small diameter. For $D = 1500$ meters, the number of clusters smaller or equal to 50 meters ranges from 40,000 to 30,000 for $\tau = 1\%$ to $\tau = 100\%$. This difference is much less marked for $D = 500$ meters, with the number of clusters smaller or equal to $50m$ ranging from 69,000 to 66,000. Also for $D = 500$ meters, the curves associated with $\tau \geq 10\%$ are overlapping. This result suggests that accounting for the number of journeys between locations becomes increasingly important as the maximum cluster diameter is increased.

A journey that started and ended in the same area may reflect two different issues. On one hand, it may indicate that a geographical area is too large, and that it includes locations aligned with two different user activities. On the other hand, some observed journeys may in reality consist of stages of a larger journey uninferred in the journey inference process. For example, consider a user who sometimes walks and sometimes catches a bus on her way to the rail station nearest to her home location. If the bus and rail stages failed to be linked, the transfer location would be misinterpreted as a location distinct from the individual’s home location. Occasions on which the individual arrives home from the station by walking and next leaves by bus would be
Figure 3-7: Journeys with O & D in same User-Area

Inferred as a location discontinuity. As such, the appropriate value of $\tau$ involves a trade-off between the number of journeys included in the same cluster and the number of location discontinuities falsely inferred.

Figure 3-7 summarizes the effect of $D$ and $\tau$ on the number of journeys started and ended in the same cluster. As expected, the number of such journeys increases with respect to both the maximum cluster diameter and the OD flow constraint. For $\tau = 100\%$, increasing the value of $D$ from 500m to 1500m results in approximately a 9 fold increase in the number of journeys started and ended in the same area. The effect of the OD flow constraint is strongest for low values of $\tau$ and tapers off as $\tau$ increases. The number of journeys started and ended in the same cluster converges to 0 for all values $D$ as $\tau$ approaches 0.

Figure 3-8 summarizes the impact of $D$ and $\tau$ on the number of individuals with at least one journey with known origin and destination for whom all observed locations where assigned to the same area. These users are those who use public transit for distances below the specific threshold $D$. For values of $\tau$ up to 10%, no individual has all locations assigned to the same area. As expected, for values of $\tau$ beyond 10%, increasing area diameter thresholds result in an increase in the number of users with a single area. For $D = 1500$ meters and $\tau = 100\%$ a single area is defined for over 1% of all users sampled.

In line with the effects described in Figures 3-6, 3-7, and 3-8, the values of $\tau = 10\%$ and $D = 1000m$ where selected for this application. The resulting user-areas is first illustrated with respect to an example passenger. Figure 3-9 shows all locations...
visited by a passenger, excluding transfer stops and stations. Geographical areas are differentiated by color. For example, the first area, represented in red, contains 6 different stops or stations. This area also contains the largest number of distinct locations. With respect to size, the user’s largest cluster, Area 2, has a diameter of 787 meters, measured as the distance between the two furthest locations within the area. The smallest cluster, Area 4, contains a single location.

A similar set of areas is defined independently for each user, such that all distinct user-locations are allocated to a user-area. This results in 714,528 user-areas, with an average of 6.8 areas per user and 2.3 locations per user-area. Figure 3-10 shows
**Figure 3-10:** Distribution of Number of Areas per User

**Figure 3-11:** Distribution of Number of Locations per User-Area
the distribution of number of areas per user. From this distribution, over 80% of users having 10 areas or less. The most frequent number of areas per user is 2. 3,171 users have a single area. Figure 3-11 shows the distribution of number of locations per user-areas. The distribution reveals that 43% of areas are constituted of a single location, while 30% contain 2 locations.

3.4.3 Activity Sequence Inference

The geographical user-areas allow for users’ location to be inferred over time. As described in section 3.3, origins and destinations of subsequent journey pairs can be compared to determine the location of an individual. In order to illustrate the activity sequence inference process, we first discuss it with respect to the example individual presented in Figure 3-9. Figure 3-12 presents the inferred activity status for this individual across the period of analysis. Each column along the x-axis is associated with a day. Time of day is indicated on the y-axis, with 12:00 AM at the top of the figure. The figure can be interpreted as the inferred activity location diary of the user. The user’s areas, mapped in Figure 3-9, are represented by the colors indexed on the right-hand side of the figure. As summarized in Table 3.1, the individual’s location during any given interval is symbolized by different statuses. Statuses below 1 represent non-area statuses. These include: public transit journeys (-2) represented in black, failure to infer location because a non-public transit trip was completed during the interval (-1) represented in dark grey, and failure to infer location because of missing journey origins and destinations (0) represented in light grey. Statuses above 0 represent specific geographical areas corresponding to clusters of stops and stations. User areas are indexed with respect to the total amount of time the individual was inferred to spend in each area. Area 1 always corresponds to the area in which the user was inferred to spend the most time over the 4-week period, and similarly for all subsequent areas.

![Figure 3-12: Activity Sequence Inference of Example User](image-url)
For example, this user completed a journey from Area 1 (in red) to Area 2 (in green) around 9:00 on Monday February 10, and another journey from Area 2 to Area 1 in the evening just before 20:00. From the first journey, the user is inferred to have been in Area 1 between 0:00 and 9:00. Then, as the destination of the first journey matched the origin of the second, she is also inferred to have stayed in Area 2 between 9:00 AM and 20:00 PM. From this user’s pattern, it appears that the first area aligns with a home location, and the second area aligns with a work location. Days on which no journeys are observed (e.g. Sunday February 23rd) are assigned status 0. Similarly, intervals bounded by unknown origins and destinations are also assigned status 0. For example, the last interval of Saturday February 22nd is bounded by an unknown destination and no journeys on the following day. Hence, the interval is assigned a status of 0. Intervals bounded by different origins and destinations are assigned status -1. For example, on Sunday March 9th, the user was seen alighting at 10:00 in an area different from the area where she was next seen boarding at 18:00. Hence the interval between 10:00 and 18:00 is assigned a value of -1. For instance, this could indicate that the user biked, walked or took a taxi between the two journeys.

Table 3.4: Distribution of Time Inferred Across Activity Status

<table>
<thead>
<tr>
<th>Status</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>≥5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>4.2</td>
<td>10.9</td>
<td>4.5</td>
<td>49.9</td>
<td>18.8</td>
<td>5.7</td>
<td>2.6</td>
<td>3.4</td>
</tr>
<tr>
<td>Average Time Spent (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This inference process is applied to the transactions of every user. The output is a sequence of intervals, characterized by an activity status, a start-time and an end-time, representing the activity pattern of each individual. Excluding days on which no journeys were observed, over 5.5 million intervals are inferred for the 99,925 users sampled. Table 3.4 summarizes the distribution of time inferred across activity statuses. Each value indicates the average percentage of time spent for a given status across all individuals. Intervals corresponding to days with no observed journeys are excluded. For example, weighting each time interval by its duration, status 1 accounts for 49.9% of time periods, while non-area statuses account for 20% of time periods.

In line with the definition of travel sequence previously introduced, each interval can be thought of as an activity. It is useful to analyze the duration of intervals associated with different statuses to gain insight on the nature of the activities associated with each status. Figure 3-13 shows the distribution of activity duration for 3 different status -1, 1 and 2. The x-axis details activity length, measured in hours. The y-axis corresponds to number of activities of a given duration across all users. Each curve shows the distribution of a different status. Status 1 is characterized by three modes. The highest mode, around 14 hours, is likely associated with spending the night at home, from 18:00 to 8:00 the following day for example. The second mode, around 8 hours is likely aligned with the 8 hours working day. The third mode,
much lower in frequency, is likely associated with users making a single journey on a given day, followed by another journey at a similar time the next day. Status 2 is characterized by a dominant mode also around 8 to 9 hours daily. This is likely reflects the fact that the secondary area corresponds predominantly with the area in which users work. All three statuses have a high-number of activities shorter than 1 hour, but short activities make up a relatively larger proportion of intervals of -1 statuses. Additionally, the proportion of activities approximately 24 hours in duration is much higher for -1 intervals. This may be explained in part by users who use public transit daily for the same unique journey. For example, some commuters may travel to work in the morning using public transit and return home in the evening through some other mode. In this case, the origin of a journey on a given day would not match the destination of the journey completed on the previous day. The resulting interval would be inferred with status -1. Overall, these results reveal that each one of these three statuses are associated with a distinct mixture of activities. The curves also appear to support that user’s home areas are primarily associated with Area 1 and that work areas are primarily associated with Area 2.

Table 3.5 summarizes the percentage of interval types. As illustrated in Figure 3-14, we categorize intervals into 4 groups. Intervals referred to as before the first journey include all intervals preceding the first journey of day, on days for which no journey was observed on the previous day. Intervals qualified as between journeys include all intervals between two journeys performed on the same day or on two consecutive
days. After Last journey includes all intervals following the last journey of the day made on days for which no journey was observed on the following day. Journey intervals include all time periods between a journey’s origin (tap-in) and destination (tap-out). As seen in the table, the majority of intervals are either journey intervals or between journeys intervals for which the destination of journey \( i - 1 \) and the origin of journey \( i \) are both known. 3.6% of intervals were between journeys interval for which the destination of journey \( i - 1 \) was observed in a different user-area than the origin of the following journey \( i \). Additionally, between journeys intervals bounded by a destination \( D_{i-1} \) which could not be inferred because the subsequent origin \( O_i \) was too far away from all stops on the bus line used are also assigned status -1.

Table 3.5: Inference by Interval Categories

<table>
<thead>
<tr>
<th>Interval Type</th>
<th>( D_{i-1} )</th>
<th>( O_i )</th>
<th>( D_i )</th>
<th>( D_{i-1} = O_i )</th>
<th>Status</th>
<th>Count</th>
<th>Prop. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before 1st Jny</td>
<td>-</td>
<td>?</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>7209</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>1+</td>
<td>344924</td>
<td>6.2</td>
</tr>
<tr>
<td>After Last Jny</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>0</td>
<td>95205</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>1+</td>
<td>256928</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>N</td>
<td>-1</td>
<td>197495</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>Y</td>
<td>1+</td>
<td>1719125</td>
<td>31.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>?</td>
<td>-</td>
<td></td>
<td>1+</td>
<td>6754</td>
<td>0.1</td>
</tr>
<tr>
<td>Between Jnys</td>
<td>?</td>
<td>✓</td>
<td>-</td>
<td>N(^a)</td>
<td>-1</td>
<td>62003</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>1+</td>
<td>234979</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>?</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>23717</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>✓</td>
<td>?</td>
<td>-</td>
<td>-2</td>
<td>328308</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-2</td>
<td>2253598</td>
<td>40.6</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>?</td>
<td>✓</td>
<td>-</td>
<td>-2</td>
<td>14300</td>
<td>0.3</td>
</tr>
</tbody>
</table>

\(^a\) Indicates intervals for which \( D_{i-1} \) could not be inferred because it was too far from \( O_i \)
3.5 Conclusion

This chapter introduced a conceptual representation of individual travel behavior based on the ordered sequence of travel events and a methodology to implement this representation using public transit fare transactions. The methodology presented is divided in two steps. First, we recognize that users may use several different stops and stations to access the same activity location. Hence, geographical clustering approaches used to identify clusters of stops and stations aligned with the activities of each user are discussed. Second, a methodology to infer the location sequence of each individual is discussed. This methodology is then applied to reconstruct the travel sequence of a sample of 99,925 users from London. The application demonstrates that the methodology described can be implemented efficiently to build a representation of each user’s travel behavior which will be analyzed in more depth in the remainder of this thesis. Preliminary analysis of user travel sequences reveal an interesting relationship between activity duration and user-areas identified for each individual.

While these results indicate that the inferred travel-sequences can adequately be used for the analysis of travel patterns, a number of potential methodological improvements may be the focus of future research. The geographical clustering approach currently implemented could be improved in two ways. First, in addition to geographical distance and origin-destination flows, network structure can potentially inform the definition of user-areas. For example, the maximum area diameter $D$ could be increased in parts of the city with low stop and station density. Second, the definition of user-areas could also be informed by the origins and destinations of consecutive journeys. For example, a stop used as the destination of journey $i - 1$ may be more likely to be associated with the same activity as the stop used for the origin of the next journey $i$. In other words, user-areas could be defined to minimize the number of time intervals in which discontinuities are observed, subject to an area diameter constraint. More generally, while the user-area definition and the location inference are implemented consecutively in the current approach, feedback between the two steps may allow for more accurate reconstruction of activity patterns.

The label assigned to each user-area is currently determined by the cumulative time inferred in each area over the period of analysis. Activity purpose associated with each area is not explicitly inferred through this approach. However, the activity duration distribution presented in Figure 3-13 suggest that different user-areas are associated with distinct mixtures of activity purpose. Hence, an interesting question for future research could focus on labeling user-areas according to explicitly inferred activity purpose. For example, land use data and point of interest (POI) data could provide useful information. Furthermore, ground truth data about the home and work location of users could be collected through an online survey and used to train a model predicting which user-area is associated with a person’s home and work location.

A number of limitations of the current location inference approach were described in section 3.3. Future research may focus on investigating the validity of the underlying
assumptions. Using other types of location traces, such as GPS location data or cell-phone data, could allow for validation of the results inferred here. Additionally, travel records from other modes, such as public bike sharing or taxi, could be integrated with smart card transactions to develop a activity sequence inference methodology based on multimodal travel data.
Chapter 4

Quantifying Regularity in Travel Sequences

4.1 Introduction

Chapter 3 presents an approach to reconstruct individual travel-activity sequences over multiple days. These sequences allow for the valuable analysis of longitudinal travel behavior. While conventional cross-sectional data, for example daily travel diary surveys, capture variability in behavior among multiple individuals, they do not capture variations in the behavior of a given user over multiple days or weeks. In contrast, multiday travel sequences derived from smart card data allow for detailed analysis of variability within the travel pattern of a single user. Such longitudinal variability, also referred to as intrapersonal (Pas and Koppelman, 1986; Schönfelder, 2006), intra-individual (Huff and Hanson, 1986), or day-to-day variability (Kang and Scott, 2010; Kitamura et al., 2006; Neutens et al., 2012) has been the focus of numerous studies over the last three decades. This section will review notions of longitudinal variability introduced by these studies in order to define the concept of regularity in travel patterns. Relying on the definition, this chapter will then introduce an approach to represent and quantify regularity in individual travel sequences.

Intrapersonal variability describes variability in the characteristics of the travel and activities made by the same individual across multiple days (Hanson and Huff, 1982; Pas and Koppelman, 1986). Consider an individual’s work related travel over several days. On a given day, she leaves for work at 8:17 AM by car via route 1. On the following day, she leaves slightly later at 8:25 AM and uses a different route 2. On some other days she uses public transit, some times from the bus stop nearest her house, sometimes from the metro station half a kilometer away. On days she uses public transit she leaves earlier around 7:00 AM. On some days she only makes 2 trips, to work and back, while on some other days she completes additional journeys for business meetings, or to drop her children at the daycare. Intrapersonal variability relates to such variations in the nature of her travel and occurs with respect to
different attributes of trips and activities (e.g. journey time, mode choice, or path choice). Dual to the fluctuations in travel attributes is the extent to which components of her behavior are repeated over time. Indeed, the individual’s activity choices, and their associated trips are not made randomly. As described by activity based travel theory, they are dictated by preferences, constraints and needs which recur over time to some degree.

Most discussion on day-to-day variability in the literature hinges on the definition of the unit of analysis for which such repetition and variation are being considered. Earliest studies focused on longitudinal variability, the seminal 35-day Uppsala travel survey studies (Hanson and Huff, 1982, 1988; Huff and Hanson, 1986) first explicitly described the choice of analysis unit. As they point out, travel behavior repeats at different levels of aggregation (e.g. trip level & activity level, daily, or weekly) and with respect to different attributes of travel (e.g. daily journey frequency, first journey departure time, mode choice, activity type). For example the number of daily journeys completed by an individual might be the same every day, but the destination or time of these journeys might change from day to day. Hence, the unit of what is being repeated (or of what varies) must be chosen in line with the aggregation level and attributes relevant to the research question considered. Huff and Hanson (1986) use the term behaviors to describe components of travel behavior characterized by combinations of attributes (e.g. activity purpose and location, or activity purpose, trip mode and time of day), for example “driving a car to work”. These behaviors are used to flexibly define a unit of analysis. Using this approach, they found that whereas individual behaviors are frequently repeated over time, the patterns in which they are observed change from day to day. In other words, their results demonstrate that an individual’s travel is poorly represented by a typical day, because the organization of trips and activities tends to fluctuate over time. This highlights that intrapersonal variability must be considered not only with respect to variability in the characteristic of a single behavior, but also with respect to the pattern in which multiple behaviors are observed.

Huff and Hanson’s behaviors align with the concept of travel events introduced in chapter 3 to discuss travel sequences. As we described, travel sequences are composed of events, characterized by attributes, which occur in an ordered series over time. The attributes of events can be considered at different levels of abstraction to define flexible units of analysis. While the term behaviors is seldom used, the same concept underlies many subsequent studies analyzing repetition and variability with respect to different aspects of travel, from the number of journeys made in a day, to first journey departure time in day, to activity duration, location and scheduling, to route and mode choice (Bhat, 2000; Buliung et al., 2008; Chikaraishi et al., 2009; Habib and Miller, 2008; Huff and Hanson, 1986; Kitamura et al., 2006; Morency et al., 2007; Pas and Sundar, 1995; Roorda and Ruiz, 2008; Stopher et al., 2008). In line with the terminology previously introduced in this thesis, we use travel events to refer to Huff and Hanson’s concept of behaviors.

While the concept of intrapersonal variability is recognized as an important dimension
of travel behavior, approaches to measure such variability remain limited in scope. Specifically, many studies measure intrapersonal variability based only on the extent to which single travel events are repeated, without consideration for how multiple events are combined. The most simplistic approaches focus on the relative frequency of events. For example, Buliung et al. (2008) propose a spatial repetition index corresponding to the percentage of activity locations which are visited more than once over a 7 day period. This measure is computed for different time periods to evaluate the spatial stability of individual’s activity-patterns at different times of the week for data collected from a 7-day activity survey in Toronto, Canada. Also relying on journey frequencies, Kieu et al. (2014) categorized origin-destination pairs of Brisbane transit users as regular or irregular based on a fixed threshold of the number of visits to origin and destinations stops in close proximity of each other. They also categorize habitual journey times using univariate clustering to identify weekday times at which over a certain number of a user’s journeys are observed. Using these categories, they measure intrapersonal variability based on the percentage of a user’s journeys completed at habitual times and on regular OD pairs. Similarly, Morency et al. (2007) evaluate the level of spatial and temporal variability of different smart card users in Gatineau, Canada, based on the frequency of journeys made to different stops at different times of day.

Other studies rely on the variance of continuous measures to quantify longitudinal variability. Pas (1987) and Pas and Koppelman (1986) evaluated the variance in number of trips per day across for a 7-day travel survey conducted in Reading, England. Their results contrasted the variance in trip generation rates associated with intrapersonal variability and interpersonal variability. Also using variance, Kitamura et al. (2006) analyzed variability in first journey departure time. Relying on the concept of individual space-time prisms, they modeled the variance of first journey time so as to differentiate variance in departure time due to randomness from variance due changes in the time constraints dictating an individual’s schedule. Chikaraishi et al. (2009) also attempted to dissect variance in first trip departure time by formulating a multilevel model for which variance in departure time is broken-down into five parts: inter-individual variation, inter-household variation, spatial-variation, temporal-variation, and intra-individual variation. All these studies focus on measuring intrapersonal variability either based on some measures of travel event frequency or variance in the characteristics of travel events. These measures are not concerned with the combinations of multiple travel events.

Another class of studies focuses on modeling travel behavior over multiple days. In doing so, they account for variability in combinations of travel events through different approaches. Some models rely on the assumption that activity and trip combinations are primarily a function of difference in days of the week. For example, using the 7-day Toronto Travel Activity Panel Survey, Habib and Miller (2008) modeled the frequency of 15 non-home/work activity categories for the 7 days of the week using 7 independent models. In contrast, some studies model the relationship between different travel events more explicitly. Roorda and Ruiz (2008) modeled preplanned and spontaneous activity duration as well as number of trips by mode using structural
equations modelling techniques on the 7-day activity survey used by Habib and Miller (2008). Their approach introduces same-day effects and next-day effects to capture the relationship between multiple activities. From a long-term perspective, (Bhat et al., 2005) examined the relationship between successive activities of the same purpose (e.g. shopping) using a 6-week travel survey from Karlsruhe, Germany. They modeled the time elapsed between successive activities using a multivariate hazard model. These studies all account for the relationship between travel events to various degrees in order to improve travel demand models. None, however, explicitly aims at measuring the variability in how travel events are combined.

Variability in the organization of travel events is rarely discussed in the literature on longitudinal variability. Specifically, while there exist indices of repetition in specific travel events (e.g. frequency or variance), no index capturing repetition in the order in which events are observed has been defined. Nevertheless, the order in which an individual completes journeys and activities is an integral component of the structure in his travel routine. To illustrate the importance of order consider the two example individuals illustrated in Figure 4-1. In this figure, the activity location sequence of two fictional individuals is illustrated over 7 days. The three locations visited by these individuals are represented by three different colors and the time spent at each location is indicated by the width of the bar. Both users visit each location exactly 7 times, always for the same duration. Measures of repetition based on the frequency of visits to each location, or on the variance of activity duration would describe these two sequences identically. However, the behavior of individual 1 is more regular than the behavior of individual 2. The component which differentiates the intrapersonal variability of these sequences is the order in which locations are visited. The first individual always visits the red location first, followed by the green then the blue. Individual 2 visits all three locations in varying order.

Figure 4-2 illustrates the importance of event order with respect to first journey time. In this case, events are defined on a daily basis. The sequences summarize the first journey time across 24 days for both individuals. Days on which the individual did not
travel are blank. Both individuals traveled on exactly 16 days. For each user, exactly 4 days were started at 8:00, 10:00, 14:00, and 16:00. Hence, the start time variance and the travel frequency of both individuals are the same. However, the intrapersonal variation of the first user is obviously different from that of the second. As for the previous example, the order in which events are observed drives this difference. The first individual’s travel is organized according to a regular pattern which repeats over the 24 days considered. In contrast, days traveled by the second individual are scattered through the 24 days and do not follow a particular pattern.

These two examples illustrate that the order in which events occur is an important component of intrapersonal variability and outline the value of variability measures sensitive to order. Many such measures explored in the human mobility literature are tied to concepts of periodicity. Eagle and Pentland (2006); Kim and Kotz (2006) were among the first studies to apply periodicity to individual mobility patterns. Both studies used the Fourier transform to identify underlying periods of repetition in travel from digital traces of location collected over multiple weeks. Kim and Kotz found daily and weekly periods to be most significant in observing individuals’ connection to WIFI access points (AP) on the Darmouth campus. Using data from MIT’s Reality Mining Experiment, Eagle and Pentland identify the same dominant periods from subject’s GPS location. Moving beyond the Fourier transform but also focusing on cyclical repetition, Williams et al. (2012) propose a measure of temporal irregularity in the intervals between a person’s visits to a given location. Their measure, based on the coefficient of variation of the interval between visits, provides an index of periodicity.
for a predetermined period of interest. For example, they apply the measure for a weekly period to smart card data, Foursquare check-in data, and to the Dartmouth WiFi AP data. They found the behavior captured from WiFi AP data to be most irregular, and the behavior captured from smart card travel data to be most regular. Song et al. (2010) present another regularity measure also based on a weekly cycle. Given hourly information of a person’s location over several weeks, they define an index of periodicity for each hour of the week, defined as the percentage of hours spent at the location most frequently visited during the corresponding hour of the week. At a more general level, Li et al. (2012) propose a probabilistic measure of periodicity in event sequences. Using GPS location records, they demonstrate that their approach is more suitable to analyze periodicity in human travel than Fourier transform measures because of its robustness to noise and missing observations. All these studies account for fluctuations or repetition in the order of events by measuring the extent to which events occurrence can be mapped to a set cycle (most often a weekly cycle).

The authors of these studies often equate periodicity and regularity, but there exists an important distinction between the two concepts. Indeed, while regularity can be associated with different forms of repetition, periodicity focuses solely on cyclical repetition. To some extent, periodicity can be considered as a special type of regularity. Regular behavior may be characterized by repetition which is not observed in a cycle. For example, in individual 1’s start time sequence (Figure 4-2), the number of days elapsed between each 4-day pattern varies. Hence, the overall pattern is composed of repeated sub-sequences which do not align with a particular cycle. Repetition occurs regardless of the shift between sub-sequences. This is especially relevant to sequences of activities, as activities are likely to be organized in a logical order. For example, visiting the doctor’s, going to the pharmacy to pick-up a prescription, and returning home is likely to occur in this logical order. Yet, the repetition of this coherent sequence is unlikely to be periodical. Visits to the doctor can be scheduled on different days of the week, and at different times of the day. What’s more, the application of periodicity measures is tightly related to a conventional calendar definition. Song et al. (2010), Williams et al. (2012), Kim and Kotz (2006), and Eagle and Pentland (2006) all discuss periodicity in the context of the most conventional cycles of repetition: the day and the week. However, regularity is an internal property of a travel sequence and should not depend on how the sequence aligns with the calendar. Some repeated patterns may repeat on non-daily or weekly cycles. For example, certain types of employment (e.g. shift-workers, firefighters, doctors) may dictate working schedules which repeat on a unit other than the week. Even though the schedule of such workers may be perfectly regular, periodicity computed on a weekly basis (as done by Song et al. and Williams et al.) would fail to capture the true level of regularity in the sequence. Certain patterns may also span more than a single day, for instance, going out in the evening, sleeping at a friend’s, then returning home the next day. The repetition of such pattern would not be captured by a measure of daily periodicity. These examples illustrate that travel patterns may not necessarily repeat periodically, or may repeat over unconventional periods, not aligned with the typical
day or week.

All the examples given above illustrate that order is an integral dimension of travel patterns. Hence, a measure of travel sequence regularity should aim at quantifying the extent of repetition in both individual events, and the order of these events. This chapter will focus on defining such a measure. To do so, a mathematical representation of the travel sequence introduced in chapter 3 will be presented in Section 4.2 and a measure of regularity based on entropy rate will be described in Section 4.3. Finally, Section 4.4 will present an example application of the regularity index on a sample of London transit users.

