Classification of London's Public Transport Users Using Smart Card Data

by

Meisy A. Ortega-Tong

Bachelor of Science in Civil Engineering
University of Chile, 2007

Civil Engineer
University of Chile, 2008

Master of Engineering Sciences
University of Chile, 2008

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

Author. . . . .
Department of Civil and Environmental Engineering
May 24, 2013

Certified by. . . . .
Nigel H. M. Wilson
Professor of Civil and Environmental Engineering
Thesis Supervisor

Certified by. . . . .
Harilaos N. Koutsopoulos
Visiting Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by. . . . .
Heidi M. Nepf
Chair, Departmental Committee for Graduate Students

Classification of London's Public Transport Users Using Smart Card Data

by
Meisy A. Ortega-Tong

Submitted to the Department of Civil and Environmental Engineering
on May 24, 2013, in partial fulfillment of the requirements
for the degree of Master of Science in Transportation

Abstract

Understanding transit users in terms of their travel patterns can support the planning and design of better services. User classification can improve market research through more targeted access to groups of interest. It facilitates planning through better survey design, as well as more detailed evaluation, through analysis of impacts based on the characterization of the affected users. Classification of public transport users can be enhanced through the use of data from smart cards. The objective of the thesis is to categorize and better understand travel patterns of London's public transport users, using an extensive database of Oyster Card transactions. Several travel characteristics related to temporal and spatial variability, activity patterns, sociodemographic characteristics, and mode choices are used to identify homogeneous clusters. Four of the groups identified represent regular users composed of workers and students who make commuting journeys during the week, and some of them make leisure journeys during weekends. The four remaining clusters are occasional users, composed of leisure travelers, and visitors traveling for tourism and business purposes.

A detailed analysis of the characteristics of each group in terms of spatial travel patterns, temporal changes in cluster characteristics, and membership is presented. Lack of temporal stability at the cluster level indicated that four clusters are more appropriate to analyze passenger behavior. The clusters were used to examine in detail characteristics of some special groups, such as visitors and registered users. Visitors belong mainly to two clusters, making it possible to identify business and leisure visitors. Registered users showed larger proportions in regular user clusters and their travel patterns were more similar to regular user behavior. The analysis of Oyster Card attrition rates showed that occasional user cards exit the system at a faster rate than cards of regular users who retain their cards for longer periods of time, explaining the high drop in the number of active Oyster Cards observed between consecutive months.

Thesis Supervisor: Nigel H. M. Wilson
Title: Professor of Civil and Environmental Engineering

Thesis Supervisor: Harilaos N. Koutsopoulos
Title: Visiting Professor of Civil and Environmental Engineering

Acknowledgements

The research presented in this thesis would not be possible without the help and support of the following parties that I would like to thank.

- To Transport for London for the funding provided for this research project, specially to Lauren Sager-Weinstein and Duncan Horne for their help and guidance during my internship in London.
- To my research advisors: Professors Haris Koutsopoulos and Nigel Wilson for their professional advice and for the time spent editing and improving this thesis.
- To my London roommates: Yonah Freemark, Laura Riegel, and Cecilia Viggiano for all the great times we spent and for cheering my life up while in London.
- To my friends: Fernanda Bravo, Anna Matias, and, Simmy Willemman for their friendship, all their support, and love.
- To my husband Gastón for all his knowledge, love and patience. Thanks for traveling more than 100 thousand miles to be with me, for encourage me to keep going, and for make me smile every day.
- To my parents and siblings in Chile for understand me, love me and bear having me so far away. Especialmente gracias a mis padres, todo lo que he logrado en la vida es fruto de su trabajo y esfuerzo.
- To anyone who indirectly contribute to this work by making me happy and being my friend.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The analysis of the travel patterns of public transportation users has always been of great interest to transit agencies, since user travel behavior has a significant impact on strategic and operational decisions. Better understanding of the characteristics and needs of passengers, such as regular travel routines, travel purposes, mandatory activities, frequency of travel, and length of trips, can provide additional tools to understand changes that could occur in ridership under particular circumstances or during unexpected events.

Technological advances in automated data collection (ADC) systems provide inexpensive means to support the analysis of passenger movements and system performance. The data obtained from Automated Fare Collection systems (AFC) and Automated Vehicle Location (AVL) systems can be used to infer users origins and destinations by matching fare and vehicle location transactions. The potential of data from ADC systems has been explored in several studies recently. A number of methods has been developed, for example to estimate origin-destination (OD) pairs and full journey from such data (see Gordon (2012)). Having access to complete journey information presents a unique opportunity to improve the study of transit users temporal and spatial travel patterns. The definition of travel pattern is usually based on various travel characteristics that need to be measured.

The research presented in this thesis develops a methodology to identify public transport passenger travel patterns using Smart Card data. The methodology is quite general and can be applied to any system with AFC and AVL data of sufficient quality to allow the estimation of OD pairs in the public transport network. The methodology is applied to the London public transport system using their automated fare collection system: Oyster Card. Oyster Card users are assigned to specific groups representing specific travel profiles, which are built using well defined travel and activity patterns.

1.1 Motivation

The identification of homogeneous travel behavior groups has been the subject of research in several prior studies. The research presented in this thesis addressed this problem in the context of public transport users and is motivated by the potential the analysis of public transport user behavior has to better inform studies in the areas of customer experience and transportation planning. The travel profile of each group provides an aggregate characterization of the users of a group as a whole, which can focus survey questions to obtain more detailed information about specific areas of interest. Understanding travel characteristics of specific groups can not only improve customer communications and surveys for customer research purposes but also provide transportation planners with richer passenger demand information in order to improve system performance or better assess network investments.

1.1.1 Customer Experience

The classification of public transport users based on their travel patterns can support the study of the representativeness of specific groups among the total population. An important group for example, includes users whose Oyster Cards are registered in the Transport for London (TfL) system. TfL has additional information about registered users, such their mailing address, email, and/or telephone number; therefore, they are a group that is relatively easy to reach. However, it is not well understood whether registered users travel behavior is representative of the whole population, if not results of any study conducted using registered users as a sample may be biased. Analyzing the travel characteristics of registered users, knowing their distribution among different travel behavior groups, and comparing their behavior with the rest of the population can determine the representativeness of this group. This allows the generalization and validation of findings from studies based on registered users. It can also help in the the design of more efficient and better targeted travel surveys for marketing research purposes. Moreover, a characterization of passengers' travel patterns is helpful to personalize email communications among registered users in order to provide them only relevant information. Information about station or line closures, unexpected events, changes in service may be specifically targeted to the affected users. This can also reduce the number of emails users receive, probably increasing the effectiveness of communications and education campaigns.

Visitors are another group of great interest to improve customer experience. 26.3 million overseas and domestic visitors arrived in London during 2011 of which an estimated 88% used the Underground during their visit. The behavior of some visitors can be identified by analyzing the Visitor Oyster Cards. This is a type of Oyster Card issued to visitors that can be purchased in advance, and delivered to any country. However, many London visitors buy normal Oyster Cards or paper tickets once they are in London, and there is no direct way to identify them. The identification of this last group of visitors can be facilitated by having a deeper knowledge of the travel behavior of visitors holding Visitor Oyster Cards and comparing it with the behavior of the travel groups identified through the classification process. Exploring the similarities between the travel behavior of Visitor Oyster Card users and other travel groups will also allow determining whether their behavior is unique or part of a broader group that also includes for example, London residents.

Another group of interest correspond to churned or inactive cards. There are a considerable number of Oyster Cards that after certain period of time become inactive and are no longer observed in TfL system. The analysis of Oyster Card attrition rates of different groups will help understanding the underlying reasons for Oyster Card attrition, which can be a first step towards understanding customer attrition. Separating the effect of customer attrition from other effects, such as seasonality, special events, and impact of different projects, could lead to more accurate estimation of passenger demand, which will improve the evaluation of strategic investments or operational planning changes, and the assessment of changes in fare policy.

1.1.2 Transportation Planning

From the standpoint of transportation planning, classifying users travel patterns allows the analysis of possible differences in level of service experienced by different passenger segments and the identification of potential biases. It can also provide better understanding of how changes in level of service affect different users and how they respond to those changes.

Knowing the main differences between groups can contribute to a better understanding of the effect of disruptions on travel behavior. Estimation of origin-destination matrices by type of user provides an opportunity to explore thc impact that users with different travel patterns have on network loads; moreover, the analysis of different groups frequent origins and destinations reveals possible geographic trends.

Finally, understanding the travel characteristics of specific groups may help establish the level of predictability of user trips. Analyzing the frequency of travel of different users allows identifying regular and occasional users, making possible to identify everyday commuters based on the consistency of their trip-making. This distinction between users can help determine the predictability of travel behavior.

1.2 Research Objectives

This thesis explores the use of automatically collected data to analyze passenger travel patterns on the London public transport system. The analysis is accomplished by developing a segmentation of the London public transport system users based on a number of descriptive variables obtained from the Oyster Card data. Oyster Card users are categorized in clusters and characterized by a specific profile, which is built based on common travel and activity patterns, and sociodemographic characteristics. Each group will have some characteristics in common to several other groups; however, each Oyster Card user belongs to only one group.

More specifically, the main goal of this thesis is achieved by focusing on the following objectives:

- Identify homogenous groups of London public transport users, based on similar travel behavior and sociodemographic characteristics.
- Distinguish geographic travel patterns of different types of users and find potential station usage trends.
- Determine travel behavior consistency over time by analyzing group membership temporal variability.
- Characterize London's visitors travel behavior based on the travel patterns of Visitor Oyster Card users and their similarities with other travel groups.
- Find a relationship between travel behavior and different card related decisions, such as registration status and card attrition. This allows the identification of possible bias in the travel behavior of specific groups of interest such as registered card users and churned card users.

1.3 Research Approach

This thesis develops and applies a passenger classification method to analyze the travel patterns of public transport users. The review of previous research provides an overview of travel pattern analysis and identifies the relevant variables used in different contexts to understand travel patterns. Given that there is no previously known information about homogeneous travel groups among the population, the most appropriate classification method is clustering. The theoretical background of the clustering method used in this thesis is presented and summarized as part of the classification methodology discussion.

The thesis pursues the objectives stated above with an extensive application using the London public transport users. The analysis is carried out using London's AFC data which is extracted from the Oyster Card database and London bus AVL system, called iBus. The information provided by both databases is joined using the origin-destination inference method developed by Gordon (2012). This inference tool estimates the most likely origins and destinations for those trips where the boarding or alighting point is not recorded in the AFC system, matching AVL and AFC records. The methodology assumes that most passengers begin their next trip close to the destination of their previous trip and their last trip of the day ends at the origin station/stop where their first trip of that day began. Using this tool and data from one week of a normal period, a network-wide origin-destination matrix of journeys is built to extract travel pattern variables for each user.

The main travel dimensions that are analyzed in this thesis to determine passenger travel profiles include all relevant (and available) user characteristics that define their travel patterns. Researchers have used different approaches to characterize travel patterns, as described in detail in Chapter 2. For this thesis, a multi-dimensional approach is used for the classification with the main variables summarized below.

- **Travel Frequency:** Travel frequency can be interpreted as an indicator of trip temporal regularity/variability and a measure of users' level of usage of the system. The travel frequency variables examined include the number of trips per day, and the number of days of travel per week.
- **Journey Start Time:** The time journeys begin may be an indicator of the purpose of the trip, and the consistency of this start time during the week can also be an indicator of trip regularity. Passengers' start time of the first and last journeys of the day are analyzed for classification purposes.
- **Travel Distance:** Travel distance is an indicator of user accessibility to different activity locations. Users' maximum and minimum travel distance are used in the analysis.
- **Activity Patterns:** Activity refers to actions users perform when they are not traveling. The activities at the end of each public transport journey impact users' travel choices; therefore, activity patterns should be considered as a travel pattern variable. The main variable explored in this thesis in terms of activity patterns is activity duration between public transport journeys.
- **Origin-Destination Frequency:** The number of times during a week that a user frequents an origin or destination is a spatial indicator of travel behavior. Spatial variability contributes toward determining user travel predictability; moreover, this dimension could also reflect travel purpose. For this thesis, users' weekly number of

different origins for the first and last journeys of the day is used to analyze travel behavior.

- **Public Transport Mode Choice:** The extensive spatial coverage of the complex London Public transport network allows users to move from one point to another using several mode combinations. Mode choices are influenced by user characteristics such as network knowledge, age, and physical ability. The number of days passengers' choose only rail or only bus is considered as a useful travel behavior characteristic.
- **Sociodemographic Characteristics:** The sociodemographic characteristics of passengers such as age, physical ability, and household income, also define their travel decisions. These characteristics are more difficult to obtain from AFC data; however, the card type can provide some information about the user. For this thesis two features were consider: whether the card is a Travelcard, which is an unlimited use pass that is payed only once and can last from 7 days to a year; or whether the card is a special discount card, which includes student cards, elderly or disabled free passes, and TfL staff passes.

Due to the large amount of data, it is necessary to establish a sampling strategy in order to use an appropriate sample. A minimum sample size is defined according to the variability in the population.

Using the distribution of the descriptive variables, a general analysis and assessment of the population travel patterns provides a broad overview of Oyster Card users travel patterns. The optimal number of clusters is determined based on a measure of the variation within groups. The clusters of users with similar travel behavior are obtained using the K-medoids clustering algorithm.

A comparative analysis between the different travel profiles is used to interpret the main characteristics of each group. Each group's spatial distribution is analyzed based on their most frequently used stations and the location of their entries and boardings over the day. A home location estimation methodology is developed to identify possible geographic trends.

The temporal stability of travel patterns is an important question. The classification methodology is applied to a sample drawn for 2012 and the characteristics of corresponding clusters are analyzed to identify possible changes. The data also allows the examination of cluster membership stability over time.

The similarities of visitor travel behavior with any particular travel group are explored. Travel patterns inferred from Visitor Oyster Cards are compared with the rest of the population behavior to provide a deeper understanding of visitor behavior and how it relates to travel patterns of the various groups. The distribution of Visitor Oyster Cards among the different groups is analyzed to identify potential tendencies and similarities.

The relationship between travel behavior and card registration status and card attrition is explored. The travel characteristics of registered card users are analyzed and compared to the rest of the population to determine representativeness. The distribution of registered users among groups is estimated and the travel behavior of registered users is tested for similarity with the travel behavior of the cluster they belong to. Oyster Card attrition rates are also analyzed to explore trends in the various groups. Attrition rates are estimated from the number of active cards observed in specific weeks of different months during 2010, 2011, and 2012. The attrition rates of each cluster are compared to identify differences and groups most likely to have higher attrition rates.

1.4 London Background

This section provides a general description of London's public transport system. The section starts with a brief overview of the city of London and its organizational authorities, and continues with a description of the principal elements of the public transport network. The section ends by describing the fare and ticketing structure, which will be helpful to understand some features of the data that will be used in following chapters for passenger classification.

1.4.1 London and the Greater London Authority

London, the capital city of United Kingdom, has a population of approximately 8.1 million inhabitants and possesses one of the largest public transport systems in Europe. Transport for London (TfL) is the government entity in charge of managing most elements of the transportation network in Greater London. TfL is part of the Greater London Authority (GLA), which was created in 1999 by an act of parliament to govern London regionally and strategically. The democratically elected Mayor of London is the primary executive arm of GLA and has wide powers over TfL (Greater London Authority, 2013). TfL's main purpose is to carry out the Mayor's Transport Strategy and manage services across London. Under this scheme, TfL is responsible for most aspects of the transport system in London, including planning, delivery, and operation of the public transport system, the roads, and the congestion charging scheme (Transport for London, 2012h).

1.4.2 The Public Transport Network

London has an extensive radial public transport network which, along with Paris, is the largest in Europe. The London public transport system includes several subsystems which are managed by TfL: bus service (London Buses), metro service (London Underground), regional rail (London Overground), light rail (London Tramlink and Docklands Light Railway (DLR)), and ferries (The River Bus). TfL directly operates only the London Underground (LU). The National Rail (NR) is not managed by TfL and is operated through franchise agreements. The remaining services (buses, Overground (LO), Tramlink, DLR and The River Bus) are operated through concession contracts.

During 2011 London's public transport system carried more than two billion trips while the ridership on a typical weekday was 3.6 million. On the busiest days of the 2012 Olympic Games more than 4.5 million trips were made. Table 1-1 summarizes the network size in terms of ridership, kilometers, and number of stations. A more detailed description of each public transportation mode in London is presented in the subsequent paragraphs.

[1] Not managed by TfL
[2] Depends on the Train Operating Company (TOC). Each TOC has its own fleet
[3] Only includes Thames Clipper fleet

Network	Fleet	Number of stations or stops	Annual Ridership (millions)
Buses	8,500 buses	20,500	2,200
Underground	525 trains	270	1,100
National Rail[1]	More than 500 trains[2]	318	883
Overground	65 trains	83	120
Tramlink	30 trams	39	29
DLR	145 trains	45	86
The River Bus	13 ferries[3]	25	3

Table 1-1: Fleet, Size and Patronage by Public Transport Network

The expansive London bus network covers most of the region, so that more than 90 percent of Londoners live within 400 meters of a bus stop. London buses traveled 486 million commercial kilometers during 2011. TfL is responsible for planning the routes, determining level of service, and controlling service quality for the 8,500 buses serving 20,500 stops on over 800 routes in Greater London (Transport for London, 2012b). The London Underground (LU) is the world's oldest metro system and one of the five most extensive, serving 1,100 million passenger trips annually on a 402 kilometer network of 270 stations on 11 lines. Waterloo and Victoria are the busiest LU stations, used by over 80 million passengers a year (Transport for London, 2012f).

National Rail (NR) provides the long distance intercity rail services, through 29 privately owned Train Operating Companies (TOC), each franchised for a defined term let by the national Department for Transport (DfT). National Rail has 318 stations that connect London with the rest of the UK and serve approximately 833 million passengers a year in London and southeast England. The London Overground (LO) is an above-ground inner suburban orbital rail service with 83 stations, 55 of which are operated by TfL. Since 2007, LO has quadrupled its annual ridership with 120 million passengers carried in 2012 (Transport for London, 2012c).

London Tramlink manages London's tram network, that connects Croydon, Wimbledon, Elmers End, Beckenham and New Addington. Tramlink ridership increased 45% in the period from 2000 to 2012, serving over 29 million passengers annually on 28 kilometers of track and 39 stops (Transport for London, 2012e). The Docklands Light Railway (DLR) is a driver-less light rail system managed by TfL's London Rail division and operated through a concession. DLR serves the Docklands, east of central London, and the newer Canary Wharf Financial District. Its ridership has increased from 8.2 million passenger journeys in 1993/94 to 86 million in 2011/12. (Transport for London, 2012d).

River Bus is the public transport service operated by London River Services Limited (LRS) that uses TfL's eight piers on the River Thames to license scheduled and chartered passenger services. The scheduled commuter river services are known as Thames Clipper. They are operated by a private company (KPMG) that offers the river's public transport service and sightseeing tours. River transport is fully integrated with the rest of the public transport network and carries approximately 3 million passenger journeys a year (Transport for London, 2012d).

London's public transport is very well integrated. London has approximately 600 stations that provide multi-modal interchanges between all modes of public transport. TfL monitors any changes in transport or land use in order to identify interchange coordination needs (Transport for London, 2013b). Fares are also integrated between bus and Underground for passengers holding unlimited Travelcards; more details about the fare structure are presented in the following section.

1.4.3 Public Transport Fares and Ticketing

The London public transport system currently has two physical payment and revenue systems: magnetic stripe tickets and Oyster Cards. Magnetic stripe tickets have been used since 1964 and presently can be used in any of the seven public transportation modes. Oyster Cards are 'contactless' Smart Cards introduced in 2003, that have long lives and can store travel value. Oyster Cards can be used on buses, Underground, trams, DLR, Overground, Riverboats, and on some National Rail journeys. These Smart Cards are used by touching the reader at the ticket gates at the start and end of their LU, LO, DLR or rail journey. On buses, trams, and river services Oyster Cards only need to be validated at the start of the journey.

The rail fare structure is based on 6 concentric zones, numbered in ascending order from Central London outwards as shown in Figure 1-1. The price of a journey depends on the starting, en route, and ending zones. For example, Zone 1 is the most congested zone; therefore, journeys that travels into, from or through this zone pay the highest fare. The zonal structure does not apply to London Buses and Tramlink that have flat rates charged on a per-boarding basis. Interchanges are free between lines of the same mode (e.g. Underground lines, Overground and DLR), but not between different rail mode lines, between rail and buses, nor between buses or trams.

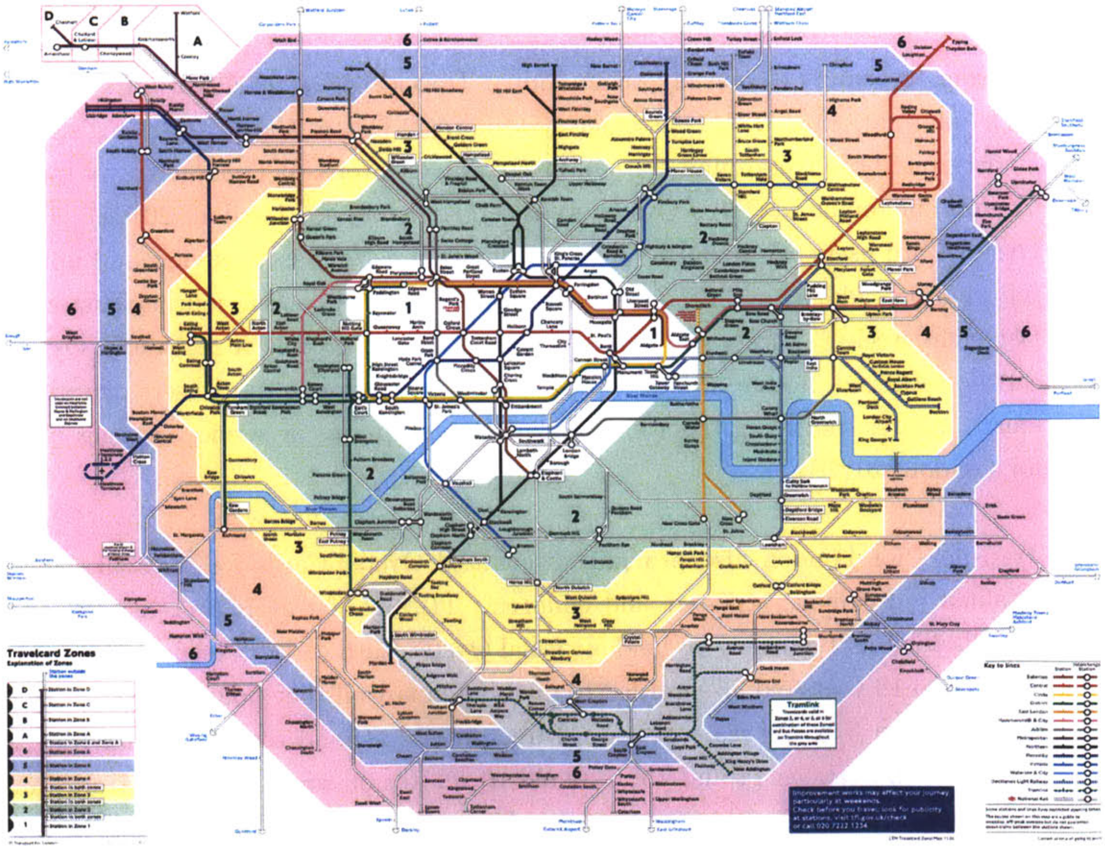


Figure 1-1: Schematic travel-zone map of London rail services (source: TfL)

Seasonal unlimited-use passes, called 'Travelcards', can be added to both Oyster Cards and magnetic stripe tickets, although one-day travel cards can only be bought as a magnetic stripe ticket. Magnetic tickets though, cannot store monthly or annual passes. Travelcards have zonal validity and allow free interchanges between any TfL and National Rail services. Users without a Travelcard can use single paper tickets or 'Pay as You Go' Oyster Cards, which allow the users to add monetary value to their Oyster Card and simply pay according to the completed journeys (Transport for London, 2012g). Single ticket or Pay as You Go fares are higher during peak hours (6:30 to 9:30 and 16:00 to 19:00 Monday to Friday). However, if an Oyster Card user makes many Pay as You Go journeys in one day, a daily price cap is applied to avoid paying more than the price of an equivalent one-day Travelcard (Transport for London, 2013a).

In order to acquire a monthly or an annual Travelcard Oyster Card users must register their information on TfL's customer system. Any Oyster Card can be registered at an Underground or Overground station, Oyster Ticket Stops, at London Travel Information Centers, or online at TfL's Oyster website (Transport for London, 2012a). Registered Oyster Card users can perform remote online transactions such as view and update their information, add money or renew Travelcards, and view their journey history for the last eight weeks. Additionally, the monetary value stored in a registered Oyster Card can be retrieved in case of loss or theft. Registered Oyster Card users can also receive email updates, notifying them about planned disruptions or service changes (Transport for London, 2012i).

Transport for London also provides discounted Oyster Cards for its staff, elderly and disabled individuals, students and children. TfL's staff travels free, as well as elderly and disabled individuals who hold a special Oyster Card called 'Freedom Pass'. Students from 16 to 18 years old are entitled to discounted fares if they hold an Oyster photocard; children from 5 to 10 years old travel free on all TfL and some National Rail services and children between 11 and 15 years old travel free on some buses and trams and have discounted fares on the rest of the system. Most of these special cards need to be registered on TfL's customer system (Transport for London, 2013c).

The Oyster Card was issued to the public for the first time in July 2003. Since then the Oyster Card penetration has grown and become the dominant fare medium for TfL services, recording more than 10 million journey transactions every day. More than 43 million cards have been issued since 2003 and over 80 per cent of all public transport journeys made in London use an Oyster Card as fare medium (Transport for London, 2011). TfL keeps records of every Oyster Card transaction for up to 8 eight weeks. Thus, the AFC database has tremendous potential for in depth analyses of the travel patterns of the users of the system, included the ones presented in this thesis.

1.5 Thesis Organization

The thesis is organized into seven chapters as follows. Chapter 2 reviews previous research on ADCS, and travel pattern and travel behavior analysis. Chapter 3 describes the classification methodology and applies it to the London public transport users using Oyster Card data. Chapter 4 describes the spatial travel patterns of each group and analyzes cluster membership stability over time. Chapter 5 provides an overview of London visitor travel patterns and relates this behavior to the identified travel clusters. Chapter 6 analyzes the travel characteristics of registered and churned card users. Chapter 7 summarizes the main research findings and discusses their implications. It also discusses the main limitation of the work and outlines future research directions.

Chapter 2

Literature Review

Automated Data Collection Systems (ADCS) provide the opportunity to study in detail individual travel patterns. Compared to manual data collection techniques, ADCS provide lower marginal costs, more detailed and disaggregate information, large sample sizes, and real-time data availability. ADCS can be classified into three categories: automatic vehicle location systems (AVL), automated fare collection systems (AFC), and automated passenger counting systems (APC). The potential of ADCS has been explored for planning, managing, and assessing the performance of public transport systems (Wilson et al., 2009). Data collected by these systems allows better understanding of public transport users' travel patterns and travel behavior.

This chapter provides an introduction to the main literature related to the study of travel patterns and travel behavior using both survey and ADCS information. The chapter first describes the two ADCS systems used for this thesis: AVL and AFC, and summarizes previous studies that have shown the benefits of AVL and AFC in Section 2.1. Finally, Section 2.2 provides an overview of previous research that has addressed the problem of analyzing travel patterns or travel behavior of public transport users.

2.1 Automated Data Collection Systems

This section provides information about the main characteristics and potential applications of two of the most commonly used ADCS: AFC and AVL. The analyses presented in this thesis, directly used data from both systems to classify passengers based on their travel patterns. The following section provides a general description of AFC and AVL and some of their most common applications.

2.1.1 Automated Vehicle Location Systems

All systems that record location information of vehicles or trains in real time are considered automatic vehicle location and tracking systems. In the case of buses, AVL systems are commonly based on Global Positioning Systems (GPS); on the other hand, urban rail AVL systems track the location of trains using track occupancy information (Wilson et al., 2009). Existing and potential uses of AVL data to improve service planning and operations management are detailed by Furth et al. (2006). They report that although AVL systems have been applied mainly for real-time operations control and monitoring, they have also been used to improve service performance, planning, and scheduling.

AVL systems have been widely used to assess and improve bus service reliability. Camus et al. (2005) used AVL data to develop a new service reliability measure based on delays and applied it to four routes of the Trieste public transport network in Italy. Similarly, Pangilinan et al. (2008) used real-time AVL data to improve reliability for a bus route in Chicago, developing a simulation model to predict the impact on service reliability when real-time AVL information is available. ElGeneidy et al. (2011) used visual means and analytical methods to analyze public transport service reliability and schedule adherence in Metro Transit, Minnesota. They used the results to show ways of identifying causes of decline in reliability. Analyzing different methods for measuring variability and presenting new visual and simulation methods to analyze different scenarios and optimize resource allocation, Sánchez-Martinez (2013) used AVL data from London buses to improve running time variability measurement and analysis tools. Most of the cited research findings demonstrate that using real-time information based on AVL data can provide significant improvements to service reliability.

In a similar manner, AVL systems have been used to estimate, analyze and predict operational variables, such as arrival times, running times, and speeds. Horbury (1999) uses AVL data from Route 18 in London and on-bus survey data to estimate ridership at stops along the route and bus speeds. It was found that the speeds and ridership estimated using AVL data were comparable with those obtained by other methods, but AVL provided larger and superior data-sets at very low cost. Similarly, Cortés et al. (2011) used AVL data from Santiago, Chile, to develop a method that allows systematic monitoring of average bus speeds. Chakroborty and Kikuchi (2004) compared bus travel times estimated using AVL data and automobile travel times, implementing a functional form that predicts the automobile travel time based on bus travel times.

Since this thesis is not focused on the analysis of bus operations but on the travel behavior characteristics of passengers over the entire public transport system, AVL systems are not used on their own. AVL data combined with AFC data are used in order to infer passengers origins and destinations when bus boarding and alighting stops are unknown. More details about this inference methodology are presented in the following paragraphs.

2.1.2 Automated Fare Collection Systems

AFC systems in public transport were introduced to replace or supplement the traditional tickets with smart cards, allowing customers to retain their cards for longer periods (Blythe, 2004). In some cases, smart card holders can be registered, providing personal information such as home location and demographic characteristics. Therefore, AFC systems open up the possibility to analyze individuals' public transport usage and learn about their travel behavior. A more detailed description of the potential benefits of AFC systems is given in Wilson et al. (2009).

Smart card data has been used by researchers in several studies with diverse objectives. Pelletier et al. (2011) provide a detailed overview of these studies and group them into three levels: strategic, focused on long-term network planning, passenger behavior analysis, and demand forecasting; tactical, related to schedule improvements, and longitudinal and individual travel patterns; and operational, focused on supply-and-demand measures and AFC system operations. This thesis is an example of a strategic-level study related to customer behavior analysis, focused on the identification of groups with distinctive travel patterns through smart card data application.

Since several AFC systems do not have exit or alighting validation records, specially in bus systems, different methods have been developed to estimate the most probable alighting point for individual trips using AFC data. Most methods are based on the two assumptions that Barry et al. (2002) proposed to estimate alighting stations in the New York subway system. First, most passengers begin their next trip close to the destination of their previous trip and second, most passengers end their last trip of the day at the origin station or stop where they began their first trip of that same day.

Zhao et al. (2007) used the same assumptions as Barry et al. (2002) with data from the Chicago CTA system to estimate bus boarding locations. Trèpanier et al. (2007) used the same approach to estimate bus alightings in Gatineau, but they also use next day transactions and historical travel data to complete missing records. Munizaga and Palma (2012) applied this method to a multimodal public transport system in Santiago, Chile, where the direction of travel is unknown.

This thesis applies the origin-destination inference methodology (ODX) developed by Gordon (2012) for the London public transport system using Oyster Card (London's AFC system) and iBus (London's AVL system) data. This inference method uses similar assumptions as Barry et al. (2002), but in this case rail exit station transactions are available therefore, origins and destinations are only estimated for bus journeys. The origin location of the stop is obtained by matching the smart card time of validation and the vehicle trip number to the AVL record of arrival time for that vehicle. To estimate destination locations, the methodology assumes that a customer's alighting location is the closest stop to the passenger's next bus boarding or station entry. More details about ODX methodology are presented in Chapter 3.

2.2 Travel Behavior and Travel Pattern Analysis

Previous researchers have employed various approaches to characterize public transport users' travel behavior and travel patterns, using either survey information or AFC data. Several characteristics of both trips and passengers have been explored to analyze travel behavior; frequency of travel, trip starting time, travel time and distance, activity[1]

[1] In this thesis, activity refers to all those actions individuals perform while not traveling

patterns, origin/destination frequency, and mode choice are the most common variables studied. Two main research threads were found on the literature: research addressing the general travel behavior problem, and research focusing on classification of travel patterns.

2.2.1 General Travel Behavior

Over the years, different approaches have been used to analyze and understand travel behavior. Travel variability, either temporal or spatial, has been commonly addressed to explore passengers travel patterns. Jones and Clarke (1988) analyze day-to-day variability in travel behavior based on three measures: a graphical representation that shows daily differences in activity purposes and duration at the individual level; a similarity index that measures individual day-to-day variability by comparing the trip purposes in the same 15-minute intervals in different days; and a graphical/numerical representation which use different codes for different trip purposes and shows them by time of day. The analysis shows that all the measures are useful for a better understanding of travel variability, all three measures use a "ceteris paribus" criteria which assumes that there are no other effects involved, requiring the introduction of other travel behavior variables to the analysis.

Pendyala et al. (2000) provide a general overview of several studies that have analyzed travel variability. Their goal is to examine and compare measures of travel behavior variability using a survey from Lexington, Kentucky, based on GPS data collection devices installed in the surveyed household vehicles. Frequency of travel (number of journeys with different purposes during different periods of time0, journey start time, travel distance and time, and purpose of the trips are the variables used to characterize travel behavior during weekdays and weekends. Travel behavior variability is explored by estimating the percentage of individuals that have similar travel variables on all reported days, all but one reported day, and all but two reported days. The results showed that there are only a small percentage of individuals who repeated their behavior on all days regardless of the travel variable being used.

Schlich and Axhausen (2003) compare different methods to measure similarity of travel behavior in order to address the question of how the similarity and variability of travel behavior can be measured. Based on data from a six-week travel diary, three similarity indeces are compared empirically. First, the repetition index developed by Hanson and Huff (1986) is explored. This index examines the proportion of individuals' activity patterns that can be considered repetitive using as attributes: mode, trip purpose, trip destination, trip distance and arrival time. The repetition is measured by comparing the deviation of the distribution of the attributes with respect to a distribution in which all possible combinations of trips are performed. Second, the similarity index developed by Pas (1983) which compares trips of different days is used. This index is flexible in the trip attributes that compares and allows using different weights for these attributes making it possible to adopt the index for different purposes. Finally, the third index analyzed is the similarity index developed by Jones and Clarke (1988) described above. The results of the comparison of these three indices indicate that daily travel patterns are more variable if the measurement index is trip-based rather than time-budget based.

Liu et al. (2009) use smart card data from Shenzhen, China with the goal of understanding collective temporal and spatial mobility patterns and their relationship to land use. They analyze three characteristics of public transport users' travel behavior: trip start times, number of station entry and exit, and most frequent origin-destination pairs. The analysis of these variables at large scale shows that the mobility patterns in Shenzhen are repetitive over time and are spatially focused in the center of the city during the peak hours. Using similar travel behavior variables, Chakirov and Erath (2011) use one full day of smart data from the entire city-state of Singapore to characterize public transport travel behavior. They analyze three main variables to describe travel behavior: the distribution of all-day journey start times; the waiting times (only in subway stations) estimated as the total recorded travel time minus the in-vehicle travel time obtained from AVL data[1]; and activity duration (time between consecutive journeys), location, and purpose. The purpose of the activities is inferred based on their durations, for example, activities lasting between 8 and 12 hours are considered work activities. The results not only show the potentials of smart card data for the characterization and analysis of travel patterns, but also present a first approach to an activity location model based on AFC records.

[1] This measure of waiting time considers only journeys with no interchanges and assumes that all passengers board the first train

With regard to activity models based on AFC data, Devillaine et al. (2012) and Lee and Hickman (2012) also developed methodologies to infer activity purpose from smart card data using activity durations, activity locations, and other smart card characteristics. The purpose of both papers is to infer passengers' activity duration, location, and purpose using AFC data. The activity duration is estimated as the time between consecutive journeys and the location corresponds to the destination of the last journey; therefore, it depends on the origin-destination inference methodology used in each case. The activity purpose (work, home, study or other) inference methodology is similar in both cases. It is based mainly on the card type (regular card, student, senior, or other), the duration of the activity, the start time of the first stage of the previous journey, and the activity location. For example, an 8 hour activity performed in the Center Business District (CBD) of the city by an adult card user whose last journey started during the morning peak is identified as a work activity. Both papers conclude that AFC data have great potential to infer journey and activity purposes, and note improvements that can be made to the activity models as more information becomes available.

Lathia and Capra (2011) analyze travel behavior in London by comparing characteristics of travel obtained using smart card data and reported in surveys. The goal is to measure the difference between perceived and actual travel behavior using trip frequency, journey start time, travel times, and public transport mode choices as travel behavior variables. They examine two hypotheses:

1. Travelers perceptions of their usage of public transport do not match their actual behavior

2. Transport operators offer incentives that do not work

The results show that users made less public transit trips than they claimed and associated the notion of regularity with repetitive time of travel and destinations rather than the amount of travel. Users also show more flexible mode choices than they reported, and claim to spend more money in travel fares than they actually do. Finally, fare incentives encouraged users to travel more in the case of holders of unlimited passes with a fixed price, but not in the case of students with special discounts.

Taking a general perspective, Nishiuchi et al. (2013) analyze variations in origin and destination frequency over a period of one month using smart card data from Kochi City, Japan to assess if there is any meaningful relationship between the daily spatial and temporal routines of public transport users. Low and high frequency passenger groups are identified from the distribution of days of travel for the analysis period. The study reveals that different card types have different journey behavior; for example, adult card users who have registered their cards in the system are likely to have more repetitive work trips.

