Mobility Insights Through Consumer Data: A case study of concessionary bus travel in the West Midlands

Ffion Carney

Thesis submitted in conformity with the requirements of

Doctor of Philosophy (Ph.D.)

Department of Geography

UCL

August 2020

Declaration

I, Ffion Carney, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: 

Acknowledgments

I would first and foremost like to thank my supervisors, Professor Paul Longley and Dr Jens Kandt, for their guidance, expertise and feedback throughout the completion of this thesis. Without their support this thesis would not have been possible. I'd like to thank everyone at TfWM who has supported this research and taken the time to collaborate and share their knowledge. I would also like to acknowledge the ESRC for providing the funding for this work. Special thanks to everyone at the CDRC and the Chorley for making it such a fun and supportive place to work. Finally, I'd like to thank my friends and family for their support, encouragement and patience over the past 4 years.

Abstract

Current transport facilities are often built around efficiency and meeting the needs of the commuting population. These can therefore struggle to provide services suited to some of the most vulnerable members of society. In order to achieve an inclusive transport system, it is vital that transport authorities have access to detailed insights into the mobility needs and demands of different groups of the population. Increasingly, these transport authorities are making use of smart technologies and the resulting data to gain greater insight into transport users, and in turn inform decision making and policy planning.

These smart technologies include automated fare collection (AFC) systems, which produce large volumes of detailed transport and mobility data from smart card transactions. To a lesser extent, retail datasets, such as loyalty card transaction data, have also been utilised. The spatiotemporal components of these data can provide valuable insight into the activity patterns of cardholders that may not be captured in traditional transport data.

This thesis presents an exploration of these two forms of consumer data, with a focus on the older population in the West Midlands. Firstly, this thesis demonstrates how smart card data can be processed and analysed to provide detailed insights into the mobility patterns of concessionary bus users and how these relate to long-term changes in bus patronage recorded in the study area. Secondly, the extent to which loyalty card transaction data can be employed to understand retail behaviours and activity patterns is explored, with a focus on how these insights can be used to supplement and enhance the understanding of mobility gained from the smart card data. Finally, these insights are discussed in terms of the capacity of the current transport network to meet the mobility needs of the older population and the potential of consumer data for future transport-related research.

Impact Statement

This thesis focuses on utilising consumer data to understand the mobility patterns of the older population in the West Midlands, and to assess the effectiveness of the current bus network in meeting these mobility needs. The consumer data used in this study were smart card transactions on the bus network provided by Transport for West Midlands (TfWM) and loyalty card data provided by a large high street retailer (HSR).

Current transportation research is largely focused on the working age population; with much attention given to understanding commuting patterns and demands. The classification of bus activity and the trip purpose inferences undertaken in this research evidenced that that older population should not be treated as a homogenous group in terms of mobility patterns and transport demand in future research. With academic research utilising these novel consumer data still in its infancy, the processes and methods developed for the analysis of 'big' consumer data could make these data more accessible for use within academia. At the time of writing, these outputs have primarily been disseminated through academic conference papers and presentations. Future journal articles are planned to publicise the methods developed for the analysis of these novel consumer data.

A large focus at TfWM is the shift towards smart technologies and transport solutions. To maximise the benefits of these technologies, it is imperative that the public transport demands of different population groups are understood. The outputs of this thesis, in particular those relating to trip purpose and the impact of public transport on activity patterns, could be implemented by TfWM to understand demand for current transport services. With COVID-19 likely to bring about a modal shift towards private transportation, walking and cycling, there is potential for a large proportion of the older population to be left behind. The method developed for the inference of trip purpose could be applied to more recent data; to understand changes in the use of bus services brought about by more people working from home and a reduction in non-essential travel. Insights from these inferences could then be used to better plan services around the needs of bus users, for example increased services to areas that attract a large number of essential trips.

Much debate has surrounded the English National Concessionary Travel Scheme (ENCTS), with questions over its value for money and capacity to meet the needs of the most vulnerable populations. Results showed that bus usage by residents of central urban areas was decreasing, whilst those living in suburban and rural areas displayed increasing usage of bus services. These insights suggested that investments in rural public transport services may bring about more benefits that the current ENCTS scheme. The outputs of this thesis will be disseminated to

TfWM to aid public transport service and policy planning and to add to the knowledge base around the use of the ENCTS. This could have a knock-on effect on the older population's quality of life; improving their inclusion in society and ability to access key services and facilities.

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

Mobility has long been linked to quality of life, with numerous studies showing that limited mobility can have a direct impact on health and wellbeing, most notably when this results in a lack of physical activity or restricts social interactions (Boniface et al. 2015). Quality of life is defined as the general wellbeing of individuals and societies. It observes whether basic needs are being met, but also expands beyond this to include social and emotional needs such as the ability of individuals to perform activities that are non-essential but pleasurable (Musselwhite and Haddad 2010). Restricted mobility can have a direct effect on an individual's ability to partake in society and perform both essential and non-essential activities and can therefore negatively impact quality of life.

Access to transportation, both public and private, is vital for individuals to maintain high levels of mobility. Current transport facilities are often built around efficiency and meeting the needs of the commuting population and can therefore struggle to provide services suited to some of the most vulnerable members of society, for example those that are older or disabled (Battellino 2007). The potential negative impacts of a lack of public transport access have meant that achieving an inclusive transport system, which is accessible for the entire population, is a fundamental element in tackling the many wider social issues that affect some of the most vulnerable groups of the population. This has become a major policy goal for many local transport authorities and operators throughout the UK.

In order to achieve an inclusive transport system, it is vital that transport authorities have access to detailed insights into the mobility needs and demands of different groups of the population. Increasingly, these transport authorities are making use of smart technologies and the resulting data to gain greater insight into transport users, and in turn inform decision making and policy planning. These smart technologies include automated fare collection (AFC) systems, which produce large volumes of detailed data that have the potential to provide insights into the efficiency of the transport network and the daily mobility patterns of travellers.

Other novel datasets, such as GPS data recorded by taxi drivers (Liu et al. 2010), mobile phone data (Jacobs-Crisioni et al. 2014; Zahedi and Shafahi 2018), and social media data (Huang and Wong 2016) have been employed to understand the spatiotemporal mobility patterns of individuals. To a lesser extent, large consumer datasets, such as retail transaction data, have been implemented for spatial analyses (Lloyd and Cheshire 2019; Birkin 2019). The majority of research utilising these data have focused on classifying different types of cardholders for market research purposes (Allaway et al. 2006) and understanding the impact of promotions on

retail behaviour (Felgate et al. 2012). However, the spatiotemporal element of these data mean they have the potential to provide valuable insight into the mobility and activity patterns of individuals.

This thesis presents an exploration of two forms of consumer data; smart card data provided by a local transport authority, and loyalty card data provided by a large high street retailer (HSR). Access to these data was facilitated by the Consumer Data Research Centre (CDRC), a big data initiative funded by the Economic and Social Research Council (ESRC). Such data are rarely available for academic research on such a scale; thus, this offered a unique opportunity to implement these data for mobility research.

1.1 Case study

In order to improve the accessibility of the transport system for vulnerable groups of the population, the Transport Act 2000 introduced the English National Concessionary Travel Scheme (ENCTS). This policy stated that those that had reached pensionable age and those that suffered from a disability were entitled to half-price bus travel within their local area. This was later extended, in April 2008, to allow for free bus travel outside of peak hours (Mackett 2014). In the UK, there are around 10.6 million disabled people and 11.9 million people of retirement age that are eligible for this scheme, corresponding to around 30% of the population (Mackett 2013). In terms of uptake, the scheme has been successful since its introduction with around 78% of older and disabled people holding an ENCTS pass (Mackett 2013). Despite this apparent success, there is some evidence that the ENCTS is not meeting the mobility needs of concessionary travellers, with bus boardings declining in all areas of the UK (Urban Transport Group 2019).

This thesis solely focuses on the travel behaviours and activity patterns of the older population that are eligible for the ENCTS, i.e. those that have reached pensionable age, in the West Midlands. The West Midlands is a county located in western central England and consists of the metropolitan boroughs of Birmingham, Dudley, Sandwell, Coventry, Walsall, Solihull and Wolverhampton. This area has a population of nearly 3 million people, of which over half a million are eligible for the ENCTS (ONS 2018a).

Public transport services in the West Midlands are the responsibility of Transport for West Midlands (TfWM), the transport arm of the West Midlands Combined Authority (WMCA). Although TfWM do not operate bus, tram or train services, they are responsible for maintaining bus stations and shelters, as well as working in partnership with transport operators to improve public transport services and deliver major public transport schemes throughout the county,

including the ENCTS. Despite the wide use of the ENCTS in the West Midlands, TfWM has observed a steady decline in bus patronage over the past several years, with concessionary boardings seen to decrease by over 20% between 2009 and 2016 (TfWM 2016). This decline in bus patronage raises the question as to whether the ENCTS has been successful in meeting the mobility needs of the older population in the West Midlands, or whether investments elsewhere may bring more success. Understanding these trends and the mobility patterns of the older population is therefore vital to determine the reasons for this recorded decline and assess the suitability of the current transport network.

1.2 Aims

Through the use of two sources of consumer data this thesis aims to assess the extent to which these novel forms of data can be used alongside other contextual, demographic and socioeconomic data to generate insights into the mobility and activity patterns of the older population in the West Midlands, as well as the ability of the current transport system to meet their mobility needs. The three primary aims of this thesis are:

1. To demonstrate how data generated by AFC systems can be processed and analysed to provide detailed insights into bus usage patterns.
2. To identify changes in the bus usage patterns of the older concessionary population, the reasons for these changes, and the potential impacts on the mobility of cardholders.
3. To understand the extent to which loyalty card transaction data can be employed to understand the activity patterns of the older population, and how these can be used to supplement and enhance current transport data and insights.

The over-arching theme of this thesis is therefore an exploratory study into the extent to which big, consumer datasets can be implemented to generate insights into mobility. As the number of local authorities and transport operators that are utilising smart card technologies and other sources of consumer data are increasing, this research could be used to inform other local authorities and transport operators that wish to employ these data to tackle a number of social issues.

1.3 Thesis structure

Chapter 2 – Mobility, social exclusion and the applications of consumer data

This chapter provides an overview of the concepts and literature relevant to the analyses and research conducted throughout this thesis. Firstly, this involves a detailed overview of the previous research that has been conducted into the mobility and activity patterns of the older population, and the links to social exclusion and wellbeing. The second section of this chapter focuses on the use of consumer data; exploring how these data have been previously utilised in spatial research and the potential for these data to be used in transportation and mobility research.

Chapter 3 – Contextual analysis of the study area and population

This chapter provides an in-depth analysis of the study area and the population. The general demographic and population trends are explored, in addition to the current transport network and urban landscape. The aim of this chapter is to outline both the study area and the study population to provide context for subsequent analyses.

Chapter 4 – Smart card data: Overview and preliminary analysis

This chapter presents a detailed overview of the smart card data utilised in this study, including the representativeness of these data and the cleaning and processing methods conducted to ready the data for subsequent analyses. Finally, the general trends that could observed in the data are presented and adjusted to account for changes in the concessionary population; primarily adjusting for a change in the age of eligibility for the ENCTS.

Chapter 5 – A classification of concessionary bus users

This chapter primarily focuses on classifying smart card holders based on their longitudinal activity on the bus network between 2010 and 2016. The aim of this was to identify groups of the population that have exhibited changing bus usage over the study period. This chapter then presents an analysis of the demographics and the spatial activity patterns of the resulting groups; identifying differences in the characteristics of different types of bus users and discussing the results in terms of the observed decline in concessionary bus boardings.

Chapter 6 – Identifying activities and inferring trip purpose

This chapter focuses on the activity of the older population throughout the study area and takes a methodological approach; implementing a machine learning method to infer the purpose of trips recorded in the smart card data. The characteristics of trips made for each purpose are then analysed and the primary activity areas of each trip purpose identified, with comparisons made with the known land uses in the study area. This chapter therefore provides insight into how the older population make use of the bus network, how this has changed over the study period, and which areas attract bus users for which purposes.

Chapter 7 – Loyalty card data: Preliminary analysis and retail behaviours

This chapter presents preliminary analyses of the loyalty card data; exploring the representativeness of the data and the potential overlap with the smart card data. The retail behaviours of the older population are compared against those of working age; including spatial and temporal store visiting patterns and purchasing behaviours. The aims of these analyses were to benchmark the general activity and retail behaviour of both groups of the population for contextualisation of subsequent results, and to provide a comparison of their spatiotemporal activity patterns.

Chapter 8 – Activity patterns and the relationship with public transportation

This chapter builds on both the smart card data and the loyalty card data insights; exploring how retail transactions can be used as a proxy for activity, and how this relates to public transport provision and accessibility. The relationships between retail behaviour, demographic and socioeconomic variables, and bus service provision and accessibility are explored. This provides a 'real-life' case study of the impact of public transport on activity, with the results discussed in terms of the importance of public transport and the efficiency of the current transport system. Additionally, the extent to which non-transport related consumer datasets can provide insight into spatiotemporal activity patterns is discussed.

Chapter 9 – Discussion, applications and research prospects

This final chapter consolidates the principal findings; discussing the key methodological and knowledge contributions of the thesis. Explicit findings that are relevant to the objectives of both the local transport provider and the high street retailer are also outlined, in addition to

outcomes that may benefit the wider academic community. Reflections are made on the methods and approaches used in this thesis, with suggestions made for future research and developments.

1.4 Notes on software and code

The majority of analyses presented in this thesis were conducted in the R Software for Statistical Computing (R Core Team 2018) environment. This is an open-source program, freely downloadable from *www.r-project.org*. Associated codes are available upon request. In addition, ESRI ArcGIS was utilised and the majority of data storage and handling operations were conducted using PostgreSQL, an open-source relational database management system.

Chapter 2 – Mobility, social exclusion and the applications of consumer data

2.1 Introduction

The primary aims of this thesis are to further the understanding of the mobility and activity patterns of the older population in the West Midlands and to explore the utility of consumer data for supplementing traditional transport data for mobility insights. To achieve this, this research focuses on a number of key areas: the concept of mobility and its relationship with wellbeing and social inclusion, the mobility and activity patterns of the older population and how these differ from those of working age, and the extent to which consumer data can be used to gain insights into these patterns. The aims of this chapter are two-fold; firstly to provide an overview of the key literature related to each of these concepts, and secondly to demonstrate that there is a need for further research into these issues and the potential that consumer data have to offer.

Section 2.2 focuses on the concept of mobility and its relationship with transport disadvantage, social exclusion and wellbeing. Additionally, the mobility patterns and travel behaviours of the older population are discussed and compared to those of working age. Section 2.3 then discusses consumer, or 'big', data and their utility for transportation research. The potential of retail transaction data for extracting mobility insights and supplementing transport-related data is then discussed.

2.2 Mobility, transport and social exclusion

2.2.1 The concept of mobility

Throughout this thesis, the concept of 'mobility' is widely discussed. In a general sense, mobility refers to the ability to move freely. Urry (2007) identified five interdependent mobilities:

1. The *corporeal travel* of people for work, leisure, family life and pleasure.
2. The physical movement of *objects*, such as food and water.
3. The *imaginative* travel effected through the images of places and peoples appearing on print and visual media.
4. *Virtual* travel, often in real time, transcending geographical and social distance.

5. *Communicative* travel through person-to-person messages via postcards, texts, emails etc.

Although all these forms of mobility play a part in how mobile one can be considered, this thesis focuses on the first mobility type: 'corporeal travel'. Mobility is therefore defined as the ability to get from one place to another, focusing specifically on movement throughout geographical space. Mobility and transportation, whilst related, do not refer to the same concept. For one to be considered mobile, one must have the ability to access the facilities and places necessary to fulfil one's needs. For example, an individual that lives in close proximity to a bus stop may have access to transportation services, however, if these services are irregular or do not service the areas that one wishes to travel to, this individual may not be considered to be highly mobile. Understanding this difference, and the relationship between mobility and transport access, is a key concept of this thesis.

The following sections explore the concept of mobility further, with a particular focus on the relationship with wellbeing and quality of life; outlining the key studies that have been conducted and discussing how these relate to the aims of this thesis.

2.2.2 Mobility, transport-related social exclusion and wellbeing

As discussed in Chapter 1, it has long been accepted that mobility is closely linked to one's independence, wellbeing and quality of life (Musselwhite 2018). As numerous studies have shown, limited opportunity to travel can have direct implications for health and wellbeing, in particular when this results in a lack of physical activity, unmet social needs, and inadequate access to key service, such as retail opportunities and health care facilities (Boniface et al. 2015; Ziegler and Schwanen 2011).

Older residents may be particularly vulnerable to restricted mobility as life-course events, such as the loss of the ability to drive or physical restrictions, often reduce access to private transport modes (Musselwhite et al. 2015; Mackett 2013). Research by Age UK (2012), found that over 600,000 older people in the UK do not leave their homes more than once a week, with 1 in every 20 members of the older population facing isolation and loneliness. If public transport services are not providing an adequate service for older residents, the risk of social exclusion and isolation is increased. Gaining an understanding of the relationship between mobility, social exclusion and wellbeing, as well as how this relates to public transport services, is therefore vital for transport authorities aiming for an inclusive transport system.

2.2.2.1 Transport-related social exclusion and disadvantage

Social exclusion refers to a wide range of processes related to poverty, deprivation and hardship (Peace 2001). There has been considerable research conducted into the concept of social exclusion, primarily from a sociological perspective. However, until the late 1990s and early 2000s little attention was paid to the mobility aspect of social exclusion and how this may relate to transport. Social issues in transport have primarily been viewed as secondary concerns, especially when compared to economic and environmental impacts. It has been stated that it is "clear that the profession is better equipped to assess economic effects than social effects" (Forkenbrock et al. 2001: 81), and that the variability and complexity of these social issues has meant that their assessment within transport is a relatively inexact science (Sinha and Labi 2007).

In response to the previous lack of research into transport-related social exclusion, The Social Exclusion Unit (SEU) released a report titled 'Making the Connections: Final report on transport and social exclusion', which was widely viewed as one of the first large-scale studies to focus on the relationship between transport disadvantage and social exclusion (SEU 2003). This report is still used as a reference for local authorities and transport operators, and was seen by many as a catalyst for future research (Cass et al. 2005; Lucas 2019).

Since this SEU report, the concept of social exclusion has evolved with a greater focus on disadvantage and inequalities. The current government has pushed geographic inequalities to the top of the policy agenda, with a focus on boosting economic performance outside of London and the South East and reviving the UK's 'left behind' towns and cities (Davenport and Zaranko 2020). Part of this 'levelling up' agenda involves investment in both regional and local transport. This aims to decrease transport inequalities, and in turn, transport disadvantage (Tomaney and Pike 2020). The following sections outline some of the causes of transport disadvantage.

2.2.2.1.1 Increasing private mobility

Although high mobility has been associated with a good quality of life and increased wellbeing (Musselwhite and Haddad 2010), policies that aim to increase mobility through access to private transportation may have the reverse effect on those that are already transport disadvantaged. This idea relates to the relational nature of transport-related social exclusion, with increasing mobility in some sectors of the population leading to decreased mobility in others. The primary link between these two concepts is the idea that as private transportation increases, local social

ties are being undermined as there is a greater geographical spread in social networks (Gray et al. 2006; Pooley 2016). Although this may be advantageous for highly mobile sectors of the population, for those that are less mobile this may result in a weakening of local social ties and increasing community severance, therefore acting to increase the social exclusion of vulnerable populations. This idea is summarised in a report by Barry (1998: 21) in which it is stated that "the private car is the enemy of social solidarity in as much as public transport is its friend".

Alongside the weakening of local social networks and ties, high levels of private mobility have also resulted in shops and services being increasingly dispersed and located in less accessible areas. This has resulted in those with a lack of access to transport services being unable to reach key facilities, with more dispersed patterns of development also resulting in a decrease in the provision of local facilities and walkable neighbourhood environments (Preston and Raje 2007; Ma et al. 2018). This idea relates to the concept of 'poverty of access', in which small groups that are reliant on public transport are no longer able to reach key facilities and services (Gray et al. 2006). The idea of 'poverty of access' is particularly relevant to rural communities, in which car owners are opting to seek goods and services from higher order retail centres which may not be accessible by public transport. This can result in the increased social exclusion of those that rely on public transport. This is an important concept to be considered when assessing transport policies and the effect that these may have on the most at-risk groups of the population.

2.2.2.1.2 Accessibility of public transport services

In addition to the issues related to an increase in private mobility, the accessibility of public transport services can also greatly influence the social exclusion of individuals and populations. This concept relates to the idea that social exclusion is not the result of a lack of opportunities, but rather a lack of access to opportunities. This idea was first conceptualised by Sen (1981) in relation to poverty and famines, in which it was argued that famines are not caused by a lack of food but instead a lack of access to food. This concept can be applied to transport-related social exclusion, with the idea that transport disadvantage and social exclusion are not just a result of a lack of public transport services, but are also due to the barriers and difficulties that individuals encounter when attempting to access these services. The SEU report emphasised the fact that it is the difficulties that individuals encounter when attempting to access public transport services that are often the most significant barriers to social inclusion, rather than a general lack of services.

Barriers to the public transport network can include the availability, accessibility and affordability of public transport services (SEU 2003). These barriers can reinforce low levels of

accessibility and can disproportionately affect vulnerable and disadvantaged individuals in the population, such as those that are older or disabled (Kenyon et al. 2002), or those that reside in peripheral regions (Binder and Matern 2019). They can also have a significant impact on access to employment, education, healthcare, shops and services, particularly among disadvantaged groups, which can contribute to social exclusion (Kilby and Smith 2012). Improving the accessibility of public transport services is therefore a key concern for many local authorities and transport operators, with increased accessibility of public transport services helping to tackle transport-related social exclusion for those individuals and communities that are most vulnerable.

2.2.2.1.3 Cost of travel

As previously identified, affordability is a major barrier to public transport use and as a result can be a cause of social exclusion. The SEU report found that the accessibility and use of a public transport system is determined by several factors, including whether people can access key services at reasonable cost, in reasonable time and with reasonable ease. The affordability of travel is likely to affect the most vulnerable members of the population to the greatest extent, including the older, disabled and those on a low income. Previous studies conducted into the relationship between transport service access and affordability have found that the most deprived segment of the population spend significantly more of their income on transport than the least deprived, in some cases up to 400% more (Bocarejo and Oviedo 2012). Bus fares in the UK rose by around a third between 1985 and 2000, which is likely to have had an adverse effect on accessibility for the most deprived and vulnerable groups of the population (SEU 2003). Since then, a number of concessionary schemes offer free travel to those that are older or disabled, such as the ENCTS. However, there have been arguments that these schemes are too widely targeted, and instead may benefit more people if concessions were provided by means testing; benefiting those outside of these groups that struggle to afford public transport services, for example students and those looking for work (Butcher 2020).

Proposed policies to improve public transport services through fare increases may exacerbate the transport disadvantage and social exclusion of the most vulnerable groups of the population. In response to this, there have been a number of studies focused on the effect of fare changes, in particular fare decreases, on the accessibility of the transport network. It has been found that amalgamating bus fare zones and offering fare discounts allowed for transport disadvantaged groups to travel more often and access employment opportunities that they may have otherwise been unable to reach (Loader and Stanley 2009). There is also evidence that lower public transport fares allowed for transport disadvantaged groups to widen their spatial search for

employment opportunities and training, which in turn could alleviate social exclusion and transport disadvantage (Butcher 2020).

2.2.2.2 The older population

The majority of studies that focus on mobility, wellbeing and social exclusion have looked at this phenomenon at a community level, however it should also been noted that social exclusion is not only clustered within deprived and isolated communities, but is also scattered among individuals in the population (Kenyon et al. 2002). Older people have been found to be particularly vulnerable, with those aged over 60 identified as being at particular risk of being transport disadvantaged (Andrews et al. 2012). Understanding the travel behaviours of the older population, and the determinants of these behaviours, is therefore an important first step in understanding the causes of, and potential solutions to, social exclusion and transport disadvantage.

There has been considerable previous research into the travel behaviours of the older population, with many studies aiming to understand their mobility patterns and possible determinants, as well as assessing how these compare to the travel behaviours of the working age population (Cui et al. 2016; DfT 2020; Busari et al. 2019). The older population's travel behaviour has undergone change over the past several decades. This is largely due to the changing lifestyle context and situation of the older population, for example a shift towards more active lives and independent living, as well as an increase in the number of older people holding driving licenses and with access to private mobility (DfT 2020).

Older transport users are often seen as homogenous, with their travel patterns and behaviours generalised to the whole older population. However, there is a need to recognise the older population as a heterogeneous group in terms of lifecycle, location and lifestyle characteristics (Hildebrand 2003). In order to account for these differences, some studies have focused on the clustering of the older population into different groups dependent on their travel behaviours. These clustering methods can be used to identify groups of the population that may be the most vulnerable to transport-related social exclusion, such as 'captive public transport users' (Haustein 2011). The groups identified in such studies may be of particular interest to policy makers and transport authorities as certain clusters are likely to be more at risk of social exclusion than others, such as those without access to private transportation, and could therefore be targeted with specific policies or interventions (Preston and Raje 2007; Lucas 2012). In addition to research focuses on the travel behaviours of the older population, research has also

been conducted into the risks of social exclusion in older age. For example, Key and Culliney (2016) social exclusion within the older population and found that those aged 85 and over were at increased risk of social exclusion than those aged 65-84. This emphasised the heterogeneity within the older population.

Other studies have focused on comparing the mobility patterns and travel behaviours of the older population to other groups, such as those that are of working age. These studies have found that travel behaviour differs both within the older population and between the older and working age populations. Using interviews and travel surveys, Mackett (2014) found that those aged between 50 to 69 made more trips than average, with people aged over 70 making fewer. Additionally, Li et al. (2012) found that travel patterns varied with age and gender, with females and those aged over 80 generally making shorter trips than average. The differences evident within populations highlight the need for granular, individual analyses into travel behaviour.

Due to the ageing population in the UK, the general travel behaviours of the older population are likely to change in the future. Better healthcare and changes in lifestyle mean that future older populations are likely to be more mobile and continue driving to a later age than the current population (Rye and Scotney 2004). Understanding what influences the travel behaviours of the older population is therefore vital in order to predict how these behaviours may change in the future and which groups may be at risk of social exclusion (Walsh et al. 2017). This can ensure that policies and interventions that are implemented to tackle social exclusion and transport disadvantage have a continued effect over many years.

2.2.3 Mobility policies

Due to the significant relationship between mobility and wellbeing, a number of policies have been introduced over the years aiming to maintain high levels of mobility for those that may be at risk of social exclusion. These policies have had varying levels of success, with some suggesting that alternative policies, often more specifically targeted at at-risk populations, may be more beneficial.

2.2.3.1 The English National Concessionary Travel Scheme

One of the main barriers to public transport that the older population experience is the affordability of fares (Titheridge et al. 2009). The ENCTS was introduced with the aim of tackling this barrier by providing free local bus travel to those that had reached pensionable age.

Although there is a general consensus between both pass holders and non-pass holders that the ENCTS is beneficial to older passengers, due to the high cost of this scheme to the taxpayer, estimated at around £1 billion per year, there have been calls to reduce or abolish the scheme (Mackett 2013). The contentious nature of the scheme has meant that there have been a number of studies that have looked at the older population's travel behaviour and the benefits of the concessionary travel scheme.

Research conducted into how the older population's travel behaviour has changed since the introduction of the ENCTS found that pass holders were more likely to report increasing trip frequency (Andrews et al. 2012). Trip patterns were also found to have changed, with concessionary pass holders making more trips stages in the same day. This suggests that the benefits of the concessionary pass extend beyond the removal of the cost barrier to offering greater flexibility in travel. Increased levels of mobility as a result of the ENCTS were also found to alleviate loneliness by allowing older people to maintain their economic and social involvement in society, in addition to improving physical and mental health (Gabriel and Bowling 2004; Mackett 2014). Research conducted by Whitley et al. (2020) found that those that were eligible for the ENCTS demonstrated increasing bus use and generally benefitted from better access to services, and Jackson et al. (2019) found that this then led to greater physical activity and less social isolation among these populations.

The positive effects that the ENCTS has had on the mobility of the older population therefore increases concerns around the decline in bus usage by older concessionary cardholders that has been recorded in the West Midlands. With the ENCTS found to have increased the frequency and flexibility of travel, and subsequently increased social inclusion, there are fears that this decline in bus usage may be indicative that segments of the population may be suffering from transport disadvantage.

2.2.3.2 Other policy solutions

The high cost of the ENCTS to the taxpayer has brought about questions as to whether this money could be more beneficial if invested elsewhere. One of the primary arguments against the ENCTS is that although the scheme benefits public transport users, these benefits are not felt by those that lack access to public transport services. The widespread variation in the benefits that the ENCTS brings mean that there are questions as to whether this money could be better invested in schemes that benefit the wider population, rather than only benefitting those that are able to access public transport services.

Due to the variation in the quality and frequency of bus services in more rural and isolated locations, it has been suggested that the money currently being spent on the ENCTS could benefit those that are arguably most in need by instead being used to improve transport services in locations where current services are limited (Age UK 2012). Rural communities are the most at risk of becoming fragmented, with older residents more likely to suffer from feelings of 'entrapment' than their urban counterparts (Palacin et al. 2016). The transport and mobility needs of older people living in rural areas are traditionally served by registered bus services, however both passenger numbers and service levels have been in steady decline as private car use increases (Gray et al. 2006; Li et al. 2019). Investing the money currently spent on the ENCTS to increase rural bus services could help to more effectively meet the mobility needs of those that are the most vulnerable to transport-related social exclusion. However, it has been found that if these schemes are not specifically targeted towards the older population they may miss out on some of the benefits (Wang 2007).

In contrast to this, there have been arguments that the mismatch that is often found between traditional public transport services and the older population's mobility needs suggest that rather than investing money in traditional services, such as registered bus services, older people may benefit more from demand responsive services. Individual needs of the older population often differ from the mobility needs of the working age population as mobility and access needs are both user specific and often local in nature (Wang 2007). The current focus of public transport services on efficiency and effectiveness is unlikely to provide the level of service that is required by the older population (Battellino 2007). Investing in non-traditional, demand responsive services, for which users can specify the times and destinations that they would like to travel and benefit from a more personalised service with specially trained drivers and home pick-ups, may better address these issues.

A number of non-traditional transport services have been implemented in urban areas around the world, including the Telebus in Victoria, Australia (Wang 2007), and Dial-a-Ride services in the UK (Davison 2014). Demand responsive transport services have grown in popularity over the past several decades (Mageean and Nelson 2003; Coutinho et al. 2020), however there are still questions surrounding whether these services will be able to meet the mobility needs of an ageing population in the UK. Traditional Dial-a-Ride services have often been criticised for the high cost of provision, the lack of flexibility in route planning, and their inability to manage high demand. Despite the issues with demand responsive services in the UK, the Telebus service in Victoria, Australia has been successful in providing transport services to those that are unable to access traditional public transport services. In addition to this, 36% of those using this service were aged 55 and over, suggesting that such services would be welcomed by older transport users. The main benefits that were identified by Telebus users were the ease of

mobility, personalised service and assurance of safety that this non-traditional transport service can provide (Wang 2007).

The success of some demand responsive services poses the question as to whether implementing similar services in the UK could more effectively tackle the mobility and social exclusion issues experienced by much of the UK's older population. With some suggesting that the current ENCTS may be too widely targeted (Mackett 2014), investments in public transport could instead be tailored towards the specific needs of those that are most at risk.

2.2.3.3 Mobility vs. accessibility

The potential causes of transport-related social exclusion, along with the current policy approaches outlined previously, highlight the important roles played by accessibility and mobility. Accessibility is related to the ease of which a place or facility can be reached, whereas mobility refers to 'ease of moving'. In this context, transport planning that is centred on increasing mobility tends to focus on the means of travel rather than the end destination. In contrast to this, transport plans that aim to increase accessibility have a greater focus on the end destination and the individual traveller rather than the transport system (Handy 2002).

In recent years there has been a transition from transport planning that focuses on facilitating mobility, to instead becoming increasingly centred on the enhancement of accessibility (Ferriera and Beukers 2012). Accessibility has been accepted as a policy goal by the UK government, in particular since the SEU report which drew attention to the previous absence of any coherent responsibility in government for accessibility (Farrington 2007). Even with this shift in focus there are still several arguments to be made as to whether transport planning and policies should primarily focus on increasing mobility or accessibility, and which of these will provide the greatest benefits to both individuals and the wider community. One of the primary arguments against a focus on mobility is the environmental, social and financial costs that occur with an increase in mobility. These include increased CO_2 emissions and the depletion of natural resources as a result of increasing car use, increased road accidents and stress caused by congestion and delays, as well as the financial cost related to the development and maintenance of infrastructure (Ferriera and Beukers 2012).

Shifting the focus of transport planners and policy makers to the enhancement of accessibility could tackle some of these costs. For example, increasing the accessibility of key services and facilities could reduce the need for travel. This could be achieved by implementing policies that ensure that future developments are easily accessible by modes of transport other than the private car, as well as improving the accessibility of the current transport network (Handy

2002). These policies would improve the accessibility of key facilities by tackling the issues related to increasingly disjointed patterns of development.

The conflict between restraining mobility and maintaining accessibility reflects the wider conflict between sustainability and growth that is ongoing in many cities globally (Marshall 1999). To ensure that new developments are easily accessible and people enjoy reasonable levels of mobility, a balance must bet struck between the two. Chapter 8 of this thesis explores this concept in relation to the bus network in the study area; exploring the impacts of transport provision and the accessibility of key facilities on the activity of the older population.

2.2.3.4 Mobility innovations and 'Mobility as a Service'

The West Midlands has been chosen as the UK's first 'Future Mobility Area' which will see investments in the transport system to make journeys quicker, cheaper and cleaner (TfWM 2018). This includes investment in new modes of transport, services and technologies, such as 'Mobility as a Service' (MaaS), car sharing, electric bikes and the increased use of data to tackle congestion and improve the efficiency of the transport network. Research into current mobility patterns can ensure that these investments and innovations benefit the whole population and do not exacerbate issues of transport disadvantage that are already present.

MaaS is a strategic solution to improve urban mobility and describes a shift away from personally-owned transport towards mobility provided as a service. The aim of MaaS is to provide an "efficient, personalised transport service, integrating a range of modes and presenting the consumer with the best options for getting from A to B" (TfWM 2018). This is facilitated by the 'Whim' app which provides flexible payment and ticketing options and allows for users to access a range of mobility options, meaning users are not tied to one mode of transport and can instead take advantage of multiple modes under the same ticket (see whimapp.com).

The increasing focus of future transport policies and innovations on 'smart transport' increases the risk of exclusion of those that are unable to access the necessary technologies to take advantage of these opportunities. The older population are particularly at risk, with lower levels of internet access and engagement with new technologies (Nimrod 2018). Understanding the ways in which the older population make use of the current transport system and which groups may currently suffer from transport disadvantage can allow for transport authorities to target these groups specifically to ensure that future innovations do not further exacerbate these issues.

2.3 Applications of consumer, or 'big', data

The recent increase of cloud computing, mobile networks and other new technologies have given rise to large quantities of data and information, which are often referred to as 'big data'. Big data has become a ubiquitous term, primarily due to its widespread use in both academia and industry. This widespread, cross-discipline use has meant that there is no formal definition of big data, however it is generally accepted that "big data represents the information assets characterised by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value" (De Mauro et al. 2015: 103). The consumer data utilised in this can be referred to as big data, and for this thesis these terms are used interchangeably.

This section discusses the emergence of big data, and the challenges and opportunities that come with this, as well as the value of these data within urban and transport studies. Smart card data has widely been identified as a potential resource for understanding urban mobility and travel behaviours of the older population, as well as the social issues that are related to this (Pelletier et al. 2011; Mohamed et al. 2014). However, other consumer data sources, such as loyalty card transactions, have received little attention in this arena. This section discusses the extent to which smart card data have been utilised within transport studies and outlines the research that have utilised loyalty card data for population studies. The potential of these data to provide insight into mobility and activity patterns is also discussed.

2.3.1 Challenges and opportunities of big data

Since 2011, the interest surrounding big data has increased exponentially; with widespread use across academia, industry and the media. These large, complex datasets can offer insight into many multifaceted issues with accuracy and timeliness. However, due to the relatively recent emergence of big data, there remain a number of issues and concerns to be addressed.

2.3.1.1 Big data challenges

The volume and velocity of data that are being produced and collected means that there are a number of ethical issues surrounding the use of big data, and there have been fears that current data use guidelines may no longer be appropriate. The emergence of big data challenged the internationally agreed guidelines that were developed by the OECD in 1980, which were intended to protect the privacy of the data subjects and ensure that personal data are not misused

by implementing limitations to data collection; ensuring that there is transparency regarding the use of the data and ensuring that there are appropriate data safeguarding and security measures in place (ITF 2015). The rise of big data challenged these original guidelines, with particular concern surrounding the data fusion and aggregation methods that could potentially de-anonymise data, with the analytical methods that are often implemented having the potential to reveal information about individuals in the datasets. In response to these concerns, the General Data Protection Regulation (GDPR) legislation was passed by the European Union 2016 and implemented in 2018. The purpose of this legislation was to give consumers better control of their personal data and to set limitations on what companies can do with personal data and how long it can be retained. With the increasing collection and implementation of these data there is a clear need for further research into the potential impacts of the emergence of big data and the evolving nature of consent that may result from this.

There tends to be an assumption in Government and industry that those that utilise big data will be more successful and productive and have a differential impact across all industries (Kaisler et al. 2013). The main concern associated with this is that there are a lack of tools and people with appropriate training to effectively exploit these data. In contrast to this, others suggest that the main concern should be that the analysis of these datasets may result in invasive and unwanted marketing, particularly from large corporations with the finance and skills to exploit large consumer datasets. There is a fear among critics that big data may be misused and could give corporations new abilities to manipulate consumers and compete unfairly in the marketplace (Bollier 2010). These arguments again highlight the need to not only develop expertise in the analysis and exploitation of big data, but also the need for new guidelines and ethical procedures to ensure data are not exploited and misused. As more data are collected, these issues will only become more prominent as there is a growing concern that organisations, in particular those related to health and social media, may know too much about individuals (Kaisler et al. 2013). This is an ongoing challenge that will need to be addressed in the near future to ensure the continued use of big data.

In addition to the ethical and privacy concerns related to the exploitation of big data, there are also a number of challenges associated with the collection and analysis of large datasets. Utilising big data and accurately interpreting the results of analyses without introducing bias has been identified as a major challenge (Bollier 2010). These datasets are often assumed to accurately reflect the social world, however there can be significant gaps, with little or no data collected from certain communities (Crawford 2013). This is particularly relevant for mobile phone and social media data, for which data will not be collected from those communities that lack access to these technologies, such as low income communities and the older population. This presents a significant challenge for those utilising big data, with decisions needing to be

made regarding which communities or populations are likely to be excluded or underrepresented. To an extent, this can be tackled by asking questions of the data, such as where the data come from and how they were collected. However, in the longer term combining big data approaches with small-scale studies may be more successful in validating both the data and the results by adding context and depth (Crawford 2013).

In addition to the bias that can be introduced during data collection, bias can also be introduced when cleaning these datasets, i.e. deciding which variables should be included in the analysis and which should be discarded. This is a much more prominent issue when analysing big data rather than conventional data sources, such as surveys and questionnaires, as the data collection methods for these conventional sources often mean that the data that are collected are already known to be relevant (Long and Thill 2015). In contrast to this, big data often contain a large number of variables that may not always be significant for the analysis and can therefore be removed. It has, however, been argued that cleaning the data to exclude certain variables removes the objectivity of the data themselves as the process is inherently subjective (Bollier 2010). Issues of bias are a major challenge for those utilising big data, with more research needed in order to minimise these risks (Ferrari et al. 2013).

2.3.1.2 Big data opportunities

Despite these concerns, big data offer new opportunities for the collection and analysis of data. Traditional data collection methods have primarily relied upon surveys, censuses and interviews (Long and Thill 2015). These data collection methods are not only costly to undertake but can also be extremely time consuming. The costs related to traditional data collection can limit who can utilise these datasets, and the time taken to collect and analyse the data can mean that once an analysis is complete the results may no longer be current (Kieu et al. 2014). Big data offer a solution to these issues with the cost of data collection greatly reduced and data often collected at or near real-time (Kaisler et al. 2013). This means that big data have the potential to provide more current and timely results, therefore improving the insights that can be drawn from these analyses.

The emergence of big data has also changed the way that scientific analyses are undertaken. Before the widespread use of big datasets, analyses were commonly limited to a small number of hypotheses that were defined before any data were collected (Mayer-Schonberger and Cukier 2013). Big data allow for a move away from the search for causality, instead letting the data 'speak for itself'. This can allow for patterns and correlations in the data to be discovered, offering novel and invaluable insights that may not have originally been anticipated. Although this new approach to data analysis may not provide insight into why a phenomenon is occurring,

the knowledge that it is occurring can provide a basis for further research into the potential causes.

This section has outlined just some of the challenges and opportunities that individuals and organisations are faced with when utilising big data. Although there are a number of concerns surrounding the emergence of big data, primarily regarding data privacy and misuse, the implementation of suitable guidelines and practices can ensure future analyses are conducted in a safe and ethical manner. Although there is still much research needed in order to ensure the safe and reliable use of big data, these datasets offer the opportunity to transform how data are collected and analysed in both academia and industry.

2.3.2 Smart card data

New ways to collect, manage and analyse large quantities of data present opportunities for local authorities and transport operators to provide a smarter, more efficient and inclusive transport system. The advent of inexpensive, often open-source analytical software has also greatly increased access to the cost-effective and near real-time processing and analysis of large datasets. Despite this, access to large-scale transport datasets remains a barrier to their widespread use for research and transport planning (Long and Shen 2013). The 'opening up' of these transport data has been identified as a significant goal for the transport sector; gaining support from the UK Government with £14 billion being invested to make the data that are routinely collected by transport operators and local governments accessible for researchers (Hobbs and Hanley 2014). As a result of this, the range of open transport data that are available is ever expanding, with the quality of available transport data also improving.

There have been many examples of big data being utilised in the urban planning and transport sectors. Data collected from handheld GPS devices have been utilised to record urban diaries (Newhaus 2009), elicit travel modes in urban centres (Gong et al. 2012), and analyse taxi drivers' travel behaviour patterns (Liu et al. 2010). In addition to this, data collected from mobile phone networks have been utilised to evaluate the density and spatiotemporal patterns of urban activities (Ratti et al. 2006). Despite these applications of big data, the technologies and methods remain immature, with most research into urban transport systems continuing to rely on traditional surveys and interviews (Long and Thill 2015).

A relatively recent technology that produces large volumes of transport data are AFC systems. These AFC systems collect payment from smart travel cards; recording all journeys that are made on the transport network. Smart card technology has been widely used in many sectors, including healthcare, banking and government, since the 1990s (Pelletier et al. 2011). Increasingly, transport authorities are investing in AFC systems to collect revenue from public

transport users as these systems are seen as a secure method of user validation and fare payment. Although the main purpose of these smart card AFC systems is to manage revenue collection, there are also several additional benefits that come from transit authorities investing in these systems. These include increased user convenience, reduced vehicle delays, increased ease in monitoring payments and the interoperability of the system i.e. cards can support different fare types at the same time (Pelletier et al. 2011). These systems also produce extremely large quantities of data on a daily basis, providing a valuable resource for studying the travel habits of transport users (Mohamed et al. 2014; Kusakabe and Asukara 2014).

The majority of studies that utilise smart card travel data from AFC systems focus on how these data can be used to optimise the effects of schedule adjustments, inform operators of the supply and demand of the transport network and for long-term planning (Pelletier et al. 2011). These studies primarily focus on improving the efficiency of the transport network. However, there is also scope for research into how smart card travel data can provide insights into the urban dynamics and social structure of the wider urban area. The remainder of this section outlines some of the previous studies that have utilised smart card travel data in this way, with a focus on population insights, urban mobility, social exclusion and transport disadvantage.

2.3.2.1 Geodemographics

The large volumes of spatial data generated by AFC systems can allow for the analysis of travel patterns and behaviours at the individual level. This can be seen as a step forward from traditional spatial targeting methods which focus on the analysis of geographic areas, often limited by the lack of heterogeneity of the population (Dalton and Thatcher 2015). Despite this, one of the most commonly used and successful spatial targeting methods is geodemographics; a process of organising areas into categories that share similarities across multiple socioeconomic attributes (Singleton and Spielman 2014). Birkin and Clarke (1998: 88) offered a definition for both demographics and geodemographics: "demography is the study of population types and their dynamics, therefore geodemographics may be labelled as the study of population types and their dynamics as they vary by geographical area". Geodemographic classifications emerged as a solution for managing highly dimensional census data, with early methods focusing on single-cities before being expanded into national coverage classifications (Singleton and Spielman 2014). Geodemographic classifications have been utilised in both public and private sector applications since the 1970s, with an increasing number of commercial classifications emerging throughout the 1980s, primarily being used as a private-sector marketing tool (Singleton 2008).

In more recent times, geodemographics have re-established a base of applications within the public sector (Longley 2005). These applications include the use of geodemographics to better

deploy resources for local policing (Ashby and Longley 2005), to investigate participation in Higher Education (Singleton and Longley 2009), as well as being used as a tool for targeting neighbourhoods in public health campaigns (Petersen et al. 2010). These examples primarily draw upon commercial geodemographic classifications. However, it has been argued that because these are often created as 'black-box' systems, with little transparency around the raw data and methods that are used (Singleton and Longley 2009), this can be problematic for public sector applications where life chances may be apportioned through their application. As such, the creation of an open source geodemographic classification by the Office of National Statistics (ONS) was an important step in developing a more reproducible and transparent geodemographic approach (Vickers et al. 2005). Despite the legacy of geodemographic applications in both public and private sectors, there have been few extensive studies linking travel behaviour and social exclusion within a geodemographic framework.

There are two main limitations associated with the use of geodemographics; ecological fallacy and the modifiable areal unit problem (MAUP). Ecological fallacy refers to the idea of "inferring characteristics of individuals from aggregate data referring to a population" (Johnson 1986). This idea can present a problem when undertaking geographical analyses, as conclusions can be drawn about individuals based on aggregated data that may not be accurate. In terms of this study, ecological fallacy is likely to have an impact if conclusions are drawn about individual travel behaviours based on aggregate travel patterns. The MAUP refers to the "geographic manifestation of the ecological fallacy in which conclusions based on data aggregated to a particular set of districts may change if one aggregates the same underlying data to a different set of districts" (Waller and Gotway 2004). This can be particularly relevant in the analysis of transport behaviours when conclusions are drawn based on custom transport boundaries, such as public transport catchment areas. It is vital that these two concepts are considered when employing geodemographic analyses.

In the assessment of transport disadvantage and social exclusion there are two commonly used approaches; the category approach and the spatial approach (Church et al. 2000; Herwangi 2013). The category approach relates to the characteristics attached to an individual, such as age, gender and race, alongside travel patterns and attitudes. Conversely, the spatial approach relates to the characteristics associated with the area in which an individual resides, for example rural or urban, the accessibility of the area and the geodemographics. Both approaches have a number of limitations. The category approach relies on the assumption that social groups are homogenous in terms of their material affluence and activity patterns, while the spatial approach tends to be too localised to capture transport-related social exclusion at the city-scale (Church et al. 2000).

The large quantities of smart card travel data produced by AFC systems can be utilised to tackle these issues. Combining both the spatial and category approaches through the use of individual travel records and geodemographic classifications could provide an alternative method of analysing transport disadvantage and social exclusion, in addition to helping transport authorities determine whether resources for tackling transport-related social exclusion should be targeted at specific social groups or geographic areas. Merging elements of both the category and spatial methods could help transport authorities and policy makers to understand individual-level exclusion better through an area-level perspective. By combining both an individual and geographical approach to the analysis of smart card travel data, a novel solution to identifying areas and groups of people that may be transport disadvantaged could be developed.

2.3.2.2 Travel patterns and behaviours

Much of the research that has been conducted into the use of smart card data focuses on understanding the travel patterns of users and clustering these users into distinct groups with similar behaviours. This can provide insights into the general travel behaviours of different groups of the population; allowing for transport operators and local authorities to identify those whose mobility needs may not currently be being met by the transport network.

Utilising smart card travel data in order to understand the temporal travel patterns of public transport users can help transport operators to identify different groups of users, for example, cardholders that primarily use public transport for commuting in the peak morning and afternoon hours, or those that predominantly use the public transport network on weekends. Studies into temporal travel patterns have been undertaken from both a station-oriented and a passenger-oriented approach, as the temporal profile of public transport stops and stations can provide insights into the role that these play in the public transport network and which temporal cluster these stations primarily cater for (Mohamed et al. 2017). These insights can contribute to a more comprehensive understanding of how the transport network is utilised. Combining this approach with a passenger-oriented approach can contribute to a better understanding of which groups of the population are utilising which transport stops and stations and at which times of the day. These insights can be particularly valuable if the temporal profiles of transport users can be combined with socioeconomic and demographic data to provide a more complete view of how the transport network is utilised and by which groups of the population (Langlois et al. 2016).

Although the temporal travel patterns of public transport users can provide valuable insight into travel behaviours, in order to fully understand passenger travel patterns, it is important to also investigate the spatial patterns of public transport use. For some smart card travel datasets,

understanding the spatial patterns of travellers can be challenging. This is principally due to the fact that many transit systems only require passengers to tap in, meaning that the destinations of individual journeys are often unknown. To tackle this issue, research has been conducted into inferring the destination of passengers by combining smart card travel data with vehicle location data and subsequent boarding locations (Gordon et al. 2013; Kieu et al. 2014). This method allows for passenger trips to be reconstructed and therefore network flows to be observed; and was found to be efficient enough to be conducted daily. This allows for transit operators to be able to observe travel behaviour on all services and at all times. This ability to observe the spatial travel behaviours of public transport users and track passenger behaviours over time can aid transport operators in identifying areas or segments of the population that may be underserved by the current transport system (Kieu et al. 2014). Understanding the spatial and temporal patterns of public transport users can also aid transport authorities in tackling issues surrounding accessibility and social exclusion; by helping to identify groups of the population that may be the most vulnerable or at risk based on a combination of their travel patterns, socioeconomic and demographic characteristics (Liu et al. 2006).

2.3.2.2.1 Travel behaviours of the older population

Although the majority of studies that utilise smart card data focus on the working-age population, a number of studies have been conducted using smart card data to understand the mobility and travel patterns of the older population. Smart card data have the potential to provide a more comprehensive understanding of the older population's travel patterns, including trip frequency and purpose. One of the main benefits of smart card data is the potential for these data to provide insights into the travel patterns of both frequent and infrequent public transport users, as most traditional survey methods may under-represent infrequent users. This is of particular importance when analysing the travel patterns of concessionary travellers, as the retired population are more likely to have irregular travel patterns than the working-age population due to a decrease in regular commuting trips (Jiang et al. 2012).

This idea was further explored in a study by Last (2010), in which the travel patterns of concessionary travellers in the North West of England were analysed, with a focus on active and inactive concessionary pass holders. It was found that of all concessionary pass holders in the study area, 56% made no trips within the 5 week study period, with only around 27% of pass holders making trips more than once per week. It was also found that 8.4% of pass holders made 57% of all concessionary trips over the study period. The large variation in trip frequency between concessionary pass holders highlights the diversity of the older population and their travel patterns. The large percentage of pass holders making little or no use of their

concessionary passes also raises questions around the effectiveness of the scheme; suggesting that only a small proportion of the older population are gaining maximum benefit.

This study also highlights how smart card travel data can be used to look at how trip frequency varies by age and geographic area. It was found that there is a gradual increase in trip frequency from age 60, with the greatest number of trips per week being undertaken by those in the 75-79 age group. Trip frequency also varied by the type of area in which the concessionary traveller lives. Trip rates were higher for those living in urban areas, and lower for those living in more rural areas. This trip variation is likely due to greater frequency and quality of transport services in urban areas, and the higher car ownership rates that are often seen in more rural areas (Gray et al. 2006).

In addition to looking at concessionary trip frequencies, studies have also utilised smart card travel data in order to gain a more detailed understanding of how the older population utilise the current transport network. Research conducted by Song et al. (2015) looked at the travel behaviour of the older population in Seoul using smart card travel data over a one week period, with a focus on travel patterns and mode choice. It was found that the majority of concessionary travellers undertook trips in off-peak hours, between 10am and 4pm. This study also found that 99% of trips undertaken by concessionary travellers consisted of single mode trips, with an average of 0.04 transfers per origin, compared to 0.43 for the working-age population. This suggests that the older population are less likely to undertake trips that involve an interchange and may suggest that the older population have difficulty navigating transfer facilities. Analyses such as these can highlight aspects of the current transport network that may act as barriers to the older population's mobility.