4.2 Sequence Representation

As introduced in chapter 3 individual travel patterns can be conceptualized as a sequence of travel events. These events unfold over time with respect to a background calendar (time of day, day of the week, month etc.). Each event is characterized by the combination of its attributes including activity purpose, location, time, duration, and trip route and mode (Figure 4-3). For example, focusing on activity sequence, each activity occurs at a certain time of day, for a certain duration, and is characterized by its purpose and location. Events can be defined in multiple ways and their dimensions can be abstracted according to the aspect of behavior considered. For instance, the sequence of locations visited by an individual could be considered without regards to the time spent at each location, or the sequence of days on which they traveled could be considered only with respect to the departure time of each day.

The introduction of this chapter argues that a key component of these sequences is the order in which events take place. An appropriate measure of regularity in a person’s travel behavior should quantify the extent of repetition in travel events and in the order in which they are performed. In order to define such a regularity index, it is necessary to develop a mathematical representation of travel sequences which captures the order of events. This section introduces a general sequence representation which can be used for various definitions of travel events. We model the mobility of each individual as a random process. This process dictates how often and in what order
travel events are generated. The notation presented follows that used by Gao et al. (2008).

Let the stochastic process corresponding to the mobility of a given individual $u$ be denoted by $X_u$ and a travel event generated by this process by random variable $X_u$. Each travel event $X_u$ can assume any discrete value $x$ contained in the set of possible travel event outcomes $E_u$ defined for individual $u$. $X_u$ has a discrete probability distribution $p(x) = Pr\{X_u = x\}$ for $x \in E_u$. For simplicity, subscript $u$ is omitted and all remaining notation is defined with respect to a single individual.

The stochastic process $X = \{..., X_{-1}, X_0, X_1, X_2, ...\}$ represents the ordered set of random variables $X_i$. Any finite sequence of this ordered set running between event $i$ and event $j$ is denoted by the ordered subset $X^j_i = \{X_i, X_{i+1}, ..., X_{j-1}, X_j\}$, with $-\infty < i \leq j < \infty$ such that $X^j_i \subset X$. Given a finite window of analysis, we observe a specific realization $x^j_i = \{x_i, x_{i+1}, ..., x_{j-1}, x_j\}$ of the finite random variable sequence $X^j_i$.

Informally, set $E$ can be thought of as an alphabet from which a string of discrete events can be constructed. Different types of sequences, or strings, can be represented based on different definitions of travel events $x \in E$. For example set $E$ could be defined as a set of geographical locations such that sequence $x^j_i$ represents the sequence of locations visited by individual $u$. Alternatively, set $E$ could represent a set of activity purposes, such that $x^j_i$ models an activity sequence. The definition of $E$ is driven by the aspects of behavior of interest and by the data used for the analysis. Different data provide information on different aspects of travel and at different aggregation levels. For instance, smart card data provides location information at the stop level, and no direct information on activity purpose. Next, we adapt this entire formulation by presenting 4 different definitions of travel events contained in set $E$. While these definitions are specifically tailored to smart card data, the general concepts they illustrate can be translated to other sources of data on individual travel.

### 4.2.1 Discrete Attribute Sequence

Perhaps the most obvious application of the formulation described above is for an alphabet of events defined based on discrete attributes of activities and journeys (i.e. purpose, location, or mode). For example, as a person engages in discrete activities throughout the city, she visits discrete locations in a sequence. This sequence can be represented by defining set $E$ as the set of all locations visited by the individual over the period of analysis. In this case, each location is assigned one of the discrete symbols in the alphabet of possible travel events, and $x^j_i$ represents the sequence of locations visited by the user.

Chapter 3 details a activity status inference approach for smart card data. In this context, user activity location is inferred between public transport journeys, based on individually defined user-areas. In order to account for unobserved journeys on other
transport modes, periods for which an individual’s activity location cannot be inferred are represented by non-area statuses. Table 4.1 summarizes the symbols assigned to different activity statuses according to the definition introduced in Chapter 3.

Table 4.1: Activity Status Summary

<table>
<thead>
<tr>
<th>Activity status</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>User is in public transit journey</td>
</tr>
<tr>
<td>-1</td>
<td>User activity location cannot be inferred because a non-public transit trip was completed during the interval</td>
</tr>
<tr>
<td>0</td>
<td>User activity location cannot be inferred because origin or destination location are not known</td>
</tr>
<tr>
<td>1</td>
<td>User is located at area 1</td>
</tr>
<tr>
<td>2</td>
<td>User is located at area 2</td>
</tr>
<tr>
<td>...</td>
<td>User is located at area ...</td>
</tr>
</tbody>
</table>

The level of detail captured by each non-area status can be adjusted. For example, activity status -2 which presents a public transit journey, could be qualified in more detail to account for variations in public transit route or mode (e.g. bus, tram, or rail). Section 4.4 will describe a detail application of the travel sequence for data from Transport for London. Other discrete activity attributes, such as purpose can also be modeled using this approach.

4.2.2 Compound Attribute Sequence

The previous sequence is composed of events based on a single discrete attribute. For certain research questions, it may be necessary to define events based on combinations of multiple attributes. For instance, the location sequence presented above abstracts the amount of time spent at each location. Event duration may provide useful information in certain contexts. Specifically, significant differences in activity duration may be an indication of differences in the nature of two activities observed in the same geographical area. For example, a user might use the same rail station to access work and to go shopping. However, shopping activities are likely to be shorter
than work activities. Incorporating duration in the event definition allows for such differences in activity characteristics to be captured.

The formulation introduced above describes travel sequences constituted of discrete events. In order to account for event duration within this formulation, duration, a continuous event characteristic, must be discretized. The activity duration distribution presented in Figure 4-5 is useful to identify a discretization approach. This distribution, derived from the location inference for the sample introduced in chapter 3, is a mixture of three distinct component distributions. The first component is associated with exponentially distributed activities of short duration. The second and third components align with two secondary modes, around 8 hours and 14 hours, respectively. The two longer duration components are likely associated with work and home activities. This distribution suggests that activity duration could be discretized to align with each component of the mixture by defining 3 activity duration categories. Then events \( x \) in set \( E \) can be defined as compound outcomes of location and duration. For example, events having occurred at a location \( l \) could be associated with three distinct symbols, \( l - \text{short} \), \( l - \text{medium} \), or \( l - \text{long} \) according to their duration (Figure 4-4).

The boundaries used to determine duration categories can be determined either at
the aggregate level, or on an individual basis. At the aggregate level, the boundaries can be defined by the aggregate activity duration distribution as seen in Figure 4-5. This aggregate approach results in consistent event definitions across individuals. A consistent boundary definition facilitates comparison of multiple individual sequences. At the individual level, a different set of boundaries is established for each person based on the individual’s activity duration distribution. For example, clustering along a single dimension could be used to identify groups of activities with similar duration. On one hand this approach provides a more customized event definition for each user, preventing a group of activities from being arbitrarily divided by aggregate boundaries and assigned to two different categories, even though they correspond to the same activity type for the specific individual (e.g. 4:57 and 5:03 for a 5:00 boundary). On the other hand, it makes for a more complex comparison of sequences across individuals.

A similar event definition can be used for combinations of attributes other than location and duration. For example, events could be defined based on different combinations of activity purpose, location, duration, and start time. This provides a flexible means of for integrating different attributes of travel events into the sequence formulation.

4.2.3 Calendar Events Sequence

As summarized in Figure 4-3, travel events unfold over time and overlay a background calendar. Both the location and the location-duration sequences abstract this calendar. However, in some contexts, it may be necessary to capture the relationship between a person’s travel events and objective calendar events. For instance, we may be interested in differentiating events which occur on different days, or during different periods of the day (e.g. peak vs off-peak). In order to integrate the background calendar into the travel sequence, we define an additional category of events, such that set $E$ includes both personal travel events, and objective calendar events.

![Figure 4-6: Calendar Events Sequence](image)

Figure 4-6 illustrates the calendar event approach with daily markers. For example, day events marking 12:00 AM are inserted between the last event started before 12:00 AM and the first event started after 12:00. Different times of the day can be marked by defining multiple calendar events, each assigned to a different symbol. Similarly, calendar events could be used to mark different days of the week, or to mark the start
and end of weekends. The calendar marker approach provides a flexible method to integrate personal events with objectively defined events.

4.2.4 Aggregated Events Sequence

The three previous sequences all focus on events defined with respect to individual activities or journeys. At a different level of temporal aggregation, multiple journeys or activities can be grouped together to define a single event. For example, travel events can be defined on a daily basis. Figure 4-2 illustrated the sequence of daily departure times for two individuals. These sequences can be modeled using the suggested formulation by grouping departure times into discrete categories. As for event duration, categories can be defined at the aggregate level or at the individual level. At the aggregate level, time thresholds are aligned to periods of interest. For example, weekday departure times could be assigned to five categories: early morning, morning peak, midday, afternoon peak, and evening. At the individual level, groups of departure times can be identified for each individual using a univariate clustering approach (e.g. DBSCAN), similar to the approach of Kieu et al. (2014).

More generally, daily travel sequence can be modeled based on different event definitions. Much as the activity compound events (e.g. location & purpose), daily events can be defined for combinations of daily attributes, for example, day type (week, week-end), number of journeys completed, and departure time.

4.3 Quantifying Regularity

As described in the previous section, we model the mobility of one individual as a sequence of events generated by a random process $X$. Through this approximation, it becomes possible to characterize the individual’s mobility by quantifying the nature of the random process $X$. A number of different properties of process $X$ may provide information about the individual’s travel. For example, consider a process $X$ representing the activity sequence of an individual. In this case, the cardinality of set $E$ informs us about the diversity of activities in which the individual engages, and the mode of probability distribution $p(x)$ reveals the individual’s most frequent activity. This section introduces such properties of $X$ which can be used to describe regularity, as defined above, in a travel sequence. As a reminder, regularity refers to the extent of repetition in the events of sequence and in the order in which these events are observed.

First, we examine the extent of repetition in the events, without consideration for the order in which they are generated. Consider two random processes generated by the travel of two different individuals, $a$ and $b$. Both processes have exactly 6 possible outcomes. For individual $a$, the 6 outcomes have uniform probability distribution $(1/6)$. For individual $b$, the first outcome occurs with a $5/6$ probability, while the
other 5 each have a probability of 1/30. On average, the first outcome of b will be repeated 5 times more frequently than any single outcome generated by a’s travel as indicated by their probability ratio. For instance, if a sequence of 12 events from each individual were considered, on average, the first outcome would account for 2 events a’s sequence and 10 events of b’s sequence. As a result, the outcome of a new event generated by the first process can be predicted with much lower accuracy than an outcome generated by the second process as the probability distribution of the second individual is dominated by one of the 6 outcomes. Generally, the second process is characterized by a higher degree of repetition and is therefore more predictable than the first. The level of randomness or unpredictability of a process can measured using entropy. Entropy measures the average information, or surprise, provided by each realization of a random variable in bits. In the case of our example, the average outcome of the first process is more ‘surprising’ than the average outcome of the second process, because the first outcome of b is more frequently repeated than each outcome in a. The entropy $H(X)$ of random variable $X$ with probability distribution $p(x) = Pr\{X = x\}$ for $x \in E$ is defined by Equation 4.1.

$$H(X) = -\sum_{x \in E} p(x) \log_2 p(x) \quad (4.1)$$

For the travel sequence formulation, $X$ represents the random variable associated with a travel event and $E$ denotes the set of all possible travel event outcomes defined for a given individual. Entropy can be thought of as a measure of variance defined for categorical probability distributions. It accounts for both the number of possible outcomes (the cardinality of alphabet $E$) and the relative frequency of outcomes. Hence, entropy equals 0 for a process with a single possible outcome (no uncertainty) and is highest when the probability distribution of a random variable with multiple outcomes is uniform (all events equally likely). Scheiner (2014) used entropy to measure and contrast the complexity of activity patterns completed by individuals of different sex. As the author point out, entropy is a good measure of the amount of heterogeneity in a categorical distribution, which is especially relevant when considering qualitative outcomes such as activities.

Because it is sensitive to both the number of possible events and to the probability of each event, entropy is a good measure of the extent to which events in the travel sequence are repeated. However, it does not capture repetition in the order in which events are observed. Random process $X = \{\ldots, X_{i-1}, X_0, X_1, X_2, \ldots\}$ is an ordered set of events, in which consecutive events are not independent of each other. It is therefore not a memoryless process. Rather, the conditional distribution of an event $X_i$ depends of the outcome of events $X_{i-1}, X_{i-2}, \ldots$ preceding it (i.e $p(X_i|X_{i-1}, X_{i-2}, \ldots) \neq p(X_i)$). In concrete terms, as introduced in a previous example, observing a visit to the doctor might significantly increase the likelihood of visiting the pharmacy in the following event. Such effects are not captured by the entropy measure which considers the probability $p(x)$ of each event independently. In order to account for the order of events in the travel sequence, or more formally for
the memory in process $X$, we rely on entropy rate. As defined by Gao et al. (2008), the entropy rate $H(X)$ of random process $X$ (Equation 4.2) is the asymptotic rate at which the entropy of sub-sequence $X^n_1$ changes with increasing $n$.

$$H(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, X_3, \ldots, X_n) \quad (4.2)$$

where, $H(X_1, X_2, X_3, \ldots, X_n)$ denotes the entropy of the joint variable $X^n_1$ defined for the subsequence $X_1, X_2, \ldots, X_n$. As stated by Gao et al. and Cai et al. (2004), this limit exists for all stationary random processes and is equal to

$$H(X) = \lim_{n \to \infty} H(X_n|X_{n-1}, \ldots, X_2, X_1) \quad (4.3)$$

$$= \lim_{n \to \infty} - \sum_{x^n_1 \in E^n} p_n(x^n_1) \log_2 \frac{p_n(x^n_1)}{p_n(x^n_{n-1})} \quad (4.4)$$

where $p_n$ denotes the joint probability distribution of a sub-sequence of length $n$. As described by equations 4.2, 4.3, and 4.4, entropy rate measures the average entropy of each new event generated by random process $X$, accounting for preceding events. It is measured as the entropy per event emitted and has units of bits per-event. The entropy rate of a random process with no memory is exactly equivalent to the entropy of this process as each new event is independent of the previous. As such, the entropy of a process is an upper bound for its entropy rate. In contrast, a process in which the outcome of an event $X_i$ is determined exactly by the previous events ($p(X_i = x|X_{i-1}, X_{i-2}, \ldots) = 1$) has entropy rate of 0, because each new event can be predicted with no uncertainty and provides no new information. Informally, entropy rate is the average measure of uncertainty, or information, or surprise, associated with each additional event generated in a sequence of events. The more memory in a random process, the more information the previous events provide about the next event, and therefore the lower the entropy rate of the process. In addition, memory in the random process is directly related to the order in which events are observed. Specifically, the more memory in a random process, the more the order of the events it generates will tend to repeat. In the context of travel, a visit to the doctor considerably increases the likelihood of a subsequent visit to the pharmacy. Therefore, the ordered combination ‘doctor-pharmacy’ will tend to repeat over time. In line with these characteristics, entropy rate is a good regularity measure of travel sequences because it is sensitive to the relative frequency of events but also to the dependencies between multiple events.

Song et al. (2010) computed the entropy rate of hourly-location sequences derived from cell phone data in order to explore predictability in individual location patterns. Hourly-location sequences are an example of sequence with very low entropy rate because the location of a person during a given hour is highly related to her location in the previous hour. Indeed, individuals tend to visit locations for several hours.
consecutively (e.g. 8 hours at work or 14 hours at home). In this sense, longer average activity duration will result in low entropy rate and shorter average activity duration will result in higher entropy rate for such a sequence. Hence, Song et al. report very high predictability in hourly-location sequences. While this result is of limited practical use because it merely reflects the tendency of individuals to visit locations for several hours at a time, their approach demonstrates how entropy rate can be used to quantify the dependencies between elements of the same sequence.

4.3.1 Entropy Rate Estimation

Having identified entropy rate as a useful measure of regularity, we discuss the approaches used to estimate entropy rate from a finite sequence. Much like the true mean or variance of a probability distribution can only be estimated given a sample from the distribution, a finite realization $x^n_j$ of process $X$ only allows for an estimation of the true entropy rate of $X$. As expressed in Equation 4.4, entropy rate is a function of the unknown joint probability distribution $p_n$ of the sequence $X^n_1$. A naïve approach consists of estimating $p_n$ from the observed frequency of combinations of symbols in $X^n_1$. This approach becomes computationally intractable as symbol combinations of increasing length are considered.

Most entropy rate estimation approaches circumvent the issue of estimating $p_n$ by relying on universal data compression algorithms (Gao et al., 2008; Cai et al., 2004). As pointed by Gao et al. (2008), these algorithms, used to compress data generated from processes with unknown probability distribution and arbitrarily long memory, are known to achieve optimal lossless compression ratio (i.e. compression ratio equal to the entropy of the generating process). Hence, they can be used to estimate the amount of redundant, or repeated information in a sequence of symbols. For instance, an English text of 100,000 bits contains a significant amount of redundant information because some of the words and combinations of words it contains are repeated. The more repetition it contains, the more the text can be compressed, by coding frequently repeated expressions. For example, if the 27-symbol long phrase "the probability distribution" occurs frequently, it can be coded by a single symbol.

Three families of lossless compression methods have been applied to entropy rate estimation. Willems et al. (1995) and Willems (1998) introduced the context-tree weighting (CTW) based entropy estimator. Cai et al. (2004) developed an estimation approach based on the Burrows-Wheeler transform (BWT), and Gao et al. (2008) reviewed and developed different estimators based on the Lempel-Ziv (LZ) compression algorithms. The estimators perform differently with respect to efficiency and bias depending on the property of the source and the sequence realization considered. Longer sequence realizations provide entropy estimates with smaller variance, and the variance of different approaches converges at different rates. The size of the alphabet also influences accuracy, with larger alphabets resulting in both higher variance and potentially higher bias for equal number of observations. Gao et al. (2008) present an extensive comparison of LZ and CTW estimators using simulation.
for binary sequences. They find that the CTW estimator consistently provides more accurate and reliable results than LZ based estimators. Cai et al. (2004) establish an upper bound on the convergence rate of the BWT estimator for finite-alphabet finite-memory processes and demonstrate that the BWT estimator performs better than an LZ-based estimator for binary sequences. No direct comparison of the CTW and BWT estimates has been reported in the information theory literature.

The BWT estimator is generally simpler to implement. Hence, the BWT entropy estimator with uniform segmentation described in detail by Cai et al. (2004) was used for the application presented in Section 4.4. The authors prove almost-sure convergence of this estimator for stationary ergodic random processes. These properties are assumed to hold for travel sequences described by the formulation previously introduced. Specifically, we assume that the underlying characteristics of an individual’s mobility do not change over the period for which the individual is observed. This assumption would be violated if a long-term change (e.g. change in residential location or job) took place during the period of analysis. While such changes may occur for a few users, we assume that such effects are negligible for the sample considered. Moreover, violating the stationarity and ergodicity assumptions may not be critical for the travel sequence application, which leverages entropy rate as an index of regularity, primarily used to compare regularity across individuals.

The BWT entropy estimator is computed in two steps. First, the Burrows-Wheeler transform is applied to the finite sequence $X^n_1$ of length $n$. Adjeroh et al. (2008) provide an in depth discussion of the transform, and its properties and implementation. Table 4.2, adapted from Adjeroh et al. (2008), illustrates how the BWT operates. The BWT is applied to sequence aardvark, resulting in transformed sequence kavraad. As Adjeroh et al. (2008) describe, first all rotations of the input sequence are listed and sorted alphanumerically. Then, the last symbol of each rotation is retained. The output of the BWT groups together outcomes (or symbols) which occur in similar contexts in the original sequence.

<table>
<thead>
<tr>
<th>All Rotations</th>
<th>Sorted Rotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>aardvarkk</td>
</tr>
<tr>
<td>ardvarka</td>
<td>ardvarka</td>
</tr>
<tr>
<td>rdvarkaa</td>
<td>arkaardv</td>
</tr>
<tr>
<td>dvarkaar</td>
<td>dvarkaarr</td>
</tr>
<tr>
<td>varkaard</td>
<td>kaardvar</td>
</tr>
<tr>
<td>arkaardv</td>
<td>rdvarkaa</td>
</tr>
<tr>
<td>rkaardva</td>
<td>rkaardva</td>
</tr>
<tr>
<td>kaardvar</td>
<td>varkaard</td>
</tr>
</tbody>
</table>

Table 4.2: BWT Illustration (adapted from (Adjeroh et al., 2008))

Formally, the BWT of any stationary process $X$ with finite memory results in a piecewise memoryless sequence. Cai et al. (2004) leverage this property of the transformed output to estimate the entropy rate of the process that generated the original
sequence. Specifically, in the second step of the estimation, the transformed sequence is segmented into $S$ segments $s$ of uniform length and the distribution of outcomes is estimated for each segment according to Equation 4.5.

$$
\hat{q}(x, s) = \frac{N_s(x)}{\sum_{y \in E} N_s(y)}
$$

(4.5)

where $N_s(x)$ denotes the number of occurrences of symbol $x$ in segment $s$. Given $\hat{q}(x, s)$, the entropy of each segment $s$ is estimated by Equation 4.6. Finally, the entropy rate of $X$ is estimated by the average entropy of all segments (Equation 4.7).

$$
\log_2 \hat{q}(s) = \sum_{x \in E} N_s(x) \log_2 \hat{q}(x, s)
$$

(4.6)

$$
\hat{H}(X) = -\frac{1}{n} \sum_{s \in S} \log_2 \hat{q}(s)
$$

(4.7)

Cai et al. (2004) recommend that the length of each segment $s$ is set as the integer value closest to $\sqrt{n}$. As mentioned above, the accuracy of the resulting estimate depends on both the length of the sequence observed and the number of different outcomes it contains.

4.4 Application

In this section we illustrate an application of the regularity concepts described above to a set of individual locations inferred from smart card data. The output of the activity sequence inference described in chapter 3 constitutes the input of this application. Table 4.3, from chapter 3, illustrates the structure of this output. Each row is associated with an activity event, inferred from the origins and destination of a user’s journeys. The activity sequences are inferred from all journeys observed in the 29-day period between February 10th and March 10th. Each activity is associated with a start-time and an end-time recorded to the nearest minute (refer to Table 4.1 for the explanation of each activity status).

**Table 4.3: Activity Sequence - Output Structure**

<table>
<thead>
<tr>
<th>User ID</th>
<th>Status</th>
<th>Start Time</th>
<th>End Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>1</td>
<td>2/10/2014 00:00</td>
<td>2/10/2014 07:30</td>
</tr>
<tr>
<td>1000</td>
<td>-2</td>
<td>2/10/2014 07:30</td>
<td>2/10/2014 08:17</td>
</tr>
<tr>
<td>1000</td>
<td>2</td>
<td>2/10/2014 08:17</td>
<td>2/10/2014 17:21</td>
</tr>
<tr>
<td>1000</td>
<td>-2</td>
<td>2/10/2014 17:21</td>
<td>2/10/2014 18:08</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>2/10/2014 18:08</td>
<td>2/10/2014 07:30</td>
</tr>
</tbody>
</table>
Sequences are translated from the form illustrated in Table 4.3 to a sequence of discrete symbols corresponding to the formulation described in Section 4.2.1. Each event is represented by its activity status, abstracting start and end times. Consecutive days with no travel are represented by a single ‘0’ status code. Figure 4-8 illustrates the resulting durationless sequence of areas visited by the user represented in Figure 4-7. Note that public transit journeys are excluded from the sequence for this particular application as they occur deterministically between every new area visited, given the current methodology. Therefore, including journey events would artificially decrease the estimated randomness of all sequences. If different journey types were differentiated using different status symbols (e.g. by mode, by route, or by duration), journey events would vary probabilistically and could be included in the travel sequence.

The entropy $H(X)$ and entropy rate $H(X)$ associated with an observed user sequence $x^*_1$ containing $n$ events is estimated as described in Section 4.3. $H(X)$ is computed according to Equation 4.1, with the probability $p(x)$ of an event $x \in E$ estimated by

$$\hat{p}(x) = \frac{n_x}{n} \quad (4.8)$$
where \( n_x \) represents the number of occurrences of \( x \) in the observed sequence \( X_1^n \) and \( n \) represents the length of the sequence observed. The entropy rate \( H(X) \) of the sequence is estimated using the BWT method described in Section 4.3.1. Figure 4-9 shows the distribution of sequence length for the sample of 99,925 users. Some users in the sample completed few journeys over the 29 days sequence. Longitudinal variability in the behavior of the same individual over time can only be measured from multiple observations of the same person. The intrapersonal variability of passengers who use public transit very infrequently cannot be analyzed from smart card data. All user-sequences shorter than 10 events long (in this case shorter than 10 activity statuses) are excluded from the regularity analysis presented next. The resulting sample contains 76,838 user-sequences.

The entropy and entropy rate distributions estimated for these sequences are presented in Figure 4-10a and 4-10b respectively. As previously discussed the value of entropy \( \hat{H}(X) \) is equivalent to the entropy rate of a sequence with no memory (or for which the order of events is ignored). The entropy distribution has a mean of 2.5 bits and a standard deviation of 0.53 bits. As a reference, a fair coin toss has entropy of 1 and a fair six-sided die roll has entropy of 2.6. Hence, on average, without accounting for the information provided by the order of events, a user-sequence is almost as random as a fair die roll. Individuals at the low-end of the distribution tend to repeatedly visit the same locations, and are therefore more predictable, while users at the high end of the distribution visit many locations and are therefore more unpredictable.

In contrast, the entropy rate distribution has mean 1.4 bits and standard deviation of 0.42 bits. The 1.1 bit difference between the two averages reflects the additional information provided by the order in which events take place. Considering the order in which events are generated, an average user-sequence is associated with only slightly
Figure 4-10: Entropy Distributions

(a) Entropy Distribution

(b) Entropy Rate Distribution

(c) $\hat{H}(X) - \bar{H}(X)$ Distribution
more uncertainty than a coin toss. In others words, on average, the next event can be predicted accurately almost 1 in 2 times when the order of events is considered, and only when 1 in 6 times when the order is not captured. Of course, the order of event does not provide the same amount of information for all individuals. As illustrated in Figure 4-10c, for some individuals, the order of events provides almost no information, while for others it reduces the uncertainty by as much as two bits. Specifically, individuals who visit many locations frequently, but always in the same order will have relatively high entropy, but relatively low entropy rate.