2.2.2 Travel Pattern Classification

This section describes research which aims at analyzing passengers' travel behavior through the classification of their travel patterns. Different classification techniques and variables have been used to identify distinct travel behavior groups, some of which are also used for the classification of passengers in this research. Examples include Hanson and Huff (1986) who used an out-of-home travel-activity survey from Uppsala, Sweden to classify individuals in homogeneous travel behavior groups. The travel behavior measures used to classify individuals were: the proportion of out-of-home time spent on different activity purposes, the proportion of single-stop trips, the number of trips per day, and the proportion of walking trips. A $K-$means clustering algorithm (Jain et al., 2000) identified five groups with different travel characteristics, that were also analyzed using sociodemographic variables (gender, household size, and number of establishments near home). The results show that even though the five clusters of individuals share distinctive travel and sociodemographic attributes, there is considerable intragroup variance with respect to the variables and substantial overlap among groups.

Ma and Goulias (1997) used different travel behavior characteristics to classify passengers from the Puget Sound Transportation Panel (PSTP). Their goal is to measure variability in activity and travel patterns over time, examining the effect of two time-scales (day-to-day and year-to-year). $K-$means clustering was used as the classification method resulting in two main groups: activity and travel clusters identified at personal and household level (using an average of all household members). Four activity clusters with different characteristics were identified. Frequency of different activity purposes, duration of activities, trip frequency by travel mode, number of trip chains, and total travel time were used as clustering variables. Another four travel clusters were identified using only trip frequency by travel mode, number of trip chains, and total travel time as clustering variables. The results showed that travel patterns have a higher degree of regularity than activity patterns, which may be explained by the transportation system constraints. It was also shown that even though person- and household-based activity and travel patterns are very similar, there is less variation at the household-level than an a person-by-person basis.

Agard et al. (2006) used smart card data to characterize user travel behavior in Gatineau, Quebec. They explore the similarities in travel patterns using the start time of trips as the main travel characteristic. The main objective of this study was to demonstrate that data mining techniques can help identify and characterize market segments among public transport users. Using a $K-$means classification algorithm, they classified users into four groups according to the repetition of the starting period of each journey: two groups of users with regular activities starting at peak hours and only during the first part of the day, two groups of users with low travel frequency and no clear travel pattern. The composition of these clusters was analyzed in terms of card type (adult, student, or elderly), showing that the regular groups are mainly composed of adults and students. The variability of cluster membership over 12 weeks was analyzed to test cluster membership stability. The clustering process was repeated for each of these 12 weeks and the share of users changing from one cluster to another was analyzed. It was shown that with the exception of students, most adults belonging to one of the two regular travel groups remain in the same cluster for all weeks.

Chapter 3

Classification of London Public Transport Users

This chapter focuses on the methodology used for the classification of London's public transport users and discusses the results from its application. Section 3.1 describes the classification methodology, describing the classification methods that were considered for application in this thesis and describing a set of travel descriptive variables to be used in the classification process. Section 3.2 describes the data needs and sources used in this thesis, including the chosen sampling strategy. Section 3.3 provides a general characterization of the samples used for the classification and section 3.4 presents clustering processes and describes the identified passenger groups. The chapter ends with a summary of the findings.

3.1 Methodology

This section describes the methodology that is used in this thesis to identify different passenger groups with similar pattern profiles. Based on the literature reviewed in Chapter 2 and on the information available, the categorization of the passengers travel patterns is performed using classification techniques without a training sample, which is commonly refereed to in the literature as unsupervised classification or clustering analysis. This section is organized as follows: first, the theoretical background of the classification methods necessary for the analysis is presented, and second, the descriptive travel variables necessary to estimate each passenger group are defined and described.

3.1.1 Classification Methods

Classification methods encompass several techniques and algorithms used to group observations based on similar qualitative or quantitative characteristics. These methods are usually divided into two categories: supervised and unsupervised classification. Supervised methods require a training sample which contains previously known information on each group membership. If a training sample is not available or there are no previously known classes, unsupervised classification methods are used. In the following paragraphs, both supervised and unsupervised approaches are described.

a. Supervised Classification

Supervised classification, also known as supervised learning, aims to predict object group membership based on input information about the object. These methods use past data as 'training' samples or previously known outputs to create and 'learn' a classification rule that allows the classification of future or new observations. In supervised learning there is always an input and an output, and the goal is to develop a mapping from the input to the output (Alpaydin, 2004).

Two simple but powerful supervised learning approaches are described below: linear regression fit by least squares and the k-nearest-neighbor prediction rule. Both methods make important assumptions about the structure of the data. While the linear regression model yields stable but possibly inaccurate predictions, the predictions obtained by k-nearest neighbors are often accurate but can be unstable (Friedman et al., 2001).

i. Least Squares

The least squares problem has been widely studied in the case of the linear regression model. This prediction model aims to predict an output $\hat{Y}$, given an input vector $X^T = (X_1, X_2, ..., X_p)$

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{p} \left(\hat{\beta}_i \cdot X_i\right). \tag{3.1}$$

Where $\hat{Y}$ is the resulting output vector, and X is the input vector. $\hat{\beta}$ is a vector of coefficients to be estimated, including the term $\hat{\beta}_0$ or intercept, also known as the *bias* in machine learning (Friedman et al., 2001). Commonly, the linear model is fitted to the training data using the least squares method. The goal is to find the values of $\hat{\beta}$ that minimize the residual sum of squares which is a quadratic function of β, and therefore always has a minimum. ANOVA can be a useful analysis tool to test the statistical significance of the estimated coefficients and to validate the model results.

ii. k−Nearest Neighbor Estimator

The k-nearest neighbors method predicts the output vector $\hat{Y}$ using those training data objects that are closest to each x in the input data. Therefore, the output $\hat{Y}$ is predicted using the model

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i, \tag{3.2}$$

where $N_k(x)$ is the set of k closest points to x in the training sample. A closeness measure must be defined, such as the Euclidean distance. In summary, the k-nearest neighbors method finds in the training sample the k observations that are closest to x in the input data, and classifies x based on the average of its neighbors classification values (Friedman et al., 2001).

b. Unsupervised Classification

Unsupervised classification methods, also known as clustering techniques, aim at categorizing the data objects without a training sample, i.e. there is no known output data. Therefore, the goal is to find clusters based on similarities of the input data. For this thesis and given the complexity of human travel behavior, it is almost impossible to have a sample of users previously labeled under a true category that can be used as training data. Consequently, unsupervised classification methods must be used to identify homogenous categories of travel patterns among users (Jain and Dubes, 1988; Jain et al., 2000; Alpaydin, 2004).

Clustering techniques are often classified as hierarchical or partitional. Hierarchical clustering groups data objects with a nested sequence of partitions using a similarity criterion for merging or splitting clusters, while partitional clustering divides data objects into a specific number of clusters optimizing a clustering criterion (Jain et al., 2000). An overview of the most common hierarchical and partitional clustering algorithms follows, including the corresponding cluster validation methods.

i. Hierarchical Clustering

Hierarchical clustering algorithms organize data into a nested sequence of groups and only require the specification of a measure of similarity –usually Euclidean distance– between each pair of data objects. A proximity matrix is built using these distance measures between objects. The hierarchical algorithm organizes the data according to the proximity matrix using a hierarchical structure usually represented by a binary tree or *dendrogram* (Jain and Dubes, 1988; Friedman et al., 2001; Xu and Wunsch II, 2005). *Agglomerative clustering algorithms* start at the bottom level of the tree with each cluster containing one object, merging similar groups to form larger groups until there is only one group that contains all the data objects. A *divisive clustering algorithm* follows the reverse process, starting at the highest level with a single group and dividing it into smaller groups (Alpaydin, 2004).

These types of clustering algorithms have several advantages. First, it is not required to know the number of clusters in advance. Second, the representation of the results in a hierarchical manner provides informative descriptions and visualization for the potential clustering structures (Xu and Wunsch II, 2005). Finally, the algorithms do not require input parameters besides the similarity measure between observations, and this allows their application to any type of data, including qualitative information.

Nevertheless, these algorithms also have a number of disadvantages. Hierarchical clustering has been criticized for low robustness and high sensitivity to noise and outliers. Since the assignment of an object to a cluster is not iterative, hierarchical algorithms are not able to correct potential misclassifications. Moreover, hierarchical clustering algorithms have high computational complexity, limiting their application to small-scale data sets; in fact, for a sample with n objects, the number of possible sub-divisions is $(2^{(n-1)} - 1)$, which may have a very high computation cost (Everitt et al., 2001).

ii. Partitional Clustering

In contrast to hierarchical clustering, partitional clustering techniques assign data objects to a number of clusters, that may or may not be specified, with no nested structure. Partitional algorithms optimize either a locally or a globally defined objective function to generate groups of observations. Finding all the possible clustering combinations to achieve an optimum value is not computationally viable; consequently, the best clustering structure is chosen after running the algorithm several times using different initial scenarios. Partitional clustering algorithms are preferred in applications that involve large data sets, where it is not computationally feasible to use the hierarchical approach. The main disadvantage is the difficulty of choosing the number of clusters and their dependency on the initialization scenario (Jain et al., 1999).

The most popular partitional clustering approaches are the *squared-error clustering* and *mixture decomposition.* Squared-error clustering methods are most commonly used. The general goal is to find the clustering structure that minimizes the squared-error for a given number of clusters (Jain and Dubes, 1988). On the other hand, mixture decomposition algorithms assume that each data object is generated according to a probability distribution associated with each cluster or population; therefore, the objective of these methods is to allocate each object to its correct population (Xu and Wunsch II, 2005).

K-means is one of the most popular squared-error clustering algorithms. It is a computationally efficient method, suitable for situations where all variables are quantitative and the dissimilarity measure is the squared Euclidean distance. Accordingly, the squared-error of each cluster in the K-means algorithm is the sum of the squared Euclidean distances of each data object $X_i^{(k)}$ with respect to the centroid of each cluster $C^{(k)}$. The centroid of each cluster k is defined as the mean of the n_k data objects belonging to that cluster,

$$c_j^{(k)} = \left(\frac{1}{n_k}\right)\sum_{i=1}^{n_k} x_{ij}^{(k)}. \tag{3.3}$$

Where $c_j^{(k)}$ is the centroid of each cluster k for the component j, n_k is the number of data objects that belong to cluster k, and $x_{ij}^{(k)}$ is the data object i for the component j. Therefore, the centroid of each cluster in a multivariate case is $C^{(k)} = (c_1^{(k)}, c_2^{(k)}, ..., c_J^{(k)})$. The squared-error of each cluster k for the component j, e_{jk}^2, is the within-cluster variation

$$e_{jk}^2 = \sum_{i=1}^{n_k}\left(x_{ij}^{(k)} - c_j^{(k)}\right)^2. \tag{3.4}$$

Therefore, the objective of the K-means algorithm is to find k clusters that minimize the sum of within-cluster variation (Jain et al., 2000). The partition of k clusters that minimizes the squared error is called the minimum variance partition. The K-means algorithm used to find this partition is summarized in Table 3-1 (Xu and Wunsch II, 2005; Jain et al., 2000).

The K-means algorithm lacks robustness against outliers that produce large distances. A generalization of the algorithm is the K-medoids algorithm that trades off robustness with computational efficiency. The K-medoids algorithm assigns one of the observations of the cluster as its center, which is known as the 'medoid' of the cluster. A summary of the K-medoids algorithm is presented in Table 3-2 (Friedman et al., 2001).

Finding the medoid for each provisional cluster requires a much higher computational

Table 3-1: Partitional Clustering: K-means

K-means Clustering Algorithm	
Step 1.	Select an initial random partition: divide the sample into k arbitrary clusters and compute the cluster centroids.
Step 2.	Assign each data object to its closest cluster centroid and generate a new partition, relocating the objects based on the minimum distance to the new centroids.
Step 3.	Compute the cluster centroids for the new partition.
Step 4.	Repeat 2 and 3 until there is no change for each cluster.

Table 3-2: Partitional Clustering: K-medoids

K-medoids Clustering Algorithm	
Step 1.	For an initial arbitrary cluster partition, find the observation in each cluster that minimizes the total distance to the rest of the cluster members and define it as the cluster medoid.
Step 2.	Assign each data object to its closest cluster medoid and generate a new partition, relocating the objects based on the minimum distance to the new medoids.
Step 3.	Find the cluster medoids for the new partition.
Step 4.	Repeat 2 and 3 until the assignments do not change.

effort than finding the centroid. Hence, several methods have been developed to reduce the computational cost of the K-medoids algorithm. Alternative strategies, such as the one proposed by Rousseuw and Kaufman (1987), have been implemented in statistical software, such as the CLARA package for *R* (CLustering Algorithms for LArge Data Sets). CLARA draws small samples from the complete data set and applies the K-medoids algorithm to obtain a set of medoids for the sample. The sampling and clustering process is repeated a pre-defined number of times to reduce bias. Subsequently, CLARA selects the final clustering result as the set of medoids that minimizes within-cluster variation. Alternatives such as CLARA make it feasible to use the K-medoids method with large data sets (Rousseeuw and Kaufman, 1990; Wei et al., 2003; Maechler et al., 2013).

The mixture decomposition approach identifies the parameters of each population (cluster) distribution as the maximum likelihood estimates of the density parameters (Jain et al., 1999). The *expectation-maximization* (EM) algorithm (Jain et al., 2000; Alpaydin, 2004; Xu and Wunsch II, 2005) is typically used for the maximization of the likelihood. The EM algorithm starts with initial estimates of the parameters and allocates the data objects according to the mixture density generated by the parameters. New parameters are estimated using the new mixture density and the procedure iteratively updates the data object allocation and the corresponding mixture density until convergence.

The EM algorithm has some disadvantages; it relies heavily on the arbitrarily chosen initial parameters, and convergence is not assured when the data set includes outliers and/or repeated samples (Xu and Jordan I, 1996; Jain et al., 2000). Nevertheless, it can be proved that under a spherical Gaussian mixture, the EM algorithm is equivalent to the K-means algorithm (Celeux and Govaert, 1992).

iii. Clustering Validation Methods

Validation methods aim at assessing the clustering results objectively and quantitatively (Jain and Dubes, 1988). As stated in Jain et al. (2000), there is no "best" clustering algorithm and several clustering methods should be used to find one that is appropriate for the data. Furthermore, the collection, normalization, and representation of the data together with cluster validation are as relevant as the clustering algorithm chosen (Jain et al., 1999).

There are several criteria that can be used to evaluate cluster validity (Jain et al., 1999) all of which aim is measure how separated the clusters are. For the purposes of this thesis, two indices will be explored to validate the clustering results: the Davies-Bouldin (DB) index (Davies and Bouldin, 1979) and the Caliñski-Harabasz (CH) pseudo F-statistic (Caliñski and Harabasz, 1974). The DB index is defined as a function of the ratio of the sum of within-cluster variation to between-cluster separation. The CH pseudo F-statistic is given by,

$$CH^{(k)} = \left(\frac{B(k)/(k-1)}{W(k)/(n-k)} \right) \tag{3.5}$$

where $B(k)$ and $W(k)$ are the sum of between-cluster and within-cluster variation respectively, k is the number of clusters, and n is the total number of objects. This index is equivalent to the F-value of a one-way ANOVA, with k representing the number of clusters. Thus, minimizing the DB index and maximizing the CH index help determine the optimal number of groups and achieve proper clustering.

ANOVA can be used to test the significance of the clustering variables. ANOVA is applied to the variables and the clustering results using either a linear model –parametric ANOVA– or a regularized discriminant analysis –non-parametric ANOVA– (Guo et al., 2005). The discriminant value of the clustering variables can also be tested by checking if certain clusters have significantly different means in these variables. ANOVA can easily support such analysis (Hanson and Huff, 1986; Ma and Goulias, 1997; Morency et al., 2007).

For this thesis, three methods were used to determine the optimal number of clusters and to validate the variables used. The within-cluster variation and the DB index are used in the following sections to determine the optimal number of clusters and to compare the significance of the results obtained with different clustering methods. ANOVA was used as a validation tool to explore the significance of the final clustering results.

3.1.2 Travel Pattern Descriptive Variables

To estimate homogeneous passenger groups based on their travel patterns using any classification method, it is necessary to have input information on travel behavior. Travel patterns can be described by looking at specific variables that together characterize each passenger's travel routines. These descriptive variables can be used as the clustering variables necessary to determine a passenger segmentation. Hence, the selected variables must include those users' characteristics that make their travel patterns distinct. As discussed in Chapter 2, previous researchers have used different approaches to characterize travel behavior. For this thesis, a multi-dimensional approach will be used. Therefore, a set of descriptive variables needs to be identified. The selected variables have been categorized into five groups: those describing temporal and spatial variability, and those capturing activity patterns, sociodemographic characteristics, and mode choices.

- **Temporal Variability**

  The temporal variability category comprises all those variables that explain travel behavior related to time. Two different dimensions are treated as temporal variability variables: travel frequency and journey start time.

  - Travel Frequency

    The frequency at which journeys are made over a day, a week (or any other period) indicates how regularly passengers use the public transport system, allowing their classification based on travel temporal variability. For this thesis, journeys are composed by all the public transport trip stages made that are necessary to reach their destination. Travel frequency is one of the travel behavior characteristics most commonly analyzed in the literature. For this thesis, travel frequency is explored using two descriptive variables:

    - *Number of Journeys per Day:* Number of complete journeys performed on each day of the week.
    - *Days of Travel:* Number of days within the period of analysis that a passenger used the public transport system.

  - Journey Start Time

    The time journeys start could indicate the journey's purpose and consistency of journey start time over a week could indicate travel regularity. For example, users that travel every weekday and their first journey of the day starts during the morning peak are more likely to be commuters, traveling either for work or study purposes. For the passenger segmentation presented in this thesis, the start times of the first and last journeys of the day are analyzed using the following descriptive variables.

- *Weekday Average First Journey Start Time:* Mean value of the starting times of thc first journcy of the day during wcekdays with travel. The time averaged is the starting time of the first travel stage of the first observed journey of the day.
- *Weekend Average First Journey Start Time:* Mean value of the starting times of the first journey of the day during weekend days with travel. The time averaged is the starting time of the first travel stage of the first observed journey of the day.
- *Weekday Average Last Journey Start Time:* Mean value of the starting times of the last journey of the day during weekdays with travel. The time averaged is the starting time of the first travel stage of the last observed journey of the day.
- *Weekend Average Last Journey Start Time:* Mean value of the starting times of the last journey of the day during weekend days with travel. The time averaged is the starting time of the first travel stagc of the last observed journey of the day.

- **Spatial Variability**

  Spatial variability variables measure passenger behavior spatially across the public transport network. The two travel dimensions that are considered in this category are: origin stop/station frequency, and travel distance.

  - Origin Stop/Station Frequency

    The frequency at which passengers use specific stops/stations to start their journeys has the potential to be a useful indicator of their mobility patterns. For example, users with the same station/stop for the last journey of the day over a week are more likely to be commuters with work or study purposes. This variable is an indicator of spatial travel variability, which could help to infer user travel predictability. The origin station/stop frequency variables that are used to identify passenger travel pattern groups include:

    - *Percentage of Different First Origins, Weekdays:* Ratio of the number of different origin stops/stations used during weekdays as the starting point for the first stage of the first journey of the day to the number of weekdays the passenger traveled.

- *Percentage of Different First Origins, Weekends:* Ratio of the number of different origin stops/stations used during weekend days as the starting point for the first stage of the first journey of the day to the number of weekend days the passenger traveled.
- *Percentage of Different Last Origins, Weekdays:* Ratio of the number of different origin stops/stations used during weekdays as the starting point for the first stage of the last journey of the day to the number of weekdays the passenger traveled.
- *Percentage of Different Last Origins, Weekends:* Ratio of the number of different origin stops/stations used during weekend days as the starting point for the first stage of the last journey of the day to the number of weekend days the passenger traveled.

- Travel Distance

The geometric distance between the start and end points of a journey can show how accessible activity locations are to a user. Travel distance variability among the journeys of a user can also show travel flexibility and user mobility around the city. In this thesis the following variables related to travel distance are used to identify passenger groups.

- *Maximum Distance Traveled:* Maximum distance traveled among all journeys made in a week. A journey distance is defined as the Euclidean distance between the starting station/stop of a journey and the ending station of the same journey.
- *Minimum Distance Traveled:* Minimum distance traveled among all journeys made in a week. A journey distance is defined as the Euclidean distance between the starting station/stop of a journey and the ending station of the same journey.

- **Activity Pattern Variability**

  The objective of making trips is to reach destinations where different activities can be performed. Activity refers to all those actions passengers perform when they are not traveling. Activities can have different purposes: work, business, study, and recreational, among others. The characteristics of the activity performed at a destination may determine passengers travel decisions. For example, to reach the destination of an 8 hour work activity passengers are likely to travel in the morning, with a fixed schedule, and choose the most reliable mode or try to minimize their travel time. For this thesis, the duration of activities performed outside home and after a public transport journey is used to identify passenger groups.

  - Activity Duration

    The length of the activities passengers perform at their destinations is a determinant of their travel choices. Longer activities far from home are usually performed with work or study purposes, while recreational activities tend to be shorter. The activity duration at the destination, typically of the first journey of the day, also indicates travel flexibility and possible tour or circuit identification[1]. The variables used in this thesis to measure activity duration are presented below. In each case days with no activities and activities performed at home are not considered in the variable estimation.

    - *Weekday Average Main Activity Duration:* Mean value of all weekday main activity durations. The main activity of the day is the longest activity performed by a passenger during that day.
    - *Weekend Average Main Activity Duration:* Mean value of all weekend days main activity durations. The main activity of the day is the longest activity performed by a passenger during that day.
    - *Weekday Average Shortest Activity Duration:* Mean value of all weekday shortest activity durations. The shortest activity of the day is the activity with the shortest duration performed during that day.
    - *Weekend Average Shortest Activity Duration:* Mean value of all weekend days shortest activity durations. The shortest activity of the day is the

[1] Tour or circuit refers to a sequence of journeys and activities that start and end at the same location

activity with the shortest duration performed during that day.

- **Sociodemographic Characteristics**

  The sociodemographic characteristics of passengers also define their travel behavior. This category tries to encompass user social, economic and demographic characteristics that could affect their travel decisions. Fare policies associated with a user based on their sociodemographic or travel characteristics are used as the travel dimension to identify users groups.

  - Fare Discounts

    The fare discounts applied to a passenger's journeys can determine travel behavior. For example, transfers are free between Underground and buses for users with a Travelcard; therefore, these users are more likely to use bus as a feeder mode to their Underground journeys. Based on the London fare structure, the descriptive variables shown below are used to characterize different travel groups.

    - *Travelcard User:* Dummy variable that indicates if the user holds a Travelcard with a 7-day duration or longer.
    - *No Special Discount Adult:* Dummy variable indicating if the user is an adult not subject to any special discount other than a Travelcard discount, i.e. the user is not a child, student, elderly, disabled or staff member.

- **Public Transport Mode Choice**

  The extent and complexity of the London's public transport network allow users to move from one point to another using several mode combinations. The public transport modes used are indicative of network knowledge, age, or physical ability. The public transport modes were grouped into two categories: bus, specifically London Buses, and rail, including all the rail-based modes (Underground, Overground, National Rail and trams (Tramlink), light rail (DLR)). For this thesis, the following descriptive variables are explored.

  - *Percentage of Bus Exclusive Days:* Ratio of the number of days during which bus was the only public transport mode used to the number of days the passenger traveled.

- *Percentage of Rail Exclusive Days:* Ratio of the number of days during which rail was the only public transport mode used to the number of days the passenger traveled.

3.2 Data Needs and Sources

This section describes the data needs and sources used for the classification of Oyster Card users, including a description of the necessary tools and processes to compute all the descriptive variables. All the descriptive variables can be estimated using information about users completed journeys. The Origin-Destination (ODX) model (Gordon, 2012) which uses AVL data from London Buses and Oyster transaction data is used to obtain complete journey information for all passengers in the system. The Oyster Card data, iBus and the ODX inference tool are described in the following paragraphs.

3.2.1 Oyster Card Data

TfL's Oyster Card database stores records for all the transactions performed on every Oyster Card in the London public transport system. These transactions include travel related information, such as entries to (or exits from) Underground, Overground, and National Rail stations, bus boardings, and fare related actions, such as adding Pay as You Go travel value or checking travel credit balance. TfL retains Oyster Card data for eight weeks and for this research two periods of Oyster data were used: one week from Monday October 17 to Sunday October 23, 2011, and one week from Monday October 1 to Sunday October 7, 2012. Table 3-3 summarizes the database statistics for both periods.

The data contains entry information for all modes, and exit information for Underground, Overground, National Rail and DLR transactions at gated stations (and at ungated stations for Pay as You Go transactions). Each card is encrypted to protect privacy. Transaction data includes the entry/exit time stamp and station for rail, boarding time for bus trips, and the type of fare discount associated with the card (Travelcard or a special discount such as student child, staff or freedom pass). More detail on the information

Table 3-3: Database Statistics

Statistic	**Oct. 17th to 23rd, 2011**	**Oct. 1st to 7th, 2012**
Total Number of Records	64,322,400	66,749,210
Average Weekday Number of Records	10,315,381	10,663,580
Average Weekend Number Records	6,372,748	6,715,655
Total Number of Oyster Cards	5,578,850	5,825,498

contained in the Oyster Card data can be found in Gordon (2012).

3.2.2 iBus

iBus is London Buses' AVL system. Every vehicle in the fleet is equipped with this location system. The system uses the GPS, tachometers, speedometers, and gyroscopes installed on the buses to track their location. The iBus goal is to record time information about bus actions near stops. Four time stamps for bus actions are needed to successfully create a record: nearing the stop, opening doors, closing doors, and pulling away from the stop. Each record stores the door opening time as the arrival time and the door closing time as the departure time. When one of the door events is not available the time approaching or departing is used as arriving or departing time. If only one of the four time stamps was recorded, this time is used as both arrival and departure times (Gordon, 2012). On a typical day, iBus collects 5 million records. For this research iBus data from one week in October 2011 (17th to 23rd) and from one week in October 2012 (1st to 7th) was used.

3.2.3 ODX Full Journey Inference

Combining the data described in this section, Gordon (2012) developed a methodology and a tool, known as ODX, to infer trip origins and destinations and to link single trips into full multi-modal journeys. A summary of the origin, destination, and full linked journey inference methodology is presented below.

a. Origin Inference Process

Trip entry stations and stops can be inferred using the methodology developed by Gordon (2012). Rail entry stations can be inferred directly from Oyster Card transactions. iBus and Oyster Card records combined allow the inference of passengers' origin bus stops. The location of the stop is obtained from the iBus record by matching the Oyster Card time stamp and vehicle trip number record to the iBus record arrival time of that vehicle. In order to have a successful match, the Oyster Card time stamp must occur within a five-minute window of an iBus arrival or departure record. This algorithm applied to London's bus network infers over 95% of bus journey origins.

The inference of the origin location and time is necessary for those modes where there are no gated stations. However, there are cases where the inference of origin stops or stations is not feasible. Oyster Card users without a Travelcard are required to validate their card at the Oyster Card readers located on rail ungated stations' platforms (such as DLR). Users that hold a Travelcard do not have to validate their card at these stations, leaving no entry record in the fare system and making it impossible to infer their origin station.

b. Destination Inference Process

The zonal fare policy for rail requires that all Oyster Cards be validated at the card readers to exit a gated station, recording the destinations of almost every Oyster Card journey in the rail system. The journeys of those passengers who did not validate their card at ungated stations are not included in these records. London bus flat fare structure does not require passengers to validate their Oyster Cards when alighting a bus, requiring a more complex process for inferring alighting locations.

The methodology presented in Gordon (2012) assumes that a passenger's alighting location is the closest stop to the user's next bus boarding or station entry. The inference is based on the assumption that passengers do not walk long distances or use non-public transport modes between Oyster Card journey stages. For the last Oyster Card record of a day, it is assumed that the alighting location of that trip is the stop closest to the origin of that day's first trip. The algorithm infers over 75% of all destinations. Most of the non-inferred destinations are due to cases where Oyster and iBus record times do not match within 5 minutes.

c. Linked Journey Inference Process

The algorithm's next stage is to use the inferred origins and destinations to link these trips into journeys, generating multi-modal journey records for each Oyster Card user's daily travel. The methodology is based on several binary, temporal, and spatial conditions that are applied to infer whether or not trip segments are linked.

All the parameters of the algorithm, such as maximum and minimum distances, times, and speeds, can be easily modified using the associated Graphical User Interface. For this thesis, the parameters were chosen based on specific operations and the geography of the London public transport network. At least 22% of all the journey segments made in a normal weekday in London can be linked using this algorithm. This percentage is directly related to the origin and destination inference rates presented in the preceding paragraphs and also reflects many journeys that have only single segments and should not be linked.

3.2.4 Sampling Strategy

Using the data sources described in 3.2.1 and 3.2.2 as an input for the ODX inference tool described in 3.2.3, full journey information was obtained for the 2011 and 2012 periods that are used for this analysis. The 5.6 million cards observed during the 2011 period represented 50.8 million completed journeys and 64.3 million individual trip stages. The 5.8 million cards active during the 2012 period represented 52.2 million linked journeys and 66.7 individual trip stages. Given this extensive database, the computation of the variables described in 3.1.2 has a very high computational cost which makes it infeasible to use the complete sample for the analysis. Therefore, a sample of the data was used which made it possible to estimate the descriptive travel variables accurately while saving resources. A simple random sample is chosen from both the 2011 and 2012 periods based on the minimum sample size estimated below. The travel behavior and sociodemographic characteristic of each sample are described in Section 3.3.

The minimum sample size is a function of the desired accuracy and level of confidence. Additionally, information about the variability of the travel characteristics within the population is required. Since the population characteristics are unknown and there is no previously known information about the variability of the descriptive variables in the population, a random sample of approximately 250,000 Oyster Cards was selected for each period (2011 and 2012). Oyster Card travel variables were computed using these two random samples. The sample size required is given by

$$N_s = \frac{Z^2_{(\alpha/2)}\left[\frac{\sigma}{\mu}\right]^2}{d^2}. \tag{3.6}$$

Where, $Z_{(\alpha/2)}$ is value at $1-\alpha$ confidence level of a standard normal distribution ($\mu = 0$, $\sigma = 1$), and d is the allowable % error (Ben-Akiva and Lerman, 1985). For each variable, the sample mean $\bar{X}$ and sample standard deviation S are used as estimators of the population mean and standard deviation (μ, σ). This is a reasonable assumption since as the sample size N_s becomes large, the sampling distribution approaches the normal distribution with mean μ and variance σ^2/N_s, which is independent of the variable's distribution in the population (Central Limit Theorem, Billingsley, 1995).

Using an allowable error of 1% and a confidence level of 95% ($Z_{(\alpha/2)} = 1.96$), N_s was estimated for each travel variable in each period sample. The minimum value of N_s required is given by the variable with the highest coefficient of variation σ/μ. For the 2011 one week sample, the minimum sample size required is 143,000 Oyster Cards. Therefore, the chosen sample size of 250,000 is more than adequate for the required accuracy.

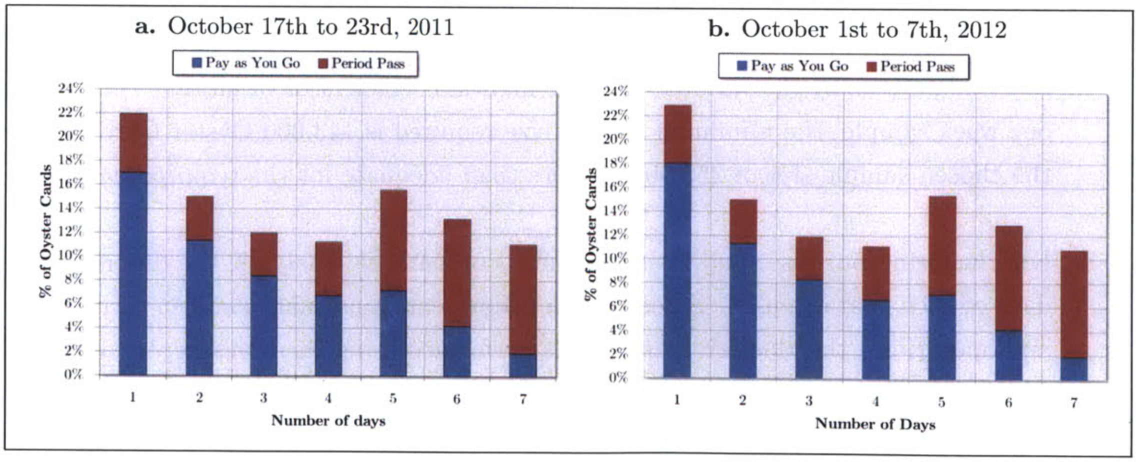
For clustering purposes and for subsequent comparative analyses, the described weeks of data of each year are used. Since October represents a normal month in terms of demand and operations, the clustering analysis is performed using random samples of size 250,000 for both years. Additionally, a three-week sample 2012 is used to analyze some of the population characteristics, where 3.2% of the whole population was randomly chosen. Below, the representativeness and characteristics of each sample are described in detail.

3.3 Sample Characteristics

The following paragraphs describe and compare the Oyster Card users' travel behavior observed during 2011 and 2012, which leads to the subsequent classification of passengers. For computational reasons, the minimum sample sizes defined in 3.2.4 are used and descriptive sample statistics of the important variables are presented. The travel characteristics can be explored by looking at the descriptive travel variables computed for each Oyster Card. Having a general knowledge of the travel patterns of the passengers as a whole can provide an initial idea of how the passenger demand is segmented.

3.3.1 Travel Frequency

The travel days frequency distribution is shown in Figure 3-1. The graphs show the number of days passengers use their Oyster Cards. The red bar represents those cards that for most of their weekly journeys used a Period Pass. For this analysis, the term Period Pass refers to any Travelcard (child, student, or adult) and all freedom passes (elderly or disabled). The results in both years show two peaks: one day and 5 days a week, similar travel behavior to that observed in other big cities such as Santiago, Chile (Coordinacion Transantiago, 2010), and Kochi City, Japan (Nishiuchi et al., 2013).


a. October 17th to 23rd, 2011

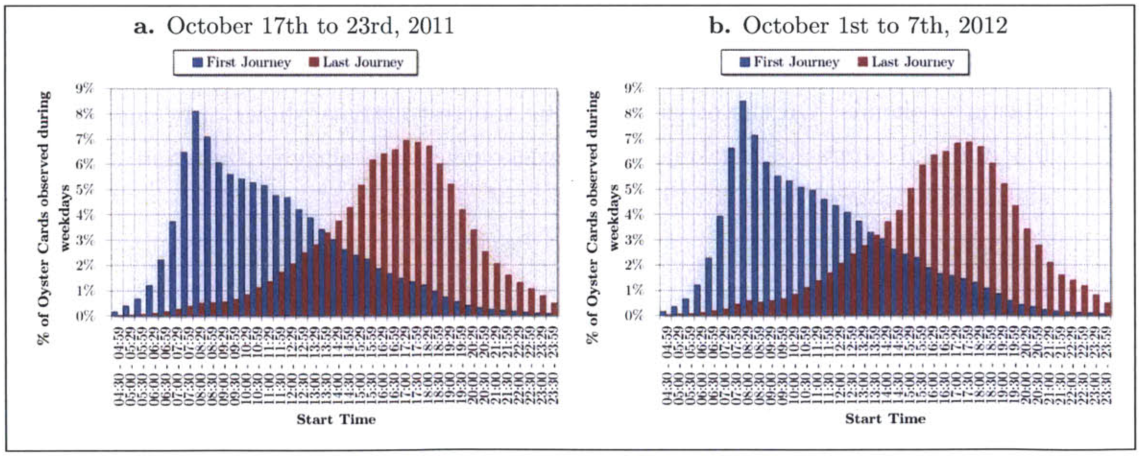
b. October 1st to 7th, 2012

Figure 3-1: Days of Travel

On average, people traveled 4 days during both analysis weeks in October 2011 and 2012. Additionally, the use of Period Passes increases with the frequency of travel. This is expected since Period Pass holders are subject to fare discounts that encourage more travel. For both 2012 and 2011 periods, the Oyster Cards users make on average 2.5 journeys a day, except for Sundays where 2.3 journeys are made.

3.3.2 Journey Start Time

Figures 3-2 and 3-3 show the 2011 and 2012 distribution of average journey start times for weekdays and weekends respectively. The blue bars show the percentage of users that on average start their first journey of the day in the half hour indicated on the x axis. The red bars indicate the start time of the last journey of the day. The graphs in Figure 3-2 clearly illustrate the peaks (from 7 to 9 am for the first journey and from 5 to 6:30 pm for the last journey). Weekends present two less sharp peaks that occur later than for weekdays, specially in the morning. Very similar behavior can be observed in 2011 and 2012, with a slightly higher morning peak peaks during weekdays for 2012.



a. October 17th to 23rd, 2011

b. October 1st to 7th, 2012

Figure 3-2: Average Weekday Journey Start Time

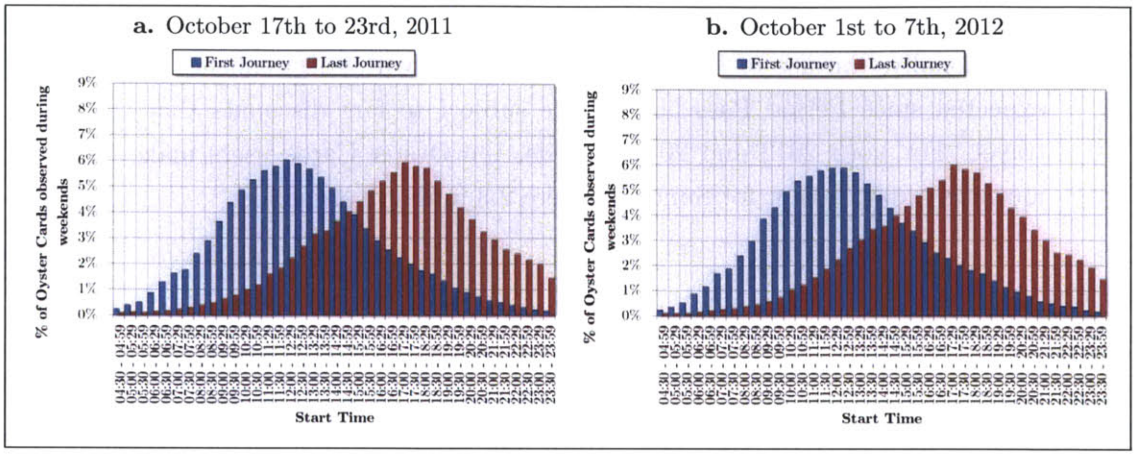


Figure 3-3: Average Weekend Journey Start Time

3.3.3 Activity Duration

The activity duration at the end of a journey is estimated as the time between the destination or exit time (inferred in the case of bus trips) and the next entry transaction. The average activity duration distribution for the main and shortest activity of the day for 2011 and 2012 is shown in Figures 3-4 and 3-5 for weekdays and weekends respectively. As can be seen from the graphs, the main activity shows two peaks during weekdays: between 1.5 and 3 hours, and between 8 and 9 hours. Only the first peak is observed during weekends. On the other hand, the shortest activity shows only one peak between 0.5 and 1.5 hours for weekends and weekdays. The activity duration does not show significant changes from 2011 to 2012.