The potential of smart card travel data to provide a more detailed view of concessionary travel patterns can allow for transport authorities to move away from relying on average characteristics, calculated from aggregated data (Last 2010). These data can allow for individual trip characteristics, such as start time, journey distance and journey time to be analysed. This can enable transport authorities to gain a more comprehensive view of how the older population are utilising the transport network.

2.3.2.3 Public transport access

Smart card data can not only be useful for examining passenger behaviours and urban movement but can also provide valuable insight into the accessibility of the transit system. As previously discussed in Section 2.2.2.1.2, accessibility is a key factor to be considered when assessing the efficiency and effectiveness of transport networks. One of the main barriers to

accessibility is the distance that one has to travel in order to access public transport services (Viggiano et al. 2016). Access distance to public transport is also an important factor to consider when evaluating the efficiency of public transport networks and identifying areas with low accessibility, although accurately measuring access distance can be difficult.

Smart card travel data can provide a novel approach to assessing public transport accessibility based on the access distance to stops and stations. Some smart card travel data, including London's Oyster card data, include the postcode of the registered home address of the users. Access distance can be inferred from these datasets as a straight-line distance or the road network distance from the registered postcode to the origin station recorded by the smart card. This method was undertaken in a study by Viggiano et al. (2016), in which it was found that when compared to the London Travel Demand Survey, this methodology was relatively accurate in estimating the accessibility of transport stops; finding that access distances tended to be longer preceding rail journeys than those preceding bus stages. These findings would appear to agree with previous research, as it has been found that people tend to be willing to travel further to access rail and subway stations than bus stops (Daniels and Mulley 2013; Burke and Brown 2007). The results of this method can highlight areas which may be underserved by the current public transport network. In addition to this, these results can also be utilised to assess the service areas of each public transport stop and station. Assessing the geodemographic attributes of each service area can provide a more comprehensive view of how the transport network is utilised and by whom, as well as providing insights into the social structure of the urban area (Smith et al. 2013).

The accessibility of public transport systems can play an important role in determining public transport use. The insights into accessibility that smart card travel data can provide offer a unique opportunity for transport operators and policy makers. Utilising these data to assess the accessibility of different transport stops and stations could improve the planning of future public transport services and be utilised in order to target areas of low accessibility.

2.3.3 Retail data

In addition to transport-related consumer datasets, there is evidence that alternative consumer data sources may be able to provide insights into mobility and activity patterns. In this thesis, loyalty card transaction data are implemented to explore this concept. Although at the time of writing there has been no research published that specifically focuses on understanding mobility and activity patterns using loyalty card transaction data, there have been a number of studies focused on the relationship between deprivation, exclusion and access to retail facilities. This

section outlines the key research that has been conducted around these concepts and highlights the potential of these data for providing insights into mobility, activity and wellbeing.

2.3.3.1 Retail geography and wellbeing

Shopping opportunities represent an important element of social wellbeing and inclusion. However, until recently the links between wellbeing and the geography of retailing had not been extensively explored (Knox 1981). Physical accessibility to retail facilities impacts wellbeing by minimising both the cost and time taken to travel to shopping opportunities, which in turn releases a greater proportion of both time and household income for recreational activities. In addition, having access to a number of retail facilities allows for greater choice, a concept that has been linked to wellbeing in a number of studies (Alsnih and Hensher 2003; Williams and Hubbard 2011). The retail choices that deprived or lower income shoppers make have been found to be largely influenced by income and, most importantly, transport availability (Robinson et al. 2000). Detailed retail transaction data, alongside demographic and socioeconomic data, can allow for this concept to be explored further.

There has been much discussion around the 'death of the high street' due to a combination of the 2008 financial crisis, a move to out-of-town shopping and an increase in e-commerce (Griffiths et al. 2008; Hughes and Jackson 2015; Hubbard 2017). It has been found that those on a lower income, those living in more deprived areas and those that lack access to private transportation are much more likely to shop locally than at out-of-town retail parks and are therefore more likely to be negatively impacted by this decline of high street shopping (Robinson et al. 2000). A study by Williams and Hubbard (2011) looking at the impacts of retail change on the shopping habits of consumers in Coventry found that deprived consumers tend to use 'traditional' shopping facilities rather than newer stores and shopping centres, with those that are more mobile and affluent deemed more able to take advantage of these new retail facilities, such as out-of-town and edge-of-town stores. Differences in shopping habits were also observed between ethnicities, genders, classes and those of different ages, with car-less, inner-city residents found to be the most disadvantaged by the move of retailers to out-of-town locations.

Given the importance of shopping and accessing retail services for social inclusion objectives, the availability of detailed transaction records can provide valuable insight into shopping habits and subsequently the social inclusion of individuals. The unique combination of both detailed public transport data and a large loyalty card transaction dataset used in this thesis presents an opportunity for these concepts to be explored through these two perspectives in parallel.

2.4 Further research potential

This chapter explored a number of key concepts which are referred to throughout this thesis. It also critically discussed some of the previous research that has been conducted into these concepts and made suggestions as to how consumer data could be utilised to further explore these ideas.

Section 2.3 outlined the challenges of working with big data. The need for further research into both the utility of consumer data and the practicalities of implementing these data for mobility insights is clear. Currently, there are few standard processes, benchmarks or methodologies for the analysis of big data, which can hinder their implementation for research within both an academic and industrial context. This thesis implements two sources of big, consumer data and, through the appraisal of these data, identifies and addresses a number of these challenges.

In terms of understanding the mobility patterns of the older population, this chapter highlighted why this is of heightened importance; with current and future investment and innovation in the transport system in the West Midlands focusing on 'smart technologies' and the implementation of data to improve the transport network. Firstly, demonstrating the utility of smart card data and the mobility insights that can be extracted are key for establishing the value of these data and the role they can play in these new mobility initiatives. Secondly, the risks of transport disadvantage and subsequent social exclusion that are associated with these investments emphasise the need for a comprehensive understanding of the mobility patterns of different populations and their transport needs and demand. In addition, this highlighted the need for research into the ways in which the current transport network may not be meeting these needs so that future policies and investments can be effectively targeted.

The potential for alternative consumer data to provide insights into mobility and activity patterns has been largely unexplored. This chapter discussed the relationship between retail activity, deprivation and wellbeing, with current research primarily relying on traditional surveys and questionnaires. The utility of large retail transaction datasets to provide detailed, granular insights into shopping behaviours and store visiting patterns would allow for these concepts to be explored in greater detail and across a larger sample size. In reference to travel diaries, Richardson et al. (1995: 151) stated that "it is well known, that persons who are less mobile or who feel themselves at a disadvantage, e.g. the unemployed or those on low income, are more reluctant to participate in travel diary exercises". The 'total-capture' benefit of these consumer data is therefore particularly pertinent when investigating issues related to deprivation and socioeconomics. With the challenges that both the retail sector and the transport sector are likely to face in the coming years; with increasing technological innovation resulting in

changing shopping and trip making behaviours, it is imperative that the potential of consumer data to provide insight into these phenomena is explored.

Chapter 3 – Contextual analysis of the study area and population

This chapter provides context to the study, focusing on the urban make up and transport network, as well as the study population. As this thesis focuses on a specific population, both in terms of residential location and age group, giving a detailed overview of this population's characteristics and the wider population in the study area helps to contextualise the results of subsequent chapters.

3.1 Study Area

As discussed previously, this thesis focuses on the older population in the West Midlands. The West Midlands is a metropolitan county and combined authority area located in west-central England, and is made up of seven metropolitan boroughs: Birmingham, Coventry, Dudley, Sandwell, Solihull, Walsall and Wolverhampton – the locations of which are shown in Figure 3.1. The primary urban centres in the study area are Birmingham, Coventry and Wolverhampton, with Solihull containing the most rural land use.

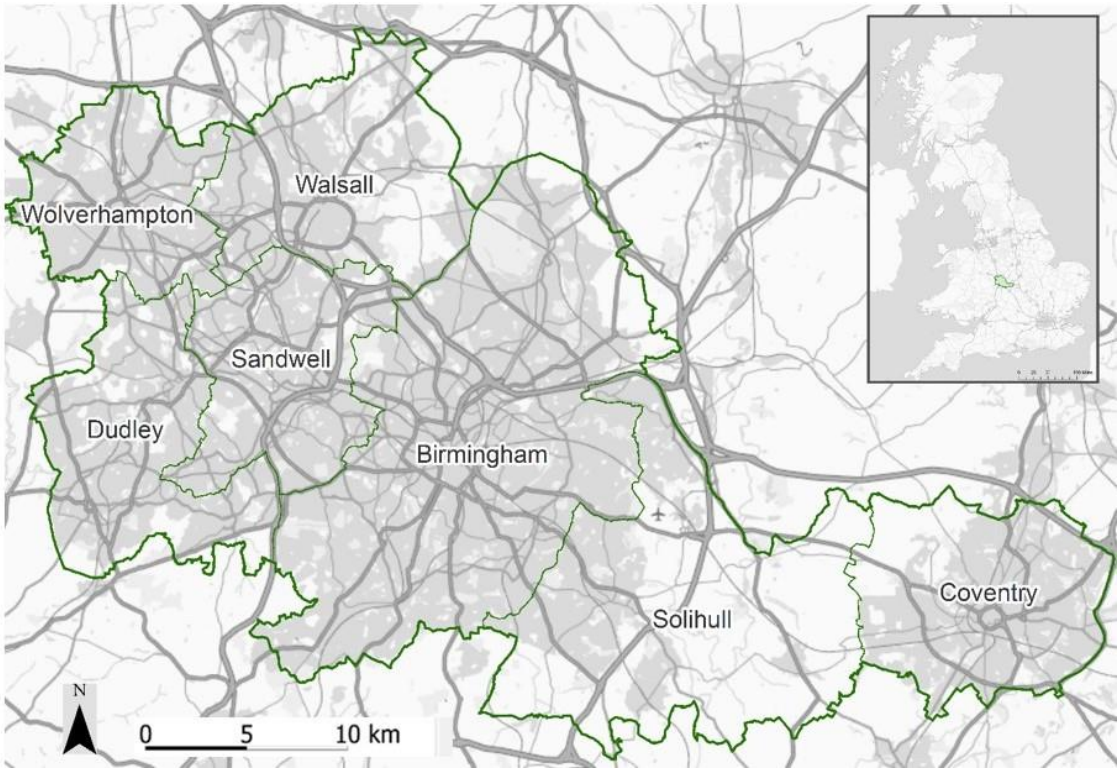


Figure 3.1 West Midlands metropolitan area

In June 2016, the West Midlands Combined Authority (WMCA) was created under the 'Local Democracy, Economic Development and Construction Act 2009'. This body has powers over transport, economic development, skills and planning. The authority's primary priorities involve co-ordinating the region to act as one on certain issues, including international promotion and investment, and improving internal and external transport links.

3.1.1 Transport in the West Midlands

The public transport network in the West Midlands consists of bus, tram and train services. The West Midlands' public transport services are utilised by large numbers of passengers each day, including commuters and visitors travelling into and throughout the urban centres. Providing an efficient, reliable and accessible public transport network throughout the whole of the West Midlands is therefore imperative for both social and economic reasons.

Smart travel cards, called 'Swift' cards, are used for ticketing on the transport network, in addition to smart ENCTS travel cards for the older and disabled populations. Smart travel cards increase user convenience and the efficiency of the transit network, as well as helping to reduce vehicle delays (Pelletier et al. 2011). The implementation of these smart cards and AFC systems in the West Midlands was therefore vital to ensure that passengers can conveniently make use of the transport network.

In addition to the high volume of public transport users that travel into the urban centres each day, Birmingham Curzon Street Station, situated close to the city centre, will be the terminus of High Speed 2 (HS2); a new high speed railway linking Birmingham to London Euston. This high speed railway will cut the journey time between London to Birmingham from around 1 hour and 20 minutes to under 50 minutes (Preston 2010). Although HS2 will improve the journey time between London and Birmingham, there has also been much discussion around ensuring that Birmingham is 'ready' for the expected opening of the line in 2026 in order for the benefits of this shorter journey time to be fully realised.

One key aspect related to this is the improvement of local public transport services throughout the city and the wider county. In order for HS2 to fulfil its potential, integration between the high speed rail network and national, regional and city-wide plans for transportation, the economy and development will be vital (Greengauge 21 2009). Investments in local metro and bus services can ensure that the economic and social benefits of HS2 can spread beyond the immediate vicinity of the Curzon Street terminus. The nature of HS2; being a large-scale engineering project, means that the benefits as well as the potential negative impacts are likely to be widespread. Those opposing the project often emphasise the need for improvements in

local and mid-scale transport services to take priority (Preston 2010). This highlights the need to ensure that resources are effectively allocated, and those that are vulnerable or disadvantaged are not disproportionately affected by investments being made in other areas of the transport network. Achieving an inclusive and sustainable transport network in the West Midlands is therefore vital, not only to ensure that large scale investments, such as HS2, are successful but also to ensure that local transport services do not suffer as a result of these investments.

3.1.1.1 Public transport services

With accessibility to public transport services being identified as an important factor in the wellbeing and social inclusion of the older population (Gray et al. 2006; Mackett 2013), understanding the accessibility of transport services is key. The primary mode of public transport in the West Midlands is the bus, which operates throughout the county, accepting both SWIFT smart card payments and concessionary passes. Although there are bus stops located throughout the county, there is clear variation in the provision of services and the location of bus stops shown in Figure 3.2.

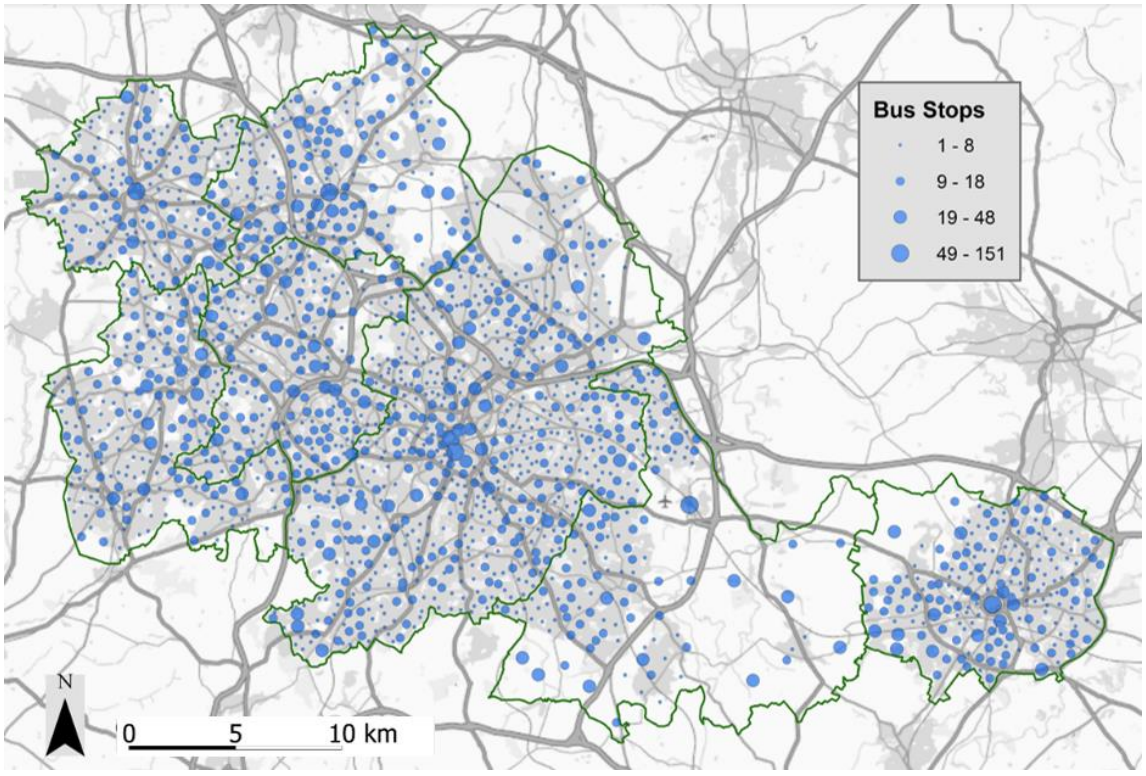


Figure 3.2 Number of bus stops in each LSOA – points proportional to number of stops and located at LSOA centroids

Although Figure 3.2 appears to show high numbers of bus stops located throughout the study area, in a report by The Guardian (2009), which utilised public transport data obtained from The National Public Transport Data Repository (DfT 2015), the West Midlands ranked last out of 143 areas in terms of the number of bus stops per 100,000 residents. This report found that the West Midlands contained just 264 bus stops per 100,000 residents, in comparison to the 1,534 found in the highest-ranked Gwynedd, Wales.

The density and location of bus stops are often used to determine the accessibility of bus services and have been shown to significantly influence the transport system's performance and level of service (Chien and Qin 2004). High concentrations of bus stops were primarily concentrated in urban centres, such as Birmingham, Coventry and Wolverhampton, with significantly lower concentrations in more rural areas. It has been found that walking distance from an origin or destination to a bus stop is a significant barrier to using public transport services, particularly for the older population (Hess 2012). Large variations in public transport provision suggest that improvements to public transport services in underserved areas and the provision of alternative transport services may be necessary to meet the mobility needs of the population (Mackett 2013).

In addition to the bus network, public transport in the West Midlands also comprises of metro and rail services. The West Midlands Metro opened in 1999 and runs on a single route, between central Birmingham and central Wolverhampton. The location of these rail and metro stations in the West Midlands and surrounding areas are shown in Figure 3.3.

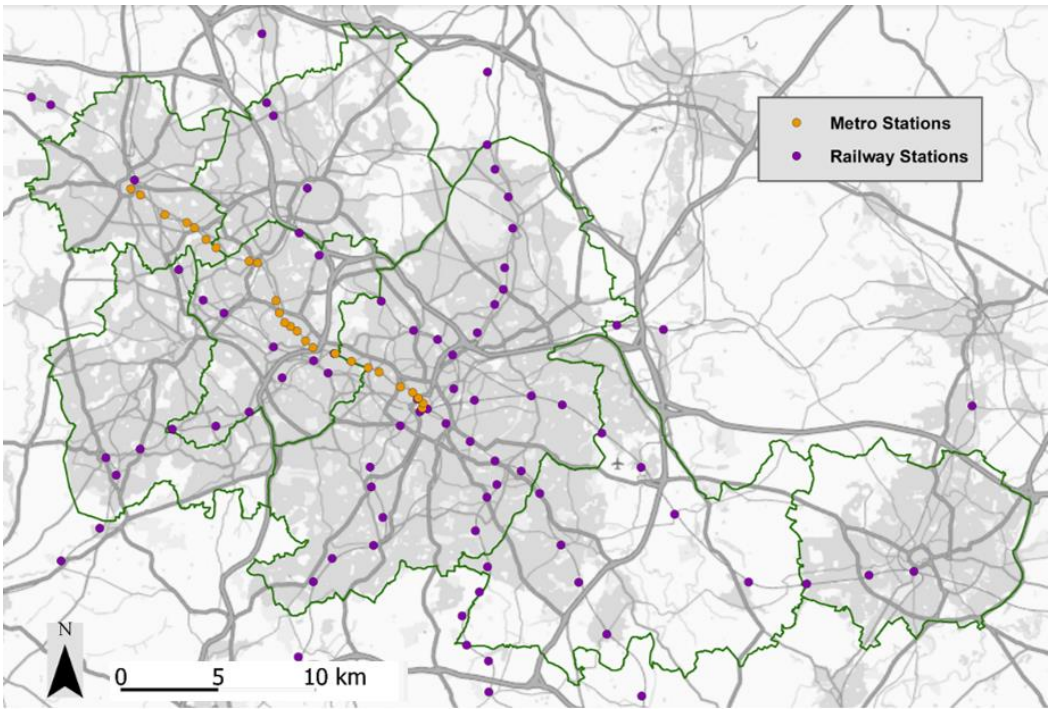


Figure 3.3 Location of railway and metro stations in the West Midlands and surrounding areas

Figure 3.4 Yearly bus, rail and metro boardings (TfWM 2017)

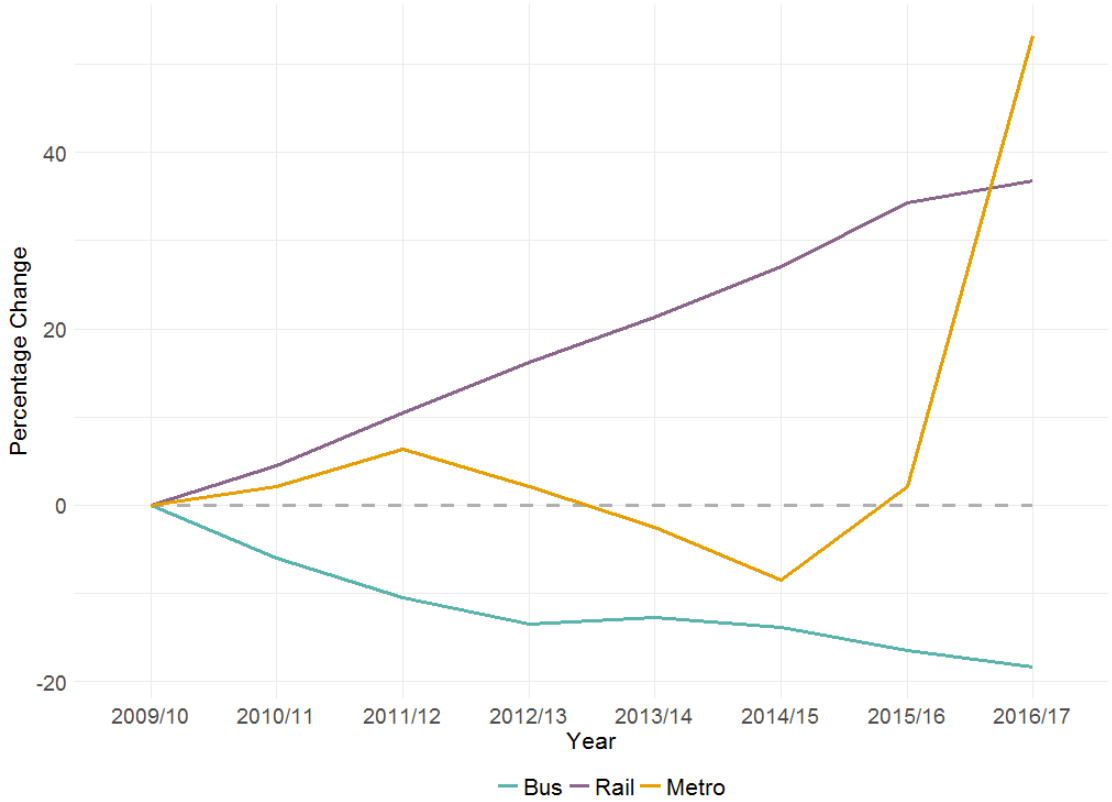


Figure 3.5 Percentage change in bus, rail and metro boardings (TfWM 2017)

Figure 3.4 shows the number of boardings per year for bus, rail and metro services from 2009/10 to 2016/17. It is clear from this figure that bus services have the majority of the modal share in the study area, with over 300 million boardings taking place over the 2009/10 period, in comparison to 40 million rail boardings and less than 5 million metro boardings. This figure does, however, show that there has been a reduction in bus boardings and an increase in rail and metro boardings since 2009/10. Figure 3.5 shows the percentage change in the number of boardings made by bus, rail and metro services over the study period. This gives insight into potential modal shifts throughout the study area. Rail journeys appear to have experienced a steady increase over the study period whilst bus boardings have decreased. The large increase in metro boardings between 2015/16 to 2016/17 is due to an extension to the metro line into Central Birmingham, which resulted in an increase in metro boardings of over 50% compared to 2009/10.

This modal shift away from bus services and towards rail and metro services may account for some of the decline observed in bus boardings, although bus journeys still made up over 80% of public transport boardings in the study area in 2016/17. Figure 3.3 shows the location of metro stations in the study area, and shows that these services only run between central Wolverhampton and Birmingham and therefore only cater to a small proportion of the West Midlands population. The locations of bus stops shown in Figure 3.2 show that these services are relatively ubiquitous throughout the study area, and although service provision will differ between areas, these services likely cater to a much greater proportion of the population than rail and metro services. The reasons behind the observed decline in bus boardings can therefore not solely be contributed to a modal shift towards other modes of public transport. With the West Midlands chosen as the UK's first 'Future Mobilities Area' (see Section 2.2.3.4) future modal shifts towards new, more sustainable modes of transport, such as electric bikes and car sharing, can also be expected. Understanding the mobility needs and demands of the population in the study area will be vital for authorities to ensure the success of these new modes of transport.

3.1.1.2 Private transport

In addition to changes in the share of public transport trips between modes, changes in private transport availability were also likely to have had a impact on the number of public transport trips and the modal share in the study area. To gain an understanding of the possible changes in private transport use, changes in the number of driving license holders were analysed. Table 3.1 shows the number of individuals holding a full driving license aged 60 and over in Great Britain. Data were available from 2012 to 2016.

Table 3.1 Proportion of the population aged 60 and over that are driving license holders in Great Britain from 2012 to 2016 (DfT 2016)

Year	Male	Female
2012	90.2	55.6
2013	90.4	56.7
2014	89.1	56.5
2015	90.0	58.5
2016	89.4	58.9

There was an increase of over 3% in the proportion of females aged 60 and over that hold a driving license between 2012 and 2016. This increase was most likely due to both the increased availability of private transport over this cohort's lifecourse and the increasing normalcy for women to be drivers over recent decades. Although the granularity of the data available meant it was not possible to deduce the exact increase in driving license holders in the study area, an increase in the proportion of female driving license holders would likely have impacted the number of trips made by both private and public transport services.

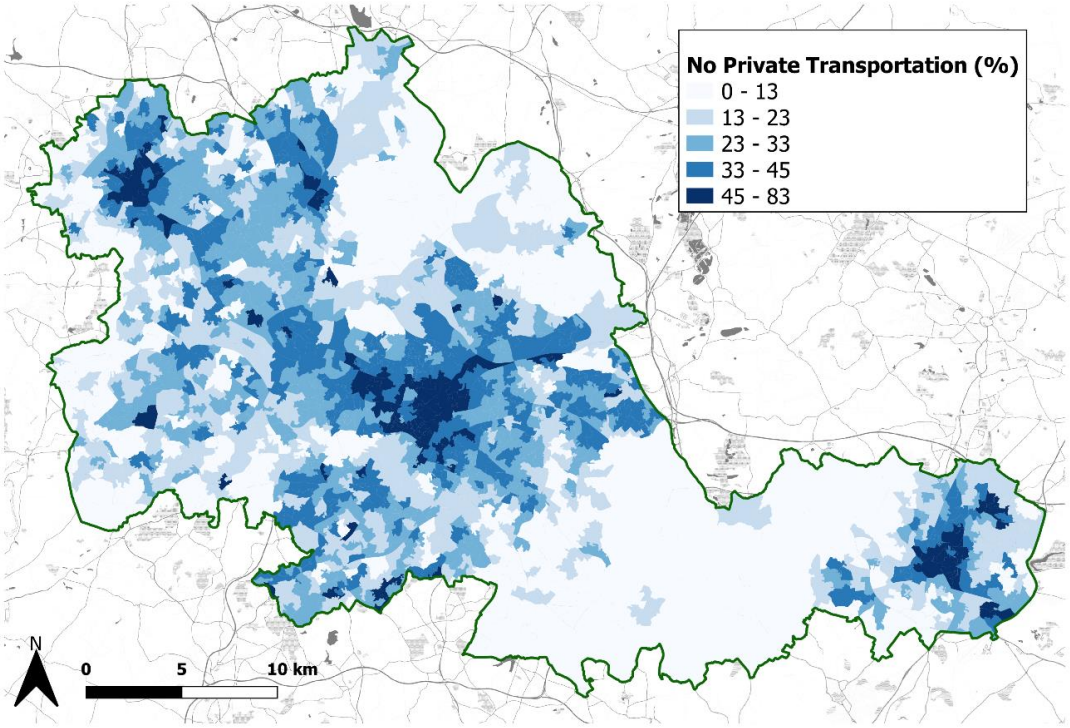


Figure 3.6 Percentage of the population with no access to private transportation at LSOA level (Census 2011; ONS 2018b)

In addition to driving license holders, access to private transport has a significant relationship with both the socioeconomics of the study area and the use of public transport services. Figure 3.6 shows the percentage of the population with no private transport availability at LSOA level. As would be expected, central urban areas generally had low levels of private transport availability, with public transport provision in these areas tending to be higher than more rural areas. Comparing this with the location of public transport services shown in Figures 3.2 and 3.3 supports this theory, with lower levels of private transport availability found in areas with a high concentration of public transport stops and stations. Additionally, the spatial patterns shown in Figure 3.6 coincide with deprivation, as shown in Figure 3.13, with those living in more deprived areas less likely to have access to private modes of transport. The relationship between deprivation and public transport use is explored in detail throughout this thesis.

3.2 Study population

3.2.1 Population structure and dynamics

Population structure and dynamics can provide insights into both the study population and the wider population in the study area.

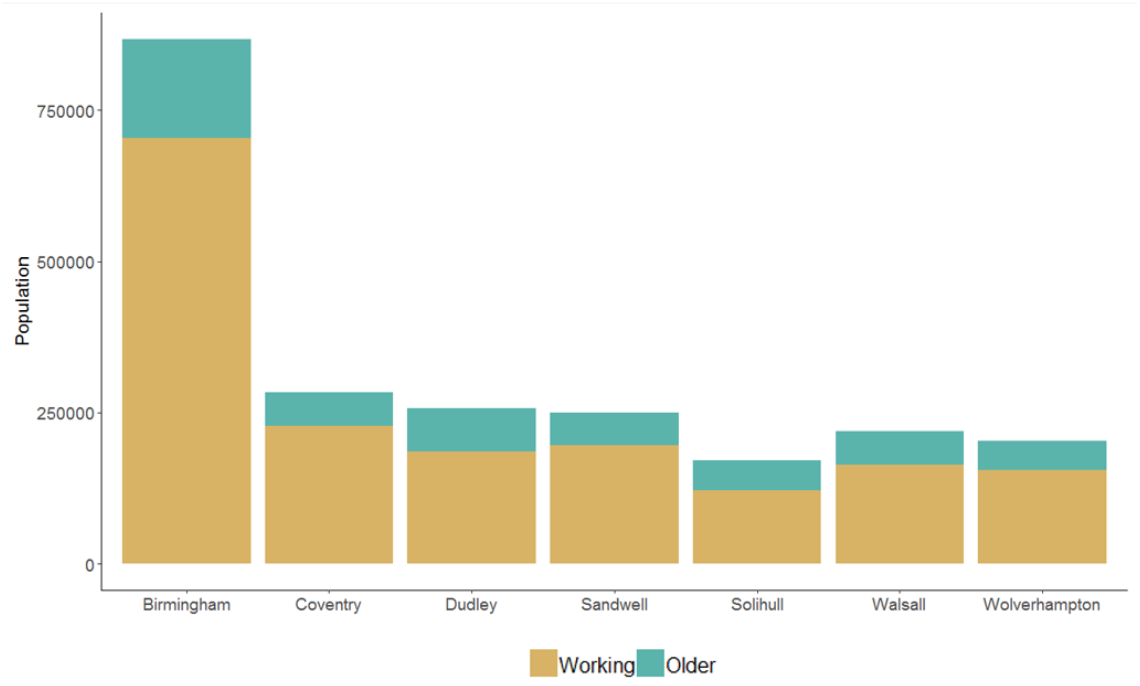


Figure 3.7 Number of older and working-age residents in each metropolitan borough (ONS 2018a)

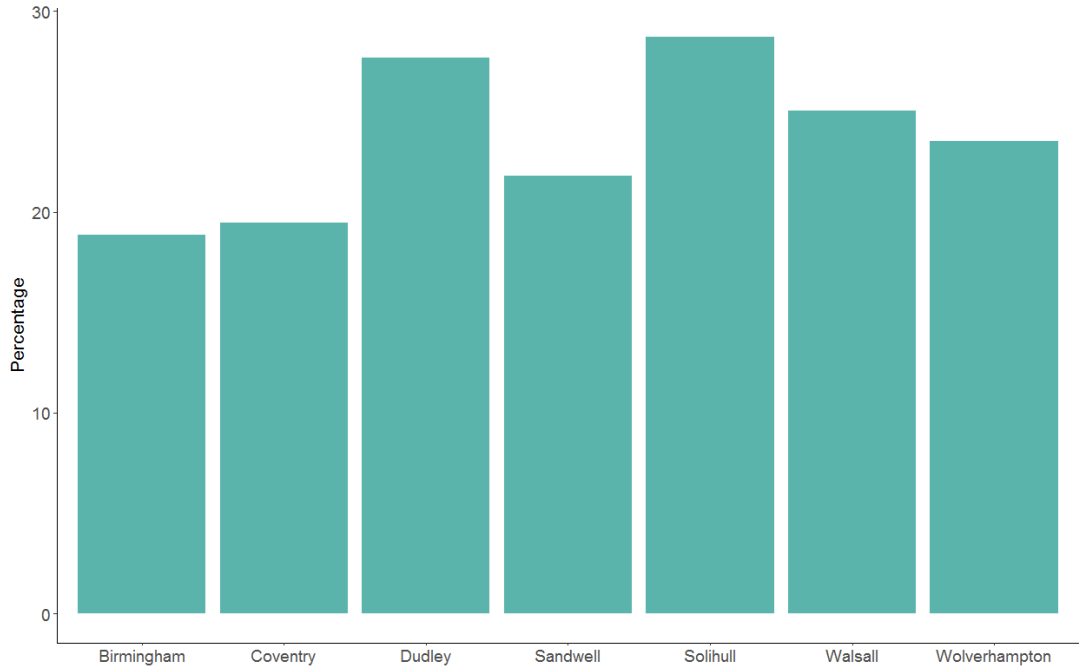


Figure 3.8 Percentage of the population aged 60 and over in each metropolitan borough (ONS 2018a)

Figure 3.7 shows the number of older and working-age residents in each metropolitan borough of the West Midlands. Birmingham had the largest population both in terms of older and working-age residents, whilst the remaining boroughs contained significantly fewer residents of both groups. Figure 3.8 shows that Birmingham and Coventry had the lowest proportion of population aged 60 and over; under 20%, in comparison to almost 30% for Dudley and Solihull. This is likely due to central urban areas attracting younger residents, with older residents more often found in suburban and rural areas, as can be seen in Figure 3.9.

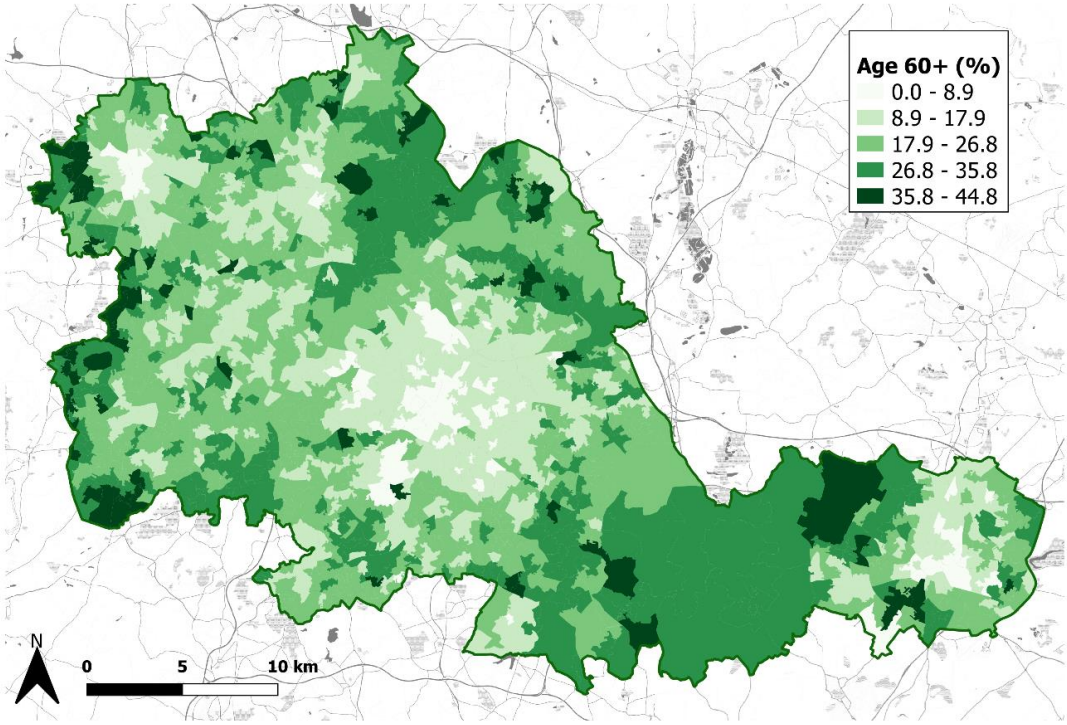


Figure 3.9 Percentage of the population aged 60 and over at LSOA level (ONS 2018a)

Figure 3.9 shows the percentage of the population that were aged 60 and over within each LSOA. The propensity of the older population to reside in suburban and rural areas may be expected to result in limited access to public transport services for some residents when compared to the locations of public transport stops and stations in Figures 3.2 and 3.3. However, when comparing the location of bus stops in the study area to the location of older residents, it was estimated that over 80% of the older population lived within a 5-minute (400m) walk of a bus stop. Physical access to public transport stops and stations may therefore not reduce significantly for rural residents, however rural and suburban areas typically have significantly lower service provision in comparison to urban centres. This may have a greater impact on public transport access for these populations. Public transport accessibility and service provision is explored in detail in Chapter 8 of this thesis.

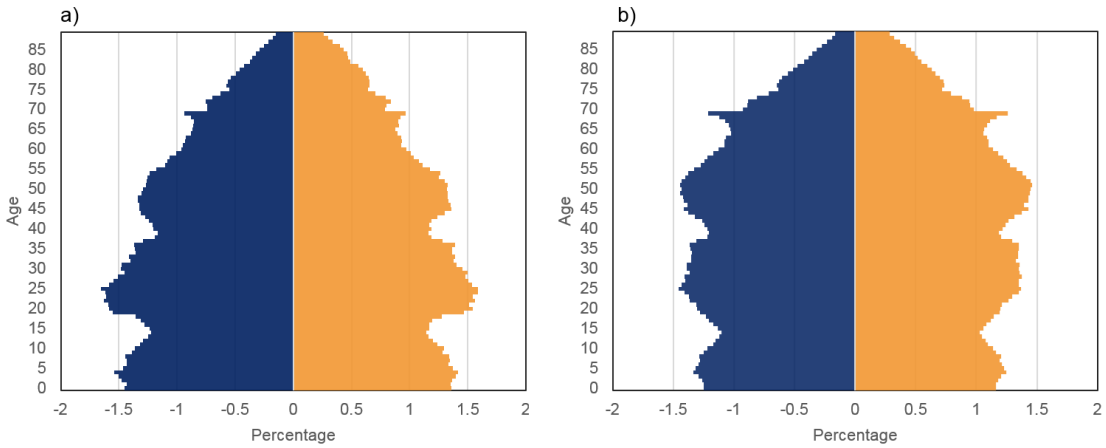


Figure 3.10 Population pyramids for a) the West Midlands and b) the UK (ONS 2018a)

Comparing the population structure of the West Midlands to the UK as a whole shows that the West Midlands contained a higher proportion of younger people, aged 20-25, and fewer older residents. This was particularly prominent for those aged 45-55 and those aged 65-70. This higher proportion of younger residents was likely due to higher population densities in urban areas, such as Birmingham and Coventry, which tend to attract younger populations, in addition to the 10 universities that are located in the area. Likewise, the lower proportion of older residents was likely due to the largely urban make-up of the study area, with around 70% of the land use being classed as urban (European Environment Agency 2017). Despite this, the West Midlands is not entirely urban, with Coventry being separated from the West Midlands conurbation by a large area of green belt land around 24 km wide, and a smaller stretch of green belt between Birmingham and Walsall. Figure 3.9 shows that this area contains a significant proportion of older residents.

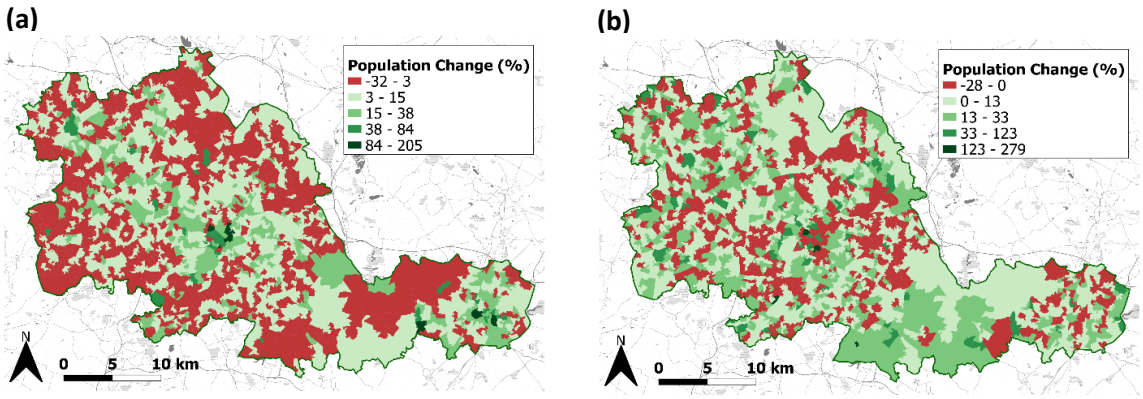


Figure 3.11 Percentage change in population (2009 – 2016) for a) the total population and b) the 60 plus population (ONS 2018a)

Figure 3.11 shows the population change between 2009 and 2016 for both the total and older populations. Focusing on the total population, the central urban areas of Birmingham, Coventry and Wolverhampton exhibited the largest population increases, with rural and suburban areas showing some areas of smaller increases and significant decreases. In terms of the 60 plus population, rural and suburban areas largely exhibited increases in older residents, in particular Solihull and Walsall. It is clear from this figure that the majority of LSOAs in the study area experienced an increase in population aged 60 and over; likely due to an ageing population rather than significant migration into the area.

3.2.2 Demographics and socioeconomics

The West Midlands is highly multicultural, with almost 30% of the population recorded as being from a non-white ethnic background in the 2011 Census. Differences have been observed in the modal choice and mobility of different ethnic groups (Schmocker et al. 2004) and therefore understanding the general distributions of these ethnic groups throughout the study area can provide contextual information when interpreting results in subsequent chapters.

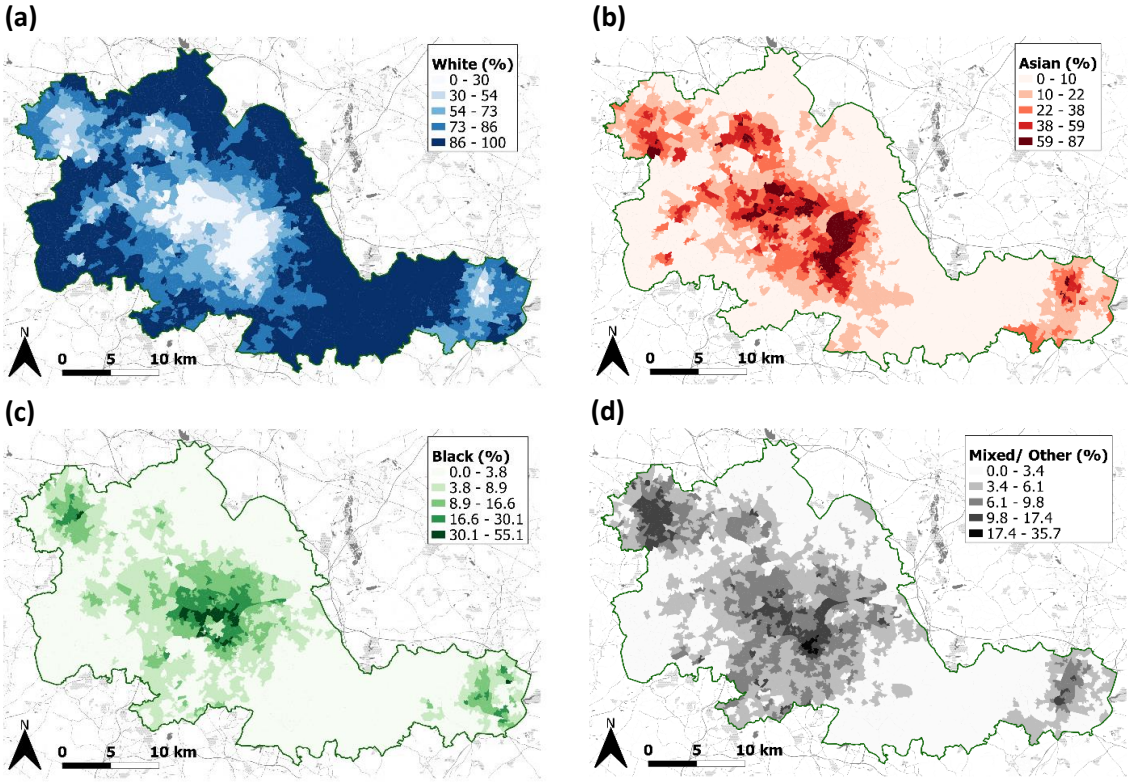


Figure 3.12 Percentage of the population that are of a) white, b) asian, c) black, d) mixed/ other ethnic background at LSOA level (Census 2011; ONS 2018b)

Figure 3.12 shows the distribution of the main ethnic groups in the study area. There tended to be lower proportions of residents from a white ethnic background in central urban areas, with rural areas, in particular Solihull and Dudley, showing the opposite. High proportions of the population in these urban centres appeared to be of an asian ethnic background, with lesser, but still significant, proportions of residents of black and other ethnic backgrounds.

The Index of Multiple Deprivation (IMD) is a government measure of deprivation. It covers seven aspects of deprivation: income, employment, health and disability, education, skills and training, barriers to housing and services, crime, and the living environment. Figure 3.13 shows which decile each LSOA in the study area was assigned to in 2015, with those assigned to Decile 1 being in the 10% most deprived LSOAs in England, and those assigned to Decile 10 being in the 10% least deprived.

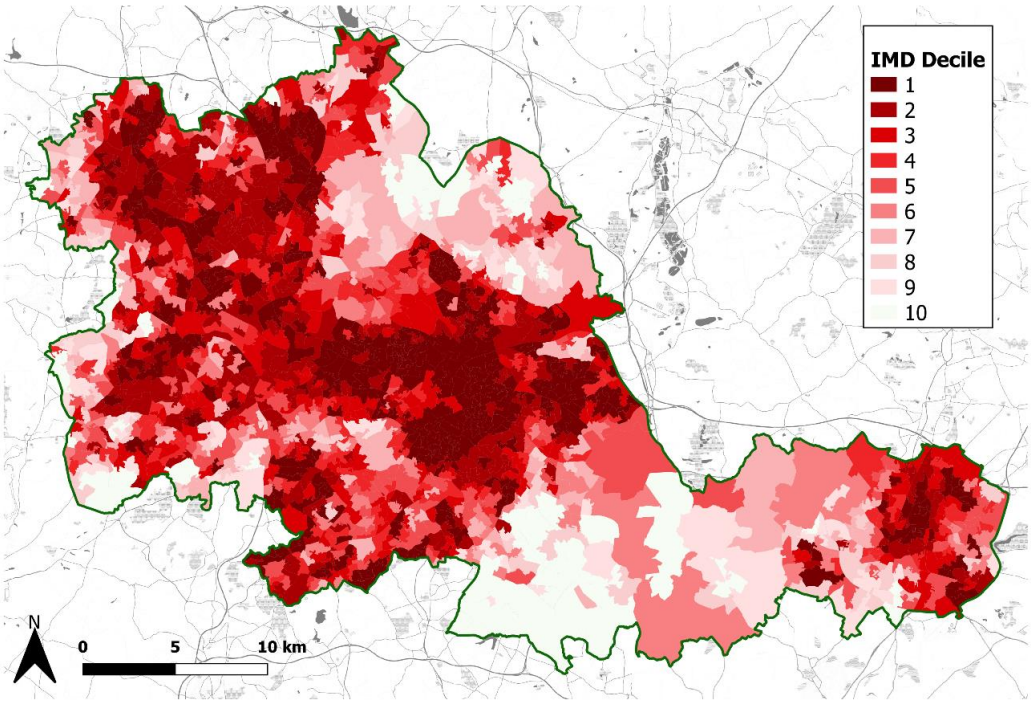


Figure 3.13 LSOA IMD deciles (1 = most deprived, 10 = least deprived) (ONS 2015)

Over 25% of LSOAs in the West Midlands belonged to Decile 1, in comparison to just 5% belonging to Decile 10. The most deprived areas appeared to be situated in central urban areas, whilst Solihull and north Birmingham contained many of the least deprived LSOAs. Comparing this to the ethnic group distributions shown in Figures 3.12, the least deprived areas also contained high proportions of residents from a white ethnic background, with those from ethnic minority backgrounds more likely to live in more deprived LSOAs. The relationship between

mobility patterns and the demographics and socioeconomics of cardholders is explored throughout this thesis, and therefore these figures provide context for the insights produced in later chapters.

3.3 Summary and conclusions

The aim of this chapter was to provide an overview of both the study area and the study population to provide context for the results of subsequent chapters. Although this thesis focuses primarily on bus travel and the bus network throughout the West Midlands, it was important to highlight the fact that metro and rail travel also account for a number of public transport journeys. It was also necessary to highlight that whilst rail and metro ridership has been increasing, the lack of stations throughout the study area, in particular for metro services, means that the bus is the primary mode of public transport for a large proportion of the population in the West Midlands. Although decreasing, in 2016 bus travel still accounted for 80% of all public transport journeys undertaken in the West Midlands. These insights were not specific to the older population, however, Chapter 4 explores trends in the bus travel of older concessionary cardholders in further detail.

In addition to public transport in the study area, providing contextual information relating to the population structure can help to explain trends that may be found in subsequent chapters, in particular differences between areas in terms of the number of older residents, and population changes that may have contributed towards changing bus usage patterns. Section 3.2 explored population structure in the study area, and highlighted differences in the proportion of older residents between each of the metropolitan boroughs and between central urban areas, suburban and rural areas. These insights are vital for interpreting bus usage patterns of older concessionary travellers, with the numbers of eligible concessionary travellers residing in areas likely to have a significant impact on the number of bus trips undertaken using ENCTS cards.

Throughout the literature review undertaken in Chapter 2, the importance of deprivation in terms of public transport usage and the risk of social exclusion was highlighted. Figure 3.13 showed the distribution of IMD deciles throughout the West Midlands. This emphasised the variation in deprivation throughout the study area. This likely has a significance impact on both the travel behaviours and the mobility levels of individuals. The insights generated in this chapter are referred to throughout this thesis in order to contextualise the results produced and explain the trends and patterns identified.

Chapter 4 – Smart card data: Overview and preliminary analysis

4.1 Introduction

The aim of this chapter is to provide an overview of the smart card data utilised in this thesis, including the representativeness and potential bias of these data and the steps taken to prepare the data for further analysis. The smart card data used throughout this thesis were provided by TfWM. These data presented a unique opportunity for the analysis of large commercial data for academic research, as data are often unavailable outside of commercial settings due to the risk of the disclosure of personal or commercially sensitive information.

Access to the data was possible through the Economic and Social Research Council (ESRC) funded Consumer Data Research Centre (CDRC), located at UCL's Department of Geography. The CDRC is a government funded big data initiative that aims to facilitate the access of commercially generated consumer datasets to academic researchers. The CDRC holds data at three access levels: open data, safeguarded data and controlled data. As the smart card data are personal in nature, relating to identifiable living individuals (GDPR 2018), these are classed as controlled data under CDRC regulations. Controlled data must be held under secure conditions with stringent access restrictions and project approval requirements. To minimise the risk of the disclosure of personal or commercially sensitive data, a number of procedures are necessary (see the CDRC User Guide 2018). As this thesis utilised both personal and commercially sensitive data, data access, handling and presentation of results had to be conducted in line with these procedures, which are outlined in Figure 4.1.