Figure 4-11 illustrates this for an actual user who visited 5 locations almost exactly the same number of times, but consistently in the same order. The estimated entropy of her sequence is 2.6, while its entropy rate is 1.0. The resulting 1.6 bit difference \( \hat{H}(X) - \hat{H}(X) \) for this individual is at the high-end of the distribution shown in Figure 4-10c. Additionally, while this individual’s routine is not conventional, it is clearly regular as both the events and the order in which they are combined are repeated over time. This is reflected by the below average entropy rate of this sequence.

Figure 4-12 illustrates another example of non-workday regularity captured by the entropy rate measure. On four separate occasions, the individual shown in the figure travels from her primary location to the her secondary location on a given day, is unobserved in the following one or two days, and travels back from the secondary to the primary location on the fourth day. In this specific example, the secondary location includes a terminal rail station, which the individual likely uses to leave London for the week-end. The user sometimes leaves on Friday, sometimes on Saturday and returns either on the following Sunday or Monday. Even though the pattern repeated spans more than single day and is not repeated periodically, its regularity is captured by the entropy rate of the sequence estimated to 0.503, below the sample average. This pattern would not accurately be captured by a periodicity measure as it does not reoccur on the same days of the week from week to week.

Examination of other individuals shows that, in general, entropy rate measures regu-
larity accurately and can serve as a useful comparison metric. Figure 4-13 compares two groups of 500 users who’s sequence is longer than 40 events. The rows of each pane represent the sequence of an individual, while the columns correspond to different times of the 29 day period. The first group is randomly selected from all users with entropy rate below 1.0 bit. These regular users fall below the 10\textsuperscript{th} percentile of entropy rate for sequences longer than 40 events. The second group is randomly selected from all users with entropy rate above 2.1 bit. These irregular users fall above the 90\textsuperscript{th} percentile of entropy rate. As expected, the sequences associated with regular users are characterized by the conventional working week structure. The irregular sequences contain much less repeated structure. It is important to note that while the dominant pattern in Figure 4-13a is associated with the typical working week, a number of non-conventional patterns, such as those illustrated above, are also qualified as regular. This demonstrates that entropy rate can be used as an indicator of regularity, capturing the extent of repetition in events and in the order in which they appear.

**Figure 4-12:** Location of User 2
(a) Regular Users

(b) Irregular Users

**Figure 4-13:** Comparison of users in the lower and upper 10\textsuperscript{th} percentile of irregularity
4.5 Conclusion

This chapter presents a discussion of intrapersonal variability and argues that the relationship between travel events must be incorporated in the analysis of intrapersonal variability. We hypothesize that the order in which an individual engages in journeys and activities constitutes an integral characteristic of her travel, and that this characteristic should be captured in the definition of travel behavior regularity. We present a measure of regularity based on entropy rate which is sensitive to the frequency of events and to the order in which events are observed. This measure is applied to a sample of over 75,000 user-sequences inferred from smart card transaction, through an entropy rate estimation procedure based on the Burrows-Wheeler transform. Results demonstrate that entropy rate accurately captures patterns of regularity for conventional working-week patterns and for more atypical routine patterns. The results also support the hypothesis that the order of events is important and captures a component of regularity not been previously considered in the literature.

The presented application focused on travel-sequences reconstructed from 4 weeks of data. In future research, it would be useful to evaluate the impact of sequence length on the convergence of the entropy rate estimate. Additionally, other extensions to the work presented in this chapter could focus on comparing the different estimation approaches mentioned in Section 4.3.1. For example, the CTW may provide better accuracy for sequences with large alphabets. The benefits and disadvantages of different estimators in the specific context of travel sequences remain to be explored.
Chapter 5

Clustering Travel Sequences

This chapter introduces an approach to identify similarities in the travel behavior of different individuals. Specifically, a user segmentation methodology based on observed travel behavior is proposed. First, examples of similarities in travel sequences are used to illustrate the type of segmentation to define. Second, an overview of the two-step segmentation methodology is given. Third, the first step, focused on categorizing the transit usage distribution of users, is described and applied to a sample of real users. Fourth, the second step, focused on categorizing the mobility and activity pattern of frequent transit users, is introduced and implemented. Finally, the robustness of the segmentation approach is examined.

5.1 Introduction

The previous chapter focused on extracting and quantifying regularity in the travel pattern of each user. For this purpose, each individual was considered independently. In this chapter, we extend the analysis to compare elements of structure in travel behavior across individuals. The objective of this chapter is to identify recurrent patterns of behavior in order to segment users according to the nature of their activity pattern. Figure 5-1 illustrates examples of such recurrent patterns.

In the figure, each sequence is associated with a different user. While each of these 6 sequences is unique at some level of detail, some shared patterns can be observed in their underlying structure. For instance, both individuals 1 and 2 follow working-day schedules, and tend to travel more rarely on week-ends. In contrast, individuals 3 and 4 to perform multiple short activities every day, and the nature of their activity pattern is similar on both week and week-end days. Finally users 5 and 6 both follow similar working-day schedules as 1 and 2, but the nature of their travel pattern changed significantly during the second week of the period considered above. During this week, they performed fewer journeys and did not follow their usual working schedule.
Figure 5-1: Six Example Individual Travel Patterns
In general, the similarities and dissimilarities in the sequences above relate to two aspects of their underlying structure: first, the daily organization of the different locations visited by a user, and second, the longitudinal organization of different daily location patterns. The first aspect pertains to the daily pattern of locations visited by users. Such pattern is defined by the locations visited, the time at which they are visited and the order in which they are organized. For example, many individuals are observed at their primary location in the morning, their secondary location during the day and return to their primary location in the evening, as they follow the conventional working day pattern. The second aspect relates to the location patterns observed across multiple days. For example, while individuals 1, 2, 5, and 6 all shared similar within day patterns, the longitudinal pattern of these days differs across the first two and the last two. The definition of a user segmentation should target both within day and day-to-day, or longitudinal, location patterns of this kind.

While identifying trends in the mobility and activity patterns of multiple users is the objective of this chapter, the framework introduced in chapter 2 reveals transit usage patterns must first be considered in order to compare such patterns. As introduced by figure 2-2, smart card transactions may provide a window into the activity and mobility pattern of each user. However, the size and shape of this window is largely related to the nature of each passenger’s level of public transit usage. Some individuals rely primarily on public transit to travel while others use it more rarely for specific types of trips. Hence, the segmentation process is designed as a two-stage process. First, users are segmented with respect to the distribution of their public transport travel over time. Second, users who achieve a large portion of their overall mobility by PT are further analyzed. For this group, a more detailed segmentation focused on the attributes described above is defined.

The remainder of this chapter is organized as follows. First, the public transport usage segmentation approach is described and applied to a sample of 99,925 users. Second, the detailed location pattern segmentation approach is described and applied to the segment of frequent users identified from the previous step. The resulting clusters of this segmentation is then described. Finally, the robustness of the clusters identified is evaluated through sensitivity analysis, and by replicating the process for second sample of users extracted from a later time period.

5.2 Transit Usage Cluster Analysis

This section describes an approach to categorize users according to transit usage. Public transportation travel represents a different proportion of each individual’s general mobility. Some passengers use transit systematically for most of their travel, while others use it only at certain times or for specific trip purposes. As introduced in Chapter 2, smart card offers a window onto the activity and mobility pattern of each user. This section focuses on classifying each user according to the portion of their
Figure 5-2: Examples of Day-to-Day Transit Usage Patterns
mobility which is captured from smart card transactions. Categorizing the activity and mobility patterns themselves will be the focus of the following section.

For illustration purposes consider PT usage of three different individuals summarized in Figure 5-2. In this figure, usage patterns are summarized as a longitudinal sequence of days. Each unit on the x-axis corresponds to a different day and each day with 1 or more PT journeys is labeled with a value of 1. Hence, the sequence summarizes both the level of PT usage and its distribution over time. For example, individual 1 was observed on almost every day of the week and traveled in this fashion throughout the entire period considered. In contrast, individual 2 also traveled on several consecutive days all concentrated within a 4-day period. Finally individual 3 traveled infrequently but recurrently over the period of analysis.

These usage patterns contrast two important dimensions: the frequency of transit travel and the distribution of this travel. In line with the window analogy, travel frequency relates to the size of the window, while the temporal distribution relates to the shape of the window. Both individual 2 and 3 were observed traveling on 5 different days, but the distribution of these days suggests that different portions of mobility patterns are observed. Any classification of usage patterns must account for differences and similarities in both these dimensions.

5.2.1 Usage Clustering Methodology

In order to capture both dimensions we use a multivariate segmentation approach. Under this approach, the usage pattern of each individual is first summarized by an attribute vector of metrics describing the distribution and frequency of public transport travel. Then patterns are compared and grouped according to their relative proximity in the space defined by the attribute vector.

Clustering Variables

Different metrics can be considered to summarize usage frequency and distribution. In general, frequency and distribution metrics rely on 5 basic elements of the days travel sequence. These elements are illustrated in Figure 5-3. The period of analysis $T$ corresponds to the length in days covered by the data used. This measure is independent of the users behavior. The spread $S$ corresponds to the number of days between the first and last day for which the user was observed. The number of gaps $G$ corresponds the number of intervals during which no travel was observed. The length of a gap $i$ in days is denoted by $g_i$. Finally, $D$ represents the number of days traveled by an individual over period $T$. Table 5.1 summarizes different metrics of frequency and distribution obtained from these 5 measures.

The suitability of a metric depends on the length of the period of observation. For short periods, frequency related attributes are most important as distribution is hard
to estimate over shorter time horizons. As the period of analysis increases, distribution related metrics become increasingly relevant because frequency is likely to change over time. Such changes in frequency can be related to any changes in travel routines, caused for example by holidays, seasonal effects, changes in employment, or changes in residence. Specifically, all gap related metrics become especially relevant as period $T$ increases, such that a higher number of gaps $G$ is observed. Given the period of 29 days considered in this application, number of days traveled and range are good candidates for clustering.

**K-Means**

In order to group users with respect to the chosen metrics, a multivariate cluster analysis approach can be used. Multivariate cluster analysis is useful to compare and group observations characterized under multiple dimensions without a priori knowledge of group structure. The unsupervised learning literature contains a number of different cluster analysis algorithms each with different strengths and weaknesses (Jain et al., 1999; Jain, 2010). One of the most commonly used of these is the K-Means algorithm. K-Means performs an iterative search of multiple data points to find the best combination of centroids for a specified number of clusters $k$ (Gan et al. (2007), for a detailed discussion of the algorithm).
Table 5.1: Metrics of Transit Usage Frequency and Distribution

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Number of Days Traveled</td>
</tr>
<tr>
<td></td>
<td>Days Traveled per Days of Spread</td>
</tr>
<tr>
<td></td>
<td>Days Traveled per Days of Analysis</td>
</tr>
<tr>
<td>Distribution</td>
<td>Spread</td>
</tr>
<tr>
<td></td>
<td>Number of Gaps</td>
</tr>
<tr>
<td></td>
<td>Average Gap Length</td>
</tr>
<tr>
<td></td>
<td>Gap Length Standard Deviation</td>
</tr>
<tr>
<td></td>
<td>Gap Length Coefficient of Variance</td>
</tr>
<tr>
<td></td>
<td>Minimum Gap Length</td>
</tr>
<tr>
<td></td>
<td>Maximum Gap Length</td>
</tr>
</tbody>
</table>
### 5.2.2 Application: Clustering Usage Patterns

Next, we analyze the dataset described in Chapter 3 according to the methodology described above. First, we examine the number of days traveled and the spread of days traveled for the sample of 99,925 users. These two measures are then used as metrics of frequency and travel distribution to perform cluster analysis.

Figure 5-4 shows the distribution of the number of days traveled for all cards in the sample considered. Two distinct components of this distribution are observed. At the low-end, the distribution of days traveled approaches an exponential distribution. For users with few days traveled, traveling by transit on any given day is approximated by a Poisson process. In contrast, the upper portion of the distribution approaches a normal distribution, with a second mode around 21 days. This corresponds to an average of 5 days traveled per week, in line with the commuter week. Overall, 44% of users in the sample were observed traveling on fewer than 5 days over the period considered. The small number of users observed on 30 distinct days are those who traveled between 12:00 am and 4:30 am on the day following the last day of the period of analysis. For transaction storage purposes, Transport for London assigns late night and early morning transactions made before 4:30 am to the calendar day directly preceding the day on which they are collected.

![Distribution of Days Traveled](image)

**Figure 5-4:** Distribution of Days Traveled

Figure 5-5 shows the distribution of the spread of days traveled across all users. Similarly to the days traveled distribution, this distribution appears to consist of two different component distributions. Each end of the distribution corresponds to a different exponential distribution: one for users with a short range and another for users with a long range. In line with this, two sharp modes are observed. The first
is associated with users who were only seen on 1 day, while the second is associated
with users who’s range spans the full period of analysis. The middle portion of the
distribution is characterized by secondary modes for ranges of 8, 15 and 22 days.
These modes align with multiples of 7 days plus 1, probably in line with the weekly
periodicity of users travel. For example, if a given user traveled once a week for 3
weeks, she would be observed for a range of 22 days.

Using the range and days traveled for each user $u$, we perform cluster analysis using
K-means. The K-Means implementation of MATLAB’s statistics toolbox with the
standard euclidean distance function was used for this purpose. When using euclidean
distance, K-means is sensitive to the relative scale of the clustering variables. If the
scale of one variable is much larger then the others, the distance between points is
dominated by the component associated with this variable. Hence, the value of days
traveled and spread was normalized for each individual as indicated in equations 5.1
and 5.2 respectively.

$$
d_u = \frac{D_u - \hat{\mu}_D}{\hat{\sigma}_D} \quad (5.1)
$$

$$
s_u = \frac{S_u - \hat{\mu}_S}{\hat{\sigma}_S} \quad (5.2)
$$

where $d_u$ and $s_u$ denote the normalized days traveled and range of user $u$ respectively,
and $\hat{\mu}_D$, $\hat{\sigma}_D$, $\hat{\mu}_S$, $\hat{\sigma}_S$ denote the sample mean and standard deviation of days traveled
and range respectively.

**Figure 5-5:** Distribution of Range of Days Traveled
K-Means, like many other clustering algorithms, is sensitive to initialization conditions. In order to limit initialization sensitivity, each initial cluster centroid is selected using the K-Means++ algorithm (Arthur and Vassilvitskii, 2007). Using K-Means++, initial centroids are selected as a function of point density in the space defined by the clustering variables. Specifically, initial centroids are selected sequentially, such that the probability of selecting any given point as a centroid is inversely related to the distance between this point and the previously selected centroids. In addition to this initialization approach, the clustering process is replicated 100 times, with a different set of initial centroids. The solution with the minimum sum of within cluster distances is retained as the final solution.

In order to evaluate the optimal number of clusters $k$, clustering solutions with a different number of clusters are compared. Solution optimality is evaluated with respect to the Davis-Bouldin (DB) index, a measure of the ratio between the sum of within cluster distances to the centroid and the sum of across cluster distances (Davies and Bouldin, 1979).

$$DB = \frac{1}{k} \sum_{i}^{k} \max_{j \neq i} R_{ij}$$

$$R_{ij} = \frac{E_{i} + E_{j}}{M_{ij}}$$

$$M_{ij} = d(C_{i}, C_{j})$$

$$E_{i} = \frac{1}{T_{i}} \sum_{l=1}^{T_{i}} d(X_{l}, C_{i})$$

where $C_{i}$ represents the centroid of cluster $i$, $X_{l}$ represents the feature vector of point $l$, and $d(a, b)$ is the distance function between two vectors $a$ and $b$ of arbitrary dimension, in this case euclidean distance. For the most appropriate clustering solutions, the distance between points contained in the same cluster is expected to be small compared to the distance between cluster centroids. Hence, low values of DB index indicate better clustering.

Results

The result of the cluster analysis of transit usage patterns is presented next. Segmenting usage patterns into 3 clusters provides the most intuitive insight and results in the most distinctive clusters.

Figure 5-6 shows the DB index of clustering solutions with $k$, the number of clusters, varying between 2 and 10. Each point in black indicates the DB index for a given solution composed of $k$ clusters computed for all 99,925 sampled users. In addition, each light grey point represents one of 20 bootstrap samples used to estimate the
variability in the DB index for a given $k$ value. Each bootstrap sample is formed of a subsample of 30,000 users randomly drawn with replacement from the 99,925 users. The DB index is lowest for the 3-cluster solutions, closely followed by the 2-cluster solution. The magnitude of differences observed in the index for different values of $k$ is significantly larger than the magnitude of internal variation estimated from the bootstrap samples. Also, the internal variability of the index increases with increasing values of $k$. Overall, we select the 3-cluster solution as it is associated with the lowest DB index and provides a more detailed segmentation than the 2-cluster solution.

The distributions of days traveled and of spread of days traveled for clusters 1, 2, and 3 is shown in Figure 5-7. These distributions reveal three distinct usage patterns. The first cluster is associated with individuals who traveled on few days over a short range of time. These users traveled fewer than 10 days, with 80% traveling between 1 and 3 days. All users in this cluster were observed for a range smaller or equal to 14 days with over 60% of observed for 3 days or less. This cluster is likely to include visitors and non-residents who visited London for a few days then left the city. In line with these characteristics, users in cluster 1 are labeled non-recurrent users. As indicated in table 5.2, 36% of users fall under this category.

In contrast with cluster 1, the second cluster includes users who were seen traveling frequently over a long period of time. Over 82% of users in this cluster traveled over a range of 28 days or more. With respect to the number of days traveled, the mode of the distribution is observed at 21 days. All users in this group traveled at least 14 days over the period considered. 33% of all sampled users are assigned to cluster 2. In line with the high number of days traveled, users of this cluster are labeled as
Figure 5-7: Distribution of Days Traveled and Range: 3 Clusters
Table 5.2: Cluster Description: 3 Clusters

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Percent Users</th>
<th>Range (Days)</th>
<th>Travel Days</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36</td>
<td>4</td>
<td>2</td>
<td>Non-Recurent Users</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>28</td>
<td>22</td>
<td>Frequent Users</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>22</td>
<td>8</td>
<td>Occasional Users</td>
</tr>
</tbody>
</table>

frequent users. These are likely to include London residents who rely primarily on public transport to move around the city.

Finally, individuals in cluster 3 were also observed over a large portion of the window of analysis, but traveled much less frequently. The spread of days traveled is between 11 and 28 days. The spread distribution appears to be constituted of three component distributions, each corresponding to ranges of over 2, 3 and 4 weeks. Individuals in this cluster traveled between 2 and 16 days. The mode of the distribution is around 4 or 5 days traveled, corresponding to roughly 1 day traveled per week. This cluster accounts for 31% of the 99,925 users sampled. We label individuals in this cluster as occasional users as they travel infrequently but throughout the period of analysis. This could include individuals who do not rely primarily on public transport but use it occasionally, for example, for shopping or social purposes.

5.3 Activity Patterns Cluster Analysis

The previous sections identified three types of public transport usage: non-recurrent, frequent, and occasional usage. This section focuses on categorizing the activity and mobility patterns of frequent users. The examples given in the introduction of this chapter illustrate the types of similarities which can be observed between the activity patterns of multiple users. Given, the focus on activity patterns, this section considers frequent public transport users. Smart card transactions capture a more significant portion of travel patterns for these users than for occasional and non-recurrent users.

5.3.1 Methodology

Identifying recurrent mobility patterns can be done through two approaches, distinguished by the type of features used to represent the pattern of each individual. The first approach is parallel to conventional market segmentation clustering methods (see for example Wedel and Kamakura (2000)). In this case, features used to summarize the pattern of each person consist of descriptive measures with an a priori identified meaning. We refer to these as descriptive features. The second approach stems from signal processing and pattern recognition. For this approach, features are extracted from dimensionality reduction of travel sequences. This type of features is defined
based on inherent elements of sequence structure identified from statistical variations across all users. We refer to these as structural sequence features.

To illustrate the differences between the two types of features, it is useful to consider an analogy between user segmentation and face recognition. Early facial recognition algorithms focused on matching faces based on predefined geometric attributes such as the distance between intuitive facial features (e.g. nose, mouth, ears and eyes). Recent and more successful approaches focus on identifying statistically meaningful variations in combinations of pixels to describe each face (Alpaydin, 2014). For example, Turk and Pentland’s (1991) seminal publication on eigenfaces described the benefits of using principal component analysis to capture variation in a set of face images without any judgment about what facial features are important. These approaches have two significant benefits. On one hand they lead to significant accuracy improvements. Additionally, they do not require a priori knowledge of meaningful facial features.

**Clustering Based on Descriptive Variables**

The descriptive approach to user segmentation is parallel to the earlier type of face recognition algorithms. A set of meaningful metrics such as the average number of activities in a day, the median first trip time in the morning, and the number of different stations visited is used to compare users. Just like the distance between the mouth, nose and eyes, these measures are determined from a priori expectations of important features. This approach is illustrated in Figure 5-8.

Most existing user segmentation research based on trip transaction data uses this approach. Ortega-Tong (2013) used 20 different clustering variables related to travel frequency, journey times, origin-destination pairs, activity duration, fare type and public transport mode choice to identify 8 different user segments. The analysis was performed using the K-medoids algorithm. She aggregated these 8 groups into four
categories: non-exclusive commuters, exclusive commuters, non-commuter residents, and leisure travelers. Kieu et al. (2014) defined measures of temporal regularity and spatial regularity to segment public transport users in South East Queensland, Australia. They manually define segment boundaries for the resulting distribution and identify four groups: irregular passengers, regular OD pair passengers, habitual time passengers, and routine OD and time passengers.

While the work of these authors highlights the potential of smart card data to classify travel patterns, a number of issues limit the applicability of their results. Ortega-Tong applied the same clustering methodology to two different samples of transactions collected on the same week of two consecutive years. The resulting clusters varied significantly between the two periods analyzed despite no significant changes in the passenger population. This temporal stability issue may in part be related to the high number of descriptive variables used to cluster passengers. As Kieu et al.’s (2014) approach segments users based on an arbitrarily chosen a threshold, it provides little insight about the inherent similarities of travel patterns in the passenger population studied. Additionally, the method can difficulty be applied to other cities or cultural contexts where a sensible boundary cannot be identified a priori.

**Clustering Based on Structural Sequence Features**

In contrast with the descriptive approach, the structural features approach is parallel to the eigenface (Turk and Pentland, 1991) techniques for face recognition. The travel and activity pattern of each individual observed over multiple days is modeled as an image. Each pixel of this image corresponds to an hour of the period of analysis (e.g. 29 days) and the value assigned to the pixel corresponds to the location of the individual during the hour. Then, the underlying sequence structure is identified by extracting statistical trends in variations of pixel values across all sequences.

Principal component analysis (PCA) is a commonly used method to identify such
statistical trends. Given a dataset of $m$ observations described by $n$ dimensions, PCA is used to identify a linear combination of the $n$ dimensions which best summarizes the variance along these dimensions. The resulting combination allows for the original $n$ dimensions to be summarized by a smaller number of dimensions. Specifically, this linear combination is derived from the eigen-decomposition of the covariance matrix of the $n$ dimensions. As illustrated in Figure 5-9, this process allows for user sequences to be clustered without having to identify a priori meaningful variables.


Both the work of Jiang et al. and Eagle and Pentland show how PCA can be used to summarize complex relationships between the multiple elements of travel sequences. The work also demonstrates that capturing these relationships can be done without a priori assumptions about their nature. In line with these benefits, the application presented in this section uses the sequence structure approach based on principal component analysis.

### 5.3.2 Application: Clustering Activity Patterns

The segmentation approach described above is next applied to the travel sequences of all 33,026 frequent users identified in the previous section as a case study. The input sequence for each user is a vector of 696 elements each representing 1 hour of the 29-day period. For each hour, the activity status in which the user spent most time is identified and assigned to the hour. In order to limit the computational complexity of the analysis, all activity statuses above the fourth geographical area are aggregated with the fifth area. Hence, user sequences are composed of at most 8 distinct activity statuses, summarized in Table 5.3. All user-sequence vectors are compiled into a large matrix $S$ of 33,026 rows, each associated to an individual, and 696 columns, each representing an hourly status.

#### Sequence Principal Components

In order to perform principal component analysis on these sequences, user patterns must be transformed from categorical sequences to numerical sequences. While each activity status is symbolized by a numerical character, these characters are used to represent distinct categories on no specific ordinal scale. Hence, each vector is transformed from a categorical vector of 696 elements to a binary vector of 5568
<table>
<thead>
<tr>
<th>Location Status</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>Public Transit Journey</td>
</tr>
<tr>
<td>-1</td>
<td>Discontinuity</td>
</tr>
<tr>
<td>0</td>
<td>No inference</td>
</tr>
<tr>
<td>1</td>
<td>Geographical user-area 1</td>
</tr>
<tr>
<td>2</td>
<td>Geographical user-area 2</td>
</tr>
<tr>
<td>3</td>
<td>Geographical user-area 3</td>
</tr>
<tr>
<td>4</td>
<td>Geographical user-area 4</td>
</tr>
<tr>
<td>5+</td>
<td>Geographical user-areas 5 and above</td>
</tr>
</tbody>
</table>

Table 5.3: Aggregated Activity Statuses

elements (696 hours × 8 categories) as seen in Figure 5-10. This vector contains 8 different sections of 696 elements, each associated with a different activity status. The 696 elements of each section are in turn associated with hours of the analysis period. An element is assigned a value of 1 if its activity status was visited by the user during the hour it represents and 0 otherwise.