3.3.4 Origin Frequency

The distribution of the number of different origin stops/stations that passengers use for their first and last journeys during weekdays is presented in Figure 3-6 for both the 2011 and 2012 periods. The graphs show that users have fewer different origins for the first journey of the day than for the last one. This may indicate that for most passengers the first journey starts at their home station/stop and the difference with the last journey number of different origins may depend on their travel purpose or travel regularity. The

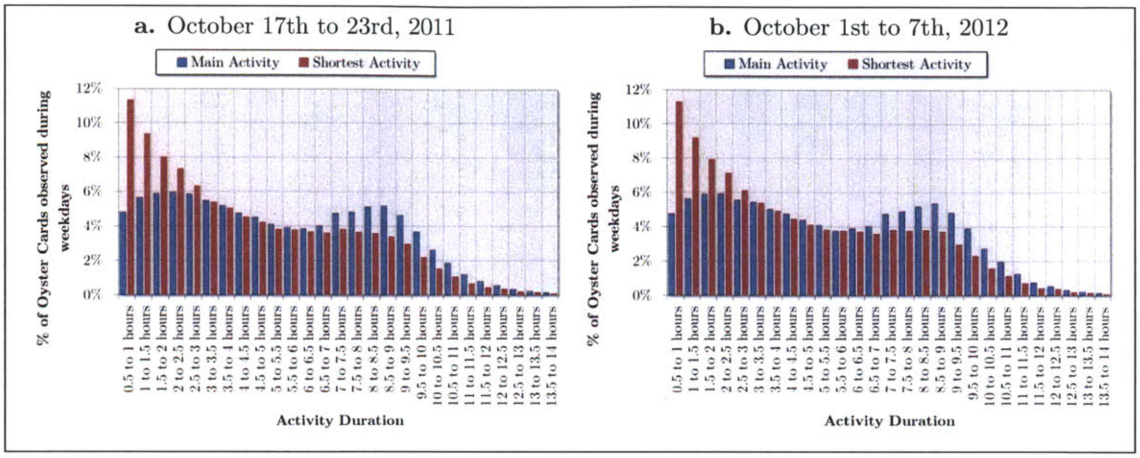


Figure 3-4: Weekday Average Activity Duration

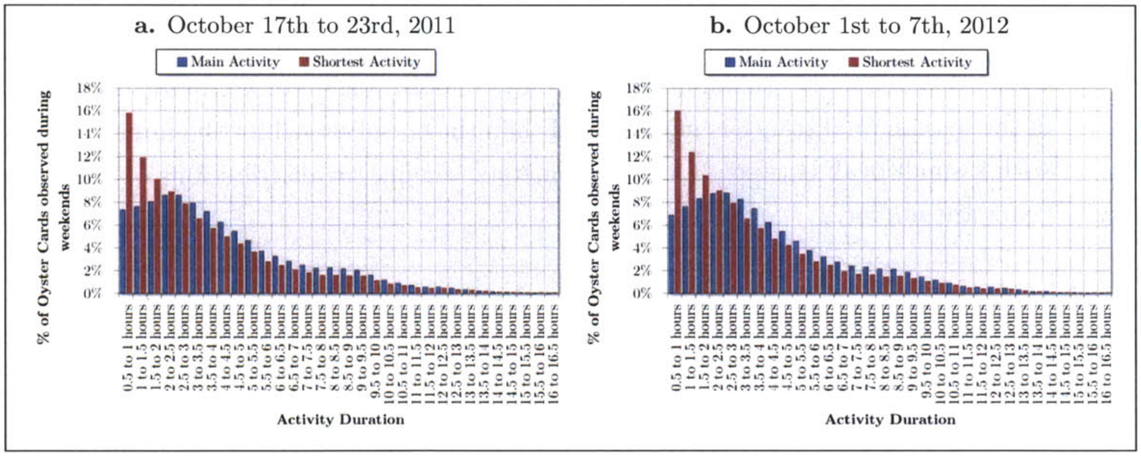


Figure 3-5: Weekend Average Activity Duration

results do not show significant change from 2011 to 2012.

3.3.5 Travel Distance

Figure 3-7 shows the distribution of the maximum and minimum distance passengers traveled during the 2011 and 2012 analysis periods. As described in 3.1.2, these values are based on the Euclidean distance between the geographic coordinates of the journey

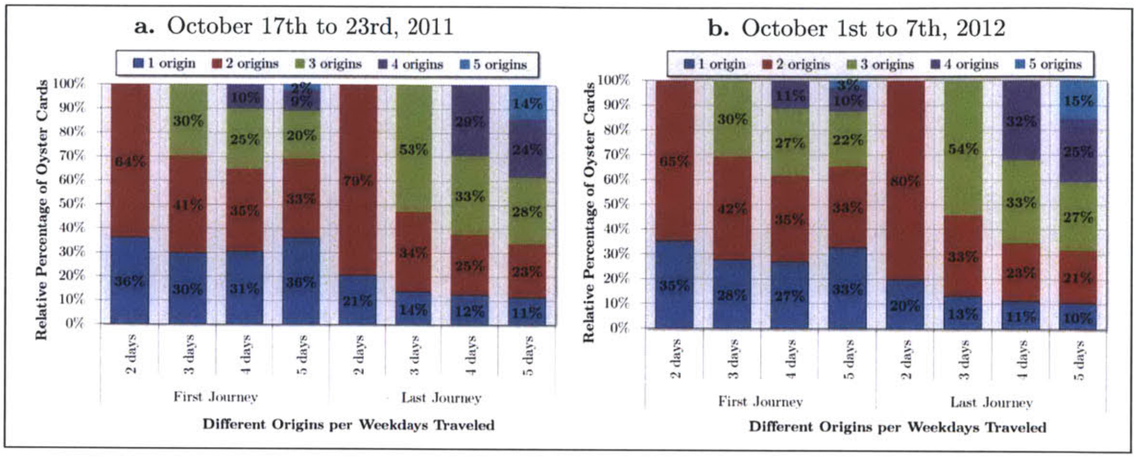


Figure 3-6: Different Origin Stops/Stations

origin and destination stop or station. As can be seen from the graphs, a high number of passengers has short (1-2 kilometer) journeys in both years. Additionally, the distribution of the maximum distance is more spread than the one observed for the minimum distance, with a small peak between 3 and 5 kilometers. Again, the distributions observed in both years are similar.

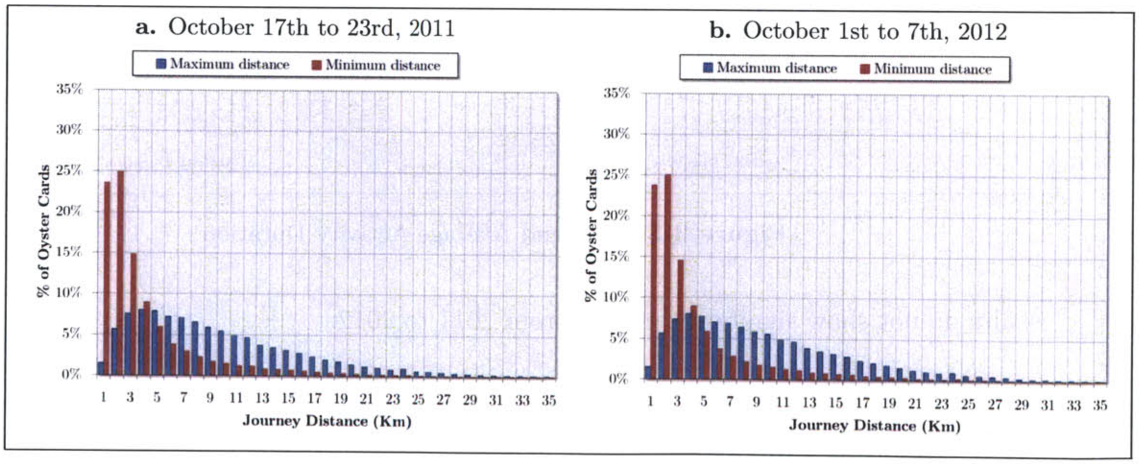


Figure 3-7: Maximum and Minimum Travel Distance

3.3.6 Mode Choice

A summary of the mode choices passengers made during the 2011 and 2012 analysis periods are presented in Table 3-4, which shows the percentage of passengers that use bus exclusively or rail exclusively for all their weekly journeys. As can be seen, the percentage of passengers that use only bus for all their journeys is slightly higher than the percentage that use only rail. During both years, a high percentage of users use bus and rail every day they travel. The percentages observed for both years are similar, with a slight increase in rail usage during 2012.

Table 3-4: Mode Choice Distribution

Weekly Mode Choices	2011	2012
Use only bus every day	34.7%	33.8%
Use only rail every day	22.7%	23.4%
Use rail and bus every day	40.6%	41.0%
Any other combination	2.0%	1.9%

3.3.7 Sociodemographic Characteristics

Table 3-5 summarizes the most relevant sociodemographic characteristics and Oyster Card features for the two samples used in the classification analysis. As can be seen, the demographic characteristics are very similar which make the samples comparable.

The results presented above show that there are more similarities than differences in travel behavior characteristics between the 2011 and 2012 analysis periods. This also occurs for all the other variables analyzed, which indicates that the 2011 and 2012 samples are comparable.

Table 3-5: Oyster Card Features

Statistic	**Oct.17-23 2011**	**Oct.1-7 2012**
Percentage of Registered Cards	47.3%	46.3%
Percentage of Travelcards	43.3%	42.6%
Percentage of Elderly Passes	10.8%	10.2%
Percentage of Disabled Passes	1.88%	1.81%
Percentage of Student/Child Passes	8.85%	8.78%
Percentage of Staff Passes	1.34%	1.31%
Percentage of Visitor Cards	0.43%	0.42%

3.4 Clustering Process

Given that there is no previously known information about passenger categories based on their travel patterns, a clustering process needs to be performed to identify travel patterns of Oyster Card users. The classification of Oyster Card users is performed in this section applying the K-medoids clustering method described in 3.1.1, using the descriptive variables obtained from the October 2011 one-week sample of Oyster journeys defined in 3.1.2. Given the large sample size, hierarchical clustering methods are not feasible due to their high computational cost. As described in 3.1.1, partitional clustering methods are the best option for applications involving large data sets. For this thesis, the K-means algorithm and its generalized K-medoids algorithm are used. These are simple but powerful clustering methods and the K-means is the most commonly used method in travel demand classification. The following paragraphs show the classification process, starting with the selection of the optimal number of clusters, continuing with cluster variables and results validation, and ending with a summary of the characteristics of each cluster.

3.4.1 Optimal Number of Clusters

Two measures were used to define the optimal number of clusters: the within-cluster variation, and the Davies-Bouldin index that measures average similarity between each cluster and its most similar one. K-means and K-medoids clustering processes were performed for different number of clusters K, using the variables described in 3.1.2 and the sample data from 2011 described in 3.2.4. For each value of K, K-medoids always had lower within-cluster variation and lower Davies-Bouldin (DB) index, indicating better cluster configuration. Indeed, K-medoids builds the clusters using a representative individual as cluster center, which is more appropriate for classifying travel patterns than using the cluster average.

The K-medoids clustering process was performed using the CLARA algorithm implemented in the R package (Maechler et al., 2013). The values of within-cluster variation and the DB index are shown as functions of the number of clusters K in Figure 3-8. The within-cluster variation decreases as the number of clusters increases; however, there is a point beyond which there is relatively little gain from further increase in the number of clusters. As can be seen from the graph, the last significant drop of the within-cluster variation occurs for $K = 7$; however, the DB index shows the last significant drop for $K = 8$. The DB index curve also indicates that $K = 10$ could be a potential optimum for the number of clusters; however, using K higher than 8 only generates smaller clusters with less distinctive characteristics. Therefore, 8 clusters were selected. For these eight clusters, the two principal components visually show that most of the clusters are separated from each other (Figure 3-9). Examining the individual medoid of each cluster is also useful to validate the number of clusters.

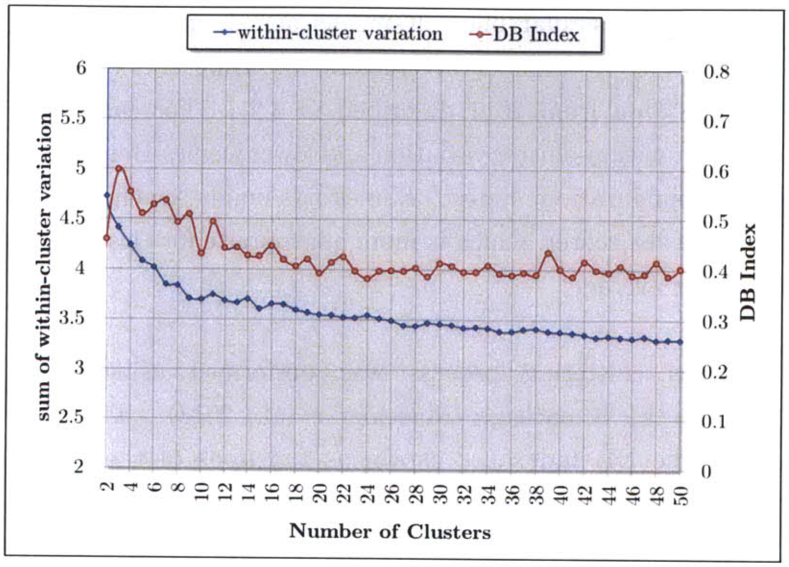


Figure 3-8: Within-Cluster Variation and DB Index per Number of Clusters - K-medoids

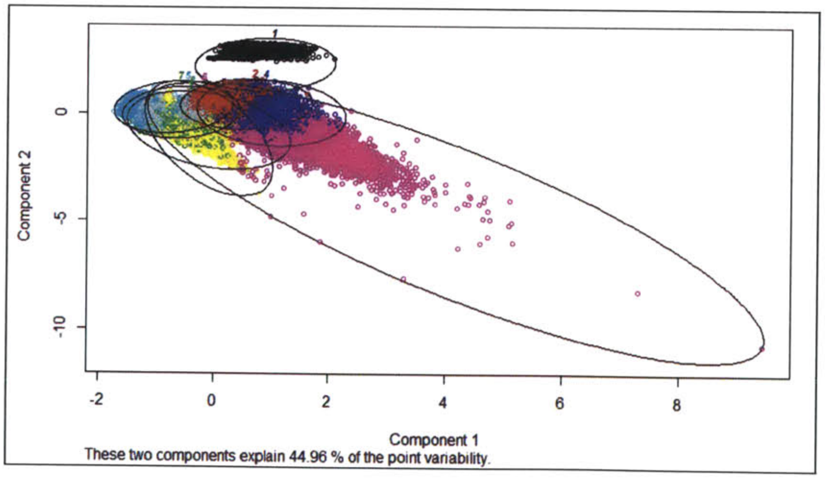


Figure 3-9: Principal Components showing K-medoids with $K = 8$

3.4.2 Clustering Analysis

The K-medoids clustering process with eight clusters provides not only information about each cluster medoid characteristics but also information about the average characteristics of each cluster. The smallest cluster contains 8% of the passengers in the sample, and the largest one contains 19%. Figure 3-10 shows each cluster as a percentage of the entire sample of 250,000 Oyster Cards.

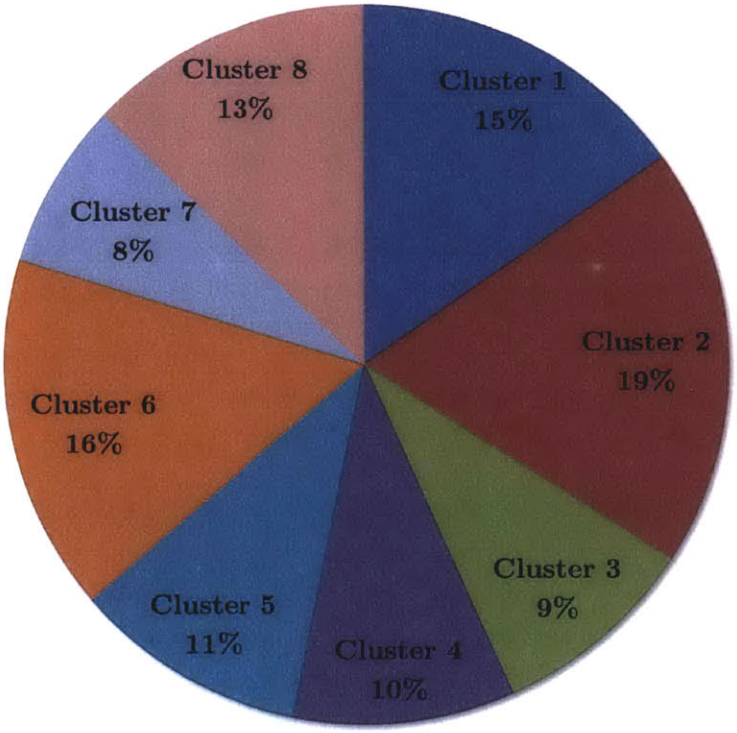


Figure 3-10: Cluster Size

Examining the characteristics of each cluster, one of the clearest differences between them is the frequency of travel. Figure 3-11 shows the distribution of the travel days per week by cluster. The graph shows that clusters 1, 2, 3 and 4 have the highest frequency of travel days, while cluster 5, 6, 7 and 8 the lowest. The first group was categorized as ***regular users (clusters 1 to 4)*** traveling 4 days a week or more, and the second group was categorized as ***occasional users (clusters 5 to 8)*** traveling less than 4 days a week.

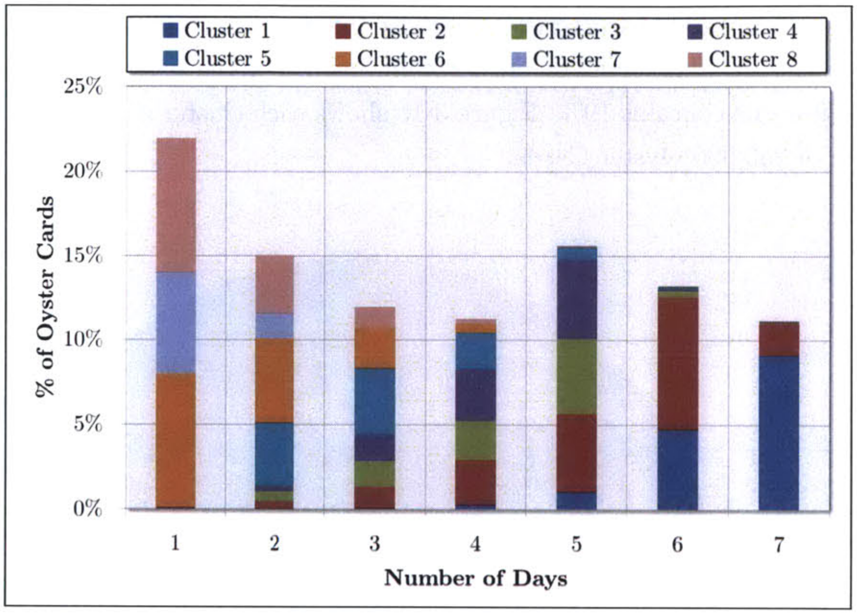


Figure 3-11: Number of Travel Days by Cluster

The weekday main activity duration distribution for members of different clusters is presented in Figure 3-12. The graph shows that the distribution of cluster 3 members is focused around 7.5 to 10 hours, and most cluster 8 members tend to have activity durations between 1.5 and 4 hours. It can also be seen, that the distributions of clusters 5 and 6 are concentrated around activities that last less than 4 hours. While clusters 2 and 4 show a distribution between 0.5 and 10 hours, cluster 1 activities are mostly between 6 and 9.5 hours. Cluster 7 does not appear in this graph since its members travel only on weekends. Their main activity has average duration of 2 hours.

Figure 3-13 shows the distribution of the start time of the first and last journeys on weekdays for different cluster members. The start time of the first journey of the day for those clusters categorized as regular users (clusters 1 to 4) is mostly focused between 7:30 and 9:30 am, although some of them present more spread distributions than others (clusters 1 and 2). Clusters 5, 6, and 8 present first journey distributions spread over the

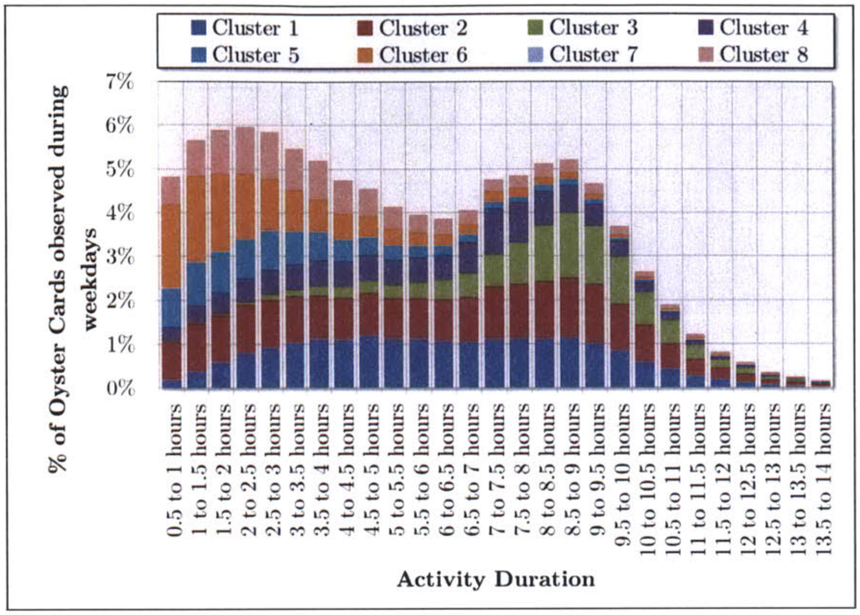


Figure 3-12: Weekday Main Activity Duration by Cluster

day; however, cluster 5 shows a tendency to afternoon journeys and clusters 6 and 8 have more midday trips. The start time distribution for the last journey of the day is more spread for all clusters. Regular user clusters tend to make their last journey between 4:00 pm and 8:00 pm. However, cluster 1 members tend to travel after 5:00 pm and cluster 4 members show a tendency to travel before 6:00 pm. Occasional user clusters last journey start times are spread over the day, with more during the afternoon. Notice that, especially for occasional users, the first and last journey of the day will be the same when only one journey per day is made. Again, cluster 7 is not included in this graph because its members travel only on weekends. On average, they start their first journey around 1:00 pm and their last journey around 5:00 pm.

The Oyster Card fare discount composition of each cluster is presented in Figure 3-14. Cluster 1 includes the highest percentage of Travelcard users (81.4%), followed by clusters 4 (54.9%) and 2 (54.3%). Cluster 8 is composed mainly of Pay as You Go users (only 13.1% rely on Travelcards). 40.3% of cluster 4 members have an Oyster Card associated with a special discount other than a Travelcard. Only 6.7% of cluster 3 members hold a

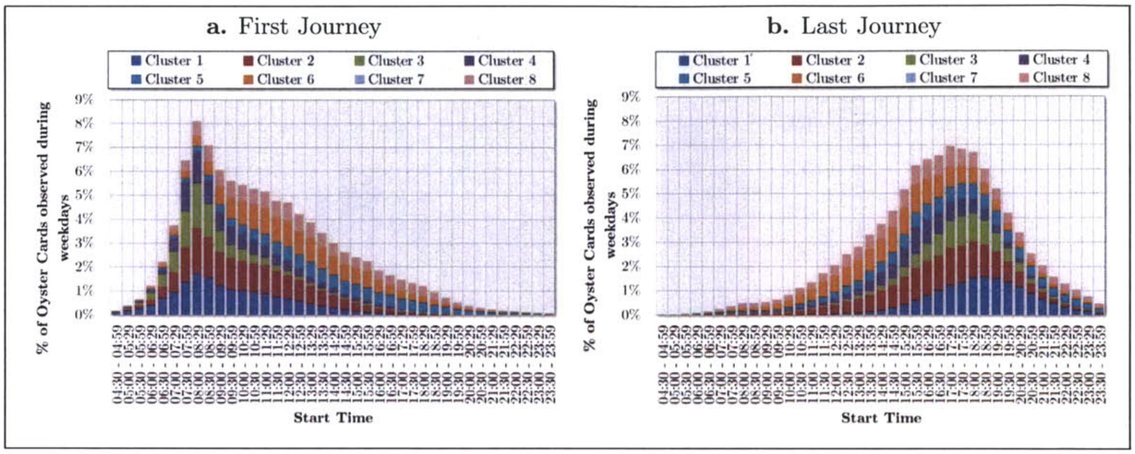


Figure 3-13: Start Time of First and Last Journey of the Day, Weekdays

discount card.

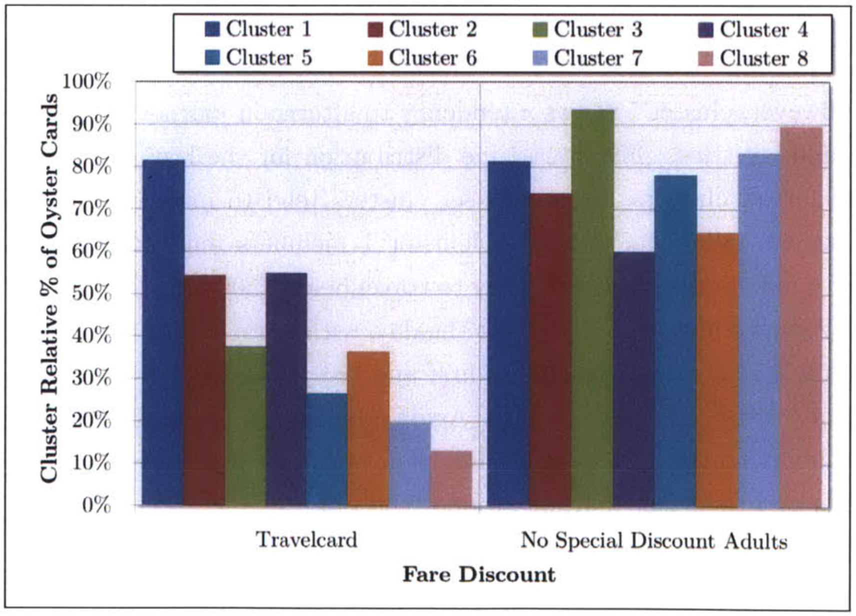


Figure 3-14: Fare Discount Distribution by Cluster

As can be seen in Figure 3-15, 73% of all Travelcard users belong to clusters categorized as regular users (1 to 4). 29% of Travelcard users belong to cluster 1 and only 3% to cluster 7. This result is expected, given that cluster 1 members travel 7 days a week and cluster 7 users only travel during weekends.

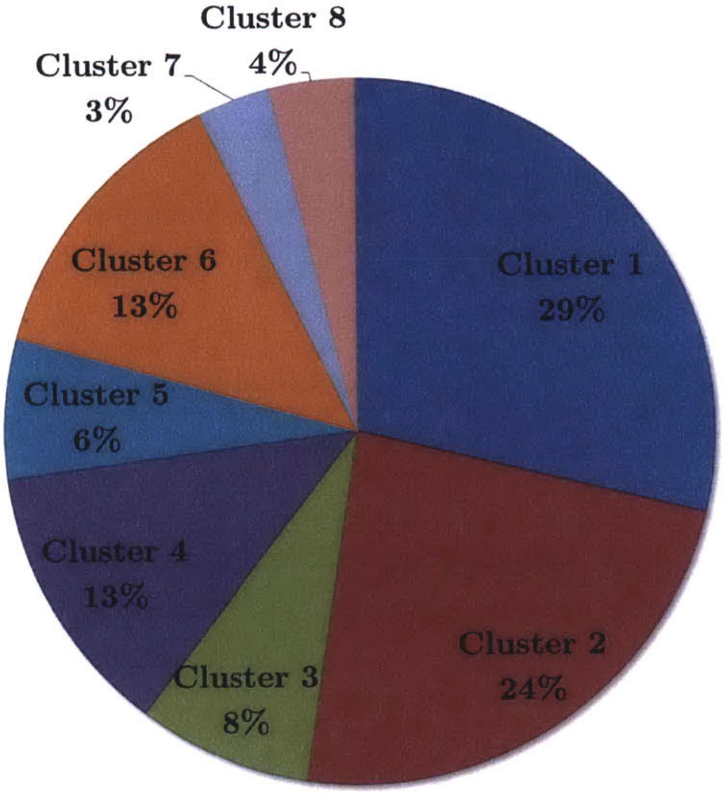


Figure 3-15: Period Pass Distribution by Cluster

Figure 3-16 illustrates the cluster distribution among different special discount holders. The graph shows that approximately 20% of each discount group are members of cluster 2, and less than 6% are members of cluster 7. While elderly freedom passes comprise 32% of cluster 6 members, 30% of child and student passes are members of cluster 4. Cluster 6 also includes a significant percentage of staff members (20%) and disabled pass holders (24%).

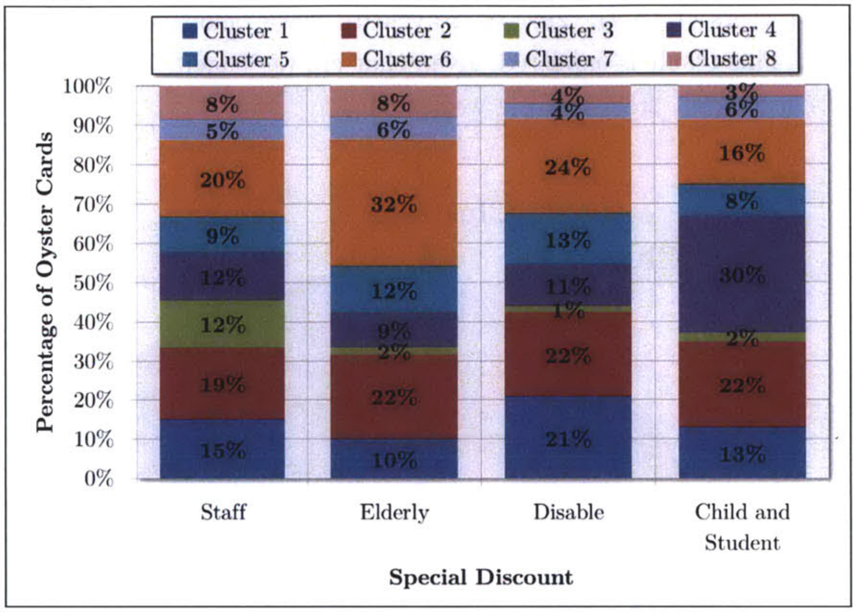


Figure 3-16: Special Discount Distribution by Cluster

The travel characteristics of each cluster are summarized below. Each cluster is described based on the travel characteristics of its medoid which is the representative individual found during the clustering process. Clusters were numbered from 1 to 8, starting from the highest to the lowest frequency of travel.

- **Cluster 1: Everyday regular users**

  The medoid of this cluster travels all 7 days of the week, making 2 or 4 journeys per day. The first journey of the day starts at approximately 8:30 am during weekdays and at 9:30 am during weekends. During weekdays, the last journey of the day starts at 7:30 pm and at 6:15 pm during weekends. During weekdays, the shortest activity of the day lasts 3.6 hours on average and the main activity lasts 5.4 hours. Additionally, the distance between the origin and destination of their journeys varies between 1 and 11 kilometers, a large range that is explained by the high number of journeys this individual performs. During four weekdays, this cluster's medoid have one origin for the first and the last journey of the day, and also presents one origin for the first and last journey on both weekend days. Only bus is used during 5 days and only rail is used for 1 day. The medoid of this cluster holds an elderly freedom pass.

- **Cluster 2: All week regular users**

  The medoid of this cluster travels 6 days a week (5 weekdays and 1 weekend day), making 1 or 2 journeys per day. The first journey of the day starts at 10:30 am during weekdays and at 1:30 pm during weekends. The last journey of the day starts at 4:30 pm during weekdays and at 5:00 pm during weekends. During weekdays, the shortest activity of the day lasts 2.5 hours on average and the main activity lasts 5.3 hours. The distance that this medoid travels is between 4 and 7 kilometers. During four weekdays, this cluster's medoid has the same origin for the first journey of the day, and only during three weekdays has the same origin for the last journey of the day. This individual uses only bus 4 days a week and uses a combination of rail and bus the remaining days. This individual is not a Travelcard holder nor has special fare discount.

- **Cluster 3: Weekday rail regular users**

  Cluster 3 medoid travels all 5 weekdays, making 2 daily journeys. On average, this individual's first journey of the day starts at 7:30 am and the last journey of the day starts at 3:30 pm. This cluster medoid performs one activity per day that last on average 7.4 hours and uses only one origin for both the first and last journey of the day for four days a week. All journeys are made using rail and the travel distance varies between 8 and 12 kilometers. The medoid of this cluster is a Travelcard holder with no special fare discount.

- **Cluster 4: Weekday bus regular users**

  Cluster 4 medoid travels all 5 weekdays, making 2 daily journeys. The first journey of the day starts at 9:30 am and the last journey of the day starts at approximately 4:00 pm. On average, the shortest activity of the day lasts 2.5 hours and the main activity lasts 7.2 hours. The medoid has the same origin for the first journey of the day for four weekdays, and during three weekdays has the same origin for the last journey of the day. This individual uses only bus, travels between 3 and 4 kilometers, and holds a child bus and tram period pass.

- **Cluster 5: All week occasional users**

  The medoid of this cluster travels 3 days a week (two weekdays and one weekend day), making 1 or 2 journeys per day. During weekdays there is only one trip per day; therefore, the first and last journey of the day are the same and starts at approximately 6:00 pm. Weekend first journey starts at 11:30 am and the last journey at 3:30 pm. The medoid of this cluster travels between 3 and 5 kilometers during the week and presents different origins for all its first and last daily journeys. Bus and rail are both used by this individual, who does not hold a Travelcard and has no special fare discount.

- **Cluster 6: Weekday bus occasional users** The medoid of cluster 6 travels only 2 weekdays, making only 1 journey per day. This individual journeys start at 2:30 pm which are only-bus journeys of 2 to 8 kilometers of distance. All journeys have different origins. The medoid of this cluster does not hold a Travelcard and has no special fare discount.

- **Cluster 7: Weekend occasional users**

  The medoid of this cluster travels 2 days a week, Saturday and Sunday, making 2 journeys per day. The first journey of the day starts at 5:30 pm and the last journey at 8:00 pm, with a maximum activity duration of 1.8 hours. This cluster's medoid presents different origins for all its first and last daily journeys and travels between 6 and 7 kilometers. This medoid uses a mix of rail and bus, is not a Travelcard nor a discount card holder.

- **Cluster 8: Weekday rail occasional users**

  Cluster 8 medoid travels only 1 weekday, performing 1 journey. This journey starts at 2:00 pm, is made using rail and is 7 kilometers long. This medoid is not a Travelcard holder and has no special fare discount.

The most distinctive characteristics of each cluster medoid described above are summarized in Table 3-6.

Table 3-6: Summary of Cluster Characteristics

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Type of User	Regular	Regular	Regular	Regular	Occasional	Occasional	Occasional	Occasional
Days of the Week	All Week	All Week	Weekdays	Weekdays	All Week	Weekdays	Weekend	Weekdays
Days of Travel	7	6	5	5	3	2	2	1
Journeys per day	2 to 4	1 to 2	2	2	1 to 2	1	1	1
Preferred mode	Mix	Mix	Rail	Bus	Mix	Bus	Mix	Rail
First Journey Start Time (weekday/weekend)	8:30 am / 9:30 am	10:30 am / 1:30 pm	7:30 am / -	9:30 am / -	6:00 pm / 11:30 am	2:30 pm / -	- / 5:30 pm	2:00 pm / -
Last Journey Start Time (weekday/weekend)	7:30 pm / 6:15 pm	4:30 pm / 5:00 pm	3:30 pm / -	4:00 pm / -	6:00 pm / 3:30 pm	2:30 pm / -	- / 8:00 pm	- / -
Main Activity Duration Time (weekday/weekend)	5.4 hrs / 4.1 hrs	5.3 hrs / 2.7 hrs	7.4 hrs / -	7.2 hrs / -	- / 4 hrs	- / -	- / 1.8 hrs	- / -
Journey Distance	1 to 11 km	4 to 7 km	8 to 12 km	3 to 4 km	3 to 5 km	2 to 8 km	6 to 7 km	7 km
Journey Origins	Mostly one	Mostly one	Mostly one	Mostly one	All different	All different	All different	All different
Type of Card	Elderly Pass	Adult PAYG	Adult Period Pass	Child Bus Period Pass	Adult PAYG	Adult PAYG	Adult PAYG	Adult PAYG

Figure 3-17 shows the normalized values of the clusters' centroids for all week, weekday and weekend variables. This normalization was made subtracting from each cluster centroid the average value of the eight centroids and dividing them by the standard deviation of all the centroids. The graph allows the visual identification of the relative travel characteristics of the clusters, for example, cluster 6 has the highest frequency of travel days, cluster 8 the shortest distances traveled, and cluster 3 shows the earliest weekday first journey start times. The standardization of the cluster centers also facilitates the identification of relative differences between clusters for different travel characteristics. For example, although clusters 7 and 8 have very close values of bus exclusive days, they have closer values on maximum distance of travel. This visual representation is very helpful for the interpretation of the clusters and for the travel analyses provided in subsequent chapters.