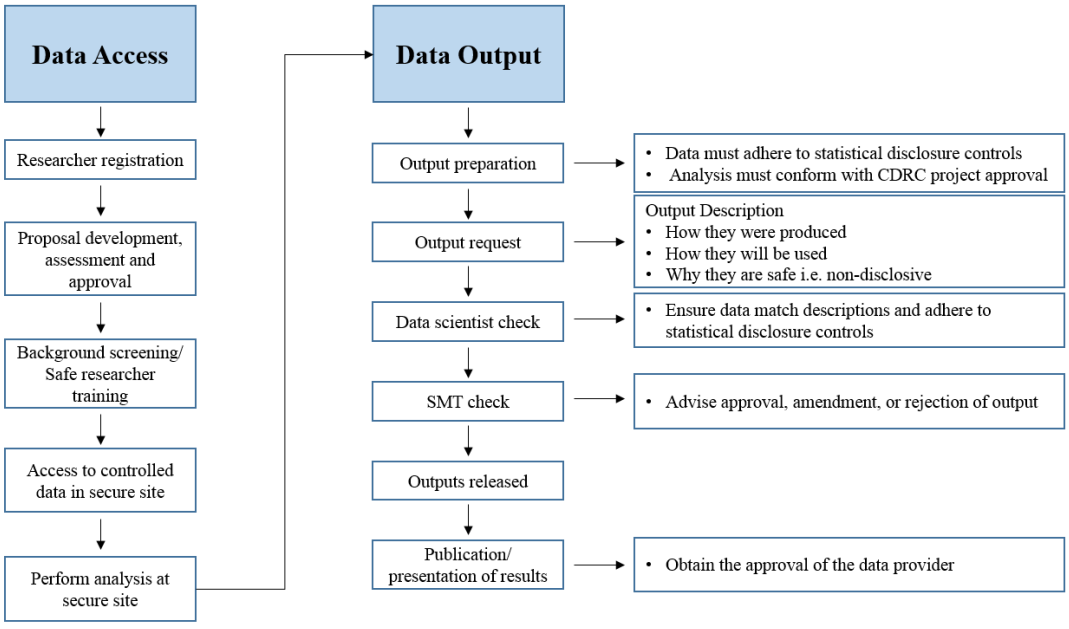


Figure 4.1 Procedures to access, analyse, output and present CDRC 'controlled' data

Access to the data was available through the CDRC's secure data service – the Jill Dando Institute Research Laboratory (JDIRL) secure facility, located at UCL. Before data could be accessed, preliminary vetting and training procedures had to be undertaken to ensure that access was only granted to trusted researchers. Researchers also had to obtain approval for all proposed uses of the data and all analyses had to be conducted within the JDIRL. Outputs to be taken out of the secure facility had to conform to a number of statistical disclosure controls, including aggregation to large geographic areas, the suppression of disclosive cells, ensuring percentages did not allow for the deduction of disclosive units, and a threshold rule of no less than 10 for counts. These disclosure controls conform to government specified regulations regarding the handling of personal or disclosive data (GDPR 2018).

Once outputs were requested to be taken from the secure facility, two CDRC data scientists were assigned to check that each output matched the description given on the output request form and that these outputs adhered to disclosure controls. The final step in conducting analyses and obtaining outputs from the JDIRL was that two members of the CDRC Senior Management Team reviewed the output requests and advised on the approval, amendment or rejection of these outputs. Once outputs were approved, any future publication or presentation of analyses had to be approved by the data provider for commercial disclosure control purposes.

In order to adhere to both statistical and commercial disclosure controls, the results presented in this thesis were constrained as necessary. A detailed overview of the specific data treatment measures undertaken, such as spatial aggregation, are outlined later in this chapter.

4.2 Data overview

The data provided by TfWM consisted of the main smart card transaction dataset, in addition to cardholder data, automated vehicle location data and bus timetable data. The first transaction in the dataset was recorded at 09:30 30/11/2009, with the last transaction recorded at 23:58 18/08/2016. Although transaction records were available from this first transaction time, it should be noted that full transaction records were not available until September 2010, in correspondence with the move from paper tickets and passes to smart cards. The cardholder data included information on the age group, sex, registered LSOA, and the type of concession (i.e. 'Over 60' or 'Disabled') for each cardholder. Each record also contained a unique account number, allowing for linkage to the transaction records. The automated vehicle location data contained timestamped GPS data of bus locations on the network, allowing for data to be linked to the transaction data and the bus timetable data. This linkage was conducted previous to this thesis by researchers at the Centre for Advanced Spatial Analysis (CASA), UCL, as part of the 'Inclusive and Healthy Mobility' project, also sponsored by TfWM

(see *cdrc.ac.uk/research/healthymobility*). An overview of the method used for this linkage is presented in Section 4.2.4.1.

These data represented a structured form of big data, with clear variables and matching IDs between data for linkage purposes. Despite the structured nature of these data, a number of errors and data quality issues were still present, in particular incompleteness of data and missing attributes. The cleaning methods used to deal with these issues are presented throughout this chapter.

4.2.1 Transaction data

The transaction data contained all bus boardings recorded by concessionary cardholders from November 2009 to August 2016. These data contained transactions made by both 'Over 60' and 'Disabled' concessionary smart card holders, however this study only utilised data recorded by 'Over 60' concessions. It should be noted that regardless of whether a cardholder had previously held a 'Disabled' concessionary smart card, once a cardholder was eligible for an 'Over 60' concessionary card, the smart card type was changed to 'Over 60'. In total, these data consisted of 308,529,869 transactions made by 463,231 'Over 60' cardholders.

Transaction variables included a unique account number and card ID, transaction timestamp (recorded as DD/MM/YYYY HH:MM), the service number and the operator code of the bus boarded. These data were built on the ITSO standard, which provides a common standard by which smart cards and automated fare collection may be administered (see *itso.org.uk*). As many unique account numbers were associated with two or more card IDs, likely due to renewals and replacements of cards, data linkages and analyses were conducted using account number only. The structure of each of the variables included in the transaction data is shown in Table 4.1.

Table 4.1 Example transaction data (account numbers and card IDs censored to ensure anonymity)

Record ID	Account number	Card ID	Transaction timestamp	Service Number	Operator Code
4563445	2******45	7*******34	10/12/2014 09:38	37	355
6654452	2******77	7*******56	02/07/2015 14:56	6a	640
5783234	2******33	7*******12	28/01/2011 19:02	244	410

Although the first transaction was recorded in November 2009, full transaction records were not available until September 2010. Table 4.2 shows the number of transactions and the number of active users (defined as having made at least one transaction) recorded for each year of the study period.

Table 4.2 Total yearly transactions and active users

Year	Transactions	Active Users
2009	16,867	5,138
2010	20,602,914	223,739
2011	53,667,606	223,507
2012	53,286,499	226,344
2013	52,728,576	223,043
2014	53,917,557	215,937
2015	48,766,303	199,597
2016	25,543,545	171,135

4.2.2 Cardholder data

The cardholder data variables included a unique account number, age group (5-year age bands), sex, registered LSOA and card type. In total, there were 463,231 smart card holders present in the study area, however there were a number of missing variables and discrepancies between attributes. For example, several cardholders were recorded as holding 'Over 60' concessionary smart cards but were recorded as being less than 60 years of age. Records such as these, where there was a clear discrepancy between attributes, were removed from subsequent analyses. In addition, a number of records were also missing data regarding sex and registered LSOAs. Although these records remained in the data, for analyses that required either of these missing variables, these records were omitted. In total, 11 cardholders displayed discrepancies between the type of concession and their recorded age group, 62,001 cardholders were missing data regarding sex and 4,611 regarding registered LSOA. With around 13% of cardholders missing 'sex' data, this would cause issues when presenting results of analyses using these data as raw counts. To remedy this, analyses using 'sex' data were limited and for results where the sex of subjects were discussed, these were presented as proportions.

4.2.3 Bus timetable data

TfWM also provided bus timetable data. These data contained information on the order of the stop along the route ('n'), the route number and direction ('I' for inbound or 'O' for outbound) that each bus runs, the NaPTAN code, i.e. bus stop code, for each scheduled stop (see Table 4.4), the journey start time, scheduled arrival and departure times, and the day of the week for which the service was scheduled. An extract of these data is shown in Table 4.3.

Table 4.3 Extract of the bus timetable data

n	Route	Journey scheduled	NaPTAN code	Arrive	Depart	Day of week	Direction
1	48	09:35:00	43000234101	09:35:00	09:35:00	1111100	O
2	48	09:35:00	43000230601	09:38:00	09:38:00	1111100	O
3	48	09:35:00	43002230102	09:39:00	09:39:00	1111100	O
			……				
1	6A	17:29:00	43001056101	17:29:00	17:29:00	0000001	I
2	6A	17:29:00	43000075302	17:31:00	17:31:00	0000001	I
3	6A	17:29:00	43001056002	17:34:00	17:34:00	0000001	I

4.2.4 Automated Vehicle Location data

The AVL data provided by TfWM contained information on the location of buses on the network using GPS and was used for linking smart card transactions to bus services. Unlike the smart card transaction data, the vehicle location data only ranged from 13/01/2014 to 18/08/2016. Attributes included a unique identifier for each bus, the timestamp at which the recording was made, and the longitude and latitude of the recording. Several analyses in this thesis relied on these AVL data, and therefore were only conducted over the time period for which AVL data were available. Where this is the case, these are clearly identified.

4.2.4.1 Origin and destination inference

The transaction data used in this study contained no locational information relating to the origins and destinations of trips made on the bus network and therefore transactions had to be linked to the AVL data to infer these locations. This analysis was conducted as part of a

previous study by Dr Alistair Leak and Dr Jens Kandt, located at CASA, UCL, as part of the 'Inclusive and Health Mobility' collaboration between the CDRC and TfWM. This section gives a brief overview of the method that was used in this study to infer the origins and destinations of the smart card transactions.

The service number and timestamp recorded in the smart card transaction data allowed for these to be linked to the AVL data, and the service on which each transaction had taken place to be determined. As the AVL data also contained information on the bus service, locational information and the timestamp that the location was recorded, the location of each transaction could also be determined. This was then linked to bus stop location data to determine the stop at which the transaction most likely took place, i.e. the origin of the trip.

Once the origin of each trip was determined, the destination of each trip could be inferred by analysing the location of the subsequent boarding that took place. This method relied on the assumption that a cardholder would alight a service at the nearest stop to which they would then board a subsequent service. Figure 4.2 shows an example of this method.

Figure 4.2 Example origin and destination inference method, with three bus routes shown

In the example shown in Figure 4.2, the cardholder firstly boarded at Location A, at 14:26. The next information provided through the AVL and transaction linkage is that the cardholder then

boarded at Location B at 14:51. From this information, the destination was inferred as the bus stop on Route A, shown in blue, that was in closest proximity to the boarding Location B. The same method was utilised for inferring the Location C alighting and boarding locations. In this example, no further transactions were recorded after the boarding at Location C, so this was the last known location of the cardholder.

In addition to the potential errors that would occur due to multi-modal trips, where a cardholder may undertake part of a trip using a mode of transport other than a bus, another limitation of this method is that it relied on a subsequent trip being undertaken for the destination to be inferred. The last journey in each trip was therefore not assigned a destination. This limited the number of transactions for which an origin and destination could be inferred to around 38 million for the 2014 to 2016 period, which equated to 27.5% of all transactions. Although this meant that a large number of transactions were not included in analyses which utilised origins and destinations, it was predicted that many of these trips were the result of a cardholder returning home having undertaken an activity and therefore the trips of interest for these analyses would still be included.

4.2.5 Supporting Data

4.2.5.1 Census geographies

Census geographies describe the subdivision of geographical areas for the purpose of the collection of population statistics for the UK Census. The main geographies associated with the Census are Output Areas (OAs) and Super Output Areas (SOAs). OAs are the base units for Census data releases, with SOAs representing larger geographical areas at which OAs can be aggregated. SOAs are divided into two geographic scales: Lower Layer Super Output Areas (LSOAs) and Middle Layer Super Output Areas (MSOAs).

Whilst many analyses presented in this thesis were initially conducted at the individual level, in order to adhere to statistical disclosure controls (see Section 4.1), spatial aggregations were necessary for the presentation of outputs. As the data included the registered LSOA of each cardholder, spatial aggregations were conducted primarily at LSOA level to allow for the linkage of these data to existing national statistics.

There are 1680 LSOAs in the West Midlands and all contained registered cardholders; with a minimum of 7 and a maximum of 631. Figure 4.3 shows the Location Quotient (LQ) of cardholders in the study area at LSOA level. The LQ shows how well represented cardholders were in comparison to underlying population volumes, obtained from 2016 ONS mid-year population estimates. This was calculated as the percentage of the cardholder population

residing in each LSOA divided by the percentage of the total West Midlands population residing in each LSOA. An LQ of 1 meant that an LSOA had an identical share of the population in comparison to the reference data, with a lower value indicating a lower share and a higher value indicating a higher share.

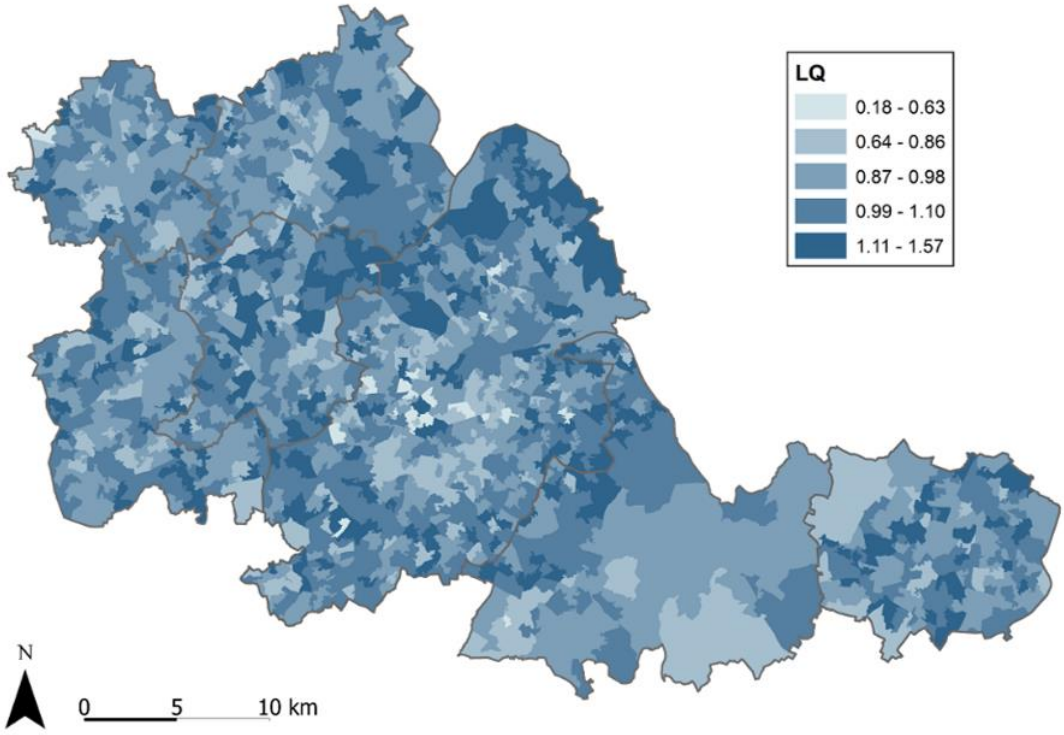


Figure 4.3 Location quotient of smart card holders

The LQ showed a relatively even distribution of cardholders throughout the study area, with the majority of LSOAs having an LQ between 0.81 and 1.12. This is unsurprising as over 80% of the older population were estimated to be cardholders. Despite this, rural areas, such as much of Solihull, appeared to have a slightly lower representation than more urban areas, in particular large areas of Birmingham and Coventry. This may be due to the lower density of bus stops and therefore likely lower bus usage in these areas (see Figure 3.2). Central Birmingham and Coventry showed areas of both high and low LQ, likely due to the generally lower number of older people living in central urban areas (Baernholdt et al. 2012) meaning any differences in representations will be magnified.

4.2.5.2 Contextual data

In addition to the primary consumer datasets analysed in this thesis, a number of other datasets such as census and travel survey data were also utilised. These are summarised in Table 4.4.

Table 4.4 Descriptions of contextual data

Data	Description	Source
ONS Mid-Year Population Estimates	Mid-year population estimates were obtained from the ONS and provided annual population estimates at LSOA level. Data were available for each year of the study period and in one-year age bands.	ONS (2018a)
2011 Census Variables	Census variables used for analysis included those relating to demographic structure and socioeconomic characteristics. These data were obtained from the ONS, and primarily used at LSOA level.	ONS (2018b)
National Travel Survey	The National Travel Survey is a yearly household survey, providing data on personal travel and changes in travel behaviour over time. Data were able to be filtered by metropolitan region and age group.	Department for Transport (2018)
National Public Transport Access Nodes (NaPTAN)	NaPTAN is the national dataset of public transport access points (e.g. bus stops, railway stations, airports). These data provided the locations of bus stops and other public transport stops and stations in the study area.	Department for Transport (2014)

These data were utilised throughout this thesis in order to draw comparisons with trends and contextualise insights. Additionally, these traditional data were used to enrich the consumer data, for example linking demographic and socioeconomic characteristics obtained from census data to the cardholder data via the registered LSOA variable.

4.3 Representativeness, bias and general trends

The smart card data utilised in this thesis were collected for the primary purpose of fare collection on the bus network. Utilising these data for reasons other than their original purpose

can present a number of issues due to the complete absence of researcher control in the data collection process. It was therefore imperative that the quality and representativeness of the data were understood prior to any analyses being undertaken. This included understanding the populations that were represented in the data and any potential bias or errors, as well as assessing the plausibility of observed trends.

4.3.1 Cardholder attributes

Common issues in research using consumer data are related to self-selection, whereby customers select themselves to participate (i.e. to obtain a concessionary smart card) and therefore represent an inherently biased sample (Heckman 1990). In order to understand the extent of this bias, the cardholder data were analysed alongside census variables and mid-year population estimates.

To understand how representative the data were in terms of the total eligible population in the study area, the age and sex attributes in the cardholder data were compared to 2016 ONS mid-year population estimates. Both the cardholder data and the census data were normalised by the total population, to account for underlying base populations and allow for comparisons between proportional age and sex distributions.

Of those cardholders with recorded sex attributes, approximately 44.7% of cardholders were male and 55.3% were female. Compared to the mid-year population estimates of 45.8% male and 54.2% female, this suggested that the distribution of sexes in the smart card data was highly representative of the actual population structure in the study area.

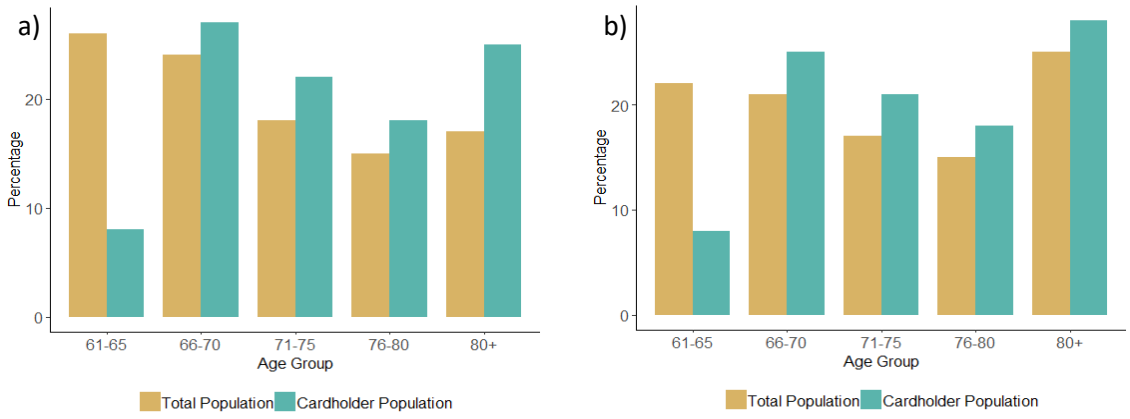


Figure 4.4 Age distributions for the a) male and b) female total and cardholder populations (5-year age bands)

Comparing the age and sex distributions of smart card holders to mid-year population estimates showed that the smart card data were under-representative of those in the 61-65 age group, and over-representative of those aged over 80, particularly for the male population. The primary reason for this over-representation of the over 80 population was likely due to mortality. Although TfWM aim to keep track of this, with some passes being sent back after a cardholder has died, this over-representation suggests that there may have been users recorded in the system who were deceased.

The under-representation of the 61-65 population is likely a result of an increase in the age of eligibility for concessionary travel, from 60 to 65 between 2010 and 2020. The effects of this on the trends found in the smart card data are analysed in detail in Section 4.4. In addition, younger age groups are more likely to have access to private transport and to be more mobile than those that are older (Kim and Ulfarsson 2004). This means that this population may be less likely to rely on public transport and therefore less likely to be ENCTS cardholders than older age groups.

4.3.2 Smart card usage

One of the main motivations behind this research was the significant decline in bus boardings by cardholders recorded by TfWM. TfWM recorded a decline in both concessionary and non-concessionary bus boardings of around 25% and 20% respectively between 2009 and 2016. Figure 4.5 shows that this decline in concessionary boardings could also be seen in the smart card data. Understanding this decline, i.e. which cardholders were making use of the bus network, which cardholders may have dropped out of the system, as well as changes in travel behaviour, is imperative for both transport authorities, operators, and the wider transport community. This decline in bus usage is not limited to the West Midlands but reflects a wider change in travel behaviour seen throughout the UK (Urban Transport Group 2019).

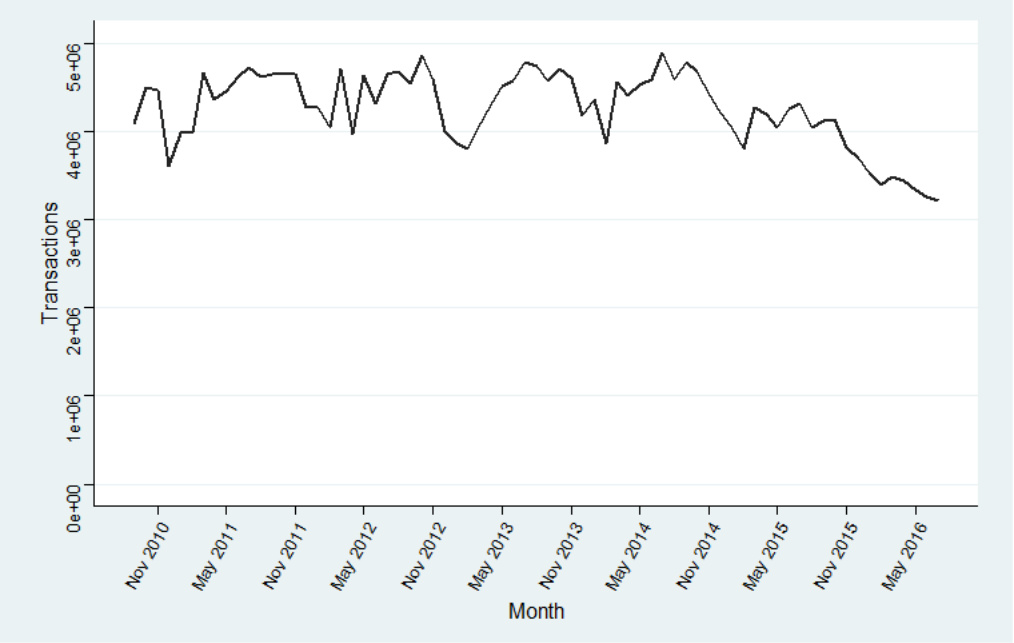


Figure 4.5 Monthly smart card transactions

In order to determine whether this decline was caused by a decrease in users making use of the bus network, or a decrease in the number of trips that cardholders were undertaking, Figure 4.6 shows the total number of active users per month over the study period. In this instance, a 'monthly active user' referred to a cardholder that had made at least one transaction that month.

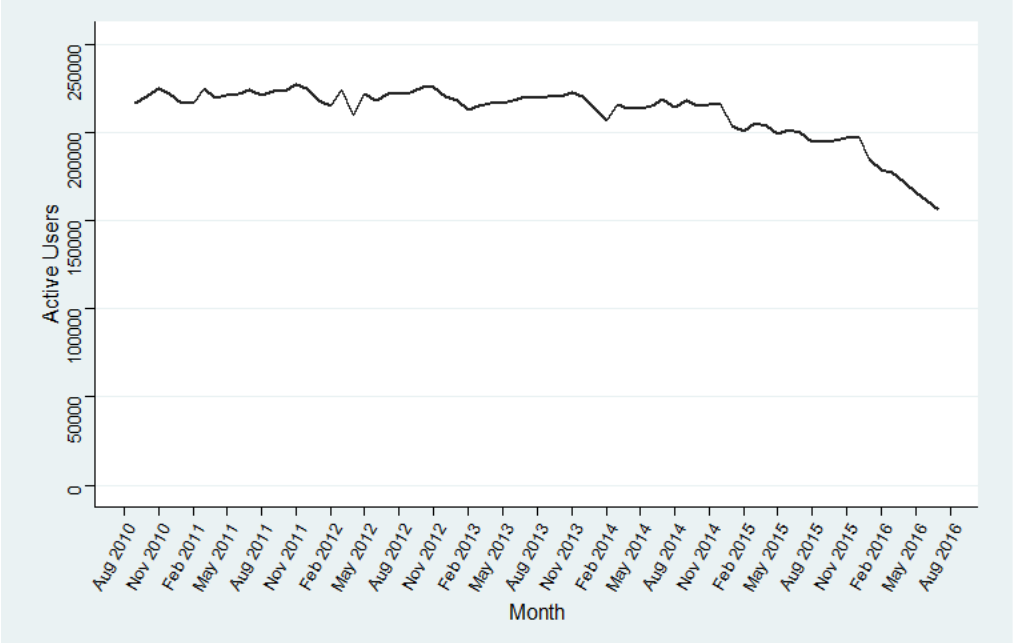


Figure 4.6 Monthly active users

There was a clear decline in active users over the study period, from 216,658 in September 2010 to 156,456 in July 2016. Taking into account only those years where full transaction records were available (2011 – 2015), the average number of transactions per cardholder did not show a significant change. This suggested that the decline in concessionary transactions made on the bus network was likely due to a decline in the number of individuals making use of the ENCTS rather than a general decline in the number of trips being made by cardholders. Chapters 5 and 6 of this thesis explore this decline in usage further; firstly, identifying which cardholders have contributed to this decline, and secondly, the kinds of trips that have declined. Additionally, the following section explores how a change in the age of eligibility for concessionary travel may have caused a decline in the number of older concessionary cardholders.

4.4 Adjusting for the change in eligibility

The age of eligibility for concessionary travel in the West Midlands is tied to the retirement age for women which, between 2010 and 2020, increased from 60 to 65 years. This increase affected the number of individuals that became eligible for concessionary travel each year, and therefore the number of cardholders entering the system. It was therefore hypothesised that the 5% greater decline in bus boardings by the concessionary population compared to the non-concessionary population was a result of this rise in the age of eligibility for concessionary travel rather than a behavioural change specific to the older concessionary population. To understand how changing eligibility, as well as general population changes, may have influenced bus boardings this section utilised two main datasets: bus patronage data published by TfWM and ONS mid-year population estimates.

While the actual age in eligibility for the ENCTS has increased at several points each year, the temporal resolution of the mid-year population estimates limited the specificity of the eligibility estimates used in this analysis. In light of this, for the following analyses, the age of eligibility was instead increased by 0.5 years each year from 2010. Table 4.5 illustrates this staged increase in the age of eligibility.

Table 4.5 Staged increase in the age of eligibility

Year	Age of Eligibility
2009	60
2010	60
2011	60.5
2012	61
2013	61.5
2014	62
2015	62.5
2016	63

4.4.1 Eligible population

The first step in understanding how the increase in the age of eligibility may have impacted the number of bus boardings was to examine how the size of the eligible population had changed over the study period. Table 4.6 shows the size of the eligible population between 2009 and 2016, in addition to the year-on-year percentage change. Eligible populations were calculated from ONS mid-year population estimates, with half-years calculated as 50% of that year's full population size.

Table 4.6 Eligible population in the study area from 2009 – 2016 (ONS 2018a)

Year		2009	2010	2011	2012	2013	2014	2015	2016
		545,446	*551,000*	*542,586*	*533,576*	*524,137*	*515,689*	*506,028*	*496,795*
2009	*545,446*								
2010	*551,000*	1.0							
2011	*542,586*	-0.5	-1.5						
2012	*533,576*	-2.2	-3.2	-1.7					
2013	*524,137*	-3.9	-4.9	-3.4	-1.8				
2014	*515,689*	-5.5	-6.4	-5.0	-3.4	-1.6			
2015	*506,028*	-7.2	-8.2	-6.7	-5.2	-3.5	-1.9		
2016	*496,795*	-8.9	-9.8	-8.4	-6.9	-5.2	-3.7	-1.8	

Before the introduction of an increase in the age of eligibility, there was around a 1% year-on-year increase in the population eligible for concessionary travel due to an ageing population in the study area. Since its implementation, the increase in the age of eligibility resulted in a decrease in eligible population of between 1.5% and 1.9% per year, and a total decrease of almost 10% from 2010 to 2016. This decrease in population eligible for concessionary travel would no doubt have impacted the number of boardings made by concessionary cardholders.

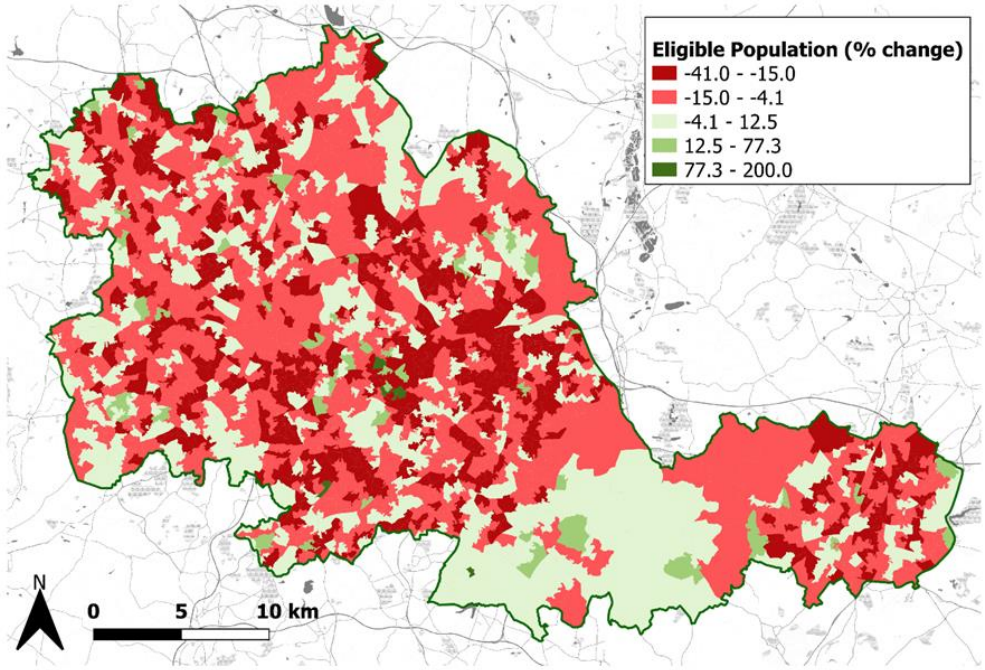


Figure 4.7 Percentage change in eligible population between 2009 and 2016 at LSOA level (ONS 2018a)

Figure 4.7 shows that the decrease in eligible population was not ubiquitous throughout the study area, with some areas instead experiencing an increase. Urban areas, in particular central Birmingham, displayed more significant changes than more rural areas. This was likely due to the higher population churn in central urban areas (Baernholdt et al. 2012). Excluding this, there was no clear spatial pattern in the areas that have experienced a decline or increase in eligible population.

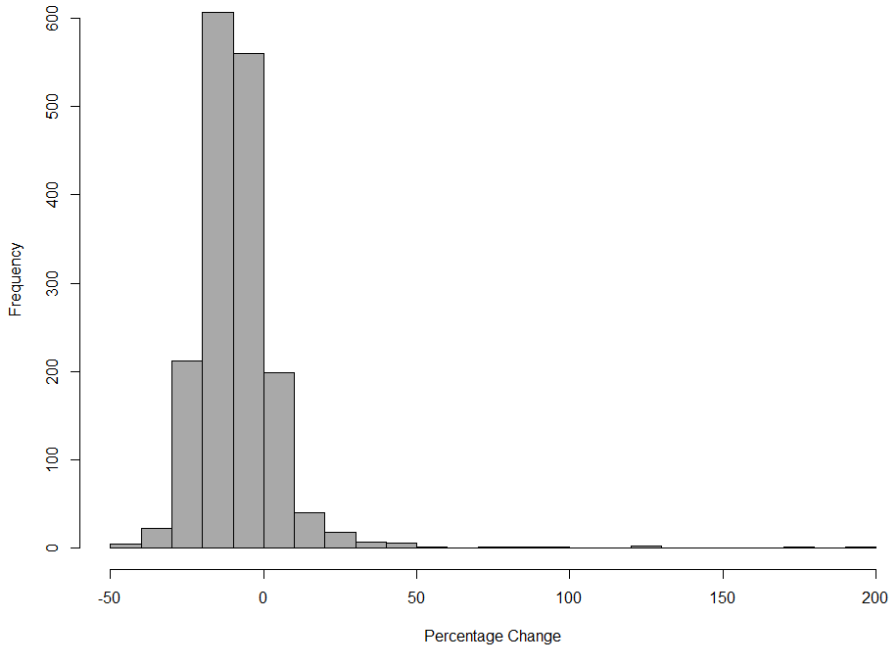


Figure 4.8 Histogram of percentage change in eligible population between 2009 and 2016 at LSOA level (ONS 2018a)

Figure 4.8 shows a histogram of the percentage change in the eligible population at LSOA level in the study area and appears to conform to a fairly normal distribution. It is also clear that the majority of LSOAs (~80%) experienced a decline in the population eligible for concessionary travel.

4.4.2 Impact on concessionary bus boardings

Understanding how this decline in eligible population may have contributed to the recorded decline in boardings was an important contextual element of this thesis. Demonstrating that the greater decline in bus usage by concessionary cardholders could be explained by an increase in the age of eligibility would evidence that there had not been a behavioural change unique to the concessionary population, and therefore insights generated into changing bus usage patterns later in this thesis could be applicable beyond just the older population.

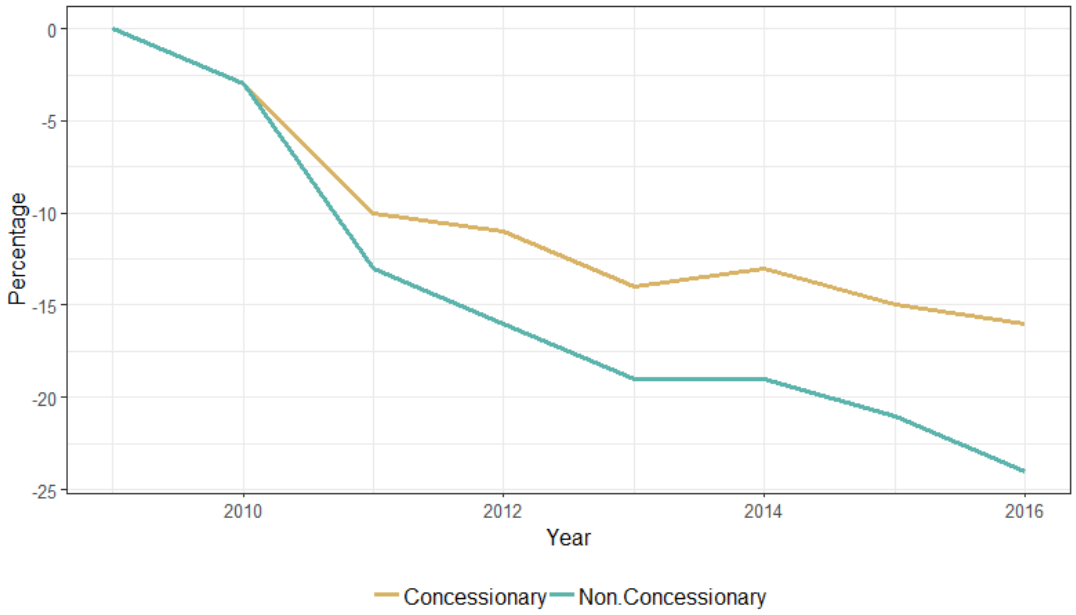


Figure 4.9 Percentage change in number of bus boardings weighted by population

Figure 4.9 shows the percentage change in the number of boardings made by concessionary and non-concessionary bus users, weighted by the population size each year. This weighting significantly altered the trends shown in Figure 4.5. Concessionary bus usage showed a decrease of 16% rather than the previously unweighted 25% and, with an increase in population not eligible for concessionary travel, bus usage by non-concessionary travellers showed a decline of 24% compared to the unweighted 20%. This suggested that usage of the bus network had in fact declined more significantly for those that were not eligible for concessionary travel than those that were. This highlighted the importance of considering changes in population sizes when interpreting trends in the data.

To see if there was a statistical correlation between the number of concessionary boardings and the size of the eligible population, the Pearson's correlation coefficient was calculated. This is a measure of the strength of association between two variables. This resulted in a Pearson's correlation coefficient of 0.94, as shown in Figure 4.10, which suggested a high correlation between the two variables. The p-value of 0.00053 also meant that this relationship was statistically significant.

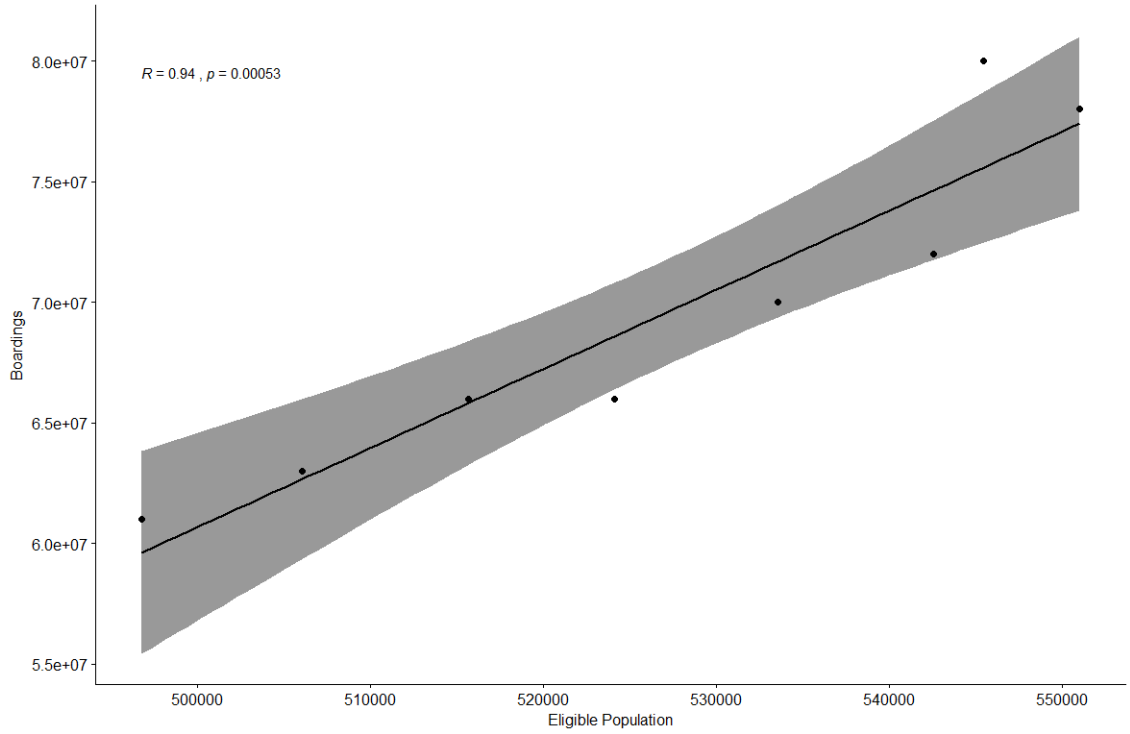


Figure 4.10 Pearson correlation between the number of concessionary boardings and the size of the eligible population

Despite this, Figure 4.9 suggests there was still a significant decline in bus usage by concessionary bus users that could not be explained by an increase in the age of eligibility. This is investigated in greater detail throughout this thesis.

4.4.3 Future trends

Previous analyses focused solely on the study period; however, the age of eligibility will continue to rise until it reaches 65 in 2020, and then again to 67 between 2026 and 2027. This increase is likely to have a significant impact on the proportion of the population that are eligible for concessionary travel in the study area, and as a result, the number of bus boardings made by those using ‘Over 60’ concessionary smart cards.

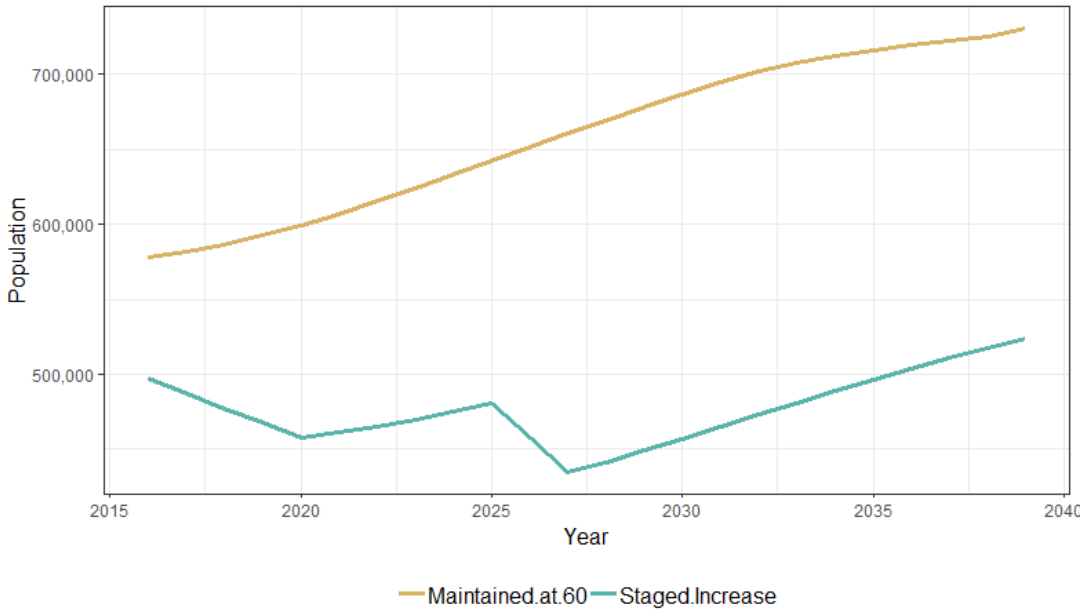


Figure 4.11 Predicted change in the size of the eligible population between 2016 and 2039 (ONS 2018a)

As illustrated in Figure 4.11, if the age of eligibility was to remain at 60 until 2039, there would be an increase in population eligible for concessionary travel of over 200,000 compared to 2016. With the staged changes in eligibility age, this number was instead only predicted to be around 50,000. As the ENCTS is funded by the government, accounting for an ageing population and the effect this will have on the eligible population in the future is vital to ensure that the scheme remains financially viable for both local authorities and transport operators.

The apparent relationship between the eligible population and the number of boardings that occur allowed for predictions to be made about the future usage of the bus network by the older population. Considering both the future predicted increase in the age of eligibility and assuming a steady rate of non-eligibility related decline in boardings by concessionary bus users, future bus patronage was predicted from 2016 to 2039; the years for which future population projections were available.

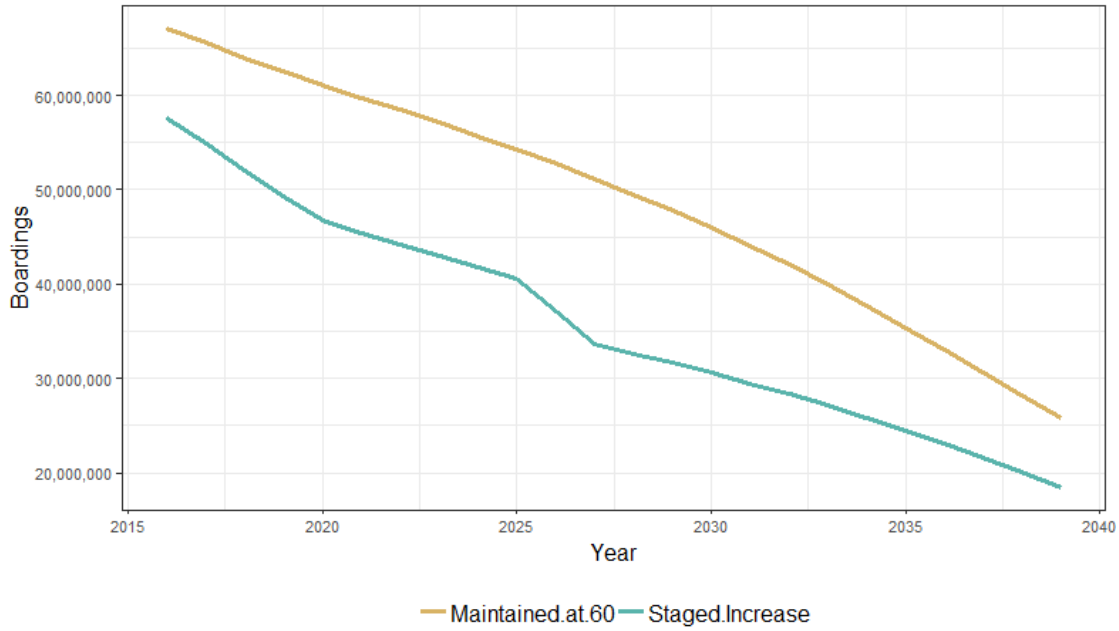


Figure 4.12 Predicted change in bus boardings between 2016 and 2039

Figure 4.12 suggests that despite a predicted increase in the population eligible for concessionary travel by 2039, if the current trend remains the same then bus usage will significantly decline over this period. Of course, it is unlikely that the rate of change in bus usage will remain constant throughout this period, however the aim of this figure was to give an indication of what future bus usage could look like in the study area with no interventions. This illustrated the importance of research aimed at identifying the cause of declining bus usage.

4.5 Summary and Conclusions

The primary aim of this chapter was to outline the smart card data to be used in this thesis, as well as to identify any potential errors or bias in the data. In addition, contextual analyses were conducted on the data, and the methods used to prepare the data for future analyses outlined. Smart card data represent a promising source of data for understanding travel behaviour; with over 80% of the pensionable population in the study area holding a concessionary smart card and all bus trips made by these cardholders recorded in the data. This is significantly higher than traditional survey methods are able to achieve.

Errors identified within the data appeared to be minor, with very few records containing missing or incorrect data, with the exception of missing ‘sex’ data. Removing these data would likely have little impact on any aggregate analyses, however, may have presented issues if individual

analyses were conducted. The near-ubiquitous nature of the smart cards meant that the data appeared to be highly representative of the older population in the study area in terms of cardholder demographics. However, the large variations in individual card usage, presented in Section 4.3.2, meant that subsequent analyses had to be conscious of the potential over-representation of more active bus users.

The method used for inferring the origins and destinations of trips made on the bus network presented promising results, however the limitation of having to rely on a subsequent trip taking place could limit its applicability in a wider context where these final trips may need to be included. As there is no suggestion that bus services in the future will require users to 'tap out' when alighting, further research is needed to develop a method for inferring the destinations of trips that are not followed by a subsequent boarding.

Accounting for the change in the age of eligibility for concessionary travel suggested that the rate of decline in bus boardings by concessionary users was less significant than the rate of decline by non-concessionary users. This was a significant finding as it was contrary to the initial trend found by TfWM and highlighted the importance of accounting for changes in the underlying population. At the time of writing, plans are in place for TfWM to provide concessionary travel to those women that have 'missed out' due to the pension age being increased (WMCA 2019). This is predicted to affect around 9,000 women in the West Midlands. Understanding how changes in eligibility may have affected the population in the study area could be of use for transport authorities when implementing future policies and schemes. The findings of this chapter were used to inform subsequent chapters of this thesis and are referred to when interpreting the results of future analyses.

Chapter 5 – A classification of concessionary bus users

5.1 Introduction

This chapter focuses on utilising the unique ability of smart card data to track an individual's transaction records over time in an attempt to group cardholders based on their long-term bus usage patterns. As outlined in Chapter 4, TfWM recorded a significant decline, since 2009, in bus boardings by older concessionary travellers. It was anticipated that utilising smart card data to conduct a longitudinal analysis of the bus usage of concessionary cardholders could provide insight into which groups of the population are making use of the bus network and which groups may have contributed to this decline. In addition, as summarised in Chapter 2, there is a need for greater research into how smart card data can be used to understand mobility patterns; the pre-processing, data cleaning and linkage methods to be utilised, and the potential of these data to supplement or replace traditional survey methods (Agard et al. 2006; Bagchi and White 2005). This chapter presents a method for classifying cardholders based on their longitudinal bus usage levels, as well as methods to process smart card data in order to extract insights into mobility.

Firstly, this research implemented a k-means clustering approach to group cardholders based on their monthly usage of the bus network; identifying active users and inactive users, as well as those that have exhibited changing usage over the study period. As discussed in Chapter 2, there is evidence that different demographic groups exhibit differing travel behaviours and trip characteristics. This study therefore subsequently compared the demographic characteristics and the spatial travel behaviours of the resulting clusters. Previous research has often treated the older population as a homogenous group with similar travel behaviours and demands (Hildebrand 2003). Exploring the differences in both the bus usage patterns, and the demographic and trip characteristics of these concessionary cardholders can demonstrate the high levels of heterogeneity found within this group of the population.

These insights helped to explain the decline in bus boardings outlined in Section 4.3.2 and provided a classification of bus users that was utilised throughout this thesis. Additionally, providing transport authorities with detailed insights into the types of cardholders that are making use of the transport network, and the types of trips that these cardholders are undertaking could aid these authorities in better targeting resources to meet the mobility needs of all users.

5.2 Method

5.2.1 Data preparation

As the aim of this chapter was to identify groups of cardholders that had exhibited similar levels of activity on the bus network over the study period, it was decided that the classification would be based solely on transaction frequency. Classifying cardholders based solely on their frequency of bus usage allowed for general groups of bus users to be identified i.e. those users that were highly active and those that rarely used bus services.

5.2.1.1 Data selection and aggregation

The first step to prepare these data for this analysis was to decide the variables to be included and the time period over which to conduct the clustering. The smart card data contained all transactions made by concessionary cardholders over the study period. Therefore, in order to obtain variables representative of bus usage, temporal aggregation of the data was necessary. In this instance, temporal aggregation refers to the process of summing the data over regular intervals. For this analysis, monthly intervals were determined to be the most appropriate as these were both representative of long-term activity levels, whilst still allowing for changes over the study period to be identified. It is acknowledged that this aggregation would have resulted in a loss of detail in the temporal activity of cardholders, for example differences in activity throughout the day, or between weekdays and weekends. As the primary focus of this chapter was to classify cardholders based on their long-term activity on the bus network rather than their granular temporal activity patterns, an aggregation to monthly frequencies allowed for this whilst maintaining a suitable number of variables for the cluster analysis.

Although transaction data were recorded from November 2009 until August 2016, initial analyses outlined in Chapter 4 showed that full transaction records were only available from September 2010. In addition, the final transaction in the data was recorded on the 18th August 2016. To ensure that only months where full transaction data were available were included in the final analysis, transactions that took place before 1st September 2010 and after the 31st July 2016 were removed.

As subsequent analyses of each initial cluster relied on the availability of a registered LSOA for each cardholder, those cardholders where there was no LSOA associated with the account were excluded from the analysis. Including these cardholders in the analysis would have skewed any results conducted at LSOA level, with differing numbers of individuals with no registered LSOA likely to be assigned to each cluster. Similar issues arose from users with missing age group and sex data. To avoid any potential bias, only cardholders with complete records were

included in the final cluster analysis, as well as only cardholders that had made at least one transaction over the study period. This initial data selection resulted in a dataset consisting of 376,208 cardholders that had made 308,529,869 transactions over the 71-month study period.

5.2.1.2 Variable preparation

Once the aggregation level and time period were determined, the data had to be prepared for the cluster analysis. Preparing the data for this analysis firstly involved using the SQL *date_part* function to extract the month and year from the transaction timestamps recorded in the smart card data. The data were then grouped by the unique account ID associated with each smart card, and the frequency of transactions per month was calculated. This dataset was then reshaped using the R function *dcast*, resulting in a dataset structured as shown in Table 5.1.

Table 5.1 Example transaction frequency dataset structure (account numbers censored to ensure anonymity)

Account Number	**Transaction Frequency**						
	Sep-2010	**Oct-2010**	**Nov-2010**	**Dec-2010**	**Jan-2011**	**Feb-2011**	**etc.**
78******35	27	24	22	31	30	27	…
78******12	12	12	9	10	11	14	…
78******92	4	2	0	2	2	1	…

Once monthly boarding variables had been calculated for each cardholder, a rate calculation was necessary to account for the differing lengths of each month. Although this would likely have a minimal effect on the cluster analysis, this step was deemed necessary so as not to overstate any monthly or seasonal trends in the transaction data. This was achieved by dividing the monthly frequencies by the number of days in each month, resulting in monthly variables representative of the average number of transactions per day.