Hence, matrix $S$ is transformed to a binary matrix of dimensions $33,026$ rows x $5568$ columns, where each column represents a status-hour. This binary matrix is standardized by subtracting the average of each column from all values in the column. This standardization ensures that the cloud of points representing users is centered around the origin of the space defined by the 696 hours analyzed Shlens (2014). The resulting standardized binary matrix is denoted by $B$. Row $u$ of $B$, representing the binary sequence of users $u$ is denoted by $b_u$. In order to compute the principal components of matrix $B$, equation 5.8 is solved to identify the eigenvectors $v$ and eigenvalues $\lambda$ of $B$’s covariance matrix $C$.

\[
C = B^T B \tag{5.7}
\]
\[
(C - \lambda I) v = 0 \tag{5.8}
\]

where $I$ denotes the identity matrix. The solutions to equation 5.8 are denoted by the eigenvalue and eigenvector sets $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $V = \{v_1, v_2, \ldots, v_n\}$ respectively, where $n$ represents the rank of $C$. The computation is implemented using the PCA function of MATLAB’s statistics toolbox. The result of this process is a set of 5568 orthogonal eigenvectors of dimension 5568. These vectors constitute the principal components (PC) of $B$. Each eigenvector is associated to an eigenvalue which is proportional to the amount of variation observed along the direction of the vector. Principal components are ordered according to their associated eigenvalue, such that the first principal components explain the most variation in $B$. For example, the first principal component is oriented in the direction along which the highest proportion of variability in $B$ is observed.

93
In order to gain an intuitive understanding of what the PCs represent, consider the first 3 PCs illustrated in Figure 5-11. Each principal component represents a weighted sum of the 5568 dimensions in matrix $B$. The weight assigned to different status-hours is related to the level of correlation between status-hours. For any one PC, status-hours with a high weight are those often observed to co-occur within the same sequence. Each principal component summarizes a different correlation pattern observed in the data. Consider, for example, PC 1 illustrated at the top of Figure 5-11. In this figure, the color of each status-hour represents its weight. The red status hours in this PC are those associated with being observed at location 2 during weekdays and at location 1 during weeknights. This indicates that individuals observed at location 2 on a given weekday were very often also observed at location 2 on other weekdays and at location 1 on weeknights. This is intuitive; for example, a user going to work on one day correlates with this user going to work on other days and returning home at night.

Status-hours with a negative weight are those which correlate negatively with other status-hours. For example, in PC 2, week-end hours at status 0 were all correlated with each other, while they were negatively correlated with week-end hours at status 1. Intuitively, observing no travel for a user on a given week-end correlates with this user not being observed at location 1 on future week-ends. Note that this relationship also holds backwards. Observing a user at location 1 on a given week-end correlates with not failing to observe this user on future weekends. PC 3 indicates that being

Figure 5-10: Sequence Binary Transformation
Figure 5-11: Principal Components 1, 2 and 3
observed at status 0 on any hour of the second week correlated with being observed at status 0 for the remaining hours of the week. This may indicate that many users took a holiday during this week.

User-sequences can be reconstructed by overlaying the correlation patterns described by a subset of PCs. A user-sequence is approximated by a weighted sum of the PCs. The weight $w_{i,u}$ of a component $v_i$ for user-sequence $b_u$ is computed by projecting $b_u$ onto $v_i$ (equation 5.9).

$$w_{i,u} = b_u \cdot v_i^T$$ (5.9)

This projection is a measure of the extent to which the pattern described by $v_i$ is observed within sequence $b_u$. For example, the first principal component illustrated in Figure 5-11 would have a high positive weight for the sequence of users who commuted on all week-days. PCs are ordered by the portion of variance they explain in the dataset. Hence user-sequences can be reconstructed up to some level of accuracy using few of the first principal components.

The correlation pattern captured by each principal component explains different portions of the variability observed in $B$. The variance explained by a principal component is loosely related to how well user-sequences can be reconstructed. Figure 5-12 shows the proportion of variance explained by the first 15 most significant PCs. The first 3 components account for over 15% of the variance. The marginal variance explained by each additional component decreases with each additional component. As shown in Figure 5-13, the first 40 principal components, (i.e. 0.7% of the original number of dimensions in $B$) capture over 50% of the variance. Almost 40% of the variance observed across all sequences is unexplained by the first 100 principal components. While this represents a significant percentage of the overall variation, not all variance observed in $B$ relates to meaningful variations in the structure of patterns. Indeed, a large portion of this variance may relate to deviations from the routine of each individual, or to idiosyncrasies specific to a certain person. Unless they are observed across multiple individuals, these specific deviations or idiosyncrasies provide no meaningful information about repeated sequence structure. By considering only the most important principal components we can exclude such noise. Exactly how many principal components should be used may depend on the interpretability and stability of the clustering results as will be exposed in the next two sections.
Figure 5-12: Principal Components - Variance Explained

Figure 5-13: Principal Components - Variance Explained
Cluster Analysis

The cluster analysis is implemented using the K-means algorithm described in section 5.2. As described above, user-sequences can be approximated from a weighted sum of principal components. A few PC weights provide a meaningful summary of the original 5568 user-sequence dimensions. This dimensionality reduction approach is used to extract clustering variables for each sequence. The projections of user-sequences onto the first 8 principal PCs is used as input clustering variables. In order to identify the optimal number of clusters $k$, the process is ran for all values of $k$ between 2 and 20. The K-means++ initialization approach is used and 150 replications are ran for each $k$ evaluated to avoid local-optimum solutions. The DB index is used to compare clustering solutions with different number of clusters.

![DB Index Index for 8 Principal Components](image)

Figure 5-14: DB Index Index for 8 Principal Components

Figure 5-14 summarizes the ratio of within-cluster distances to between cluster distances, as measured by the DB-index, for values of $k$ between 2 and 20. Lower values of DB indicate a better solution. This figure suggests two different $k$ values result in compact clusters. The first important drop in DB is observed for $k = 4$ and the second at $k = 11$. The optimal solution may depend on the application considered. On one hand, for a high-level aggregation of users, a solution with fewer clusters may be more useful. On the other hand, solutions with more clusters may be better suited to applications requiring more homogeneous and detailed groupings. Both the 11-group and the 4-group solutions are presented next. Figure 5-15 shows the distribution of PC weights for the first 8 principal components. Each distribution is associated to one of the 8 clustering dimensions used.
Figure 5-15: Distribution of PC Weight - PC 1 to 8
5.3.3 Clustering Results - 4 Clusters

The solution with 4 clusters is illustrated in Figure 5-16. In this figure, each cluster is illustrated by a random sample of 500 user-sequences. Within each pane, the sequences are aligned with respect to time, shown on the x-axis. The activity status of each user is indicated for each hour of the 29-day period by a different color. Each group of sequences is associated with distinctive visual characteristics which correspond to the underlying structure of the sequences it contains. Table 5.4 presents descriptive statistics of each cluster. All values reported in Table 5.4 are cluster averages, except for the number of journeys performed over the analysis period which is summarized by cluster median. $G^k_i$ denotes cluster $i$ of the clustering solution with $k$ clusters.

<table>
<thead>
<tr>
<th>Descriptive Variable</th>
<th>$G^4_1$</th>
<th>$G^4_2$</th>
<th>$G^4_3$</th>
<th>$G^4_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Users</td>
<td>24%</td>
<td>32%</td>
<td>29%</td>
<td>15%</td>
</tr>
<tr>
<td>Med. Nbr. of Journeys</td>
<td>49</td>
<td>43</td>
<td>73</td>
<td>41</td>
</tr>
<tr>
<td>Area 1 (% of time inferred$^a$)</td>
<td>53%</td>
<td>43%</td>
<td>57%</td>
<td>59%</td>
</tr>
<tr>
<td>Area 2 (% of time inferred$^a$)</td>
<td>28%</td>
<td>18%</td>
<td>18%</td>
<td>19%</td>
</tr>
<tr>
<td>Status 0 (% of time inferred$^a$)</td>
<td>2%</td>
<td>6%</td>
<td>1%</td>
<td>5%</td>
</tr>
<tr>
<td>Status -1 (% of time inferred$^a$)</td>
<td>9%</td>
<td>18%</td>
<td>11%</td>
<td>8%</td>
</tr>
<tr>
<td>Nbr. of Days with Single Journey</td>
<td>1.9</td>
<td>4.5</td>
<td>2.6</td>
<td>3.2</td>
</tr>
<tr>
<td>Nbr. of Distinct Locations</td>
<td>7.3</td>
<td>9.5</td>
<td>11.8</td>
<td>7.2</td>
</tr>
<tr>
<td>Med. Departure Time (weekdays)</td>
<td>8:08</td>
<td>10:33</td>
<td>9:14</td>
<td>9:01</td>
</tr>
<tr>
<td>Std. Dev. Departure Time (weekdays)</td>
<td>1:24</td>
<td>3:01</td>
<td>2:36</td>
<td>2:10</td>
</tr>
<tr>
<td>Med. Activities Per Day</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Nbr. of Weekend Days Traveled</td>
<td>3.0</td>
<td>4.3</td>
<td>6.7</td>
<td>2.8</td>
</tr>
<tr>
<td>Nbr. of Weekdays Traveled</td>
<td>19.1</td>
<td>15.5</td>
<td>19.2</td>
<td>15.8</td>
</tr>
<tr>
<td>Entropy Rate (BWT)</td>
<td>1.0</td>
<td>1.4</td>
<td>1.3</td>
<td>1.2</td>
</tr>
</tbody>
</table>

$^a$Percentage of inferred time which the user spent in given status. Days on which no journeys are observed are excluded.
Cluster $G_4^1$: User-sequences in this cluster share two dominant characteristics. First, clear working days are observed, with users spending most weekdays between 9:00 and 17:00 in their secondary location. This is reflected by the vertical green bands delineating weekdays in Figure 5-16. Second, travel for users in this cluster is significantly reduced during weekends. On average, these individuals traveled on only 3 weekend days over the 4-week period analyzed. This is illustrated by the vertical white bands delineating on weekend days in Figure 5-16. As indicated in Table 5.4, the average median departure time is earliest for users in cluster $G_4^1$ on weekdays, at 8:08. Additionally, users in this group are also most regular, as indicated by the lowest average entropy rate. These attributes suggest that these individuals frequently and regularly use public transport to commute. The limited extent of their travel on weekends and the low number of distinct locations they visited indicates that they use public transit primarily to commute. Hence, users in this cluster are labeled as exclusive commuters in line with Ortega-Tong (2013). This cluster accounts for nearly one fourth of frequent users.

Cluster $G_4^2$: The dominant feature of user-sequences in this cluster is the high number of hour bins for which location cannot be inferred. On average, excluding days where no travel was observed, 24% of hours in sequences assigned to cluster $G_4^2$ were inferred as status 0 or -1. Users in this cluster often complete only 1 journey per day traveled. This may suggest that they use other transport modes between transit journeys. For example, this group may include users who use public transit to commute to work in the morning and return home by carpooling or bicycle in the afternoon. For this reason, users in this group are labeled as multi-modal users. Multi-modal users account for nearly a third of all frequent users.

Cluster $G_4^3$: Users in this cluster completed the highest number of journeys and traveled on the highest number of both week and weekend days. On average, these individuals completed 4 activities per day. This group seems to include two types of sequences: users with distinct working-days who also used transit for non-work related activities and users who traveled on most days, performing multiple short activities within the same day. These users could include workers who travel around London after work and on weekends for leisure purposes, as well as university students and seniors who travel a lot but do not operate under a structured work schedule. This group is labeled multi-purpose transit users. Multi-purpose transit users make up 29% of all frequent users.

Cluster $G_4^4$: The travel patterns of individuals in cluster $G_4^4$ changed significantly during the second week of the analysis period, between Monday February 17\textsuperscript{th} and Sunday February 23\textsuperscript{rd}. This week corresponds to the school half-term in London. On the first, third, and fourth week, $G_4^4$ users follow a pattern similar to exclusive commuters, spending the middle portion of weekdays in their secondary location. In contrast with cluster $G_4^1$, the average activity duration in the secondary location is shorter for cluster $G_4^4$. This is reflected by the higher average percent of time spent in the primary location.
Figure 5-16: Clusters $G^4_1$ to $C^4_4$
In the second week, the working day pattern is replaced by more occasional travel scattered throughout the week. The effect of the second week is likely an indicator that users in this cluster operate on a schedule driven by the school calendar. For example, pupils and their parents would be included in this group. These users are referred to as school-driven users.

### 5.3.4 Clustering Results - 11 Clusters

For applications requiring more detailed and homogeneous sequence patterns, a partition with 11 different solutions is considered. Table 5.6 summarizes the descriptive statistics of each cluster. Figures 5-17 to 5-19 provide a visual representation of the underlying structure for these 11 clusters similar to Figure 5-16. Table 5.5 shows the cross tabulation of these two clustering solutions.

<table>
<thead>
<tr>
<th>Cluster (k = 4)</th>
<th>$G_{11}^{11}$</th>
<th>$G_{21}^{11}$</th>
<th>$G_{31}^{11}$</th>
<th>$G_{41}^{11}$</th>
<th>$G_{51}^{11}$</th>
<th>$G_{61}^{11}$</th>
<th>$G_{71}^{11}$</th>
<th>$G_{81}^{11}$</th>
<th>$G_{91}^{11}$</th>
<th>$G_{101}^{11}$</th>
<th>$G_{111}^{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_{1}^{4}$</td>
<td>99</td>
<td>35</td>
<td>0</td>
<td>15</td>
<td>21</td>
<td>1</td>
<td>0</td>
<td>25</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$G_{2}^{4}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>30</td>
<td>0</td>
<td>100</td>
<td>55</td>
<td>88</td>
<td>46</td>
<td>74</td>
</tr>
<tr>
<td>$G_{3}^{4}$</td>
<td>0</td>
<td>65</td>
<td>99</td>
<td>24</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>7</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>$G_{4}^{4}$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>20</td>
<td>99</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>23</td>
</tr>
</tbody>
</table>

Cluster $G_{11}^{11}$: The first cluster corresponds largely to the exclusive commuter cluster described in the 4-segment solution, with 99% of its members also assigned to cluster $G_{1}^{4}$. The effect of weekends is more pronounced for this cluster. Users in this group only traveled 2.2 weekend days on average.

Cluster $G_{21}^{11}$: The second cluster is characterized by clear working days during the week combined with weekend travel. This cluster contains daily commuters who also tend to travel on weekends. Whereas these users were mixed with exclusive-commuters and multi-purpose transit users in the 4-cluster solution, they are assigned to a separate group in the 11 cluster solution. 65% of users in cluster $G_{21}^{11}$ also belong in cluster $G_{3}^{4}$, while the remaining 35% belong in cluster $G_{4}^{4}$. Separating these passengers results in the more distinct effect of weekends for group $G_{11}^{11}$. In line with their commuting and leisure travel patterns, these users are referred to as non-exclusive commuters Ortega-Tong (2013).
### Table 5.6: 11 Clusters - Descriptive Statistics

<table>
<thead>
<tr>
<th>Cluster Average</th>
<th>$G_1^{11}$</th>
<th>$G_2^{11}$</th>
<th>$G_3^{11}$</th>
<th>$G_4^{11}$</th>
<th>$G_5^{11}$</th>
<th>$G_6^{11}$</th>
<th>$G_7^{11}$</th>
<th>$G_8^{11}$</th>
<th>$G_9^{11}$</th>
<th>$G_{10}^{11}$</th>
<th>$G_{11}^{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Users</td>
<td>15%</td>
<td>14%</td>
<td>10%</td>
<td>5%</td>
<td>7%</td>
<td>10%</td>
<td>6%</td>
<td>5%</td>
<td>9%</td>
<td>9%</td>
<td>9%</td>
</tr>
<tr>
<td>Med. Nbr. of Joursneys</td>
<td>46</td>
<td>68</td>
<td>79</td>
<td>47</td>
<td>51</td>
<td>42</td>
<td>48</td>
<td>47</td>
<td>39</td>
<td>70</td>
<td>35</td>
</tr>
<tr>
<td>Area 1 (% of time inferred(a))</td>
<td>52%</td>
<td>53%</td>
<td>65%</td>
<td>51%</td>
<td>70%</td>
<td>59%</td>
<td>34%</td>
<td>50%</td>
<td>49%</td>
<td>34%</td>
<td>42%</td>
</tr>
<tr>
<td>Area 2 (% of time inferred(a))</td>
<td>30%</td>
<td>26%</td>
<td>15%</td>
<td>19%</td>
<td>11%</td>
<td>21%</td>
<td>24%</td>
<td>19%</td>
<td>16%</td>
<td>18%</td>
<td>17%</td>
</tr>
<tr>
<td>Status 0 (% of time inferred(a))</td>
<td>2%</td>
<td>1%</td>
<td>1%</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
<td>3%</td>
<td>4%</td>
<td>6%</td>
<td>2%</td>
<td>12%</td>
</tr>
<tr>
<td>Status -1 (% of time inferred(a))</td>
<td>9%</td>
<td>10%</td>
<td>8%</td>
<td>12%</td>
<td>7%</td>
<td>8%</td>
<td>24%</td>
<td>13%</td>
<td>13%</td>
<td>26%</td>
<td>15%</td>
</tr>
<tr>
<td>Nbr. of Days with Single Journey</td>
<td>1.6</td>
<td>1.9</td>
<td>2.8</td>
<td>3.3</td>
<td>3.6</td>
<td>2.7</td>
<td>3.7</td>
<td>3.2</td>
<td>4.7</td>
<td>4.1</td>
<td>4.9</td>
</tr>
<tr>
<td>Nbr. of Distinct Locations</td>
<td>6.4</td>
<td>10.5</td>
<td>12.1</td>
<td>9.9</td>
<td>9.0</td>
<td>6.8</td>
<td>8.0</td>
<td>9.7</td>
<td>9.7</td>
<td>13.1</td>
<td>8.1</td>
</tr>
<tr>
<td>Med. Nbr. of Activities Per Day</td>
<td>3.2</td>
<td>3.6</td>
<td>4.1</td>
<td>3.4</td>
<td>3.5</td>
<td>3.3</td>
<td>3.2</td>
<td>3.5</td>
<td>3.1</td>
<td>3.8</td>
<td>3.0</td>
</tr>
<tr>
<td>Nbr. of Weekend Days Traveled</td>
<td>2.2</td>
<td>6.0</td>
<td>7.1</td>
<td>4.7</td>
<td>3.5</td>
<td>2.8</td>
<td>3.5</td>
<td>4.1</td>
<td>5.4</td>
<td>6.4</td>
<td>2.3</td>
</tr>
<tr>
<td>Nbr. of Weekdays Traveled</td>
<td>19.2</td>
<td>20.1</td>
<td>19.2</td>
<td>14.7</td>
<td>17.7</td>
<td>16.1</td>
<td>18.9</td>
<td>15.1</td>
<td>12.5</td>
<td>19.2</td>
<td>15.0</td>
</tr>
<tr>
<td>Entropy Rate (BWT)</td>
<td>0.9</td>
<td>1.2</td>
<td>1.3</td>
<td>1.4</td>
<td>1.3</td>
<td>1.2</td>
<td>1.2</td>
<td>1.3</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

\(a\) Percentage of inferred time which the user spent in given status. Days on which no journeys are observed are excluded.
Cluster $G_{11}^1$: The effect of working-days and week-ends is much weaker for users in cluster $G_{3}^1$. Rather, these passengers are observed making a high number of journeys on both weekdays and weekends. On average, users in this cluster performed 4.1 activities per day traveled and tended to start traveling later in the day, with a median departure time around 10:00 on both weekdays and weekends. Figure 5-17 also reveals that this cluster is characterized by generally shorter activities outside the primary area than other groups. This is reflected by the high percentage of time these users spent in the primary area. As seen in Table 5.6, cluster $G_{3}^1$ is associated with the highest average number of week-end days traveled. In line with these characteristics, passengers in this cluster likely follow a looser schedule not driven by standard working hours. For example, individuals not employed full-time who make journeys for multiple purposes at different times of the day could be included in this group. 99% of these users were also assigned to cluster $G_{4}^2$. Members of this group are referred to as 7-day week flexible-schedule users.

Cluster $G_{11}^2$: The common characteristic of users in cluster $G_{4}^2$ is more occasional travel in the first week of the period of analysis. Users in this cluster traveled on average 4.7 week-end days and 14.7 week-days. This cluster includes users from all groups of the 4-cluster solution, with 49% of users also included in cluster $G_{4}^1$ and the rest evenly distributed among groups $G_{2}^1$, $G_{3}^1$, and $G_{4}^1$. Group $G_{4}^1$ could include, for instance, commuters who were on holiday for the first week. While the frequency of travel was lower during the first week, a few journeys were still observed for some users indicating that those were still present in London during this time.

Cluster $G_{11}^3$: Users in cluster $G_{5}^1$ spend the most time at their primary location; on average of 70% of inferred time periods. The pattern emerging from this cluster in Figure 5-18 is similar to that of cluster $G_{3}^1$. Its members start traveling late with an average median first journey time of 10:30 on weekdays. They travel travel on most week-days and on Saturdays, but unlike users in $G_{3}^1$, rarely on Sundays. This could include individuals who spend most of their time at home and perform short daily activities. For example, retirees who travel daily for shopping, lunch or other activities could be included in this group. This group is referred to as 6-day week flexible-schedule users.

Cluster $G_{11}^4$: As for cluster $G_{4}^2$, the dominant feature characterizing cluster $G_{6}^1$ is a change in travel pattern for 1 of the 4 weeks analyzed. In this case, the second week of the period considered is marked by reduced travel. Clear working-days, early first journey start time on weekdays and decreased travel on week-ends are observed for the remaining three weeks. In comparison with $G_{4}^1$, the effect of working days is more distinct for $G_{6}^1$. As described above, the second week of the period of analysis aligns with the school-half term in London. 99% of users in cluster $G_{6}^1$ are also classified in group $G_{4}^1$ in, while the remaining 1% were grouped with exclusive commuters in cluster $G_{1}^4$. Passengers in $G_{6}^1$ are referred to as school-driven users.

Cluster $G_{11}^5$: Figure 5-18 illustrates that individuals in cluster $G_{7}^1$ spent most working hours in their primary area and non-working hours in their secondary area. This indicates that cluster $G_{7}^1$ might include individuals who were inferred to spend less
time at their home location than at their work location. This would result in a user’s primary area being associated with work and secondary area being associated with home. Frequent journey-sequence discontinuities may explain this pattern as suggested by the 24% average of time period inferred as status -1. In part, commute journeys using multiple modes may explain this pattern. For example, consider a user whose journey to work is composed of two-stages: first a non-PT trip to a rail station, followed by a rail journey to the center of London. If this passenger returns home in the evening through a different station for the last PT stage of the journey, a discontinuity would be observed from the end of one commuting day to the start of the next journey on the following day. This would result in failure to infer the users location at home overnight. If this pattern is repeated a few times over the period of analysis, the total time inferred in the home area will be smaller than the time inferred in the work area. Another possible sequence of journeys explaining this pattern could be associated with users who follow a reverse commute pattern, for example night workers. A third journey sequence explaining the inversion could relate to users spending their home time in more than one location. For example, individuals who spend nights in more than one location, (e.g. at their home and at a partner’s home) would be inferred spending more time at work than at one home location. Passengers in this group are referred to as reverse pattern commuters.

Cluster $G_8^{11}$: Similarly to group $G_4^{11}$, this group includes individuals whose travel was reduced during one of the weeks analyzed. As observed in Table 5.6 the attributes of both clusters are very similar. The only distinction between the groups is the specific week during which the pattern was observed to change.

The remaining three clusters are all characterized by large proportions of time periods for which location could not be inferred. Hence, the patterns identified by these clusters are more noisy and less clearly defined than the previous groups. While this may impact the interpretability of the clusters, some underlying differences between these three groups may still provide some insight about how users combine transit with other modes.

Cluster $G_9^{11}$: This group includes frequent users who traveled on fewest weekdays and performed the second smallest number of journeys. Additionally, their average start-time is around midday and their travel pattern is similar on weekdays and weekends. Despite making few journeys, they visit multiple locations, and are less regular than other frequent users. These passengers are associated with the second highest average number of days on which a single journey was observed. These characteristics suggest that this cluster includes users who use public transport on certain days for few non-work related activity. For example, individuals are not in full-time employment but who leave home less frequently than individuals in clusters $G_3^{11}$ and $G_5^{11}$ could be included in this group.

Cluster $G_{10}^{11}$: This cluster includes users who traveled on almost every day of the period of analysis, weekdays and weekends. Their journey sequence are characterized by multiple discontinuities. On average, 24% of time periods for passengers in this group are assigned status -1. These users perform multiple activities per day and
are associated with the highest irregularity index. This suggests that cluster $G_{10}^{11}$ is composed of multi-modal users who entwine public transport journeys within a complex activity schedule on almost every day they travel. For example, this could include individuals who combine active modes, such as walking and cycling, with public transport. Additionally, no clear working days are observed, suggesting users in this group do not tend to be employed full-time.

Cluster $G_{11}^{11}$: Users in this cluster share similar characteristics to those in cluster $G_{9}^{11}$. The most important difference between the two clusters is the frequency of travel on weekdays and weekends. While users in cluster $G_{9}^{11}$ traveled equally throughout the week, the travel of individuals in cluster $G_{11}^{11}$ is concentrated on weekdays. Additionally, passengers in cluster $G_{11}^{11}$ visited fewer distinct locations than those in cluster $G_{9}^{11}$ and are associated with the smallest number of days traveled. These differences, along with the high number of single journey days, may suggest these passengers use public transport for work related journeys but not to commute. For example, self-employed individuals working from home who sometimes need to travel to business meetings may be included in this cluster.

5.4 Robustness

We next evaluate the impact of different factors and assumptions on the definition of clusters. In order to ensure the approach is robust, sensitivity is considered with respect to 3 factors: principal component stability, the number of clusters, and the period of analysis.