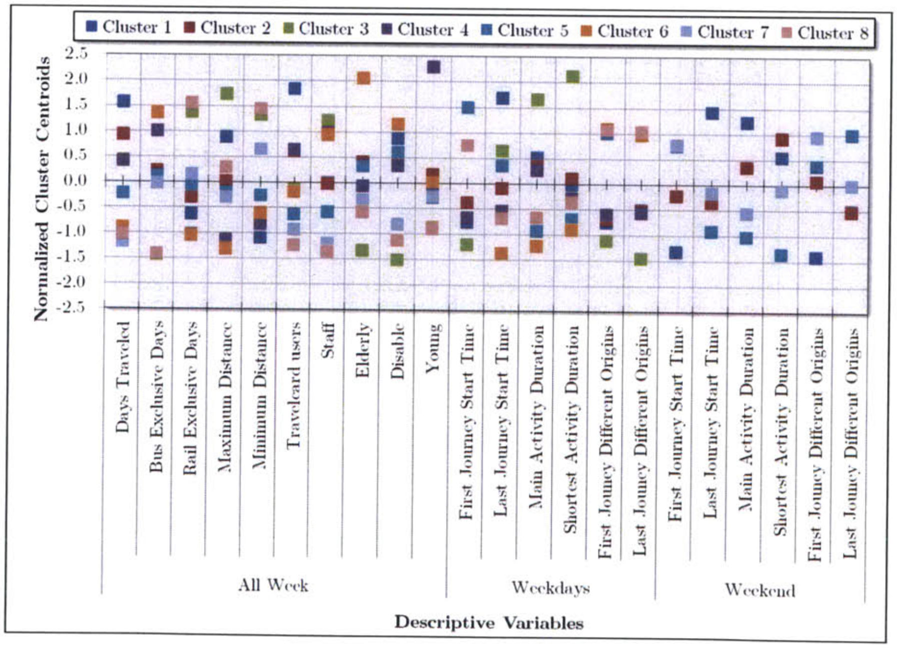


Figure 3-17: Normalized Cluster Centers

3.4.3 Cluster Initial Interpretation

An initial interpretation of the resulting clusters is provided below based on the analysis of the descriptive variables presented above.

- **Commuters**

  Regular clusters 3 and 4 seem to be mostly composed of commuters. Most characteristics of clusters 3 and 4 are very similar with exception of the mode choice (cluster 3 users prefer rail and cluster 4 users prefer bus) and the percentage of special discount cards (40% of cluster 4 members vs. 6% of cluster 3 members). Additionally, cluster 4 members' activities are shorter (5.8 hours) than cluster 3 members' activities (8.1 hours). This travel behavior is consistent with commuters behavior; however, cluster 4 travel behavior is suggestive of student travel behavior, while cluster 3 members behavior evidences typical worker travel patterns. Students are more likely to have more bus trips because of lower fares and fixed fare structure, and usually school days are shorter than workdays. These two clusters could be merged into one group composed of both students and workers that use the system only during weekdays with the main purpose of commuting (exclusive commuters). They are likely to be Pay as You Go users who benefit from daily fare capping.

  During weekdays, clusters 1 and 2 travel behavior also shows commuter travel behavior characteristics, but during weekends their travel behavior seems to be for leisure purposes. These are the clusters with the highest frequency of travel (more than 6 days per week) and they have very similar characteristics. Both clusters have main activities that last approximately 6 hours. However, cluster 1 members perform more than 1 daily activity (between 1 and 3) and the average difference between the start time of their first and last journey of the day is 8.5 hours, which is consistent with work activity. Their weekend journeys seems to be for leisure, with shorter activity durations (approximately 4 hours maximum) and starting the first journey of the day around the midday (12:00 to 1:00 pm). The main difference between these two clusters is that most of cluster 1 members are Travelcard holders (81%), while 54% of cluster 2 are Travelcard holders. Cluster 2 activity durations are shorter and similar to the values observed for cluster 4, which might be indicative of student travel patterns, but with less certainty given that only 10% of the members are student card holders. These two clusters could be merged into one group composed of students and workers who use the system during weekdays with the main purpose of commuting and during weekends for leisure purposes, probably taking advantage of the unlimited travel that most of the member enjoy.

Figures 3-18 and 3-19 show the difference during weekdays in main activity duration and journey start times between the two types of commuters described above. Exclusive commuters show a sharp peak for activities that last between 6.5 and 10 hours. Non-Exclusive commuters show a more spread distribution of activity durations, which is consistent with a higher number of daily activities. Potential-student clusters (2 and 4) have similar activity duration distribution. They show a uniform distribution for activity duration lasting less than 6.5 hours and peaks for activity duration of approximately 8 hours. As expected, first and last journey start times show morning and afternoon peaks. Potential-worker clusters (1 and 3) show sharper peaks than the other clusters (2 and 4), which is consistent with the expected worker and student travel hours.

The non-commuting behavior shown during weekends by cluster 1 and 2 members is illustrated in Figure 3-20. The main activity duration distribution show peaks for less than 4 hours and the first and last journey of the day start times are normally around midday.

- **Non-Commuters**

  Clusters 5 to 8 show travel characteristics that are not typical of commuters. The duration of their activities are between 2 and 4 hours and their journeys start during off-peak hours, which is consistent with activities for leisure, recreational, or sporadic work purposes. Analyzing similarities between clusters, clusters 5 and 6 seem to have common characteristics as well as clusters 7 and 8.

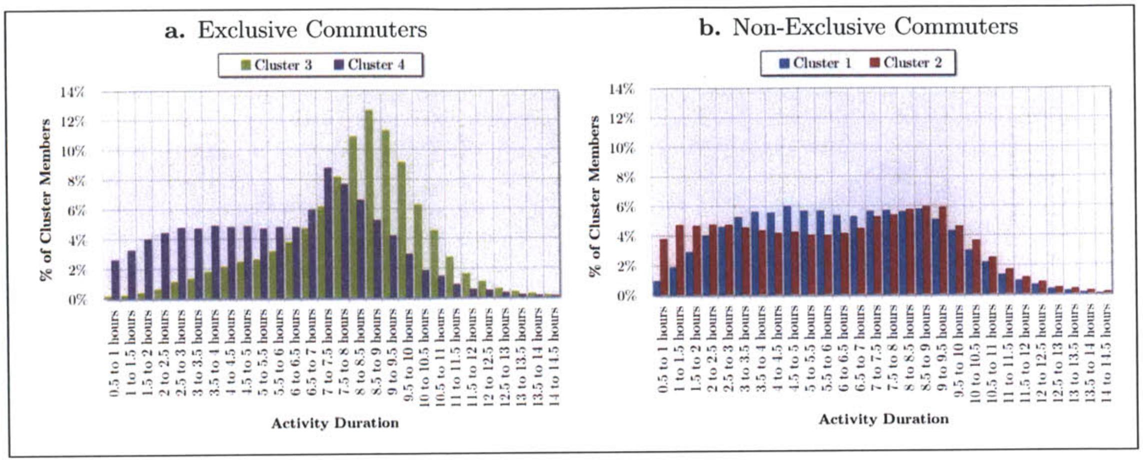


Figure 3-18: Commuters Main Activity Duration

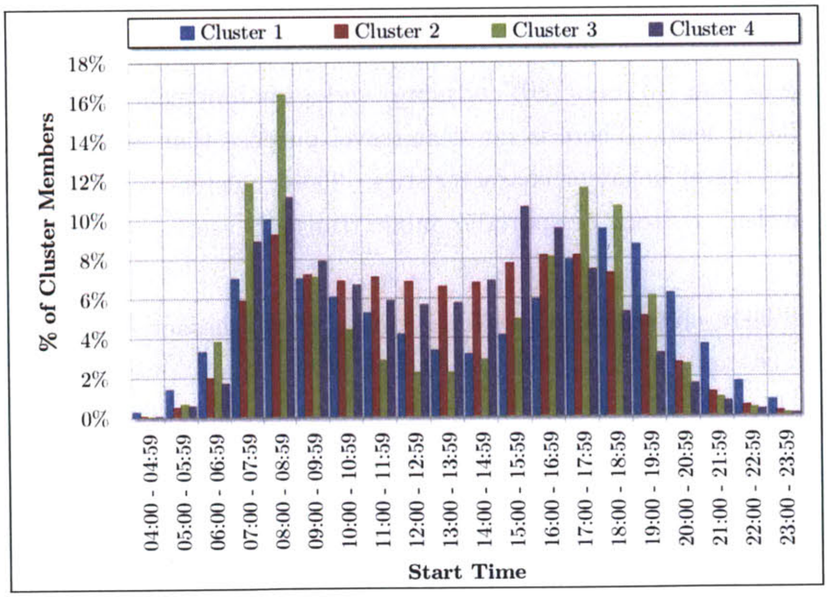


Figure 3-19: Commuters First and Last Journey Start Time

Despite the mode they use and the weekday they travel, clusters 7 and 8 travel characteristics are very similar. Both clusters have few travel days (less than 2), performing on average 1 journey per day. Their journeys start during the afternoon and the difference between their minimum and maximum travel distance

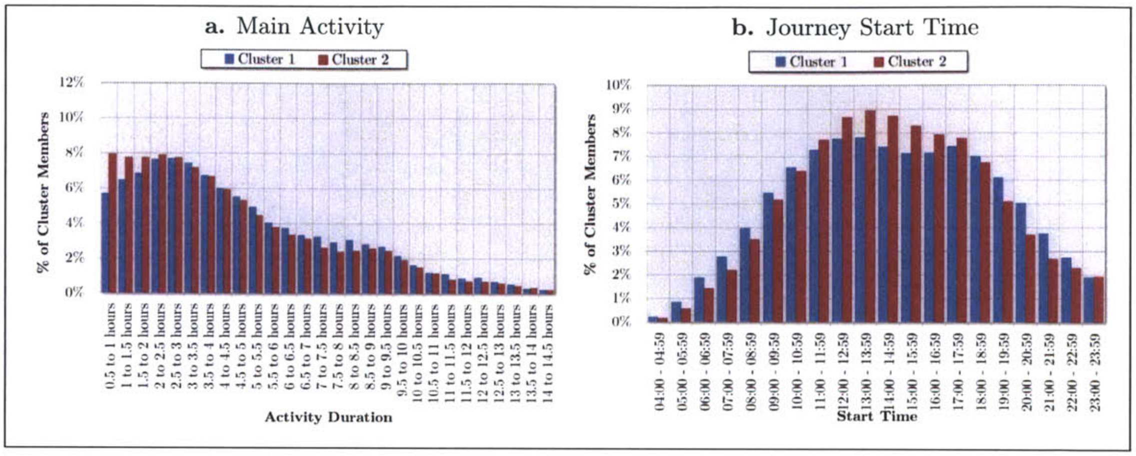
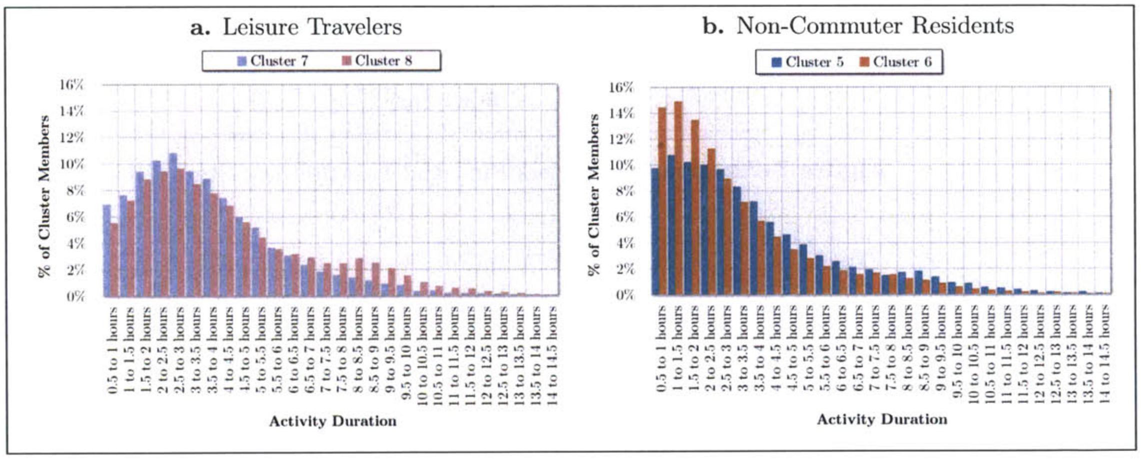


Figure 3-20: Non-Exclusive Commuters Weekend Activity Duration and Start Times

is 3 kilometers (a smaller difference than for commuters). Most of their members are Pay as You Go users (80% or more) and a small percent hold special discount cards (17% or less). There is no clear travel purpose that could be inferred using only these travel behavior characteristics. These clusters could be composed of leisure travelers, visitors, or sporadic public transport users.

Similarly, clusters 5 and 6 show common travel characteristics in spite of their mode choices and the days of the week they travel. They travel between 2 and 3 days per week performing no more than 2 journeys per day. They journeys are performed during the afternoon and between 64% and 74% of their members use Pay as You Go. They have a higher percentage of special discount card holders, especially cluster 6 (35%), who are mostly Freedom Pass holders (14% of cluster 5 and 25% of cluster 6). As for clusters 7 and 8, there is also no clear travel purpose that could be inferred using only these travel behavior characteristics. However, the higher percentage of special discount clusters indicates that there a significant percentage of London residents in this group, especially elderly (12% and 22% of cluster 5 and 6 respectively).

The analysis of other travel characteristics such as spatial travel patterns may improve the interpretation of these clusters, which is addressed in subsequent chapters. Figures 3-21 and 3-22 show the distributions of the main activity durations and the first and last journey start times respectively. The activity durations have peaks at less than 2.5 hours and the start time of the first and last journey of the day are normally distributed. Cluster 8 shows a more uniform distribution of journey start times starting earlier than the other clusters (8:00 am); cluster 5 is normally distributed around 3:00 pm, with clusters 6 and 7 around midday. The distributions in Figures 3-21 and 3-22 show great similarity with those for clusters 1 and 2 during weekends, which support the hypothesis that the main travel purpose of these clusters is leisure.


a. Leisure Travelers

b. Non-Commuter Residents

Figure 3-21: Non-Commuters Main Activity Duration

3.4.4 Cluster Validation: ANOVA Analysis

Analysis of variance (AVOVA) is a commonly used statistical procedure that compares the mean values of different variables between groups established in the data. The simplest form of ANOVA, called one-way ANOVA, uses only one variable or *factor* to form the groups to be compared. Since for this case there is only one classification partition, one-way ANOVA is the appropriate method to analyze the validity of the clusters obtained. The results of the one-way ANOVA applied to the 2011 clustering data are presented in Table 3-7.

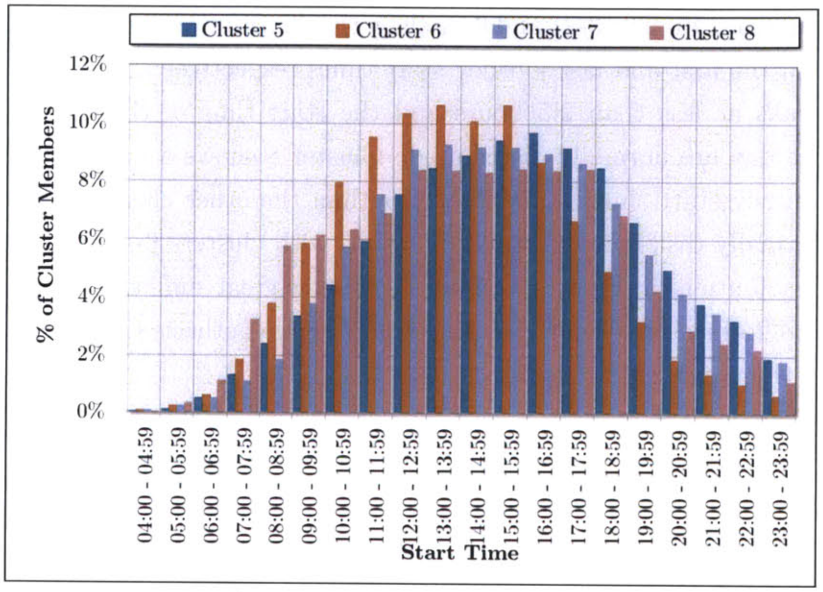


Figure 3-22: Non-Commuters First and Last Journey Start Time

Table 3-7: One-way ANOVA Results for the 2011 Cluster Data

	Sum of Squares	**Degrees of Freedom**	**Mean Square**	**F**	**Sig.**
Between Groups	$1.146 \cdot 10^{11}$	7	$1.6 \cdot 10^{10}$	2,933	$< 2 \cdot^{-16}$
Within Groups	$3.742 \cdot 10^{13}$	6,702,499	$5.58 \cdot 10^{6}$		
Total	$3.753 \cdot 10^{1}3$	6,702,506			

These results are aggregate for the complete vector of variables used for the clustering process. At the 95% confidence level, these are significant results ($\alpha < 2 \times 10^{-16}$); therefore, the null hypothesis that the means of the eight clusters are equal can be rejected. A more detailed validation analysis was done repeating the one-way ANOVA for each of the variables used. The results obtained also showed that the means of the clusters for each of the clustering variables are significantly different. Table 3-8 shows the one-way ANOVA results for the variable days of travel.

Table 3-8: One-way ANOVA Results for the 2011 Clustering Data - Variable: Days of Use

	Sum of Squares	Degrees of Freedom	Mean Square	F	Sig.
Between Groups	853,027	7	38,832	30,792	$< 2\times10^{-16}$
Within Groups	206,579	248,233	1		
Total	1,059,606	248,240			

Figures 3-23 and 3-24 show the variance of the days of travel and the maximum distance per cluster respectively. As can be seen from the box plots, there is substantial variation within clusters and the means are different between groups, results that are supported by the significant value of the associated F-statistics obtained with one-way ANOVA. The same significant results were found for the rest of the variables. Based on these one-way ANOVA results, the Oyster Card users' clusters obtained with the K-medoids algorithm can be validated.

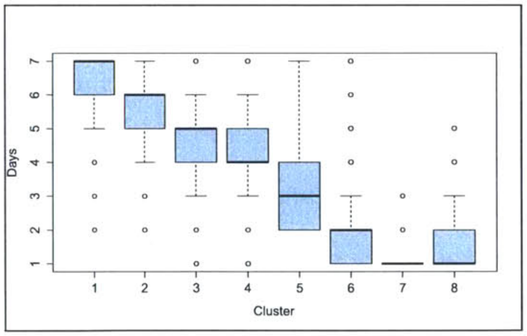


Figure 3-23: Days of Travel per Cluster

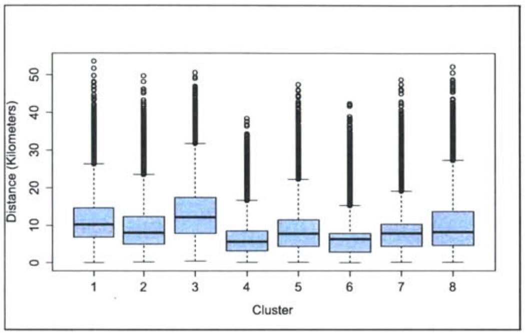


Figure 3-24: Maximum distance traveled per Cluster

3.5 Summary

The clustering methodology applied to London's Oyster Card users identifies eight passenger groups with similar travel characteristics. Four of these groups are regular users traveling four days or more per week, and four are occasional users. Three of the clusters have members traveling any day of the week, four of them have members traveling only during weekdays and only one travels during weekends exclusively. There are two groups that prefer bus over rail, and only one that prefers rail. The 5 remaining clusters use bus and rail with no particular preference.

Four major groups were identified matching pair of clusters with similar characteristics: exclusive commuters, non-exclusive commuters, leisure travelers, and non-commuter residents. For the first two groups, it was possible to identify student and work commuters based on their activity duration, journey start times, and mode preferences. The identification of travel purpose was less clear for the last two groups. The characteristics were very similar among these two groups with the difference that the non-commuter resident group has a higher percentage of special discount cards, especially elderly freedom passes. Subsequent analyses of other travel characteristics help to improve this initial interpretation of the clusters.

The within-cluster variation, the Davies-Bouldin index and the ANOVA analysis validated the clustering results significance. Several other characteristics of the clusters can be analyzed to provide a better understanding of the behavior of each group. The analysis of non-clustering characteristics are presented in detail in subsequent chapters.

Chapter 4

Cluster Spatial Distribution and Temporal Stability

This chapter analyzes the spatial travel patterns of cluster members and explores the variability of cluster characteristics over time and passengers membership stability. The first goal of this chapter is to determine whether spatial travel patterns are related to other travel behavior, activity patterns or sociodemographic variables. The development of a simple home location methodology aims at finding a relationship between cluster members home location and their travel characteristics. This chapter's second goal is to determine the consistency of cluster travel behavior characteristics over time and cluster membership stability. Distinguishing the less variable travel characteristics may allow the identification of group of users with a higher level of travel behavior predictability, which could support the assessment of potential transport planning improvements.

The chapter is organized in two sections. Section 4.1 describes and analyzes the spatial distribution of users with different travel profiles. The section first identifies the stations most commonly used by each group and ends by analyzing the home location estimated for the members of each cluster. Section 4.2 repeats the clustering process performed in 3.4 using the 2012 Oyster Card data sample described in 3.2.4. In addition, a comparative analysis between 2011 and 2012 cluster results is performed. The stability of the characteristics of similar clusters is analyzed and passengers' membership is tested using those Oyster Cards observed in both periods.

4.1 Passenger Groups Spatial Distribution

It is important to have a deeper understanding of where different groups of passengers move around London. This understanding can help not only to analyze location variation and geographic trends for the different type of users but also to understand how geographic constraints shape the travel characteristics of different clusters. The following paragraphs analyze the rail stations most commonly used by different passenger groups and present a methodology to estimate the home location of most cluster members.

4.1.1 Most Frequent Stations

The thirty-five most frequently used rail stations in London during weekdays and weekends are shown in Figures 4-1 and 4-2. The black line in both graphs shows the total number of entries at the corresponding station as a percentage of the number of rail users that travel during weekdays and weekends respectively. The color squares show the number of entries as a percentage for each cluster membership. These graphs help identify geographic differences between clusters and compare them against the behavior of the whole population.

As can be seen in Figures 4-1 and 4-2, the most frequently visited stations are Waterloo, London Bridge and Victoria. During weekdays, occasional clusters 5 and 8 exhibit high concentration at stations such as Victoria, King Cross, Paddington, and Euston, all National Rail terminal stations. This may be explained because these are non-commuter clusters making leisure journeys, and there is a high probability that a significant proportion of them are visitors. Cluster 3 shows a higher concentration than the whole population mainly in Waterloo and Canary Wharf. For the remaining stations for this cluster seems to behave similarly to the entire population. Clusters 1 and 2 also show behavior consistent with the behavior of the whole population. These are expected results given that most of clusters 1, 2, and 3 members are commuters. Clusters 4 and 6 (weekday bus regular and occasional respectively) are mainly composed by bus users and do not show any significant trends compared to other clusters.

Figure 4-2 shows that during weekends cluster 7 not only has a large percentage of members using National Rail terminals (Waterloo, Victoria, King Cross, Euston, and

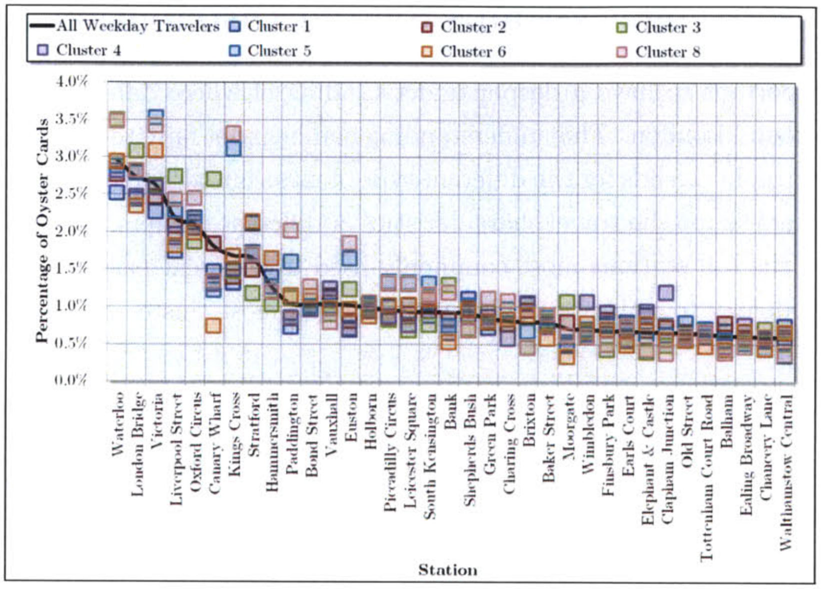


Figure 4-1: Most Frequent Weekday Stations by Cluster

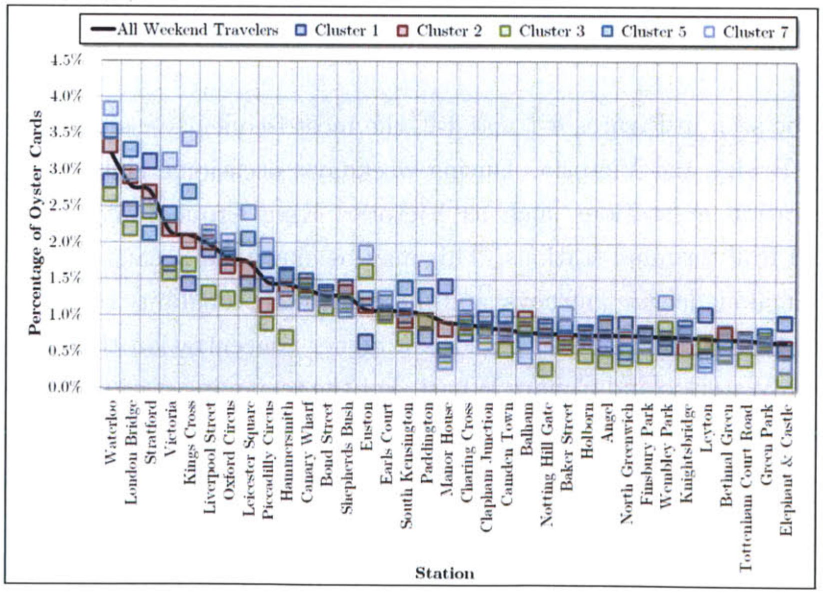


Figure 4-2: Most Frequent Weekend Stations by Cluster

Paddington) but also shows a high concentration around tourist locations such as Leicester Square, Piccadilly Circus, Baker Street, and Wembley Park. Cluster 5 has similar behavior to cluster 7 but with smaller percentages specially for the tourist locations. Cluster 3 has lower percentages compared to the total population. As for weekdays, cluster 2 does not show significant differences from the total population during weekends. Cluster 1 also has similar behavior to the total population, with the exception of a few stations such as Kings Cross and Euston, where the cluster distribution is lower than the total population.

Figure 4-3 shows the forty next most frequently used stations (following the 35 top stations shown in Figure 4-2) during weekends. This figure shows that cluster 7 members clearly use the Heathrow Terminals 1,2, 3 station more than the general population, which might indicate that a significant part of this cluster members are overseas visitors. This is consistent with the travel characteristics of cluster 7, that as analyzed before, perform short duration activities and travel during off-peak hours.

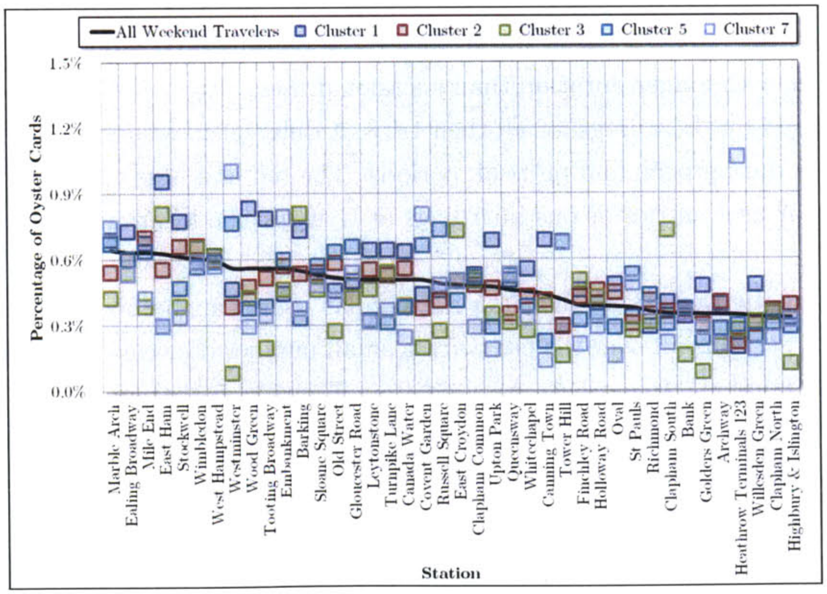


Figure 4-3: 40 Next Most Frequent Weekend Stations by Cluster

Most station entries correspond to members of clusters 1 to 4, which represent regular users, specially at high ridership stations. On the other hand, occasional users are a major part of the ridership at international terminals. Figure 4-4 shows the cluster distribution for the four airport stations where more than 40% of the entries are occasional users. Cluster 8 uses London City Airport more than Heathrow Airport. London City Airport is a small international and domestic airport having high demand from business travelers due to its proximity to London's financial industry centers (City of London and Canary Wharf). This could indicate that a significant proportion of cluster 8 members are business visitors, which is consistent with their travel behavior; despite their travel characteristics which suggest that they are leisure travelers, they show earlier and longer activities than other occasional clusters and some similarities with commuter clusters. Cluster 7 members, on the other hand, show a higher percentage of entries in Heathrow terminals, which is London's major international airport and the world's busiest airport in terms of international passenger demand (Airports Council International, 2013). This suggest that cluster 7 is mainly composed of overseas visitors.

The location of station entries and bus boardings of occasional and regular users varies over the day. Regular users travel during the morning and afternoon peak and perform activities with longer duration than occasional users. Figures 4-5 and 4-6 show passenger entries and bus boardings for all their trips at different locations in London for 15-minute intervals during peak and off peak periods. The red and blue circles represent occasional or regular user entries or bus boardings at that specific location.

Note that between 8:00 and 8:15 am regular users can be observed boarding buses or entering stations all over the city with a small number of occasional users moving around Central London. It can also be seen from Figure 4-5 that in the 5:30 pm to 5:45 pm period there are a large number of regular users making entries or boardings in Central London and occasional users are again focused in Central London in higher numbers than during the morning peak. This may be explained because regular users make commuting journeys that start during the morning peak, stay in the same location working or studying (concentrated in Central London), and make journeys home during the afternoon peak. Occasional users on the other hand perform mostly leisure activities and a significant proportion of them are visitors making trips near tourist locations (Central London) during off-peak hours.

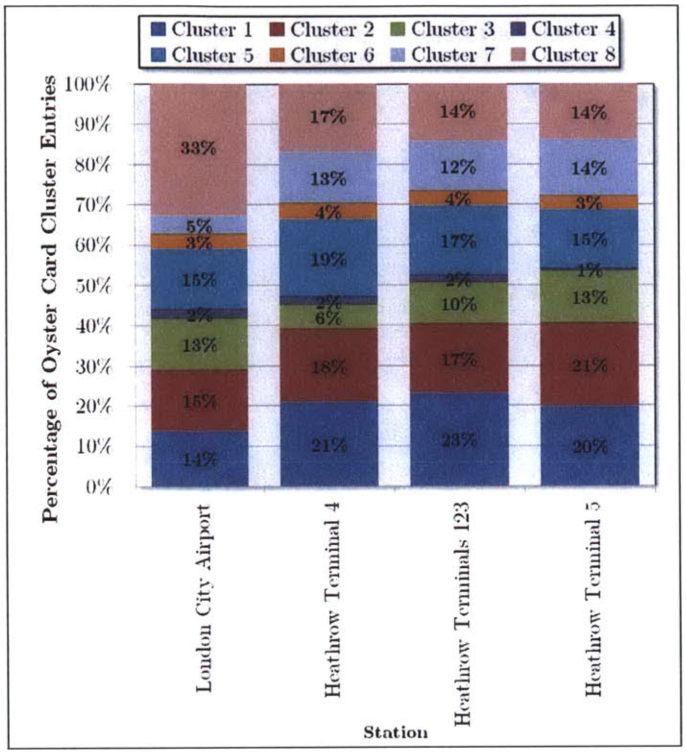


Figure 4-4: International Terminal Entries Cluster Distribution

Figure 4-6 shows the difference of the location of station entries and bus boardings between regular users and occasional users during off-peak periods. There are a large number of occasional users distributed through the city, mostly in Central London, between 12:00 and 12:15 pm. Between 9:45 and 10:00 pm, there are some areas of the city near Central London where the number of occasional user entries and boardings is higher than the number of regular users. This is again explained by the fact that occasional users perform leisure and recreational activities especially during off-peak hours.

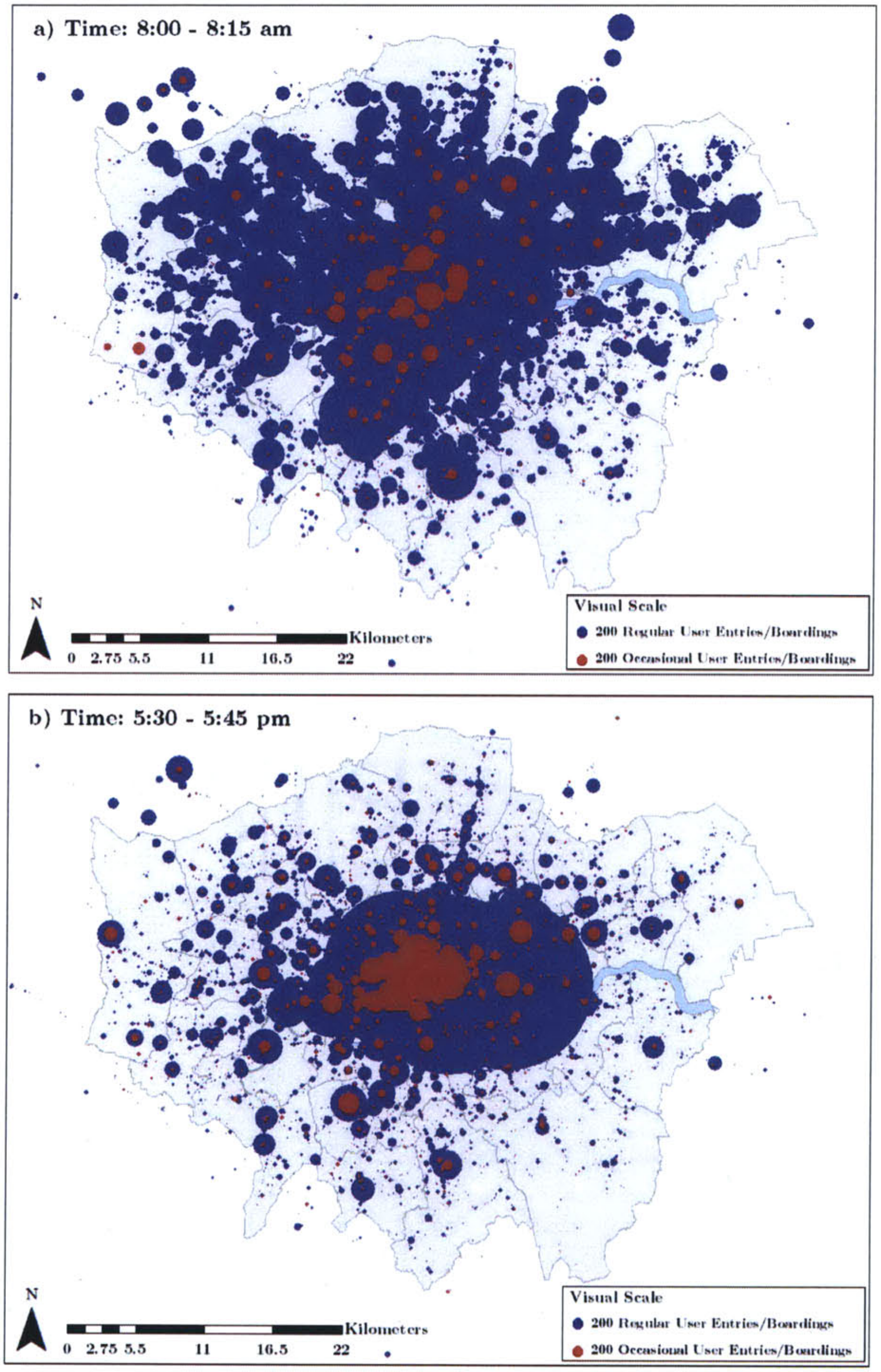

a) Time: 8:00 - 8:15 am
N
Kilometers
0 2.75 5.5 11 16.5 22
Visual Scale
● 200 Regular User Entries/Boardings
● 200 Occasional User Entries/Boardings
b) Time: 5:30 - 5:45 pm
N
Kilometers
0 2.75 5.5 11 16.5 22
Visual Scale
● 200 Regular User Entries/Boardings
● 200 Occasional User Entries/Boardings

Figure 4-5: London Regular and Occasional User Entry Locations

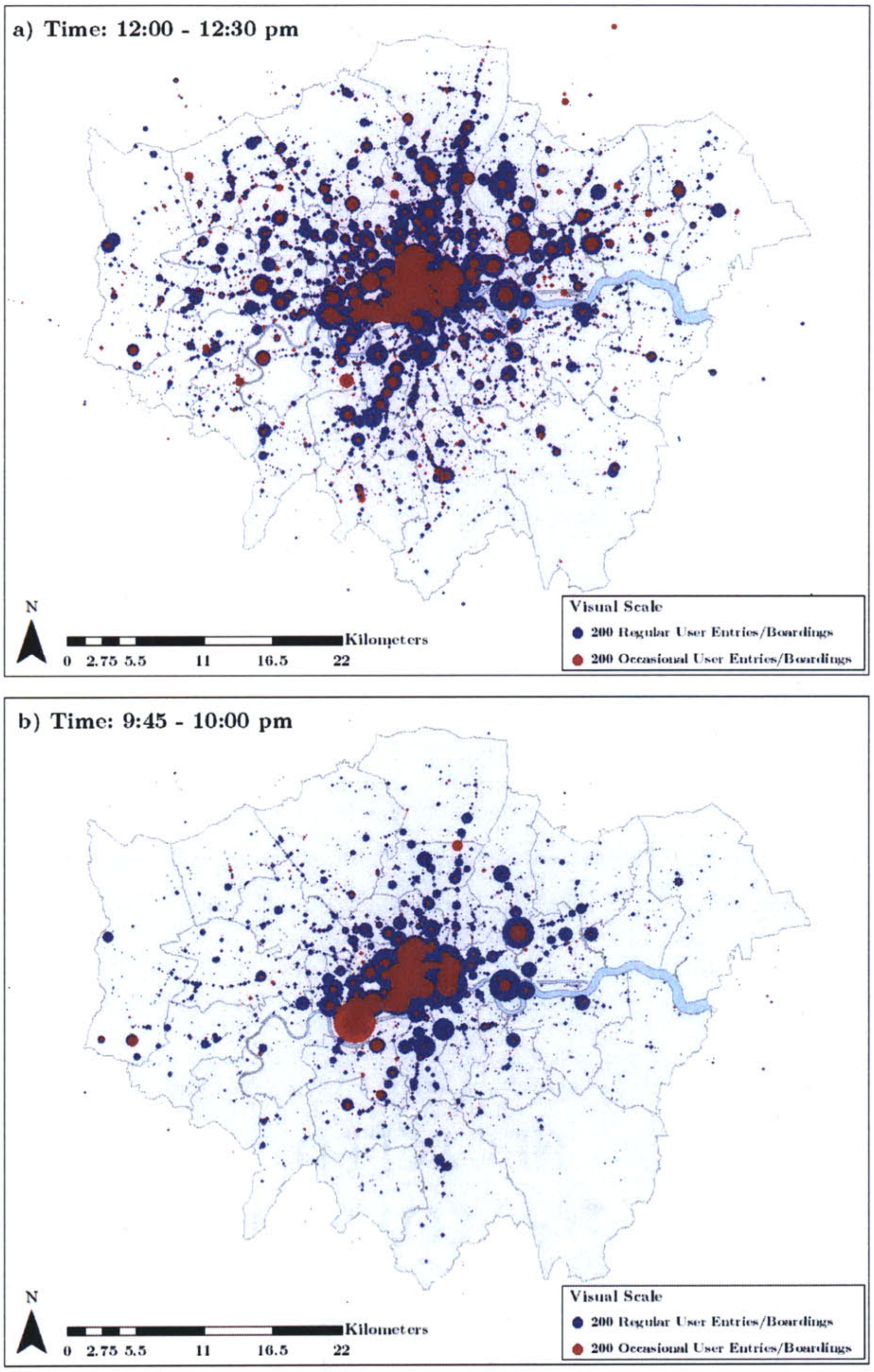

a) Time: 12:00 - 12:30 pm
N
0 2.75 5.5 11 16.5 22
Kilometers
Visual Scale
● 200 Regular User Entries/Boardings
● 200 Occasional User Entries/Boardings
b) Time: 9:45 - 10:00 pm
N
0 2.75 5.5 11 16.5 22
Kilometers
Visual Scale
● 200 Regular User Entries/Boardings
● 200 Occasional User Entries/Boardings

Figure 4-6: London Regular and Occasional User Entry Locations

4.1.2 Home Location Estimation

A general picture of where the London population lives is provided in Figure 4-7. The first map shows the population density of each borough in persons per hectare. The second map shows the number of persons living in each borough as a percentage of London's total population. The first map shows that the highest densities are around Central London, but the second map shows that a large percentage of people live far from the center in areas such as Croydon and Barnet.