5.2.2 Clustering methods

For this classification, cluster analysis was chosen as the most appropriate approach. Cluster analysis is a multivariate method, which aims to classify a sample of subjects into a number of different groups on the basis of a set of measured variables with the result that similar subjects are classified together (Everitt et al. 2011). These analyses can generally be divided into two subgroups; hard clustering and soft clusters. In hard clustering, each data point is allocated to a single 'best fit' cluster, whereas soft clustering instead assigns each data point a probability of belonging to each cluster (Bora and Gupta 2014). As the aim of this chapter was to identify distinct groups of cardholders based on their activity levels, a hard clustering method was implemented; assigning each cardholder to a single cluster.

Cluster analysis itself is not one specific algorithm but can instead be achieved by various algorithms that differ significantly in what constitutes a cluster and how to efficiently allocate them. The most appropriate algorithm for a cluster analysis depends on the data being used and the intended use of the results. A number of clustering algorithms were considered for this analysis, including k-means, hierarchical, and density-based spatial clustering of applications with noise (DBSCAN). There were both benefits and limitations to each of these methods, with no algorithm likely to have resulted in a perfect clustering solution. Instead, it was important to consider the positives and negatives of each method in regard to the data to be used for the analysis, in order to select the best method.

- Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. This method can be split into two types: (1) agglomerative and (2) divisive. The agglomerative method is a 'bottom-up' approach where every observation starts in its own cluster and pairs of clusters are then merged as one moves up the hierarchy. Conversely, the divisive method is a 'top-down' approach where all observations start in one cluster and splits are performed as one moves down the hierarchy (Sasirekha and Baby 2013). These methods require a distance matrix to be created, i.e. a matrix which shows the distance between every observation. For large volumes of data this can be computationally expensive. It is therefore less appropriate for clustering data of the size used in this analysis (Larose 2005) so was not deemed suitable for this study.

- DBSCAN is a density-based clustering non-parametric algorithm. This algorithm groups together points that are close to one another, and marks as outliers those points that lie alone in low-density regions (Khan et al. 2014). The distance between points is usually measured as the Euclidean distance, i.e. a straight line distance. DBSCAN

begins by dividing data in *n* dimensions, starting at a random point, and will count how many points are nearby – forming a cluster. This cluster then iteratively expands until there are no other points that are nearby. This method is effective at separating clusters of high density and low density and for identifying outliers in the data but is less effective at identifying clusters of similar densities (Tran et al. 2013). Additionally, this method relies on the maximum distance between two points for them to be determined neighbours and the minimum number of members per cluster to be predetermined. As it was predicted that the analysis may result in clusters of similar densities, this algorithm was deemed inappropriate for the analysis.

- K-means is an iterative relocation algorithm that assigns each data point to one of *k* clusters based on a standardised Euclidean minimum distance metric. Firstly, the initial number of clusters ('*k*') is selected. Each observation is then assigned to its closest cluster centre, based on the distance to the cluster mean, with the cluster means updating after each iteration to be the mean of its constituent observations (Wagstaff et al. 2001). The algorithm converges when there is no further change in the allocation of instances to clusters. It is a popular clustering method for the analysis of large, multivariate data with the aim of findings groups of homogenous observations (Everitt et al. 2011).

Weighing up the positives and negatives of each of these methods, described above, left k-means as the most appropriate technique for this analysis. A k-means clustering method was determined to be effective at grouping data of the volume used in this research into distinct groups. Due to its relative simplicity and extensive use within social science research, this approach would also allow for easy replication on future data, for example more up-to-date smart card data.

5.2.2.1 K-means algorithm

K-means clustering is a type of unsupervised learning which aims to group observations based on their similarity. The aim of this algorithm is to partition *n* observations into *k* clusters so as to minimise the within-cluster sum of squares (WCSS), i.e. the variance. This can be expressed as:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left| x - m_i \right|^2$$

where x represents the sum distance between each data point, m_i represents their respective cluster centres, k represents the number of clusters, and E represents the objective function.

This algorithm uses an iterative refinement technique, consisting of four main steps (Jain and Dubes 1988):

1. Select an initial partition with k clusters.
2. Generate a new partition by assigning each observation to its closest cluster centre.
3. Compute new cluster centres.
4. Repeat steps 2 and 3 until cluster membership stabilises.

Figure 5.1 shows an illustration of the k-means algorithm on a 2-dimensional dataset with three clusters.

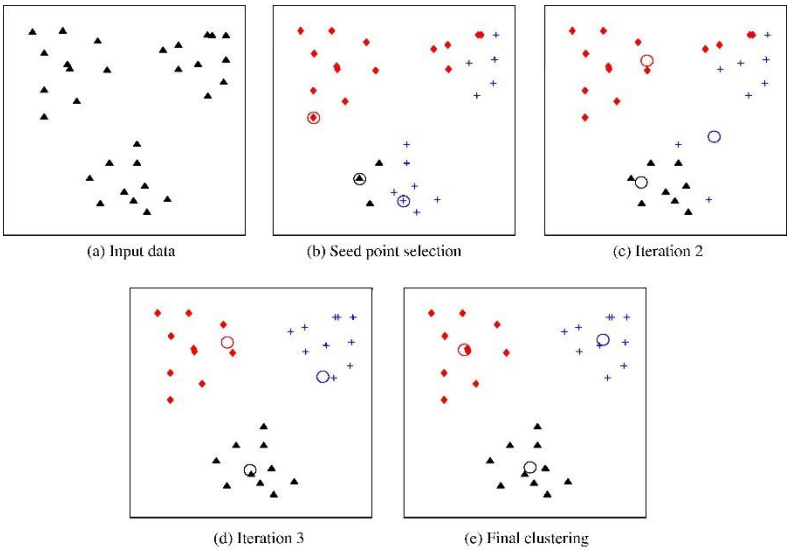


Figure 5.1 Illustration of K-means algorithm. (a) Input data with three clusters; (b) initial partition and assignment of data points to clusters; (c) and (d) intermediate iterations; (e) final cluster solution (Jain 2010)

There are a number of variations of k-means algorithms, such as 'Hartigan-Wong', 'Lloyd' and 'MacQueen' (Morissette and Chartier 2013). For this analysis, the R default algorithm 'Hartigan-Wong' (Hartigan and Wong 1979) was selected as this focuses on minimising the total WCSS, and therefore maximising the similarities between clusters. Given the random allocation of initial seeds in the k-means algorithm, a number of iterations had to be performed

to achieve an optimum solution. Singleton and Longley (2009) suggest that 10,000 iterations are appropriate to achieve this, and this was therefore used as the maximum threshold for this analysis.

5.2.2.2 Number of clusters

Unlike other clustering methods, such as Hierarchical clustering, k-means clustering requires a predetermined number of clusters to be specified before the analysis is run. The number of clusters, often labelled as *k*, used for a k-means classification can greatly affect the outcome of the clustering and the interpretation of the results. There is no set method for determining the optimal number of clusters, as this often depends on the shape and scale of the data, as well as the desired clustering resolution (Ray and Turi 1999). Increasing *k* will always reduce the amount of error in the resulting clusters, as smaller clusters will contain more similar data points. A large number of clusters, however, may not always be appropriate for the analysis. It was therefore important that a trade-off was struck between a suitable number of clusters and the amount of error in the resulting clusters.

In order to determine the optimal number of clusters to be used in this classification, the 'Elbow Method' was used. This method consisted of plotting the average total WCSS for different values of *k*. The WCSS is a measure of the variability of the observations within each cluster. In general, the smaller the WCSS is, the lower the variability between observations in each cluster. The aim of this method was therefore to identify a point at which increasing the number of clusters further did not significantly decrease the average total WCSS (Hughes and Moreno 2013). The results of this analysis, showing the average total WCSS plotted against values of *k* ranging from 2 to 10, are shown in Figure 5.1.

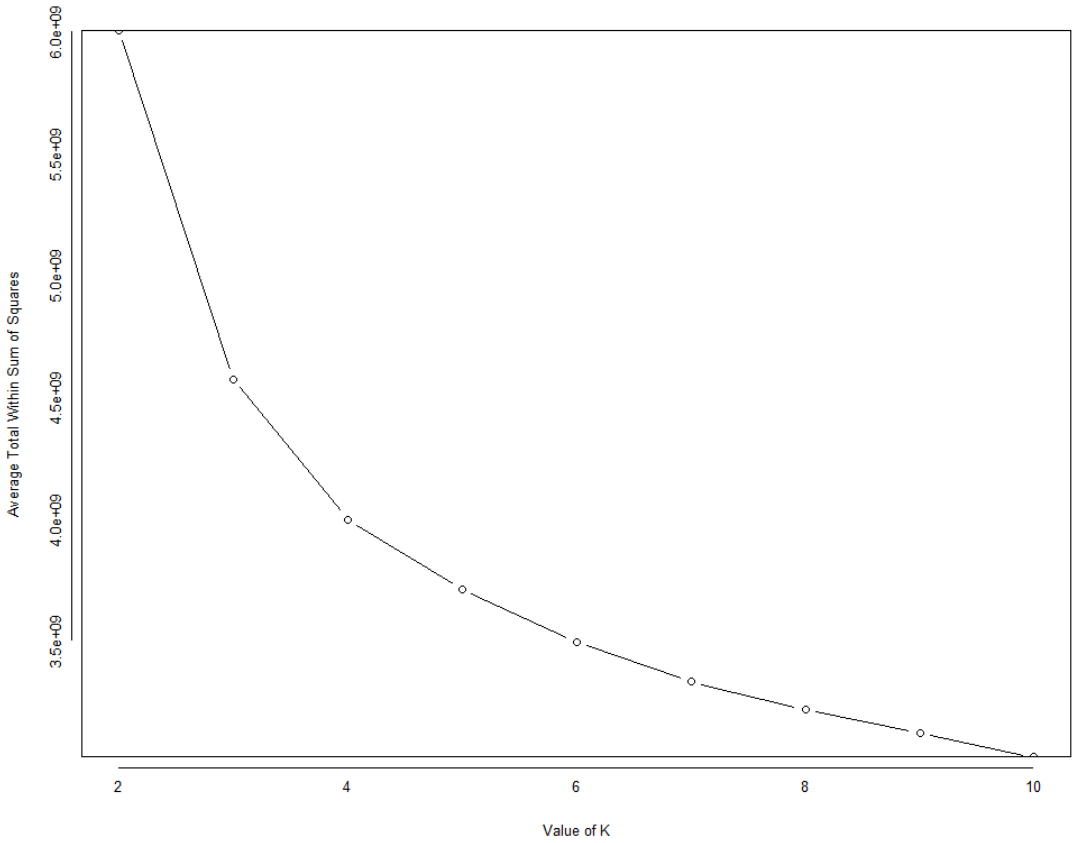


Figure 5.2 Average total WCSS for values of *k* ranging 2-10

Although there was no clear point, or elbow, in this graph the rate of reduction in the WCSS values decreased significantly at a *k* value between 5 and 7. As there was no clear optimal value of *k* shown in this plot, the number of members in each cluster was also analysed. As the aim of this analysis was to identify groups of cardholders that exhibit similar bus activity patterns, the optimal number of clusters would not necessarily have equal cardholder membership. It was important, however, that no clusters should have a membership low enough that the initial clusters, or any resulting analyses, did not adhere to the statistical disclosure controls outlined in Chapter 4.

Table 5.2 Cluster membership for *k* values ranging from 5 to 7

k	**5**	**6**	**7**
Cluster membership	339,862	266,455	266,490
	3,454	73,498	73,708
	15,540	3,373	2,982
	17,171	15,546	15,596
	29,789	17,336	17,355
	-	29,608	29,294
	-	-	391
Range	336,408	263,082	266,099
Average distance from the mean	103,480	68,227	64,071

Comparing both Figure 5.2 and Table 5.2 it was determined that 6 clusters would be optimal for the analysis; striking a balance between a low WCSS value and suitable cluster memberships. 5 clusters resulted in one predominant cluster, whereas 7 clusters resulting in one cluster that contained significantly fewer cardholders and therefore the outputs of analyses conducted on this cluster may not have adhered to statistical disclosure controls.

Once the number of clusters had been determined, the k-means clustering algorithm could be run on the data, with the aim of grouping cardholders into clusters that displayed similar bus activity patterns across the 71-month study period.

5.2.3 Cluster characteristics

5.2.3.1 Data linkage: demographic and socioeconomic characteristics

The cardholder data, outlined in Section 4.2.2, contained the age, sex and registered LSOA of each smart card holder, and could be linked to the resulting clusters using each cardholder's unique account number. This information was used for calculating the age and sex distributions of each cluster, as well as for inferring additional demographic characteristics.

As the smart card data only contained information relating to the age and sex of cardholders, additional demographic characteristics had to be inferred. This was achieved through cross-validation with census data. These census data were available at LSOA level and were also able to be grouped by the age and sex of respondents. This allowed for the proportions of individuals of a certain age and sex belonging to a specific demographic group to be calculated for each LSOA. Linking the smart card data to these census data therefore allowed for these proportions for each cluster population to be calculated, and therefore inferences of the demographic make-up of each cluster to be made.

Selection of the data to be included in the analysis was subjective and was therefore based on insights generated from previous research. This was outlined in Chapter 2. Table 5.3 outlines the data that were selected for the analysis, as well as providing a justification as to why these data were used.

Table 5.3 Demographic and socioeconomic data: descriptions and justifications

Data	Source	Description	Justification
IMD	ONS (2015)	Official measure of relative deprivation for small areas in England, based on seven domains of deprivation: - Income deprivation - Employment deprivation - Education, skills and training deprivation - Health deprivation and disability - Crime - Barriers to housing and services - Living environment deprivation LSOAs are classified into 10 deciles, with Decile 1 describing areas that are most deprived, and Decile 10 describing those that are least deprived i.e. LSOAs in Decile 1 fall among the most deprived 10% of all areas.	Understanding the relationship between deprivation and use of the public transport network can aid local transport authorities in the targeting of resources and ensuring deprived areas have adequate transport access (Palacin et al. 2016).

Car or Van Availability	Census 2011 (ONS 2018b)	These data apply to the number of cars or vans that are owned, or available for use, by one or more members of a household. This includes company cars and vans that are available for private use but do not include motorbikes or scooters. Cars or vans used by residents of communal establishments were also not included. Data were available at LSOA level.	Access to private modes of transport, have a significant impact on both public transport usage and the types of trips undertaken on public transport services (Kim and Ulfarsson 2004). Understanding the relationship between access to private cars or vans and usage of concessionary smart cards can provide insights into the mode choice of different groups of cardholders, in addition to the potential to identify areas that may lack transport options.
Ethnic Group	Census 2011 (ONS 2018b)	These data describe the ethnic group of usual residents in England and Wales. Data were available at two tiers, with the first tier describing general ethnic groups and the second tier describing more detailed ethnic groups. For ease of interpretability, this analysis utilised data at the first tier. These ethnic groups were: - White - Mixed/ Multiple - Asian/ Asian British - Black/ African/ Caribbean/ Black British - Other Data were available at LSOA level.	Differences have been found in the travel behaviours of different ethnic groups, including the number of public transport trips undertaken and the purpose of these trips (Schmocker et al. 2004). Understanding differences in the ethnic make-up of each cluster can allow for further exploration of these insights and aid transport authorities in meeting the mobility needs of all concessionary travellers.
General Health	Census 2011 (ONS 2018b)	These data relate to the general health of the population in England and Wales. These data are a self-assessment of health and consisted of three variables: - Very good or good health - Fair health - Bad or very bad health Data were available at LSOA level.	The links between mobility and health are well documented e.g. the relationship between improved mobility opportunities and increased health (Mackett 2013), and improved health through better access to healthcare services (Cass et al. 2005). Understanding the relationship between health and bus usage of the concessionary population can provide insights into the potential impact that increased mobility can have on the health of concessionary users.

NS-SeC	Census 2011 (ONS 2018b)	The National Statistics Socioeconomic Classification (NS-SeC) is the official socioeconomic classification in the UK. The NS-SeC is a nested classification and is most commonly used in the eight-class version: - Higher managerial and professional occupations - Lower managerial and professional occupations - Intermediate occupations - Small employers and own account workers - Lower supervisory and technical occupations - Semi-routine occupations - Route occupations - Never worked or long-term unemployed This can also be reduced into a hierarchical, three-class version: 1. Higher occupations 2. Intermediate occupations 3. Lower occupations Data were available at LSOA level.	Evidence suggests that socioeconomic background is strongly linked to travel behaviour and mode choice, with almost every National Travel Survey identifying significant inequalities in the travel patterns and access to transport services of lower income populations (Lucas 2012). Comparing the NS-SeC distributions of each cluster can provide further insight into these differences.
Age Group	Cardholder Data	The date of birth of each cardholder was supplied when applying for a concessionary travel card. As part of the anonymization process, TfWM converted these dates of birth to age groups, with 5 year age bands: - 61-65 - 66-70 - 71-75 - 76-80 - 80+ These age groups were available at the individual level.	Travel behaviours change with age, with the older population likely to make fewer, shorter trips than the working age population (Mackett 2014). Despite this, mobility patterns of the older population are not homogenous, and there are likely to be differences in travel behaviours between age groups within this population. Understanding the age distributions of each cluster is particularly important for identifying individuals that have entered the system, recently become eligible for concessionary travel, or exited the system due to old age.

Sex	Cardholder Data	The sex of the cardholder (either 'Male', 'Female' or 'Unknown') was supplied when applying for a concessionary travel card, and was provided by TfWM. Data were available at the individual level.	Contextual analyses presented in Chapters 2 and 4 suggested that there may be significant and changing differences in the travel behaviours of men and women, e.g. the increase in the proportion of women with driving licenses and the increase in retirement age for women are likely to have resulted in changing travel behaviours.

5.2.3.2 Spatial profiles

The spatial profiles of the resulting clusters depicted the geographic distribution of cardholders throughout the study area, both in terms of where they lived and the areas that they visited. These profiles, therefore, were comprised of two elements: the residential locations of cardholders allocated to each cluster and the destinations of their trips made on the bus network.

Analysis of the residential locations of each cardholder provided insights into differences in the distribution of the clusters throughout the study area. In order to generate these profiles, the registered LSOAs recorded in the cardholder data were linked to mid-year population estimates. The proportion of the eligible population allocated to each cluster within each LSOA was then calculated, allowing for areas with high and low concentrations of each cluster to be identified.

Understanding the activity areas of each cluster throughout the study area not only provided insights into how different types of bus users made use of the network, but also into the mobility needs and demands of concessionary cardholders. Apart from limits on use imposed by the bus companies, concessionary bus users are likely to have different mobility needs and demands than the working age population (Kim and Ulfarsson 2004) and understanding these are vital for transport authorities to be able to meet these needs and provide adequate transport services for different groups of the population. These spatial profiles utilised the destinations of trips made on the bus network, inferred using the method outlined in Section 4.2.4.1, to identify differences in the activity areas between the clusters.

In order to understand and visualise the differences in the activity areas of each cluster, a hotspot analysis was conducted. Hotspot analysis uses vectors to identify statistically significant hot and cold spots in the data, i.e. areas where high and low values cluster together. This was chosen over other methods of spatial analysis, such as heat maps, as this method identifies areas that are statistically significant and the results are therefore less subjective.

Hot spot analysis works by looking at each feature in the dataset in the context of its neighbouring features. To determine whether an area is a statistically significant hot or cold spot, the local sum for a feature and its neighbours is compared proportionally to the sum of all features (Songchitruksa and Zeng 2010). A statistically significant z-score will result when this local sum is very different from the expected local sum, and when this difference is too large to be the result of random chance.

To prepare the smart card data for this analysis, the data were separated into transactions made by cardholders belonging to each cluster. As full trips can consist of several interchanges between bus services, the final destinations of each trip had to be extracted. Bus timetable data provided by TfWM, outlined in Section 4.2.3, were analysed to establish a maximum dwell time, after which a transaction was determined to have resulted in an activity taking place rather than an interchange between services.

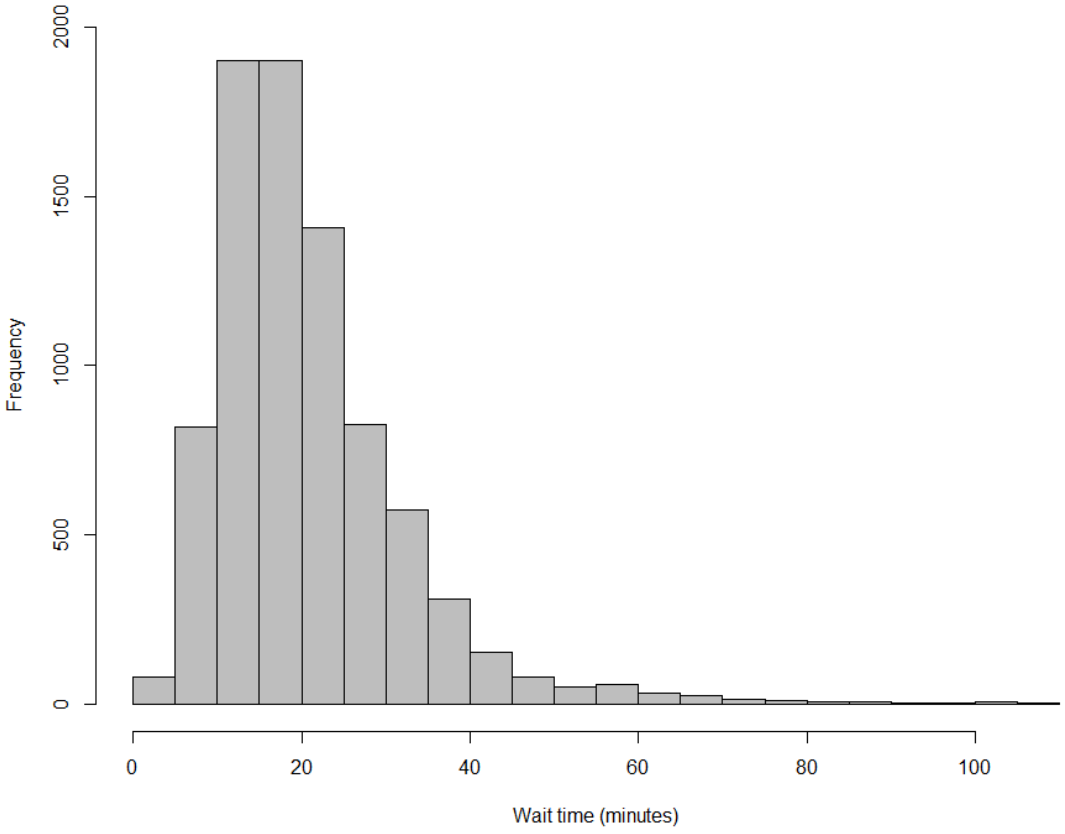


Figure 5.3 Histogram of average off-peak wait times

Figure 5.3 shows the frequency of wait times at 'interchange' bus stops. These interchange stops were defined as stops on the bus network that service more than one route. Wait times were calculated as the average time difference between a bus servicing one route arriving at the

interchange stop and the next bus servicing a different route leaving the interchange stop. These wait times were calculated as the average off-peak, weekday wait time for 8,260 bus stops in the study area. Analysis of the distribution of wait times showed that over 90% of bus stops had an average wait time of less than 30 minutes. A dwell time threshold of 30 minutes was therefore implemented for this analysis, with any dwell time beyond this threshold determined to have been the result of an activity taking place rather than an interchange between services.

In order to remove interchange stages, firstly the dwell time between subsequent bus boardings was calculated using the *datediff* function in SQL to calculate the difference, in minutes, between an alighting and a subsequent boarding. Once these were determined, those stages with a dwell time of less than 30 minutes were identified as interchange stages and removed from the analysis. This resulted in a dataset containing only stages where an activity was determined to have taken place. The frequency of alightings at each bus stop was then calculated. In order to determine activity areas rather than singular bus stops, the data were aggregated to LSOA level; summing the frequencies of all bus stops within each LSOA.

The hotspot analysis was conducted using the 'Hot Spot Analysis' tool in ArcGIS. This tool calculates the Getis-Ord G_i^* statistic for each feature in the dataset (Songchitruksa and Zeng 2010) – in this instance each LSOA. The resultant z-scores and p-values signify whether features with high or low values cluster spatially, and whether this spatial clustering is statistically significant. This G_i* statistic is an example of a local indicator of spatial association (LISA) (Anselin 1995).

The Getis-Ord G_i* statistic is expressed as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}}$$

where x_j is the attribute value for feature j, $w_{i,j}$ is the spatial weight between features i and j, n is equal to the total number of features, and:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{X}\right)^2}$$

The G_i^* statistic is a z-score representing standard deviations, so no further calculations were required. The larger the z-score is, the more intense the clustering of high values is and therefore the more statistically significant the hotspot. Likewise, the smaller the z-score, the more intense the clustering of low values, and the more statistically significant the cold spot. The resultant hotspot maps showed the areas that attract either statistically significant high or low numbers of visits from each cluster, and therefore provided insights into differences in the activity areas and the spatial distribution of destinations of different types of bus users.

5.2.4 Within-cluster changes

Initial results suggested that analysis of more granular, within-cluster differences would be valuable in understanding the overall decline in bus usage recorded by TfWM. It was anticipated that the initial k-means clustering would be effective at grouping cardholders that exhibited similar general bus activity patterns but would be less effective at identifying these more granular changes in usage. This was largely due to the initial value of *k* selected for the analysis, with a higher value allowing for more clusters and therefore more granular differences between cardholders to be identified. Rather than increasing the initial value of *k*, which would have resulted in a number of clusters displaying similar general bus activity, it was determined that each initial cluster would be clustered again based on their change in bus usage over the study period.

To generate variables that were indicative of changes in bus usage within each initial cluster, the monthly transaction frequencies calculated for each cardholder in Section 5.2.1.2 were transformed into ‘percentage difference from average’ variables. This firstly involved identifying the first month in which each cardholder transacted as, it was acknowledged that, including months prior to a cardholder entering the system in the calculation would skew the results. Once these first transactions were identified, the average monthly transaction frequency ranging from the first full ‘active month’ to the end of the study period was calculated for each

cardholder. In order to identify changes in transaction frequency from these averages, the monthly percentage change from this average was then calculated for each cardholder.

Re-clustering each initial cluster, following the same method as outlined in Section 5.2.2 but with a *k* value of 3, allowed for cardholders within these clusters that had exhibited increasing, decreasing and steady bus usage to be identified. This allowed for a more detailed view of bus activity than would have been generated had a larger initial value of *k* been used. The initial clustering grouped cardholders based on their general bus usage. With this re-clustering method grouping these cardholders again based on their granular changes in bus activity, this allowed for a multi-dimensional view of each cluster and sub-cluster. The demographic and socioeconomic characteristics of these cardholders were analysed following the method outlined in Section 5.2.3.1.

5.3 Results

5.3.1 Clusters

5.3.1.1 Bus Activity

The initial k-means clustering algorithm resulted in 6 clusters, each exhibiting distinct patterns of activity on the bus network over the study period. Figures 5.4 shows the average monthly transaction frequency of each of these resulting clusters.

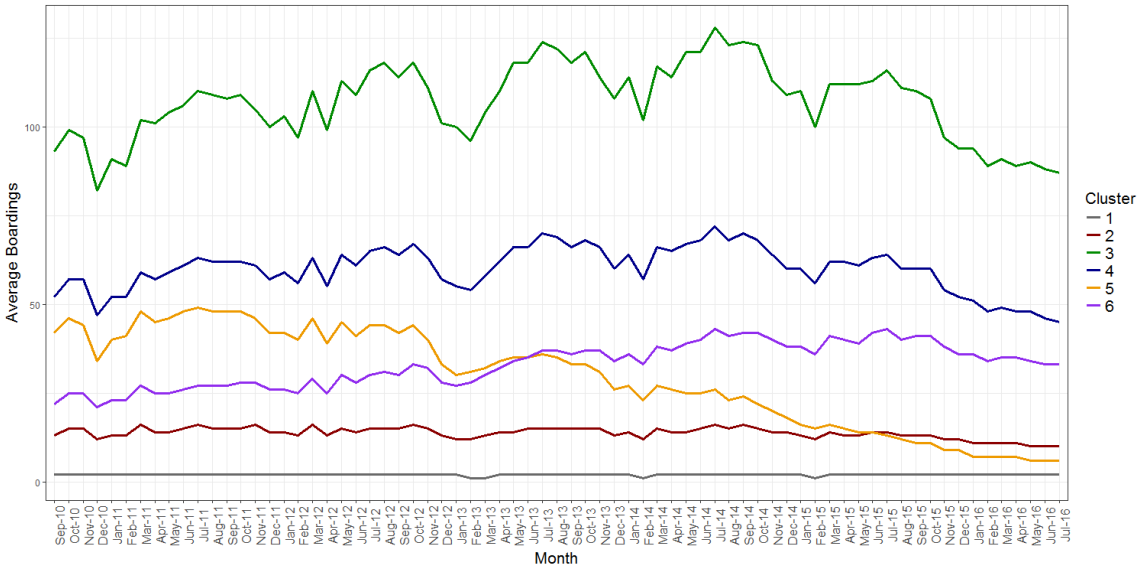


Figure 5.4 Average monthly transaction frequency of each cluster

As can be observed in Figure 5.4, each of the clusters displayed unique patterns of activity on the bus network. Cluster 1 exhibited very low use of the bus network, with few transactions occurring over the study period. Similarly, Cluster 2 also exhibited relatively low bus usage, with an average frequency of around 14 transactions per month. In contrast, Clusters 3 and 4 exhibited significantly higher activity on the bus network.

Clusters 5 and 6 displayed changing patterns of bus usage over the study period. Cluster 5 exhibited a decrease in transaction frequency, with average frequency decreasing from 42 transactions in September 2010 to just 6 transactions in July 2016. Despite the overall trend of decreasing bus patronage outlined in Chapter 4, Cluster 6 consisted of cardholders that had displayed increasing bus usage over the study period from an average of 22 monthly transactions in September 2010 to 33 in July 2016.

All clusters, with the exception of Cluster 1, appeared to exhibit a significant decline in boardings towards the end of the study period, beginning around July 2015. Analysing the number of monthly active users, i.e. the number of cardholders that have made at least one transaction, helped to determine whether this decline was due to a decrease in transactions per user or a decline in the total number of users.

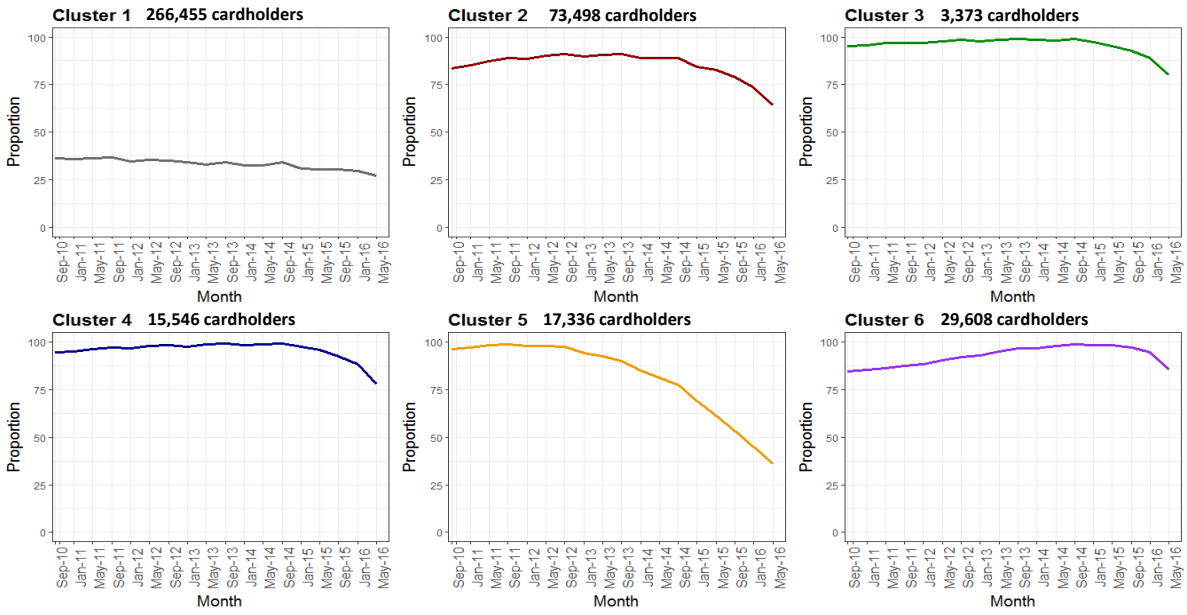


Figure 5.5 Proportion of active users each month

All clusters exhibited a significant decline in monthly active users, including Cluster 6 which was characterised by increasing bus usage. This decline in active users coincided with the decline in transactions seen in Figure 5.5, suggesting that cardholders may have dropped out of the system rather than undertaken fewer trips. Unsurprisingly, this decline was most prominent in Cluster 5, as this cluster was characterised by decreasing bus usage. Cluster 1 showed a less significant decline. This was, however, likely due to the low number of transactions and active users that this cluster contained. The patterns shown in these figures suggest a progressive withdrawal from the concessionary scheme and could be indicative of a cohort effect, with the majority of declines beginning around the same time; the end of 2014 and the beginning of 2015. The significant declines in both transactions and active users highlighted the need to analyse each sub-cluster in addition to the differences between clusters.

The distinct patterns of bus activity between clusters allowed for cluster names and descriptions to be developed; outlined in Table 5.4:

Table 5.4 Cluster names and descriptions

Cluster	Name	Members	Description
1	*'Rarely Seen'*	266,455	Characterised by low levels of activity on the bus network: - Average frequency of <2 transactions per month - On average, 33.7% of cluster members were active each month
2	*'Irregular'*	73,498	Characterised by irregular use of the bus network: - Average frequency of 14 transactions per month - On average, 60.1% of cluster members were active each month
3	*'Highly Active'*	3,373	Characterised by particularly high bus usage: - Average frequency of 107 transactions per month - On average, 95.4% of cluster members were active each month

4	*'Active'*	15,546	Characterised by high bus usage: - Average frequency of 59 transactions per month - On average, 95.1% of cluster members were active each month
5	*'Decreasing Usage'*	17,336	Characterised by decreasing activity on the bus network: - Average frequency of 29 transactions per month - On average, 80.1% of cluster members were active each month - An average 85% decrease in monthly transaction frequency over the full study period
6	*'Increasing Usage'*	29,608	Characterised by increasing activity on the bus network: - Average frequency of 33 transactions per month - On average, 92.4% of cluster members were active each month - An average 50% increase in monthly transaction frequency over the full study period

5.3.1.2 Cluster demographics

The following figures show the demographic and socioeconomic distributions within each cluster; inferred using the method outlined in Section 5.2.3.1.

Index of Multiple Deprivation

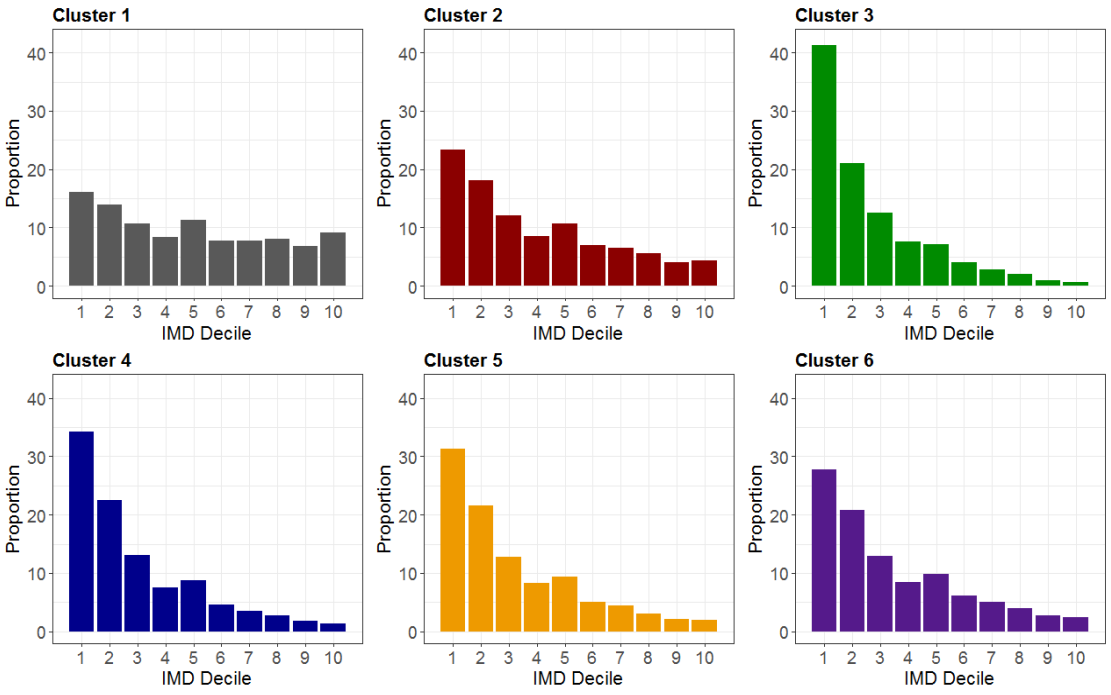


Figure 5.6 Proportion assigned to each IMD decile (1 = most deprived, 10 = least deprived)

Car or Van Availability

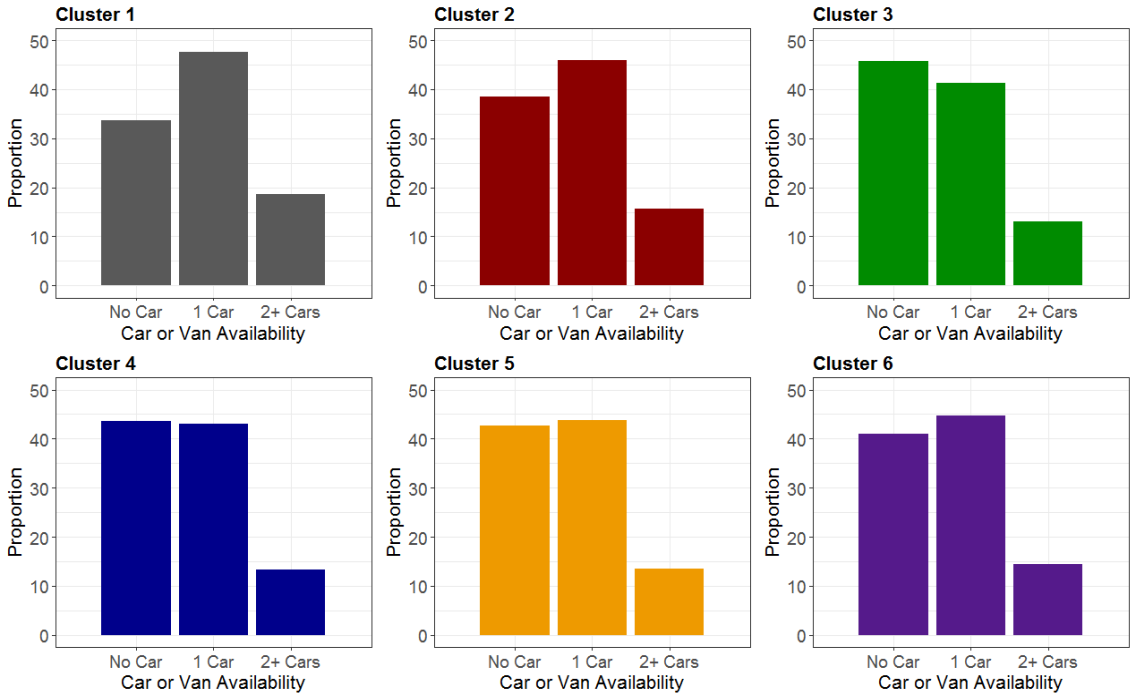


Figure 5.7 Proportion of cardholder with access to a private car or van

Ethnic Group

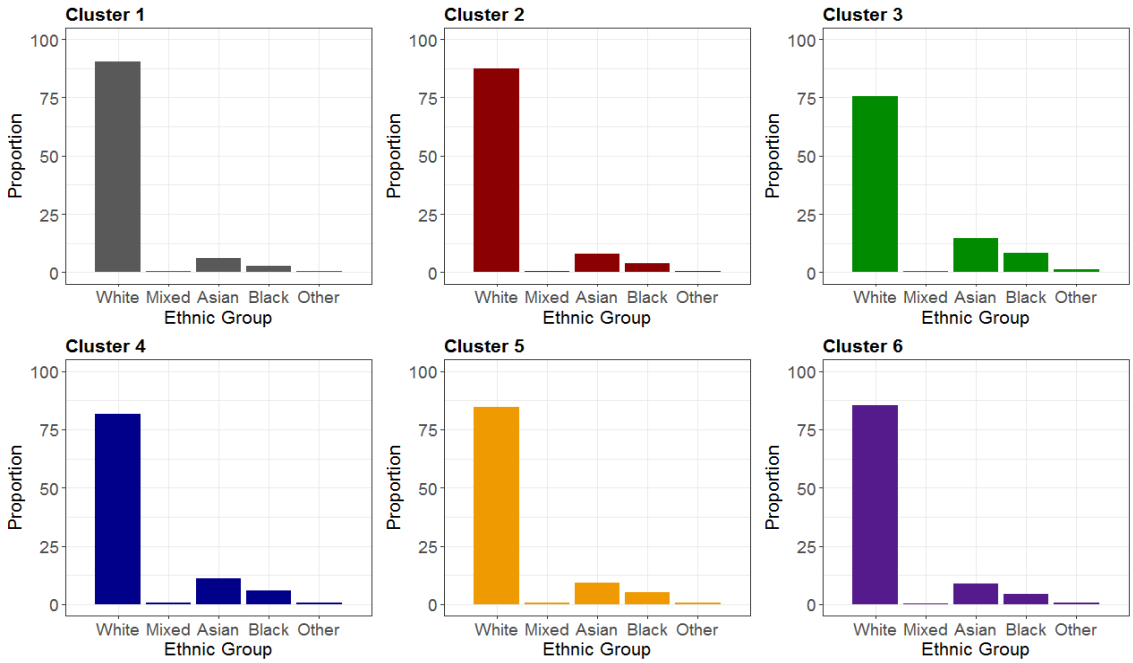


Figure 5.8 Proportion of cardholders belonging to each ethnic group

General Health

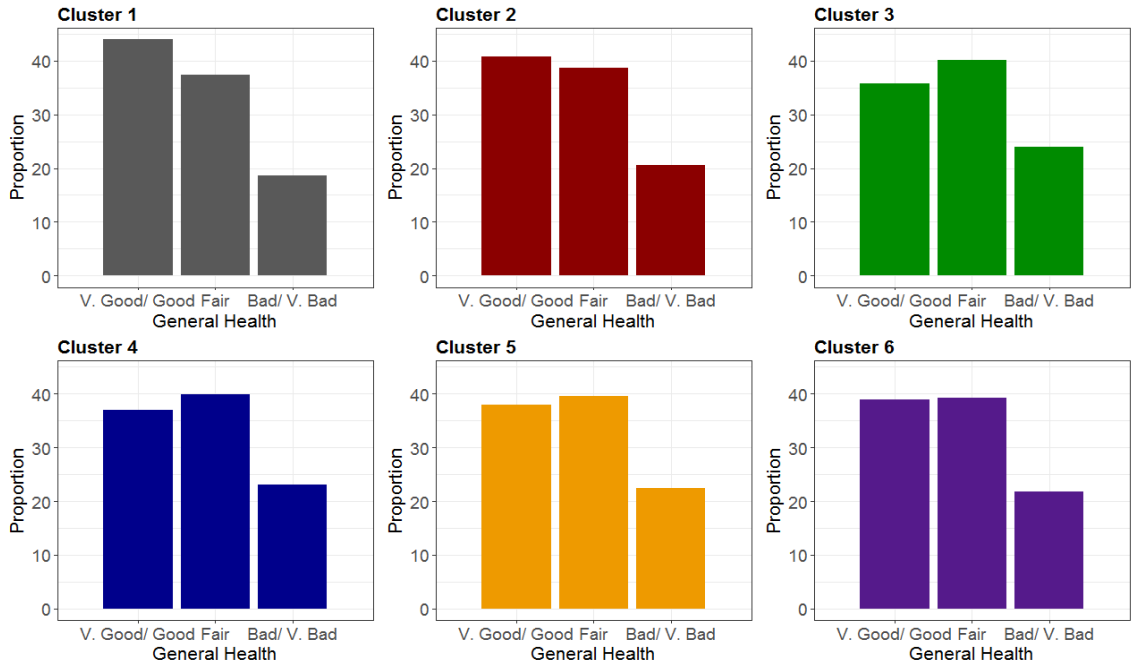


Figure 5.9 Proportion of cardholders with 'Very Good or Good', 'Fair' and 'Bad or Very Bad' health

National Statistics Socio-Economic Classification

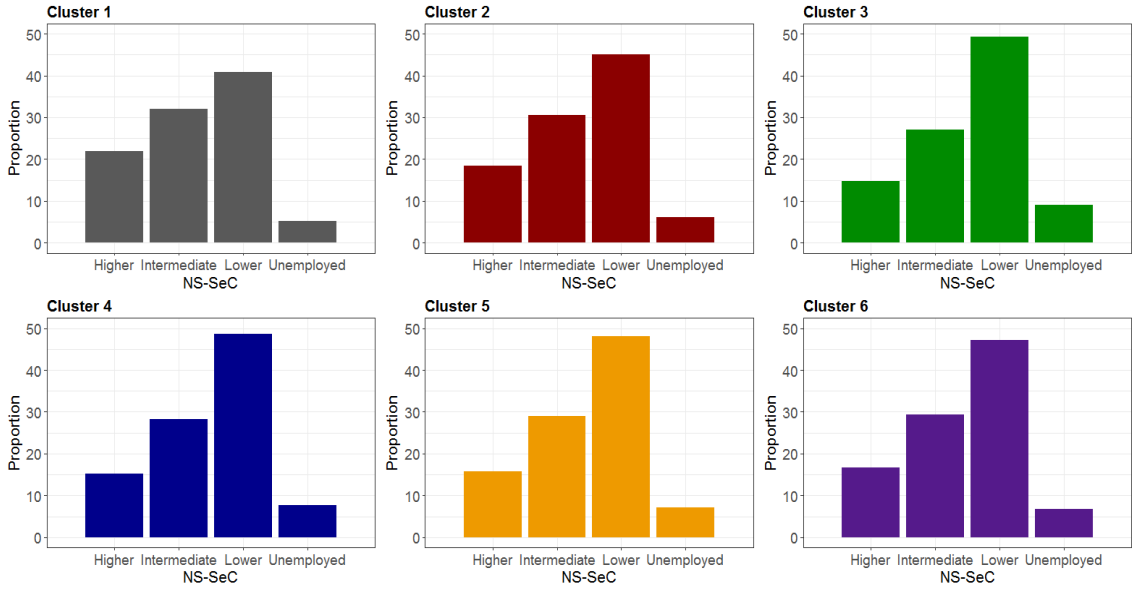


Figure 5.10 Proportion 'higher', 'intermediate' or 'lower' occupations or long-term unemployed cardholders

Age Group

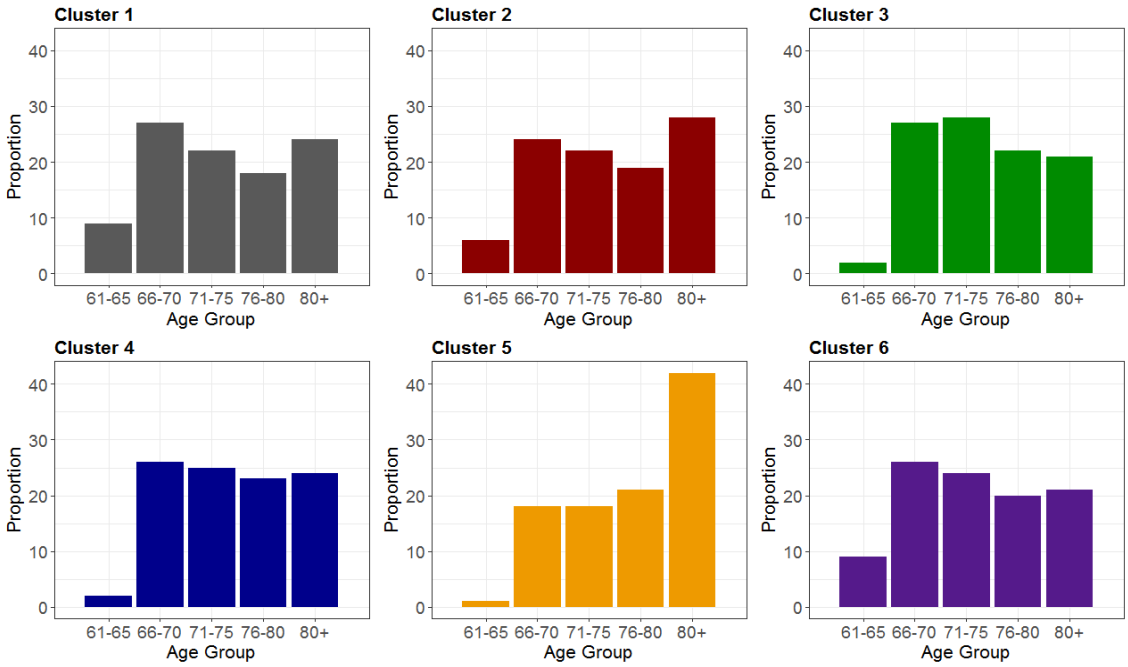


Figure 5.11 Proportion of cardholders belonging to each 5-year age group

Sex

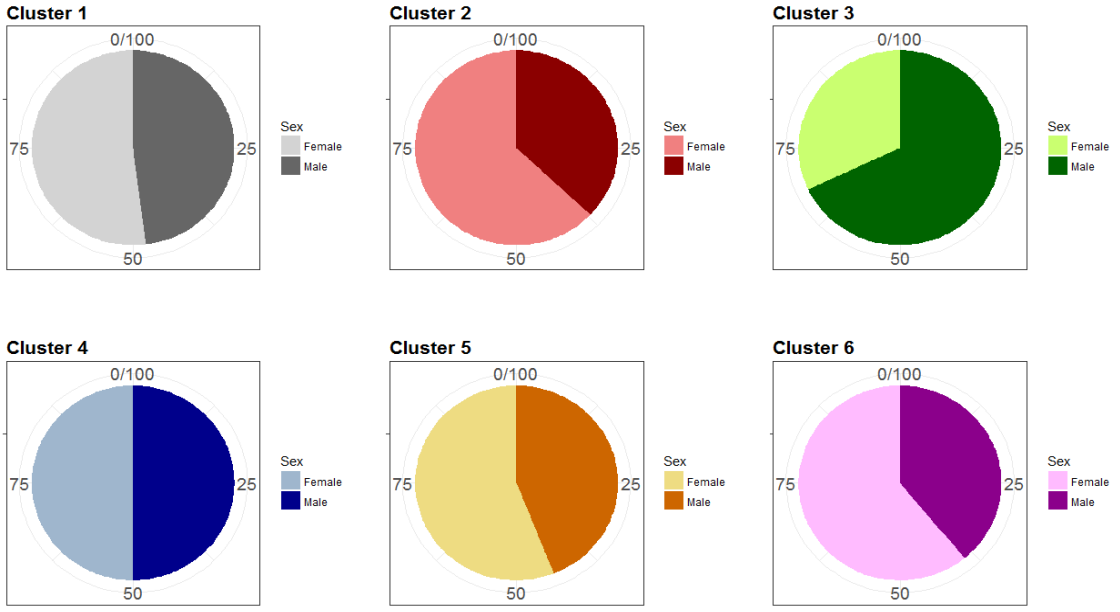


Figure 5.12 Proportion of male and female cardholders

Results suggested that there were clear differences in the demographic and socioeconomic makeup of each cluster, in particular between those that were most active on the bus network and those that exhibited infrequent usage. This was particularly evident in the analysis of IMD. Those that were most active on the bus network (Clusters 3 and 4) contained significantly higher proportions of cardholders living within the most deprived areas, with 41% and 34% respectively. In comparison, those in Clusters 1 and 2, characterised by low and irregular bus usage, contained 16% and 23% of cardholders living in areas of Decile 1 respectively. Little difference was evident in the IMD distributions between Clusters 5 and 6; those that exhibited changing activity levels on the bus network. The spatial distributions of the residential locations of each cluster are presented in Section 5.3.1.3 and suggested a stark difference in the spatial distributions of active and less active bus users.

Analysis of general health and NS-SeC categories showed similar distributions, with the most active clusters containing lower proportions of cardholders with 'good/ very good health', as well as lower proportions of cardholders inferred to be in the 'higher occupations' NS-SeC category than those that used the bus network less regularly. These results were unsurprising given the IMD distributions, as data relating to health and employment are used to build the IMD classification. The relationship between these characteristics and bus usage provides insights into the groups of the population that rely on the bus network.