5.4.1 Principal Component Stability

As principal components are determined from the variance observed between user-sequences without a priori assumptions, they may be sensitive to specific attributes of the sample which are not representative of the general population. Hence, it is crucial to evaluate their robustness. This can be done by comparing the principal components obtained from different samples. As extracting and processing additional samples from Transport for London’s database is computationally intensive, new subsamples can be defined by bootstrapping the original 33,026 frequent-user sample. In line with this approach, these users are sampled with replacement to define 20 subsamples of 10,000 users.
Figure 5-17: Clusters $G_{11}^{1}$ to $G_{411}^{1}$
Figure 5-18: Clusters $G_5^{11}$ to $G_8^{11}$
Figure 5-19: Clusters $G_9^{11}$ to $G_{11}^{11}$
For each of these 20 samples, we run the principal component analysis described in section 5.3.2 and compare the resulting PCs. The average correlation between pairs of matching components \(i, \rho_i\), is used as metric of PC similarity (Equation 5.10).

\[
\rho_i = \frac{1}{||P||} \sum_{k,l \in P} |v_{i,k}^T \cdot v_{i,l}|
\]  

(5.10)

where \(P = \{(k, l) : k, l \in \mathbb{N} \land k < l \leq 20\}\) denotes the set of 190 sample pairs defined from the 20 sub-samples, and \(v_{i,k}\) denotes the \(i^{th}\) principal component of sample \(k\). Figure 5-20 shows these results for the first 13 principal components. The x-axis indicates the principal component number. The y-axis shows the average correlation between all pairs of PCs defined for the 20 samples.

As seen from Figure 5-20 the average correlation between principal components is above 0.9 for the first 8 principal components, indicating high stability. As the variance explained by the number of principal components decreases, so does the stability across samples. This is intuitive as principal components explaining less variance are more likely to be associated with noise.

### 5.4.2 Number of Clusters

Having established the stability of the first principal components, we can evaluate the robustness of the \(k\) values selected. As detailed in section 5.3.2 the DB index was used to select the number of clusters. In order to ensure the DB index profile established
would generalize to another sample, it is useful to get a sense of variability in the index for different samples. As for the principal component stability, we define 20 sub-samples of 10,000 users by sampling the original sample with replacement. For each of the 20 samples, values of k from 2 to 20 were evaluated. Figure 5-21 shows the results for all 20 sub-samples. For a given value of k, each point represents the DB index for 1 of the 20 sub-samples. The results are consistent with the DB profile for the original sample. Solutions with 4 and 11 clusters consistently exhibit lower DB index values. Additionally, the results reveal that the variance in one DB-index is especially high for less compact clusters. For example, the standard deviation of DB for $k = 7$ is 0.033 while the standard deviation for $k = 11$ is 0.016.

5.4.3 Number of Principal Components

Figure 5-20 demonstrates that the first 10 principal components are stable across different bootstrap samples. While principal component stability is a necessary requirement for clustering, it does not follow that all stable PCs need be used as clustering variables. The principal components should be included based on the additional
discriminatory power they provide. In order to evaluate the marginal effect of each additional component on cluster definition, we repeat the clustering analysis for $k = 11$ using different number of principal components. Figure 5-22 shows how the stability of clusters evolves for different number of principal components. The $y$-axis shows the percentage of frequent users who were assigned to the same cluster in two different clustering solutions. For example, 32% of users were classified in the same clusters when comparing the clustering defined based on 1 component to the clustering defined based on 2 components. Comparing the 8-component clustering to the 9 component clustering reveals that 98% of frequent users are assigned to the same group in both solutions. The subsequent components are associated with similar stability, up to some level of natural variation. This indicates that PCs above the eighth component do not significantly affect which users are grouped together.

As the stability between different clustering solutions is low when few principal components are used, matching equivalent clusters is not always straightforward. Identifying which cluster of one solution corresponds to which cluster of another solution may not be obvious if users in each cluster of the first solution are assigned to multiple clusters of the other solution. For example, consider Table 5.7 which shows the cross tabulation of the two clustering solutions based on 1 PC and 2 PCs. The columns of this table show how individuals of a given cluster in the 2-PC solution are assigned to clusters of the 1-PC solution. For instance, of the 2967 users assigned to cluster 1 in the 2-PC solution, 1104 were also assigned to cluster 11 of the 1-PC solution. While this may suggest that these two clusters should be matched, the table also reveals that cluster 11 of the 1-PC solution is also strongly associated with clusters 2 and 7 of the 2-PC solution. In order to analyze cluster stability, it is necessary to
identify the cluster matching which maximizes the number of individuals assigned to the same clusters across both solutions. This is equivalent to finding the row permutation of Table 5.7 which maximizes the sum of the diagonal elements. Table 5.8 shows how the rows of Table 5.7 can be rearranged such that the sum of diagonal elements is maximized. Identifying this permutation is done by solving the standard assignment optimization problem, with clusters corresponding to nodes and cluster overlap corresponding to arc weights (Bertsekas, 1991).

### Table 5.7: Cross Tabulation of 1-PC and 2-PC Clusters ($k = 11$)

<table>
<thead>
<tr>
<th>Cluster Index (2 PCs)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>1095</td>
<td>839</td>
<td>0</td>
<td>836</td>
<td>1015</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>973</td>
<td>1097</td>
<td>26</td>
<td>116</td>
<td>0</td>
<td>3</td>
<td>70</td>
<td>1015</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>123</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>575</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1150</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>253</td>
<td>142</td>
<td>823</td>
<td>718</td>
<td>0</td>
<td>614</td>
<td>959</td>
<td>102</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>792</td>
<td>1165</td>
<td>0</td>
<td>0</td>
<td>384</td>
<td>453</td>
<td>451</td>
<td>295</td>
<td>172</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>927</td>
<td>908</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>973</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>321</td>
<td>657</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>736</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>95</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1426</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1025</td>
<td>809</td>
<td>0</td>
<td>959</td>
<td>1061</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1003</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1104</td>
<td>1550</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1003</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 5.8: Reordered Cross Tabulation of 1-PC and 2-PC Clusters

<table>
<thead>
<tr>
<th>Cluster Index (2 PCs)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>925</td>
<td>1100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>825</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>234</td>
</tr>
<tr>
<td>11</td>
<td>1104</td>
<td>1550</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1003</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>927</td>
<td>908</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>973</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>973</td>
<td>1097</td>
<td>26</td>
<td>116</td>
<td>0</td>
<td>3</td>
<td>70</td>
<td>1015</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>1095</td>
<td>839</td>
<td>0</td>
<td>836</td>
<td>1015</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>792</td>
<td>1165</td>
<td>0</td>
<td>0</td>
<td>384</td>
<td>453</td>
<td>451</td>
<td>295</td>
<td>172</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>123</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>575</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1150</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1025</td>
<td>809</td>
<td>0</td>
<td>959</td>
<td>1061</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>253</td>
<td>142</td>
<td>823</td>
<td>718</td>
<td>0</td>
<td>614</td>
<td>959</td>
<td>102</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>321</td>
<td>657</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>736</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>95</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1426</td>
</tr>
</tbody>
</table>
5.4.4 Temporal Stability

The factors evaluated in the previous sections were both concerned with stability of the clustering solution within the period of analysis considered. Beyond this, it is important to evaluate how the clustering identified generalizes across other time periods. As described in section 5.3.1, when comparing user clusters defined from data extracted on the same week of two different years, Ortega-Tong (2013) identified important temporal stability issues. They found that different clusters emerged from the two time periods. In order to examine this issue with respect to the approach we present, the transactions of another random sample of 18,712 cards are extracted from the period between October 20th and November 19th 2014. Similarly to the February-March period, this period spans 29 days and is aligned with the school half-term on the second week. For simplicity, we refer to the February-March sample as the February sample, and to the October-November sample as the October sample.

Figure 5-23: Temporal Stability Evaluation Approach

Figure 5-23 provides an overview of the approach used to compare the results of the February and October samples. First, the methodology applied to the February sample in sections 5.2.2 and 5.3.2 is independently applied to the October sample, as illustrated on the right hand side of Figure 5-23. This results in a set of independently
defined clusters for the October period. Second, October users are classified with respect to the February clusters, as illustrated in the central portion of Figure 5-23. This is done by projecting October travel sequences onto the February defined principal components and by mapping the resulting weights to the February cluster centroids.

October users are first clustered with respect to public transport usage patterns. The characteristics of the three resulting clusters closely match those of the February clusters as summarized in Table 5.9. The centroid of each cluster is the same across both periods. Overall the proportion of users assigned to each cluster is consistent, with small variations which may be explained by seasonal effects. Notably, the percentage of non-recurrent users is larger in October than in February. For example, this could reflect differences in visitor flows between the two seasons.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Period</th>
<th>Percent Users</th>
<th>Range (Days)</th>
<th>Travel Days</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Oct.</td>
<td>40</td>
<td>4</td>
<td>2</td>
<td>Non-Recurrent Users</td>
</tr>
<tr>
<td>2</td>
<td>Oct.</td>
<td>31</td>
<td>28</td>
<td>22</td>
<td>Frequent Users</td>
</tr>
<tr>
<td>3</td>
<td>Oct.</td>
<td>29</td>
<td>22</td>
<td>8</td>
<td>Occasional Users</td>
</tr>
<tr>
<td>1</td>
<td>Feb.</td>
<td>36</td>
<td>4</td>
<td>2</td>
<td>Non-Recurrent Users</td>
</tr>
<tr>
<td>2</td>
<td>Feb.</td>
<td>33</td>
<td>28</td>
<td>22</td>
<td>Frequent Users</td>
</tr>
<tr>
<td>3</td>
<td>Feb.</td>
<td>31</td>
<td>22</td>
<td>8</td>
<td>Occasional Users</td>
</tr>
</tbody>
</table>

Principal component analysis is then applied to the 5,724 frequent users identified from the October sample. The correlation between principal components from the October and February period is used to examine the temporal stability of components. Figure 5-24 illustrates the correlation coefficient between pairs of October and February components. Warm colors indicate high correlation. As expected, the correlation is highest for pairs of corresponding principal components, illustrated on the diagonal of Figure 5-24. Excluding the 4th and 5th components, the first 8 component pairs all have correlation near 1, indicating high stability. The 4th and 5th pairs have slightly lower correlation of 0.88. However, the correlation between the 4th and 5th off-diagonal elements matching these two components is significantly higher than other off-diagonal elements. This may indicate that the 4th and 5th PCs combined together capture the same component of variability across both time periods. The correlation between the 10th and 11th components is higher along the off-diagonal (i.e. between the 10th component of October and the 11th component of February and vice versa) than along the diagonal (e.g. between the 10th component of October and the 10th component of February). This indicates that the 10th October component is most similar to the 11th February component, and vice versa for the 11th October component. This switch could be explained by differences in the amount of variability explained by the two components.

Figure 5-25 compares the correlation of the diagonal elements of Figure 5-24 to the av-
Figure 5-24: PC Correlation Across Time Periods

Figure 5-25: PC Temporal Stability
verage principal component correlation obtained from the February bootstrap samples illustrated in Figure 5-20. The figure shows that the across period component stability is generally consistent with the within period stability for the first 8 components. The decrease in correlation for the 4th, 5th, 10th, and 11th components is explained by the increase in correlation of the off-diagonal elements as described above.

Having established that the principal components across both time periods are consistent, we examine the stability of the 11-cluster solution. As illustrated in Figure 5-23, each October user is classified twice. First, users are clustered with respect to October principal component weights using the K-means algorithm. Second, the same users are classified with respect to the February cluster centroids. Denote by $G_{Feb}^k$ the $k$th cluster of the February 11-cluster solution (superscript 11 is omitted for simplicity). Each user $u$ is assigned to the cluster $G_{Feb}^k$ which minimizes the distance between the cluster centroid $C_{Feb}^k$ and point $X_u$ representing the user (Equation 5.11).

$$k_u = \arg \min_{k \in K} \|C_{Feb}^k - X_u\|$$

(5.11)

where $X_u$ denotes the attribute vector describing user $u$, composed of the first 8 February principal component weights, $k_u$ denotes the cluster assigned to user $u$. This approach is equivalent to classifying October users according to the voronoi regions associated with cluster centroids $C_{Feb}^k$ lying in the two-dimensional space defined by the first 8 February principal components. As a result, each October user is assigned to two clusters, one from the February clustering, and another from the October clustering. Stability can then be evaluated from the overlap between the two assignments. The percentage of individuals which are assigned to matching clusters across both periods reveals how stable the clusters are.

Table 5.10 summarizes the distribution of users over the February and October clusters. Each column is associated with a cluster of the October partition and each row with a cluster of the February partition. October clusters are ordered so as to maximize the number of users assigned to the same cluster across both periods through the approach described in Section 5.4.3.

The value reported in the table corresponds to the percentage of users in a given October cluster also assigned to a given February cluster. For example, considering the intersection of the second row and the first column, 1.1% of all users assigned to cluster $G_{Oct}^1$ for the October clustering were assigned to cluster $G_{Feb}^2$ in the February clustering. Diagonal values indicate the degree of cluster stability. For instance, 99.4% of users in cluster $G_{Oct}^6$ were also classified in the equivalent February cluster, indicating that cluster 6 is 99.4% stable. Overall, 91% of all frequent users in the October sample were allocated to the same cluster across both periods. This high overlap indicates that the clustering is stable over time.

Some clusters are less stable than others. Notably, clusters 5 and 9 are least stable with 81% and 78.1% overlap, respectively. Clusters 6 and 7 are most stable, with around 99% of users in these clusters being assigned to the same cluster in February.
and October.

The overall size of each cluster is also stable over time. Table 5.11 summarizes the size of each cluster for different clusterings and user samples. Cluster size is indicated as a percentage of frequent users in the sample considered. The first row corresponds to the original clustering solution presented in Section 5.3.4. The second row shows the cluster size for the February clustering of October users, while the third shows the cluster size for the October clustering of October users. The proportion of frequent users included in each cluster is generally consistent across all three partitions.

### Table 5.10: February-October Clustering Overlap

<table>
<thead>
<tr>
<th>October Clustering of Oct. Users (%)</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Cluster 5</th>
<th>Cluster 6</th>
<th>Cluster 7</th>
<th>Cluster 8</th>
<th>Cluster 9</th>
<th>Cluster 10</th>
<th>Cluster 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>G^{Feb}_1</td>
<td>95.7</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>1.4</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
<td>1.3</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>G^{Feb}_2</td>
<td>1.1</td>
<td>95.8</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.9</td>
<td>0.3</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>G^{Feb}_3</td>
<td>0.0</td>
<td>2.0</td>
<td>96.0</td>
<td>1.1</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>1.4</td>
<td>0.2</td>
<td>5.3</td>
<td>0.0</td>
</tr>
<tr>
<td>G^{Feb}_4</td>
<td>0.6</td>
<td>0.1</td>
<td>0.0</td>
<td>87.9</td>
<td>1.2</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.4</td>
<td>0.3</td>
<td>2.9</td>
</tr>
<tr>
<td>G^{Feb}_5</td>
<td>1.1</td>
<td>1.6</td>
<td>3.3</td>
<td>0.3</td>
<td>81.0</td>
<td>0.0</td>
<td>0.3</td>
<td>1.7</td>
<td>5.0</td>
<td>1.2</td>
<td>0.0</td>
</tr>
<tr>
<td>G^{Feb}_6</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.6</td>
<td>2.3</td>
<td>99.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>3.4</td>
</tr>
<tr>
<td>G^{Feb}_7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.4</td>
<td>0.0</td>
<td>98.7</td>
<td>0.0</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>G^{Feb}_8</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>1.4</td>
<td>0.0</td>
<td>0.0</td>
<td>91.0</td>
<td>6.1</td>
<td>0.5</td>
<td>0.9</td>
</tr>
<tr>
<td>G^{Feb}_9</td>
<td>0.0</td>
<td>0.1</td>
<td>0.3</td>
<td>9.9</td>
<td>0.5</td>
<td>0.2</td>
<td>0.0</td>
<td>4.5</td>
<td>78.1</td>
<td>3.3</td>
<td>7.9</td>
</tr>
<tr>
<td>G^{Feb}_{10}</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.3</td>
<td>1.0</td>
<td>0.7</td>
<td>88.3</td>
<td>1.8</td>
</tr>
<tr>
<td>G^{Feb}_{11}</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>9.5</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>6.8</td>
<td>0.0</td>
<td>82.6</td>
</tr>
</tbody>
</table>

### Table 5.11: Relative Cluster Size Comparison - 11 Clusters

<table>
<thead>
<tr>
<th>Cluster (% of Frequent Users)</th>
<th>Clustering 1</th>
<th>Clustering 2</th>
<th>Clustering 3</th>
<th>Clustering 4</th>
<th>Clustering 5</th>
<th>Clustering 6</th>
<th>Clustering 7</th>
<th>Clustering 8</th>
<th>Clustering 9</th>
<th>Clustering 10</th>
<th>Clustering 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>February</td>
<td>14.6</td>
<td>14.2</td>
<td>10.4</td>
<td>5.2</td>
<td>7.4</td>
<td>9.9</td>
<td>5.9</td>
<td>5.3</td>
<td>8.6</td>
<td>9.1</td>
<td>9.4</td>
</tr>
<tr>
<td>October</td>
<td>13.8</td>
<td>13.9</td>
<td>11.1</td>
<td>6.0</td>
<td>7.5</td>
<td>9.6</td>
<td>5.6</td>
<td>5.5</td>
<td>8.3</td>
<td>9.4</td>
<td>9.4</td>
</tr>
<tr>
<td>October</td>
<td>14.1</td>
<td>14.2</td>
<td>10.5</td>
<td>6.2</td>
<td>7.5</td>
<td>9.0</td>
<td>5.5</td>
<td>5.1</td>
<td>8.0</td>
<td>10.2</td>
<td>9.7</td>
</tr>
</tbody>
</table>

119
5.5 Conclusion

In conclusion, this chapter describes a methodology to identify recurrent patterns of travel behavior from user’s travel-sequences. This methodology is organized in two steps. First users are grouped according to the distribution of their transit usage over time. This categorization results in three groups of usage: non-recurrent users, frequent users, and occasional users. Second, common elements of mobility and activity patterns are identified for frequent users. The extent to which these elements are observed in each user sequence is used to compare individuals and form clusters of behavior. Two stable sets of clusters are identified: one with 4 clusters and another with 11 clusters. These groupings are revealed to be stable across different time periods.
Chapter 6

Socio-Demographic Analysis

6.1 Introduction

The clusters identified in Chapter 5 describe groups of users with similar travel patterns. These similarities are defined with respect to within day and day-to-day components of realized travel. Realized traveled, as conceptualized by activity based travel theory (ABTT), is the result of a decision process connecting short-term trip and activity decisions to long-term life-choices. As described in Chapter 2 and summarized in Figure 6-1, ABTT implies important dependencies between long-term life decisions and realized travel, as long-term decisions control the constraints dictating short-term travel decisions. In line with those dependencies, we aim to examine the degree of association between the recurrent travel patterns observed in Chapter 5 and long-term life decisions captured through socio-demographic qualities. Specifically, we focus on socio-demographic attributes related to household attributes, occupation status and vehicle ownership, three broad categories of long-term life-choices reflecting elements of an individual’s lifestyle as discussed by Krizek and Waddell (2002) and Walker and Li (2007).

Understanding the association between the clusters identified in Chapter 5 and user demographics may serve two purposes. First, socio-demographic attributes may provide more information about the nature of activity schedule driving the travel patterns observed in each clusters. This would provide additional insight about the pattern associated with the different clusters. This insight could allow for hypotheses about the users contained in each cluster to be strengthened. For example, cluster 6 of the 11-cluster solution is supposed to contain primarily pupils who were on holiday of the second week of the analysis period. Demographic information about the occupation of users in this cluster could help confirm or counter this hypothesis. Second, in addition to enhancing the understanding of each cluster, evaluating the demographic composition of each cluster could highlight socio-demographic attributes most highly correlated with travel decisions captured by smart card transactions. From a data-mining perspective, this may point to useful demographic characteristics which can
accurately be inferred from smart card transactions. For example, observing that the travel of high-income young professionals tends to follow a given pattern would point to ways of identifying such individuals from smart card data. In turn, being able to infer different demographic attributes from smart card transactions may provide opportunities to improve customer information or marketing campaigns targeted at smart card users.

In order to relate demographics to clusters of travel pattern, we use data which allows for the smart card transactions of a given individual to be connected to her socio-demographic attributes. Under this disaggregate approach, the demographics of each cluster are derived from the characteristics of each individual contained in the cluster. We use two different data sources containing disaggregate information about socio-demographics and public transport travel. Section 6.2 examines the distribution of different smart card types used by individuals of specific demographic segments to access discounted fares in Transport for London’s (TfL) network. Section 6.3 connects information from a large scale annual travel diary survey in London, the London Travel Diary Survey (LTDS), to the smart card transactions of its respondents.

### 6.2 Oyster Card Type

As part of Transport for London’s complex pricing scheme, users of certain socio-demographic status are eligible for fare discounts of different types. In order to benefit from these discounts, users must provide proof that they belong to the eligible socio-demographic group to obtain a special fare card. For every smart card transaction, the card type and the discount applied for the transaction are recorded. These records
Table 6.1: Distribution of Card Types by Transit Usage Cluster

<table>
<thead>
<tr>
<th>Card Type</th>
<th>Non-Recurrent</th>
<th>Frequent</th>
<th>Occasional</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oyster</td>
<td>31087</td>
<td>22653</td>
<td>22014</td>
<td>75.8</td>
</tr>
<tr>
<td>Disabled Freedom Pass</td>
<td>270</td>
<td>684</td>
<td>561</td>
<td>1.5</td>
</tr>
<tr>
<td>Elderly Freedom Pass</td>
<td>2342</td>
<td>2905</td>
<td>4429</td>
<td>9.7</td>
</tr>
<tr>
<td>Visitor Oyster</td>
<td>753</td>
<td>24</td>
<td>82</td>
<td>0.9</td>
</tr>
<tr>
<td>Student Photocards</td>
<td>1936</td>
<td>6760</td>
<td>3425</td>
<td>12.1</td>
</tr>
</tbody>
</table>

Total 36,388 33,026 30,511

offer useful information about the socio-demographics of users subscribed to such discounts. We focus on three card types requiring proof of identity:


- Elderly Freedom Passes awarded to senior London residents also allowing for free travel across London (London Councils, 2015b).

- Oyster photocards for students and pupils are awarded to students in full-time education. As students under 18 are entitled to specific discounts, student photocards differentiate between students 18 and above and students younger than 18.

In addition to these card types, TfL also offers a card service for tourists, which allows for special visitor cards to be ordered online and mailed to overseas addresses. These cards offer no special discount and can be used by any users, including non-visitors, but are primarily used by tourists.

Table 6.1 summarizes the distribution of different card types by cluster for the 99,925 user sample previously described. As illustrated by the table, the large majority of cards observed during the 4-week period of analysis are ordinary Oyster cards. Students and elderly cards respectively make up the second and third largest proportion of cards, with a share of approximately 10%. Visitor Oyster cards make up less than 1 percent of cards sampled. The low number visitor cards is expected for the period of analysis, which does not align with a high tourism season.
6.2.1 Measure of Association

In order to evaluate correlation between cluster membership and different socio-demographic indicators, such as those summarized in Table 6.1, it is necessary to account for variations in cluster sizes and in demographic segment sizes. All else being equal, an individual in a given demographic segment is more likely to belong to a larger cluster because this cluster contains a larger portion of all users, irrespective of the card they use. Similarly an individual in a given cluster is more likely to belong to a larger demographic segment because the segment contains a larger percentage of the population. For example, in Table 6.1, only 2% of non-recurrent users used a visitor oyster card, but 88% of visitor cards are non-recurrent users. Focusing only on non-recurrent users, visitor cards would appear to make up a negligible portion of non-recurrent users unless the proportion of users in each card type segment was accounted for. Conversely, focusing only on the percentage of visitor cards classified as non-recurrent users would fail to account for the fact that non-recurrent users make up a larger proportion of the overall population, irrespective of card type. Hence, association should be measured using a metric which controls for the probability of belonging to a cluster, independent of card type, and for the probability of belonging to a demographic segment, independent of cluster membership.

The odds ratio provides such a metric. It is a measure of association between two factors in a population which controls for the relative probability of both factors across the population. As indicated in Equation 6.1, the odds ratio between a cluster \( k \) and a discrete demographic characteristic \( i \) is the ratio of the odds of having characteristic \( i \) given membership to \( k \) over the odds of having characteristic \( i \) given membership to any other cluster.

\[
OR_{i,k} = \frac{P(i|k)/P(i'|k)}{P(i|k')/P(i'|k')} \tag{6.1}
\]

where \( P(i|k) \) denotes the probability of having characteristic \( i \) given membership to cluster \( k \). All clusters other than \( k \) are aggregated as \( k' \), indicating the user does not belong to \( k \), and all demographics other than \( i \) are aggregated as \( i' \), indicating the user does not have demographic attribute \( i \). The ratio measures how much more (or less) likely a user who belongs to cluster \( k \) is to have characteristic \( i \) compared to a user who does not belong to \( k \). In line with Bayes' theorem, this ratio is also equivalent to the ratio of the odds of belonging to \( k \) given characteristic \( i \) over the odds of belonging to \( k \) given the absence of characteristic \( i \), such that \( OR_{i,k} = OR_{k,i} \). In other words, the odds ratio also measures how much more likely a user with characteristic \( i \) is to belong to cluster \( k \) compared to a user without characteristic \( i \). An odds ratio below 1 indicates negative association (e.g. less likely to belong to cluster \( k \) given characteristic \( i \)), while an odds ratio above 1 indicates positive association (more likely to belong to cluster \( k \) given characteristic \( i \)). An odds ratio equal to 1 indicates no association (the probability of belonging to cluster \( k \) is independent of characteristic \( i \)).
Given a sample from the user population, the odds ratio can be estimated according to Equation 6.2.

\[
\hat{OR}_{i,k} = \frac{N_{i,k} \cdot N_{i',k'}}{N_{i',k} \cdot N_{i,k'}}
\]  

(6.2)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Characteristic</th>
<th>( i )</th>
<th>( i' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )</td>
<td>( N_{i,k} )</td>
<td>( N_{i',k} )</td>
<td></td>
</tr>
<tr>
<td>( k' )</td>
<td>( N_{i,k'} )</td>
<td>( N_{i',k'} )</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2: Odds Ratio - Cluster and Characteristic

where \( N_{i,k} \) denotes the number of users with characteristic \( i \) in cluster \( k \), and \( N_{i',k'} \) denotes the number of users in any cluster other than \( k \), and with any characteristic other than \( i \). Table 6.2 illustrates the counts used in Equation 6.2. For a population where \( i \) and \( k \) are binomially distributed, the odds ratio \( OR_{i,k} \) follows a log-normal distribution. Hence, the natural logarithm of \( OR_{i,k} \) is normally distributed (Rice, 2006). The test statistic \( Z_{i,k} \) defined by Equation 6.3 is approximately normally distributed and is used to test whether the null hypothesis that \( OR_{i,k} \) equals 1 can be rejected for a given confidence level (Morris and Gardner, 1988).

\[
Z_{i,k} = \frac{\ln (\hat{OR}_{i,k})}{1/N_{i,k} + 1/N_{i',k} + 1/N_{i,k'} + 1/N_{i',k'}}
\]  

(6.3)

The odds ratio is used to measure the association between card type and public transit usage cluster membership, as well as between card type and cluster membership for the 11 cluster solution presented in Chapter 5.