Based on the main ODX inference assumption that passengers both start and end their daily public transport journeys near home, a methodology to estimate approximate locations for Oyster Card users' homes was developed. For each Oyster Card ID (*id*) the steps detailed below were followed to determine the home location of the card user:

1. Determine the coordinates (X^d_A, Y^d_A) for the origin station or stop location of the first stage of the first journey of each day (d).

2. Determine the coordinates (X^d_B, Y^d_B) for the destination station or stop location of the last stage of the last journey of each day (d).

3. If the geometric distance between the two coordinates (A and B) is shorter than 1 kilometer, estimate the coordinates for the mid-point (MP) between those two coordinates. If the distance is longer than 1 kilometer, discard that day since no consistent home location can be identified.

$$MP^d_{id} = \begin{cases} \left(\frac{X^d_A + X^d_B}{2}, \frac{Y^d_A + Y^d_B}{2}\right) & \text{if } \sqrt{\left(X^d_A - X^d_B\right)^2 + \left(Y^d_A - Y^d_B\right)^2} < 1 \text{ km} \\ \nexists & \text{otherwise} \end{cases} \tag{4.1}$$

4. Compare the mid-points estimated for all the days that Oyster Card user traveled. Establish as home location the geographic point that is most common during the week.

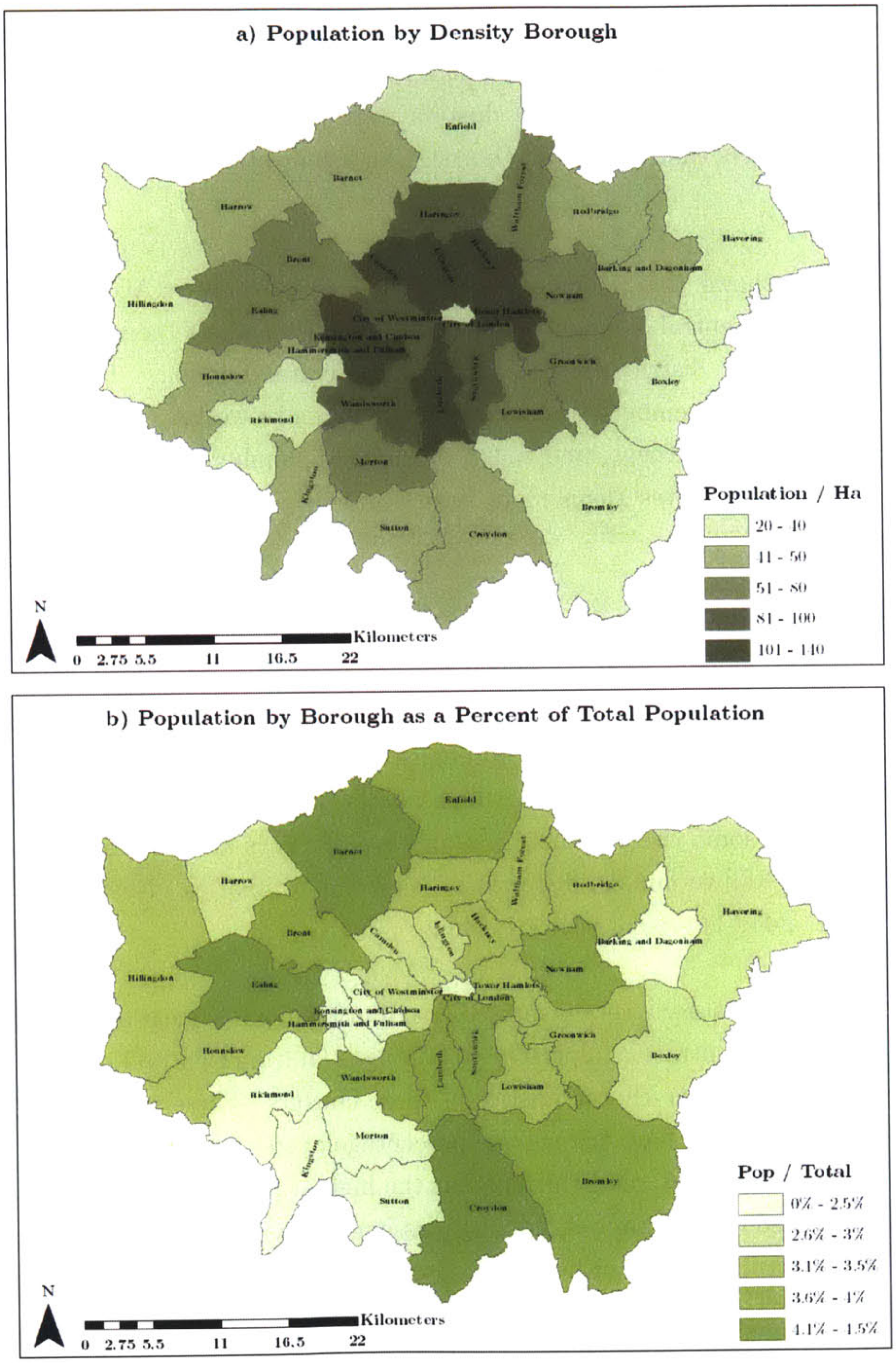


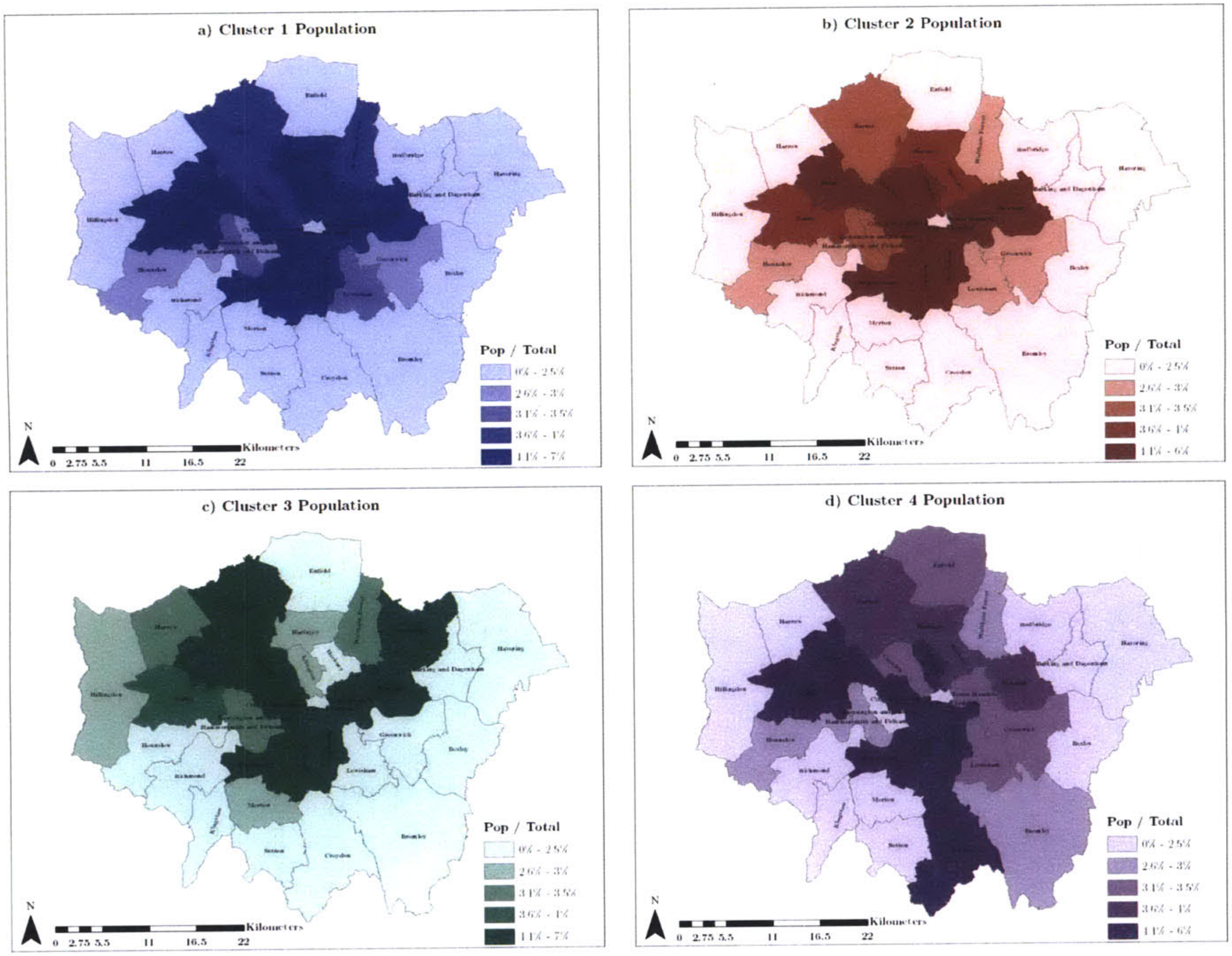
Figure 4-7: Greater London Population

Applying this method to the October 2011 sample of Oyster Card data, 91% of home locations were estimated. Figures 4-8 and 4-9 show the home location of cluster members, for regular and occasional user clusters respectively. The maps show the number of cluster members that live in each borough as a percentage of that total number of cluster members.

Figure 4-8 shows that members of cluster 3 live mostly in West London and in the periphery of Central London. Clusters 1 and 2 members' homes locations show very similar pattern, distributed mostly around Central London. This is consistent with clusters 1 and 2 members travel behavior, who make weekday commuting journeys and weekend leisure journeys, and make the highest number of trips per day. Living near Central London, gives them more travel options which may explain their high travel frequency.

Figure 4-9 on the other hand shows that the homes of occasional users clusters 5, 7, and 8 are located mostly around Central London, especially cluster 8. Additionally, leisure travelers (clusters 7 and 8) show similar home location patterns for those members whose homes are in the West London boroughs. It is important to note that these clusters travel one to two days per week, which makes their home estimation less accurate and, in the case of some visitors, their first and last journeys are performed in airport stations or National Rail terminals, distorting their home estimation. Nevertheless, staying near Central London is typical visitor behavior.

Clusters 4 and 6 show a very dispersed distribution of home locations across the boroughs, with higher concentration in peripheral boroughs far from Central London. This dispersion is consistent with the characteristics of these clusters. These two clusters are composed mainly of bus users, have the highest percentage of special discount cards (40% of cluster 4 and 35% of cluster 6), and cluster 6 has the highest percentage of Freedom passes (25%). Therefore, reduced or limited mobility passengers trying to reduce interchanges or avoid stairs may explain the high bus usage.



a) Cluster 1 Population

b) Cluster 2 Population

c) Cluster 3 Population

d) Cluster 4 Population

Figure 4-8: Home Location for Regular User Clusters

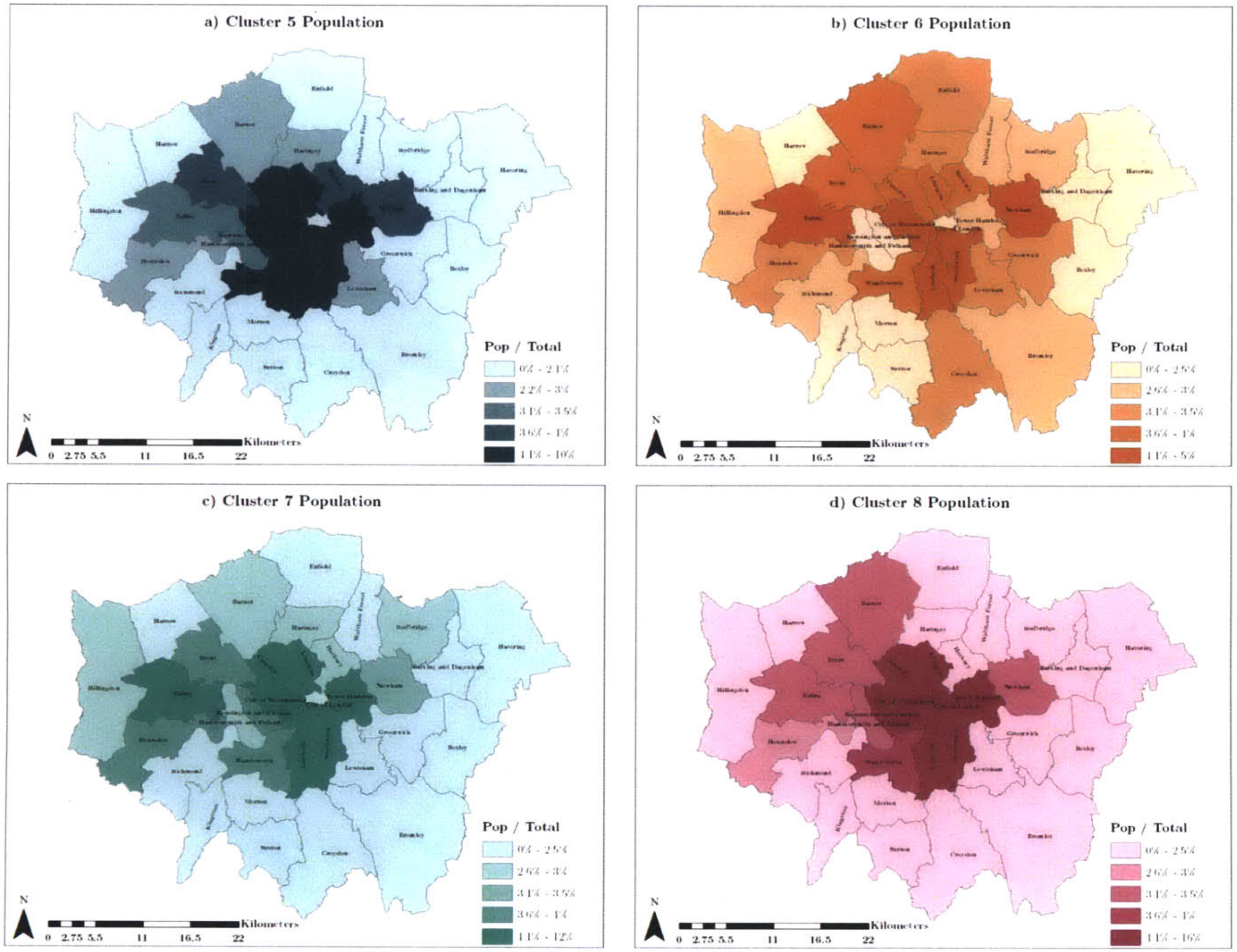


Figure 4-9: Home Location for Occasional User Clusters

4.2 Cluster Temporal Stability

The main goal in this section is to examine the level of consistency of travel behavior over time, and if there is any relationship between travel behavior temporal stability and the travel behavior itself. Having deeper knowledge of the temporal consistency of the travel behavior of different groups can help assess the predictability of each group's behavior, which can help in the assessment of strategic or operational planning changes.

The classification process described in 3.4 was applied using one week of Oyster Card data from 2012. The data is based on the same 250,000 random sample of Oyster Cards from October 1-7th, 2012 described in 3.2.4. The K-medoids clustering algorithm with K=8 was used. This was the best cluster configuration according to the within-cluster variation values and the Davies-Bouldin index (see Appendix A).

As described in Section 3.3, both 2011 and 2012 samples have very similar travel and demographic characteristics overall hence the classification results obtained using each sample should be comparable. However, it is important to know that 98% of the 250,000 Oyster Cards observed in the 2012 sample were not included in the 2011 sample, which is caused because each sample was drawn independently, but it also reflects the dynamics of the oyster Card system. Nevertheless, some of the 2012 clusters have very similar characteristics to those obtained for 2011 data while others showed significant differences in some of the characteristics. For those 2012 clusters that maintained most of their travel characteristics, matching them with the corresponding 2011 cluster was straightforward. On the other hand, the identification of the corresponding 2011 cluster for the 2012 clusters showing major changes was not clear, so they were matched using judgment. A brief description of each 2012 cluster is presented below, highlighting the major differences and similarities with 2011. The names of 2011 clusters were retained to facilitate their identification, despite the fact that some characteristics had changed.

- **Cluster 1: Everyday regular users**

  As in 2011, this cluster's members travel on average 6 days a week, making either 2 or 3 journeys per day. The difference with 2011 for the first journey of the day is only 30 minutes for weekdays (9:30 am) and weekends (12:30 pm), which is not significant given the sample size. The last journey of the day starts at the same time as in 2011 (6:30 pm during weekdays and 6:00 pm during weekends). The travel distance varies between 2 and 12 kilometers (1 kilometer more than in 2011). As in 2011 the members use a mix of bus and rail for their journeys. 77.4% of the members are Travelcard holders (4% less than in 2011) and 98.2% do not hold special discount cards (17% more than in 2011).

- **Cluster 2: All week regular users**

  As in 2011, members of cluster 2 travel on average 5 days a week (any day of the week), making 1 or 2 journeys per day. During weekdays, the difference with 2011 for the first and last journey of the day is only 30 minutes (11:00 am and 4:00 pm respectively), which is not significant given the sample size. During weekends, the first journey of the day starts at 1:00 pm as in 2011 but the last journey of the day starts one hour earlier (4:00 pm). The members of this cluster have activities that last between 3 and 4 hours during weekdays (1 to 2 hours less than in 2011) and between 1 and 2 hours during weekends (1 hour less than in 2011). The travel distance varies between 1 and 7 kilometers, which is less than the distance observed during 2011 by 2 kilometers. Unlike 2011, the members of this cluster have the same origin for the first journey of the day during three weekdays, and on two weekdays have the same origin for the last journey of the day. The last journey starts at the same time as in 2011 (6:30 pm during weekdays and 6:00 pm during weekends). The travel distance varies between 2 and 12 kilometers (1 kilometer more than in 2011). As in 2011 the members use a mix of bus and rail for their journeys. The highest difference is the percentage of Travelcard holders (88.9%), which is 34.5% more than in 2011. This difference is due to this cluster's higher percentage of special discount holders (73.9%, 47.7% higher than in 2011).

- **Cluster 3: Weekday rail regular users**

  This cluster characteristics are very similar to those observed in 2011. As in 2011, members of cluster 3 travel on average 4 weekdays, making 2 journeys per day. Their first journey starts at 9:00 am and their last journey at 5:30 pm (same as in 2011). The activities of cluster 3 members last between 6 and 8 hours and the distance they travel is between 5 and 12 kilometers (± 2 kilometers compared to 2011). As in 2011, most of this cluster members prefer rail (92% use rail at least once every day). The number of travel card holders is very similar to 2011 (39.8% are Travelcard holders and 95.5% do not hold special discount cards).

- **Cluster 4: Weekday bus regular users**

  The members of cluster 4 travel on average 3 weekdays (one day less than in 2011) making 2 journeys per day. The first trip of the day starts at noon (2 hours later than in 2011) and the last journey starts at 2:30 pm (1.5 hours earlier than in 2011). Their activities last approximately 4 hours (2 hours less than in 2011) and the journey length varies between 2 and 5 kilometers (2 kilometers less than in 2011). As in 2011, most of this cluster's members prefer bus (89% uses bus every day).

  The greatest difference in this cluster compared with 2011 is that almost all the members (99.2%) hold a especial discount card with a free pass or special discount Travelcard (59% more than in 2011). 54.5% are elderly, 32.3% students or children, 7.5% disabled, and 4.8% staff. These characteristics could indicate that the travel patterns of this cluster are typical of users benefiting from free travel or significantly lower fares.

- **Cluster 5: All week occasional users**

  Cluster 5 members travel on average 4 days per week (one day more than in 2011), making 1 or 2 journeys per day, during weekdays and weekends. The first journey of the day starts one hour earlier than in 2011 for weekdays and weekends (1:30 pm on weekdays and 1:00 pm on weekends). The last journey of the day starts at the same time as in 2011 for weekdays (5:00 pm) and approximately one hour later for weekends (5:30 pm). This cluster's members travel further than in 2011: between 3 and 10 kilometers (3 kilometers more than in 2011). As in 2011, this cluster's members prefer a mix of rail and bus for their journeys. Only 12.6% of the members hold a Travelcard (13.8% less than in 2011) and 98.5% do not hold special discount cards (20.5% more than in 2011).

- **Cluster 6: Weekday bus occasional users**

  Cluster 6 members travel on average 2 weekdays, making 1 journey per day. Their first and last journeys of the day starts at approximately the same time as in 2011 (12:30 pm and 3:30 pm respectively), and their journey length is only 1 kilometer longer than in 2011 (between 3 and 7 kilometers). As in 2011, almost all their journeys have different origins and most of the members prefer bus (89% use bus every day). Unlike 2011, 99.8% of the members do not hold special discount cards (35.3% more than in 2011) and only 9.3% are Travelcard holders (27.1% less than in 2011).

- **Cluster 7: Weekend occasional users**

  As in 2011 this cluster's members travel on average 1 or 2 days a week during weekends, making 1 or 2 journeys per day. The first and last journey's start times do not change compared to 2011 (2:00 pm and 5:00 pm respectively). They have almost the same travel distance as in 2011: between 5 and 9 kilometers (1 km more than in 2011). As in 2011, 73% of this cluster's members use rail and bus at least once for at least half of the days they travel (same as in 2011). Only 19.8% of this cluster's members are Travelcard users (approximately the same as 2011) and 86.2% do not hold special discount cards (3% more than in 2011).

- **Cluster 8: Weekday rail occasional users**

  As in 2011, cluster 8 members travel on average 2 days a week, making one journey per day. As in 2011, the journeys start between approximately 1:00 pm and 4:00 pm and their journey length varies between 7 and 10 kilometers. The members of this cluster have different origins for all their journeys and 94% of them use rail every day (1% more than in 2011). Only 10.2% of this cluster's members are Travelcard holders (2.9% less than in 2011) and 95.5% do not hold special discount cards (5.8% more than in 2011).

As can be seen, most differences between 2011 and 2012 are not significant given the sample size and might be explained by the sampling error. Nevertheless, some differences are significant specially those related to the percentage of Travelcards and special discount cards. Clusters 4 and 5 showed the greatest differences with respect to 2011. They present small differences in almost all their characteristics, but the most significant differences are the number of days of travel and the type of Oyster Card. Cluster 5 increased the number of days of travel from 3 to 4, and cluster 4 decreased the number of days from 4 to 3. In addition, almost all cluster 4 members hold special discount cards, which was not observed for any cluster in 2011. Clusters 2 and 6 on the other hand showed similar characteristics to their analogous 2011 clusters, but their main difference was the percentage of Travelcard holders, which increased for cluster 2 and decreased for cluster 6. Clusters 1, 3, 7 and 8 kept most of their 2011 characteristics.

This results might indicate that even though the variables Travelcard and special discount card are related to travel behavior, they are also related to other external factors (such as monthly budget or school registration), that can change over time changing the individual Travelcard or special discount card status without modifying their travel behavior. Therefore, it is probable that some Pay as You Go users acquire Travelcards or obtained access to special discount cards, specially for those clusters with high travel frequency. However, it is important to note that this could also be affected by the fact that the 2011 and 2012 samples were drawn independently and only 2% of the cards are common to both samples.

It was also observed that groups with the highest and the lowest frequency of travel presented the most stable travel behavior. This could be because passengers traveling every day have no other option than to use public transport, either for home location or income reasons, or accessibility to other modes. This could also be the case for passengers traveling one or two days, who are more likely to be visitors.

Figure 4-10 shows the size of the clusters in each period. The sizes of the clusters are similar to those observed in 2011 (3% average absolute difference), showing absolute differences between 1% and 4%, with the exception of cluster 2 that is 8% smaller than in 2011.

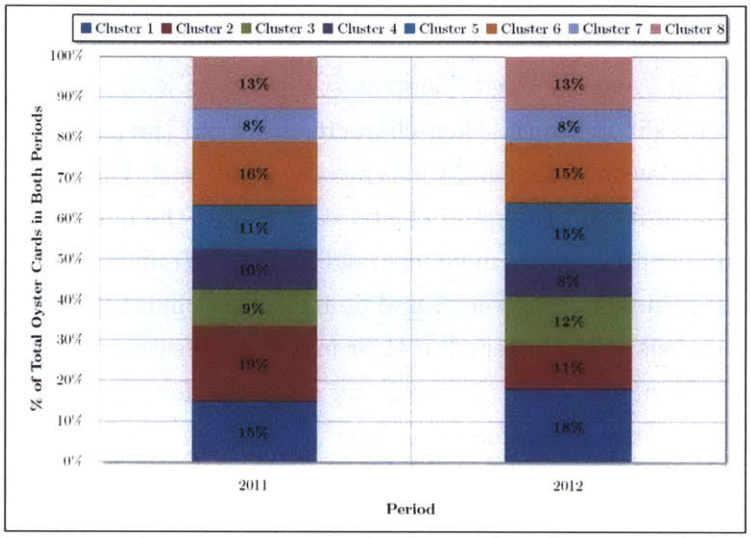


Figure 4-10: Cluster Size Comparison 2011-2012

Grouping the clusters based on the groups identified in Section 3.4.3 (exclusive commuters, non-exclusive commuters, leisure travelers, and non-commuter residents), the observed travel characteristics are more similar when comparing 2011 and 2012. Table 4-1 shows the values for the average variables of each group in 2011 and 2012. As can be seen, most groups maintain their characteristics from one year to the next, specially exclusive commuters and leisure travelers. The major differences are observed in the percentage of Travelcards and special discount cards. The group of non-Exclusive commuters increased their percentage of members with Travelcards and special discount cards by 17% and 19% respectively. This could indicate that some Pay as You Go users acquired a Travelcard or had access to a Freedom pass (17% increase) in 2012, which would be consistent with this group high frequency of travel.

Table 4-1: Group Average Characteristics Comparison

Group	**Exclusive Commuters**		**Non-Exclusive**		**Non-Commuter**		**Leisure Travelers**	
(Clusters)	**(1 &2)**		**Commuters (3 & 4)**		**Residents (5 & 6)**		**Commuters (7 & 8)**	
Variable / Year	2011	2012	2011	2012	2011	2012	2011	2012
Days of the Week	Weekdays	Weekdays	All Week	All Week	All Week	All Week	All Week	All Week
Days of Travel	6	6	4	4 2	3	1	1	
Journeys per day	2 to 3	2 to 3	2	2	1	1	1	1
Exclusive Bus Days (%)	49%	50%	38%	40%	69%	63%	17%	17%
% Exclusive Rail Days (%)	22%	19%	40%	39%	15%	19%	68%	69%
First Journey Start Time (weekday/weekend)	10:15 am / 12:30 pm	10:00 am / 12:40 pm	9:30 am / -	10:00 am / -	1:00 pm / 1:50 pm	1:20 pm / 1:15 pm	12:40 pm / 2:00 pm	12:50 am / 2:00 pm
Last Journey Start Time (weekday/weekend)	5:30 pm / 5:20 pm	5:30 pm / 5:10 pm	4:40 pm / -	4:30 pm / -	4:00 pm / 5:00 pm	4:20 pm / 5:00 pm	4:00 pm / 4:30 pm	4:00 pm / 5:30 pm
Main Activity Duration Time (weekday/weekend)	6 hrs / 3.5 hrs	6.1 hrs / 3.1 hrs	6.9 hrs / -	6.2 hrs / -	3.5 hrs / 1.7 hrs	4.2 hrs / 3.4 hrs	4.2 hrs / 2.2 hrs	4.4 hrs / 2.3 hrs
Journey Distance	2 to 10 km	2 to 10 km	4 to 10 km	4 to 9 km	3 to 7 km	3 to 8 km	6 to 9 km	7 to 10 km
Journey Origins	Mostly one	Mostly one	Mostly one	Mostly one	All different	All different	All different	All different
Travelcard Holders	66%	82%	46%	64%	32%	11%	15%	14%
Special Discount Card Holders	23%	29%	24%	43%	30%	1%	13%	8%
Freedom Pass Holders	12%	19%	7%	26%	8%	6%	21%	1%
Student or Child Pass Holders	9%	8%	15%	14%	8%	0%	4%	2%
Staff Pass Holders	1%	2%	2%	3%	1%	0%	1%	1%

Approximately 2% of the Oyster Cards in the random sample in 2011 were also included in the random sample in 2012. Of these users, only 28.3% belong to the same cluster in both years; however, 66.4% belong to a cluster of equal frequency of travel, i.e. 66.4% of the sample belonged to a regular (1 through 4) or occasional (5 through 8) cluster in 2011 and was in the same type of cluster in 2012 (59.5% occasional and 70.7% of regular users). Figure 4-11 shows each clusters' members temporal stability. The blue bars illustrate the percentage of each cluster that remained in the same cluster, the red bars show the percentage that remained in the same frequency category (occasional for clusters 5 through 8, or regular for clusters 1 through 4), and the green bars show the percentage of users that moved to another cluster with a different frequency category. Clusters 1, 3, and 8 show the highest temporal stability (45%, 48%, and 38% remained in the same cluster respectively). These clusters travel characteristics did not changed significantly from 2011 to 2012, which shows consistent behavior of the members of these clusters. Clusters 1, 2, and 3 (commuter clusters) show the highest percentage of members that remained in the same frequency category from 2011 to 2012 (67%, 65%, and 62% respectively).

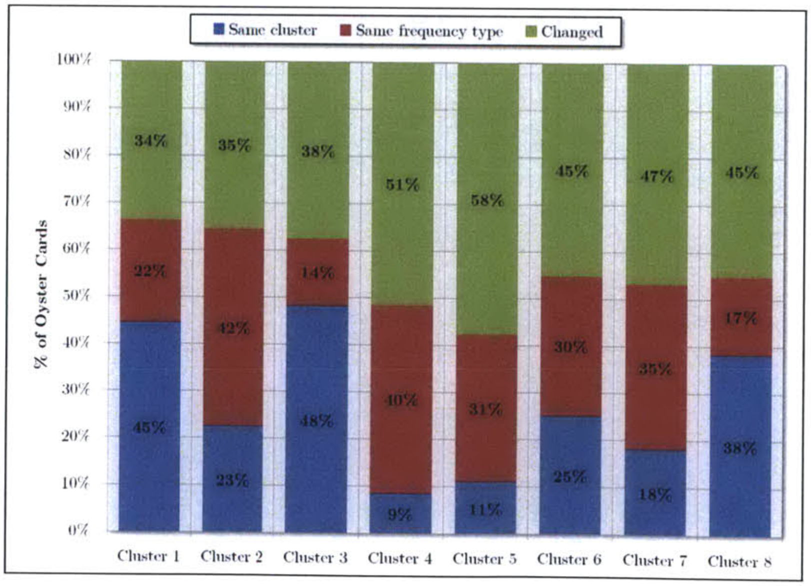


Figure 4-11: Temporal Stability per Cluster

Figure 4-12 shows the changes in cluster membership considering the cluster categories discussed in Section 3.4.3. Considering these groups, 45.3% of the total population remain in the same group from 2011 to 2012. The graph shows that at least 37% of the members had temporal stability at the group level, which is higher than the percentages observed considering the eight clusters. Non-exclusive commuters showed the highest temporal stability (55%), which indicates that high frequency travelers show more consistent behavior over time. Non-commuter residents show the lowest temporal stability (37%) and 42% of them exhibit commuter behavior in 2012, which could indicate that the leisure behavior shown by some of this group members during 2011 was particular to the analysis week.

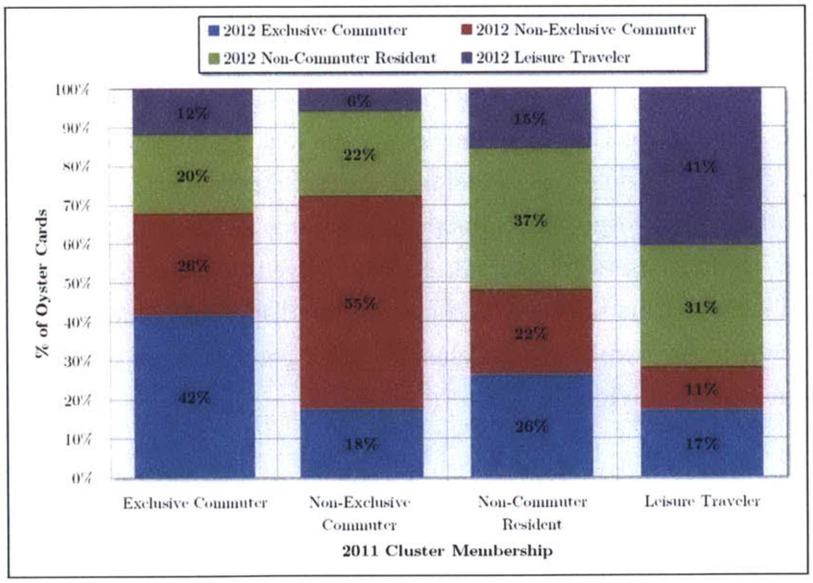


Figure 4-12: Temporal Stability per Cluster Category

4.3 Summary

The analysis of the most frequently used stations showed that occasional user clusters 4, 7, and 8, show high percentage of members using National Rail terminal stations such as

Victoria, Kings Cross, Paddington, and Euston, which suggest that these clusters have significant percentages of visitors or non-residents. Leisure travelers (particularly clusters 7 and 8) show high percentage of members using airport stations, which could indicate a high percentage of international visitors. Additionally, the most frequent stations for regular user clusters 1, 2, and 3 are very similar to the full population, which indicates that the behavior of the total population is highly influenced by regular users travel behavior (regular clusters are 53% of the total population).

Regular users' morning journeys start all over Greater London, during the peak period (6:30 - 9:30 am). Regular users do not show more movement during off-peak hours, when occasional users' journeys start, mainly focused in Central London. During the afternoon, regular users' journeys start in Central London during the peak hours (16:00 - 19:00 pm), and occasional users travel into late night hours (past 10:00 pm).

A methodology to estimate home locations was developed to analyze the differences between cluster. Homes of occasional users (clusters 1, 2 and 5) are located mostly in Central London, especially cluster 5. Staying in Central London is a typical visitor behavior. Regular users (clusters 3, 4, and 6) live mostly outside and on the periphery of Central London. Clusters 7 and 8 showed more dispersed home locations far from Central London which may explain these clusters' high percentage of bus users and of reduced or limited mobility individuals.

In order to determine cluster temporal stability, the clustering process was repeated using an independent 250,000 Oyster Card random sample from 2012. The results showed that there is a lack of temporal stability from 2011 to 2012, specially at the cluster level. The comparison of clusters characteristics showed that only clusters with the highest and lowest frequency of travel remain most of their characteristics from 2011 to 2012. The characteristics of clusters 1, 3, 7 and 8 showed the greatest similarities with the clusters obtained in 2011. These clusters have high (clusters 1 and 3) and low frequency of travel (clusters 7 and 8), and they favor rail users (clusters 3 and 8). This could indicate that passengers traveling every day have no other option than to use public transport for home location, income, or accessibility reasons, which could also be the case for passengers traveling only one or two days (probably visitors).