The analysis of access to private modes of transport suggested a strong relationship between access to private transport and use of bus services, with higher proportions of regular bus users living in areas with low private transport availability. Despite this, there were still large proportions of irregular and rarely seen cardholders in Clusters 1 and 2 residing in areas with low proportions of private car or van availability. Although this does not necessarily mean that these cardholders themselves lack access to private modes of transport, these cardholders are of particular interest when assessing transport disadvantage and access to both private and public transport services.

When considering distributions between age groups, the starkest differences existed between Clusters 1, 5 and 6. Cluster 5, consisting of those that had decreased bus activity over the study period, contained a significantly higher proportion of those aged over 80, whilst Clusters 1 and 6 contained a higher proportion of those in the youngest age group, 61-65. These results suggested that bus usage may decline with age, and although younger age groups displayed increasing usage of the bus network, these users were more likely to use bus services irregularly than older age groups. This irregular usage may be due to younger age groups being more likely to also have access to private transportation (Musselwhite et al. 2015) and therefore make use of both private and public transport. Although TfWM aim to keep track of cardholders who have

died and remove their details from the system, there is also the possibility that Cluster 5 still contained a number of these cardholders.

Trends in the distribution of sexes within clusters were also apparent, with Clusters 3 and 4 the only clusters containing over 50% male cardholders. The remaining clusters contained significantly greater proportions of females, in particular Cluster 2, characterised by irregular bus usage, and Cluster 6, characterised by increasing usage. A high proportion of male cardholders regularly using the bus network may suggest that these cardholders use bus services for commuting purposes, with men more likely to work to an older age (Department for Work and Pensions 2015). This is supported by the age distributions shown in Figure 5.11.

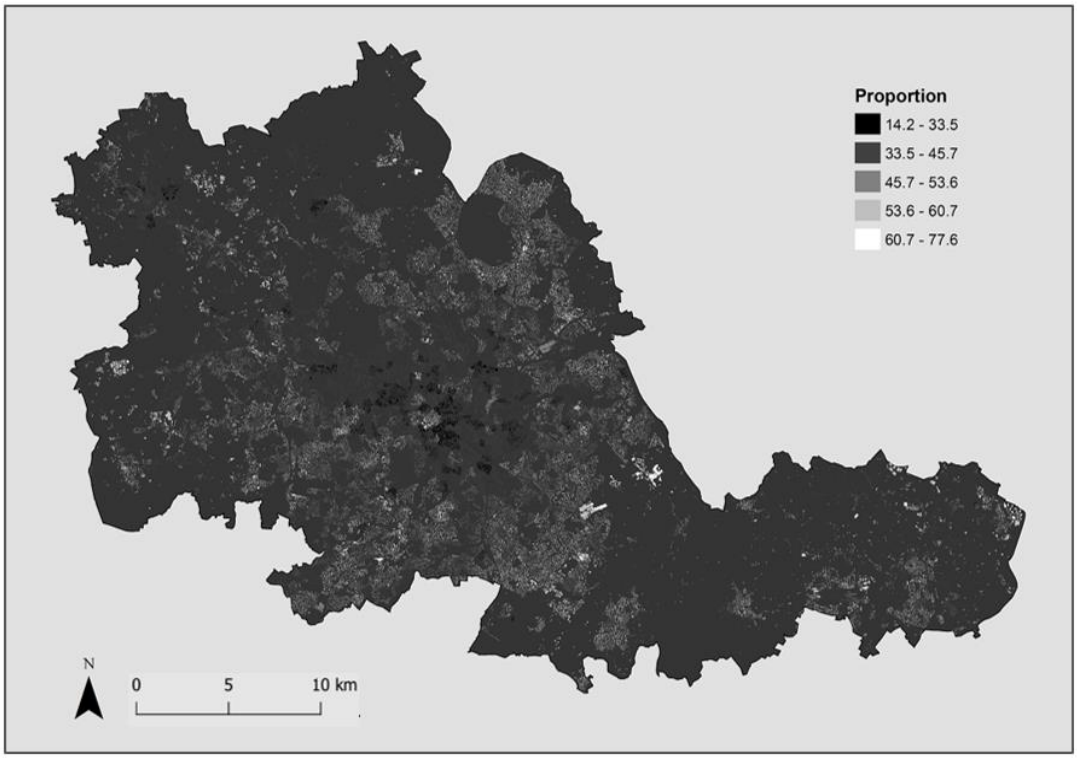
These results suggested that bus usage varied significantly between different groups of the older population. One of the aims of this thesis was to understand the declining bus patronage of the older population. These insights evidenced that the only significant difference between those that had dropped out of the concessionary scheme and the other clusters was age, with cardholders allocated to Cluster 5 containing much greater proportions of individuals aged 80 and over. This would be expected, with mobility and trip frequency found to decline with age (Mackett 2014). Comparing these results with those shown in Figures 5.4 and 5.5 suggested that the re-cluster method was needed to fully understand the types of cardholders that had contributed to the decline in bus boardings.

5.3.1.3 Spatial profiles

The spatial profiles shown in Figures 5.13 to 5.18 show where the cardholders in each cluster resided, presented as the proportion of LSOA population belonging to each cluster, and the areas to which they undertook trips, shown as hotspot activity areas at LSOA level.

Cluster 1 – 'Rarely Seen'

a)



b)

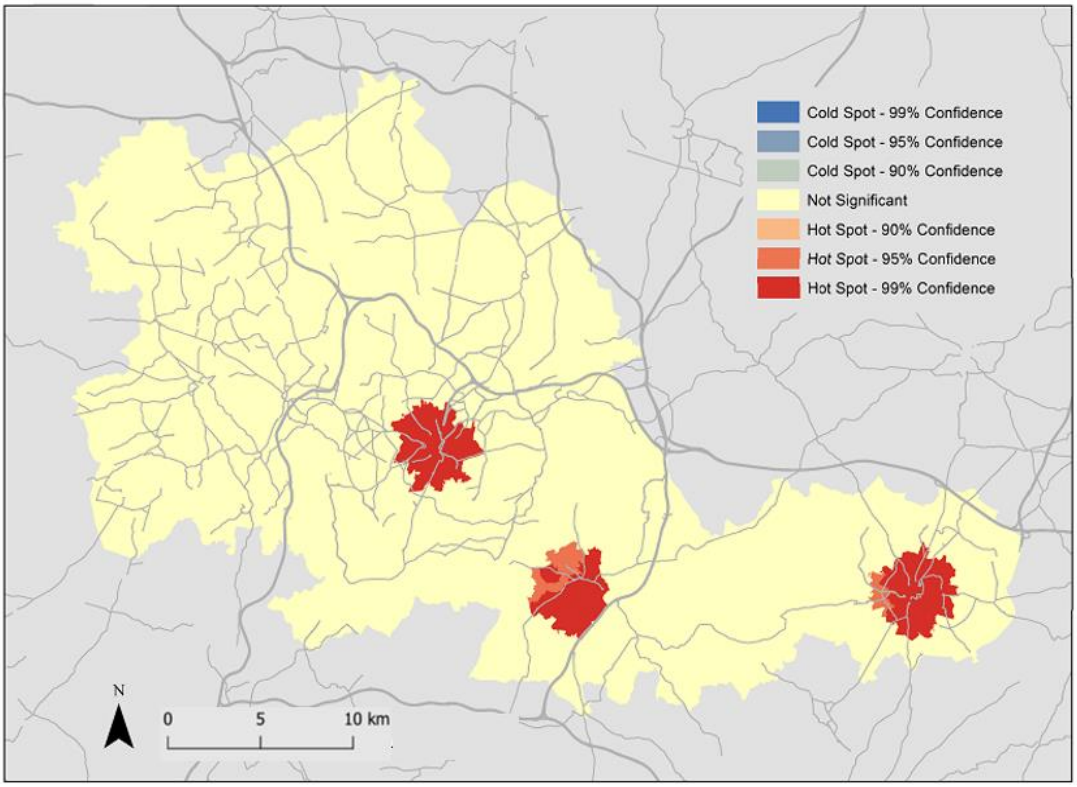
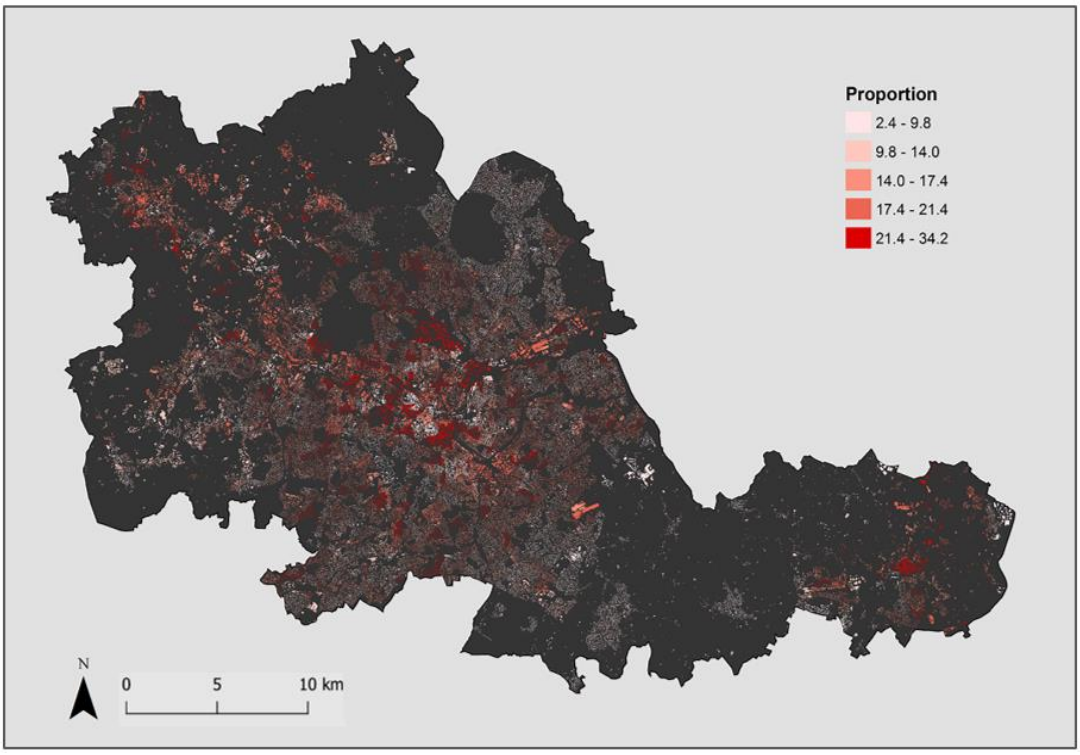


Figure 5.13 Spatial profiles for Cluster 1 showing a) the distribution of residential locations and b) the hotspot activity areas

Cluster 2 – 'Irregular'

a)



b)

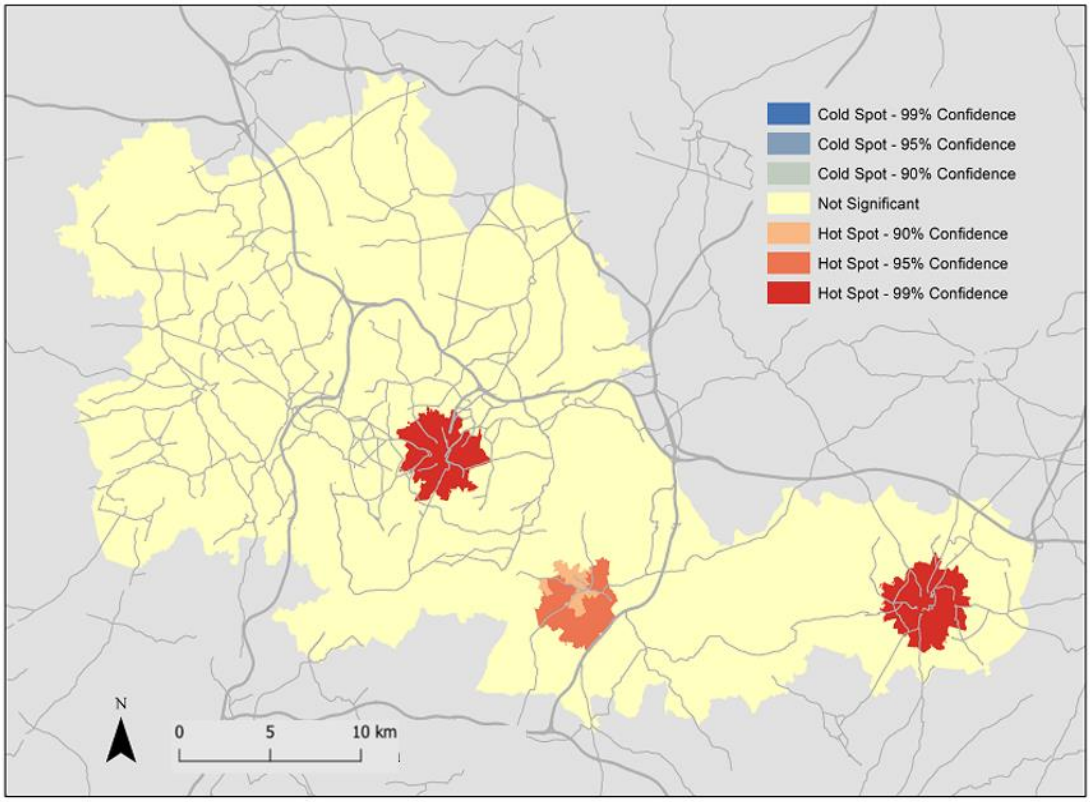
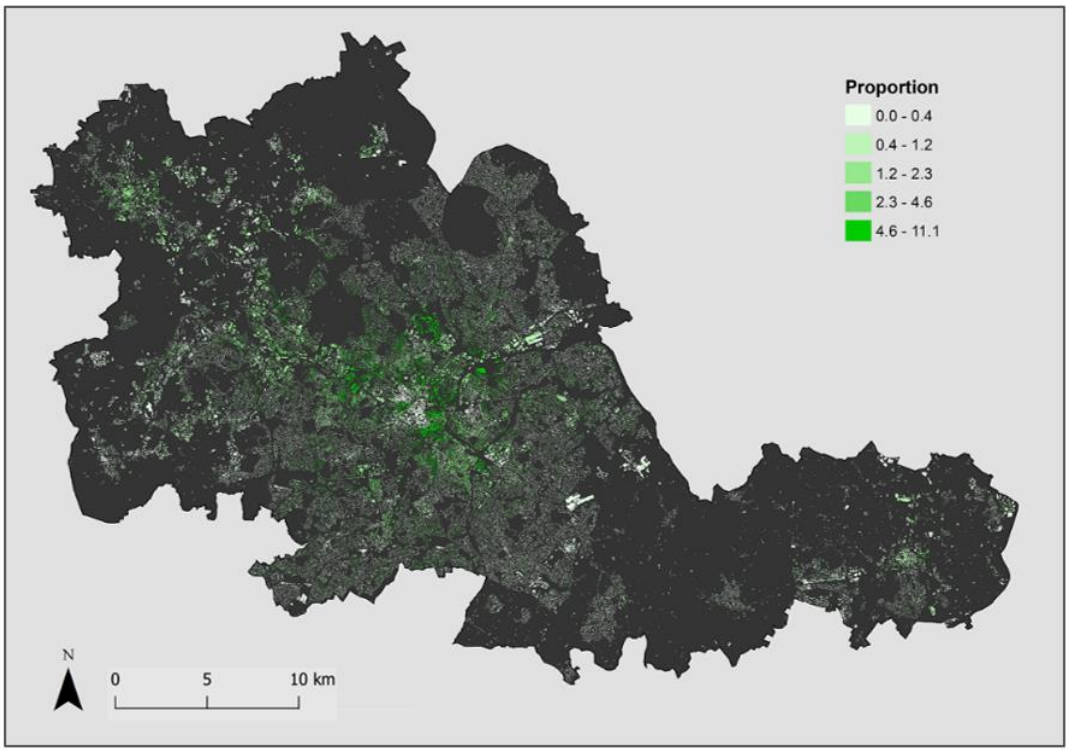


Figure 5.14 Spatial profiles for Cluster 2 showing a) the distribution of residential locations and b) the hotspot activity areas

Cluster 3 – 'Highly Active'

a)



b)

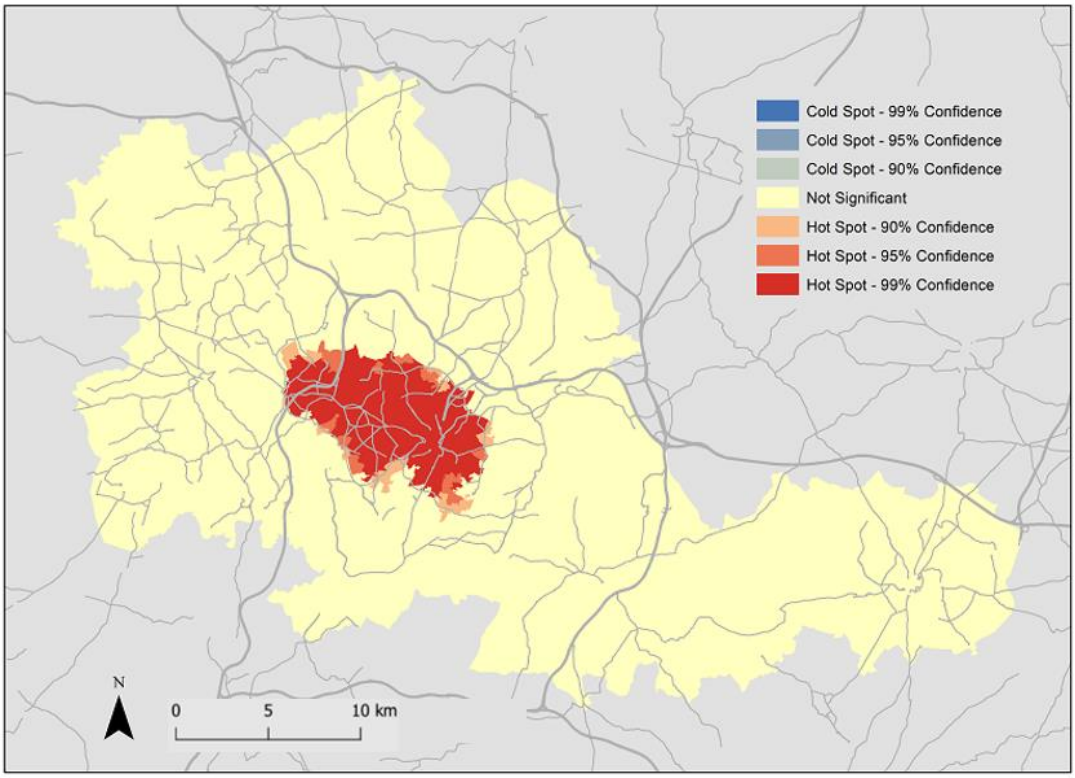
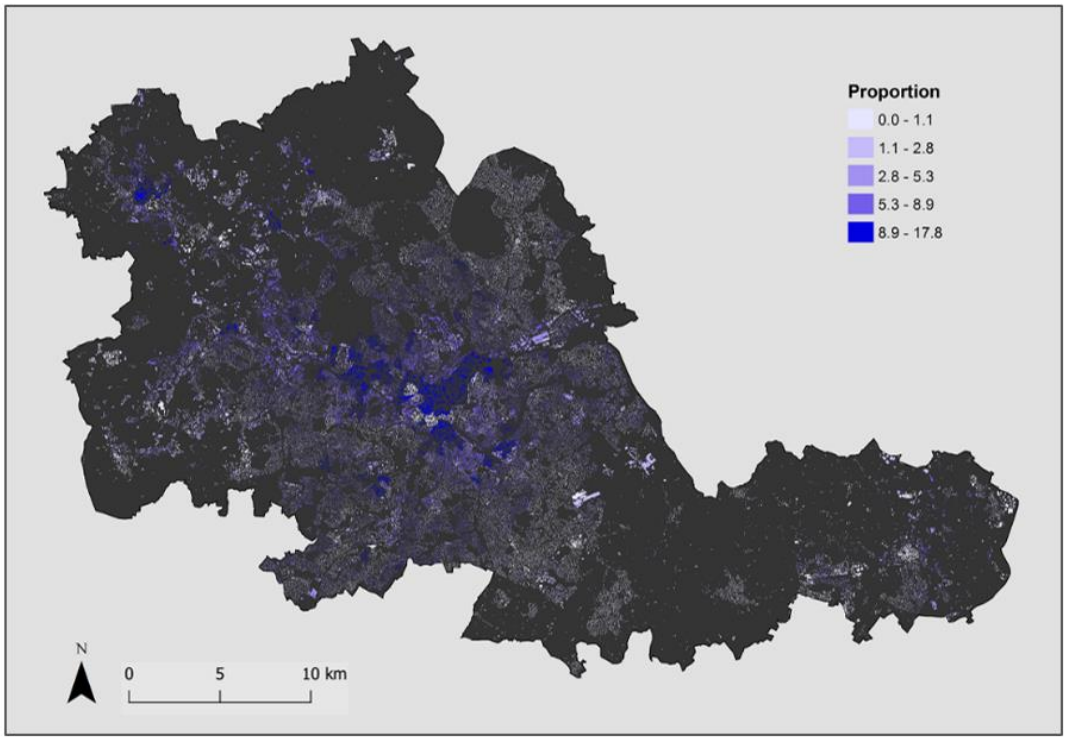


Figure 5.15 Spatial profiles for Cluster 3 showing a) the distribution of residential locations and b) the hotspot activity areas

Cluster 4 – 'Active'

a)



b)

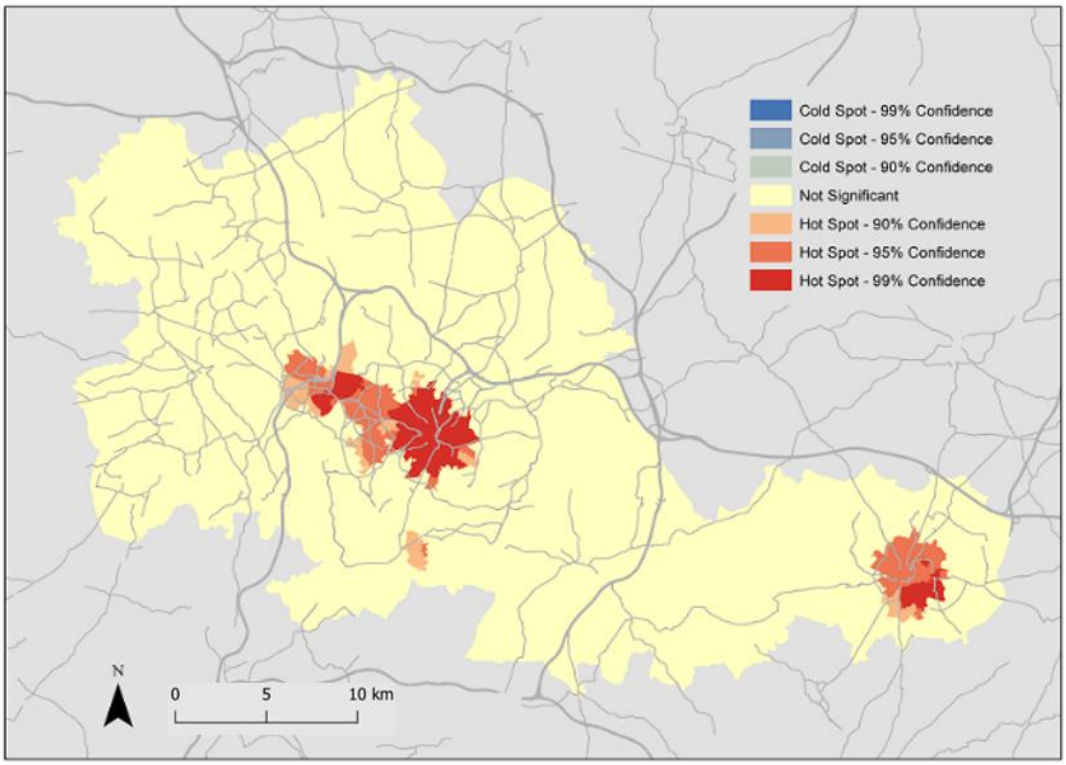
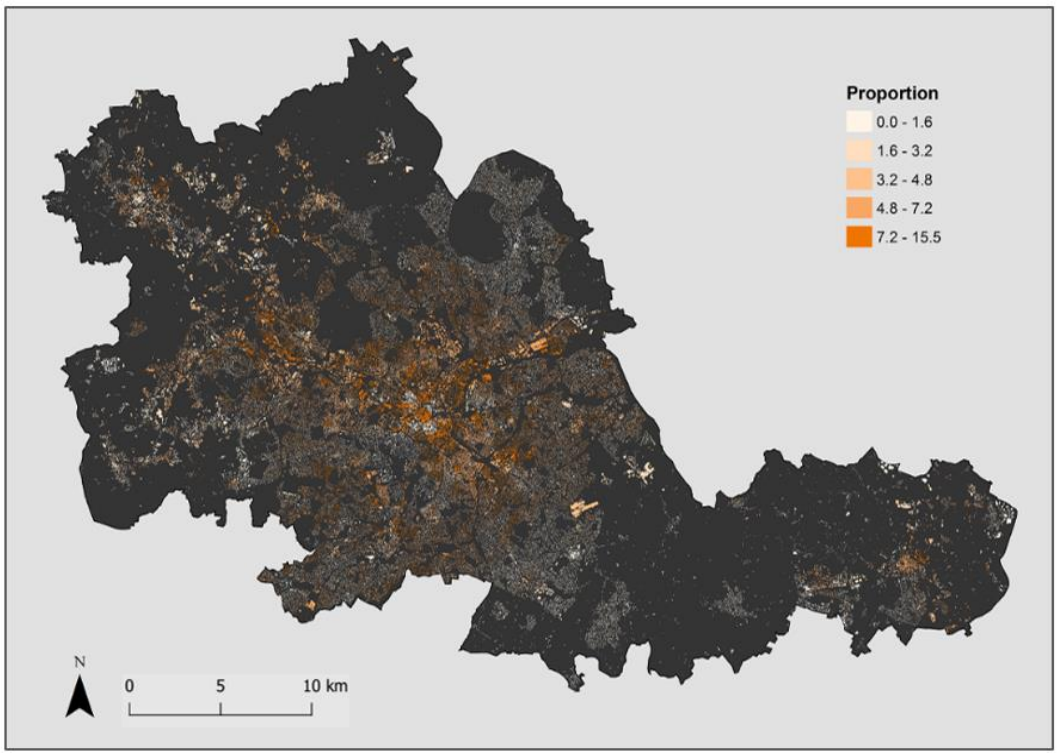


Figure 5.16 Spatial profiles for Cluster 4 showing a) the distribution of residential locations and b) the hotspot activity areas

Cluster 5 – 'Decreasing Usage'

a)



b)

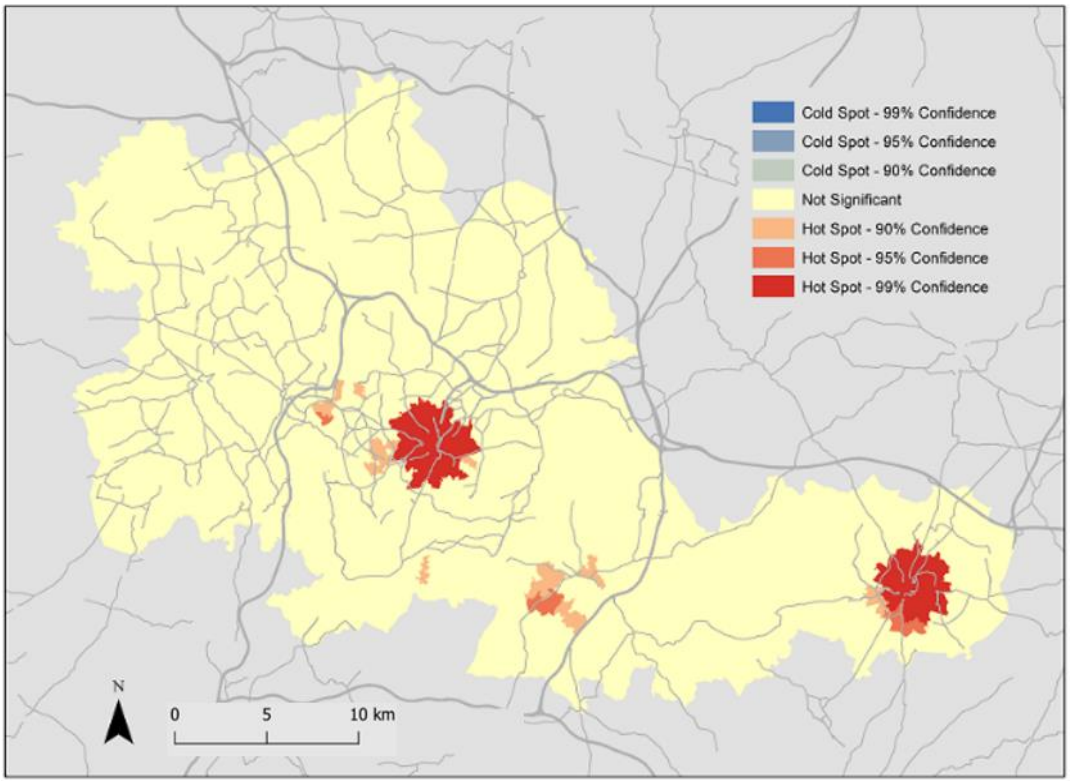
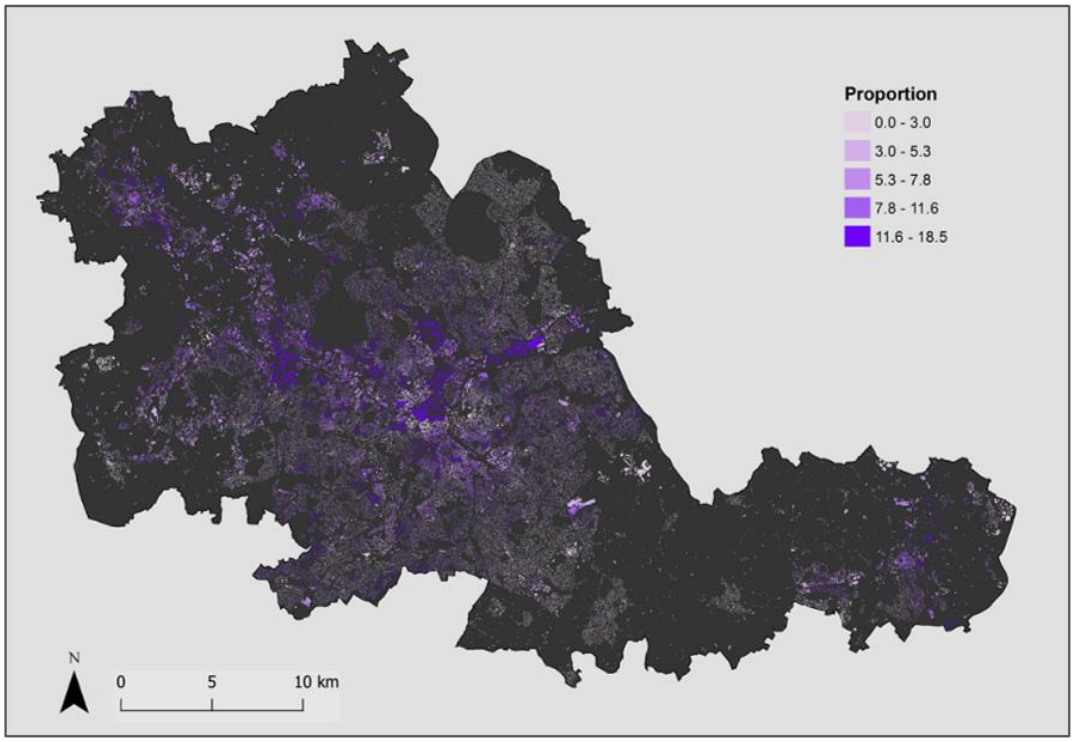


Figure 5.17 Spatial profiles for Cluster 5 showing a) the distribution of residential locations and b) the hotspot activity areas

Cluster 6 – 'Increasing Usage'

a)



b)

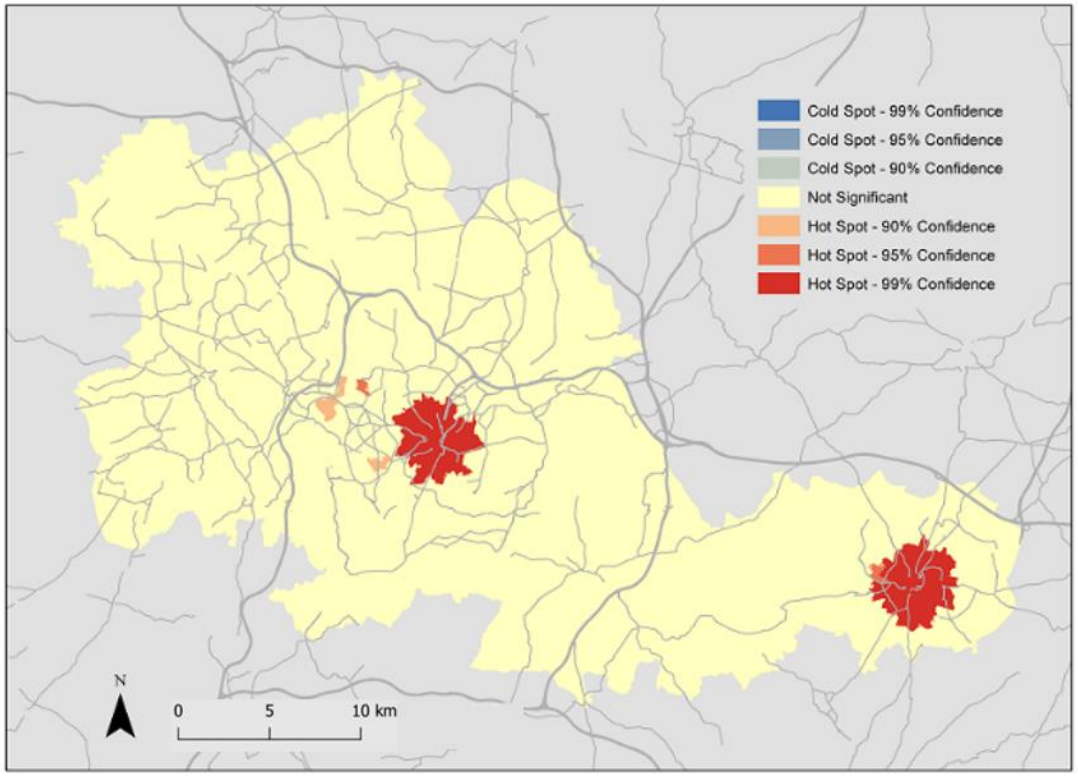


Figure 5.18 Spatial profiles for Cluster 6 showing a) the distribution of residential locations and b) the hotspot activity areas

Comparing the residential locations, it was clear that the distribution of cardholders in Cluster 1 was significantly different to the other clusters. High concentrations of these cardholders were primarily located in more rural and suburban areas throughout the study area. Significant low concentrations were found in highly populated urban areas, in particular central Birmingham. In comparison, Clusters 3 and 4, consisting of the most active cardholders, contained significantly higher concentrations of cardholders residing in central urban areas, in particular Birmingham and Wolverhampton.

Clusters 5 and 6 showed more dispersed spatial distributions, with high concentrations of cardholders distributed throughout the study area and no clear differences evident between those cardholders that had exhibited increasing and decreasing bus usage. This was likely due to these clusters representing cardholders entering and exiting the system, with the age profiles in Figure 5.11 exhibiting the most significant difference between these clusters rather than any other demographic or socioeconomic factors.

These insights suggested that cardholders residing in urban areas were more likely to be highly active bus users. This trend would be expected considering the layout of the bus network, outlined in Chapter 4, with higher concentrations of stops located in central urban areas. Further, these insights were of particular importance for the identification of those cardholders that may lack access to transport services, which is explored further in Chapter 8.

The activity areas represented by the hotspot maps displayed unique patterns of activity between clusters. Comparing the activity areas of all 6 clusters, four main hotspots were evident; central Birmingham, west Birmingham and Sandwell, central Solihull, and central Coventry. The majority of hotspot areas identified for each cluster fell within these areas, however differences in their distributions were evident between clusters.

Cluster 3 showed the most distinctive activity areas in comparison to the other clusters. All statistically significant hotspots for these cardholders were located in and around central and west Birmingham, with no other hotspot areas identified. In comparison, all other clusters showed statistically significant hotspots in central Coventry, which were not evident in Cluster 3.

Those cardholders that were less active on the bus network, contained in Clusters 1 and 2, displayed statistically significant hotspots in central Solihull, where there is a large retail area, in addition to central Coventry and Birmingham. Again, few differences were evident between cardholders with increasing and decreasing bus usage.

There were no cold spots, i.e. areas with significantly fewer visits than would be expected, identified for any of the clusters. This may be due to the aggregation of bus stops to LSOAs. As

most LSOAs contained a number of bus stops, and many of these stops were located in close proximity to LSOA boundaries, the difference between neighbouring areas may not have been as pronounced as had this analysis been conducted at bus stop level.

These spatial profiles clearly demonstrated how smart card data can provide detailed insight into the spatial patterns of bus usage in the study area. The findings suggested that cardholders that exhibit different levels of activity on the bus network may also use the bus network for different purposes. Providing transport authorities with insights into how the bus network is being used by different groups of the population, and the areas that these groups visit, can aid these authorities in identifying areas where demand for services may be high. Supplying an adequate public transport service, not only for the time that users wish to travel, but also to the areas that they wish to travel, is vital for transport authorities and operators. Chapter 6 and Chapter 8 explore these concepts further; analysing the purpose and spatial distribution of trips on the bus network and the accessibility of key facilities via bus services.

5.3.2 Sub-clusters

Trends shown in Figures 5.4 and 5.5 suggested that each cluster displayed both decreasing transaction frequency and decreasing active users over the study period. The second phase of this analysis looked at changes in transaction frequency from each cardholder's monthly average; re-clustering each initial cluster using the 'percentage difference from average' variables outlined in Section 5.2.4. This re-clustering approach was able to identify cardholders that had exhibited more granular changes within each cluster, including those that exhibited increasing, decreasing and steady bus usage over the study period. Focusing on those cardholders within each initial cluster that exhibited either decreasing or increasing usage, Table 5.5 outlines each of these sub-clusters. As Clusters 5 and 6 contained cardholders entering and exiting the system, these were not included in the following analyses.

Table 5.5 Sub-cluster names and descriptions

	Name	Cardholders	Proportion (%)
Cluster 1			
1a	Increasing Usage	44,014	16.5
1b	Decreasing Usage	12,751	4.8
Cluster 2			
2a	Increasing Usage	11,628	18.8
2b	Decreasing Usage	1,277	2.1
Cluster 3			
3a	Increasing Usage	396	11.7
3b	Decreasing Usage	656	19.4
Cluster 4			
4a	Increasing Usage	2,039	13.1
4b	Decreasing Usage	3,604	23.2

Surprisingly, given the trends evident in Chapter 4 and previously in this chapter, a number of cardholders exhibited increasing bus usage over the study period, in particular those cardholders in Clusters 1 and 2. A limitation of the method that was implemented to identify these changes was the disparity that was likely to have manifested between the percentage change of highly active users and less active users. For irregular and rarely seen users, a minimal increase in transaction frequency would have resulted in a much larger percentage change than for those users that were initially more active on the bus network. The greater proportion of cardholders shown to have exhibited decreasing bus patronage in Clusters 3b and 4b are therefore more likely to have contributed to the overall decline in bus boardings recorded by TfWM. Figures 5.19 to 5.22 explore these trends in more detail to validate the results shown in Table 5.5; displaying both the average monthly boarding frequency and the proportion of monthly active users.

5.3.2.1 Sub-cluster bus activity

Cluster 1 – 'Rarely Seen'

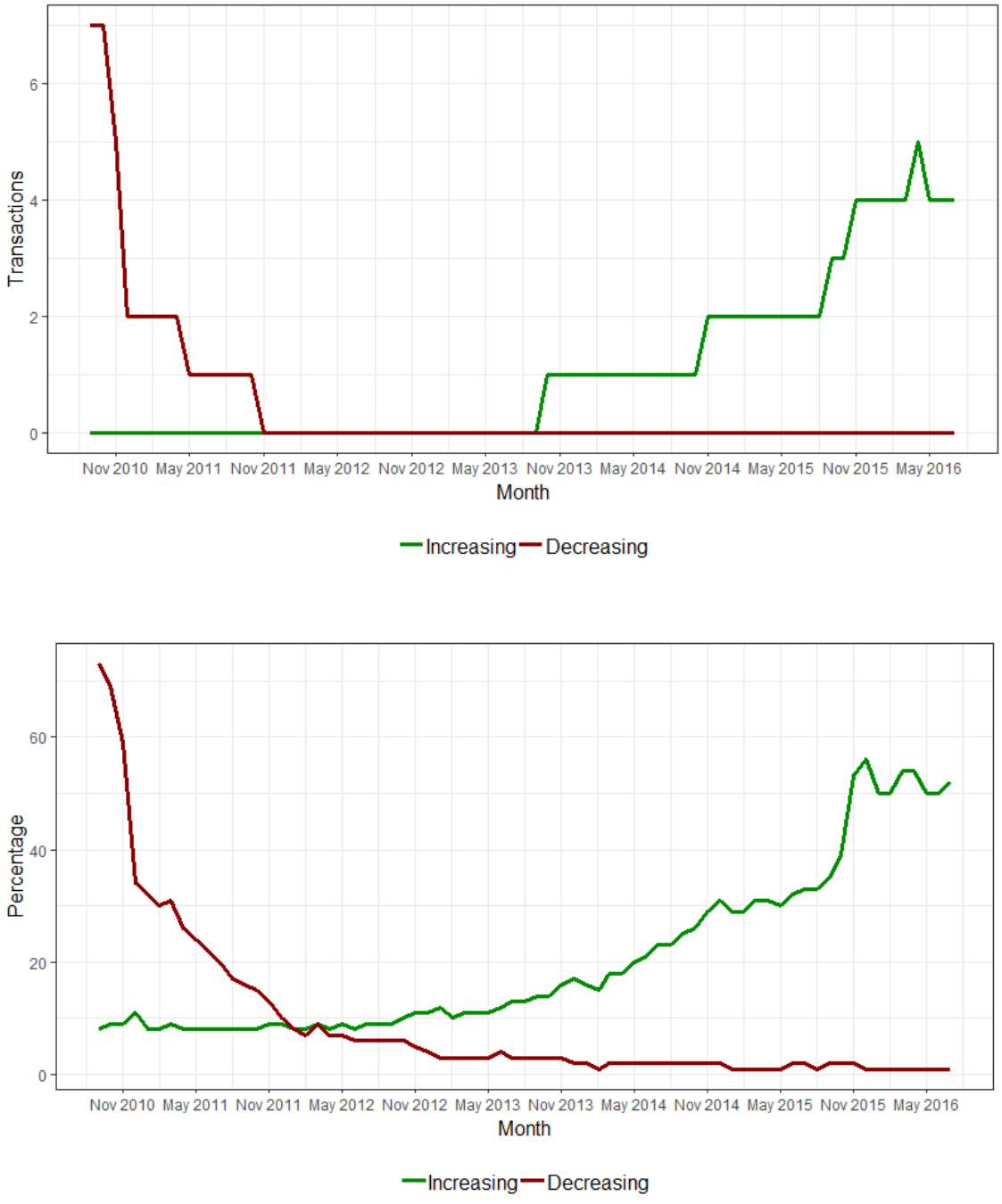


Figure 5.19 Activity level plots for Cluster 1 showing a) the average monthly boardings and b) the proportion that are active users each month

Cluster 2 – 'Irregular'

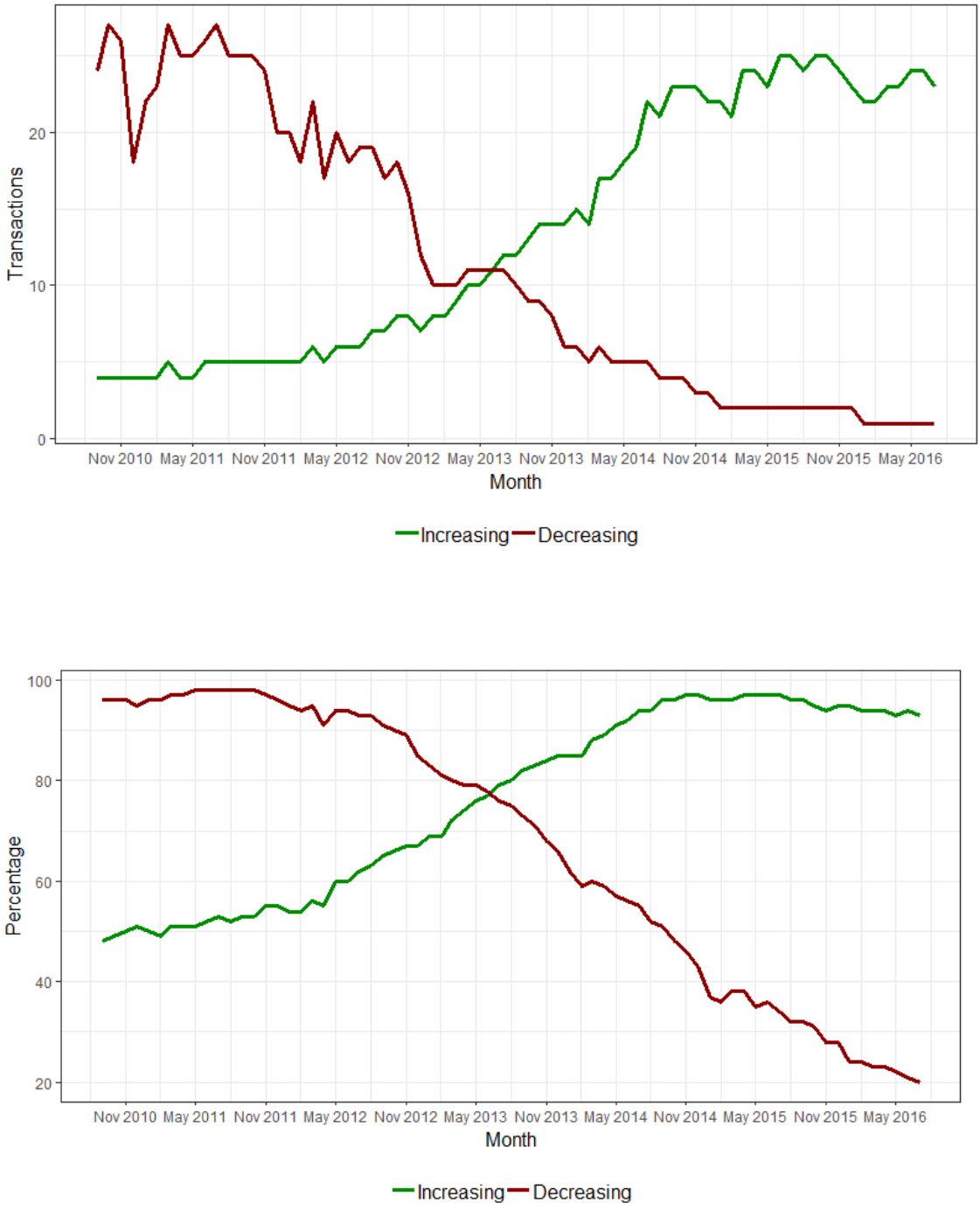


Figure 5.20 Activity level plots for Cluster 2 showing a) the average monthly boardings and b) the proportion that are active users each month

Cluster 3 – 'Highly Active'

Figure 5.21 Activity level plots for Cluster 3 showing a) the average monthly boardings and b) the proportion that are active users each month

Cluster 4 – 'Active'

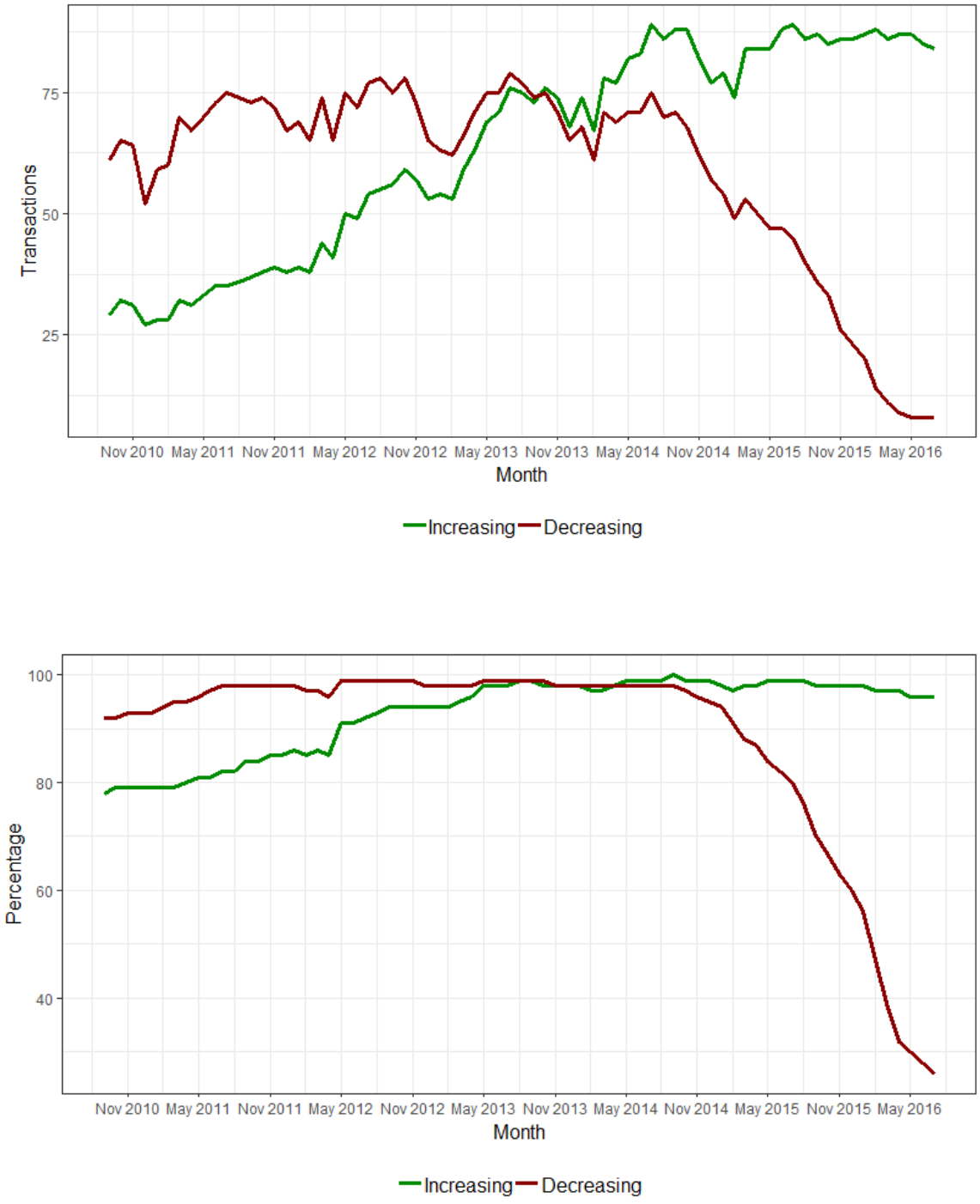


Figure 5.22 Activity level plots for Cluster 4 showing a) the average monthly boardings and b) the proportion that are monthly active users

Figures 5.19 to 5.22 show that distinct groups of cardholders exhibiting both increasing and decreasing bus usage could be identified within each initial cluster. The likely reason that these cardholders were not initially identified in the k-means cluster analysis as exhibiting changes in usage, and therefore assigned to either Cluster 5 or 6, is two-fold; the value of *k* selected for the analysis, and differences in the patterns of bus usage.

Selecting a higher initial *k* value may have resulted in these sub-clusters being assigned to unique initial clusters, with the possibility of multiple initial clusters containing cardholders that exhibited increasing or decreasing usage. Although this may have improved the initial clustering in regard to those cardholders with changing bus usage patterns, a trade-off was struck between gaining detailed insights into bus usage patterns and ensuring the initial clustering resulted in a reasonable number of clusters for ease of interpretation and subsequent analyses. Identifying sub-clusters within the initial cluster assignments also had the benefit of allowing for comparisons between cardholders assigned to different initial clusters yet exhibiting similar trends in changing bus usage.