6.2.2 Card Type Odds Ratio Analysis

Table 6.3 summarizes the odds ratio between card types and transit usage clusters. Values highlighted in blue indicate a positive association, significant at the 95% confidence level. Values highlighted in red indicated a significant negative association. All odds ratios reported in bold are statistically significant at the 95% confidence level. As seen in the table, the strongest association observed is between visitor oysters and non-recurrent users. The odds ratio of 12.64 indicates that visitors are 12.64 times more likely than non-visitors to be classified in the non-recurrent cluster. The odds ratio of 2.48 between normal Oyster cards and the non-recurrent cluster indicates that standard Oyster card users are more likely than users of disabled, elderly, and student cards to be classified as non-recurrent users. This is intuitive as disabled, elderly, and student cards users must hold on to their card in order to retain their fare...
Table 6.3: Odds-Ratio Card Type - Transit Usage Clusters

<table>
<thead>
<tr>
<th>Card Type</th>
<th>Cluster</th>
<th>Non-Recur.</th>
<th>Frequent</th>
<th>Occasional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oyster</td>
<td></td>
<td>2.48</td>
<td>0.57</td>
<td>0.76</td>
</tr>
<tr>
<td>Disabled Freedom</td>
<td></td>
<td>0.37</td>
<td>1.68</td>
<td>1.34</td>
</tr>
<tr>
<td>Elderly Freedom</td>
<td></td>
<td>0.53</td>
<td>0.86</td>
<td>2.08</td>
</tr>
<tr>
<td>Visitor Oysters</td>
<td></td>
<td>12.64</td>
<td>0.06</td>
<td>0.24</td>
</tr>
<tr>
<td>Student Photocards</td>
<td></td>
<td>0.29</td>
<td>2.95</td>
<td>0.88</td>
</tr>
</tbody>
</table>

OR_{i,j} indicates significance at the 95% Confidence Level

Table 6.4: Odds-Ratio Card Type - 11 Clusters

<table>
<thead>
<tr>
<th>Category</th>
<th>Cluster</th>
<th>$G_1^{11}$</th>
<th>$G_2^{11}$</th>
<th>$G_3^{11}$</th>
<th>$G_4^{11}$</th>
<th>$G_5^{11}$</th>
<th>$G_6^{11}$</th>
<th>$G_7^{11}$</th>
<th>$G_8^{11}$</th>
<th>$G_9^{11}$</th>
<th>$G_{10}^{11}$</th>
<th>$G_{11}^{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oyster</td>
<td></td>
<td>3.87</td>
<td>2.66</td>
<td>0.59</td>
<td>1.46</td>
<td>0.36</td>
<td>0.18</td>
<td>4.04</td>
<td>1.89</td>
<td>0.84</td>
<td>1.12</td>
<td>0.70</td>
</tr>
<tr>
<td>Disabled</td>
<td></td>
<td>0.18</td>
<td>0.16</td>
<td>3.10</td>
<td>1.24</td>
<td>2.69</td>
<td>0.23</td>
<td>0.25</td>
<td>0.70</td>
<td>1.73</td>
<td>1.35</td>
<td>1.25</td>
</tr>
<tr>
<td>Elderly</td>
<td></td>
<td>0.36</td>
<td>0.29</td>
<td>2.31</td>
<td>0.86</td>
<td>5.21</td>
<td>0.22</td>
<td>0.32</td>
<td>0.67</td>
<td>1.68</td>
<td>0.60</td>
<td>1.56</td>
</tr>
<tr>
<td>Students</td>
<td>&lt;17</td>
<td>0.14</td>
<td>0.35</td>
<td>0.52</td>
<td>0.37</td>
<td>0.85</td>
<td>20.15</td>
<td>0.20</td>
<td>0.32</td>
<td>0.61</td>
<td>0.67</td>
<td>1.48</td>
</tr>
<tr>
<td>Students</td>
<td>≥18</td>
<td>0.64</td>
<td>0.91</td>
<td>1.91</td>
<td>1.08</td>
<td>1.17</td>
<td>0.48</td>
<td>0.47</td>
<td>0.95</td>
<td>1.31</td>
<td>1.61</td>
<td>0.85</td>
</tr>
</tbody>
</table>

OR_{i,j} indicates significance at the 95% confidence level

discount. In contrast, a large portion of normal Oyster cards are only used on a single day. Both disabled and student cards are positively associated with the frequent user cluster. Users of student cards are 3 times more likely than other users to be classified as frequent users. Disabled cards are most strongly associated with frequent users and have some positive correlation with occasional users. Finally, elderly cards are positively associated to the occasional user cluster.

Table 6.4 summarizes the odds ratio between card types and the 11 cluster solution presented in Section 5.3.2 for the 33,026 users classified in the frequent user cluster. Student cards used by users 18 and over and under 18 are differentiated for this more detailed comparison. As summarized in the table, clusters 1, 2, and 7, marked by clear working days all have strong positive correlation to the normal oyster card category and negative correlation with all other card types. Clusters $G_3^{11}$, $G_5^{11}$, and $G_9^{11}$ are all positively associated with disabled, elderly and 18+ student cards. Cluster $G_3^{11}$ is most strongly associated with disabled freedom passes while cluster 5 is most strongly
correlated to elderly passes. The odds ratio of 20.15 between cluster \(G_0^{11}\) and students under 17 indicates that students under 17 are 20 times more likely to be classified in cluster \(G_0^{11}\) than users of other card types. The correlation pattern for clusters \(G_4^{11}\) and \(G_8^{11}\) is less marked, with some degree of positive association with standard oyster cards. Cluster \(G_{10}^{11}\) has highest positive association with users of adult student cards, and some degree of association with disabled freedom cards. Finally, cluster \(G_{11}^{11}\) has mild association with elderly freedom cards and under 18 student cards.

These associations generally align with the travel pattern characteristics of each cluster. In order to move beyond the coarse demographics provided by card type, the next section presents a more detailed source of demographic information which is used in subsequent sections to explore the relationship between individual characteristics and corresponding travel patterns.

### 6.3 London Travel Diary Survey

The London Travel Diary Survey (LTDS) is a continuous household survey focused on the travel of Greater London residents. The survey, conducted since April 2005, is operated on an annual basis, with each survey-cycle running from April to March of the following year (Transport for London, 2011). On each survey year, a random sample of approximately 8,000 households, including approximately 19,000 individuals, is interviewed face-to-face. The interview covers questions related to characteristics of the household and household members above 5 years old, as well as questions regarding one day of travel for each household member. These questions provide socio-demographic information about the household and its members, and about the mode, travel time, duration, purpose, origin, and destination of each journey-stage completed by household members. Personal information about household members below 5 years old is collected by proxy through an adult household member.

Since April 2011, LTDS respondents over 16 years old are asked, on a voluntary basis, to provide the ID number of the two Oyster cards they use most frequently. This allows for the reported public transport travel of a respondent to be compared to the transaction data recorded by Transport for London for the corresponding card. Riegel (2013) analyzed the discrepancies between the reported and observed travel for a sample of 4,000 LTDS respondents interviewed between July 2011 and March 2012. She found that approximately half of journey stages observed from smart card data had matching LTDS journey stages and that less than half of LTDS stages had matching smart card records. Her work highlighted the value of combining objectively collected transaction records with individual survey responses. Beyond journey-level information, LTDS also allows for useful socio-demographic information to be related to the observed travel behavior of a respondent. While smart card transactions provide objective data on the characteristics of public transport journeys, the information they provide about user socio-demographics is limited, as detailed in Section 6.2. In this section we relate the clusters identified in Chapter 5 to the socio-demographics...
of LTDS respondents who volunteered their smart card ID number.

6.3.1 Sample Description

The data available for this research includes the 2011-2012 and the 2012-2013 LTDS surveys. The 2011-2012 survey was answered by 8053 households containing 19632 individuals, while the 2012-2013 survey was answered by 8156 households and 19438 individuals. Of these 39,070 individuals, 12,089 provided at least one Oyster card number. In line with the period of analysis previously defined, we focus on cards with at least one observed transaction between February 10th and March 10th 2014. This sample excludes all card IDs which were provided by more than one respondent (e.g. same card used by multiple members of the same household). Where more than one card ID was provided by the same respondent, we select the first ID provided or the ID not provided by other members of the same household where relevant. The resulting sample includes 5,713 individuals distributed over 4,732 households.

Figures 6-2a and 6-2b compare the distribution of days traveled and spread of days traveled, respectively, for the LTDS sample and the random oyster sample used in Chapter 5. Overall, the distribution of days traveled for the LTDS respondents is similar to that of the general oyster population, with a smaller percentage of LTDS users traveling on fewer than 5 days. This is expected as the LTDS survey only includes London Area residents, excluding non-residents who are likely to travel on fewer days than residents. This is also reflected in the distribution of the spread of days traveled, with a larger percentage of LTDS users seen over a range larger than 20 days and a higher percentage of users from the random oyster sample having a spread lower than 5 days.

These two figures hint at issues of representativeness encountered with the sample available for this analysis. Specifically, bias in the sample of Oyster IDs available for LTDS respondents is introduced at two levels. First, the population of individuals sampled for the survey does not align with the population of oyster users. Indeed, a number of Oyster users reside outside the catchment area of LTDS. This could include users living outside of the Greater London area who frequently use London’s public transport, or visitors from the UK or from overseas who occasionally visit London. Additionally, LTDS samples individuals for all modes of transportation, and is not limited to public transport users. Second, the portion of LTDS interviewees who provide their Oyster ID is not representative of the overall LTDS sample. Respondents younger than 17 years old are not asked to provide their ID, and only a portion of those asked agree to disclose this personal information. The demographic characteristics of a respondent may correlate with her willingness to share personal information. Figure 6-3, for example, shows that the percentage of elderly respondents who agree to share their card ID is disproportionately high compared to the percentage of elderly interviewed for the survey. This may reflect the fact that elderly respondents are more open to sharing their personal information than respondents in other demographic segments.
Figure 6-2: Days Traveled and Spread of Days Traveled Distribution - LTDS Sample and Random Oyster Sample
Bias with respect to the general Oyster user population and with respect to LTDS respondents willing to provide their card ID will result in specific portions of Oyster users being systematically over or under-represented. Such biases are likely to affect user clusters differently, as certain clusters contain different proportions of individuals being over or under represented. For example, a cluster presumed to contain mostly students and pupils would be significantly under-represented, as Oyster IDs are unavailable for users under 17. As a result, the analysis presented in this section should be interpreted in light of possible biases related either to the discrepancy between Oyster users and the population sampled for LTDS, or to the discrepancy between general LTDS respondents and LTDS respondents who accept to provide their smart card information. In line with the exploratory nature of the analysis, these biases are not corrected for explicitly (for example through weighting).

### 6.3.2 Classification of LTDS Respondents

In order to relate the socio-demographic information provided by LTDS to transaction data, the 5,713 respondents for whom smart card transactions were observed during the period of analysis are classified according to the clusters identified in Chapter 5. First, each respondent is assigned to one of the three public transport usage clusters: non-recurrent, frequent, and occasional users (see Section 5.2.2). For this purpose the attribute vector $X_u$ used in Section 5.2.2 is computed for each user $u$, and users are assigned to a cluster $k \in K$ by comparing cluster centroids $G_k$ to the attributes of each user (Equation 6.4).
where $X_u$ denotes the attribute vector composed of the number of days traveled $D_u$ and the spread of days traveled $S_u$ by user $u$, and $k_u$ denote the cluster assigned to $u$.

This approach is equivalent to classifying LTDS respondents according to the voronoi regions associated with cluster centroids $G_k$ lying in the two-dimensional space defined by $D$ and $S$. Table 6.5 presents the results of this assignment. As compared to the random oyster sample used in Chapter 5, a larger proportion of LTDS respondents are classified as occasional users and correspondingly, a lower proportion are assigned to the non-recurrent user cluster. This is expected as the population sampled for LTDS only includes residents of the London area as previously discussed. 1,973 LTDS respondents are classified as frequent users, corresponding to a similar proportion of the random Oyster sample.

Table 6.5: LTDS Transit Usage Classification

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Non-Recurrent</th>
<th>Frequent</th>
<th>Occasional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Oyster (%)</td>
<td>36</td>
<td>33</td>
<td>31</td>
</tr>
<tr>
<td>LTDS Oyster (%)</td>
<td>25</td>
<td>34</td>
<td>41</td>
</tr>
</tbody>
</table>

The 1,973 frequent users are then mapped onto the detailed clusters identified in Section 5.3.2. Each individual’s binary travel sequence $b_u$ is computed and projected onto the space defined by the first 8 principal components $v_1, v_2, \ldots, v_8$ computed from the 99,925-user random sample. Hence, each individual is represented by an 8 element feature vector $W_u$. This vector is compared to the centroid of each cluster according to Equation 6.4, substituting $X_u$ by $W_u$. As for the public transport usage clusters, each respondent is assigned to the cluster centroid $G_k$ it lies closest to. Table 6.6 and Table 6.7 summarize the distribution of LTDS respondents for the 4-cluster and 11-cluster solutions respectively.

The difference between the proportion of LTDS respondents and the proportion of users from the random sample is largest for groups $G^4_2$ and $G^4_5$, and for groups $G^{11}_2$, $G^{11}_5$ and $G^{11}_6$. These differences reflect the biases previously discussed.

Having classified the observed travel behavior of each LTDS respondent, we present a descriptive overview of the socio-demographic characteristics of each cluster. For the reminder of this chapter, we focus on the 11-cluster solution to analyze demographic associations. These clusters contain more homogeneous activity patterns which are more likely to align with particular demographic segments than the more aggregated 4-cluster solution.
Table 6.6: LTDS Travel Sequence Classification (4 Clusters)

<table>
<thead>
<tr>
<th>Sample</th>
<th>$G_1^4$</th>
<th>$G_2^4$</th>
<th>$G_3^4$</th>
<th>$G_4^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Oyster (%)</td>
<td>23.7</td>
<td>32.4</td>
<td>29.1</td>
<td>14.8</td>
</tr>
<tr>
<td>LTDS Oyster (%)</td>
<td>25.9</td>
<td>36.9</td>
<td>26.4</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Table 6.7: LTDS Travel Sequence Classification (11 Clusters)

<table>
<thead>
<tr>
<th>Sample</th>
<th>$G_1^{11}$</th>
<th>$G_2^{11}$</th>
<th>$G_3^{11}$</th>
<th>$G_4^{11}$</th>
<th>$G_5^{11}$</th>
<th>$G_6^{11}$</th>
<th>$G_7^{11}$</th>
<th>$G_8^{11}$</th>
<th>$G_9^{11}$</th>
<th>$G_{10}^{11}$</th>
<th>$G_{11}^{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Oyster (%)</td>
<td>14.6</td>
<td>14.2</td>
<td>10.4</td>
<td>5.2</td>
<td>7.4</td>
<td>9.9</td>
<td>5.9</td>
<td>5.3</td>
<td>8.6</td>
<td>9.1</td>
<td>9.4</td>
</tr>
<tr>
<td>LTDS Oyster (%)</td>
<td>17.5</td>
<td>10.9</td>
<td>9.7</td>
<td>4.4</td>
<td>13.4</td>
<td>5.1</td>
<td>5.9</td>
<td>4.9</td>
<td>9.2</td>
<td>6.7</td>
<td>12.2</td>
</tr>
</tbody>
</table>

Table 6.8 summarizes the average socio-demographic characteristics of the 11 clusters. The table includes 5 variables, each related to a different socio-demographic dimension.

- **Median age** The highest median age of 64 is associated with cluster $G_5^{11}$, closely followed by cluster $G_2^{11}$ with a median age of 56 years old. Clusters $G_2^{11}, G_7^{11},$ and $G_{10}^{11}$ are associated with the youngest median ages of 35, 33, and 35 respectively. Cluster $G_2^{11}$ and Cluster $G_7^{11}$ are characterized by distinct working days and by more complex activity patterns than group $G_1^{11}$. Notably, users $G_2^{11}$ are observed to perform activities outside their secondary area on weekends. Group $G_{10}^{11}$ is characterized by the most complex travel pattern, with a high percentage of discontinuities and a high number of journeys on both weekdays and weekends. These travel patterns align with younger demographics.

- **Percent of full-time employed users** Clusters $G_1^{11}, G_2^{11}$ and $G_7^{11}$ have the highest percentage of full-time employed users, while clusters $G_3^{11}$ and $G_5^{11}$ contain the smallest proportion of full time workers. These two clusters are characterized by the absence of distinct working days and by short-activities outside the primary area, which aligns with the low proportion of full-time employed users they contain. In contrast, $G_1^{11}, G_2^{11}$ and $G_7^{11}$ were associated with distinct working days.

- **Median Income Bracket** In line with the low rate of full-time employment, Clusters $G_3^{11}$ and $G_5^{11}$ are also associated with the lowest median income, between £10,000 and £15,000. Clusters $G_1^{11}$ and $G_7^{11}$ have highest median income bracket, between £50,000 and £75,000.
Table 6.8: Descriptive Overview of Cluster Demographics

<table>
<thead>
<tr>
<th>Variable</th>
<th>$G_{1}^{11}$</th>
<th>$G_{2}^{11}$</th>
<th>$G_{3}^{11}$</th>
<th>$G_{4}^{11}$</th>
<th>$G_{5}^{11}$</th>
<th>$G_{6}^{11}$</th>
<th>$G_{7}^{11}$</th>
<th>$G_{8}^{11}$</th>
<th>$G_{9}^{11}$</th>
<th>$G_{10}^{11}$</th>
<th>$G_{11}^{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>37</td>
<td>35</td>
<td>56</td>
<td>40</td>
<td>64</td>
<td>42</td>
<td>33</td>
<td>36</td>
<td>52</td>
<td>35</td>
<td>48</td>
</tr>
<tr>
<td>Emp. Full-Time (%)</td>
<td>78</td>
<td>69</td>
<td>20</td>
<td>43</td>
<td>14</td>
<td>50</td>
<td>83</td>
<td>54</td>
<td>26</td>
<td>46</td>
<td>34</td>
</tr>
<tr>
<td>Children (%)</td>
<td>27</td>
<td>20</td>
<td>18</td>
<td>24</td>
<td>13</td>
<td>41</td>
<td>25</td>
<td>18</td>
<td>18</td>
<td>19</td>
<td>27</td>
</tr>
<tr>
<td>Vehicle (%)</td>
<td>70</td>
<td>38</td>
<td>18</td>
<td>24</td>
<td>13</td>
<td>41</td>
<td>25</td>
<td>18</td>
<td>18</td>
<td>19</td>
<td>27</td>
</tr>
</tbody>
</table>

- **Percent of users in households with children.** The proportion of users in households with children is about twice as high for cluster $G_{6}^{11}$ compared to other clusters. 41% of individuals in this cluster belong to households with one or more children. This group was characterized by a reduction in travel during the school half-term week, indicating some affiliation with the school calendar.

- **Percent of users in households with access to a vehicle** Around 70% of respondents in clusters $G_{1}^{11}$ and $G_{7}^{11}$ live in households with access to one or more vehicles, in contrast to only a quarter of those in clusters $G_{3}^{11}$ and $G_{10}^{11}$. This aligns with the low number of weekend days traveled by users in $G_{1}^{11}$ and $G_{7}^{11}$, who may use car modes for non-work-related activities.

These characteristics suggest that interesting trends may be exist with regards to the demographics of each cluster. In order to investigate these associations in more detail, the next five sections each focus on one of the five different demographic dimensions explored in Table 6.8. Section 6.3.3 discusses occupation status, Section 6.3.4 focuses on age, Section 6.3.5 discusses household income, Section 6.3.6 discusses household composition, and Section 6.3.7 discusses vehicle ownership and access.

### 6.3.3 Occupation Status

In order to examine the correlation between cluster membership and occupation status, LTDS respondents are grouped into 9 distinct occupation categories: full-time employees, part-time employees, self-employed individuals, students and pupils, unemployed individuals, individuals who are unable to work due to disability or illness, retirees, individuals looking after the home or family, and other occupations. The self-employed category includes both part-time and full-time self-employed individuals. Unemployed individuals include both those looking for employment and waiting to take-up employment. The ‘other’ category includes a range of other occupations stated by respondents, for example, ‘regular voluntary work’, ‘caring for ill family member’, or ‘awaiting visas’.
Table 6.9: Distribution of Respondents by Occupation Status

<table>
<thead>
<tr>
<th>Occupation</th>
<th>LTDS (%) 2011-2013</th>
<th>LTDS Oyster (%) All Users</th>
<th>LTDS Oyster (%) Frequent Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emp. Full-Time</td>
<td>38</td>
<td>36</td>
<td>46</td>
</tr>
<tr>
<td>Emp. Part-Time</td>
<td>8</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Self-Emp.</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Student/Pupil</td>
<td>9</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Unemployed</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Unable to work</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Retired</td>
<td>19</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>Stay-Home</td>
<td>7</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.9 shows the occupation distribution for all LTDS respondents in the first column, for the 5,713 LTDS respondents with Oyster ID in the second column, and for the 1,973 LTDS respondents classified as frequent users in the third column. The first column includes all LTDS respondents 16 or older interviewed between 2011 and 2013. As previously discussed, the LTDS Oyster sample includes only respondents 17 or older. The table shows that students 16 or older are under-represented in the LTDS Oyster sample. Furthermore, respondents below 16 were not asked for their occupation status in the survey. Hence, while 9% of LTDS respondents 16 and older are students or pupils, pupils likely make up a larger percentage of all LTDS interviewees as a large proportion of respondents below 16 are probably pupils. Retired respondents are generally over-represented in the LTDS Oyster sample, but less for the portion of this sample classified as frequent users.

Table 6.10 summarizes the odds ratio between occupation status and cluster membership for the 11-cluster solution. For simplicity clusters are denoted by their index only (i.e. $G_i^{11}$ is denoted by $i$). Values of odds ratio significant at the 95% confidence level are indicated in boldface. Significant positive association is highlighted in light blue, while significant negative association is highlighted in light red. Clusters 1, 2, and 7, have strong positive correlation with full-time employment and strongest negative correlation with being retired. Cluster 1 is also negatively correlated with all occupation categories which exclude some form of employment. The odds ratio of 5.26 indicates that full-time employed individuals are 526% more likely to be classified in cluster 1 than individuals in other occupation statuses. Clusters 3 is most strongly correlated with being unable to work, and is also correlated with part-time employees and retirees. Clusters 4 and 8 have no significant association with occupation other than a negative association with retired users for cluster 8. Cluster 5 is
### Table 6.10: Odds-Ratio Occupation Status - 11 Clusters

<table>
<thead>
<tr>
<th>Occupation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emp. Full-Time</td>
<td>5.26</td>
<td>2.85</td>
<td>0.26</td>
<td>0.87</td>
<td>0.15</td>
<td>1.14</td>
<td>6.14</td>
<td>1.35</td>
<td>0.36</td>
<td>0.97</td>
<td>0.54</td>
</tr>
<tr>
<td>Emp. Part-Time</td>
<td>0.67</td>
<td>0.73</td>
<td>1.76</td>
<td>1.03</td>
<td>1.11</td>
<td>2.30</td>
<td>0.62</td>
<td>1.16</td>
<td>0.76</td>
<td>0.72</td>
<td>1.06</td>
</tr>
<tr>
<td>Self-Emp.</td>
<td>1.06</td>
<td>1.15</td>
<td>0.51</td>
<td>0.67</td>
<td>0.57</td>
<td>1.22</td>
<td>0.35</td>
<td>1.66</td>
<td>1.42</td>
<td>0.91</td>
<td>1.67</td>
</tr>
<tr>
<td>Student/Pupil</td>
<td>0.46</td>
<td>0.92</td>
<td>1.19</td>
<td>1.46</td>
<td>0.90</td>
<td>1.45</td>
<td>0.84</td>
<td>1.04</td>
<td>1.27</td>
<td>1.84</td>
<td>0.90</td>
</tr>
<tr>
<td>Unemployed</td>
<td>0.45</td>
<td>0.90</td>
<td>0.65</td>
<td>1.92</td>
<td>0.99</td>
<td>1.60</td>
<td>0.35</td>
<td>0.89</td>
<td>1.26</td>
<td>2.27</td>
<td>1.22</td>
</tr>
<tr>
<td>Unable to Work</td>
<td>0.17</td>
<td>0.00</td>
<td>4.47</td>
<td>0.00</td>
<td>2.51</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.28</td>
<td>1.42</td>
<td>0.87</td>
</tr>
<tr>
<td>Retired</td>
<td>0.04</td>
<td>0.08</td>
<td>2.70</td>
<td>1.15</td>
<td>5.60</td>
<td>0.24</td>
<td>0.07</td>
<td>0.45</td>
<td>2.34</td>
<td>0.67</td>
<td>1.40</td>
</tr>
<tr>
<td>Stay-Home</td>
<td>0.34</td>
<td>1.06</td>
<td>1.52</td>
<td>1.06</td>
<td>1.25</td>
<td>1.13</td>
<td>0.18</td>
<td>0.68</td>
<td>1.15</td>
<td>1.04</td>
<td>1.80</td>
</tr>
<tr>
<td>Other</td>
<td>0.20</td>
<td>0.74</td>
<td>0.84</td>
<td>0.95</td>
<td>1.71</td>
<td>0.00</td>
<td>0.00</td>
<td>2.81</td>
<td>1.43</td>
<td>0.61</td>
<td>2.42</td>
</tr>
</tbody>
</table>

\( \text{OR}_{i,j} \) indicates significance at the 95% confidence level

Most strongly correlated with retirees, and is also positively associated with users who are unable to work due to illness or disability. It is also negatively associated with full-time employment. Cluster 6 has strong negative correlation with retired users and some level of positive correlation with part-time employed users. Clusters 9 has almost equal positive association with retired users and users who are unable to work. Cluster 10 is positively associated with unemployed users. Cluster 11 has strongest positive correlation with stay-home users, with slightly weaker positive correlation with self-employed and retired individuals.