Various differences observed from 2011 to 2012 were not significant for the sample size analyzed and might be explained by the sampling error. However, for most clusters, the variables that showed the greatest temporal differences were those related with the type of Oyster Card (Travelcard or special discount card). This implicates that a further analysis of these variables is required, and it will probably be better not include them for future classification analyses.

More stability was observed comparing frequency of travel (regular or occasional) and grouping the clusters in four categories: exclusive commuters, non-exclusive commuters, non-commuter residents and leisure travelers. In general, the travel average travel variables for the four groups were similar in 2011 and 2012. As at the cluster level, the major differences from 2011 to 2012 were observed for the percentage of Travelcards and special discount cards. This may be explained because holding Travelcards or special discount cards is related to external factors (such as monthly budget or school registration) that can change over time, causing changes in the type of card acquired without affecting travel behavior. These effects can cause some Pay as You Go users to switch to Travelcards or obtain access to special discount cards, which is plausible specially for high frequency cluster members. However, this can also be caused by the fact that the 2011 and 2012 samples were drawn independently and only 2% of the cards belong to both samples. Again, these results suggest that it may be appropriate either to omit these variables from the clustering process or consider a longer period of analysis.

Only 28% of the cards observed in both 2011 and 2012 belong to the same cluster, however 66% of them belong to a cluster with the same frequency of travel, specially regular user clusters (71%). Grouping the clusters in the 4 aggregate categories: exclusive commuters, non-exclusive commuter, non-commuter resident and leisure traveler, the temporal membership stability increases to 45%. Non-exclusive commuters show the highest temporal stability (55%), supporting the hypothesis that high frequency travelers show more consistent behavior over time.

Cluster characteristics and membership showed greater stability when aggregating the cluster into four homogenous travel groups. Given that each year's sample was drawn independently and represents no more than 4.5% of the complete Oyster Card population, this result it may be indicating that the eight clusters are over-fitted to the sample and may be better to consider only these four travel groups. This also raises the question of whether one week is sufficient for travel behavior analysis, and what would be the appropriate trade-off between sample size and number of analysis days.

Chapter 5

Visitor Travel Behavior

The goal of this chapter is to characterize London visitors' travel patterns using Oyster Card data. The direct way to identify visitors in London using Oyster Card data, is to analyze Visitor Oyster Card users which is a special Oyster Card available to visitors. However, many London visitors do not use this card, instead using either normal Oyster Cards or paper tickets. This last group of visitors could be identified exploring the travel behavior similarities between Visitor Oyster Cards and other travel groups. Therefore, this chapter analyzes Visitor Oyster Card travel behavior and explores its correlation with the travel behavior of the different passenger clusters identified in Chapter 3, to understand not only the travel patterns of Visitor Oyster Cards holders but also of visitors overall and potentially of other non-visitors with similar behavior.

The chapter is divided into four sections. Section 5.1 provides a description of London visitors, summarizing the visitor characteristics captured by two UK visitor surveys and describing the expected visitor travel behavior. Section 5.2 analyzes London visitor travel patterns using data from Oyster Cards specially designed for visitors (Visitor Oyster Card). Section 5.3 uses the results from the clustering analysis performed in Chapter 3 to analyze the visitor membership among different travel profile groups. The chapter ends with a summary of the findings in Section 5.4.

5.1 London Visitors

The United Nations International Recommendations for Tourism Statistics (IRTS), define a visitor as "a traveler taking a trip to a main destination outside his/her usual environment, for less than a year, for any purpose (business, leisure or other personal purpose) other than to be employed by a resident entity in the country or place visited", where the usual environment is defined as "the geographical area (though not necessarily a contiguous one) within which an individual conducts his/her regular life routines" (United Nations, 2008).

London, as the capital city of the UK, is the commercial, financial, and cultural heart of the country. As such it attracts a large number of visitors every year, both for business and tourism. During 2011, 26.3 million overseas and domestic visitors arrived in London, spending more than 118.1 million nights in the city (London & Partners, 2011). For the 2012 Olympic and Paralympic Games, an estimated 590,000 overseas visitors arrived in the UK during the months of July and August (UK Office for National Statistics, 2012). The high number of visitor arriving every year and the differences in travel behavior compared to local residents makes London's overseas and local visitors an interesting group to analyze and study the impact they have on the public transport system.

This section first provides a summary of visitor characteristics based on two different visitor surveys carried out periodically in London and the UK. The section ends by distinguishing between the expected travel behavior of visitors and London residents.

5.1.1 United Kingdom Visitor Surveys

Two visitor surveys are carried out periodically in the UK with the goal of collecting information about overseas and local visitor characteristics: the International Passenger Survey, performed across the UK, and the London Visitor Survey, carried out only in Greater London.

The United Kingdom Office for National Statistics conducts the International Passenger Survey (IPS), which targets passengers entering or leaving the UK at all major airports, sea ports, and train terminals. This survey has been conducted continuously since 1961 with the results mainly used for national economic measures, and for tourism and migration statistics (Office for National Statistics, 2013).

IPS collects information from 700,000 to 800,000 interviewees annually and classifies them as:

- UK resident visitors
- Foreign resident visitors:
    - Short stay visitor: Stay less than 3 months
    - Medium stay visitor: Stay 3 to 6 months
    - Long stay visitor: Stay 6 to 12 months
    - Migrant: Stay more than 12 months

IPS allows public access to the data collected up to the last available quarter. Using data from 2011, Figures 5-1 and 5-2 summarize overseas and UK resident visitor characteristics. During 2011, 42% of UK visitors were overseas residents.

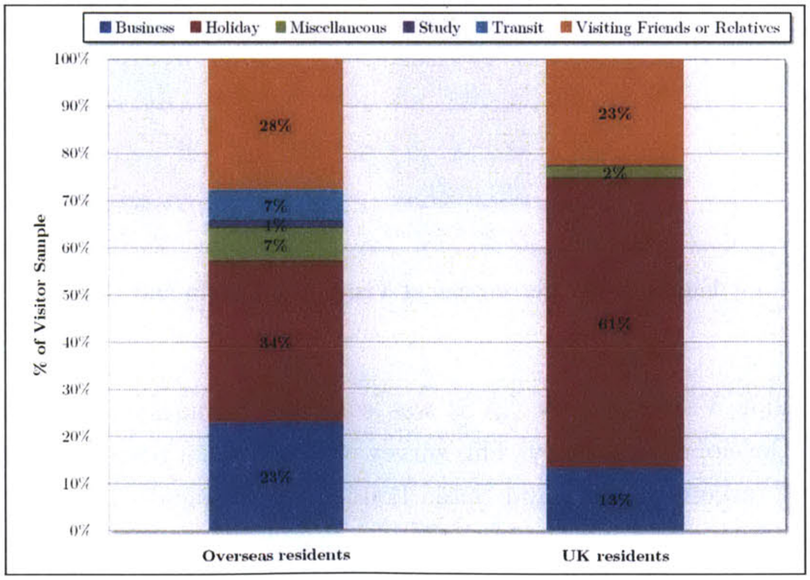


Figure 5-1: Percentage of Visitors by Origin and Purpose

Figure 5-1 shows the trip purpose distribution for overseas and UK residents. As can be seen, 23% of overseas residents and 13% of UK residents are business visitors, and most overseas and UK residents purpose of travel is holiday or visiting friends and relatives (62% and 84% respectively).

The number of nights visitors spend in the UK are summarized in Figure 5-2. Most overseas residents and UK residents stay less than 14 nights (86% and 74% respectively). According to the foreign resident visitor definition above, 99% of overseas resident visitors are short stay visitors (stay less than 3 months).

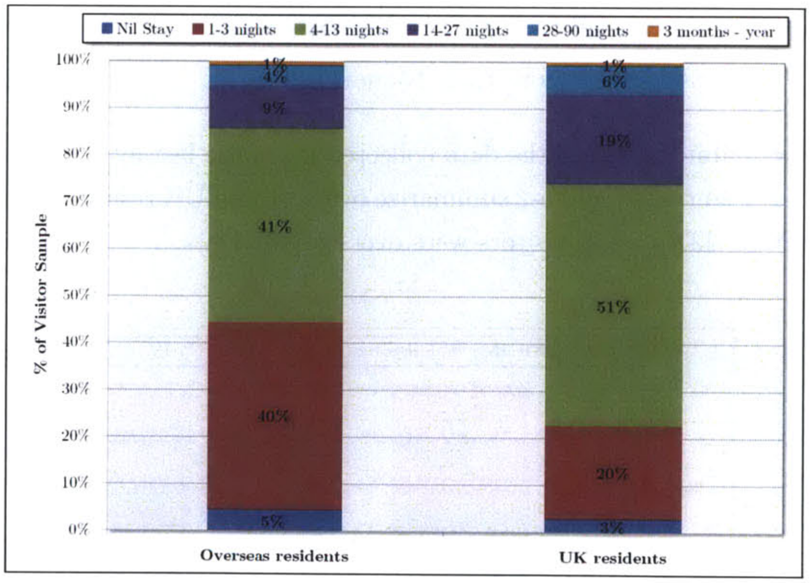


Figure 5-2: Percentage of Visitors by Origin and Length of Stay

The London Visitor Survey (LVS) was conducted annually from 2006 to 2009 by the London Development Agency. This survey was carried out throughout the year at different locations[1] around central and outer London, with slightly larger sample sizes in the summer (July and August). The goal was to collect information from different visitors, identify the strengths and weaknesses of London as a visitor destination, and track visitor

[1] The specific locations are not listed in LVS reports.

satisfaction over time (London Development Agency, 2009).

The LVS uses a sample of approximately 5,000 interviewees per calendar year, grouped into the following categories:

- Overseas visitors
- UK overnight visitors (UK residents who live outside Greater London and are staying at least one night in the capital)
- Day visitors (those on trips between 3 and 24 hours not taken on a regular basis), including:
  - UK day visitors (UK residents who live outside Greater London and are not staying overnight)
  - London residents (living in one of the 33 London boroughs)[1]

According to the 2009 LVS report, most UK day visitors and London residents purpose of the trip was holiday/leisure (68% and 72% respectively). Almost half of overseas visitors (46%) were visiting London for the first time and two thirds of UK day visitors had visited London more than 10 times in the past 5 years. It is important to note that train was the most common mode of transport to arrive in London and the Underground/DLR was the main public transport mode while in London, regardless of visitor type. Overseas visitors were most likely to use the Underground/DLR (88% of them used it while in London), and their next most frequent modes of transport were bus (55%), walking (53%), train (25%) and taxi (10%). Similarly, among London residents the preferred modes of transport were Underground/DLR (51% used LU while in London), walking (43%), bus (38%), train (17%) and car (their own or as a passenger) at 7%. A high percentage of the visitors stay in Central London during their visit (55% of overseas visitors, 45% of UK visitors). The City of Westminster is the most frequently visited place regardless of the trip purpose or visitor place of residence.

[1] London residents are only considered when they are visiting cultural or tourist locations around London and some of the survey questions are not asked of them.

5.1.2 Expected Visitor Behavior: a priori Hypothesis

Visitors can be classified according to their differences with frequent users of London's public transport system. Frequent users are assumed to be London residents who use public transportation on a regular basis with the main purpose of work or study. Based on this hypothesis the following visitor types are defined:

1. *Business Visitors*

   Visitors who come to London with the main purpose of work, study or short term business. They are likely to have modest knowledge of the public transport system and can be first time or returning visitors. Making multiple work or study trips on a regular basis might generate substantial knowledge of the system. According to LVS, business visitors constitutes 7% of UK overnight visitors and 5% of overseas visitors.

2. *Returning Leisure Visitors*

   Visitors whose main travel purpose is holiday/leisure. Visitors in this group have been in London before or visit London regularly, therefore they are likely to be UK residents or overseas visitors from neighboring countries. Additionally, they are likely to have some knowledge of the public transport system. According to LVS, they are 48% of UK overnight visitors and 6 to 12% of overseas visitors.

3. *First Time Leisure Visitors*

   First time visitors whose main purpose of travel is holiday/leisure. They are likely to have little knowledge of the public transport system and to be overseas tourists. They are approximately 36% of UK overnight visitors and 89% of overseas visitors according to LVS.

It is expected that London residents who use public transport on a regular basis for work or study purposes perform one long-duration activity daily at a medium-to-long distance from home. In addition, they may perform short-duration activities such as shopping, recreational, or social activities. Business visitors are likely to behave similarly to London residents, probably having a higher number of short-duration recreational activities during the day, near or far from their base (hotel or friends/relatives home) but focused in Central London. Returning and first time leisure visitors may have completely different patterns of activities than London residents; they are more likely to have short-to-medium duration activities in Central London, and perform more non-public transport trips (notably walking). Additionally, given the difference in travel purpose and schedule flexibility, leisure visitors may travel during off-peak hours and their destinations are likely to be more concentrated. It is also expected that visitors with little knowledge of the system likely prefer different routes and use the most reliable public transport mode (Underground or rail modes). Expected leisure visitor travel behavior is summarized below.

- High number of short-to-medium duration activities
- Trips start during off-peak periods
- Activities focused in Central London. Depending on the origin of the trips, visitor will probably make short trips
- Long walking trips between public transportation trips
- High number of rail trips: the most visible and easy-to-understand public transport mode

5.2 Visitor Oyster Card Travel Behavior

Transport for London issues special Oyster Cards for visitors. Visitor Oyster Cards have the same features as normal adult Oyster Cards with the only difference being that they can be used only for Pay as You Go travel. Visitors can buy a Visitor Oyster Card online on TfL's website[1], or on the Visit Britain website[2], from their home countries and the card will be delivered before they travel to London (Transport for London, 2013d).

Of the 23 million Oyster Cards used during 2012, only 527,000 (2.3%) were Visitor Oyster Cards. The travel characteristics of Visitor Oyster Cards were analyzed, using one week of Oyster Card data from April 15-21st, 2012. 24,857 Visitor Oyster Cards were used that

[1] More details about the Visitor Oyster Card at http://visitorshop.tfl.gov.uk/

[2] http://www.visitbritainshop.com, which is the official shop of British Tourism Authority

week. Similar statistics were also obtained from a random sample of 220,297 non-visitor Oyster Cards used during the same period[1]. This data contained all transactions made on the cards; however, the information was not enough to estimate bus origin and destination locations and to link trip stages into complete journeys[2]. Hence, only journey stages were analyzed and, in some cases, only rail trips were considered.

The distribution of days of travel by visitors and non-visitor Oyster Cards is presented in Figure 5-3. As can be seen, the distribution of travel days for non-visitor cards is consistent with the characteristics of the 2011 and 2012 samples analyzed in Section 3.3 (Figure 3-1), with two peaks at one day and 5 days. On the other hand, most visitors show only one travel day a week, and the frequency decreases as the number of days increases. It is important to note that this results could be affected by the period of analysis. Using only one week of data does not allow to analyze travel continuity between consecutive weeks, especially for leisure travelers such as visitors who are more likely to visit London over a weekend.

Figure 5-4 shows the distribution of the start time of the trips for weekdays for visitor and non-visitor trip stages. As can be seen, there are significant differences between visitor and non-visitor Oyster Card trip starting times. Visitor trips start later than non-visitor trips and do not have as sharp peaks. It is interesting to see that, unlike non-visitors, visitors have a small night peak between 10:00 - 11:00 pm. This tendency is consistent with the expected visitor travel behavior as it is clearly associated with evening leisure activities.

The weekday average activity duration after rail trips is presented in Figure 5-5. As discussed in Chapter 3, the activity duration at any destination is estimated by calculating the time between the exit time and the subsequent entry transaction. Since journey stages were not linked in this analysis, activities shorter than 30 minutes were assumed to be interchanges, and only activities that started and ended at the same station were included. The activity duration used was the average of all day activities. While non-visitors have

[1] For ease of reference this sample of non-visitor Oyster Cards will simply be refer to as non-visitors, even though it certainly includes some (unknown) number of visitors

[2] The data misses some flags necessary to identify interchanges, making it infeasible to apply the ODX tool to the database

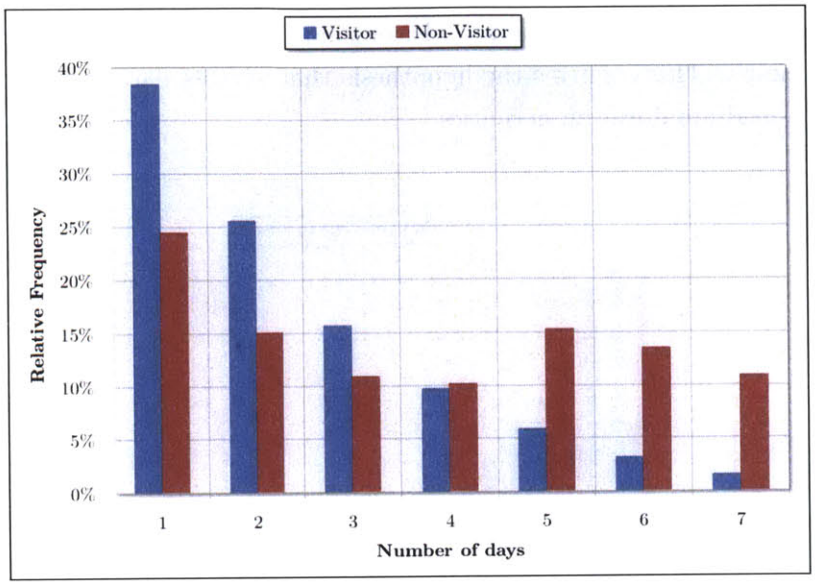


Figure 5-3: Days of Travel

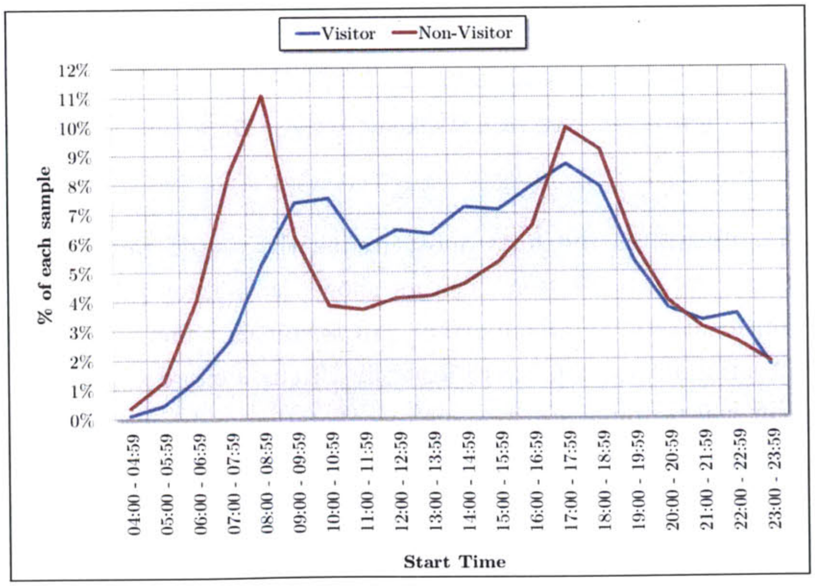


Figure 5-4: Average Weekday Journey Stage Start Time

a peak between 8 and 9.5 hours, visitors have a higher peak for activities lasting less than 2 hours. This confirms the hypothesis that visitors participate in a high number of short-to-medium duration activities.

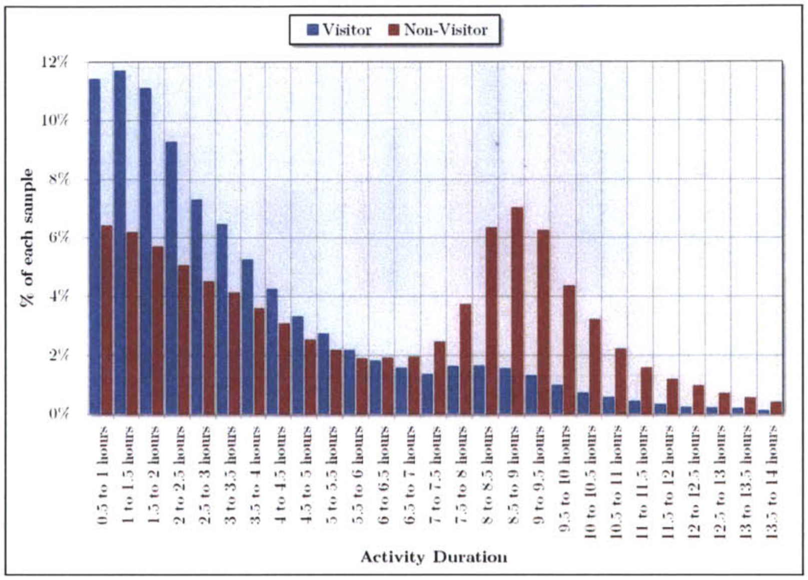


Figure 5-5: Average Weekday Activity Duration (After Rail Trips)

Figure 5-6 illustrates the forty most frequently visited rail stations for visitors and non-visitors. As can be seen, visitors show the highest percentages at either tourist or cultural locations, or at important rail terminals. The percentage of non-visitors is lower at tourist stations such as Piccadilly Circus, Oxford Circus and Westminster, but is higher at rail terminal stations such as Waterloo and London Bridge. Additionally, it was estimated that at least 32% of visitor cards used some non-public transport modes to move between activities (distances longer than 2 kilometers between the destination of one trip and the start of the next). This percentage drops to 13% for non-visitor card users.

Figures 5-7 and 5-8 show the visitors twenty-five most common non-public transport movements between activities for visitors and non-visitors respectively. The bars represent the percentage of Oyster Cards that make a non-public transport movement and the circles

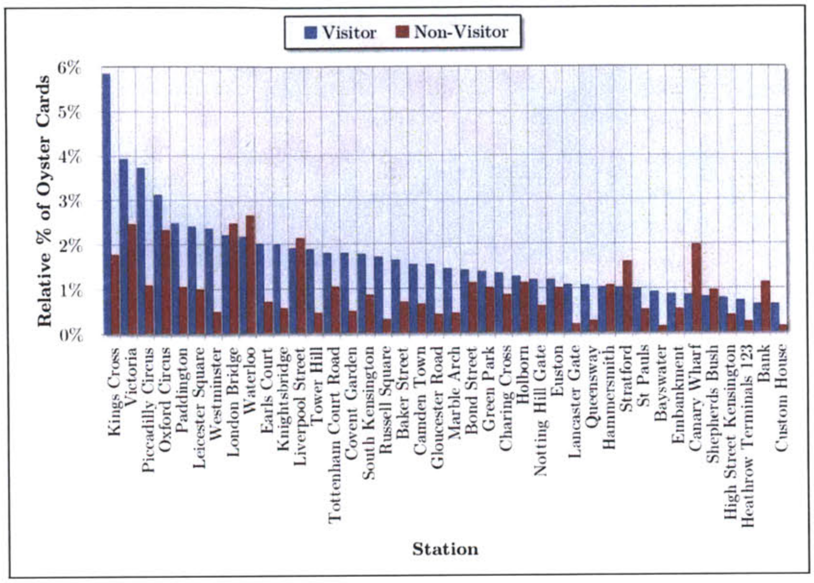


Figure 5-6: Most Frequently Visited Rail Stations

represent that movement distance[1]. For example, approximately 3.3% of non-public transport movements are made between Piccadilly Circus and Oxford Circus, and the distance between these stations is 0.8 kilometers. Note that for visitors, most such movements are made within walking distance and take place in Central London. For non-visitors, the most common movement between activities is in the shopping area between Oxford Circus and Bond Street with 3.2%. Unlike visitors, this first high percentage decreases sharply to under 2% for the following non-visitor movements. Additionally, non-visitor movements shown in Figure 5-8 are not greater than 1 kilometer. These results indicate that visitors make more and longer non-public transport trips between activities, findings that are consistent with the expected visitor behavior described in Section 5.1.2.

Additionally, it was estimated that while 53% of Visitor Oyster Card users only take rail and 11% only take bus when using public transportation, 31% of non-visitor Oyster Card users only take rail and 24% only bus. This result may be indicating that, as expected,

[1] Euclidean distance between stations

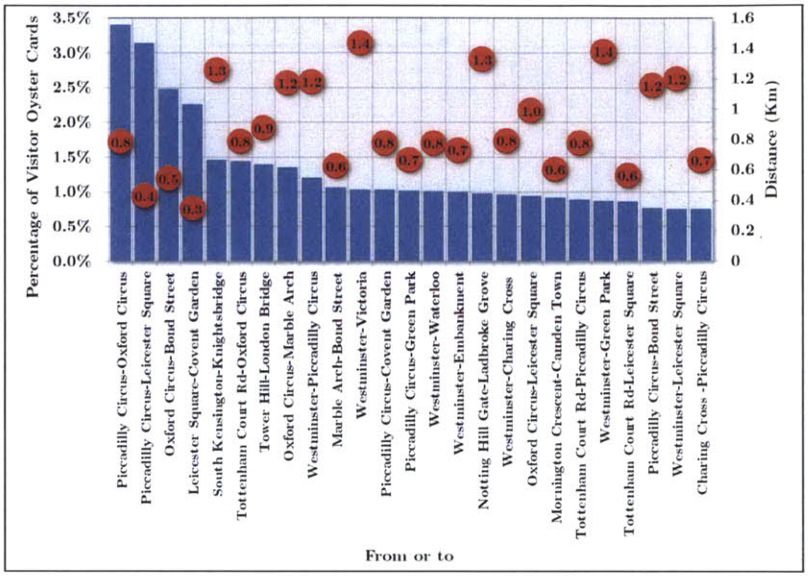


Figure 5-7: Most Frequent Visitor Non-Public Transport Movements

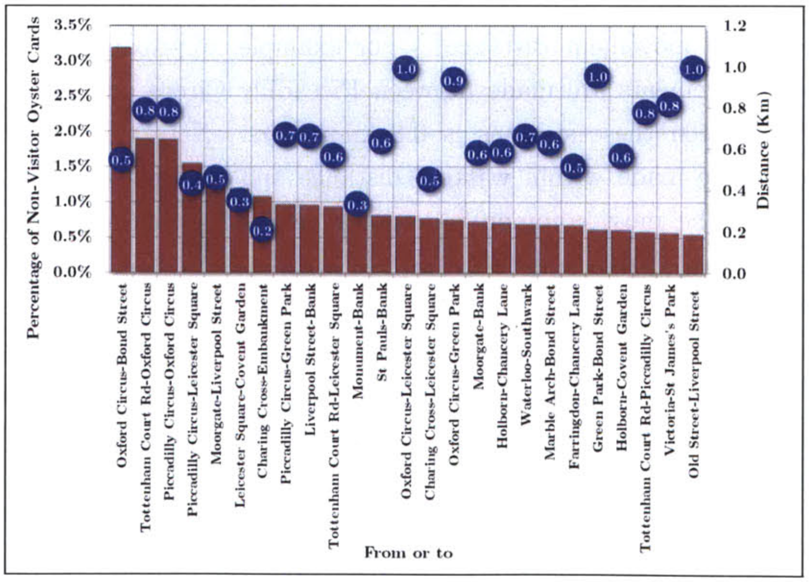


Figure 5-8: Most Frequent Non-Visitors Non-Public Transport Movements

visitors prefer the most visible and easy-to-understand public transport mode, but it is also probably related to their locations (i.e. Central London).

5.3 Visitor Travel Profile

Complementing the visitor travel behavior analysis presented above, this section studies visitor travel patterns by studying the characteristics of the travel groups to which they belong. Since there are an unknown number of visitors that holds non-visitor Oyster Cards, this section aims to analyze if Visitor Card users behavior is similar to other travel groups in order to identify potential visitors groups that do not hold Visitor Oyster Cards.

From the 248,241 Oyster Cards in the 2011 sample used for the cluster analysis (Section 3.4), only 1,072 are Visitor Oyster Cards which represents 0.43% of the sample. This small number of Visitor Cards are distributed among the eight clusters in different proportions, as indicated in Figure 5-9.

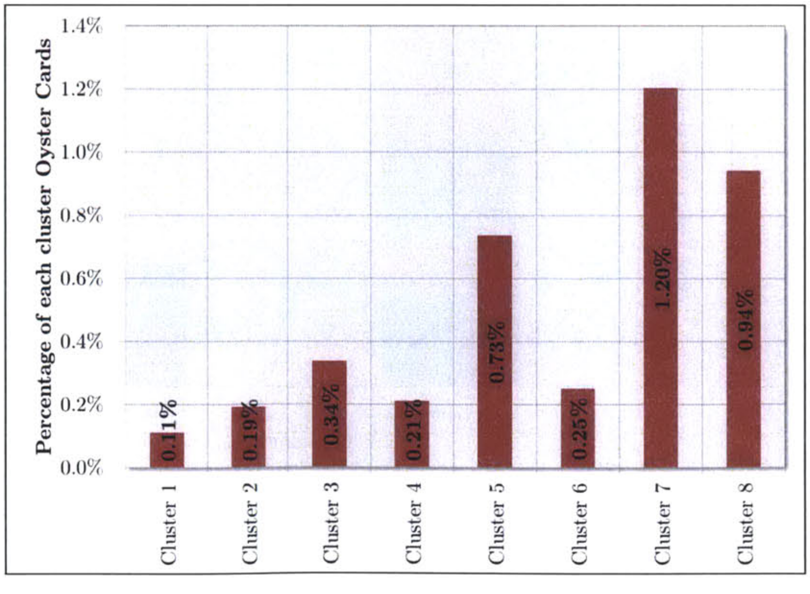


Figure 5-9: Visitor Cluster Membership

As you can see, Visitor Oyster Card holders represent no more than 1.2% of any cluster. Clusters 4, 7, and 8, show the highest percentage of Visitor Oyster Card holders, a result that was expected given that these are occasional user clusters traveling for leisure purposes. As expected, the highest travel frequency cluster (cluster 1) has the lowest percentage of Visitor Oyster Cards (0.1%).

The distribution of Visitor Oyster Cards among the clusters is shown in Figure 5-10. Clusters 4, 7, and 8 have the highest percentages, with cluster 8 having the largest proportion (28%). The bar chart also shows that the occasional user clusters (5 through 8) represent the largest portion of the Visitor Oyster Card population (76%). This is consistent with this group's travel behavior, whose activities are of short duration, during off-peak hours and made in Central London. The highest travel frequency group (cluster 1) has the lowest share of Visitor Oyster Cards (4%), as expected.

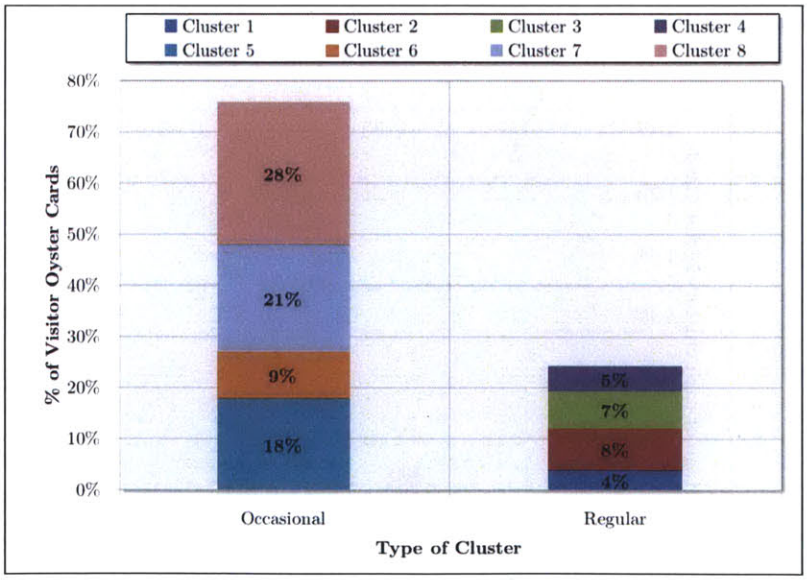


Figure 5-10: Visitor Oyster Cards Distribution by Cluster

The results are consistent with the Visitor Cards travel behavior analyzed using the sample of April 2012. Clusters 5, 7, and 8 seem to represent travel profiles most similar to visitor behavior. Indeed, all three cluster members make their first trip during the midday, travel no more than 3 days a week, and perform activities of no more than 4.5 hours. Cluster 8 (weekday rail occasional users) has the highest share of the visitor population and cluster 7 (weekend occasional users) shows the next highest share. Both clusters 7 and 8 are leisure travelers whose travel characteristics are similar to the expected visitor behavior described in Section 5.1.2: its members travel one day a week during off-peak hours, engage in short to medium duration activities, prefer rail over bus (in the case of cluster 8) and show high usage of international terminal stations, particularly Heathrow Airport (cluster 7) and London City Airport (cluster 8). Therefore, these results support the hypothesis that these clusters include a high percentage of visitors (probably holding Pay as You Go cards), which according to their temporal and spatial travel characteristics are business (cluster 8) and leisure visitors (cluster 7).

Figure 5-11 shows the percentage of cluster members that belong to each user type. As can be seen, clusters with high percentage of visitors (5, 7, and 8) do not show the highest percentage of other types of users (elderly, disabled, or young), but do have a high combination of these three groups, especially cluster 5 (25%). These results indicate that there may be some similarities between the behavior of these three sociodemographic groups and visitors.

5.4 Summary

This chapter provided an overview of London visitor travel behavior characteristics based on Oyster Card data analysis and visitors cluster membership. More than 26 million visitors arrived in London during 2011 and 590,000 visited during the Olympic Games in 2012. According to UK's International Passenger Survey 42% of the visitors that arrived to the UK during 2011 were overseas residents and most of the visitors stayed for short periods for leisure or holidays purposes. In addition, 88% of overseas visitors use the Underground while in London according to the London Visitor Survey.

The analysis of Visitor Oyster Card travel transactions and the subsequent visitor cluster membership analysis support the hypotheses about expected visitor travel behavior. It was observed from the Visitor Oyster Card transactions that visitors travel few days a week (mostly one to three days), during off-peak hours performing short to medium duration

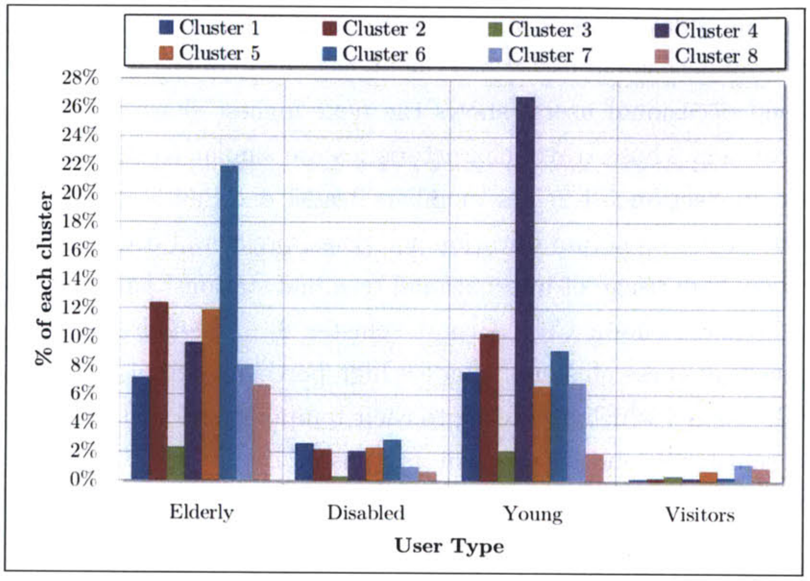


Figure 5-11: User Types Cluster Membership

activities (up to 2 hours). They perform most of their activities in Central London, making non-public transport trips between public transport journeys and preferring rail, the most reliable public transport mode, over bus (70%).

The distribution of Visitor Oyster Cards among clusters showed that occasional user clusters present similar travel characteristics, which could indicate that some members of these groups are visitors holding non-visitor Oyster Cards. Additionally, the sociodemographic distribution of the clusters indicated that visitor behavior may have some similar characteristics with elderly, disabled, and student groups. Those clusters with a high percentage of Visitor Oyster Cards, especially leisure traveler (clusters 7 and 8), showed travel behavior characteristics similar to visitor behavior: low travel frequency, traveling during off-peak hours, with short-to-medium duration activities, preferring rail over bus, and using international terminal stations. Previously analyzed spatial and temporal travel characteristics indicate that these are likely to be business visitors (cluster 8) and leisure visitors (cluster 7). The identification of these clusters could be a first step to identify visitors that hold non-visitor Oyster Cards. A recommended future research line is to analyze the Oyster Card "life" of these clusters' members (time between the last use and issue date of the card) to refine the identification of visitors.

Chapter 6

Oyster Card Registration and Churn

The goal of this chapter is to understand how users travel patterns can affect the decisions they make about card ownership over time. These decisions relate to actions such as the card registration in TFL's system, keeping the same card for long periods of time, using multiple cards, and having one card used by more than one household member. In this thesis the registration status and the Oyster Card attrition rates (known as Oyster Card churn) are explored. Registered cards are a sample within the Oyster Card population that can be reached more easily than other users given that TfL has contact information for them (address, telephone, and/or email). Hence, cheaper, more focused and more efficient surveys can be undertaken with registered users. The analysis of registered users travel behavior can allow determining whether these users' behavior is representative of the population's behavior which could validate them as a focus group. On the other hand, understanding the reasons behind Oyster Card attrition, can be a first step to understanding customer attrition. Separating the effect of customer attrition from other effects, such as seasonality, special events and impact of internal or external projects, could lead to more accurate predictions of passenger demand, which would improve the evaluation of strategic or operational planning changes, and the assessment and improvement of fare incentives.

This chapter includes three sections: Section 6.1 provides an analysis of the travel behavior of registered and unregistered cards, Section 6.2 explores the travel characteristics of Oyster Card users whose cards are no longer active in the system, and Section 6.3 summarizes the chapter's findings.

6.1 Registration Status

An Oyster Card is registered in the system, when the card user provides contact information to TfL associated with the unique Oyster Card number. Registration can be done in person at an Underground or Overground station, Oyster Ticket Stops, at London Travel Information Centers, or online on TfL's Oyster website. As reviewed in Section 1.4.3, registered users can review their Oyster Card transaction history online, and are protected against card loss or theft. Users need to be registered in order to acquire a monthly or annual Travelcard.