Differences in the bus usage trends between the sub-clusters and Clusters 5 and 6 may have also contributed to these cardholders not being initially assigned to clusters showing changes in usage. Cardholders exhibiting decreasing bus usage in Clusters 3b and 4b showed a significant decline in both transactions and active users towards the end of 2014. This is in comparison to the trend seen in Figure 5.4, where Cluster 5 shows declining usage throughout the majority of the study period. Figures 5.19 to 5.22 also exhibit sharper and more significant changes when compared to Figure 5.4. These sudden changes in bus usage are particularly significant in Sub-Clusters 1b, 3b and 4b. Understanding the characteristics of these sub-clusters, both in this chapter and following chapters, may go some way to explaining the trends in bus usage seen over the study period.

5.3.2.2 Sub-cluster characteristics

Conducting a two-tier approach to the cluster analysis, i.e. identifying sub-clusters within each initial cluster, allowed for the within-cluster characteristics to be compared. For example, following the initial cluster analysis, it was not known whether the characteristics of highly active users that exhibited decreasing bus usage were similar to those highly active users that exhibited increasing usage, and vice versa. This section discusses these sub-cluster characteristics, with the aim of identifying both the differences between sub-clusters and the similarities in those cardholders that exhibited either increasing or decreasing usage. Analysis of the characteristics of each of the sub-clusters revealed a relatively high uniformity between the

cardholders that displayed increasing and decreasing usage, with some distinct exceptions. For ease of interpretation, only these significant differences are presented and discussed.

The age profiles of each sub-cluster varied significantly, with sub-clusters displaying decreasing bus usage generally containing high proportions of cardholders aged 80 and over. These trends were unsurprising given the results of the initial cluster analysis shown in Figure 5.11. For both Clusters 1b and 2b, those rarely seen and irregular users that have decreased bus usage, the proportion of cardholders aged over 80 was significantly higher than other age groups. In comparison to this, Clusters 3b and 4b showed more dispersed distributions between age groups. Only minimal differences in the proportion aged 80 plus and the younger age groups were found. This was not found to be the case with those cardholders aged 61-65. These results suggested that although old age may result in decreasing mobility and use of the bus network for some cardholders, this was not the sole reason for declining bus usage, in particular for those that were most active on the bus network.

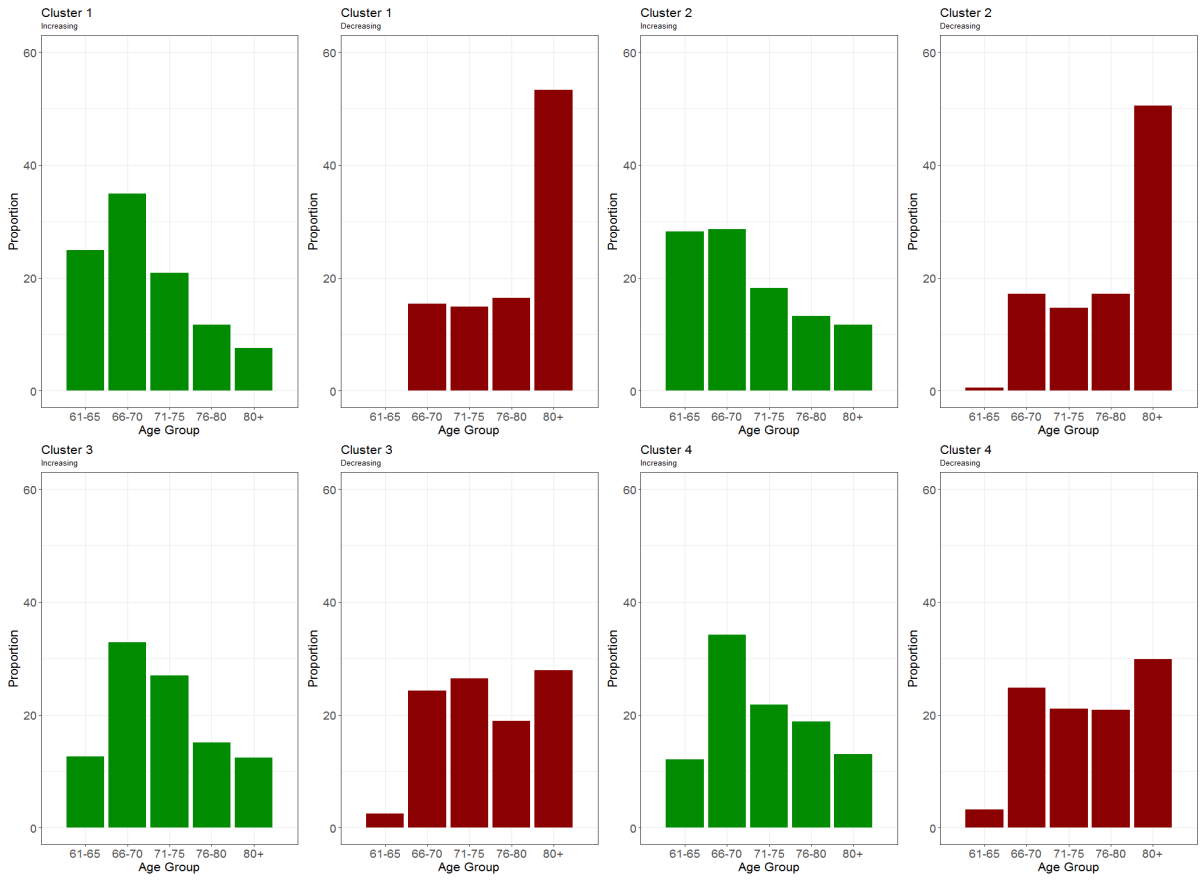


Figure 5.23 Sub-cluster age distributions (5-year age bands)

Despite the significant variances in the age distributions of each sub-cluster, the majority of the demographic and socioeconomic characteristics, outlined in Table 5.3, were similar. The largest differences were evident in the distribution of cardholders living within each IMD decile. Figure 5.24 shows the distribution of cardholders residing in LSOAs of each IMD decile for all 'increasing' and 'decreasing' sub-clusters.

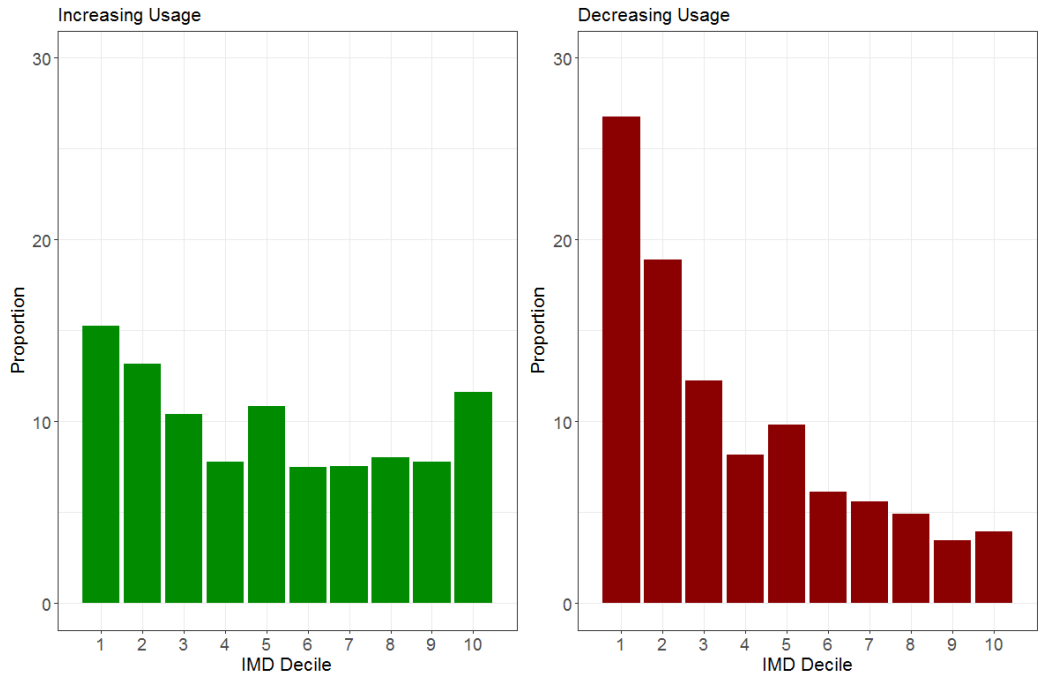


Figure 5.24 'Increasing' and 'Decreasing' sub-cluster IMD distributions

Those cardholders that showed decreasing use of the bus network appeared to be more likely to reside in deprived areas than those that had increased bus usage. On average, 27% of those displaying decreasing bus usage lived in areas assigned to Decile 1, in comparison to 15% of those displaying increasing use of the bus network. These results are of particular relevance for transport authorities aiming to provide an inclusive and accessible transport network for the entire population. More deprived populations are often more at risk of transport disadvantage due to lower levels of private transport availability and less disposable income to take advantage other modes of transport, such as taxis (Lucas 2003). These results therefore suggested that there may be groups of deprived cardholders whose mobility needs are not being met by the current bus network. The reasons for this, including changes in service provision and access to key facilities, are explored further throughout this thesis.

The differences in IMD distributions suggested that there may be significant differences in the spatial distributions of cardholders that had displayed increasing and decreasing activity over the study period. Figure 5.25 shows the total proportion of the eligible population that exhibited increasing and decreasing bus usage for all sub-clusters.

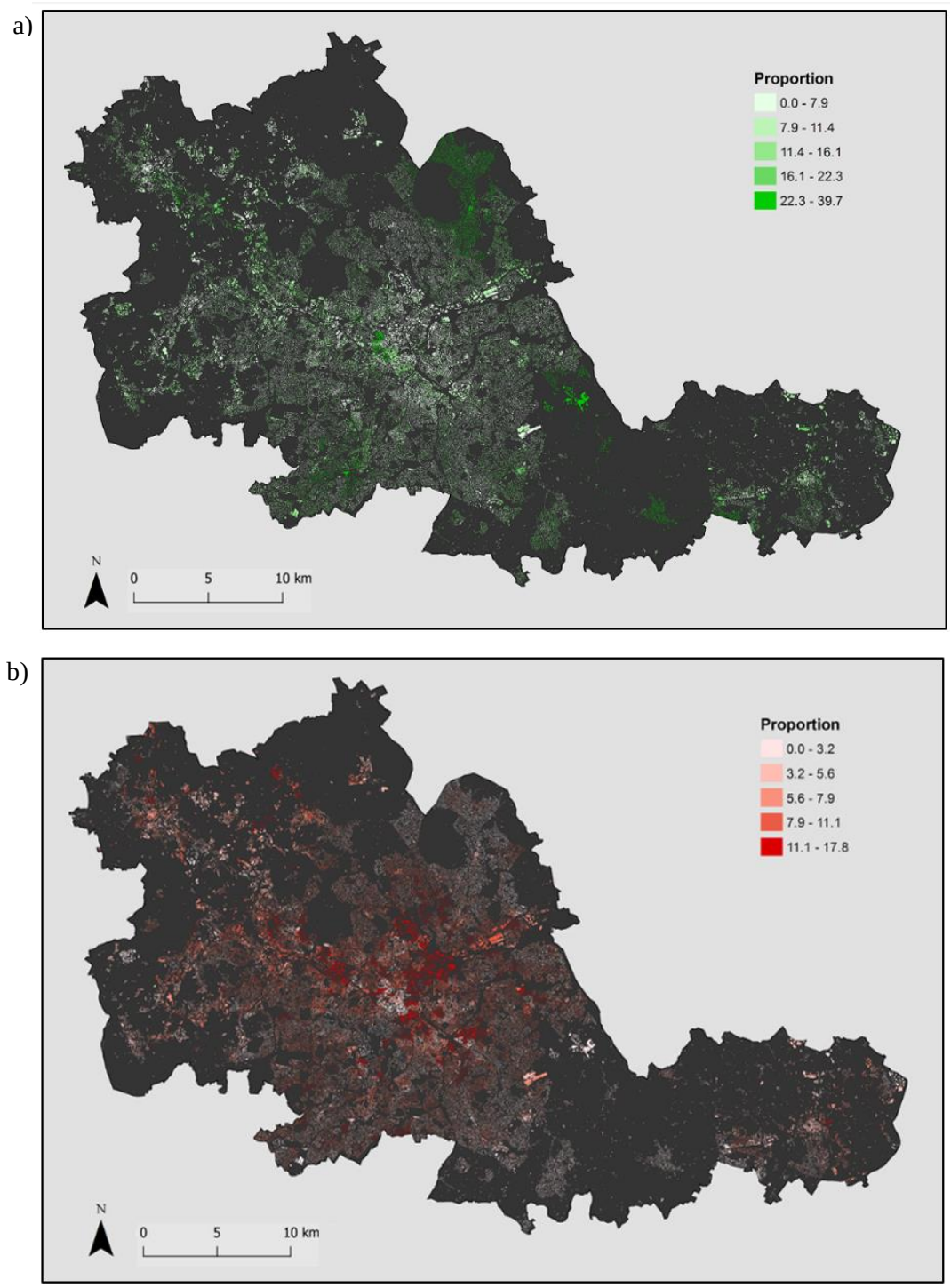


Figure 5.25 Distribution of residential locations for cardholders exhibiting a) increasing bus usage and b) decreasing bus usage

These results suggested that significant proportions of cardholders residing in more rural and suburban locations, such as Sutton Coldfield and Solihull, exhibited increasing bus usage. In comparison, large proportions of cardholders residing in central Birmingham had decreased bus usage over the study period. These differences were not as evident when comparing the spatial profiles of Clusters 5 and 6. Differences in the spatial distributions of cardholders that have increased or decreased their bus usage could help in the targeting of schemes to increase mobility and tackle social exclusion.

5.4 Discussion and conclusions

The aim of this chapter was, firstly, to group cardholders based on their longitudinal activity on the bus network. Secondly, this chapter aimed to identify cardholders that had exhibited more granular increasing and decreasing bus usage over the study period. In addition, the results of this chapter provided a first step in understanding the mobility trends of the older population and were therefore utilised further throughout this thesis.

Insights into the characteristics of different groups of public transport users have, until recently, been largely inferred from qualitative surveys and questionnaires. This chapter demonstrated how smart card data, along with census and administrative data, can be used to effectively infer the characteristics of public transport users. Clear differences were evident in the demographic and socioeconomic characteristics of cardholders that made regular use of bus services and those that had undertaken fewer trips. Understanding these characteristics can assist transport authorities in ensuring that bus services are meeting the transport demands of different groups of users, for example when determining price capping and potential concessionary schemes.

The ‘total-capture’ benefit of smart card data means that cardholders that make little or irregular use of the bus network, or those users that have decreased bus patronage, are still captured in the data. These cardholders are of particular interest for transport authorities aiming to understand trends in public transport usage but are often unable to be captured by traditional data collection methods. The cluster analysis implemented in this chapter was able to identify 6 main groups of cardholders that exhibited similar travel behaviours. These clusters, and their subsequent characteristics, emphasised the heterogeneity of concessionary travellers in the study area. These variations highlight the need for public transport services to cater to the mobility demands of different groups of users. With smart card data able to provide detailed insights into both the characteristics of users that undertake regular trips on the bus network and those that use bus services less regularly, transport authorities will be able to design services to better meet the travel demands of all users.

The ENCTS was introduced with the aim of increasing public transport usage by the older population, and in turn improving their access to services and increasing social inclusion (Mackett 2014). This study, however, identified a significant number of cardholders exhibiting decreasing bus usage for all clusters, suggesting that these aims may not have been realised. If bus services are not catering to the concessionary population in terms of where and when they wish to travel, the success of the scheme is likely to be limited.

Results presented in Section 5.3.2.2 suggested that age and deprivation were the driving forces behind declining bus usage, with older age groups and cardholders living in more deprived areas more likely to have exhibited decreasing use of bus services. Although studies have found that mobility often declines with age (Rye and Scotney 2004), these results may be of concern to transport authorities. With older and more deprived populations most at risk of social exclusion (Lucas 2003), a large decline in bus usage by these groups indicates that current bus services are not meeting their mobility needs. This may have wider implications for groups that are already vulnerable.

Although the general trends presented in Chapter 4 suggested a dramatic decline in bus usage by concessionary travellers, this section identified a significant number of cardholders that had displayed increasing bus usage over the study period. These cardholders were generally in the youngest of the age groups included in this study; likely a result of these users becoming eligible for concessionary travel. Additionally, 'rarely seen' and 'irregular' clusters contained a higher proportion of cardholders exhibiting increasing, as opposed to decreasing, bus usage. This may reflect an increase in the older population making use of multiple modes of transport, with increasing metro and rail boardings (see Figure 3.4) together with an increasing number of driving license holders (see Table 3.1). This suggests an increase in irregular bus usage but a decrease in those that use the bus network as their primary mode of transport. These insights indicate that the decline in bus usage outlined in Chapter 4 and earlier in this chapter may not have been as extreme as initially anticipated, with users both entering and exiting the system, and changes in the age of eligibility largely affecting the slowing rate of users entering the system (see Chapter 4).

5.4.1 Method limitations and future research prospects

Although this chapter demonstrated the utilities of smart card data for understanding mobility patterns, there were a number of limitations to the methods and insights presented. The k-means method implemented for the clustering, although effective, had a number of limitations. Firstly, the initial number of clusters selected greatly impacted the results. As this was largely subjective, an alternative initial *k* value could have been selected which could have produced

different resulting clusters both in terms of membership and characteristics. Additionally, as previously discussed, the monthly temporal aggregations limited the granularity of results and therefore the clustering approaches were unable to detect lesser changes in activity, for example changes in weekday and weekend activity. Unlike a DBSCAN approach, k-means does not identify outliers; allocating all cardholders to the nearest cluster regardless of how well they fit. This may have resulted in some outliers being included in clusters, however, due to the large number of cardholders allocated to each cluster, these outliers were unlikely to have had a significant effect on the aggregate results presented in this chapter. If future analyses were to focus on the individual level, outliers would have to be identified and removed prior to the clustering.

The spatial profiles presented in Section 5.3.1.3, were limited to those trips for which a destination could be inferred by the method outlined in Chapter 4. This is likely to have resulted in a number of trips being excluded from this analysis, and therefore more detailed spatial patterns of activity being overlooked.

In addition, this chapter focused only on those cardholders that had undertaken at least one bus trip over the study period. The cardholder data provided by TfWM contained 501,145 unique accounts, however this analysis was conducted only on the 376,208 cardholders that had transacted over the study period. This means that there were 124,937 individuals that held a concessionary smart card but had never transacted on the bus network. Understanding the characteristics of these cardholders could provide transport authorities with further insight into the groups of the population that do not make use of the bus network. In addition, smart card data from other transport modes, such as rail and metro, could reveal whether these cardholders were active on other public transport services other than the bus. These further analyses could enable the identification of potentially transport-deprived groups of the population; allowing for transport authorities to implement targeted schemes and policies.

This chapter provided an initial classification of bus users; identifying groups of cardholders that had exhibited similar patterns of bus usage over the study period. Further chapters in this thesis utilise these results to gain insights into the mobility patterns of each cluster, and the potential reasons for the evident changes in bus patronage. Whilst this research was limited to only the study area and population, it clearly demonstrated the value of using smart card data to aid understanding of changes in mobility patterns and public transport usage. This can provide valuable insights for the wider transport-research community. The decline in bus usage found throughout Chapters 4 and 5 was not unique to the study area, with the majority of local authorities in the UK experiencing similar trends (White 2009). This chapter demonstrated a method for identifying cardholders exhibiting declining usage, as well as methods for the

inference of their demographic and socioeconomic characteristics, and travel behaviours. The implementation of the ITSO standard on all smart card ticketing throughout the UK means that data preparation and linkage methods are transferable to smart card data held by other transport authorities and therefore this method can be replicated on other study populations or areas.

Chapter 6 – Inferring trip purpose

6.1 Introduction

Chapter 5 identified the cardholders that had contributed to the reduction in bus usage discussed in Section 4.3.2. In addition to understanding which cardholders exhibited decreasing bus usage, understanding the types of trips being undertaken and how this has changed over time can provide valuable insight into how bus services are used by the older population and whether there have been any identifiable changes that may have contributed to the recorded decrease in boardings. The smart card data used in this study can be utilised to answer these questions.

The spatial patterns in the activity areas shown in Section 5.3.1.3 showed distinct differences between clusters. This chapter further explores the idea of activity areas and the spatial distribution of trips in terms of the purpose of trips undertaken on the bus network. The aims of this chapter are therefore two-fold; firstly, to develop a methodology for inferring the purpose of trips recorded in the smart card data, and secondly to identify any changes in the purpose of bus travel over the study period.

When alternative transport data sources, such as smart card data and GPS data, first became available for research, it was expected that these data could substitute for conventional travel surveys. However, a number of limitations have prevented this from being fully realised. One of the main limitations is the lack of information relating to the purpose of trips; information which is usually recorded in traditional surveys and questionnaires. Han and Sohn (2016) stated that smart card data will not replace household surveys until trip purpose can be identified in a reliable manner.

In light of this, a number of studies have attempted to infer the purpose of trips recorded by alternative data sources. Although a number of data sources have been implemented for trip purpose inference (e.g. GPS data, mobile network data), this section focuses on those that have utilised AFC smart card data. The majority of these studies have implemented machine learning algorithms trained on household or travel survey data containing information pertaining to trip purpose, to predict the purpose of trips recorded by AFC systems. Essentially, these methods compare the characteristics of trips recorded in these data, for example trip start time and activity duration, to make predictions on the likely purpose of the trip. Liu et al. (2018) utilised this method; comparing boarding times, activity durations and frequency of visits to the same place between smart card data and household travel survey data. A naïve Bayes probabilistic

model was then used to predict trip purpose. This method was able to predict commuting trips with an accuracy of over 95% and other activities with an accuracy of around 60%.

Chakirov and Erath (2012) used a similar method to detect work, home and 'other' activities using only activity duration. A simple rules-based method was trained on household travel survey data and then used to predict the trip purpose of smart card records. As part of this same study, Chakirov and Erath (2012) also explored the benefit of including land use data in the prediction of trip purpose. A discrete choice modelling approach using activity start time, recorded in both the household travel and smart card data, alongside land use information was implemented to ascertain whether the inclusion of land use information produced more accurate predictions. These methods were able to detect home, work and 'other' activities with a mean accuracy of close to 90%, with the inclusion of land use data found to result in a slight increase in this prediction accuracy.

The majority of studies into trip purpose inference have focused on identifying work, education and home trips (Jun and Dongyuan 2013; Zhou et al. 2014; Chakirov and Erath 2012; Devillaine et al. 2012). Although these previous studies have presented promising results, the focus of this thesis on the older population means that far fewer trips are likely undertaken for work and education purposes and therefore the method used in this chapter takes into account other purposes that may not be as easily distinguishable.

This chapter presents a machine learning method for inferring the purpose of smart card trips using National Travel Survey (NTS) data, with a specific focus on the kinds of trips that are more likely to be undertaken by the older population. A validation method utilising land use data is also implemented, and the spatial distribution of activity areas for each purpose analysed. The purpose of this validation method is to further analyse the accuracy of the trip inference method and assess whether there is a relationship between the urban makeup of the study area and the locations of the inferred trip purposes. Finally, the characteristics of each trip purpose are compared, and the longitudinal trends are analysed.

6.2 Method

6.2.1 Research design

Due to the number of different machine learning algorithms and data variables that could be utilised for this analysis, these had to be tested and analysed first to ensure that the optimum model was selected for the trip purpose inferences. This section presents the analytical design; highlighting the methods and variables that were tested, before subsequent sections outline these in more detail.

1. *Selection of machine learning algorithms to be tested*

Firstly, the machine learning algorithms to be tested on the data were selected. There are many different methods that can be implemented to infer trip purpose, however variations in these methods mean that the accuracy of these inferences can also vary. In order to ensure that the final model implemented the optimum machine learning algorithm, a number of algorithms were tested beforehand. This first section presents each of the algorithms that were selected for comparison; outlining how each algorithm works and the differences between them.

2. *Selection of trip purposes to be predicted*

At the most granular level, the NTS data contained 14 trip purposes. These included purposes such as 'holiday' and 'other including just walk' which were unlikely to be relevant for the smart card data recorded on the bus network. Another variable in the NTS data aggregated these purposes into 8 categories, including 'shopping', 'commuting' and 'education'. Due to the focus of this study on the older population, it was expected that some purposes, in particular commuting and education, would contain few records and therefore there would not have been enough data for these to be included in the final model. This section analyses the amount of data that was available for each trip purpose before deciding and justifying which would be included in the final model.

3. *Selection of the scale at which to conduct the analysis*

The NTS data could be filtered at a number of scales: all regions, metropolitan regions, and the West Midlands metropolitan area. Although data pertaining only to the West Midlands would have been expected to produce the most accurate trip purpose predictions; due to the relatively small sample sizes in the NTS data, expanding the scale of the study would have allowed for more data to be included. This would be expected to lead to improved predictions. To test this, all the selected machine learning algorithms were run on the NTS data at each scale to determine which scale produced the most accurate predictions.

4. *Selection of the variables to be included in the analysis*

The variables to be included in the final model were primarily dictated by the common variables that were present in both the NTS data and the smart card data, for example trip start time and trip distance. However, as not all of these variables would significantly improve the model, the

importance of each variable was tested to determine which should be included in the final model.

5. *Selection of the optimum machine learning algorithm*

Once the trip purposes to be predicted, the scale of the data and variables to be included had been selected, the machine learning algorithm to be run on the smart card data was selected. This was selected by comparing the prediction results that each algorithm produced when run on a validation dataset.

6.2.2 Machine learning algorithms

Machine learning refers to the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions; instead relying on patterns and inferences (Bishop 2006). Machine learning algorithms build a mathematical model based on sample data, or 'training data', in order to make predictions or decisions without being explicitly programmed to perform the task. In this case, NTS data were implemented as the training data and were used to make trip purpose predictions for the transactions recorded in the smart card data.

For this analysis, six different commonly used machine learning algorithms were implemented and compared; linear discriminant analysis (LDA), classification and regression trees (CART), k-nearest neighbours (kNN), support vector machines with linear kernel (SVM), random forest (RF) and naïve Bayes (NB). These algorithms were selected as they provided a mixture of simple linear (LDA), non-linear (CART, kNN), complex non-linear (SVM, RF) and probabilistic (NB) methods.

The aim of each of these algorithms is the same; to use an object's characteristics to identify which class it belongs to. However, there are key differences between them. Linear classifiers achieve this aim by making a classification decision based on the value of a linear combination of the characteristics, whereas non-linear classifiers are generally used for classifying non-linearly separable data. Further, probabilistic classifiers predict a probability distribution over a set of classes, rather than the most likely class that each observation belongs to (Guo-Xun et al. 2012).

Linear discriminant analysis (LDA)

LDA is a method used to find a linear combination of features that separates two or more classes of objects, in this case trip purpose (Bishop 2006). This method works by calculating the statistical properties of each variable, for each class. In the case of this research, the variables were the characteristics of each trip e.g. trip start time, and each class was the trip purpose e.g. shopping. The first step was to calculate the separability between different classes in the training data, i.e. the distance between the means of different trip purposes. This is called the between-class variance. The second step was to calculate the distance between this mean and the samples of each class in the training data. This is called the within-class variance. The third step was to construct a lower dimensional space which maximises the ratio of the between-class and within-class variance, i.e. maximises the between-class variance and minimises the within-class variance (Tharwat et al. 2017). In the case of this study, these steps were run on the NTS data. The results were then used to classify the purpose of trips recorded in the smart card data by allocating them to their closest class based on the mean of the trip characteristic variables. This method is most appropriate for large sample sizes and normally distributed variables.

Classification and regression trees (CART)

CART is a commonly used method for predicting the value of a target variable based on several input variables. CART methods work by creating a decision tree that aims to correctly classify members of a population by splitting it into sub-populations based on several dichotomous independent variables, i.e. each "branch" is split into two (Breiman et al. 1984). For the analysis in this chapter, the branches were the trip characteristic variables whilst the class labels, also known as the "leaves", were the trip purposes. This algorithm was trained on the NTS data, with all input variables and all possible split points evaluated so that the best split point was chosen each time. To determine the best split point, the Gini index was used which provides an indication of how mixed the training data assigned to each node are (Berk 2008). A node that has all classes of the same type would have G=1 whereas a node where there is a perfect split between two classes would have G=0.5. Once the CART algorithm had been trained on the NTS data, this was then run on the smart card data; classifying trip purpose by assigning trips to the nodes. CART algorithms require no specific data preparation to be undertaken, however, they can often be unstable with a small change in data leading to a large change in the structure of the decision tree which can produce inaccurate results (Hayes et al. 2015).

K-Nearest Neighbours (kNN)

kNN is a non-parametric method used for classification. For this method, an observation is classified by a similarity measurement; with the observation being assigned to the class most common among its *k* nearest neighbours measured by a distance measure, for example Euclidean. '*k*' refers to the number of nearest neighbours, so for a *k* value of 1, the observation is assigned to the class of its single nearest neighbour (Everitt et al. 2011). In the case of this research, the purpose of trips recorded in the smart card data were assigned based on their Euclidean distance to the trip purpose classes extracted from the NTS data. This method is sensitive to the local structure of the data and data measured on different scales must be standardised before the algorithm is run (Gou et al. 2012).

Support vector machines with linear kernel (SVM)

SVM is a linear model used for the classification of data. The first step of an SVM model is to find a separating line between different classes, in this case trip purposes. The SVM algorithm then finds the points of each class that lie closest to the line and measures the Euclidean distance. This distance is called the margin, with the goal of the SVM algorithm to maximise this margin and therefore maximise the difference between classes (Ben-Hur 2009). For data that are not initially linearly separable, the SVM algorithm converts these data to linearly separable data in a higher dimension. For this research, the SVM algorithm found a separating line between the trip purposes recorded in the NTS data. When the smart card data are inputted, these lines were used to separate the data into trip purpose classes.

Random forest (RF)

Random forests (RF) are an ensemble learning method for classification. Although random forests are part of the CART category of algorithms, these differ from the CART method applied in this study as the RF method builds multiple decision trees. One of the drawbacks of learning with a single tree is the problem of overfitting; single trees tend to learn training data too well resulting in poor prediction performance on unseen data (Steinberg 2009). RFs can correct for this by building many independent trees that are learned from the same training data, with each tree containing its own 'out-of-bag' sample, i.e. data that were not used during construction (Breiman 2001). These data are then used to get an unbiased estimate of the classification error as trees are added. The fundamental idea of the RF algorithm is to combine many decisions trees into a single model to increase the accuracy of predictions.

Naïve Bayes (NB)

NB classifiers are a type of probabilistic classifier. These are based on Bayes' Theorem which describes the probability of an event, based on prior knowledge of conditions that might be related to the event, i.e. the probability of *A* happening given that *B* has occurred (Joyce 2003). NB classifiers calculate the probability of an observation belonging to each class given each feature of the observation, whilst assuming that the value of a particular feature is independent of the value of any other feature (Rish 2001). In this case, the NB algorithm used the NTS training data to calculate the probability of a trip belonging to each trip purpose based on the individual characteristics of the trip.

This study tested each of these algorithms in terms of their accuracy in predicting the trip purpose of records in the training data. The aim of comparing each of the algorithms was to ensure that the final model produced the most accurate trip purpose predictions. To compare this, both the accuracy and the Cohen's kappa coefficient of each algorithm were analysed. The accuracy is the measure of how 'correct' the results are, i.e. the percentage of inferences that match the known data. The Cohen's kappa coefficient is a measure of inter-rater reliability, and therefore tells us the degree of agreement between results when the algorithm is run on the data several times (Landis and Koch 1977). Both of these measures were compared between each of the algorithms to construct an optimum final model to be run on the smart card records.

6.2.3 National Travel Survey data

Like many previous studies, this analysis utilised travel survey data to train the machine learning algorithm. In this case, NTS data were used (outlined in Table 4.4). The NTS is a yearly household survey capturing data on personal travel and changes in travel behaviour over time. It was first conducted in 1965 and has been conducted on an annual basis since 1988. However, in order to correspond with the smart card data, this analysis used all data recorded between 2009 and 2016. This NTS data contained information on the purpose of trips that respondents had undertaken, in addition to a number of variables relating to the characteristics of these trips and the demographics of respondents. The NTS defines trips as a 'one-way course of travel with a single main purpose'. Trips can consist of several stages which are defined as when there is a change in the form of transport or a change of vehicle requiring a separate ticket (DfT 2016). At the highest level, the NTS classified trips into 8 different purposes: 'commuting', 'business', 'education/ education escort', 'shopping', 'other escort', 'personal

business', 'leisure' and 'other'. General definitions of these trip purposes are shown in Table 6.1.

Table 6.1 NTS 'trip purpose' variable definitions

Trip Purpose	Definition
Commuting	Trips from home to a usual place of work or from usual workplace to home
Business	Personal trips in course of work
Education/ education escort	Trips to school or college, or trips to accompany someone to school or college
Shopping	Trips to the shops or from the shops to home
Other escort	Trips made to accompany someone else somewhere, excluding school or college
Personal business	Visits to services, such as banks and hairdressers, and medical consultations etc.
Leisure	Mostly entertainment, sport, holidays and day trips
Other	Any other trips

Although a number of variable combinations and aggregations could have been implemented, for the model to be run on the smart card data the NTS variables that were able to be used were limited to those that were contained within both datasets. These variables and their descriptions are outlined in Table 6.2.

Table 6.2 Description of variables contained in both the smart card and NTS data

Trip Characteristics	Description	Measure
Travel day	Day of the week that the trip took place	Measured from 0-6 corresponding to Sunday-Saturday
Trip distance	The total trip distance recorded in kilometres	12 bins ranging 1 to 12 corresponding to 'Under 1 km' to '200+ km',
Start time	The time of the day that the trip began	Hourly time bands, recorded in bins ranging from 1 to 24
End time	The time of the day that the trip ended	Hourly time bands, recorded in bins ranging from 1 to 24

Travel time	The total time spent travelling	14 bins ranging from 1 to 14 corresponding to ‘Less than 3 minutes’ to ‘6+ hours’
Main mode of transport	The main mode of transport used for the trip	13 categories including local bus, recorded as numerical values ranging from 1 to 13 with each corresponding to a mode of transport
Trip purpose	The main purpose of the trip	8 categories (shown in Table 6.1), recorded as numerical values ranging from 1 to 8 with each corresponding to trip purpose
Demographic Characteristics		
Age	The age of the respondent	5 year age bands ranging from ‘60-64’ to ‘85+’ and recorded as numerical values ranging from 16 to 21
Sex	The sex of the respondent	Either ‘male’ or ‘female’, recorded as 1 for males and 2 for females
Home location	The region where the respondent is a resident	Broken down into metropolitan and non-metropolitan areas, includes 15 regions (e.g. ‘West Midlands metropolitan’) with numerical values ranging from 1 to 15 corresponding to each region

In addition to these variables, the dwell time between trips recorded in the NTS data could also be calculated. Trips undertaken by an individual on a specific travel day were given the same ‘Day ID’ and a ‘Journey Sequence’, referring to the order in which this trip took place over the travel day. These data, along with the journey start and end times outlined in Table 6.2, allowed for the time difference between subsequent trips to be calculated. Firstly, data were grouped by their unique ‘Individual ID’ and ‘Day ID’ and sorted by ‘Journey Sequence’. Then, using the *DATEDIFF* function in SQL, the difference, in minutes, between the end time of a trip and the subsequent start time of the next trip in the sequence was calculated. This resulted in a new ‘dwell time’ variable to be created for all trips except the last trip undertaken each day, for which a subsequent trip start time was not available. This variable was then aggregated into the same bins as the ‘travel time’ variable outlined in Table 6.2.

As the smart card data only relates to bus trips, the NTS data were filtered to only include those trips where the main mode of transport was recorded as 'local bus'. It should be noted that a separate variable, 'Bus in London', was not included in the analysis due to the focus of this study on the West Midlands and the uniqueness of London's transport network. In addition, due to the unique travel behaviour of the older population, data were filtered to only include those that were recorded as being age 60 plus and eligible for concessionary travel. Table 6.3 shows the number and percentage of trips for each purpose contained in these data.

Table 6.3 Total number and percentage of trips for each trip purpose

Trip Purpose	**Total trips**	**Percentage**
Commuting	3,070	4.29
Business	298	0.42
Education/ escort	237	0.33
Shopping	38,142	53.32
Other escort	619	0.87
Personal business	11,512	16.09
Leisure	17,657	24.68

Due to the low number of 'business', 'education/ escort' and 'other escort' trips contained in the NTS data, these trip purposes were removed from the analysis. Although it is acknowledged that this would result in no trips being inferred for these purposes, to optimise the accuracy of the final model, this trade-off was deemed necessary. As this method relied upon identifying common characteristics between trips, including those low recorded trips would limit the algorithm's ability to identify these similarities. An alternative approach could have been to group these purposes into a singular 'other' category. However, as these different trip purposes likely had distinct characteristics, grouping these into a single category would have increased the within-category differences and therefore limited the machine learning algorithm's ability to determine the common characteristics of these trip types in order to classify them.

The final NTS dataset therefore consisted of all recorded trips for commuting, shopping, personal business and leisure purposes. Table 6.4 shows an example of the final NTS dataset, with variables shown in the bins recorded in the NTS data.

Table 6.4 Formatted National Travel Survey data

Travel Day	Trip Distance	Start Time	End Time	Total Time	Dwell Time	Age Group	Sex	Trip Purpose
7	4	15	15	6	8	16	2	Shopping
5	7	12	14	9	3	16	2	Commuting
5	7	17	18	8	11	19	1	Leisure
1	5	7	8	6	7	17	2	Personal Business

6.2.4 Model accuracy and optimisation

Selecting the optimum algorithm for the final model, as well as the data to be included in the model, was vital to ensure that the trip purpose inferences were as accurate as possible. This section presents the method used for this optimisation, in addition to the results of each analysis. This section is therefore divided into three: an outline of the method used to compare each model, the spatial scale to be used, and the selection of the optimum variables to be included in the final model.

6.2.4.1 Algorithm comparison method

The method used for assessing the accuracy of each model relied on the NTS data outlined in the previous section. As the NTS data contained information on the purpose of each trip, these data could be used to assess how accurate each model is at correctly predicting trip purpose. In order to test the accuracy of each of the models, training and validation datasets were created from the NTS data. The training data were used for 'training' the machine learning algorithms, with the validation dataset used to assess how accurate each of the models were when predicting trip purpose. For this analysis, 80% of the NTS data were used as the training dataset and 20% used as the validation dataset. This 80/20 split is a commonly used ratio in machine learning; derived from Pareto's principle which states that for many events roughly 80% of the effects come from 20% of the causes (Newman 2004).

The training data were split into 'input' data, consisting of the trip characteristic variables, and 'output' data, consisting of the trip purpose variable. This analysis then used a 10-fold cross validation, which splits the training data into 10 parts; training on 9 and testing on 1. This

method was repeated 3 times for each algorithm with different splits of the data in an effort to get the most accurate inferences.

The accuracy of each model was assessed using the 'accuracy' metric, where the ratio of the number of correctly predicted instances was divided by the total number of instances in the dataset to give a measure of how 'correct' the results were. Additionally, Cohen's kappa coefficient, which measures inter-rater agreement, i.e. the degree of agreement between results when the algorithm is run on the data each of the 3 times, was used. Landis and Koch (1977) characterised values of <0 indicative of no agreement, 0 – 0.2 as slight, 0.21 – 0.4 as fair, 0.41 – 0.60 as moderate, 0.61 – 0.8 as substantial, and 0.81 – 1 as almost perfect. Alternatively, Fleiss (1981) characterised kappa values of >0.75 as excellent, 0.4 – 0.75 as fair to good, and below 0.4 as poor. It should be noted that these guidelines are by no means universally accepted and are instead based on personal opinion.

Once this method was run on the training dataset and the algorithm with the highest accuracy was identified, predictions could be made on the validation dataset. This validation data consisted of 20% of the original NTS data and gave an independent final check of the accuracy of the 'best' model before it was run on the smart card data. The validation data were inputted the same way the smart card data were, and the algorithm that had been trained on the training data then predicted the trip purpose of the records included in the validation data. As these validation data were taken from the NTS data the trip purpose is known. Therefore, the accuracy of the models could be deduced by comparing these inferences to the known trip purposes.

6.2.4.2 Scale selection

The NTS survey data included the details of trips undertaken by residents in England. As part of the demographic information collected, the respondent's region and statistical region were recorded. This means that the data could be filtered for different regions. For example, data could be limited to just residents of the West Midlands metropolitan area. Ensuring that the population included in the NTS data was as similar as possible to the population in the smart card data was expected to result in the most accurate results. However, due to limitations in the number of respondents included in the NTS data, it was also recognised that increasing the study area to include all metropolitan regions or to include all respondents in England could produce more accurate predictions.

As the optimum variables to be included in the analysis had not yet been determined, the following analyses included all common variables for the NTS and smart card data, outlined in

Table 6.5. Three scales were analysed; all regions, all metropolitan regions and just the West Midlands metropolitan region.

Table 6.5 Number of respondents and trips at each spatial scale

Scale	Respondents	Trips
All regions	8,016	35,435
All metropolitan regions	2,434	12,514
West Midlands metropolitan region	538	2,978

The validation method was run on the data for each of these regions to compare the accuracy and the Cohen's kappa coefficient of each, shown in Table 6.6.

Table 6.6 Maximum accuracy and kappa values for data at each spatial scale

Scale	Accuracy	Kappa
All regions	0.59	0.21
All metropolitan regions	0.61	0.27
West Midlands metropolitan region	0.67	0.41

The results shown in Table 6.6 clearly suggested that predictions based only on data from residents of the West Midlands metropolitan area resulted in both the highest accuracy and the highest Cohen's kappa coefficient. An accuracy of 0.67 meant that 67% of the predictions were correct when compared to the known trip purposes, and the Cohen's kappa coefficient of 0.41 suggested that there was moderate agreement. As the optimum variables to be included in the analysis and the optimum algorithm to be used had not yet been determined, these results were likely to improve. Although limiting the data to just the West Midlands metropolitan area reduced the amount of data that could be used for training the model, due to the significant difference between the accuracy and kappa when compared to the other scales, it was determined that further analyses would only utilise data from the West Midlands metropolitan area.

6.2.4.3 Variable selection

To test which variables should be included in the analysis, the importance of each variable was calculated. "Importance" refers to the score that can be assigned to features based on how useful they are at predicting a target variable (Kuhn 2019). Measures of importance differed between each of the algorithms:

ROC Curve Analysis – LDA, kNN, SVM, NB: A receiver operating characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied (Kuhn 2019). This is created by plotting the true positive rate, i.e. the proportion of actual positives that are correctly identified as such; also known as the *sensitivity*, against the false positive rate, i.e. the proportion of actual positives that are incorrectly identified as such; known as the *probability of false alarm* (Kuhn 2019). ROC analysis uses the maximum area under the ROC curve as a measure of importance for each variable, as the greater the area under the ROC curve, the greater separation between positive and negative results. The greater the distance between these positive and negative results is equivalent to bringing closer together samples belonging to the same class and separating groups belonging to different classes (Hanley and McNeil 1982). In the case of this analysis, these classes were trip purpose and the ROC curve analysis therefore computed how important each variable was for distinguishing between purposes.

Recursive Partitioning – CART: Recursive partitioning is a statistical method for multivariable analysis, such as CART. These methods work by creating decision trees that split populations into sub-populations with the aim of correctly classifying members. To calculate the importance of each trip characteristic variable the reduction in the mean squared error, i.e. the average squared difference between the estimated values and the actual value (Schluchter 2005), attributed to each variable at each split is tabulated and the sum returned. The greater the reduction in the mean square error, the more important the feature. As there may also be variables that are important but not used in the split, this method also tabulates the top competing variables at each split (Kuhn 2019).

Random Forest – RF: RFs are an ensemble learning method for classification similar to CART but are made up of multiple independent decision trees. To calculate the variable importance in RFs the prediction accuracy on the out-of-bag data of each tree is recorded, and then the same is done after removing each predictor variable in the out-of-bag sample. The difference between

the two accuracies are then averaged over all trees and normalized by the standard error to give an importance value for each variable (Kuhn 2019); in this case, each trip characteristic variable.

All measures of importance are scaled to have a maximum value of 100 and a minimum of 0. The results of these analyses are shown in Table 6.7.

Table 6.7 Variable importance for each machine learning algorithm (variables of high importance are highlighted)

	ROC Curve				Recursive Partitioning	Random Forest
Variable	**LDA**	**kNN**	**SVM**	**NB**	**CART**	**RF**
Dwell Time	91.8	91.8	91.8	91.8	90.7	100.0
Start Time	81.4	81.4	81.4	81.3	100.0	82.0
End Time	80.5	80.5	80.5	80.5	83.3	69.3
Age Group	69.5	69.5	69.5	69.6	11.4	42.4
Travel Time	35.5	35.5	35.5	35.6	22.9	40.3
Trip Distance	34.1	34.1	34.1	34.2	0.0	42.6
Sex	23.4	23.8	23.8	23.8	0.0	0.0
Travel Day	2.8	2.8	2.8	2.8	0.0	60.5

These results show significant differences between both the variables and the algorithms analysed by each importance measure. ‘Dwell time’, ‘start time’ and ‘end time’ had relatively high importance for each of the algorithms, suggesting that these variables should be included in the final analysis. The importance of each of the other variables differed significantly between algorithms. For example, ‘travel day’ had the lowest importance for all algorithms with the exception of RF, and ‘sex’, ‘age group’ and ‘trip distance’ also varied greatly between algorithms. These results helped inform which variables should be included in the final model.

6.2.5 Final model

This section presents the results when each algorithm was implemented on the optimised data i.e. data only for the West Midlands metropolitan region, containing only the variables that were

shown to have high importance for each algorithm. Considering the results shown in Section 6.2.4.3, to build the final model each of the optimised algorithms were run on the NTS data. Table 6.8 shows the accuracy and Cohen's kappa coefficient for each algorithm, defined in Section 6.2.2.

Table 6.8 Results of the final model selection (for algorithm acronyms see Section 6.2.2)

Algorithm	**Accuracy**	**Kappa**
LDA	0.67	0.33
kNN	0.66	0.33
SVM	0.64	0.26
NB	0.63	0.25
CART	0.63	0.26
RF	0.70	0.48

The RF algorithm produced the best results, with an accuracy of 0.70 and a kappa of 0.48. This suggests that 70% of the predictions were correct and there was moderate agreement between iterations. The accuracy and kappa of the RF model were higher than CART, despite the similarities between methods, likely due to the potential overfitting of the CART algorithm, as discussed in Section 6.2.2. The approach of the RF to produce more than one tree (and often several hundred) increased the accuracy where there may have been variations within each prediction variable.

Running the RF model on the validation dataset gave an overview of accuracy for each of the individual trip purposes. Table 6.9 shows the *prevalence*, *detection prevalence* and *accuracy* of the RF model. The prevalence is the percentage occurrence of each trip purpose in the NTS data. The detection prevalence is the percentage occurrence of each predicted trip purpose. The accuracy is the accuracy computed for each trip purpose individually.

Table 6.9 Detection prevalence and accuracy of the final model run on the validation dataset

	Commuting	*Leisure*	*Personal Business*	*Shopping*
Prevalence	0.03	0.22	0.18	0.57
Detection prevalence	0.05	0.17	0.05	0.73
Accuracy	0.89	0.69	0.57	0.70

These results show that the model was particularly effective at identifying commuting trips, with an accuracy of 89%, but much less effective for identifying personal business trips. Leisure and shopping trips showed promising results, with a balanced accuracy of around 70%. Comparing the results for prevalence and detection prevalence, it was clear that the model significantly over-estimated shopping trips and underestimated personal business trips. This was considered when interpreting the results of the smart card predictions, in particular when discussing any raw counts.

6.2.6 Smart card data preparation

In order to infer the purpose of trips and to validate these inferences against the known locations of key facilities in the study area, the origins and destinations of trips had to be known. The method for inferring these origins and destinations was outlined in Section 4.2.4.1. The data used in this analysis therefore consisted of transactions recorded between January 2014 and August 2016 for which the origin and destination could be inferred. In total, this resulted in over 36 million records being included in the following analyses.

6.2.6.1 Trip chaining and activity identification

A singular transaction on the bus network does not necessarily constitute a trip on its own but may instead make up part of a full trip chain. This meant that transactions recorded in the smart card data that were the result of a transfer between bus services, rather than an activity having taken place, should not have been considered as an individual trip but rather as a single stage that makes up part of a full trip chain. Joining these individual transactions into full trip chains was therefore vital to ensure that trips that result in an activity taking place were accurately

identified and the trip characteristic variables calculated for the smart card data matched those in the NTS data.

6.2.6.1.1 Home Trips

The first trips which needed to be identified prior to the analysis taking place were 'home trips'. For this analysis, these trips were defined as "records which are the result of a cardholder returning home between trips". Essentially, a cardholder may have undertaken several activities throughout a single day, returning home between these activities. If these home trips were not identified before being input into the algorithm, the trips that cardholders took to return to their home residence would have been incorrectly allocated a trip purpose of either 'shopping', 'personal business', 'leisure' or 'commuting'.

The first step in identifying home trips was to assign each cardholder their 'home stops'. These home stops were the stops which service a cardholder's home address and were therefore their likely destination stops when returning home. The only information contained in the smart card data relating to a cardholder's home location was a registered LSOA. These registered LSOAs could not solely be used for inferring the home stops of cardholders for a number of reasons. LSOAs in the study area vary greatly in size, from 0.06 km^2 in central Birmingham to 18.7 km^2 in the more rural Solihull. Additionally, the number of bus stops within each LSOA also ranges from 1 in some of the smallest LSOAs to 117 in central Wolverhampton. In addition, as LSOAs were originally built around neighbourhoods, their boundaries often correspond with the road network and therefore not only do a number of bus stops lie in close proximity to 2 or more LSOAs, but many paired bus stops, i.e. those that service the same route for inbound and outbound services, are located within different LSOAs. In light of this, a method had to be developed to identify the likely home stops of each cardholder.

A number of previous studies have utilised the home locations of smart card holders, for example Viggiano et al. (2016). These studies rely on the assumption that individuals begin their first journey of the day at home; an assumption which was also used in this analysis. The first stops used by cardholders each day was referred to as the 'first origin stop'. In addition, although previous studies have often only allocated a single home stop or station to a cardholder, this study allowed for the allocation of more than one stop. Each LSOA contained an average of 9 bus stops, so cardholders were likely to live within close proximity to more than one stop. Therefore, they could utilise multiple stops depending on route choice. Allowing for multiple home stops to be allocated accounted for this. However, to prevent every first origin stop being allocated as a cardholder's home stop, a threshold of 20% was implemented, i.e. for a stop to be identified as a home stop it must have been the first origin for at least 20% of the

cardholder's transactions. This therefore limited the number of home stops for each cardholder to 5. This 20% threshold was determined following an analysis of first origins, with nearly half of cardholders having 5 distinct first origin stops, as shown in Figure 6.1.

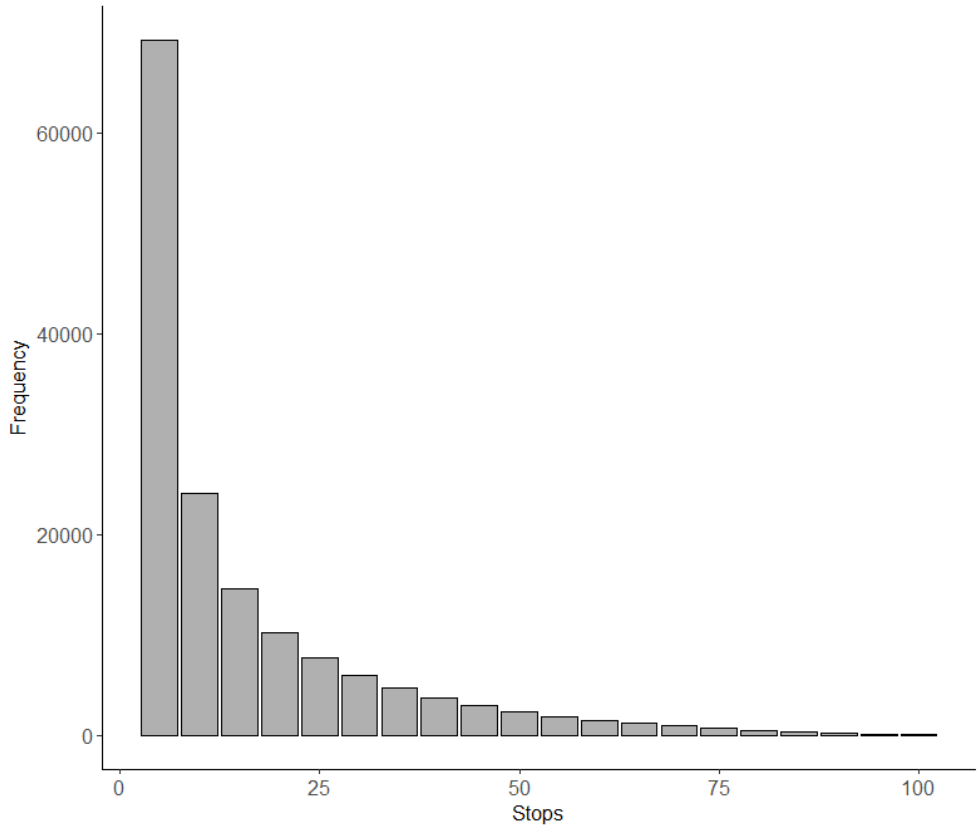


Figure 6.1 Histogram showing number of distinct first origin stops visited by each cardholder

Considering these assumptions and restrictions, the following method was implemented for inferring each cardholder's home stops and subsequently their home trips:

1. The first origin stop for each cardholder for each 'active day' (i.e. a day when a cardholder had transacted) was identified.
2. The proportion that each of these stops had been a cardholder's first origin was calculated.
3. These stops were filtered for a threshold of >20%.
4. All qualifying stops were assigned as a 'home stop' for each cardholder.