#### 6.3.4 Age

The association between user age and cluster membership is evaluated by discretizing respondent age in 10 categories spanning 5 years each. LTDS data records respondent age to the nearest year at the time at of the interview. In order to account for time elapsed between the interview year and the 2014 period of analysis considered, the difference between the survey year and the period of analysis is added to the age of each respondent. Hence, the age of all respondents interviewed in year 2011-2012 is incremented by 2 years, and the age of all respondents interviewed in year 2012-2013 is incremented by 1 year. This coarse adjustment may result in the age of respondents being over and under estimated. This is assumed to be negligible with respect to the age categories selected. Note that the adjustment will shift the age of the youngest respondents from 17 to 18.
Table 6.11: Odds-Ratio Age Categories - 11 Clusters

<table>
<thead>
<tr>
<th>Age Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 19</td>
<td>0.78</td>
<td>1.36</td>
<td>1.55</td>
<td>3.68</td>
<td>0.00</td>
<td>3.10</td>
<td>0.00</td>
<td>0.00</td>
<td>1.66</td>
<td>2.35</td>
<td>0.00</td>
</tr>
<tr>
<td>20-24</td>
<td>0.87</td>
<td><strong>2.15</strong></td>
<td>0.47</td>
<td>1.37</td>
<td>0.50</td>
<td>0.62</td>
<td>1.13</td>
<td>1.40</td>
<td>0.89</td>
<td><strong>2.11</strong></td>
<td>0.70</td>
</tr>
<tr>
<td>25-29</td>
<td>1.25</td>
<td>1.79</td>
<td>0.71</td>
<td>1.05</td>
<td><strong>0.31</strong></td>
<td>1.08</td>
<td>1.60</td>
<td>1.62</td>
<td>0.82</td>
<td><strong>2.00</strong></td>
<td><strong>0.35</strong></td>
</tr>
<tr>
<td>30-34</td>
<td>1.28</td>
<td><strong>1.54</strong></td>
<td><strong>0.49</strong></td>
<td>1.27</td>
<td><strong>0.39</strong></td>
<td>0.76</td>
<td><strong>3.04</strong></td>
<td>1.09</td>
<td><strong>0.57</strong></td>
<td>1.46</td>
<td>0.75</td>
</tr>
<tr>
<td>35-39</td>
<td><strong>2.49</strong></td>
<td>1.42</td>
<td><strong>0.36</strong></td>
<td>1.03</td>
<td><strong>0.27</strong></td>
<td>0.95</td>
<td>0.97</td>
<td>1.58</td>
<td>0.63</td>
<td>0.77</td>
<td>0.93</td>
</tr>
<tr>
<td>40-44</td>
<td><strong>1.63</strong></td>
<td>0.98</td>
<td><strong>0.50</strong></td>
<td>0.58</td>
<td>0.63</td>
<td><strong>2.19</strong></td>
<td>1.69</td>
<td>0.62</td>
<td>0.73</td>
<td>0.69</td>
<td>1.13</td>
</tr>
<tr>
<td>45-49</td>
<td>1.44</td>
<td>1.29</td>
<td>0.90</td>
<td>0.74</td>
<td>0.70</td>
<td>1.50</td>
<td>0.62</td>
<td>1.29</td>
<td>0.82</td>
<td>0.63</td>
<td>1.00</td>
</tr>
<tr>
<td>50-54</td>
<td>1.40</td>
<td>0.94</td>
<td>1.39</td>
<td>1.06</td>
<td>0.78</td>
<td>1.57</td>
<td>0.54</td>
<td>0.43</td>
<td>0.65</td>
<td>0.94</td>
<td>1.10</td>
</tr>
<tr>
<td>55-59</td>
<td>1.01</td>
<td>1.17</td>
<td>0.50</td>
<td>0.70</td>
<td>0.66</td>
<td>2.00</td>
<td>1.69</td>
<td>0.61</td>
<td>1.75</td>
<td>0.60</td>
<td>0.92</td>
</tr>
<tr>
<td>≥ 60</td>
<td><strong>0.10</strong></td>
<td><strong>0.12</strong></td>
<td><strong>3.04</strong></td>
<td>1.03</td>
<td><strong>5.39</strong></td>
<td><strong>0.30</strong></td>
<td><strong>0.12</strong></td>
<td>0.72</td>
<td><strong>2.02</strong></td>
<td><strong>0.61</strong></td>
<td><strong>1.75</strong></td>
</tr>
</tbody>
</table>

OR\(_{ij}\) indicates significance at the 95% confidence level

Table 6.11 summarizes the odds ratio between age categories and cluster membership. The first cluster is positively correlated with ages between 35 and 44 years and is negatively correlated with ages above 60. In contrast, Cluster 2 correlates positively with younger age categories, between 20 and 34. The correlation strength decreases with increasing age. Clusters 3, 5, and 9 are all positively correlated with ages above 60. This association is considerably stronger for the fifth cluster. Cluster 10 is positively associated with individuals in their twenties. Cluster 11 also has some level of association with the category including ages above 60, but this association is weaker than for clusters 3, 5, and 9. Cluster 6 is positively associated with the age category covering the youngest half of the forties, while cluster 7 has strong positive correlation with the category covering the youngest half of the thirties. Clusters 4 and 8 have no significant association with age. The age category including ages 19 and below has no significant associations with any of the clusters, likely due to the low number of observations it contains.

### 6.3.5 Income

LTDS data provides information about respondent income at the household level. Over 35% of households interviewed between 2011-2013 refused or were unable to provide information about their annual revenue. In order to mitigate the missing information, the LTDS process involves an inference of the missing income values.
This process relies on other stated characteristics of the household such as household structure, household tenure, number of household members employed full-time and car accessibility. We use the result of this inference. Figure 6-4 shows the income distribution for all LTDS respondents interviewed between 2011 and 2013 and for all LTDS respondents classified as frequent users. The household income distribution of frequent users is roughly representative of the general LTDS distribution. The proportion of frequent users in income categories above £50,000 is slightly larger than that of the general LTDS sample. Correspondingly, the proportion of users in income categories below £35,000 is smaller for frequent users than for general LTDS respondents.

Table 6.12 summarizes the odds ratio between the income and cluster membership for the 11 cluster solution. The first cluster has negative association with the three lowest income categories and positive association with the four highest income category. It has strongest association with the highest income category. Cluster 2 is also negatively correlated to the two lowest income categories, but positively correlated to mid-range incomes between £20,000 and £49,999. Cluster 3 is positively correlated with low-income categories, and negatively correlated to higher-income categories. This cluster has a higher level of association with the lowest income category than any other cluster. Cluster 5 displays a pattern of association reverse to cluster 1, with strongest positive correlation for the second lowest income category and negative correlation with all income categories above £35,000. Cluster 6 is significantly associated with annual income between £50,000 and £74,999. Cluster 7 has a similar association pattern as cluster 1, with no significant association to the highest income category. Cluster 9 has positive association with the second lowest income category and negative association with incomes in the third highest category. Cluster 10 only has significant positive correlation to incomes between £10,000 and £19,999. Clusters 4, 8, and 11
Table 6.12: Odds-Ratio Income Categories - 11 Clusters

<table>
<thead>
<tr>
<th>Income (£)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 5,000</td>
<td>0.23</td>
<td>0.25</td>
<td>4.04</td>
<td>1.02</td>
<td>2.22</td>
<td>0.34</td>
<td>0.29</td>
<td>0.75</td>
<td>1.40</td>
<td>1.29</td>
<td>0.68</td>
</tr>
<tr>
<td>5,000-9,999</td>
<td>0.12</td>
<td>0.31</td>
<td>2.27</td>
<td>1.23</td>
<td>2.96</td>
<td>0.66</td>
<td>0.23</td>
<td>0.87</td>
<td>2.06</td>
<td>0.88</td>
<td>0.97</td>
</tr>
<tr>
<td>10,000-19,999</td>
<td>0.32</td>
<td>1.17</td>
<td>1.30</td>
<td>0.97</td>
<td>1.77</td>
<td>0.98</td>
<td>0.42</td>
<td>0.77</td>
<td>1.15</td>
<td>1.53</td>
<td>1.19</td>
</tr>
<tr>
<td>20,000-34,999</td>
<td>0.96</td>
<td>1.45</td>
<td>0.76</td>
<td>1.35</td>
<td>0.99</td>
<td>0.81</td>
<td>0.52</td>
<td>1.00</td>
<td>0.94</td>
<td>1.44</td>
<td>0.96</td>
</tr>
<tr>
<td>35,000-49,999</td>
<td>1.54</td>
<td>1.88</td>
<td>0.55</td>
<td>0.78</td>
<td>0.40</td>
<td>1.18</td>
<td>1.81</td>
<td>0.95</td>
<td>0.83</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>50,000-74,999</td>
<td>2.30</td>
<td>1.08</td>
<td>0.28</td>
<td>0.80</td>
<td>0.21</td>
<td>1.68</td>
<td>2.20</td>
<td>1.46</td>
<td>0.51</td>
<td>0.81</td>
<td>1.04</td>
</tr>
<tr>
<td>75,000-99,999</td>
<td>2.24</td>
<td>1.21</td>
<td>0.58</td>
<td>0.85</td>
<td>0.14</td>
<td>0.87</td>
<td>2.54</td>
<td>1.08</td>
<td>0.71</td>
<td>0.42</td>
<td>0.98</td>
</tr>
<tr>
<td>≥ 100,000</td>
<td>2.92</td>
<td>0.51</td>
<td>0.06</td>
<td>0.92</td>
<td>0.35</td>
<td>1.54</td>
<td>1.74</td>
<td>1.27</td>
<td>0.62</td>
<td>0.37</td>
<td>1.43</td>
</tr>
</tbody>
</table>

ORi,j indicates significance at the 95% confidence level.

are not significantly correlated to any income category.

6.3.6 Household Composition

The London Travel Diary Survey also records information about the composition of each household. Households are categorized according to one of six categories: couple with children, couple without children, lone parent, single adult, single pensioner, and other. The ‘Other’ category includes all other household composition not covered by the first 5 categories. For example it could include multiple single adults sharing housing, or multigenerational households.

The association between household composition and cluster membership is summarized in Table 6.13. Cluster 1 is positively associated to households including couples, both with and without children. Cluster 2 has no significant positive association, and negative association with single pensioner households. Cluster 3 is positively associated to both lone and single pensioner households, with a stronger correlation with single pensioner households. Cluster 5 and 9 are strongly correlated with single pensioner households, and more so than cluster 3. Cluster 4 and 10 are positively associated with the ‘other’ household category. Cluster 6 has a strong positive association with both types of households including children. Additionally cluster 10 is positively associated with single adult households.
Table 6.13: Odds-Ratio Household Composition - 11 Clusters

<table>
<thead>
<tr>
<th>Household</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Couple &amp; Children</td>
<td>1.73</td>
<td>0.97</td>
<td>0.54</td>
<td>1.11</td>
<td>0.40</td>
<td>2.14</td>
<td>1.41</td>
<td>0.74</td>
<td>0.76</td>
<td>0.56</td>
<td>1.39</td>
</tr>
<tr>
<td>Couple</td>
<td>1.99</td>
<td>1.08</td>
<td>0.68</td>
<td>0.57</td>
<td>0.87</td>
<td>0.62</td>
<td>1.11</td>
<td>1.11</td>
<td>0.66</td>
<td>0.67</td>
<td>1.20</td>
</tr>
<tr>
<td>Lone Parent</td>
<td>0.41</td>
<td>0.63</td>
<td>1.78</td>
<td>1.21</td>
<td>0.83</td>
<td>2.56</td>
<td>0.50</td>
<td>0.83</td>
<td>0.90</td>
<td>1.89</td>
<td>1.13</td>
</tr>
<tr>
<td>Single Adult</td>
<td>0.75</td>
<td>1.39</td>
<td>1.37</td>
<td>0.61</td>
<td>0.68</td>
<td>0.57</td>
<td>1.18</td>
<td>1.51</td>
<td>1.28</td>
<td>1.57</td>
<td>0.73</td>
</tr>
<tr>
<td>Single Pensioner</td>
<td>0.06</td>
<td>0.14</td>
<td>2.55</td>
<td>1.40</td>
<td>3.70</td>
<td>0.25</td>
<td>0.00</td>
<td>0.65</td>
<td>2.67</td>
<td>0.46</td>
<td>1.03</td>
</tr>
<tr>
<td>Other</td>
<td>0.63</td>
<td>1.39</td>
<td>0.75</td>
<td>1.82</td>
<td>1.13</td>
<td>1.29</td>
<td>1.25</td>
<td>0.98</td>
<td>0.81</td>
<td>1.77</td>
<td>0.63</td>
</tr>
</tbody>
</table>

$\text{OR}_{i,j}$ indicates significance at the 95% confidence level.

6.3.7 Vehicle Ownership and Access

LTDS focuses on travel in London across all modes. Hence, it records information about vehicle ownership and vehicle usage. Vehicle ownership or access is recorded at the household level, while usage is recorded at the individual level for the reported day of travel. The number of cars, motorcycles, small vans, vans and other vehicles owned by or accessible to each household is recorded for each vehicle type separately. The survey makes no distinction between ownership of and access to a vehicle. We aggregate all vehicle types and classify households in two categories: households with 0 accessible vehicles and households with one or more accessible vehicles. For the reminder of this section we use the terms car and vehicle, and access and ownership interchangeably.

Table 6.14: Household Access to Vehicle and Transit Usage

<table>
<thead>
<tr>
<th>Cluster</th>
<th>No. Vehicles</th>
<th>Non-Recurrent</th>
<th>Frequent</th>
<th>Occasional</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>609</td>
<td>1522</td>
<td>1289</td>
</tr>
<tr>
<td></td>
<td>≥ 1</td>
<td>819</td>
<td>446</td>
<td>1028</td>
</tr>
</tbody>
</table>

First, we consider the association between vehicle ownership and the three transit usage clusters. These clusters are defined based on the level of transit usage of each individual, which we expect to be correlated with car availability. Tables 6.14 and 6.15 summarize the distribution and association, respectively, between household car ownership and transit usage. As expected, the cluster with the largest percentage of
users having no vehicle in the household is the frequent user cluster. Only 23% percent of users in this cluster live in households which have access to a vehicle. The non-recurrent user group has the strongest positive association with vehicle access. Users living in a household with at least one car are 2.7 times more likely to be assigned to this than those living in households without access to a vehicle. The occasional user cluster is also positively associated with household vehicle ownership, but the magnitude of this association is weaker than for the non-recurrent user group.

Focusing on the frequent user group, Table 6.16 summarizes the odds ratio between vehicle access and the 11 activity pattern clusters. Clusters 1 and 7 have strongest positive association with household access to a vehicle, with members of these clusters being over 3 times more likely to live in households with cars. Clusters 6 and 11 also have significant positive correlation with access to 1 or more vehicles. In contrast, clusters 3, 5 and 10 are most positively associated with households without vehicles. Clusters 2 and 9 also have positive, thought weaker, association to households without vehicles.

### 6.3.8 Result Interpretation

The 5 previous sections discuss correlation with respect to 5 demographic dimensions independently: occupation status, age, household income, household composition, and vehicle access. In order to establish a comprehensive analysis of the relationship
between cluster membership and demographic characteristics, this section presents a summary of associations across the different demographic dimensions presented above. The associations observed suggest certain similarities in the demographic characteristics of certain sets of clusters. We organize the summary of associations according to these similarities. The dominant travel pattern characteristics associated with each cluster are summarized to support the interpretation of demographic associations.

**Clusters 1 and 7**

Travel Pattern of Cluster 1: Distinct working days spent in the secondary area, infrequent travel on weekends, evenings and mornings spent in the primary area. Likely includes users who use public transport to commute on most weekdays.

Travel Pattern of Cluster 7: Distinct working days spent in the primary user-area, occasional week-end travel, evenings and mornings spent in area other than primary area. Likely includes users who use public transport to commute on most weekdays, but use more than one home area, or non-PT modes for certain stages of their commute.

For most demographic dimensions examined in sections 6.3.3 through 6.3.7, clusters 1 and 7 present very similar patterns of association. Individuals in either of these clusters are over 5 times more likely to be employed full-time than any other occupation status, over 3 times more likely to belong to a household with at least 1 vehicle, and 3 times more likely to belong to a household with annual income above £50,000. Both clusters have strong negative association with individuals above 60 years old. This aligns with the travel characteristics of these clusters, marked by clear working days and limited travel on week-ends.

The two clusters differ most noticeably with respect to their activity pattern. Where as users in cluster 1 spend mornings and evening in their primary area and working hours in their secondary area, users in cluster 7 spend working hours in their primary area. This inversion may be caused by two different patterns. First, users who frequently spend the night outside their home, for example at a partner’s home, are inferred to spend more time in the work area than in any single home area. Hence, their primary area is associated with work. Second, journey sequence discontinuities between the last trip of the day and the first trip of the following day result in failure to infer location over night. For instance, the travel sequence of individuals who occasionally use non-public modes (e.g. car-pooling or taxi) for the first stage of their morning commute would be characterized by such discontinuities.

The differences in travel patterns of cluster 1 and 7, may relate, in part, to age differences between the two clusters. While the first cluster has significant positive correlation with users between 35 and 45, cluster 7 has strong positive correlation with younger users between 30 and 35. Additionally, cluster 1 has significant positive association with households made up of couples or couples with children, whereas
no significant household composition correlation is observed for cluster 7. Younger individuals and individuals not living in a household composed of a couple are more likely to spend nights in multiple locations. The demographic associations observed for cluster 7 support this hypothesis.

Overall, these results suggest that users in clusters 1 and 7 tend be full-time employed individuals in higher-income households with access to cars who use public transit almost exclusively to commute.

**Cluster 2**

Travel Pattern of Cluster 2: Distinct working days spent in the primary user-area, frequent travel on weekends, likely includes individuals who use public transport for both work and non-work related activities.

Similarly to cluster 1 and 7, cluster 2 is marked by clear working days and has significant positive correlation with full-time employment. However, in contrast with the two previous groups, this cluster is positively associated with ages between 20 and 35 and with incomes between £20,000 and £49,999. Additionally, unlike clusters 1 and 7, cluster 2 is characterized by a significant negative association with vehicle ownership. These demographic characteristics indicate that users in cluster 2 tend to be younger, middle-range income individuals who use public transport to commute but also for other social and leisure purposes. This is supported by the high number of weekend journeys and the high number of distinct stations visited by these users. For example, young professionals without family obligations who perform frequent leisure and social activities would be most likely to be included in this cluster.

**Clusters 3, 5, and 9**

Travel Pattern Cluster 3: No distinct working days, short activities, high number of journeys on both weekdays and weekends, late first journey departure time. Likely includes individuals not employed full-time who make journeys for multiple purposes at different times of the day.

Travel Pattern Cluster 5: No distinct working days, short activities, high number of journeys on weekdays, infrequent travel on Sundays, late first journey departure time. Likely includes individuals not employed full-time who spend most of their time at home.

Travel Pattern Cluster 9: No distinct working days, short activities, fewer journeys and days traveled than average, similar patterns on both weekdays and weekends. Likely includes individuals not employed full-time who leave home on fewer days than individuals in cluster 3 and 5.

The odds ratio analysis also revealed that clusters 3, 5 and 9 shared a number of attributes across multiple socio-demographic dimensions. Most notably, they are posi-
tively associated with retired and disabled users in low-income categories. While these positive associations are observed across all three clusters, their magnitude varies from cluster to cluster. Cluster 3 is most strongly associated with disabled users, while cluster 5 is most strongly associated with elderly and retired users. Specifically, users in cluster 5 are most likely to be retired, to be older than 60, and to live in a single pensioner household with income below £20,000 and without access to a car. Users in cluster 3 are most likely to be unable to work due to illness or disability, to be in a household with income below £5000, and not to have access to a car. Some level of association is also observed between lone parent households and cluster 3. Cluster 9 shares more similarities with cluster 5 in terms of income and household structure, but is almost equally associated with disabled and retired occupations categories.

These association patterns are consistent with the correlation patterns observed with respect to card type. As previously shown in Table 6.4, disabled freedom cards are most strongly associated with cluster 3 and elderly freedom cards with cluster 5. The table also reveals some level of significant association between these clusters and adult students, particularly for cluster 3. Overall, this suggest that these three clusters are primarily composed of non-working, mostly disabled or retired, low-income individuals without access to cars.

**Cluster 6**

Travel Pattern of Cluster 6: Reduction in travel and change in activity pattern over the second week of the analysis period, which coincides with the school half-term in London. Likely composed of individuals who were on holiday during the school half-term.

Cards associated with students under 18 are 20 times more likely to be in cluster 6 than other card types (Table 6.4). As no LTDS respondent under 17 was asked to provide an Oyster card ID, the correlation between pupils and cluster is not observed from the LTDS data presented in Section 6.3.3. However, the exclusion of pupils reveals another interesting association. LTDS respondents assigned to cluster 6 are over 2 times more likely to belong to a household with children and to be between 40 and 45 years old. Users in cluster 6 also tend to belong in households with income between £50,000 and £75,000 and with access to a vehicle. Cluster 6 is also positively correlated with part-time employments. Overall, these characteristics suggest that users in cluster 6 tend to be not only pupils but also parents of pupils on holiday during the half-term.

**Cluster 10**

Travel Pattern of Cluster 10: Multiple journey sequence discontinuities, no distinct working days, high number of journeys on both weekdays and weekends, complex
activity schedule. Likely composed of multi-modal users who entwine public transport journeys within a complex activity schedule not primarily driven by full-time employment.

Cluster 10 is characterized by significant positive correlation with individuals between 20 and 29, living in single adult and ‘other’ households with incomes between £10,000 and £20,000. Unemployed individuals are 2.3 times more likely to be assigned to cluster 10. The analysis of card types used by individuals in cluster 10 also reveals a mild, but significant association with adult students and disabled freedom pass user. These demographic attributes and travel characteristics supports the hypothesis that this cluster includes primarily lower-income youth living alone or in shared housing, with complex and irregular activity schedules.

Cluster 11

Travel Pattern of Cluster 11: No distinct working days, fewer journeys and days traveled than average, infrequent travel on weekends, late first journey departure time. Likely includes home bound passengers who use public transport for non-commute journeys conducted on weekdays (e.g business related journeys).

Cluster 11 has positive association with self-employed, retired and stay-home LTDS respondents. Additionally, it is positively correlated with access to at least one vehicle and to individuals above 60 years old. All three occupation categories associated with this cluster correspond to individuals who are more likely to be home bound. Overall, the demographic associations provide some insight regarding users contained in cluster 11, but further information on activity purpose would be required to confirm the hypothesis presented above.

Clusters 4 and 8

As detailed in Section 5.3.2, clusters 4 and 8 are marked by reduced travel on the first and last week, respectively, of the period of analysis. As observed in the analysis presented above, no clear socio-demographic trends are observed for these two clusters. This is likely due to the fact that the only characteristic shared by members of each cluster is the change in activity pattern for the first or last week. Unlike cluster 6, this change does not appear to correlate to meaningful demographic attributes. The change in travel pattern observed on the first and last week may also reflect card churn of the finite analysis period considered. For example, cards may have come in use during the first week or may have stopped being used after the third week.
6.3.9 Multivariate Association

The associations described above hint at consistent multidimensional demographic trends for the different clusters. In order to verify that these trends are not the result of correlation between demographic variables, it is necessary to evaluate the associations between clusters and multiple variables simultaneously. For example, we observed that users in cluster 3 are more likely to be disabled and to belong to low-income households. Of course, the income of an individual is highly correlated to her occupation status. We expect some level of correlation between being disabled and belonging to a low-income household. Hence, the association observed between disabled users and cluster 3 could be the result of confounding with income. In other words, it would be possible that, controlling for income, individuals in cluster 3 are no more likely to be disabled than users in other clusters, or vice versa. In order to verify whether one of these two variables is associated to cluster 3 only through its correlation to the other variable, correlation must be estimated by controlling for the effect of the other variable.

Regressing cluster membership on the demographic variables of interest allows for such correlation between explanatory variables to be controlled for. This section introduces a multinomial logit model relating the probability of an individual being assigned to a given cluster to the individual’s demographic characteristics. The model parameters resulting from the estimation reflect the level of association between cluster membership and the explanatory demographic variables. In order to limit the number of parameters being estimated, we group the 11 clusters into 7 groups according to the interpretation presented in the previous section. Additionally, a limited number of the variables analyzed in sections 6.3.3 through 6.3.7 are included, based on the strength of their association. The probability that a user \( u \) belongs to a cluster set \( i \) is modeled according to Equation 6.5.

\[
P_u(i|\mathcal{K}) = \frac{\exp(\beta_0^i + \beta^i_X u)}{\sum_{j \in \mathcal{K}} \exp(\beta_0^j + \beta^j_X u)}
\]  

(6.5)

where \( \mathcal{K} \) denotes the groups of cluster sets defined in Section 6.3.8, \( \beta_0^i \) an alternative specific constant of cluster group \( i \), \( \beta^i \) the vector of alternative specific coefficients for cluster group \( i \), and \( X^i_u \) the vector of demographic variables describing individual \( u \) included in the utility of cluster set \( i \). Vector \( X^i_u \) includes both, variables specified across all cluster groups and cluster group specific variables.
The variables included in the utility of the different cluster groups are summarized below:

\[ X_{\text{age}} = \begin{cases} 
1 & \text{for users younger than 35} \\
0 & \text{for users 35 and older} 
\end{cases} \]

\[ X_{\text{car}} = \begin{cases} 
1 & \text{for users with access to at least 1 car} \\
0 & \text{for users with access to no car} 
\end{cases} \]

\[ X_{\text{ret-dis}} = \begin{cases} 
1 & \text{for disabled or retired users} \\
0 & \text{for users in other occupation categories} 
\end{cases} \]

\[ X_{\text{full-time}} = \begin{cases} 
1 & \text{for users employed full-time} \\
0 & \text{for users in other occupation categories} 
\end{cases} \]

\[ X_{\text{Unemployed}} = \begin{cases} 
1 & \text{for unemployed users} \\
0 & \text{for users in other occupation categories} 
\end{cases} \]

\[ X_{\text{SE-SH}} = \begin{cases} 
1 & \text{for self-employed or stay-home users} \\
0 & \text{for users in other occupation categories} 
\end{cases} \]

\[ X_{\text{Kids}} = \begin{cases} 
1 & \text{for users in households with children} \\
0 & \text{for users in households without children} 
\end{cases} \]

\[ X_{\text{Income}} = \text{midpoint of household income range in £1,000} \]

The disabled category includes users unable to work due to illness or disability. Unemployed users include individual looking for work or waiting to take-up new employment. Stay-home users include all respondents looking after the home or family.