TfL has a range of policies and processes to control and safeguard access to, and use of, personal information associated with Oyster cards. The registered user information that TfL stores in their system includes:

- Title (Mr/Mrs/Ms/Miss)
- First name, middle initial and surname
- Address
- Telephone number and email address (only if the user applied online)
- Password
- Encrypted bank card details of customers who purchase Oyster products using a debit or credit card
- History of automatic payment transactions including location, date and time
- History of Oyster Card transactions (up to eight weeks)

In the following sections, an overview of the registered Oyster Cards characteristics is provided. First, some current general statistics about registration status are presented, to continue with a description of registered users travel patterns.

6.1.1 Registered Oyster Cards

During 2012, approximately 23 million Oyster Cards were used in London's public transport system. 80% of these cards are normal or retail Oyster Cards, which means that the card has no special features. The remaining 20% is composed by Photocards (10%), staff passes (0.4%), freedom passes (7%), credit cards[1] (0.2%), and visitor cards (2%).

Of the Oyster Cards used during 2012 26.7% are registered, which correspond to approximately 6 million cards. Most of these cards are retail Oyster Cards (67%), with the remaining 33% composed mainly of Photocards (31%), that require user registration, and credit cards (2%). Table 6-1 provides a summary of registration statistics for 2012.

Table 6-1: 2012 Oyster Card Registration Statistics

Type of Oyster Card	Total Number (percentage) of cards	Number (percentage) of registered cards	Percentage of all Oyster Cards
Retail	18,330,173 (80.0%)	4,100,832 (67.0%)	22.37%
Photocard	2,347,844 (10.2%)	1,910,176 (31.2%)	81.36%
Staff	114,384 (0.5%)	81 (0.001%)	0.07%
Freedom	1,509,362 (6.6%)	19,953 (0.3%)	1.32%
Credit card	91,609 (0.4%)	91,565 (1.5%)	99.95%
Visitor	527,055 (2.3%)	1,593 (0.03%)	0.30%
Total	22,920,427 (100%)	6,124,200 (100%)	26.72%

Currently, TfL has information about more than 6 million registered cards (active or inactive). Of that total, 2.7 million cards were active between October 17th and October 23rd, 2011, which corresponds to a 46.4% of all the Oyster Cards used in that period. Approximately 117,000 registered cards were observed in the random sample extracted from the 2011 data (see Section 3.2.4), which is 47.3% of the total random sample. This data will be used for the registered users travel behavior analysis presented in 6.1.2.

[1] In 2007 the British bank Barclays launched a card that combines standard Oyster card with credit card functionality

6.1.2 Registered User Travel Behavior

Analyzing the distribution of registered users among different travel groups, it is possible to characterize the travel behavior of registered Oyster Card users. Determining whether registered users are representative of the complete Oyster Card population can either validate the devclopment of morc efficient, more focused and less expensive surveys, or can provide knowledge of the travel characteristics of the sample which provides sampling strategy improvement opportunities.

Using the results from the clustering process described in Section 3.4.2 using the 2011 data samples, the distribution of registered users among clusters was explored. Between 35% and 62% of each cluster members are registered in TfL Oyster Card system, representing an average of 48% across clusters. Figure 6-1 shows the percentage of registered cards observed in each cluster. The bar chart shows that the registered users of clusters 5 through 8 (occasional users) are 41% or less of the total size of each cluster (39% average), while registered users represent 51% or more (56% average) of clusters 1 through 4 (regular users). This difference (17% on the average) implies that regular or frequent users are more likely to register their Oyster Cards than occasional or sporadic users. One of the reasons for this is that regular user clusters have a higher percentage of members that hold Travelcards or special discount passes, which encourages (or requires in the case of monthly and annual Travelcards) the registration of their cards.

Figure 6-2 compares the cluster size within the complete sample and within the registered users. The two distributions are very similar with the differences varying between 1% and 4%. Cluster 6, weekday bus occasional users, presents the highest difference between total and registered cluster size (4%). This cluster has the highest percentage of elderly (22%), which may explain the lower registration rate.

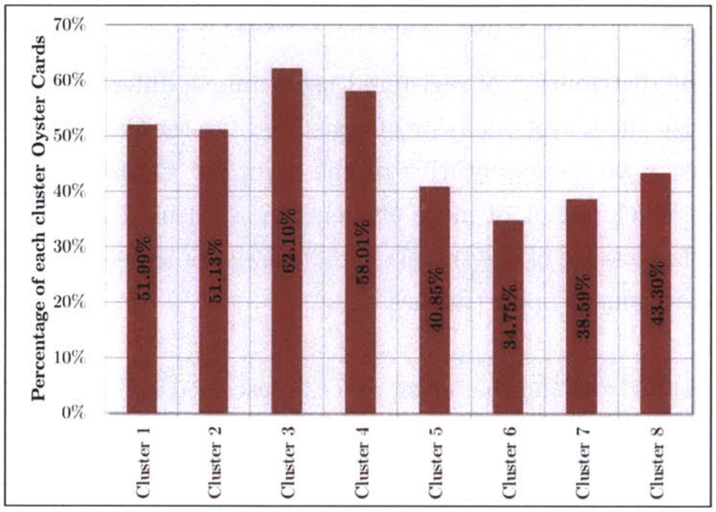


Figure 6-1: Percentage of Registered Oyster Cards per cluster

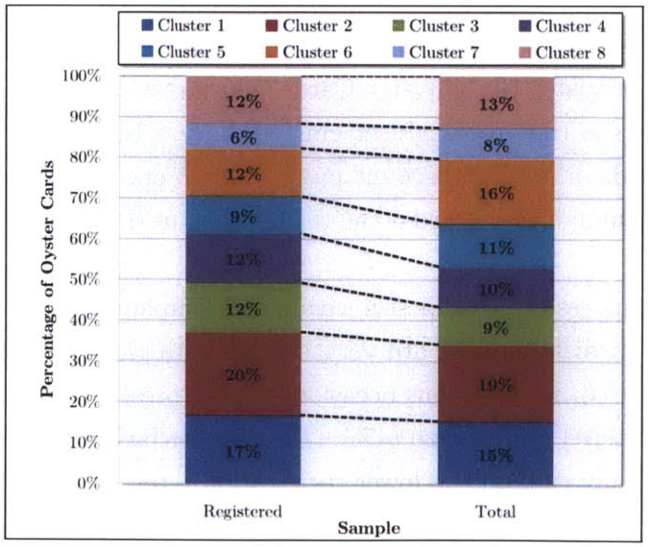


Figure 6-2: Registered and Total Oyster Cards Cluster Size Percentage

The differences between the cluster centroids for all the members and the average travel characteristics of registered members were estimated in order to determine whether or not registered users are representative of each cluster. The differences between the standard deviations of all the members and registered users were also computed. The resulting differences as a percentage of each cluster total population are shown in Tables 6-2 and 6-3 (see Appendix B for absolute differences). The colored cells show the relative difference for clusters and variables (from green for minimum difference, to red for maximum difference).

Table 6-2: Centroid Differences as Percentage of Total Cluster Centroid

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Average	Maximum	Minimum
Days of Use	0%	3%	1%	3%	1%	3%	1%	2%	2%	3%	0%
Weekdays Main Activity Duration	4%	11%	1%	8%	8%	16%		6%	8%	16%	1%
Weekends Main Activity Duration	0%	2%			4%		2%		2%	4%	0%
Weekdays Shortest Activity Duration	8%	13%	1%	13%	10%	16%		8%	10%	16%	1%
Weekends Shortest Activity Duration	1%	2%			8%		5%		4%	8%	1%
Weekdays First Journey Start Hour	1%	3%	1%	3%	2%	2%		0%	2%	3%	0%
Weekends First Journey Start Hour	2%	3%			2%		0%		2%	3%	0%
Weekdays Last Journey Start Hour	1%	2%	0%	0%	2%	3%		1%	1%	3%	0%
Weekends Last Journey Start Hour	1%	2%			2%		0%		1%	2%	0%
Percentage of Bus Exclusive Days	8%	11%	11%	2%	7%	2%	4%	0%	6%	11%	0%
Percentage of Rail Exclusive Days	14%	13%	1%	11%	9%	13%	5%	1%	8%	14%	1%
Weekly Maximum Travel Distance	2%	5%	2%	4%	5%	7%	3%	1%	3%	7%	1%
Weekly Minimum Travel Distance	3%	4%	2%	2%	8%	5%	6%	1%	4%	8%	1%
Percentage of Different First Origins, Weekdays	3%	4%	2%	4%	1%	0%		1%	2%	4%	0%
Percentage of Different First Origins, Weekends	2%	3%			1%		0%		2%	3%	0%
Percentage of Different Last Origins, Weekdays	1%	1%	1%	2%	0%	0%		0%	1%	2%	0%
Percentage of Different Last Origins, Weekends	3%	3%			1%		1%		2%	3%	1%
Percenage of Travelcards	3%	2%	8%	3%	27%	37%	22%	54%	19%	54%	2%
Percentage of No Special Discount Cards	5%	8%	3%	10%	7%	15%	1%	6%	7%	15%	1%
Average	3%	5%	3%	5%	5%	9%	4%	6%			
Maximum	14%	13%	11%	13%	27%	37%	22%	54%			
Minimum	0%	1%	0%	0%	0%	0%	0%	0%			

Clusters 2, 4, 5, and 6 show the highest differences between registered users and the total cluster for both the average and standard deviation. These clusters show some of their highest differences for activity duration and mode choice. The highest differences are observed for the percentage of Travelcards, especially for occasional clusters 5 through 8. It is important to note that except for Travelcards and Special Discount cards the differences are no greater than 16%.

Table 6-3: Standard Deviation Differences as Percentage of Total Cluster Standard Deviation

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Average	Maximum	Minimum
Days of Use	5%	6%	1%	7%	3%	4%	3%	3%	4%	7%	1%
Weekdays Main Activity Duration	6%	10%	6%	11%	2%	9%		2%	7%	11%	2%
Weekends Main Activity Duration	2%	3%			1%		1%		2%	3%	1%
Weekdays Shortest Activity Duration	2%	6%	4%	4%	4%	11%		2%	5%	11%	2%
Weekends Shortest Activity Duration	2%	2%			4%		1%		2%	4%	1%
Weekdays First Journey Start Hour	6%	2%	8%	4%	1%	6%		1%	4%	8%	1%
Weekends First Journey Start Hour	0%	1%			1%		1%		1%	1%	0%
Weekdays Last Journey Start Hour	8%	8%	5%	7%	3%	0%		1%	4%	8%	0%
Weekends Last Journey Start Hour	2%	2%			0%		1%		1%	2%	0%
Percentage of Bus Exclusive Days	1%	1%	4%	0%	1%	5%	0%	0%	2%	5%	0%
Percentage of Rail Exclusive Days	5%	3%	2%	7%	1%	6%	1%	5%	4%	7%	1%
Weekly Maximum Travel Distance	0%	0%	1%	1%	0%	3%	2%	2%	1%	3%	0%
Weekly Minimum Travel Distance	3%	2%	2%	1%	4%	0%	4%	1%	2%	4%	0%
Percentage of Different First Origins, Weekdays	1%	4%	2%	3%	4%	1%		2%	2%	4%	1%
Percentage of Different First Origins, Weekends	1%	2%			2%		2%		1%	2%	1%
Percentage of Different Last Origins, Weekdays	2%	3%	2%	2%	0%	2%		5%	2%	5%	0%
Percentage of Different Last Origins, Weekends	0%	8%			2%		4%		4%	8%	0%
Percenage of Travelcards	4%	0%	1%	0%	10%	13%	9%	29%	8%	29%	0%
Percentage of No Special Discount Cards	10%	9%	26%	2%	11%	8%	2%	31%	12%	31%	2%
Average	3%	4%	5%	4%	3%	5%	2%	6%			
Maximum	10%	10%	26%	11%	11%	13%	9%	31%			
Minimum	0%	0%	1%	0%	0%	0%	0%	0%			

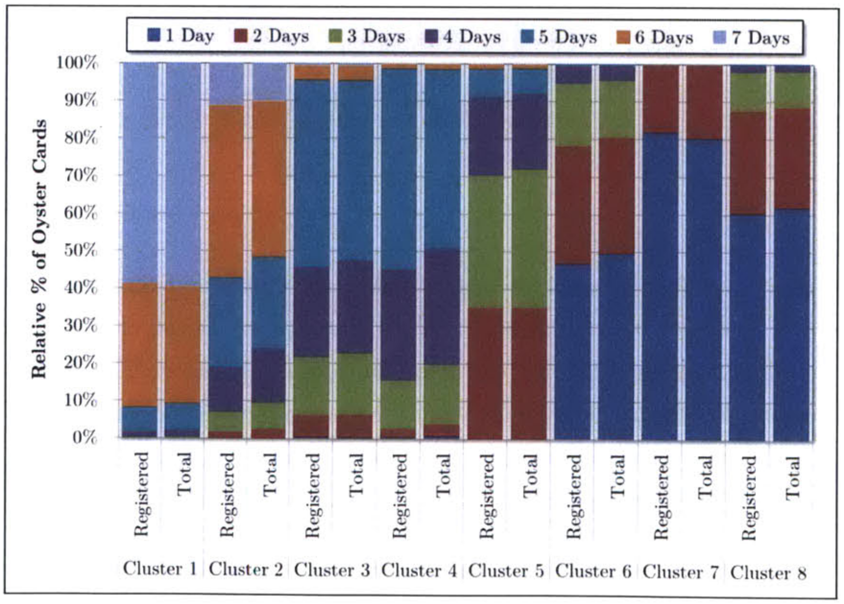


Figure 6-3: Days of Travel per cluster

The distribution of days with travel for registered users and the complete population of each cluster is shown in Figure 6-3. Most clusters show very similar distributions. Clusters 2 and 4 seem to show some differences between the distributions, with registered users traveling more days than the complete population.

As can be seen in Tables 6-2 and 6-3, the journey start time shows one of the travel variables smallest differences and the activity duration shows the highest. Figures 6-4 and 6-5 show for regular and occasional user clusters, a comparison between the registered users distribution of the start time of both the first and last journeys of the day and the complete population distribution for the same variable[1]. Figures 6-6 and 6-7 show the same comparison but for the distribution of the main activity duration.

Occasional clusters 5 and 6 show more differences when comparing the distribution of registered users and the total cluster. Registered users of these clusters seem to travel later than the total cluster which is more noticeable for cluster 6 that shows a sharp peak between 3:00 and 4:00 pm. The activities of registered users of clusters 5 and 6 have longer duration than the total cluster activities and their distribution presents small peaks around 8 hours. The distributions of the registered users of clusters 7 and 8 are very similar to the total population distribution.

Doing the same analysis for regular clusters, clusters 2 and 4 show more differences between the distributions. Registered users of these clusters have sharper peaks than the total cluster, especially for the morning peak. Clusters 1 and 3 registered users also show sharper morning peaks but the difference with the total population distribution is smaller. Clusters 2 and 4 show more registered users performing activities of more than 6 hours than the total cluster population. In general, the distributions of the registered users of clusters 1 and 3 are very similar to the total population distribution.

[1] The distribution aggregates the number of first and last journeys made by each Oyster Card.

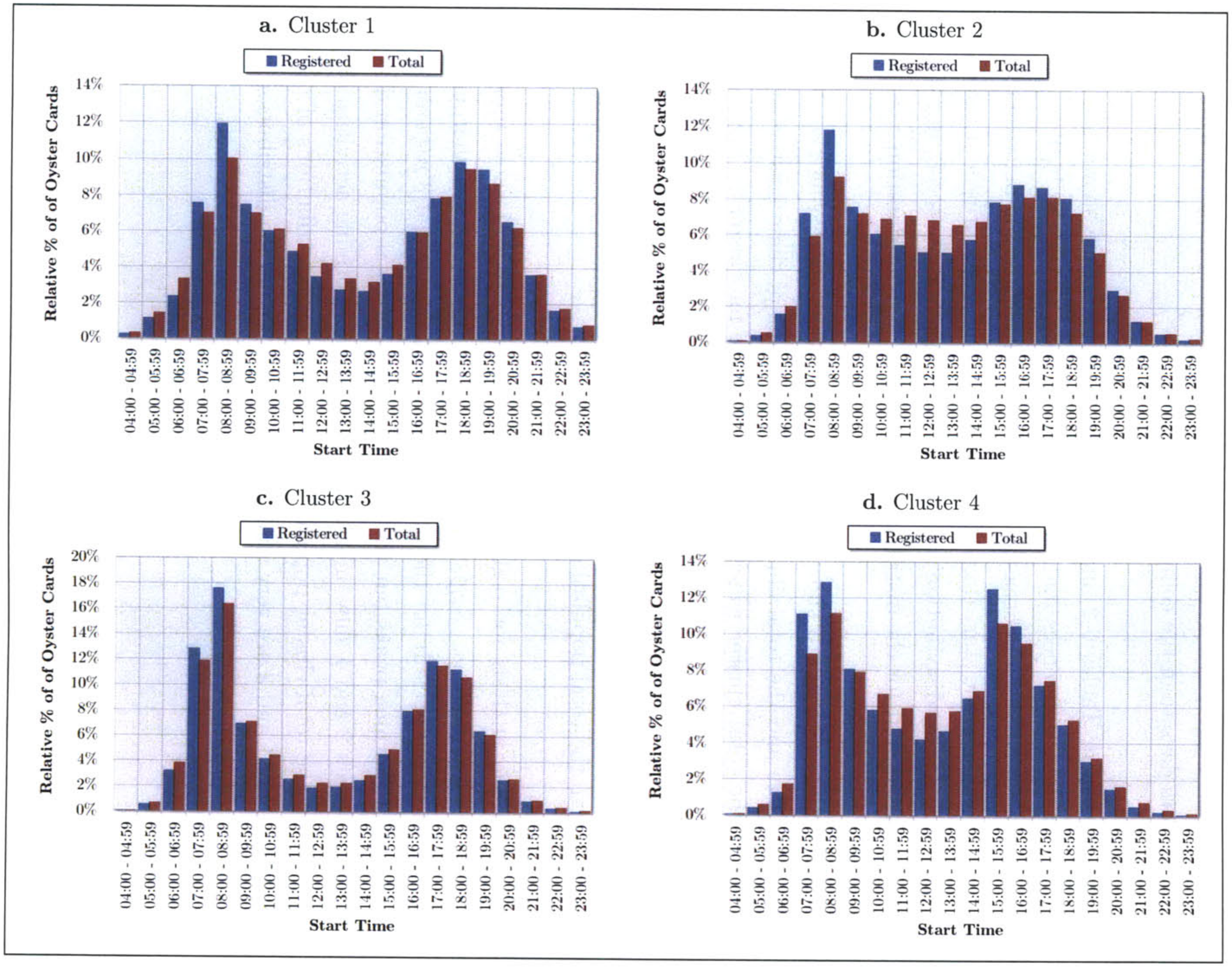


Figure 6-4: Journey Start Time - Regular Users

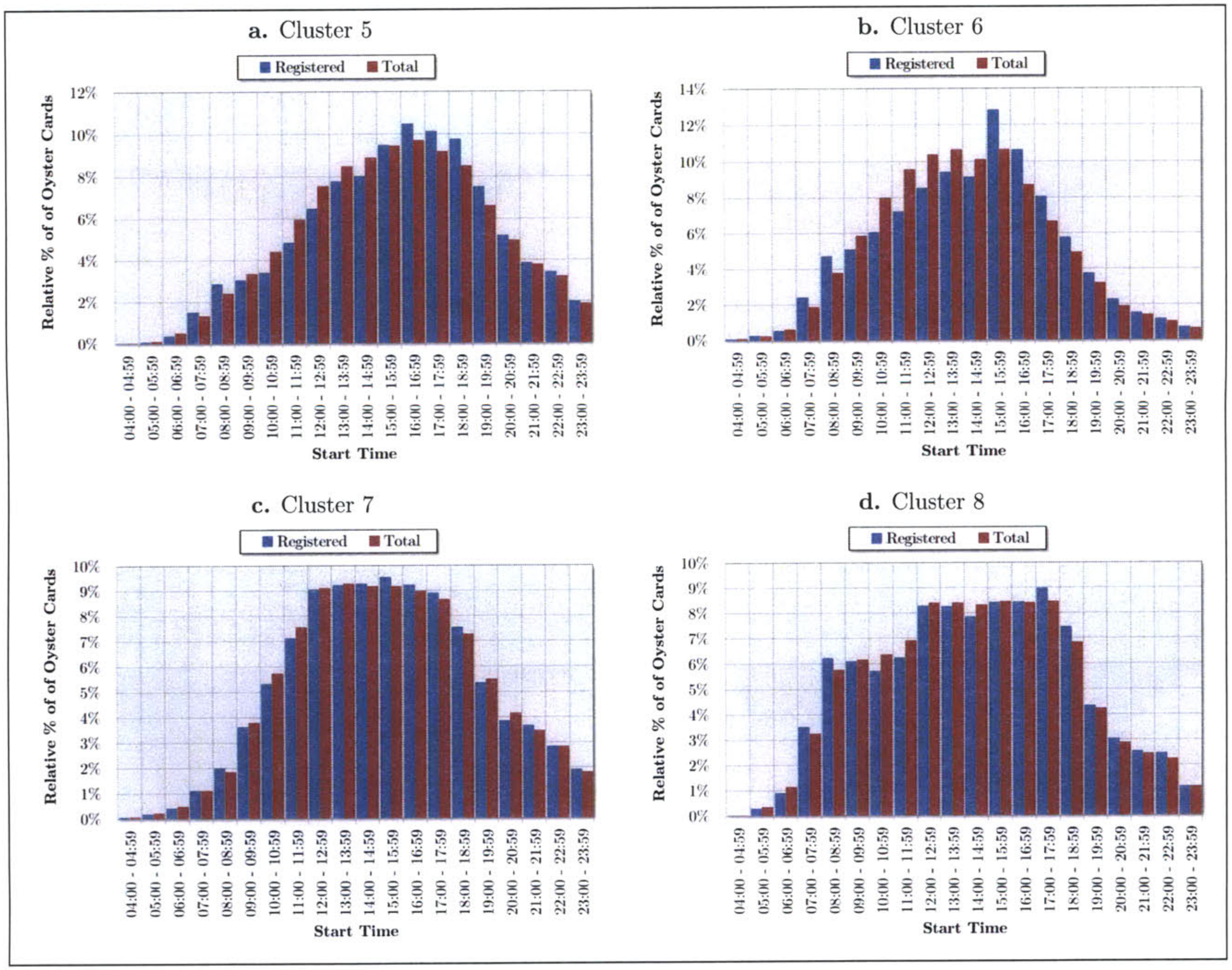


Figure 6-5: Journey Start Time - Occasional Users

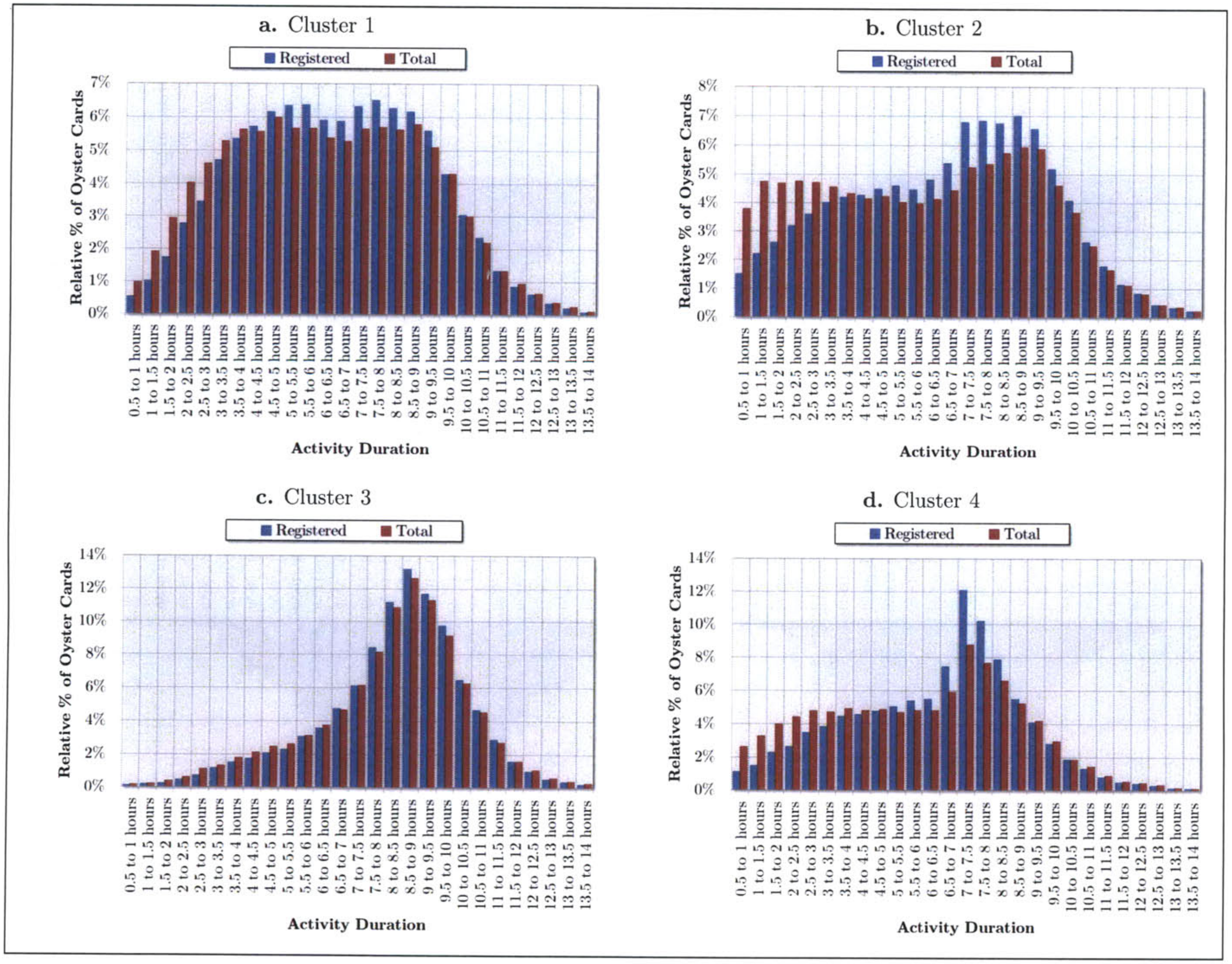


Figure 6-6: Main Activity Duration - Regular Users

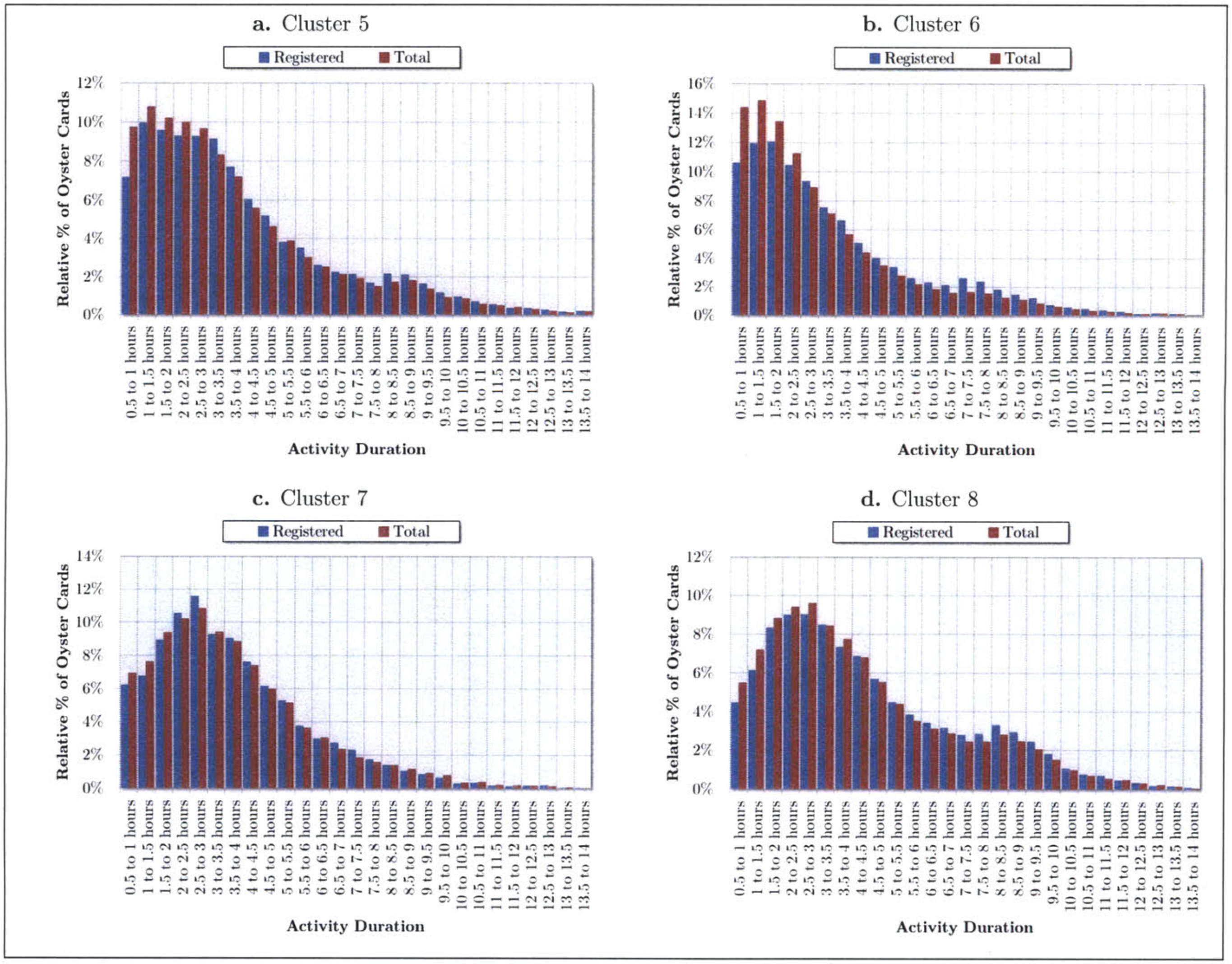


Figure 6-7: Main Activity Duration - Regular Users

6.2 Attrition Rates: The Churn Problem

Customer attrition, also called churn, addresses the turnover rate of current customers of a specific system. In the case of Oyster Cards, where an Oyster Card is not necessarily associated with a unique user and a unique user does not necessarily use only one Oyster Card, customer attrition and Oyster Card attrition are not the same. Despite the fact that Oyster Cards are intended to be personal and non-transferable, because of London Underground fare structure, an Oyster Card can be used by more than one user at the same time on buses, while it is possible that one Oyster card is used by more than one user for trips at different times on buses and the Underground. Additionally, it is possible for a single user to have more than one Oyster Card and alternate their use.

Having better knowledge about a public transport system attrition rates and the reasons for customer attrition can help to improve fare incentives to retain customers, and to understand changes in ridership. User attrition allows measuring different effects on the system, separating user attrition from other possible impacts, such as impacts of line extensions on demand over time. However, for a public transport system, customer attrition is not easy to measure directly. Public transport users have no contractual relationship with the system provider; therefore, they can change their mode choice at any time, even to make a single trip. Despite the fact that user attrition analysis is more helpful to understand changes in the demand structure, the analysis presented in this section focuses only on Oyster Card attrition. Estimating card attrition establishes a starting point for user attrition analysis. It is important for time series analysis to characterize card attrition over time, and identify possible patterns that could describe it. Understanding Oyster Card churn behavior can lead to more accurate estimates of demand, and separates user attrition from other effects, such as seasonality, special events and impact of internal or external projects.

To measure Oyster Card attrition over time it is necessary to determine which cards are no longer in use as a function of the time since they were last observed in the system. Oyster Card data allows the identification of periods when an Oyster Card is inactive, i.e. the card is not being used in the system. Oyster Card attrition can result from several events including card loss, migration, change of mode choice or death. An Oyster Card can become inactive from one period to another whether or not the card owner is using the system. The holder of an inactive card can be active using a different Oyster Card because s(he) owns multiple cards, or because the card was lost and s(he) is using a new card. In this case card attrition rate will be higher than user attrition rate.

This section aims to analyze Oyster Card attrition and characterize churned Oyster Card travel patterns. Simple measures are developed to estimate Oyster Card attrition rates in 2010 and 2011, using monthly Oyster Card records. Additionally, the travel group membership of churned Oyster Cards is also analyzed.

6.2.1 2010-2011 Oyster Card Attrition

Oyster Cards with transactions over a specific period of time are termed *Active Cards*. Oyster Cards with no observed transactions over that time are termed *Inactive Cards*. Therefore, active and inactive card status depends on the analysis time period. The time periods analyzed cover one week of daily records for each month from March 2010 to October 2011. The weeks selected correspond to the first available weeks of every month that did not have important holidays or special events. Table 6-4 summarizes the data used.

Figure 6-8 shows the number of active cards in each week over the two-year period. The continuous line shows the number of active cards growing steadily over time, except for declines in August in both years. These drops are likely due to the holiday period, when many regular Oyster Card users are away from London. Hence, for this analysis the August data was excluded. The results from the linear model fitted to the data show that the number of active cards (in a week) has been increasing over time at an average rate of about 50,300 Oyster Cards per month. The OLS results present a correlation coefficient close to 1 ($R^2 = 0.95$) which indicates a high goodness of fit.

By analyzing the number of active cards in each period, it is possible to obtain the number of these cards remaining active in subsequent periods. Figure 6-9 shows the status of the active cards observed in each week (Table 6-4) over subsequent periods as a percentage of the initial number of active cards. The number of active cards diminishes over time,

Table 6-4: Weeks of Data Analyzed - 2010/2011

Year	Period	Week of the Year
2010	March 1st to 7th	10th
	April 12th to 18th	16th
	June 7th to 13th	24th
	August 16th to 22nd	34th
	September 13th to 19th	38th
	October 11th to 17th	42nd
2011	April 10th to 16th	16th
	May 8th to 14th	20th
	June 6th to 12th	24th
	July 11th to 17th	29th
	August 15th to 21st	34th
	September 19th to 25th	39th
	October 17th to 23rd	43rd

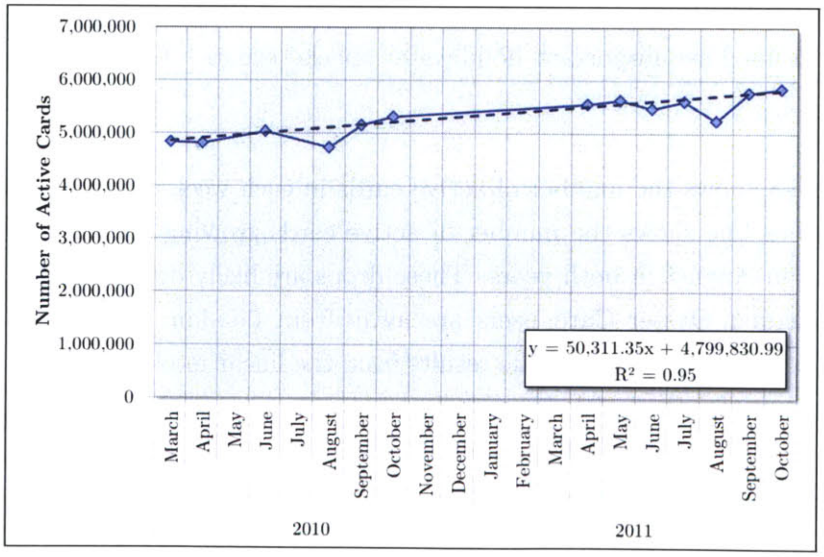
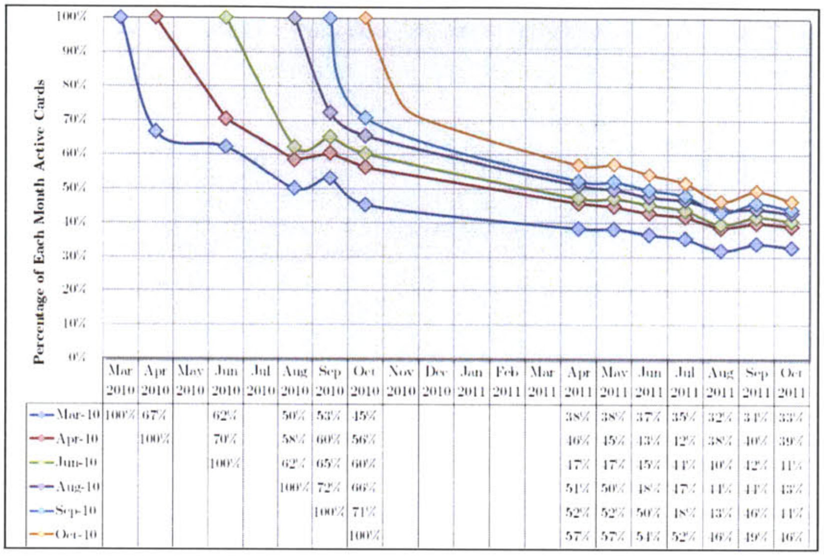


Figure 6-8: Number of Active Oyster Cards over Time

which is a way to measure Oyster Card attrition rates. All curves exhibit a similar trend with a sharp decrease in the first month, while later months present a slower rate.



	Mar 2010	Apr 2010	May 2010	Jun 2010	Jul 2010	Aug 2010	Sep 2010	Oct 2010	Nov 2010	Dec 2010	Jan 2011	Feb 2011	Mar 2011	Apr 2011	May 2011	Jun 2011	Jul 2011	Aug 2011	Sep 2011	Oct 2011
Mar-10	100%	67%		62%		50%	53%	45%						38%	38%	37%	35%	32%	34%	33%
Apr-10		100%		70%		58%	60%	56%						46%	45%	43%	42%	38%	40%	39%
Jun-10				100%		62%	65%	60%						47%	47%	45%	44%	40%	42%	41%
Aug-10						100%	72%	66%						51%	50%	48%	47%	44%	44%	43%
Sep-10							100%	71%						52%	52%	50%	48%	43%	46%	44%
Oct-10								100%						57%	57%	54%	52%	46%	49%	46%

Figure 6-9: Oyster Card Attrition over Time

Two distinct behaviors can be observed in Figure 6-9. First, all curves show dips for August and slight increases for September in both years, which reflects the temporary decline in the number of active Oyster Cards during August. Second, the March 2010 curve is much lower than the rest of the months. This phenomenon could be explained by a change in the number of short time visitors; though there is no clear explanation such a change in March 2010.