5. For all smart card transactions where the destination stop was the same as one of the cardholder's assigned home stops, the trip purpose for these trips was assigned as 'Home'.
6. Method repeated for each year to consider potential house moves.

6.2.6.1.2 Trip chaining

In addition to identifying home trips, a method also had to be developed for linking individual transactions into full trips. Many of the variables used for the trip purpose inference relied on this linkage and the subsequent aggregation of their trip characteristic variables. For example, in the NTS data 'travel time' was recorded as the total time between departure and arrival including any transfers between services. Therefore, full trips were needed to be able to calculate this variable in the smart card data. Additionally, linking transactions into full trips allowed for activities to be identified and therefore included in the final trip purpose inference analysis. It was therefore important that 'transfer stages' were identified so as not to be included in this analysis and incorrectly allocated an activity.

The method used for identifying transfer stages was outlined as part of the spatial profiling method presented in Section 5.2.3.2, with the same method employed in this section. This method identified stages with a dwell time of less than 30 minutes as being a transfer between services. These transfer stages were labelled as 'transfer' in the smart card data.

Once these transfer stages had been identified, these could then be linked to subsequent stages undertaken by a cardholder on the same day. Firstly, data were sorted by 'account ID', 'date' and 'timestamp'. Individual transactions, i.e. singular transactions that were not part of a trip chain, were then identified and labelled as 'individual'. The final transaction of each trip chain was also identified by analysing whether the prior transaction by that cardholder was labelled as a transfer stage, and if so, these were labelled as 'end' stages. For example, if a cardholder had undertaken three transactions on the bus network in a single day; two that were previously identified as transfers and another unlabelled subsequent transaction, this transaction would be the end of the trip chain and therefore the stage that resulted in an activity being undertaken.

Once the 'transfer', 'individual' and 'end' stages had been identified, these could be linked into full trip chains by assigning unique IDs to every full trip, whether this was a single 'individual' trip or a trip chain consisting of multiple 'transfers' and an 'end' stages. An algorithm was designed to go through the sorted data and assign a new unique ID each time a 'transfer' stage followed an 'end' stage, or each time an 'individual' trip was encountered, as illustrated in Table 6.10.

Table 6.10 Example output of the trip chaining algorithm

record_ID	account_ID	trip_type	chain_ID
52248580	78******35	transfer	1
52248581	78******35	transfer	1
52248582	78******35	end	1
52248583	78******12	individual	2
52248584	78******92	transfer	3
52248585	78******92	end	3
52248586	78******12	transfer	4
52248587	78******12	end	4

Each full trip, or trip chain, having been assigned a unique ID ('chain_ID') allowed for aggregate trip characteristics to then be calculated.

6.2.6.2 Variable calculation and recoding

Before the trip purpose analysis could be conducted on the smart card data, the same trip characteristic variables that were used to train the model were calculated. Each full trip having been previously assigned a unique ID allowed for these trip characteristics to be aggregated. Unlike the unchained individual transaction data, these aggregate characteristics were comparable to the NTS data and could therefore be used for the trip purpose analysis. Table 6.11 outlines the methods used for calculating each variable, with the methods designed to optimise the comparability between these variables and the NTS variables.

Table 6.11 Methods for calculating trip characteristic variables

Trip Characteristics	Method
Travel Day	Day of the week extracted from the transaction timestamp of each trip chain using the *DATEPART* function in SQL.
Trip Distance	Road distance (km) calculated between the origin and destination of each trip chain.
Start Time	Time (hour) extracted from the boarding timestamp variable for the first transaction of each trip chain.
End Time	Time (hour) extracted from the alighting timestamp variable for the last transaction of each trip chain.
Total Travel Time	Time difference between the start time of the first 'transfer' transaction in the trip chain and end time of the 'end' transaction calculated using the *DATEDIFF* function in SQL.
Dwell Time	Time between the alighting time of first trip and the boarding time of the next trip calculated using the *DATEDIFF* function in SQL.

In addition to calculating each aggregate trip variable, these then had to be recoded to match the NTS format outlined in Table 6.2; aggregating variables into the same bins. The result of this was the smart card data reformatted to exactly match the NTS example data shown in Table 6.4, minus the trip purpose variable to be inferred.

6.2.7 Spatial patterns and validation method

Once the trip purpose inferences had been made, the spatial distribution of each purpose was analysed. Not only could this provide insight into how the locations of different activities varied, but the results could also be further validated against land use in the study area, the methods for which are outlined in the following sections.

6.2.7.1 Spatial distribution of trips

Understanding the spatial distribution of different trip purposes can provide insight into how the older population utilise the bus network for undertaking activities, and how this relates to the function of different areas. As the trip purpose inference method was unlikely to have correctly

predicted the purpose of all bus trips, it was determined that identifying hotspots for each trip purpose would produce the most accurate representations of the spatial distribution of each purpose. Identifying only areas with a statistically significant number of trips for each purpose would, therefore, reduce the impact of any inaccuracies in the trip purpose predictions. This hotspot analysis followed a similar method as outlined in Section 5.2.3.2, however was conducted at the bus stop rather than LSOA level. The aim of this was to identify clusters of stops that attracted a significant number of trips for each inferred purpose.

To calculate these hotspots, firstly the number of trips for each inferred trip purpose was calculated for each bus stop in the study area. The same hotspot analysis method as outlined in Section 5.2.3.2 was then run on these data, resulting in z-scores for each bus stop. These z-scores indicated whether there was a statistically significant clustering of trips for each purpose, i.e. whether an area attracted a significant number of visitors for shopping, commuting, personal business or leisure purposes. This method therefore identified the areas to which the older population travelled using bus services, whilst also reducing the impact that any incorrect inferences had on the resulting spatial distributions.

6.2.7.2 Land use validation method

To further validate the results of the trip purpose inferences, the spatial distribution of trips in the area were compared to detailed land use data, as a high correlation between available land uses and purposes of trips undertaken to these areas would be expected (Boarnet and Crane 2001). For example, trips made for shopping purposes would be expected to primarily end in retail areas, high streets and shopping centres, whereas those undertaken for commuting purposes may be expected to end in central urban areas. Understanding the relationship between urban land use and trip purpose can also provide insights into how the older population interact with the study area, i.e. which areas attract high numbers of concessionary bus users and which do not.

The European Environment Agency provides land use data for the West Midlands, shown in Figure 6.2. These data contain a high level of detail, with 20 different land uses identified; including industrial and commercial areas, densely populated urban areas, sports and leisure facilities, and green urban areas.

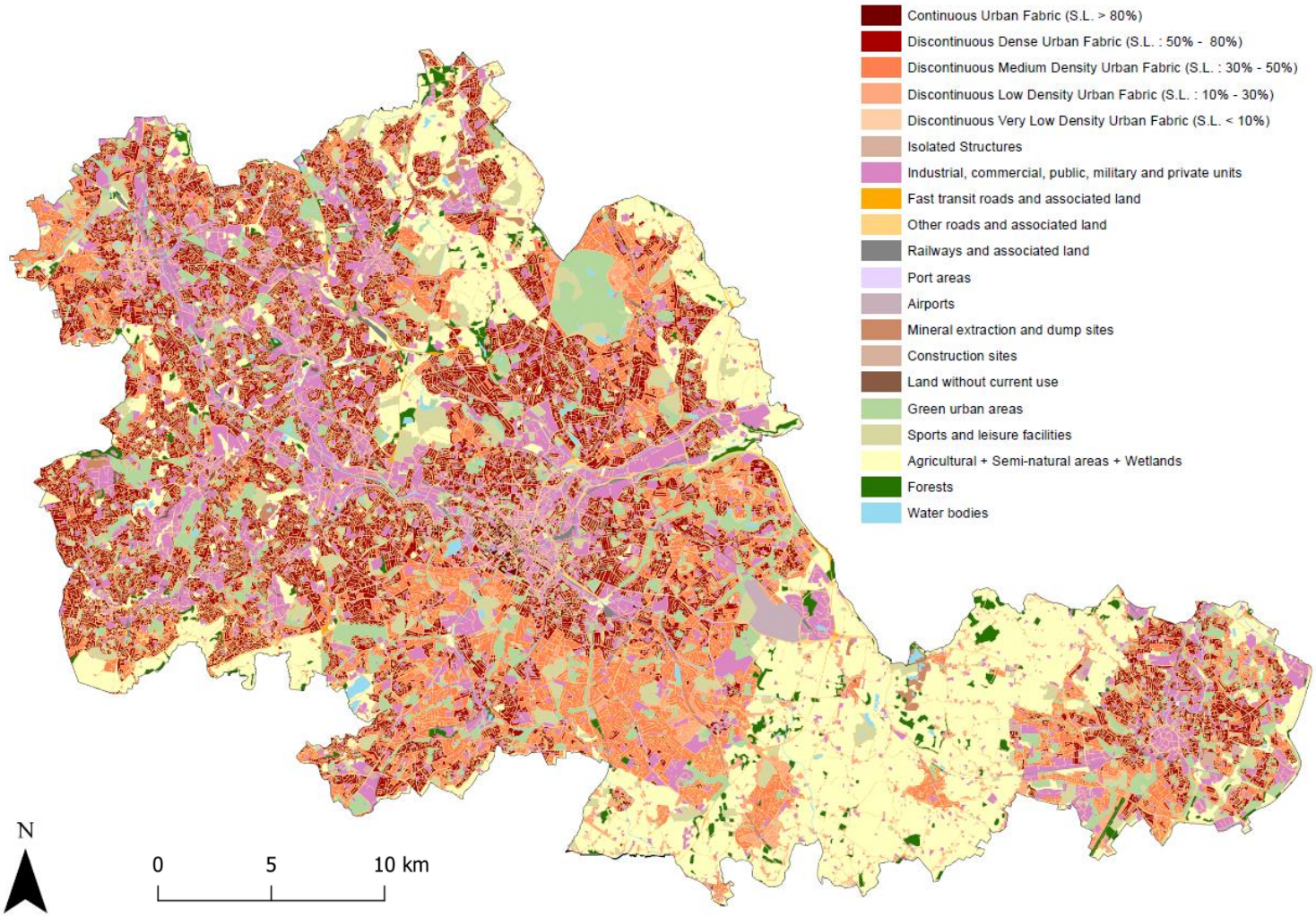


Figure 6.2 Land use map of the study area (European Environment Agency 2017)

The high level of detail shown in Figure 6.2 allowed for comparisons between the inferred trip purposes at each bus stop and the known land use in the areas surrounding these stops. This validation method relied on the construction of 'service areas' for each bus stop. A service area was defined as a region that encompasses all accessible streets within a specified impedance, for example road distance. A number of studies have looked at the service areas of transport stops and stations by analysing the distance that individuals are willing to walk to access transport services (Hess 2012). Although many of these studies highlight the likely variances within and between populations, 400m service areas are the general 'rule of thumb' used in transport service planning (Kimpel et al. 2007). To allow for comparison to previous studies, 400m service areas were implemented in this analysis. In this instance, these 400m service areas were therefore the estimated distance that cardholders would be willing to walk to access a facility or activity once alighting a bus service.

The service areas used in this analysis were constructed as 400m walking distances along the road network, using the *Network Analyst* tool in ArcGIS. This tool utilises a network dataset, in this case the road network, and the point locations of the bus stops in the study area. Impedances are usually measured as either distance or time. As this analysis focused on walking trips, distance was used as the impedance. This was to minimise the impact of different walking speeds which are likely to vary greatly within the eligible population. The *Network Analyst* tool then computed the service areas as 400m polygons built along the road network. Examples of the bus stop service areas generated are shown in Figure 6.3.

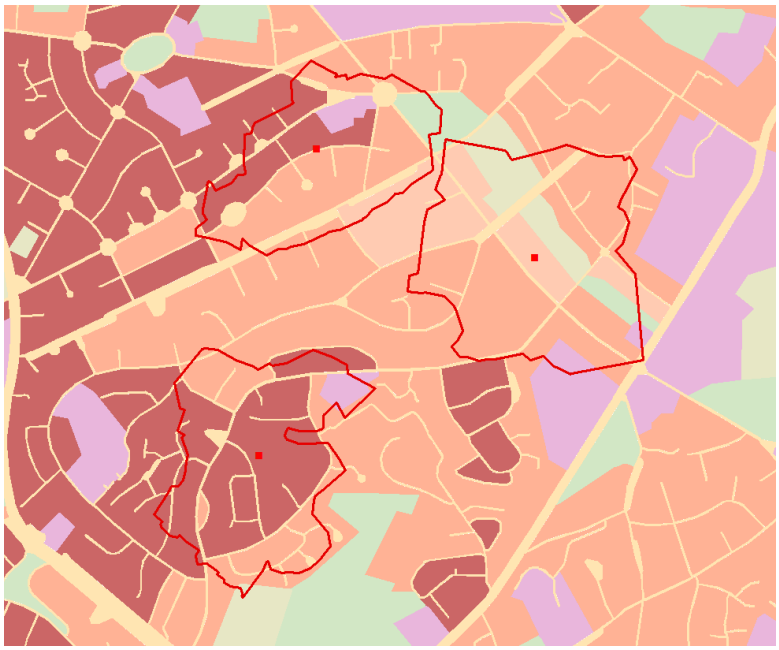


Figure 6.3 Service areas for three bus stops in the study area overlain on the land use map

These service areas allowed for the available land uses within 400m of each bus stop to be identified. The percentage of each bus stop service area that intersected with each land use was then calculated. This quantification meant that the inferred trip purposes could then be compared to the available land uses; acting as a validation for the trip purpose analysis, as well as providing insight into the areas which attract high numbers of visitors for each purpose. The method for this quantification firstly involved intersecting the bus stop service areas with the land use layer in ArcMap. The area of overlap between each service area and each land use category was then calculated and the results converted into percentages. These land use percentages were then compared between the statistically significant hotspots for each trip purpose. The aim of this was to validate the trip purposes inferences and to understand whether there is a relationship between land use and the types of activities undertaken in an area.

6.3 Results

This section presents the results of the analyses; firstly focusing on the trip chaining method and the inferred trip purposes, and then exploring the spatial distribution and validation against land use. The results of the trip purpose inference are also compared to the results of the classification conducted in Chapter 5 in order to understand if different clusters made use of the bus network for different purposes.

6.3.1 Activity identification and trip purpose inferences

The trip chaining and activity identification method outlined in Section 6.2.6.1 resulted in close to 30 million full trips being included in the trip purpose inference analysis, ranging from January 2014 to August 2016. Of these, over 1 million were identified as ‘home trips’ and therefore excluded from the trip purpose inference analysis. Although these 1 million home trips may be lower than expected as the majority of trips would have been expected to result in a cardholder undertaking a bus journey to return home, this low number of home trips was likely due to the uncertainty surrounding both the inferred destinations of each trip and the allocation of home stops to each cardholder.

Having excluded home trips, the trip purposes of around 28.5 million trips were inferred. Of these trips, 20% were identified as being part of a trip chain, i.e. either a ‘transfer’ or ‘end’ stage, with the remaining 80% identified as ‘individual’ trips. The results of the trip purpose inference, showing the proportion of trips allocated to each trip purpose, are shown in Figure 6.4 alongside a comparison to the NTS data.

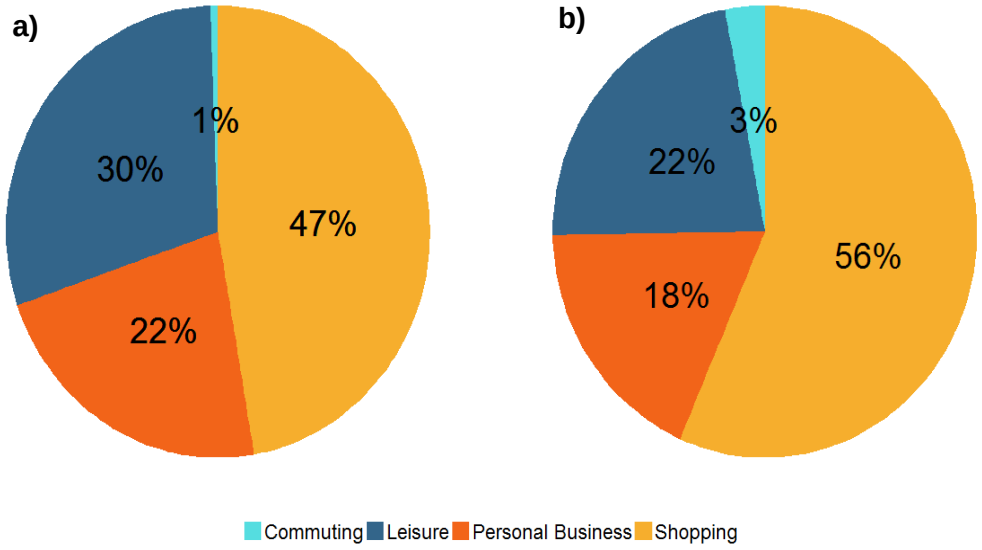


Figure 6.4 Trip purpose prediction proportions for a) smart card data and b) NTS data

It is clear from Figure 6.4 that the distribution of trip purposes was relatively similar between the smart card data and the NTS data, suggesting the random forest method produced realistic inferences. From the smart card data, fewer trips were inferred as being made for shopping and commuting purposes, with a higher proportion predicted to have been made for leisure and personal business. The detection prevalence shown in Table 6.9 suggests that the algorithm may have significantly over predicted shopping trips and under predicted personal business trips. This was considered when interpreting these results, with potentially even fewer trips being made for shopping purposes and a greater number of personal business trips. Very few trips were inferred as being for commuting purposes, which may be expected due to the population being of pensionable age and therefore the vast majority would have likely retired. The raw counts for each trip purpose are shown in Table 6.12.

Table 6.12 Inferred trip purpose counts

Commuting	Leisure	Personal Business	Shopping
164,107	8,468,287	6,388,852	13,464,848

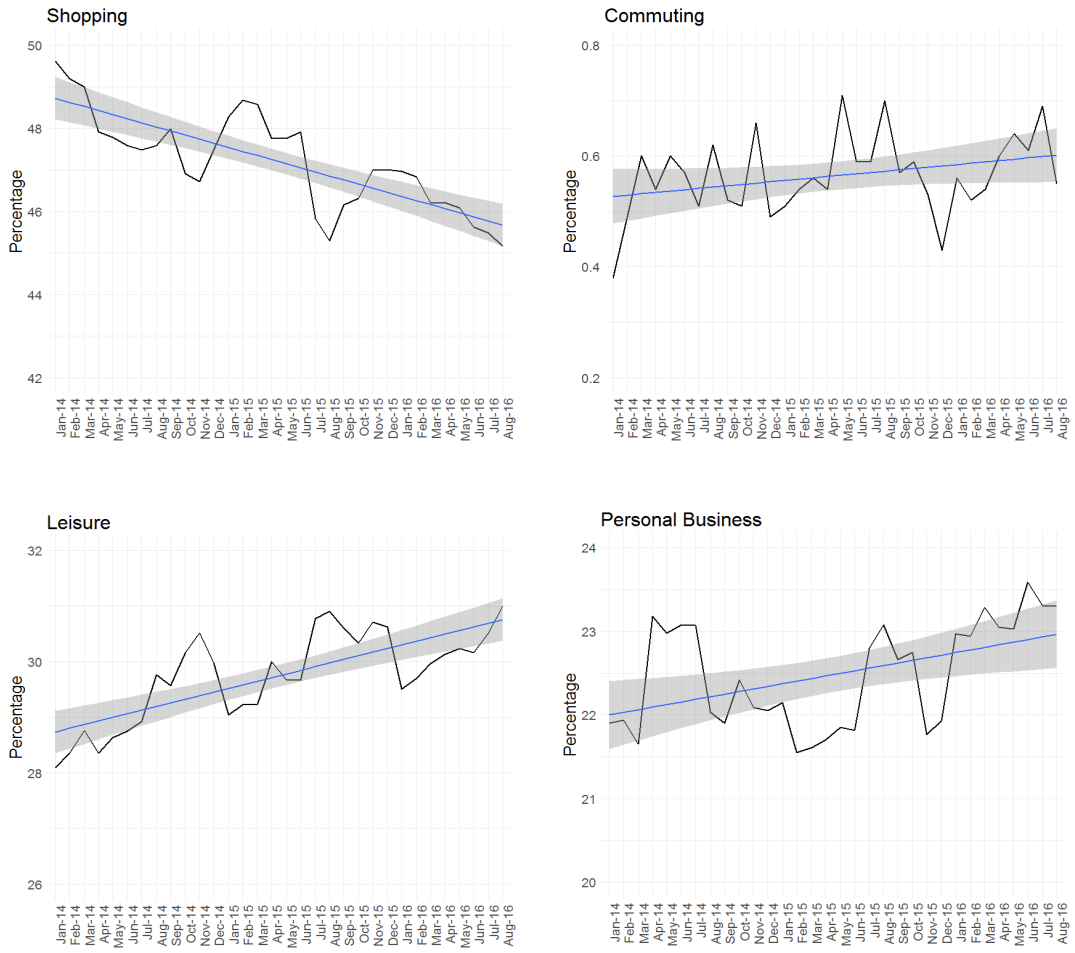


Figure 6.5 Monthly percentage of trips and trend line for each inferred purpose

Figures 6.5 shows the monthly proportions of each trip purpose. It should be noted that for visualisation and interpretation purposes, the scales for each trip purpose differ. As the percentage of trips varied from month to month, a trend line was overlaid on each graph. With the exception of commuting trips, which showed only a slight increase over the study period, trips made for other purposes showed changing distributions. Shopping trips showed a decline of over 4% between January 2014 and August 2016, with a decline in the yearly average of around 2%. Conversely, leisure and personal business trips underwent an average yearly increase of around 1%. Due to the likely inaccuracies in the inference of trip purposes, conclusions could not be drawn on the specific number of trips made for each purpose and how these changed, however general patterns in the distributions do suggest that the share of bus trips made for different purposes shifted over the study period.

The characteristics of each trip purpose were also compared. This allowed for further validation of the results, as certain trips may have been expected to have certain characteristics. For example, commuting trips would be expected to primarily take place on weekdays and during peak times. As well as validating the results, these characteristics also tell us about demand for bus services and differences in the use of the bus network for different trip purposes, for example when shopping trips were most likely to take place in comparison to commuting trips. Figures 6.6 to 6.12 show the distributions of the characteristics included in the inference analysis for each trip purpose.

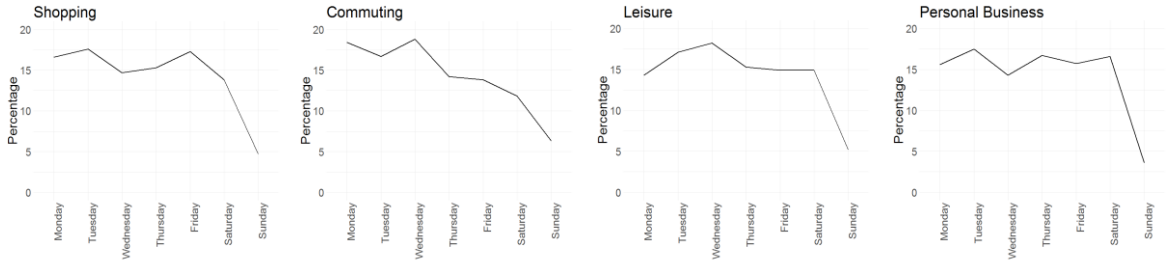


Figure 6.6 Travel day distributions

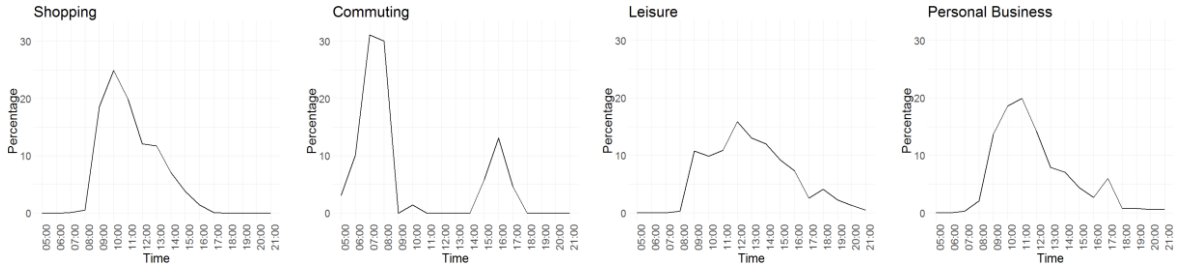


Figure 6.7 Start time distributions

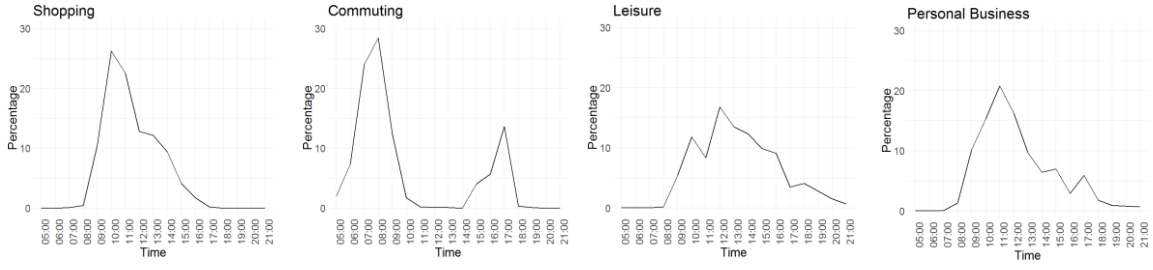


Figure 6.8 End time distributions

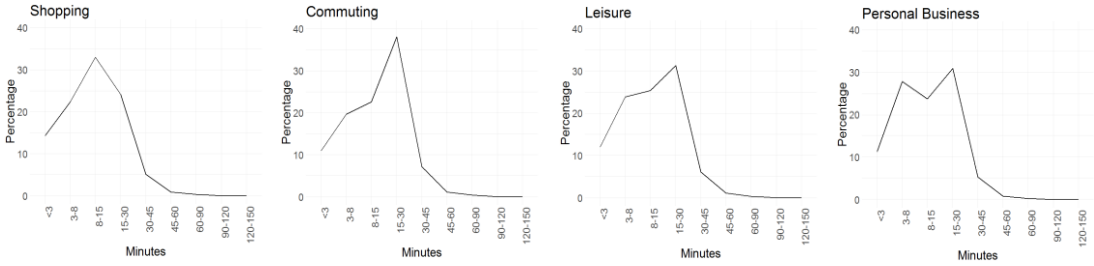


Figure 6.9 Travel time distributions

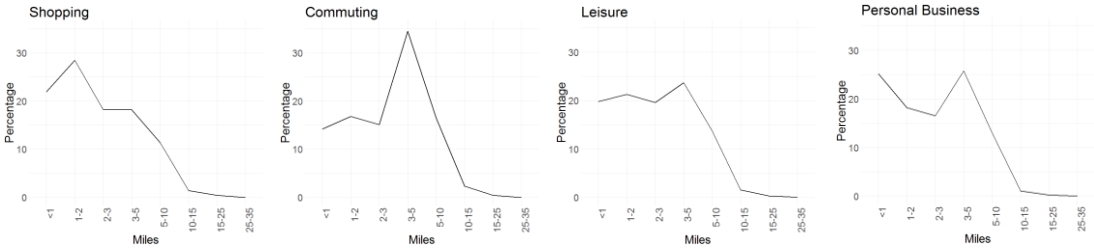


Figure 6.10 Trip distance distributions

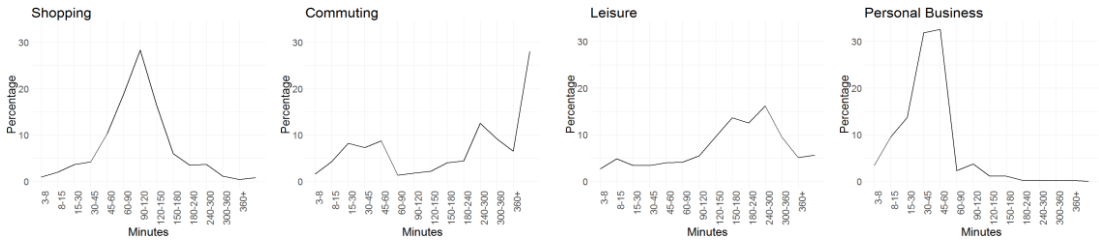


Figure 6.11 Dwell time distributions

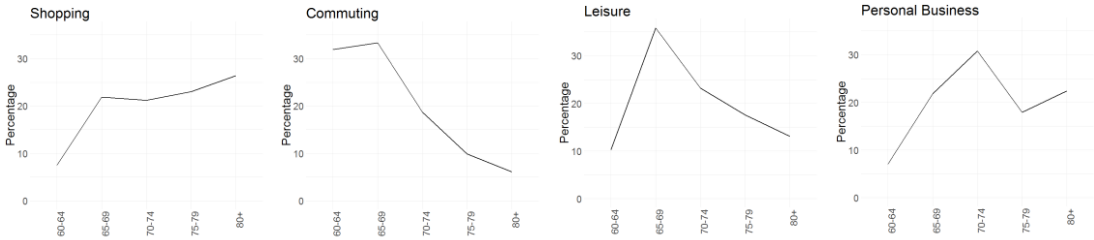


Figure 6.12 Age group distributions

These distributions showed differences between the characteristics of trips undertaken for different trip purposes and can provide insight into the demand for both bus services and facilities in the study area. In terms of when trips are undertaken, all trip purposes showed fewer trips on Sundays with relatively even distributions throughout the rest of the week, with the exception of commuting trips which also showed fewer trips across the weekend. These results would be expected, with public transport operators often running reduced services on Sundays and work trips more likely to take place during weekdays.

The distributions in 'start time' and 'end time' for each trip purpose also showed clear differences. Commuting trips were more likely to start and end around the morning and evening peak hours; which coincides with typical 9am to 5pm working hours. Personal business trips also showed a peak at around 5pm, which may correspond with those that are still in work undertaking these personal business trips after working hours. Shopping trips were most likely to take place in the morning, peaking at 10am, whereas leisure trips were found to be more distributed throughout the day with many taking place later than the other trip purposes. This is likely due to the nature of leisure trips, such as visits to friends and entertainment facilities, which are unlikely to be constrained by the same facility opening hours as shopping and personal business trips. Analysis of these trip characteristics can tell us about demand on the bus network. The 'start time' and 'end time' distributions suggest that demand for bus services by the older population was high during the morning period, with only commuting trips, which make up just 1% of inferred purposes, showing a significant peak in the afternoon.

The 'travel time' distributions showed similar patterns between trip purposes, with very few trips taking longer than 45 minutes. This is likely a reflection of the size of the study area and bus network. In general, commuting, leisure and personal business trips showed longer travel times than shopping trips. Again this may be expected, with shopping trips likely undertaken locally, whereas commuting, personal business and leisure trips may be undertaken to a specific location, such as a work location, medical facility or entertainment facility respectively.

These distributions were also reflected in 'travel distance', with shopping trips generally taken over a shorter distance than the other trip purposes. Like travel time, this suggested that shopping trips were undertaken more locally than those for other purposes. Commuting trips peaked at 3 – 5 miles, which is shorter than the average commuting distance found by the DfT (2017) of 5 – 10 miles. Again, this is likely a reflection of the size of the study area and the fact that it is largely urban; suggesting trips to a work location would not be as far as when travelling from a rural area.

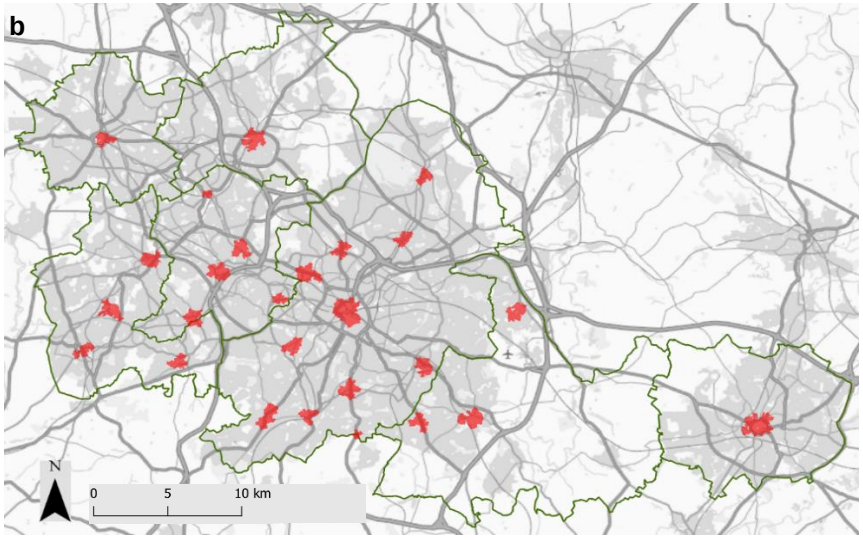
The ‘dwell time’ distributions showed the largest differences between each trip purpose. As expected, the majority of commuting trips showed a dwell time of over 6 hours, which corresponds with typical working hours. Shopping trips mostly lasted between 1 and 2.5 hours, which again is what would realistically be expected for these types of trips. Personal business trips displayed the most significant peak, between 30 minutes and 1 hour. As these trips involve activities such as visiting a bank or picking up a prescription, this shorter dwell time would be expected. Leisure trips were more evenly distributed, however, and contained a large number of trips with a dwell time of over 2 hours. These distributions are realistic considering these trips include visiting family and friends or undertaking sports activities.

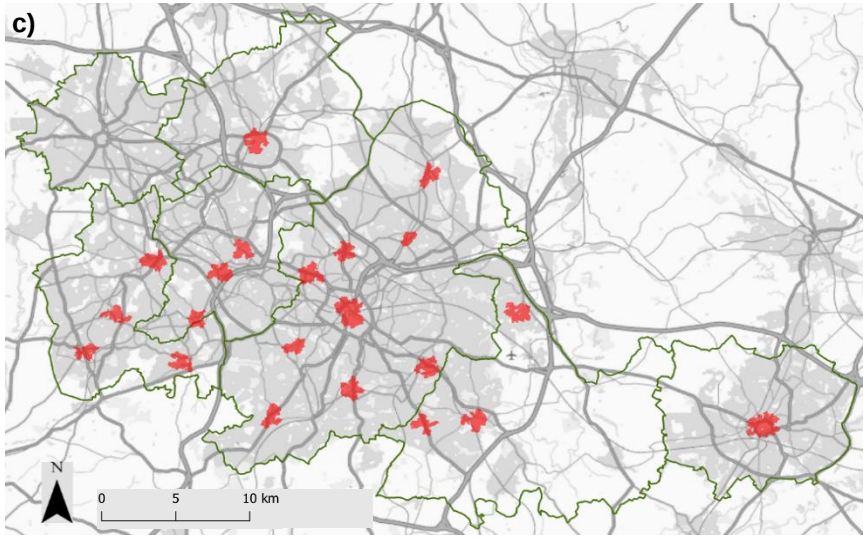
Comparisons between the ‘age group’ distributions for each trip purpose evidenced that younger age groups were much more likely to undertake commuting trips than those that were older. In addition, leisure trips also appeared to decline with age. Shopping and personal business trips increased with age, which could be due to some of the population losing the ability to drive (Andrews et al. 2012) and therefore relying on public transport to undertake those trips which may previously have been undertaken by private modes of transport.

6.3.2 Spatial distribution

The hotspot analysis resulted in the identification of bus stops that attracted a statistically significant number of trips for each purpose. These were constructed at bus stop service area level using counts of the destinations of trips made for each purpose. There were no statistically significant cold spots found for each trip purposes, so Figures 6.13 to 6.16 show only the hotspots found for each year of the analysis at bus stop service area level. The lack of significant cold spots means that there were no areas that attracted significantly fewer trips than would have been expected for each purpose. This again validates the results of the trip purpose inferences.

Figure 6.13 Shopping activity areas showing statistically significant hotspots at bus stop catchment level in a) 2014, b) 2015 and c) 2016





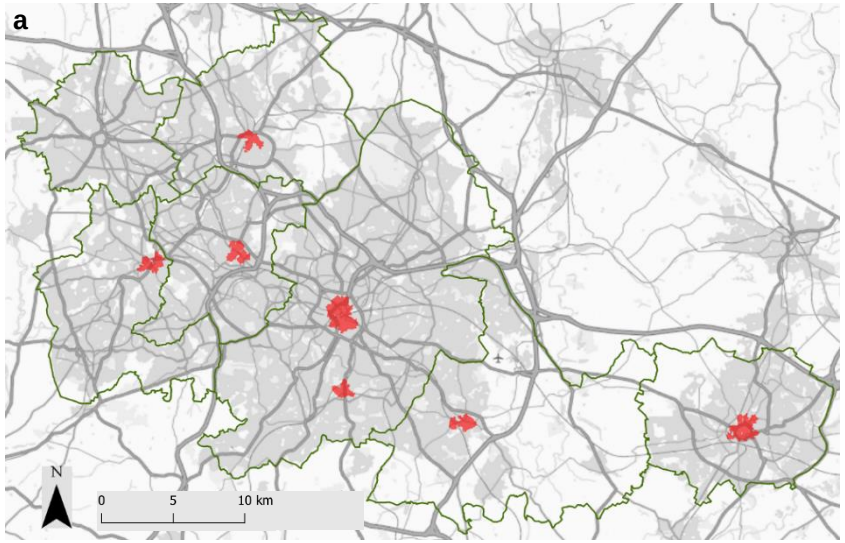
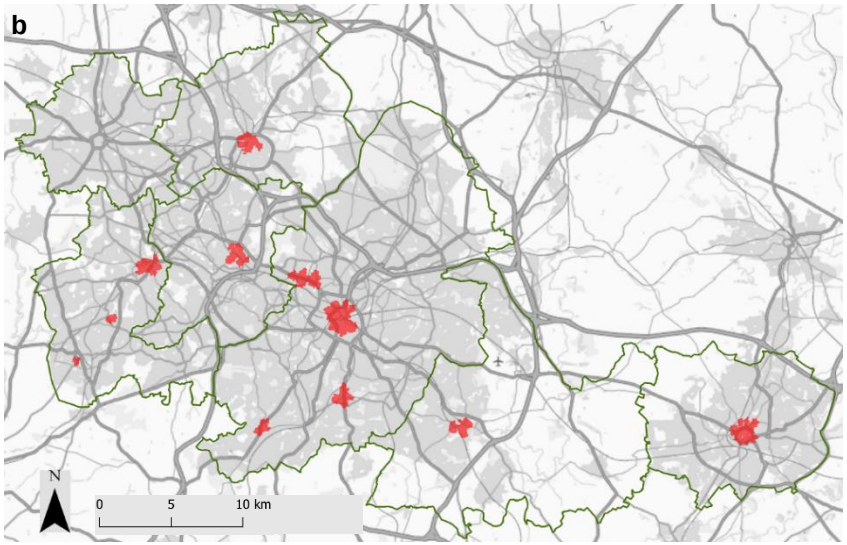
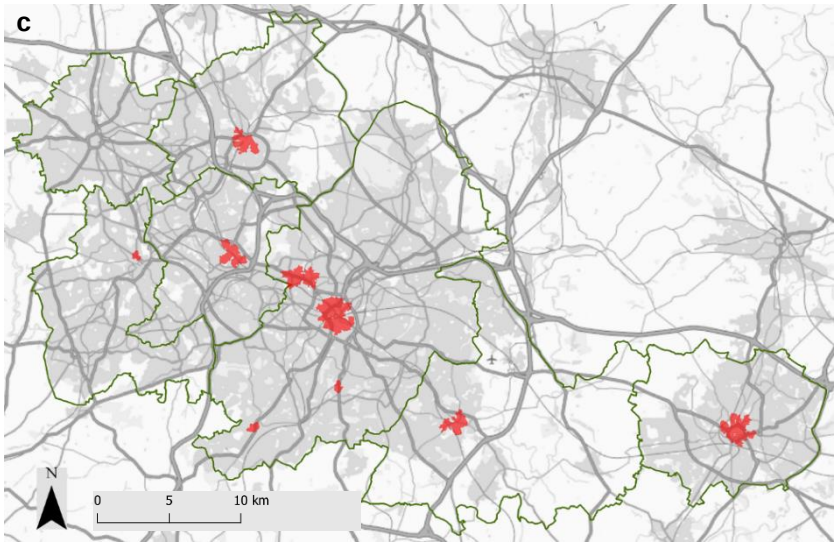


Figure 6.14 Commuting activity areas showing statistically significant hotspots at bus stop catchment level in a) 2014, b) 2015 and c) 2016





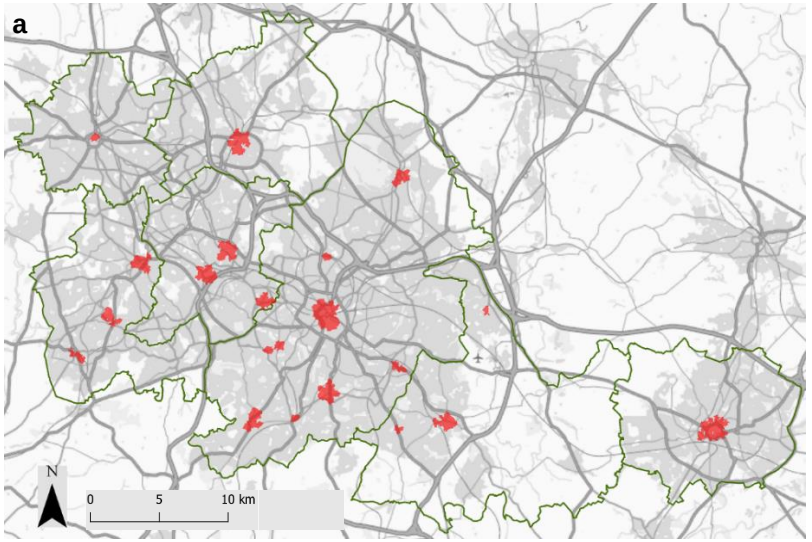
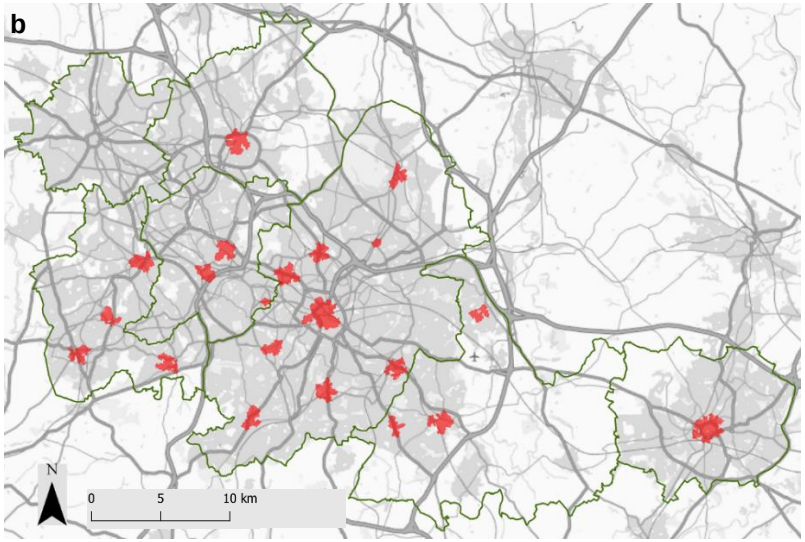
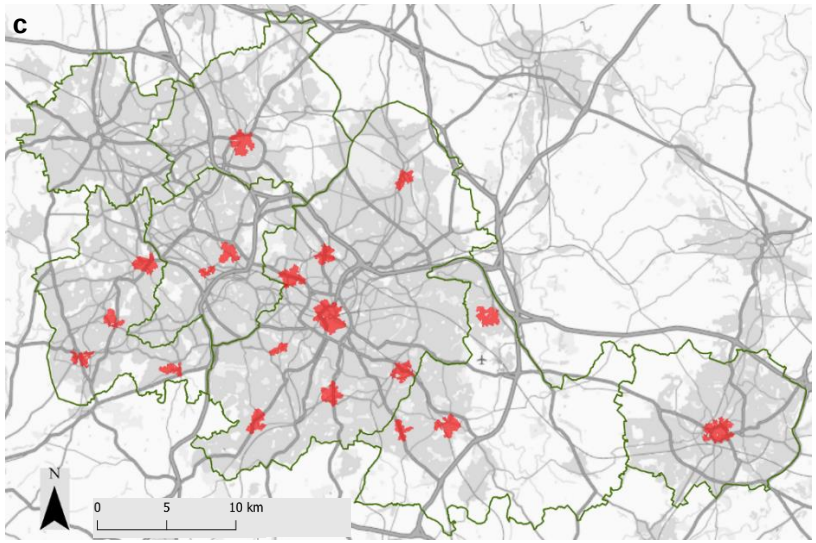


Figure 6.15 Leisure activity areas showing statistically significant hotspots at bus stop catchment level in a) 2014, b) 2015 and c) 2016





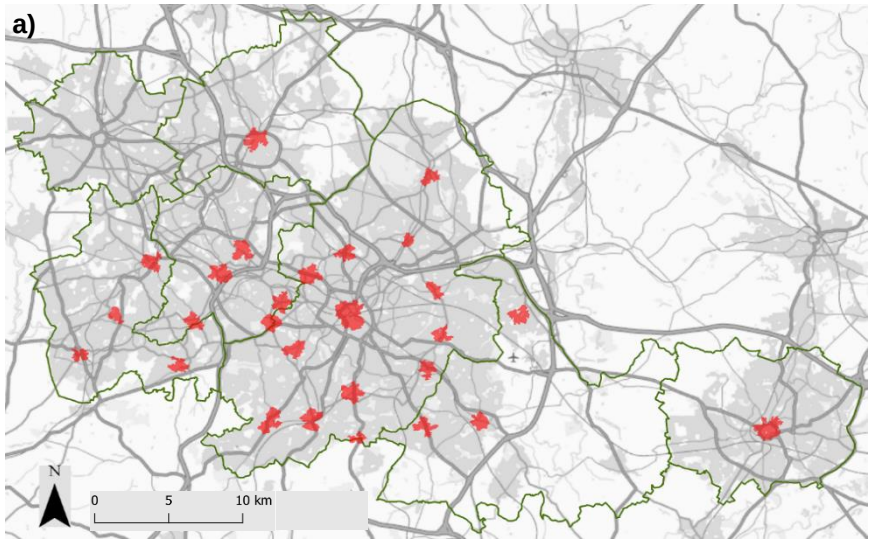
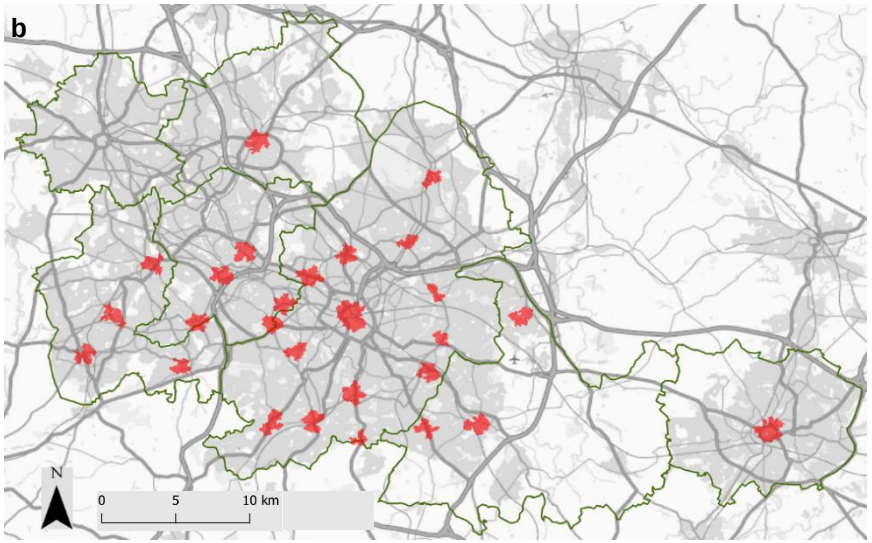
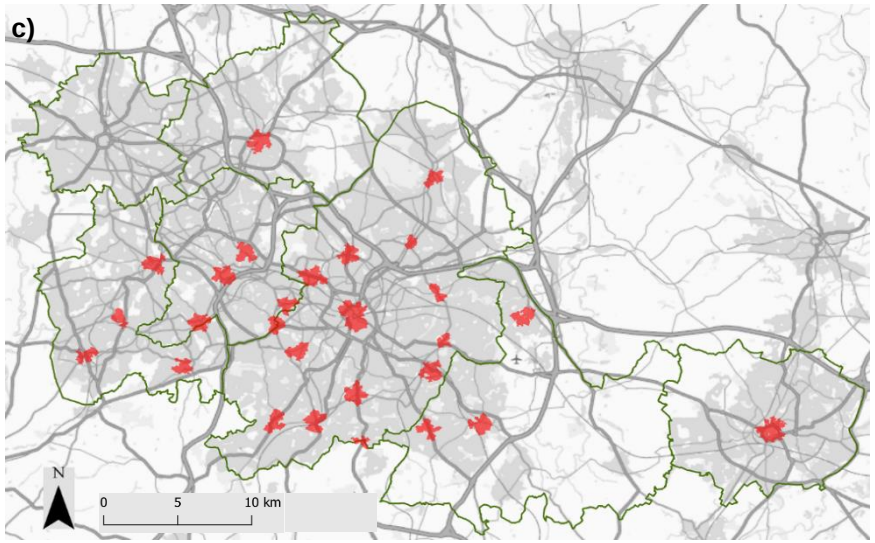


Figure 6.16 Personal business activity areas showing statistically significant hotspots at bus stop catchment level in a) 2014, b) 2015 and c) 2016





The results of the hotspot analysis indicated that there were not significant variations between each of the trip purposes. This may be expected, as the majority of trips for each purpose would likely be undertaken in retail centres and central urban areas. These similarities also suggest that the service network in the area is suitable for a range of purpose. Similar spatial distributions for each trip purpose show that the areas that were visited by the older population, and therefore accessible via the bus network, provided services and facilities to meet a range of purposes.

The proportional distributions of trip purpose shown in Figure 6.4 do, however, suggest that the share of activities undertaken in these areas differed. For example, although both commuting trips and shopping trips showed hotspots in central Birmingham and Coventry, there would still be expected to be a significantly higher number of shopping trips undertaken to these areas due to the higher overall number of trips made for shopping purposes (see Table 6.12 and Figure 6.4).

Despite similarities in these hotspot locations, slight variations were identifiable both between trip purposes and between years. The shopping hotspots shown in Figure 6.13 suggested that bus trips made for shopping purposes had moved away from urban centres, towards smaller retail areas and out-of-town retail parks, with fewer and smaller hotspots located in central urban areas such as central Birmingham and Coventry when comparing 2014 to 2016. Although central Wolverhampton contained shopping hotspots in both 2014 and 2015, these were not present in 2016. These changes in shopping hotspots suggest that how the bus network is being used by the older population may have changed and may be indicative of other wider social and behavioural changes, such as a shift in retail activity to out-of-town retail parks and a growth in e-commerce. This reduction in shopping trips to central Wolverhampton may also be indicative of a modal shift. As discussed in Section 3.1.1.1, the metro line in the area runs between Birmingham and Wolverhampton and was extended in 2015 to terminate in central Birmingham rather than further out of the city centre. The reduction in shopping trips to Wolverhampton may therefore be indicative of a shift to the use of metro services to undertake these trips rather than an overall reduction in the number of shopping trips made via public transportation.

Commuting trips made up the minority of inferred trip purposes, with only around 1% of all trips, and appeared to cluster around the main urban centres in the study area. The clustering of these trips around urban centres was unsurprising as the majority of employment opportunities are located in these central areas. The spatial distributions of these trips remained relatively constant over the study period, with the exception of additional hotspots in south and west Birmingham, and south Dudley in 2015 and 2016. These new commuting hotspots could be indicative of land use change, with new developments resulting in more job opportunities.