**Estimation Results**

The model coefficients are estimated via maximum likelihood estimation using BIOGEME (Bierlaire, 2003). The 1968 LTDS respondents classified as frequent users are included for the estimation. The group with clusters 4 and 8 is set as the baseline, with all coefficients equal to 0. Table 6.17 summarizes the variables included in each cluster group and the resulting coefficient estimates. Coefficients significantly different than 0 at a 90% confidence level are indicated in boldface.
Table 6.17: Cluster Membership Model - Coefficient Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cluster</th>
<th>1, 7</th>
<th>2</th>
<th>3, 5, 9</th>
<th>4, 8</th>
<th>6</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC</td>
<td>-0.50**</td>
<td>0.08</td>
<td>1.61***</td>
<td>-0.75***</td>
<td>-0.14</td>
<td>-0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age &lt; 35</td>
<td>0.05</td>
<td>0.13</td>
<td>-0.43**</td>
<td>-0.53*</td>
<td>0.47*</td>
<td>-0.69***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HH Car</td>
<td>0.97***</td>
<td>-0.17</td>
<td>-0.18</td>
<td>0.39</td>
<td>-0.60**</td>
<td>0.56***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income (£1,000)</td>
<td>0.006*</td>
<td>-0.007*</td>
<td>-0.010***</td>
<td>0.001</td>
<td>-0.008*</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retired or Disabled</td>
<td>-1.31***</td>
<td>-1.68***</td>
<td>0.95***</td>
<td>-1.02**</td>
<td>0.21</td>
<td>0.68***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full-Time Employed</td>
<td>1.02***</td>
<td>0.82***</td>
<td>-0.53***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kids</td>
<td></td>
<td></td>
<td></td>
<td>0.64***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployed</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-Emp. or Stay-H.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.52**</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* indicates significance at the 90% confidence level (t-statistic ≥ 1.64)

** indicates significance at the 95% confidence level (t-statistic ≥ 1.96)

*** indicates significance at the 99% confidence level (t-statistic ≥ 2.58)

\( n = 1968, k = 36 \)

\( LL_0 = -3829.55 \)

\( LL_{cst} = -3459.51 \)

\( LL = -3005.98 \)

\( \hat{p}^2 = 0.206 \)
The model results are overall consistent with the associations observed from the odds ratios analysis. The coefficient of each variable indicates the direction and the magnitude of its association to each cluster, controlling for the effect of other variables included in the model. The explanatory power of socio-demographics on cluster membership is limited, as indicated by the adjusted rho squared of 0.206, but the trends revealed by the model coefficients are aligned with the travel characteristics of each cluster.

In addition to confirming the trends observed through the odds ratio analysis, the model reveals the following patterns. Income remains positively associated with clusters 1 and 7 even after controlling for employment status, suggesting that the observed association with income is not only a result of correlation with employment status. In contrast, controlling for full-time employment, cluster 2 is negatively associated with income. This suggests that full-time employees in cluster 2 tend to earn less than full-time employees in clusters 1 and 7. Along with the positive, but insignificant, coefficient for being younger than 35, this further supports the hypothesis that cluster 2 is composed of younger, mid-range income users. The model confirms that cluster 6 is positively associated with users living in households with children. Controlling for the low income of users in cluster 10 reveals that the association between unemployment and this cluster is positive, but insignificant. Finally, the model also confirms that home-bound users including self-employed, stay-home and retired users with access to cars are more likely to be in cluster 11.

In order to provide more intuition about the model, we illustrate the results in the context of two examples. First consider an individual living alone with no kids and no car, with a gross annual income of £30,000. For this individual, transitioning from full-time employment to retirement would result in an increase from 23% to 69% in the probability of being assigned to cluster 3. This 3 fold increase is representative of the effect of the retired and disabled variable on the membership to cluster 3. As another example, consider an user living in a household composed of a couple with a car and with £60,000 annual income. For this user, the probability of being assigned to cluster 6 would be below 6%. In contrast, a user with the same characteristics but with at least one child would have a probability of being assigned to cluster 6 equal to 11%. While the absolute probability remains low, this increase represents almost a two-fold change in the probability of being assigned to cluster 6. Given that cluster 6 is significantly under-represented in the LTDS sample, it is expected that the model underestimates the probability of being in this cluster.

Overall, while the model results do not support that socio-demographic variables alone are sufficient to predict the travel pattern of an individual, the direction and magnitude of the associations suggested by the model coefficient are coherent with the hypothesis discussed about each cluster. These associations provide useful insight about the activity schedules and constraints which dictate the 11 travel pattern clusters identified. The results are limited to individuals defined as frequent users, and for whom public transport represent a significant portion of mobility across all modes.
6.4 Conclusion

This chapter examines the demographic characteristics of the travel pattern clusters identified in Chapter 5 using two different data sources. First, demographic information related to card type is considered for the 99,925 individuals originally used to define the travel pattern clusters. Second, 5,713 respondents of a travel diary survey are assigned to one of the three public transit clusters. Of these respondents, 1,973 frequent users are further classified into one of the 11 clusters defined in Section 5.3.2. The demographic composition of these clusters is then analyzed.

The demographic analysis is broken down into two parts. First occupation status, age, household income, household composition, and access to vehicle are considered independently across each cluster. Associations were measured using odds ratios. This analysis revealed that different clusters could be grouped according to the demographic characteristic of the individuals they contained. Cluster 1 and 7 are found to contain primarily individuals employed full-time living in high-income households with access to at least 1 vehicle. Cluster 2 is found to be positively associated with younger users also working full-time, with mid-range income and no access to a car. Retired and users unable to work due to illness or disability are significantly more likely to be assigned to clusters 3, 5, and 9. These clusters are also associated with individuals in lower income households. Cluster 6 is found to be positively associated with users living in households with children. The card type analysis also reveals that students under 18 are 20 times are most likely to be assigned to cluster 6. These associations align with the hypothesis that cluster 6 contains students and parents of students who were on holiday during the school half-term. Cluster 10 was found to be associated with lower income youth and individuals living in single adults or non-family households. The card type analysis also revealed some positive association between cluster 10 and adult students. Finally, cluster 11 was found to include a larger proportion of home-bound users, either self-employed, stay-home, or retired, with access to vehicles.

In order to confirm these demographic trends, the second part of the analysis focused on estimating the correlation between multiple demographic variables and cluster membership simultaneously. For this purpose, a multinomial logit model was estimated to explain cluster membership as a function of 8 different demographic variables. The estimated coefficients are consistent with the associations observed from the odds ratio analysis, but socio-demographic variables available from the LTDS survey alone are also found not to be an accurate predictor of cluster membership. This indicates that the clusters identified in Chapter 5 capture elements of activity patterns not captured by the socio-demographic variables available for this analysis.

These results raise interesting questions for future research. The strength of the associations between travel patterns and demographic characteristic suggests that it would be worth investigating the value of smart card data to predict certain demographic characteristics. This research focused on identifying recurrent travel patterns without a priori knowledge of user demographic characteristic. As such, the clusters
identified for this research reflect the inherent correlation between different travel behaviors (captured from the principal components of user travel sequences). They are not explicitly defined to capture the correlation between travel behaviors and demographic characteristics. Despite being defined without knowledge about socio-demographics, we find that the identified clusters are strongly associated with certain demographic segments (e.g. retirees, students etc.). Explicitly correlating the variables used for the cluster analysis to socio-demographic segments of interest could reveal even stronger association patterns. In other words, while this research applied unsupervised learning methods to identify recurrent patterns of travel behavior, supervised learning methods could be applied to define travel pattern based explicitly on demographic attributes. For example, the travel sequence vectors described in Section 5.3.2 could be used as input to train a prediction model with known labels about the demographic characteristic of each user sequence.

Certain limitations of the modeling approach used for the LTDS data may also justify further research on the association between the clusters we identified and demographic attributes. Issues of sample size and representativeness may be associated with the sample used for the model described in Section 6.3.9. The limited number of observations may be problematic considering the number of parameters estimated. Further research could focus on model specification and estimation with a larger and more presentative sample. For example, sample size could be increased through an online survey distributed widely to all Oyster users with registered email addresses. An appropriately designed survey including only questions relevant to this application would reach a larger and more representative sample of Oyster users compared to the portion of LTDS respondents having provided their card ID. Additionally, other relevant demographic variables not captured in the LTDS survey could be integrated in this type of survey.
Chapter 7

Conclusion

7.1 Summary

As the economic opportunities fostered by large cities become more diverse, the travel patterns of public transport (PT) users become more heterogeneous and more complex. From better personalized customer information, to improved travel demand models used to inform long-term network expansions, understanding these patterns at the disaggregate level is crucial to address developments in urban travel behavior. In order to further this disaggregate understanding, this thesis explored how the structure of individual travel patterns observed over several weeks can be represented, analyzed and compared from public transport smartcard transactions. Specifically, the contributions of the research are organized around three objectives. First, we introduced a representation of individual travel patterns and developed a measure to quantify regularity in the represented patterns. Second, we developed an approach to identify clusters of travel patterns with similar structure, considered with respect to public transport usage and activity patterns. Finally, we presented an exploratory evaluation of the associations between the identified clusters and elements of lifestyle captured from socio-demographic characteristics. These three objectives are considered in the context of a practical application using the transactions of a sample of approximately 100,000 users collected between February 10th and March 10th 2015 in London. Findings related to each objectives are summarized in the following sections.

7.1.1 Activity Sequence Inference

A representation of activity-travel patterns which captures the characteristics of journeys and activities, as well as their organization is introduced in the form of a travel sequence. The travel sequence is composed of ordered events (e.g. journeys, activities, days traveled), each characterized by attributes such as purpose, time, duration, or location, such that the temporal relationship between events is preserved. Such
sequences are reconstructed from the public transport trip transactions of each user. First, all stops and stations are geographically clustered for each user to identify areas of the city aligned with the user’s activities. User-areas are defined such that the distance between the two furthest stops or stations contained in the same area is no larger than 1,000 meters and such that no pair of stops or stations accounting for over 10% of a user’s journeys are grouped in the same area. These areas are then used to infer the activity status of users, based on consecutive pairs of journeys. The destination of a journey $i$ is compared to the origin of the user’s next journey $i + 1$ to infer her location in the interval between the two journeys. User-areas are ordered with respect to the total time spent in each area over the 4-week period considered. The area in which a user is inferred to have spent most time is referred to as her primary area, and similarly for the remaining areas.

Analyzing the duration of intervals for different user areas suggests that users’ primary and secondary areas are associated with specific activity types. The distribution of interval duration for the primary area is characterized by a mode for activities around 14 hours long, and another mode for activities around 8 hours long. These two modes likely align with nights and morning spent at home and with 8-hour working days, respectively. The distribution of interval duration for the second area is characterized by a distinct mode for intervals between 8 and 10 hours long. This likely corresponds to workday activities. Additionally, intervals shorter than one hour make-up a significant percentage of activities across all user areas. The relative frequency of activities shorter than one hour increases for user-areas beyond the secondary area.

7.1.2 Travel Sequence Regularity

Relying on the concept of travel sequence, we develop a measure of regularity which captures the extent to which the order of events characterizing a user’s travel, such as activities or journeys, is repeated over time. Each individual’s mobility is modeled as a stochastic process with memory, from which each new travel event is generated. In this model, the probability of the next event taking on a given value depends on the sequence of previously observed events.

Regularity is broken down in two components: the extent of repetition in individual events and the extent of repetition in the order in which the events occur. These two components are captured based on the conditional probability of observing an event given knowledge of the sequence of events preceding it. Specifically, the entropy rate of the process is used to estimate the average amount of randomness associated with each new event, accounting for memory in the process. Highly irregular mobility processes are associated with high randomness (low predictability) and highly regular processes are associated with low entropy randomness (high predictability). In contrast with entropy rate, the entropy of a process measures the average amount of randomness associated with each generated event without consideration for the order of events (i.e. assuming no memory). As such, entropy measures exclusively the first component of
regularity (repetition of events), entropy rate measures regularity with respect to both components, and the difference between the entropy and entropy rate of a process measures exclusively the second component (repetition in the order of events).

We implement these concepts using the sample of user travel-sequences previously described. The Burrows-Wheeler transform algorithm is used to estimate the entropy rate of each sequence as described by Cai et al. (2004). The results reveal that, without accounting for the order of events, the average travel sequence is almost as random as rolling a six-sided die, as indicated by an entropy of 2.5 bits. Accounting for the order of events, the average travel sequence is slightly more random than a fair coin toss, as indicated by the average entropy rate of 1.4 bits. This indicates that failing to account for the order in which areas are visited omits a significant component of the structure of travel patterns. As suggested by the distribution of the differences between entropy and entropy rate across all users in the sample, the importance of event order varies across the population, from nearly 0 bits to as much as 2 bits. Analysis of specific users suggests that this difference is highest for users who visit a wide variety of areas (high entropy), but always in the same order (low entropy rate).

### 7.1.3 Travel Sequence Clusters

In order to compare the travel-sequences of all users in the sample, a two-step clustering methodology is implemented. First, users are clustered with respect to the distribution of their public transport usage over the analysis period. The resulting clusters are used to categorize users according to the portion of their activity-travel pattern observable from smart card transactions. Three transit usage groups are identified: non-recurrent users, frequent users, and occasional users. Non-recurrent users traveled on few days all concentrated in a short time period, occasional users traveled few days spread over the period of analysis, and frequent users traveled the most.

Frequent users are further classified according to the nature of their activity-travel pattern. In order to compare user patterns, sequences are represented as hourly location vectors. Each hour of the 4-week period is assigned the user-status inferred as previously described. Principal component analysis is used to extract common elements of structure across all frequent-user sequences, and the linear projection of each sequence vector onto the first 8 principal components is used as input to the cluster analysis. Using the $k$-means algorithm, clustering solutions with different number of clusters are compared. The 4-cluster and 11-cluster solutions, resulting in the most compact clusters, are selected based on the DB index. Variability in the principal component definition and cluster compactness is evaluated from bootstrap samples. Cluster temporal stability is also evaluated for the 11-cluster solution based on the transactions of a second sample of users collected between October and November 2015. Cluster are independently defined based on the October-November sample, and compared to the February-March clusters. The analysis reveals that 91% of
October-November users are classified in the same cluster across the February-March and October-November clustering.

Using a annual travel survey of greater London residents for whom smart card transactions are available, the socio-demographic characteristics of different clusters are evaluated. Socio-demographic variables related to elements of lifestyle, including household car ownership, household income, employment status, household structure, and age are used to evaluate associations between clusters and demographic segments. The significance and magnitude of associations between clusters and demographic attributes is tested using the odds-ratio measure. As described below, these associations provide useful insight into the patterns identified from clustering. A simple multinomial logit model for cluster membership was estimated. The results confirm the trends observed from the odds ratio analysis, and indicate that demographic attributes alone are not sufficient to predict cluster membership.

The 4-cluster solution revealed the following four patterns. Sequences assigned to the first cluster are characterized by distinct working-days and reduced travel on weekends. This pattern is associated with passengers who use public transport primarily to commute. The second cluster contains sequences with a high percentage of intervals for which user area status could not be inferred and the highest number of single journey days. This suggests that individuals in this cluster use public transport in combination with other modes. The third cluster includes users who completed the highest number of journeys and the traveled on the highest number of days on average. The group includes passengers who use transit to commute and for other purposes on weekends, as well as individuals who do not follow the conventional working-day schedule. The fourth group is characterized by reduction in travel on the second week of the analysis period corresponding to the half-term school holiday, and by distinct working days on the remaining weeks.

The 11-cluster solution provides groups composed of more homogeneous patterns, for which demographic associations are evaluated in more detail. The first cluster is characterized with clear working days, and little travel on weekends. This pattern is strongly associated with full-time employed individuals living in high-income households with access to at least 1 vehicle. The second cluster is characterized by clear working days and non-work related travel on week-ends. This pattern is associated with younger full-time employed individuals living in medium-income households with limited vehicle access or ownership. The third cluster is characterized by short activities across all days of the week and a large proportion of time spent in the primary area. This pattern is most strongly associated with lowest-income individuals who are unable to work due to disability, and also strongly associated with elderly users. The fourth and eight clusters are characterized by a reduction in travel on the first and last week of the analysis period respectively. No distinct demographic associations are identified for those clusters. The fifth cluster is characterized by short non-work activities, a high percentage of time spent in the primary area, and reduced travel on weekend and specifically on Sundays. This cluster is most strongly associated with elderly retired users and also strongly associated with disabled users.
The sixth cluster is characterized by reduced travel on the second week of the analysis period. This cluster contains primarily pupils or parents of pupils who took a holiday during the school half-term. The seventh cluster is characterized by clearing working days, more time inferred in the work area than in the home area, and reduced travel on week-ends. This pattern is associated with younger full-time employed individuals living in high-income household with access to at least 1 vehicles. The ninth cluster is characterized by the least number of weekdays traveled among frequent users, no distinct working days and late first journey departure time. This groups is equally associated with retired and disabled passengers. The tenth group is characterized by complex activity patterns, multiple journeys performed on both weekdays and weekends and frequent location discontinuities. This pattern is associated with lower-income adults in their twenties, for example students or unemployed passengers. Finally, the eleventh cluster is characterized by travel concentrated on weekdays, and by the lowest number of journeys among frequent users. This pattern is positively associated with individuals with home-bound occupations, such as self-employment, taking care of the home, and being retired.

7.2 Limitations & Future Work

The methods and analysis used to investigate disaggregate travel patterns in this research can be improved and expanded in several ways. Each step of the research framework reveals different opportunities for further research. These are described in the following sections.

7.2.1 Inferring Activity Patterns

Chapter 3 focuses on the representation of travel behavior over time. Specifically, it introduces an approach to partially reconstruct the activity-travel patterns of users from smart card transactions. As illustrated in the layers analogy (Figure 7-1), only a limited portion of each individual’s full travel sequence, across all modes and purposes, can be inferred from PT journeys. In the current approach, non-public transport journeys are only inferred when they result in discontinuities in PT journey sequences. For example, if a user travels to a given area and back using a non-public transport mode between the PT journeys, her non-public transport travel is not captured. Additionally, the current definition of user-areas relies on the amount of time spent in each area. As such, the activity purpose associated with each user-area (e.g. home area or work area) is not explicitly inferred. For example, user home area appear to be associated with the primary area for most users but not for all (e.g. primary area appears to correspond to the work area of individuals in cluster $C_{11}$). A number of research extensions described below would allow for a fuller picture of individual travel sequences to be reconstructed from smart card data.
• **User-area activity purpose.** In order to ensure that user-areas are defined consistently across the user sample, future research could focus on the explicit inference of home, work, and other user-areas. For example it would be useful to ensure that primary areas align with home areas across all users. Instead of labeling user areas according to the amount time spent in each area, collecting data about true activity purpose and developing a model to infer purpose would allow for more consistent area definition. An online survey distributed by email to TfL's registered users could ask respondents to label the stops and stations they visited according to activity purpose. For example, a personalized survey display tailored according to the smart card transactions of the respondent could be used to show users only stops and stations they have recently visited (see Chow, 2014, for example of prompted recall survey). A model inferring activity type according to attributes of user travel patterns and land use data could then be developed and estimated based on the labeled stops and stations. This model could then be applied across all users to infer the activity purpose associated with different user areas.

• **Location inference assumptions.** Comparing the travel sequences inferred
from smart card data to complete travel diary data would allow the validation of the assumptions used in the location inference process. Data from GPS enabled devices or smartphone applications (e.g. Moves) can capture individual mobility across all transport modes. This data would be useful to calibrate the inference methodology presented in Chapter 3. For instance, the current approach assumes that, given that the destination of a journey $i$ matches the origin of the following journey $i + 1$, the user remained in the area over the interval bounded by $i$ and $i + 1$. While this is likely to be true for short intervals, longer intervals may contain unobserved journeys completed using other modes. Given data about the complete mobility of users, it would be possible to identify the types of intervals which are most likely to contain unobserved journeys.

- **Geographical clustering.** As described in Chapter 3, user-areas are defined by clustering the stops and stations visited by users with constraints on maximum cluster diameter and on the proportion of journeys completed within the same cluster. This approach could be improved in two ways. First, network structure could be used to inform the clustering process. For example, the cluster diameter could vary by network area according to stop or station density. Second, the origins and destinations of consecutive journeys could be considered so as to minimize the number of time intervals in which discontinuities are observed, subject to a stop distance constraint.

### 7.2.2 Travel Behavior Regularity

In order to define the regularity measure introduced in Chapter 4, the mobility of each individual is modeled as a stochastic process which generates each new travel event (e.g. activities or journeys). The index is then computed based on an estimate of the entropy rate of the stochastic process. A number of future research opportunities relate to this stochastic representation of travel patterns.

- **Estimator variance and bias** The variance and bias of this estimate depends on the length of the sequence used for the estimation and on the distinct number of possible events. As the length of the observed sequence decreases, and as the number of possible events increases, the variance and bias of the estimate increase. An important extension to this research should focus on evaluating the effect of these two factors on the convergence of the estimator. Other estimation approaches, such as the CTW estimator (Willems, 1998), may provide estimates with lower variance and bias depending on the nature of the sequence. The benefit of these other approaches could be evaluated by comparison with the BWT estimator currently used.

- **Non-discrete sequence applications.** The application presented in this thesis focused on the simple location sequence for which events are defined based on a single discrete attribute. Other sequence definitions introduced in Section 4.2 could be implemented for different event attributes, including, for ex-
ample, activity duration and time of day. This would allow for regularity to be investigated along different dimensions, beyond user location.

- **Mobility as probabilistic sequence** Further research could also utilize the probabilistic conceptualization of mobility developed for purposes other than quantifying regularity. For example, focusing on location events, the probability of the next visited location given the preceding sequence of locations could be used to predict the future locations which an individual will visit. Also, the transition probability between events could be used to identify similarities across different patterns. For example the probability of observing a visit to the home location, given the two previous locations visited were home and work (resulting in the sequence home-work-home) is likely to be similar for individuals who commute regularly. The transition probabilities \( p(X_i = x | X_{i-1}, X_{i-2}, ... ) \) for different sequences of interest, such as home-work-home, could even constitute the input to passenger clustering.

### 7.2.3 Travel Pattern Classification

The repeated activity-travel patterns identified and analyzed in Chapters 5 and 6 highlight interesting future research questions.

- **Unbounded Window of Analysis** As the current methodology focuses on identifying recurrent travel patterns within a finite window of analysis of 4 weeks, further research is necessary to adapt the approach to the operational context. In the context of every day operations, the analysis window constantly increases as new transactions are recorded. Hence, the cluster to which each user is assigned must be updated as new information becomes available. For example, this dynamic clustering problem could be implemented using a rolling window of analysis. Alternatively, users for whom new transactions are observed could be reclassified at fix time intervals. In general, future research could evaluate the effect of window length on the results of this research. On one hand, a longer window of analysis provides more information about each user’s pattern over time and may be useful to address boundary effects related to card churn. On the other hand, the larger number of transactions collected over long time periods may require more computational resources. The trade-offs between the benefits and disadvantages of increased window of analysis could be evaluated in further research.

The dynamic implementation of user clustering may provide interesting research opportunities. First, it would allow for changes in the travel patterns of individuals to be tracked over time. Certain changes in the demographic characteristics, such as loosing employment or having children, may be reflected by changes in the travel of a user. Updating cluster assignment as new transactions are observed would allow for such changes to be recorded.

In the longer term, changes in the aggregate characteristic of clusters could also
be monitored. As shifts in demographics occur over time, the nature and size of each cluster may vary. For example, as employment rate varies, the percentage of users with distinct working days may also change. Similarly, seasonal effects may influence the relative size of each cluster. For instance, the proportion of non-recurrent users may increase in highly touristic seasons. Investigating aggregate fluctuations in cluster attributes over longer time horizons is likely a rich subject for future research.

- **Non-recurrent and occasional users** The clustering methodology described in Chapter 5 was organized in two steps. First, users were classified with respect to their public transport usage, then the activity-travel patterns of frequent public transport users were analyzed in further detail. Detailed analysis of the two remaining public transport usage groups, namely non-recurrent users and occasional users, may reveal additional patterns of interest. As the passengers assigned to these two groups traveled on fewer days than frequent users, a different representation of a user’s travel may be more appropriate. Future research could focus on identifying such a representation and on the analysis of these two groups.

- **Predicting Demographic Characteristics** As demonstrated in Chapter 6, distinct associations exist between the recurrent travel patterns identified from cluster analysis and user sociodemographic attributes. While demographics alone were not sufficient to explain cluster membership, these associations suggest that certain travel patterns may be a strong indicator of some demographic characteristics. The approach presented in this research did not focus on defining groups of travel patterns based specifically on socio-demographic attributes. Rather, recurrent travel patterns were identified independently from demographics, and associations were evaluated for those patterns once they were identified. In the machine learning literature, this is referred to as an unsupervised approach, as the groups are not defined based on labels known a priori. In contrast, it would be possible to extract travel patterns of interest based on demographic attributes explicitly. This supervised learning approach could be used to predict the demographic attributes of a user based on her public transport transactions.
Adams, S. (2013). Rush hour Tube passengers left shocked after being advised to walk 2.5 miles rather than wait for train. *Mirror*.


Fare Collection. *Transportation Research Record: Journal of the Transportation Research Board*, 2351:133–141.