Excluding data from March and August, a logarithmic regression analysis was conducted. The dashed line in Figure 6-10 shows the logarithmic curve and the parameters obtained. The data follows the expression

$$P = -0.158 \cdot \ln(m) + 0.863 \tag{6.1}$$

where P is the percentage of Oyster Cards observed in the month of analysis and m is the number of subsequent months. The correlation coefficient is close to 1 ($R^2 = 0.97$), which indicates that the logarithmically decreasing function is the best fit to the Oyster

Card attrition rate data.

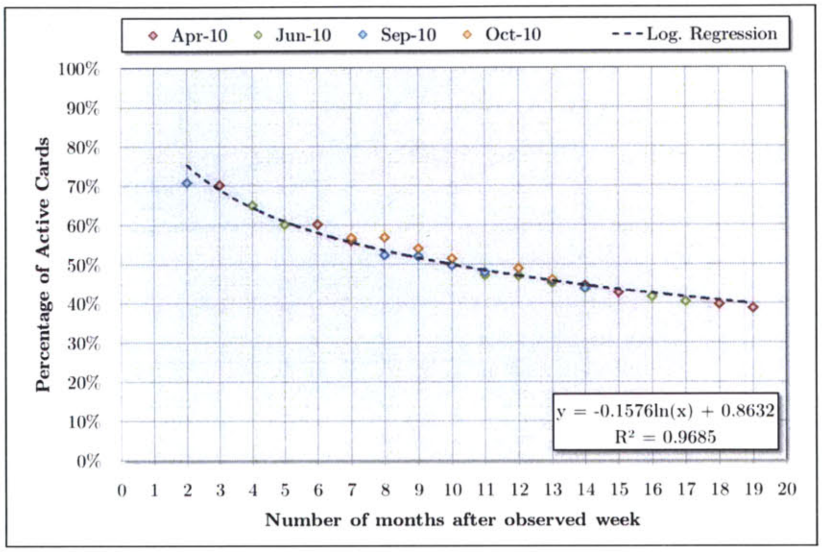


Figure 6-10: Oyster Card Attrition over Time

6.2.2 Churned Oyster Cards Travel Characteristics

Using Oyster Card data from 2011 and 2012, the characteristics of churned Oyster Cards can be obtained from the cluster's travel profiles obtained in Chapter 3. Besides the data described in 3.2, one week of Oyster Card record is available for the months of November 2011, December 2011, and January 2012. By obtaining the number of active cards in each period, an analysis analogous to the one presented in 6.2.1 can be done. Table 6-5 summarizes the periods used for this analysis.

Figure 6-11 shows how the number of active Oyster Cards observed during October 17th to 23th, 2011 decreases over time. The dashed line shows the logarithmic regression obtained from the 2010/2011 data (see Section 6.2.1). As it can be seen, the regression curve is very close to the new data points, which implies a tendency similar to that observed in the 2011/2012 data.

Table 6-5: Weeks of Data Used - 2011/2012

Year	Period	Week of the Year
2011	October 17th to 23th	43rd
	November 14th to 20th	47th
	December 12th to 18th	51st
2012	January 16th to 22nd	3rd
	September 19th to 25	38th
	October 1st to 10th	40th

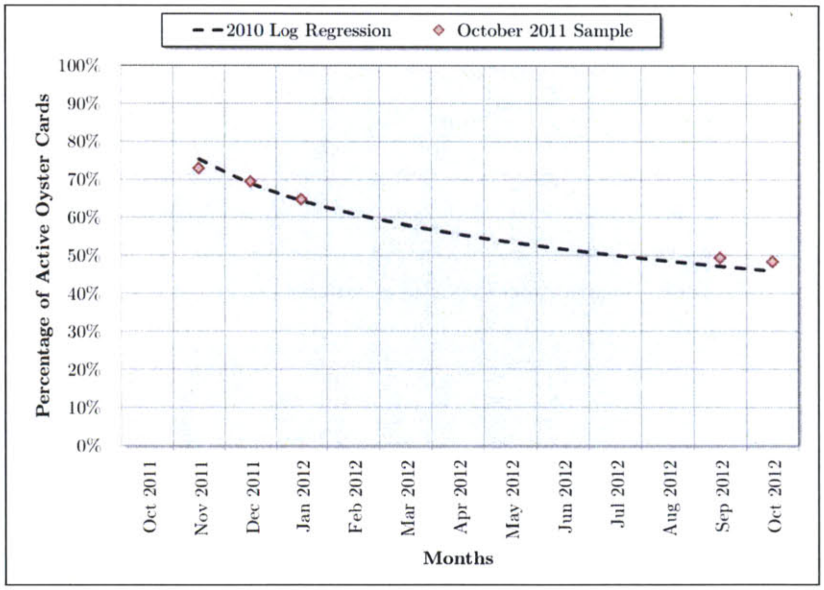


Figure 6-11: October 2011 Attrition over Time

A similar analysis can be done for each of the clusters defined in Chapter 3. Figure 6-12 shows the attrition rates over time as a percentage of each cluster's active cards. The black squares shows the attrition rate for all the cards observed in the October 2011 sample and the dotted line shows the logarithmic regression obtained from the 2010/2011 data. The graph shows that those clusters defined as regular users (1 through 4) are above the total sample attrition rate curve (dotted line), and those defined as occasional users (5 through 8) are below the curve. The attrition curve of regular clusters does not show the sharp dip observed for the first month in the 2010/2011 and in the 2011/2012 analyses. Occasional users though show a much sharper dip, especially for clusters 7 and 8, which indicates that the drop in active cards observed from the first to the second month could be explained by sporadic users of the system, especially visitors. This also indicates that people who travel more and have more regular travel patterns retain their Oyster Cards longer than those who travel only occasionally.

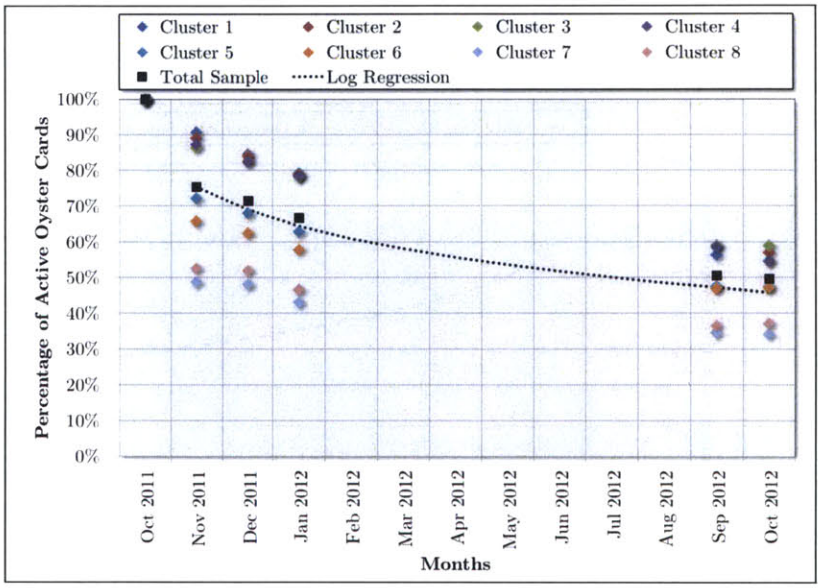


Figure 6-12: Cluster Attrition Rates over Time

6.3 Summary

This chapter presented an overview of the travel characteristics of users that took different Oyster Card ownership decisions over time. Oyster Card user decisions, such as registering their cards and maintaining their card for long periods of time, are related to user travel behavior characteristics that can help to explain why these decisions are made.

The registration status analysis showed that the percentage of registered users varies among clusters. Regular user clusters showed higher percentage of registered cards than occasional user clusters. The characteristics of registered users were compared with the characteristics of the cluster they belong to explore representativeness at the cluster level. Clusters 5 and 6 (non-commuter residents) showed the largest differences among occasional user clusters when comparing the behavior of registered users against the total cluster population, where registered users travel later than the cluster and performed longer activities. In the case of regular user clusters, clusters 2 (student non-exclusive commuter) and 4 (student exclusive commuter) showed the largest differences. In this case, registered users presented sharper morning and afternoon peaks and their activities were longer, which is a behavior observed for work commuters. These differences indicate that registered users have a tendency to behave as regular users (traveling during peak hours and performing longer activity duration), and that registered user behavior is not representative of clusters with high with high percentage of special discount cards or with high variability in travel behavior.

Register users belonging to non-commuter resident clusters (5 and 6) showed travel characteristics more similar to regular or commuter clusters. This result may indicate that some of these clusters members are regular users who traveled few days that specific week and therefore, they were wrongly classified as occasional users. In the case of student commuter clusters, the differences can be explain by the fact that student cards required registration, which in this case makes registered users's characteristics close to actual student behavior. In general, the travel characteristics of registered users were closer to regular users behavior.

The differences in registered users percentages indicate that the application of an expansion process is necessary using registered users as analysis sample. It is important to consider and know the main observed differences when designing focused surveys and developing sampling strategies based on registered users. Regular users are 53% of the 2011 analysis sample, and they make 81% of the week journeys. Registered cards are 47% of the analysis sample and they comprise 51% of the week journeys. From these registered cards, 61% are regular users who make 86% of the registered user journeys. These general statistics provide a first idea to develop an appropriate sampling expansion methodology.

In order to avoid possible biases, adjustments and further analysis is necessary to address registered user lack of representativeness among occasional users clusters and clusters with high percentage of special discount cards. Registered users' travel behavior was representative of the cluster only in the case of regular users, specially exclusive commuters; therefore, registered users can be used for analysis targeting only these type of groups.

The number of active Oyster Cards diminished at a logarithmic rate over time, showing large drops during the first subsequent month analyzed. Regular clusters (1 through 4) show lower attrition rates than occasional users clusters (5 through 8) and the number of active cards that belong to regular user clusters decreases at a slower rate than those that belong to occasional users clusters. For the first month analyzed, occasional user clusters show a larger drop in the number of active cards than the complete population; therefore, their intermittent use of the system over time explain the large drop in active cards observed for the total population after the first month. The churn analysis made for 2010/2011 and 2011/2012 shows that attrition rates are similar for different periods, showing a consistent tendency which may allow better inference for future periods.

Chapter 7

Summary and Conclusions

This chapter concludes the thesis by summarizing the main results, findings, and recommendations. It discusses limitations in the methodology, and identifies possible future research directions. The chapter is organized in three sections: Section 7.1 summarizes the analyses presented in this thesis, focusing on the main findings, Section 7.2 provides some recommendations based on the analysis of the findings, and Section 7.3 describes this thesis limitations and proposes future research in this area.

7.1 Summary and Findings

A classification of London public transport users was developed using the K-medoids clustering algorithm applied to a sample of Oyster Cards. Several travel variables were used to characterize the travel behavior associated with each Oyster Card. Variables related to temporal variability (travel frequency and journey start time), spatial variability (origin frequency and travel distance), activity pattern variability (activity duration), sociodemographic characteristics (Travelcards and special discount cards), and public transport mode choices were estimated using Oyster Card origin-destination travel data inferred using ODX (Gordon, 2012). The clustering analysis was performed using a random sample of 250,000 Oyster Cards observed during the week of October 17th to 23rd, 2011.

Spatial travel patterns, home locations, membership temporal stability, visitor travel characteristics, and Oyster Card management behavior of each cluster were analyzed in detail. Travel patterns were studied by analyzing the most frequently used stations and comparing them with the complete population behavior. A temporal-spatial analysis of the location of regular and occasional users trip entry station and boardings was performed. A home location estimation methodology was developed based on the location of passengers first origin and last destination of the day. The clusters obtained with 2011 data were compared to corresponding clusters obtained using data from a similar week in 2012. The stability of cluster membership for the Oyster Cards observed in both years was tested. The travel characteristics corresponding to a special card available to visitors (Visitor Oyster Card) were analyzed and compared with the cluster characteristics. Finally, the travel behavior of users with different Oyster Card management behavior was studied. Users who registered their cards with TfL and those who hold the same card for long periods of time were compared with respect to the rest of the population. The main results found for all these areas are summarized below.

7.1.1 Cluster Analysis

Eight passenger groups with similar travel characteristics were identified. Four of them represent regular users that travel 4 days a week or more, and four are occasional users traveling less than 4 days per week. The clusters were characterized as every day regular users (traveling all days of the week), all week regular users (traveling 6 days), weekday rail regular users (traveling 5 weekdays and preferring rail), weekday bus regular users (traveling 5 weekdays and preferring bus), weekend occasional users (traveling one weekend day), weekday rail occasional users (traveling one weekday and preferring rail), all week occasional users (traveling 3 days a week, during weekends and/or weekdays), and weekday bus occasional users (traveling 2 weekdays and preferring bus).

Regular user clusters show travel patterns very similar to the whole population, indicating that the behavior of total population is strongly influenced by regular users. Regular users' morning journeys start throughout Greater London, specially during the morning peak (6:30-9:30). They commute to Central London then showing little movement during the off-peak hours and start their last daily journey near Central London during the afternoon peak (16:00-19:00). Occasional users travel behavior showed more variability over time than regular users behavior, making few journeys during the week, specially during off-peak hours.

The analysis of the cluster characteristics suggested that clusters could be further aggregated into four logical groups: exclusive commuters, non-exclusive commuters, leisure travelers, and non-commuter residents. Exclusive commuters showed regular use of the system only during weekdays, behavior typical of workers or students. Non-exclusive commuters show similar behavior to exclusive commuters during the week, but they also make leisure journeys during the weekend. Leisure travelers travel few days during the week, making journeys with leisure purposes. Members in the non-commuter resident group show behavior similar to leisure travelers, but their high number of special discount cards implies that a proportion of this group were residents.

7.1.2 Spatial Travel Patterns

The spatial patterns analysis showed that regular users spatial travel patterns were similar to the whole population. On the other hand, occasional users use National Rail terminal stations such as Victoria, Kings Cross, Paddington, and Euston, more than the rest of the population, indicating that they are mostly visitors for leisure or business. From the leisure travelers group, two clusters were identified as leisure and business visitors based on the high percentage of entries they have at airport stations, specifically at Heathrow and London City Airport.

Occasional user journey start times are normally distributed around the midday period, with high temporal variability during the day, and their journeys are generally made around Central London. Unlike regular users, occasional user residences are located mostly in Central London which is consistent with visitor behavior. Regular user residences were concentrated mainly in the periphery outside Central London.

7.1.3 Temporal Stability

The results of the temporal stability analysis showed that there is a lack of temporal stability from 2011 to 2012, specially at the cluster level. The comparison of cluster characteristics showed that only clusters with the highest and lowest frequency of travel maintained most of their characteristics from 2011 to 2012. Only 28% of the cards observed in both 2011 and 2012 belong to the same cluster in both years. For most clusters, the variables that showed the greatest temporal differences were those related to the type of Oyster Card (Travelcard or special discount card). Further analysis of these variables is required to better understand their role in the cluster process and possible exclude them in future analysis. At the group level (exclusive commuters, non-exclusive commuter, non-commuter resident and leisure traveler) most of the characteristics of these groups were maintained in 2012. The exception again was the percentage of Travelcards and special discount cards.

The differences in temporal stability at different group aggregation levels suggest that eight clusters may represent a highly granular classification with significant overlap among these clusters. This may due to the large number of explanatory variables used for classification. Maintaining only those variables that have the highest explanatory power such as frequency of travel, activity duration, journey start times and mode choice, has the potential to result in more robust cluster identification.

7.1.4 Visitor Travel Patterns

The analysis of Visitor Oyster Cards observed over a week in 2010 showed that visitors travel few days a week (mostly one to three days), during off-peak hours, performing short to medium duration activities at their destinations (up to 2 hours). They perform their activities mostly in Central London, making non-public transport trips between public transport journeys and preferring rail, the most reliable public transport mode, over bus (70%). These characteristics are consistent with the expected visitor travel behavior defined based on the reports of two important visitor surveys conducted in London.

A small percentage of visitors exhibit behavior similar to regular users. The distribution of Visitor Oyster Cards among clusters revealed that visitor have similar characteristics as occasional users: low travel frequency, traveling during off-peak hours, with short-to-medium duration activities, preferring rail over bus, and using international terminal stations, which supports that an important proportion of these groups' members are visitors.

7.1.5 Registration Status

The study of the registration status of Oyster Cards showed that the proportion of registered card users varies among different groups, showing the largest percentages (between 51% and 62%) for regular user clusters. This is expected because card types such as student cards and monthly and annual Travelcards, require registration. Additionally, the registered card users characteristics was compared to the characteristics of the cluster they belong. Occasional user clusters and clusters with high percentage of special discount cards showed the greatest difference in travel behavior. The behavior of registered users belonging to the exclusive commuter group behavior was representative of the cluster.

The differences in the percentages of registered users among clusters implies that a expansion process is needed in order to use registered card users as analysis sample. For those clusters where registered user behavior was not representative of the cluster behavior, additional adjustments are necessary to avoid introducing biases to the analysis. This research represents a first step to understand the representativeness of registered card users and how to manage their differences in behavior with the rest of the population. It is important to consider the differences highlighted here when designing surveys based on registered users.

7.1.6 Oyster Card Attrition Rates

It was observed that active Oyster Cards decrease logarithmically over time. The attrition rates observed for the periods 2010/2011 were similar to the ones observed for 2011/2012, showing a consistent trend. The number of active cards has a big drop after one month. The analysis of attrition rates by cluster showed that this drop is explained by the behavior of occasional users, that contributes the most to this drop. This is an expected occasional user behavior explained by intermittent use of the system over time or by visitors that only use the system once. Regular user clusters showed lower attrition rates than occasional user clusters, and the number of active cards that belong to regular users decreases at a slower rate than those that belong to occasional users.

7.2 Recommendations

This thesis showed that it is possible to analyze the travel characteristics of public transport users and identify passenger groups with similar travel behavior using AFC and AVL data. Computing the travel characteristics of a 250,000 Oyster Card sample using a Python script takes about 3 hours (on a workstation), and the classification can be performed using any powerful statistical software such as R or MatLab.

Given the findings discussed in Section 7.1, some recommendations are provided below to enhance and complement the work presented in this thesis.

- The analysis of the eight clusters identified led to logical aggregations into four larger groups. These groups showed more stability over time, in terms of both group travel characteristics and group membership. Given the sample size analyzed, it is possible that the resulting eight clusters were over-fitted, and four groups seem to be more realistic and appropriate. Therefore, it is recommended to consider these four groups for further studies that require general travel patterns of the population.

- The temporal stability analysis also showed that at the cluster level, the percentage of Travelcards and special discount cards showed the highest differences from 2011 to 2012. It is not clear whether this is caused by Pay as You Go users switching to Travelcards or obtaining access to special discount cards, or because of the sampling strategy. Further examination of the temporal scope of the analysis is recommended, along with an assessment of the role of different Oyster Card products: Are they a cause or a consequence (or both) of travel behavior?.

- The last point also suggest that the number classification variables need further analysis. The sensitivity of the results to the type and number of travel variables used in the classification process need to be further investigated. It is recommended to reduce the number of variables and maintain only those that help to identify more distinctive behavior between groups: frequency of travel, activity duration, journey start times and mode choice.

- The analysis showed that visitors that hold Visitor Oyster cards are a small percentage but have very similar travel behavior with other groups, indicating that a high percentage of visitors do not use the Visitor Oyster Card. Additional validation can be done using survey information focusing on visitors public transport usage. An evaluation of the Visitor Oyster Card product based on the observed visitor travel behavior is recommended, which could help finding the best strategy to increase their usage among different type of visitors or developing an alternative mean visitor identification.

- The analysis of registered users showed that their travel characteristics are more similar to regular users travel patterns. According to these results, if registered card users are used for survey purposes, it is recommended to interpret and use these results according to their registered users share. For instance, the 2011 clustering analysis showed that 63% of registered users are regular users (travel more that 4 days) while 53% of the complete population are regular users, a difference that must be considered when designing a sampling strategy and evaluating the survey responses.

7.3 Limitations and Future Research

The research presented in this thesis has a number of limitations that need to be addressed. Most of them lead to interesting future research opportunities. The limitations found and the possible recommended future research directions are listed below.

- The current analysis for computational reasons, focused only on one week of data both in 2011 and 2012. The chosen weeks were representative of the year with no holidays or special events. However, given the observed variability and the card attrition rates, future analysis may benefit by considering a longer period.

- Further validation of the classification results can be done using information from the London Travel Demand Survey (LTDS). LTDS has information about some interviewees Oyster Card number, which can help identify to which cluster surveyed individuals belong to. LTDS provides other information about users, such as work status, student status, age, income, and other sociodemographic characteristics, which can be used to validate the interpretation of each group.

- The thesis provides a first approach to determine activity types using Oyster Card data through the analysis of activity durations. This approach was rather simple, considering only activities between public transport trips. Several authors have developed different methods to infer activity purposes using smart card data and applied it to other public transport systems (Devillaine et al., 2012; Lee and Hickman, 2012). An improved activity purpose inference methodology can provide useful information for further analysis and complement other studies, for example related to land use.

- The research analyzed visitors based on the travel characteristics from Visitor Oyster cards and groups with similar travel behavior were identified. However, a further step could be taken by refining the identification of visitors that hold retail Oyster Cards. The Oyster Card issuance information could be analyzed among those groups whose members behave more similarly to visitors. This will help to verify if these cards were used for the first time that week, and where they were issued (cards issued at international ports are more likely to be visitors). Analyzing the periods of activity and inactivity of cards will also help in identifying visitors.

- A study of the impact of each cluster or group on the network loads can be performed based on this thesis results. Origin-destination matrices by cluster or group can be estimated and this information can be used to better understand how the behavior of the different groups affects the system.

Appendices

A 2012 Within-Cluster Variation and Davies-Bouldin Index

Figure A-1 shows the values of within-cluster variation and the DB index as a function of the number of clusters K. The within-cluster variation decreases as the number of clusters increases; however, there is a point at which there is relatively little gain from further increase of the number of clusters. As in 2011, the first significant drop of the within-cluster variation occurs for $K = 7$; however, the DB index shows the first significant drop for $K = 8$, which was the number of cluster selected.

Two principal components illustrated in Figure A-2 visually show that most of the clusters are separated enough from each other. As in 20111, using a higher number of K only generates smaller clusters with less distinctive characteristics.

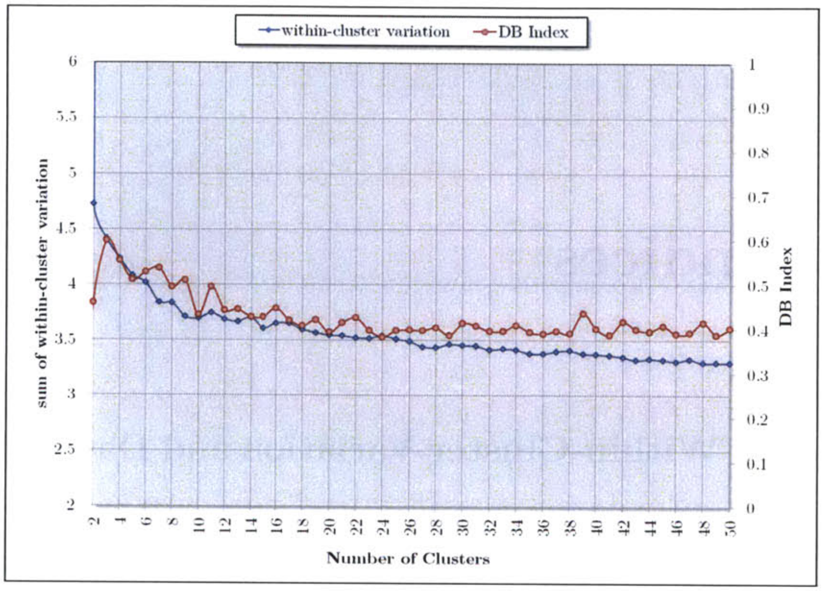


Figure A-1: Within-Cluster Variation and DB Index per Number of Clusters - K-medoids

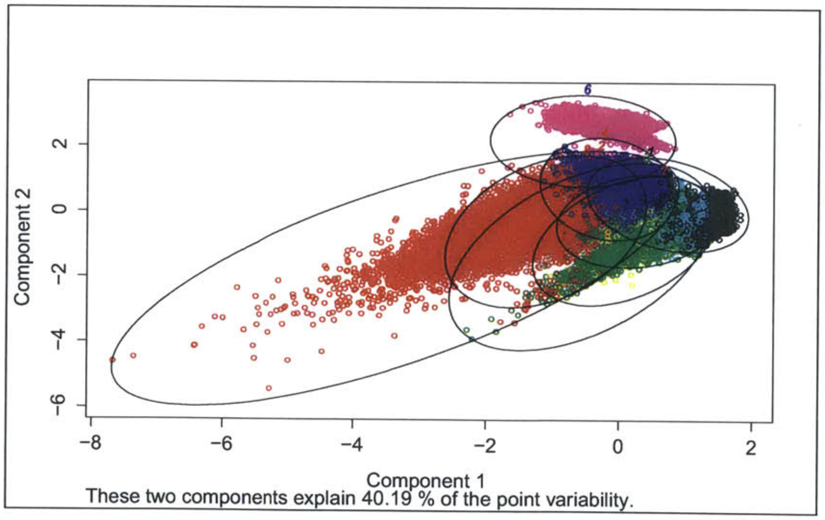


Figure A-2: Principal Components showing K-medoids with $K = 8$

B Registered Users and Total Population Differences

Table B-1: Absolute Centroid Differences

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Days of Use (Days)	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1
Weekdays Main Activity Duration (minutes)		16.4	7.2	38.5	15.7	16.7	28.6	26.1
Weekends Main Activity Duration (minutes)	4.7	7.5		5.3		0.6		
Weekdays Shortest Activity Duration (minutes)		16.3	5.5	37.4	16.7	17.2	31.1	19.0
Weekends Shortest Activity Duration (minutes)	8.3	10.1		4.9		2.5		
Weekdays First Journey Start Hour (minutes)		13.0	6.6	18.4	0.9	5.8	16.6	17.7
Weekends First Journey Start Hour (minutes)	2.7	15.0		21.2		16.9		
Weekdays Last Journey Start Hour (minutes)		20.8	5.0	22.4	12.3	8.1	1.3	25.5
Weekends Last Journey Start Hour (minutes)	4.3	16.0		22.6		10.3		
Percentage of Bus Exclusive Days	2%	3%	0%	6%	0%	1%	1%	2%
Percentage of Rail Exclusive Days	2%	3%	1%	4%	1%	2%	0%	0%
Weekly Maximum Travel Distance (meters)	278.3	107.6	244.8	125.8	64.6	248.4	231.9	383.7
Weekly Minimum Travel Distance (meters)	105.1	252.1	163.0	88.9	62.7	43.3	11.1	125.9
Percentage of Different First Origins. Weekdays		1%	1%	2%	4%	1%	2%	0%
Percentage of Different First Origins. Weekends	0%	1%		2%		4%		
Percentage of Different Last Origins. Weekdays		0%	1%	1%	0%	4%	2%	0%
Percentage of Different Last Origins. Weekends	1%	0%		2%		2%		
Percentage of Travelcards	4%	7%	3%	1%	7%	2%	2%	13%
Percentage of No Special Discount Cards	1%	6%	3%	6%	6%	4%	6%	9%

Table B-2: Absolute Standard Deviation Differences

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Days of Use (Days)	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0
Weekdays Main Activity Duration (minutes)		4.0	7.8	19.2	3.3	10.9	18.9	12.5
Weekends Main Activity Duration (minutes)	2.0	1.1		5.8		4.6		
Weekdays Shortest Activity Duration (minutes)		6.1	6.6	12.2	3.4	4.2	7.1	14.2
Weekends Shortest Activity Duration (minutes)	0.9	4.5		3.3		3.7		
Weekdays First Journey Start Hour (minutes)		2.0	9.8	3.7	2.5	8.8	5.9	11.6
Weekends First Journey Start Hour (minutes)	1.7	1.5		1.1		0.8		
Weekdays Last Journey Start Hour (minutes)		6.6	6.4	13.0	2.5	10.1	10.5	0.5
Weekends Last Journey Start Hour (minutes)	2.1	1.0		3.9		4.1		
Percentage of Bus Exclusive Days	0%	0%	0%	0%	0%	0%	0%	2%
Percentage of Rail Exclusive Days	0%	1%	1%	1%	1%	1%	1%	1%
Weekly Maximum Travel Distance (meters)	121.6	11.3	69.9	27.0	107.7	20.9	65.4	158.9
Weekly Minimum Travel Distance (meters)	256.3	139.6	149.8	42.8	95.6	17.0	20.2	5.3
Percentage of Different First Origins, Weekdays		1%	0%	1%	0%	0%	1%	0%
Percentage of Different First Origins, Weekends	0%	1%		1%		0%		
Percentage of Different Last Origins, Weekdays		0%	0%	1%	1%	0%	0%	0%
Percentage of Different Last Origins, Weekends	1%	1%		2%		0%		
Percentage of Travelcards	4%	5%	1%	0%	10%	2%	0%	6%
Percentage of No Special Discount Cards	1%	5%	7%	4%	10%	4%	1%	4%

Bibliography

Agard, B., Morency, C., and Trépanier, M. (2006). Mining public transport user behaviour from smart card data. In *Proceedings of the 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM).*

Airports Council International (2013). 'International Passenger Rankings'. Retrieved May 9, 2013, from `http://bit.ly/10JmkdR`.

Alpaydin, E. (2004). *Introduction to machine learning.* MIT press.

Barry, J., Newhouser, R., Rahbee, A., and Sayeda, S. (2002). Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board*, 1817(-1):183–187.

Ben-Akiva, M. and Lerman, S. (1985). *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press.

Billingsley, P. (1995). *Probability and measure.* Jown Wiley & Sons.

Blythe, P. (2004). Improving public transport ticketing through smart cards. In *Proceedings of the Institution of Civil Engineers, Municipal Engineer*, volume 157, pages 47–54.

Caliñski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Camus, R., Longo, G., and Macorini, C. (2005). Estimation of transit reliability Level-of-Service based on Automatic Vehicle Location data. *Transportation Research Record: Journal of the Transportation Research Board*, 1927:277–286.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3):315 – 332.

Chakirov, A. and Erath, A. (2011). Use of public transport smart card fare payment data for travel behaviour analysis in Singapore. In *Proceedings of the 16th International Conference of Hong Kong Society for Transportation Studies, Hong Kong.*

Chakroborty, P. and Kikuchi, S. (2004). Using bus travel time data to estimate travel times on urban corridors. *Transportation Research Record: Journal of the Transportation Research Board*, 1870(-1):18–25.

Coordinacion Transantiago (2010). 'Usos de Tarjetas Bip Informe Generacion de Tablas Consulta'. Retrieved May 9, 2013, from `http://bit.ly/10zhlQf`.

Cortés, C., Gibson, J., Gschwender, A., Munizaga, M., and Zuñiga, M. (2011). Commercial bus speed diagnosis based on GPS-monitored data. *Transportation Research Part C: Emerging Technologies*, 19(4):695–707.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1(2):224–227.

Devillaine, F., Munizaga, M., and Trèpanier, M. (2012). Detection of public transport user activities through the analysis of smartcard data. In *Proceedings of the 91st Annual Meeting of the Transportation Research Board, Washington, DC.*

ElGeneidy, A., Horning, J., and Krizek, K. (2011). Analyzing transit service reliability using detailed data from automatic vehicular locator systems. *Journal of Advanced Transportation*, 45:66–79.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2001). *Cluster analysis.* Edward Arnold, London.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer Series in Statistics.

Furth, P., Hemily, B., Muller, T., and Strathman, J. (2006). *TCRP Report 113: Using Archived AVL-APC Data to Improve Transit Performance and Management.* Transportation Research Board of the National Academies.

Gordon, J. (2012). Intermodal passenger flows on London's public transport network. Master's thesis, Massachusetts Institute of Technology.

Greater London Authority (2013). 'Mayor & Assembly'. Retrieved May 9, 2013, from `http://bit.ly/12hBXNT`.

Guo, Y., Hastie, T., and Tibshirani, R. (2005). Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 1(1):1–18.

Hanson, S. and Huff, J. (1986). Classification issues in the analysis of complex travel behavior. *Transportation*, 13:271–293.

Horbury, A. (1999). Using non-real-time Automatic Vehicle Location data to improve bus services. *Transportation Research Part B: Methodological*, 33(8):559 – 579.

Jain, A. and Dubes, R. (1988). *Algorithms for clustering data.* Prentice Hall advanced reference series. Prentice Hall.

Jain, A., Duin, R., and Mao, J. (2000). Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Jones, P. and Clarke, M. (1988). The significance and measurement of variability in travel behaviour. *Transportation*, 15(1):65–87.

Lathia, N. and Capra, L. (2011). How smart is your smartcard?: measuring travel behaviours, perceptions, and incentives. In *Proceedings of the 13th international conference on Ubiquitous computing, Beijing.*

Lee, S. and Hickman, M. (2012). Trip purpose inference using automated fare collection data. The 4th Transportation Research Board Conference on Innovations in Travel Modeling, Tampa, FL.

Liu, L., Hou, A., Biderman, A., Ratti, C., and Chen, J. (2009). Understanding individual and collective mobility patterns from smart card records: a case study in Shenzhen. In *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO.*

London & Partners (2011). 'Key Visitor Statistics'. Retrieved May 9, 2013, from `http://bit.ly/17fqAZs`.

London Development Agency (2009). London Visitor Survey Report. Retrieved May 9, 2013, from `http://bit.ly/17b92xJ`.

Ma, J. and Goulias, K. (1997). A dynamic analysis of person and household activity and travel patterns using data from the first two waves in the Puget Sound Transportation Panel. *Transportation*, 24(3):309–331.

Maechler, M., Struif, A., Hubert, M., and Hornik, K. (2013). Package 'cluster'. Retrieved May 9, 2013, from `http://bit.ly/16kCBhM`.

Morency, C., Trépanier, M., and Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203.

Munizaga, M. and Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, 24(0):9–18.

Nishiuchi, H., King, J., and Todoroki, T. (2013). Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data. *International Journal of Intelligent Transportation Systems Research*, 11:1–10.

Office for National Statistics (2013). 'Overseas Travel and Tourism, Q3 2012'. Retrieved May 9, 2013, from `http://bit.ly/YIE04b`.

Pangilinan, C., Wilson, N. H. M., and Moore, A. (2008). Bus supervision deployment strategies and use of real-time Automatic Vehicle Location for improved bus service reliability. *Transportation Research Record: Journal of the Transportation Research Board*, 2063:28–33.

Pas, E. (1983). A flexible and integrated methodology for analytical classification of daily travel-activity behavior. *Transportation Science*, 17(4):405–429.

Pelletier, M., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568.

Pendyala, R., Parashar, A., and Muthyalagari, G. (2000). Measuring day-to-day variability in travel characteristics using GPS data. In *Proceedings of the 79th Annual Meeting of the Transportation Research Board, Washington, DC.*

Rousseeuw, P. J. and Kaufman, L. (1990). *Finding groups in data: An introduction to cluster analysis.* John, John Wiley & Sons.

Rousseeuw, L. and Kaufman, P. (1987). Clustering by means of medoids. *Statistical data analysis based on the L1-norm and related methods*, 405.

Sánchez-Martinez, G. (2013). Running time variability and resource allocation: A data-driven analysis of high-frequency bus operations. Master's thesis, Massachusetts Institute of Technology.

Schlich, R. and Axhausen, K. W. (2003). Habitual Travel Behaviour: Evidence from a six-week travel diary. *Transportation*, 30:13–36.

Transport for London (2011). 'Mayor unveils design of the royal wedding Oyster card'. Retrieved May 9, 2013, from `http://bit.ly/eLJAhe`.

Transport for London (2012a). 'Getting Around with Oyster'. Retrieved May 9, 2013, from `http://bit.ly/12APsHL`.

Transport for London (2012b). 'London Buses factsheet'. Retrieved May 9, 2013, from `http://bit.ly/YGrHyy`.

Transport for London (2012c). 'London Overground factsheet'. Retrieved May 9, 2013, from `http://bit.ly/ZLuC6O`.

Transport for London (2012d). 'London River Services factsheet'. Retrieved May 9, 2013, from `http://bit.ly/Y8gnHB`.

Transport for London (2012e). 'London Tramlink factsheet'. Retrieved May 9, 2013, from `http://bit.ly/10lBSon`.

Transport for London (2012f). 'London Underground factsheet'. Retrieved May 9, 2013, from `http://bit.ly/QQpQmJ`.

Transport for London (2012g). 'Oyster Card factsheet'. Retrieved May 9, 2013, from http://bit.ly/131EVnl.

Transport for London (2012h). 'Transport for London factsheet'. Retrieved May 9, 2013, from http://bit.ly/Zvn3pv.

Transport for London (2012i). 'Using Oyster Online'. Retrieved May 9, 2013, from http://bit.ly/Jh0mKW.

Transport for London (2013a). 'About Oyster'. Retrieved May 9, 2013, from http://bit.ly/ab2Ibl.

Transport for London (2013b). 'TfL Interchange'. Retrieved May 9, 2013, from http://bit.ly/ZviuYg.

Transport for London (2013c). 'Tickets'. Retrieved May 9, 2013, from http://bit.ly/7sGccI.

Transport for London (2013d). 'Visitor tickets'. Retrieved May 9, 2013, from http://bit.ly/a5HMXJ.

Trépanier, M., Tranchant, N., and Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1):1–14.

UK Office for National Statistics (2012). 'London 2012 Games attract over half a million overseas visitors in July and August'. Retrieved May 9, 2013, from http://bit.ly/18x4KRY.

United Nations (2008). International Recommendations for Tourism Statistics. *United Nations Publications*, (83).

Wei, C. P., Lee, Y. H., and Hsu, C. M. (2003). Empirical comparison of fast partitioning-based clustering algorithms for large data sets. *Expert Systems with applications*, 24(4):351–363.

Wilson, N. H. M., Zhao, J., and Rahbee, A. (2009). The potential impact of automated data collection systems on urban public transport planning. In *Schedule-Based Modeling* *of Transportation Networks*, volume 46 of *Operations Research/Computer Science Interfaces Series*, pages 1–25. Springer US.

Xu, L. and Jordan I, M. (1996). On convergence properties of the EM algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151.

Xu, R. and Wunsch II, D. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678.

Zhao, J., Rahbee, A., and Wilson, N. H. M. (2007). Estimating a rail passenger trip Origin-Destination matrix using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5):376–387.