The spatial distribution of leisure trips largely corresponded with the location of shopping trips, likely due to these kinds of trips taking place in both central urban areas and retail parks. Comparing the distribution of leisure hotspots with retail areas showed a large number of these hotspots were located in close proximity to shopping centres. The spatial distribution of these leisure trips remained relatively constant over the study period; primarily concentrated in retail areas. This suggests that while shopping trips appeared to have experienced a shift to out-of-town shopping centres, the number of trips taking place to central urban areas for leisure purposes remained significant. These trips made up 30% of overall inferred trip purposes, which corresponds to almost 8.5 million trips.

Personal business trips differed from shopping and leisure trips in that their spatial distribution appeared to be concentrated around smaller retail areas and high streets. According to the NTS, examples of personal business trips include visits to banks, health centres and hairdressers (DfT 2016). These trips being concentrated around smaller high streets would therefore be expected, with many of these facilities available on most typical high streets in the UK. Figure 6.11 showed that on average these trips had the lowest dwell time, peaking around 30 to 60 minutes. It would therefore be expected that these short duration trips would likely take place more locally than shopping and leisure trips and therefore have higher concentrations along local high streets and within retail areas.

6.3.2.1 Validation against available land uses

Comparing inferred trip purposes against known land uses in the study area further highlighted the differences in activity areas between each trip purpose, in addition to validating the results of the inference method. These land uses were calculated as the percentage of each hotspot area, constructed as 400m bus stop service areas, covered by each land use (shown in Figure 6.2). Table 6.13 shows the primary land use makeup of the hotspot areas for each trip purpose. Primary land uses were defined as those that made up the majority of land uses for the service areas.

Table 6.13 Land use make-up for the inferred trip purpose hotspot areas

Land Use	Shopping	Commuting	Leisure	Personal Business
Industrial, commercial, public, military and private units	42.53	57.17	45.44	40.06
Continuous Urban Fabric	3.65	7.75	4.53	3.61
Discontinuous Dense Urban Fabric	22.36	18.36	21.60	26.15
Discontinuous Medium Density Urban Fabric	16.34	7.39	14.67	16.90
Discontinuous Low Density Urban Fabric	1.56	2.94	1.25	0.82
Agricultural, Semi-natural areas and Wetlands	5.69	0.98	4.98	3.45
Green urban areas	5.88	3.85	6.04	7.10
Sports and leisure facilities	1.99	1.57	1.49	1.90

These results showed differences in the land use in areas that were frequented for different purposes, in particular between commuting trips and the other trip purposes. 'Industrial, commercial, public, military and private units' are classed as land units that are under industrial or commercial use or are used for public service facilities. Most urban centres and retail areas contain a large proportion of land classified as this land use. It was therefore unsurprising that for all trip purposes, this land use was the most common. Commuting hotspots were most likely to contain this land use; as can be seen in Figure 6.14, these trips were primarily taken to urban centres. Personal business hotspots contained the lowest proportion of this land use. Comparing this to Figure 6.16, these trips were more likely to take place in smaller retail centres and high streets and therefore were likely to contain higher proportions of 'urban fabric' land use than industrial and commercial units. Personal business trips also contained trips to hospitals and medical centres which were less likely to be located in urban centres, and therefore may explain the higher proportion of 'Discontinuous Dense Urban Fabric'.

Although unexpected, the slightly higher proportion of shopping and leisure hotspots that contain 'Agricultural, Semi-natural areas and Wetlands' land use areas can likely be explained by the location of out-of-town retail parks. These out-of-town shopping areas are usually

located away from urban centres and, as the 400m service areas used in this analysis likely contain a proportion of the surrounding land, an overlap with some rural land uses may be anticipated.

The high proportion of hotspots containing 'Industrial, commercial, public, military and private units' land uses, and the relatively low proportions containing land uses that would be found in more rural and suburban areas suggested that the trip purpose inferences and hotspot analyses produced realistic results in terms of the spatial distribution of trips throughout the study area. Additionally, the differences between trip purposes further validated the inferences, with work locations more likely to be located in built up, urban areas than those for shopping, personal business and leisure. This analysis was solely focused on validating the inferences rather than to further extrapolate any insights into the changing distributions of trips.

6.3.3 Comparison to cardholder classification

Chapter 5 presented a classification of concessionary cardholders based on their activity on the bus network. Understanding the reasons these cardholders undertake bus trips can provide insights into the different demands of cardholders. Table 6.14 gives a brief outline of each of the clusters and Figure 6.17 shows the proportion of trips undertaken for each inferred trip purpose by each cluster.

Table 6.14 Overview of cardholder clusters generated in Chapter 5

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Members 266,455 **Bus usage** Very low **Characteristics** ○ Low deprivation ○ High car/ van availability ○ High proportion from white ethnic background	**Members** 73,498 **Bus usage** Low **Characteristics** ○ Low deprivation ○ Relatively high car/ van availability ○ Majority female	**Members** 3,373 **Bus usage** Very high **Characteristics** ○ High deprivation ○ Low car/ van availability ○ High proportion from non-white ethnic backgrounds ○ Majority male	**Members** 15,546 **Bus usage** High **Characteristics** ○ Relatively high deprivation ○ Relatively low car/ van availability ○ Even male/ female split	**Members** 17,336 **Bus usage** Decreasing **Characteristics** ○ Relatively high deprivation ○ High proportion of cardholders aged 80+ ○ Majority female	**Members** 29,608 **Bus usage** Increasing **Characteristics** ○ Majority female ○ Representative of people entering into the ENCTS scheme

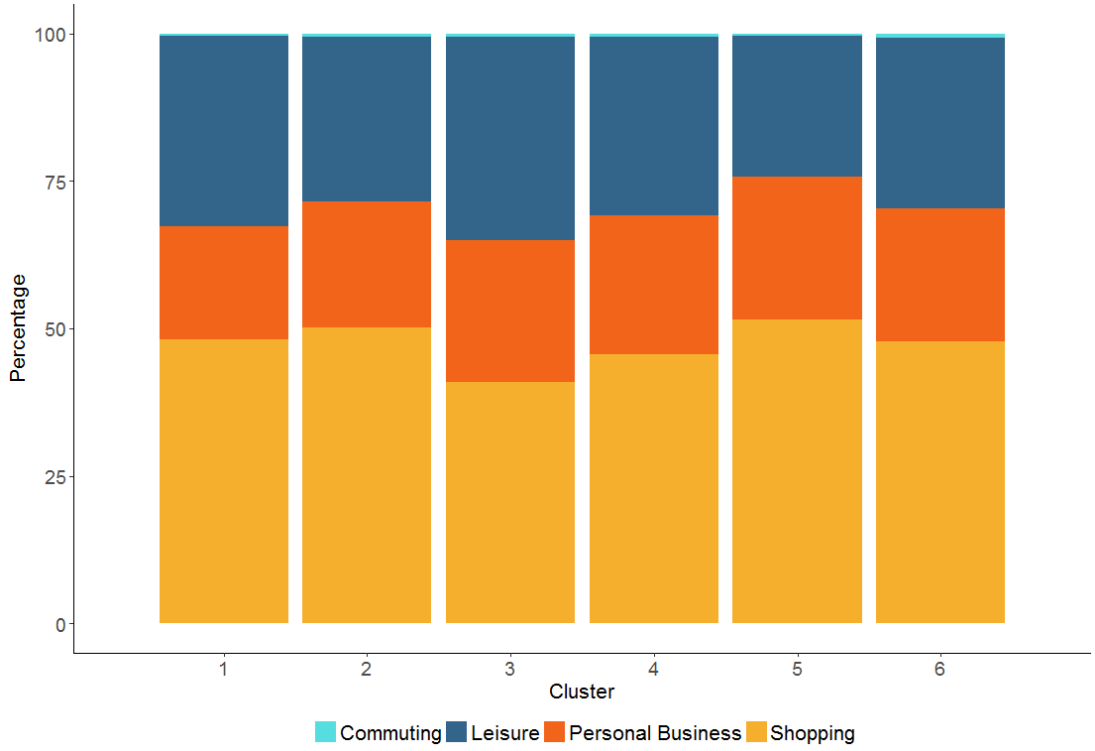


Figure 6.17 Trip purpose percentages for each mobility cluster

The general trip distributions were similar between clusters; however, some differences were evident. Cluster 3, those that were categorised as the most active bus users, made a smaller proportion of shopping trips compared to the other clusters and a larger proportion of trips for leisure purposes. No clusters made a large number of trips for commuting purposes. However, Cluster 6, containing those cardholders that had increased bus usage over the study period, contained almost double the proportion of trips for commuting purposes than those that had exhibited decreasing usage in Cluster 5. This is likely due to the younger age groups entering the system that are more likely to still be in employment and therefore undertaking commuting trips. The least active bus users (Clusters 1 and 2) made a higher proportion of trips for shopping than those that were more active (Clusters 3 and Clusters 4), and a lower proportion of personal business trips.

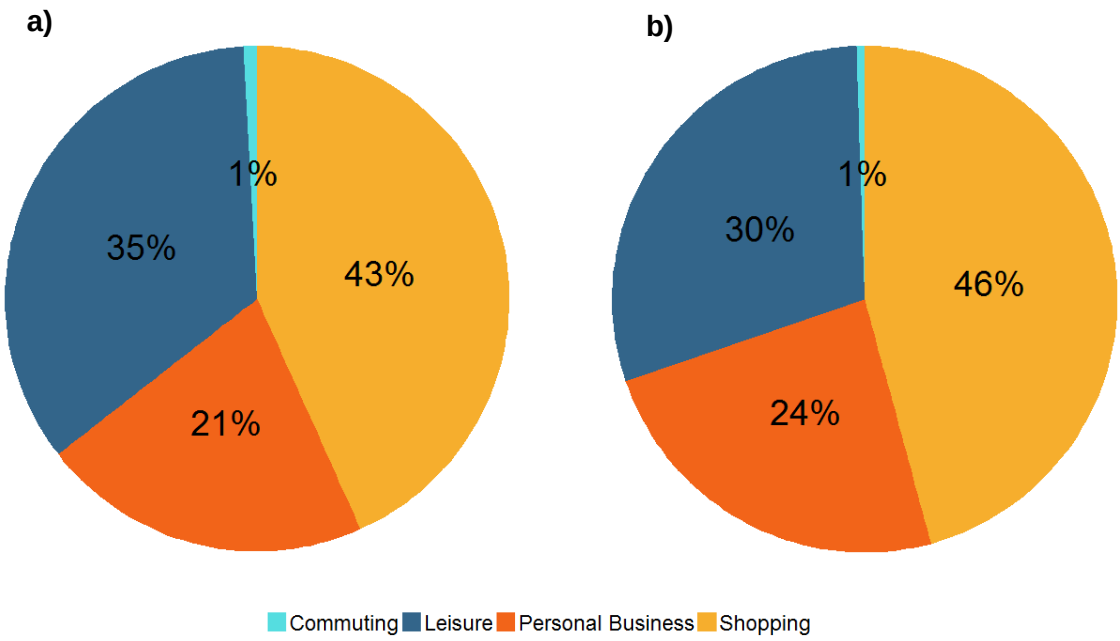


Figure 6.18 Average trip purpose proportions for cardholders that exhibited a) increasing and b) decreasing bus usage

Figure 6.18 shows the proportion of trips made for each purpose for those sub-cluster cardholders who were identified as exhibiting changing bus usage over the study period in Section 5.3.2.2. Comparisons of these distributions to those presented in Figure 6.5 can validate these results and highlight changes in the purpose of bus usage by concessionary travellers. Figure 6.18 shows that those who exhibited increasing bus usage tended to make more trips for leisure purposes and fewer trips for shopping purposes than those who had decreased usage. These same patterns were also shown in Figure 6.5, with the number of trips made for shopping purposes showing the largest decrease over the study period, and those for leisure purposes showing the largest increase. If those who increased their usage of bus services also undertook a higher proportion of trips for leisure purposes, then this could explain the changes in trip purpose distributions shown in Figure 6.5.

These insights show an apparent change in demand for bus services. As discussed in Chapter 3, an ageing population and changing modal share in the study area, for example an increase in the number of driving license holders of the older population (see Table 3.1) and an increase in metro and train boardings (see Figure 3.5), were likely to have changed how bus services were used by the older population. These insights suggest that shopping trips were less likely to be made using bus services, whereas leisure trips were increasing. These insights can aid transport

authorities and operators in planning future transport services, ensuring that demands are met in terms of where and when populations are travelling.

6.4 Discussion and Conclusions

6.4.1 Insights into trip purpose and activity areas

The general trends found throughout this chapter, in particular the changes presented in Figure 6.5, provide insight into travel behaviours of the older population. The most significant change in trip purpose throughout the study period was the decrease in shopping trips and increase in leisure trips found between 2014 and 2016. When compared to observed retail behaviours, such as a shift away from city centres and high streets to out-of-town retail centres (Scheiner 2006) and the increase in e-commerce in recent years (Falk and Hagsten 2015), a decrease in bus trips made for shopping purposes may be expected. Out-of-town shopping centres and retail parks often lack adequate public transport services (Gray et al. 2006), and therefore cardholders may seek alternative transport to undertake trips to these areas. An increase in online shopping, although likely not as significant for the older population than the working age population, may also explain this decline in the percentage of trips undertaken for shopping purposes.

The activity areas identified in Figure 6.13 to 6.16 also suggested changes in the spatial distributions of bus trips in the study area. The apparent shift in shopping trips to smaller high streets and out-of-town retail parks could reflect a modal shift towards tram and train travel to larger urban centres, with both undergoing an increase in patronage over the study period, as discussed in Section 3.1.1.1. In 2016, an extension to the West Midlands metro into central Birmingham was opened whereas previously, metro services ran between Wolverhampton and Snow Hill located to the North of the main city centre. This decline in shopping trips in both central Birmingham and central Wolverhampton, as seen in Figure 6.13, could therefore be indicative of a modal shift towards metro services.

Comparing the trip purposes of the cardholder clusters created in Chapter 5 showed that different types of cardholders made use of the bus network for different purposes. In general, those that were most active on the bus network undertook a higher proportion of trips for leisure purposes, whilst those that were less active tended to make most trips for shopping purposes. Although these differences were not large, these insights did suggest that the most active bus users have more diverse trip purpose distributions, which in turn suggests that bus services may be relied upon as a main mode of transport for a greater proportion of these cardholders. In addition, the trip purpose distributions of those users who exhibited increasing and decreasing

bus usage reflected those changes found in Figure 6.5, with shopping trips decreasing and leisure trips increasing over the study period.

6.4.2 Reflection on methods, inferences and future research potential

The apparent correlation of these insights with other observed changes in behaviour suggests promising results. Currently, there is no single accepted method for inferring trip purpose from novel big datasets. This chapter drew upon previous studies to construct both an initial trip purpose inference method and a validation method for trips recorded by AFC systems. Conventional travel surveys, which have long been one of the only sources of trip purpose data, are often time consuming and costly to conduct (Kieu et al. 2014). Although not an appropriate method to accurately predict the purpose of individuals trips, the success of the methods implemented in this chapter are promising for the substitution of these conventional surveys with novel consumer data.

Previous studies have had success in identifying work, education and home trips, or trips made for single activities (see Deng and Ji 2010; Liao et al. 2007), with accuracies often around 80-90%. However, identifying trips made for purposes other than these have produced significantly lower accuracies. These previous studies, such as Lu et al. (2012), have mostly reported accuracies of around 60-70% depending on which purposes are being predicted. The 70% accuracy for this chapter was therefore comparable, and in many cases exceeded, those found in these previous studies. This relatively high accuracy may be due to the similarities in the populations represented in both datasets, with the NTS data able to be filtered to contain only concessionary bus users residing within the West Midlands metropolitan area.

Despite the success of the methods applied in this study, there were a number of uncertainties and limitations. Firstly, there were a number of limitations to implementing NTS data to train the models. The NTS data and smart card data covered different time periods, with the smart card data spanning from 2014 to 2016 whereas the NTS data were collected between 2009 and 2016. This non-overlap was largely the result of the focus of this study on only the pensionable population in the West Midlands. This limited the amount of NTS training data that were available and so including data collected over a 7-year period rather than a 2-year period was deemed necessary to maximise the amount of data to train the models. Implementing the same 2014 to 2016 time frame but increasing the amount of training data through access to TfWM's own survey data may have improved the accuracy of inferences. As the methods presented in this chapter are transferable between data, this is likely to improve in the future as more data become available.

In addition to the limitations attached to utilising NTS data, the smart card data also had a number of limitations. Although the NTS data contained a large number of trip characteristic variables, only those that were also present in the smart card data could be implemented in this study. This limited the number of variables able to be inputted into the algorithm to make predictions. Additionally, only smart card records for which the destination was able to be inferred (outlined in Section 4.2.4.1) could be included in the study. This limited the amount of data available and meant that the purposes of a large proportion of recorded trips were not able to be inferred. Although this study did infer the trip purpose of over 30 million trips, when analysing the more granular results presented in Figure 6.5, these missing data may have influenced some of these trends.

As discussed in Section 6.1, one of the main limitations to the smart card data is the lack of information relating to trip purpose. Although this chapter presented a viable method for inferring this, the reliance on survey data to train the model is a major limitation. If smart card data are to provide a viable alternative to traditional survey data, a method which does not firstly rely on these surveys may be necessary. Currently, this method is promising for using alongside and enhancing survey data, in particular for the analysis of spatial distributions, but would not yet be able to replace these traditional methods.

In addition to the limitations and uncertainties of both the NTS and the smart card data, there were a number of limitations to the methods implemented throughout this chapter. The method for inferring the home stops of cardholders, although generating promising results, produced far fewer 'home' trips than would have been expected. This is likely to be due to inaccuracies in both the inference of destinations, presented in Section 4.2.4.1, and also the inference of cardholders' home stops. When a bus user applies for a concessionary travel pass a home address must be provided. If data were available at a more granular spatial scale, such as postcode or OA level, these could have been included in the analysis to produce more accurate results.

The trip chaining method relied on the assumption that, for an activity to have taken place, the dwell time must exceed 30 minutes; deduced by analysing the average wait times at bus stops. As wait times at different stops throughout the study area are likely to vary, undertaking this analysis at an individual bus stop level may have improved the accuracy. For the purposes of this analysis, with over 14,000 bus stops in the study area, this was considered too computationally expensive. Future analyses could compare the dwell times of individuals with the average wait times at each bus stop to better improve the identification of activities.

The random forest method implemented to infer the purpose of trips also had a number of limitations. The 70% accuracy of the random forest model still meant that there were likely a

large number of trips purposes that were inferred incorrectly. These incorrect inferences were unable to be detected, meaning that whilst this method was able to produce aggregate results, this approach would not be appropriate if individual trips were to be analysed. Although less likely when using CART models (see Section 6.2.2), overfitting can often occur in random forest models. This means that the model learns the training data too well which can negatively impact the performance of the model on new data, especially where there may be large variations within variables (Breiman 1984). This may have been a particular issue in this study as the training data and the smart card data were from two different sources. There were therefore likely to be greater differences between the variables recorded in each dataset and therefore an overfitted model may not have produced accurate predictions. As more data become available, through both TfWM's own survey data and increased availability of smart card data, these issues may be addressed.

Despite these limitations, this chapter presented a viable method for inferring the purposes of trips recorded in smart card data. Larger volumes of training data and more granular spatial data may have improved the accuracy of the method. However, this method is transferable and can therefore be implemented in the future as more data become available. These methods and results could be used to further understand the transport needs and demands of different population groups; increasing understanding of how the bus network is used and where and when different types of trips are undertaken. Planning future transport services around these needs and demands would increase both the efficiency and effectiveness of the public transport network.

Chapter 7 – Loyalty card data: Preliminary analysis and retail behaviours

7.1 Introduction

This thesis has primarily utilised smart card data collected on the bus network. However, other non-traditional sources of data can also be used to explore the mobility and activity patterns of the older population. The most significant limitation of the smart card data is their limited scope to just one mode of transport. Although this means that these data can be of use for specific analyses into mobility on the bus network, without the implementation of other data sources, conclusions reaching beyond this cannot be deduced. Due to the now ubiquitous nature of consumer data, it is hypothesised that non-traditional, non-transport related consumer datasets may be able to provide valuable insights into mobility and activity patterns. The following chapters utilised loyalty card data provided to the CDRC by a large high street retailer (HSR) to further explore mobility and activity patterns beyond the scope of the bus network. This is centred around the idea of 'triangulation', i.e. using more than one method or source to collect data on the same topic (Carter et al. 2014). It was hypothesised that this data triangulation could enrich traditional transport data and supplement insights.

'Loyalty card data' refer to the data produced when a transaction is conducted using a card linked to a retailer's loyalty card scheme. These schemes have been prominent since the early 1990s, when retailers began to recognise the benefits of retaining rather than obtaining customers (Kotler 2002). Since then, loyalty card schemes have increased in popularity among retailers, with almost all major retail chains in the UK operating some form of scheme. In their most basic form, these schemes reward customer loyalty by awarding points when a purchase is made which can then be redeemed as discounts on future purchases. As well as benefitting cardholders, these schemes provide retailers with detailed data on the purchasing habits of members. Data are firstly collected through an application process, where a cardholder provides basic demographic information, for example age, sex and a home address. Customers are then provided with a loyalty card that records their purchasing habits at the point of sale. These loyalty cards can therefore provide retailers with a detailed understanding of their customers, through both their transaction behaviour and their demographics.

Chapter 6 inferred the purpose of trips made by cardholders on the bus network and found that there was a change in both the number of retail trips and the spatial distribution of these trips over the study period. The HSR loyalty card data allow for the characteristics of retail trips to be explored in greater detail; understanding how these differ between population groups and what

this can tell us about the mobility of cardholders. These loyalty card data contain both spatial and temporal information, with the location and timestamp of each transaction recorded in the data. This information can allow for the activity patterns of cardholders to be deduced, both in terms of the times at which cardholders undertake retail activities and the areas that they visit. Additionally, the product information recorded with each transaction can allow for the consumption patterns of cardholders to be explored, including the types of products purchased and how these differ between groups.

This chapter gives a detailed overview of these loyalty card data, including the representativeness and potential bias of the data, the general activity and consumption patterns of cardholders, and how these compare between different groups of the population. The primary aims of this chapter were therefore to critically examine the data in terms of the capacity to offer insights into mobility and activity patterns, and to provide contextual analysis of the general activity patterns of cardholders which informed the statistical analyses undertaken in Chapter 8.

7.2 Data overview

The loyalty card scheme run by the HSR allows customers who are registered to earn points for each transaction conducted, which in turn can be used for discounts on future purchases. Each HSR loyalty card is registered to an individual, rather than a household, and the HSR also operates schemes which offer double the number of points per transaction than the general loyalty scheme for those aged 60 and over and for new or expecting parents.

The data provided by the HSR comprised of loyalty card transactions between 04:17 01/04/2012 and 23:57 30/09/2014, i.e. all transactions where a loyalty card was used either to collect points or redeem a discount. Additional data were provided for customers (sex, year of birth and postcode), stores (store location and retail structure descriptions), and products (product categories over 6 hierarchical levels). To contextualise these data in terms of the smart card data and the administrative data used throughout this study, Figure 7.1 shows the time span of each dataset.

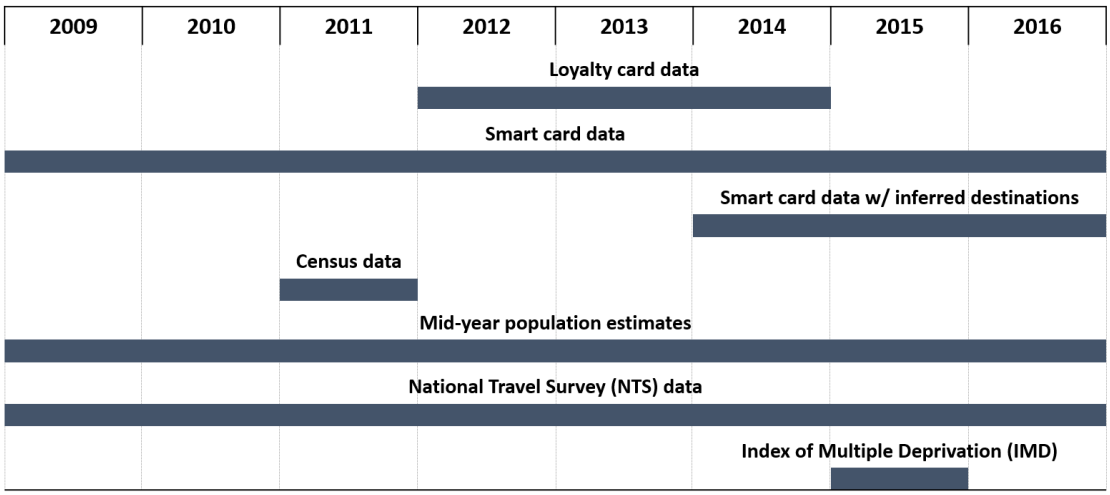


Figure 7.1 Time spans for data implemented throughout the thesis

The loyalty card and smart card data overlapped between 2012 and 2014, however inferred destinations for the smart card data were only available from 2014 onwards and therefore there was minimal overlap between the data implemented in Chapter 6 and the loyalty card transactions. Additionally, much of the research in this thesis relied upon census data, which was collected by the ONS on the 27/03/2011, over a year before the first recorded loyalty card transaction. Where population estimates were implemented rather than census data, mid-year population estimates were utilised as these are undertaken every year and therefore give a more accurate estimation of the size and makeup of the population. Population estimates were available over the whole study period and where these were used, the year is stated.

For analyses at the individual level the lack of overlap between data may impact the accuracy of results, however, for this study the focus was primarily on the aggregate mobility and activity patterns of populations and therefore the different time spans of the data were unlikely to have had a significant impact. Although analyses using the smart card data focused on changes in mobility patterns over time, the HSR data were primarily implemented throughout this chapter and the next to provide a static view of activity, as general retail behaviours are unlikely to have significantly changed over the study period. Although an increase in e-commerce has been recorded over recent years (Statista 2020b), the focus of much of these analyses on the older population reduces the impact this may have had on retail behaviours; as older consumers are much less likely to make use of e-commerce than the younger generation (Statista 2020c).

Additionally, this non-overlap between data is generally a bigger issue for traditional data collection methods, such as surveys and questionnaires, that are often collected during a single

time period. Unlike these traditional methods, consumer data are continually collected and can therefore provide real-time insights into phenomena, or data can be filtered for specific periods of interest. Although this study was limited by the data available to the CDRC for research, the principles and methods applied throughout this study are applicable to the data that are continually collected by the HSR and local transport authority. These methods can therefore be implemented on data collected outside of the study period of this thesis.

As the focus of this thesis was on residents of the West Midlands, the HSR data were filtered to only include those cardholders that had a registered postcode within the West Midlands metropolitan region. This resulted in a customer dataset of 670,939 cardholders, which translated to around 24% of the West Midlands population. As each loyalty card is allocated to a single account via an email address and home address, there was little likelihood of individuals holding more than one card.

One limitation of these data were potential errors in the registered addresses of cardholders. Much of the analyses conducted in this chapter and Chapter 8 relied on accurate information of the residential location of cardholders. Identifying the presence of address errors was vital. These address errors were likely to be the result of either an error in the recording of a cardholder's address or an address change since registering for the scheme which had not been updated. These errors may have resulted in transactions being recorded at a significant distance from a cardholder's registered address. Research by Lloyd and Cheshire (2019), using the same loyalty card dataset utilised in this study but for the whole of the UK, found discrepancies between store visiting patterns and registered addresses in approximately 3.6% of all cardholders; around half of which were estimated to be due to address errors and the other half due to address changes. Of this 3.6%, the highest proportion of discrepancies were found in the 18-20 population, with almost negligible proportions found for the older population. Although these relatively low proportions suggested that address changes and errors were unlikely to affect the results of aggregate analyses, cardholders with registered addresses within the West Midlands that were found to have never transacted at a store located either within the West Midlands metropolitan area or at stores within 10 miles of the study area were removed from future analyses. This resulted in 14,151 cardholders being removed from the West Midlands HSR data, which translated to 2.1% of cardholders; comparable to the results found by Lloyd and Cheshire (2019).

The following sections give a detailed overview of each of the datasets provided by the HSR, including analysing the quality and representativeness of the data, and the general trends that can be detected. Additionally, the activity and consumption patterns of both the older and

working age populations that could be extracted from these data were analysed in terms how these relate to mobility.

7.2.1 Customer data

The customer data provided by the HSR included a unique account ID for each cardholder, as well as their year of birth, gender (M, F or U for undisclosed) and postcode at the time of registration. Postcodes were provided in 'postcode units'; the smallest geographical units available. These areas cover 15 properties on average, although this ranges from 1 to 100. In the West Midlands there are 84,327 postcode units, of which 47,810 contained at least one cardholder. As a number of the analyses undertaken in this chapter utilised administrative data, these postcode units were aggregated to LSOAs to allow for this linkage. This resulted in all 1680 LSOAs in the West Midlands containing registered cardholders, with a minimum of 143 and a maximum of 1,730 cardholders per LSOA. Figure 7.2 shows the Location Quotient (LQ) of cardholders in the study area at LSOA level, following the same method as outlined in Section 4.2.5.1. This figure therefore shows how representative the data were of the population in each LSOA.

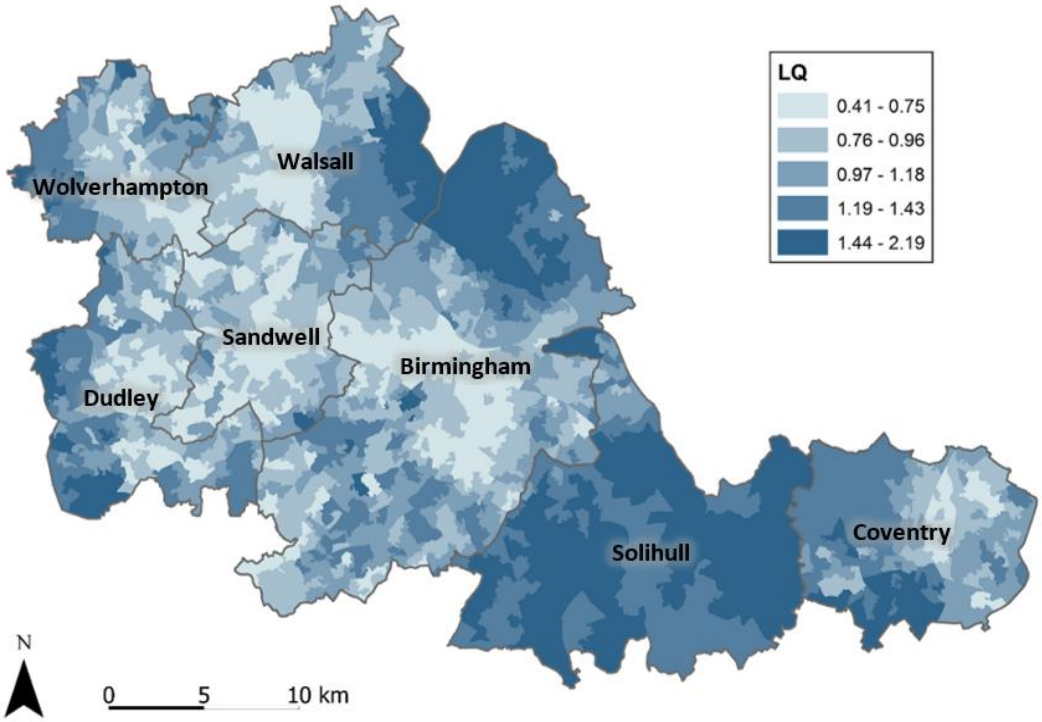
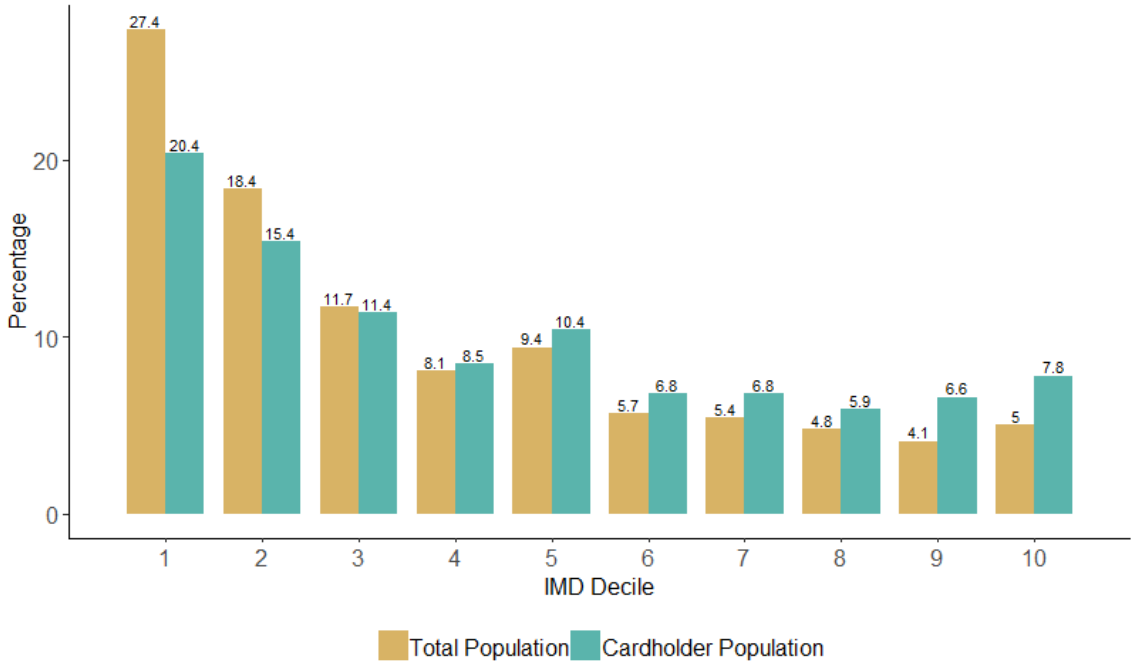


Figure 7.2 The location quotient of loyalty cardholders at LSOA level

This figure shows that there was not an even distribution of cardholders throughout the study area, with urban areas generally having lower LQs than suburban and rural areas. This suggests that the HSR data were more representative of certain groups of the population than others. Within Solihull, the most rural authority, the vast majority of LSOAs were overrepresented in the data. In comparison, central Birmingham and large areas of Coventry, Sandwell, Walsall, Wolverhampton and Dudley were underrepresented. The area of Sutton Coldfield in north Birmingham also contained large areas that were overrepresented. This area is largely suburban and one of the least deprived in the West Midlands. The LQ distribution shown in Figure 7.2 therefore suggests that suburban and rural residents were overrepresented in the data whilst those living in central urban areas were underrepresented. This was likely to have resulted in different socioeconomic and demographic groups having varying representations in the data, which is explored further in this section.

The variation in LQ throughout the study area suggested that IMD decile distributions varied between the cardholder population and the West Midlands population. Deprivation is widely discussed throughout this thesis in terms of the relationship with mobility and activity (see Section 2.2.2), and it has also been found in previous research that retail behaviour varies between those of different deprivation levels (Robinson et al. 2000). Comparing the distribution of cardholders residing in each IMD decile to the West Midlands population was therefore an important step in determining whether the retail behaviours extracted from these data were likely to be representative of the behaviours of the wider West Midlands population. Figure 7.3 shows the percentage of the cardholder population and the percentage of the total population that reside in each IMD decile. The distribution of IMD deciles throughout the study area can be found in Figure 7.12.



Figures 7.3 Percentage of the total West Midlands and HSR cardholder populations residing in each IMD decile (1 = most deprived, 10 = least deprived)

These IMD distributions showed similar patterns between the total West Midlands population and the HSR cardholder population, with the majority of cardholders residing in LSOAs of Deciles 1 and 2, and these proportions declining as deprivation decreases. As discussed in Chapter 4, the majority of LSOAs in the West Midlands are in Deciles 1 and 2. Additionally, these most deprived LSOAs were generally located in the most central urban areas and therefore had higher population densities than suburban and rural LSOAs. This means the higher proportion of cardholders residing in these areas would be expected.

Unlike the underlying West Midlands population, the HSR population had a greater share of cardholders residing in Deciles 4-10 and a smaller share in Deciles 1-3. This is likely due to the propensity of loyalty card schemes, and this specific HSR, to attract more affluent customers (CrowdTwist 2018). The transaction data used in this thesis were therefore likely to be under-representative of more deprived segments of the population. As previous results (see Chapter 5) have suggested that IMD may influence mobility levels, it was important that this was considered when interpreting results as these would likely be representative of only a sub-section of the population. To account for this, many analyses throughout this chapter and Chapter 8 included IMD data.

In light of GDPR implemented in 2018 (outlined in Section 2.3.1.1), the date of birth that was provided to the HSR by the cardholder was converted to a year of birth when these data were supplied to the CDRC. This presented an issue when attempting to identify cardholders that would also be eligible for concessionary travel. To combat this, ages were inferred using the mid-point of each year (1st July). The year of birth variable was converted to a date of birth variable, with every date of birth recorded as the 1st July. The age of each cardholder at both the time of sale and the start date of transactions (1st April 2012) was then calculated. Although this method likely resulted in a number of cardholder ages being inferred incorrectly, using the mid-point of each year was predicted to result in an equal number of overestimates and underestimates of age and would therefore minimise the effect this would have on aggregate results. Throughout this chapter, comparisons are made between the 'older' and 'working age' populations. The older population refers to cardholders who are eligible for concessionary travel, i.e. were at least 61 years old in 2012, with the working age population referring to cardholders aged between 16 and 60. Figure 7.4 shows the number of cardholders in 5-year age bands for male and female loyalty card holders.

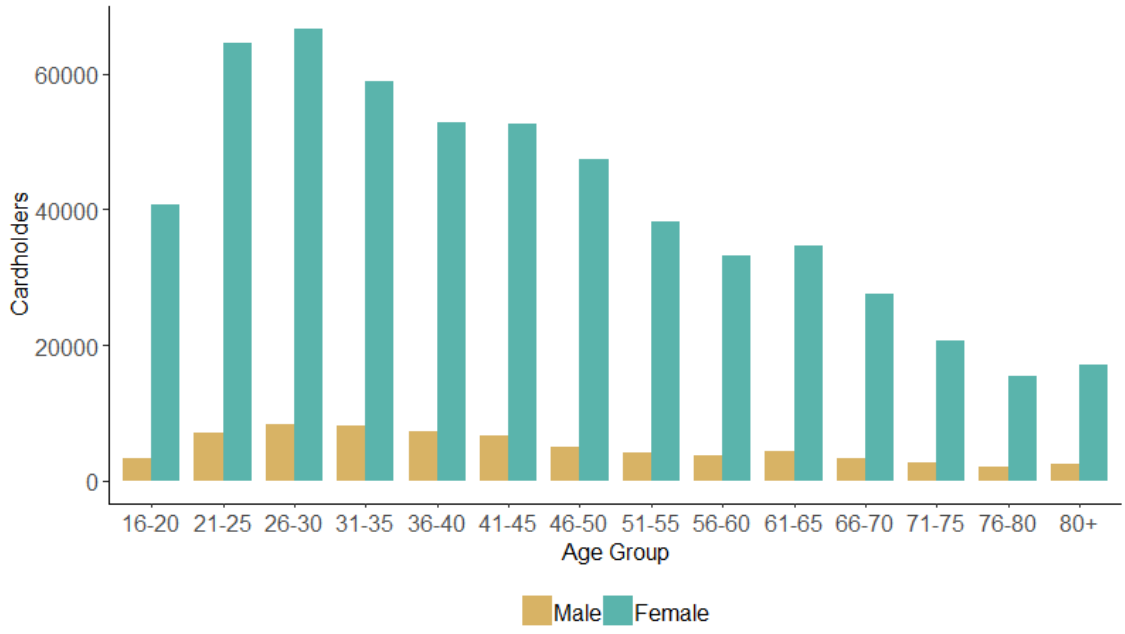


Figure 7.4 Number of male and female cardholders (5-year age bands)

Evidently there were a far greater number of female cardholders than male (569,490 compared to 67,844). This was to be expected, with previous research finding that women are more likely to belong to loyalty card schemes than men (YouGov 2018). The highest number of female

cardholders were aged between 21-30, with the number of cardholders then decreasing with age after the 26-30 age group. Similar patterns were seen with male cardholders, with the number of cardholders peaking between 26-35.

Unlike the smart card data, of which over 80% of the older population were cardholders, the HSR data covered a smaller proportion of the population. Only 22.7% of the working age population and 26.6% of the older population belonged to the scheme. Comparing the proportion of male and female cardholders with the underlying population, calculated using 2014 mid-year population estimates, it was found that approximately 89.3% of cardholders were female and just 10.7% male. This is in comparison to 50.7% female and 49.3% male for the total population of the West Midlands. This translated to an approximate representation of 40.7% of the total female population and 5.0% of the total male population. Comparing the age distributions of cardholders to mid-year population estimates can tell us how representative these data were of the population in the study area, i.e. whether the age group distributions shown in Figure 7.4 were an accurate depiction of the wider population. To achieve this, Figure 7.5 shows the percentage of the total population that were cardholders (5-year age bands) and Figure 7.6 shows the age group distributions of the total and cardholder populations, i.e. the percentage of each population that belonged to each 5-year age band.

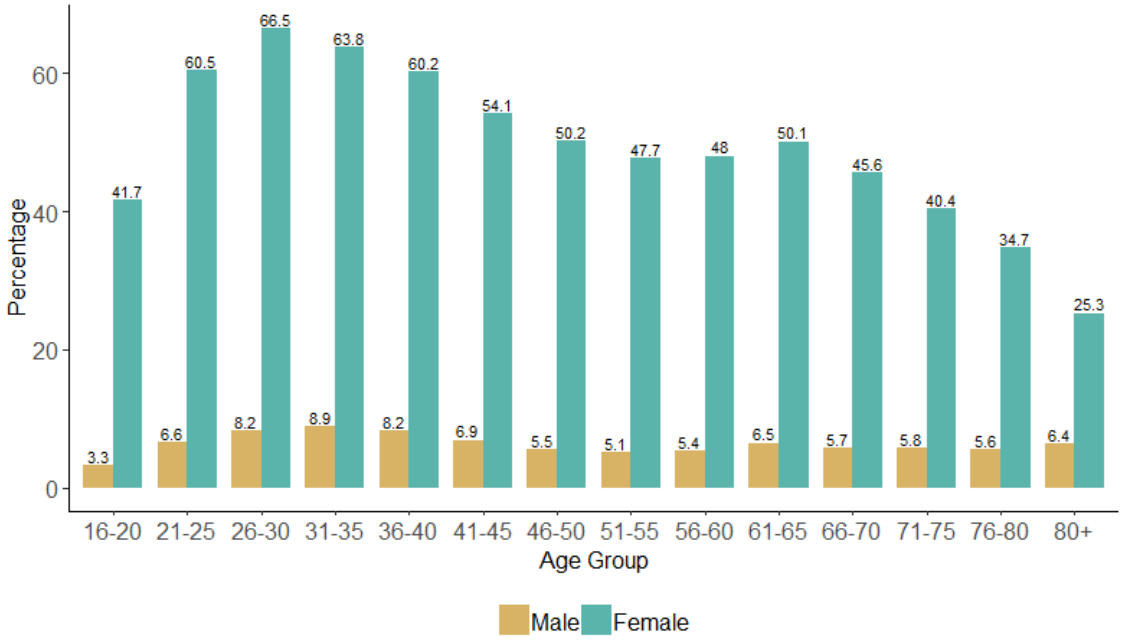


Figure 7.5 Percentage of the total male and female West Midlands population that are HSR cardholders (5-year age bands)

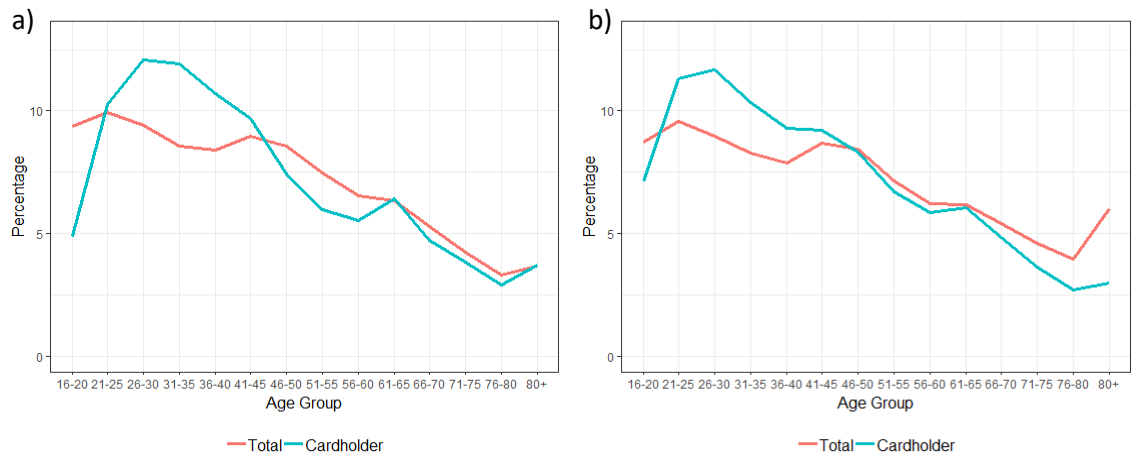


Figure 7.6 Age distributions for the a) male and b) female total and cardholder populations (5-year age bands)

Figure 7.5 shows a similar pattern to Figure 7.4, with a peak in the percentage of the total population that were HSR cardholders in the 26-30 age group for women and the 31-35 age group for men before again declining with age. Figure 7.6 tells us how the age distributions of the HSR cardholders compared to the wider population. These graphs show that the distribution of cardholders between age groups was similar to that found in the general population, however younger age groups, between 21-45, were marginally overrepresented. Additionally, there were fewer 60 plus males that were cardholders than would be expected when compared to the wider population. Understanding these differences was vital when interpreting the results of analyses and determining which groups of the population were likely to be represented in these results; and, therefore, the extent to which these could be applied to the wider population. Figure 7.5 tells us that the results of any analyses were likely to be dominated by the retail behaviours and activity patterns of female cardholders. The following analyses therefore only focus on the female population, unless stated otherwise. Although this focus on the female population may limit the relevance of the results to the wider population, results shown in Chapter 5 suggested that those smart card holders that exhibited changing bus usage were more likely to be female. Additionally, as discussed in Section 4.4, the female population are of particular interest due to the increase in retirement age for women meaning that a large number of women were "missing out" on concessions that they would have previously been eligible for.

In addition, analyses that did not consider age were likely to be dominated by the retail behaviour of younger populations, with significantly more cardholders than the older population and over-representations of these age groups in the data. To remedy this, the majority of

analyses undertaken throughout this chapter and Chapter 8 included information on the age of cardholders, either differentiating between 'older' and 'working age' cardholders or including cardholders' age groups in the analyses.

7.2.2 HSR store data

The store data included the unique store number, store name (i.e. the store location), store type (defined by an HSR classification), store format describing its primary retail function, the opening date and closing date (if applicable), postcode, and location coordinates for each HSR store. In total, there were 107 HSR stores located in the West Midlands metropolitan area. Stores could be broadly separated into 'Destination', 'Chemists', 'Community' and 'Convenience' types. 'Destination' store types represented the largest number of stores in the study area, with 38, followed by 35 'Chemist' stores, 24 'Community' stores and 10 'Convenience' stores. 'Chemist (Health centre)' represented a unique store type which provide a pharmaceutical service, often located within GP surgeries. 'Destination (EOT)' stores, described as 'edge-of-town' locations, were larger stores located primarily in retail parks and out-of-town shopping centres. 'Community' stores were typically smaller stores oriented around providing pharmacy services and serving local communities, whilst 'Convenience' stores were smaller stores located in urban areas, or in and around transport hubs in the case of 'Convenience (Travel)' stores. The number of each store type and store format in the study area are shown in Tables 7.1 and 7.2.

Table 7.1 HSR store types

Store Type	Count
Destination (Large high street)	9
Destination (Small high street)	18
Destination (EOT)	11
Convenience (High street)	5
Convenience (Travel)	5
Chemist (High street)	28
Chemist (Health centre)	7
Community	24
Total	107

Table 7.2 HSR store formats

Format	Count
Local Pharmacy	60
Health and Beauty	40
Flagship	4
Airport	3
Total	107

Due to 'Airport' format stores being located within Birmingham airport and therefore accessible only to airport users these were removed from further analyses, however the locations are still shown and clearly labelled in the figures shown in Section 7.2.2.1 for contextual purposes. Additionally, due to the low number of loyalty card transactions conducted at 'Chemist (Health centre)' store types, with these accounting for just 0.2% of total transactions, these stores were also removed.

7.2.2.1 Spatial distribution of HSR stores

HSR stores in the West Midlands serve different functions, both in terms of the services that they provide and the customers that they attract. The spatial distribution of stores and store types throughout the study area can provide insight into these functions. For example, stores serving local communities may be expected to be located on smaller high streets and in suburban areas, whereas large flagship stores may be expected to be located in central urban areas. Additionally, the spatial distribution of stores can also tell us about the accessibility of stores to different population groups, for example in terms of IMD. The following analyses aimed to summarise the distribution of stores throughout the study area. Subsequent sections then analysed how store patronage and consumption patterns differed between store types and populations.

To understand the spatial distribution of stores, a similar service area analysis method as outlined in Section 6.2.7.2 was implemented. This consisted of constructing 400m service areas along the road network for each HSR store; representative of an average 5-minute walk time. 400m thresholds were implemented as it has been found that this is the average distance that individuals are willing to walk to access public transport services (Kimpel et al. 2007) and therefore this can be assumed to also be the average distance an individual would be willing to

walk to access a facility from a public transport stop; ensuring consistency throughout this research. Figure 7.7 shows an example of this method, with the road network within 400m of the HSR store locations highlighted in red.

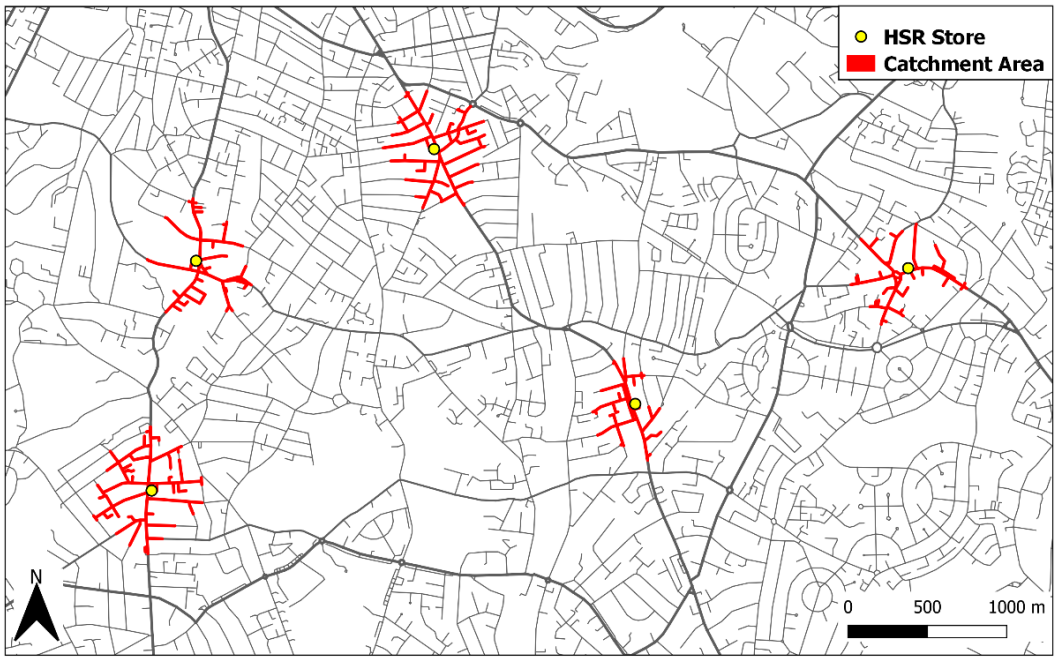


Figure 7.7 Example HSR store service areas

These service areas allowed for the areas in which HSR stores are located to be analysed in terms of both land use and the surrounding population. Conducting an intersect between the HSR store service areas and contextual data, for example population density recorded at LSOA level or land use recorded at land parcel level, allowed for the makeup of each service area to be deduced. This was calculated as the area of overlap between each service area and each contextual data unit. These were then converted to percentages to allow for comparability between store types.

7.2.2.1.1 Land use

The land use surrounding each store type tells us the function of these stores, i.e. whether they are likely to serve local communities or attract customers from further afield, as well as the saturation of HSR stores throughout the study area. Figure 7.8 shows the location of HSR stores as proportional points located at LSOA centroids (to protect the anonymity of the data provider) and the land use in the study area. Figure 7.9 shows the percentage of each store type's service areas that intersected with the four most common land uses.

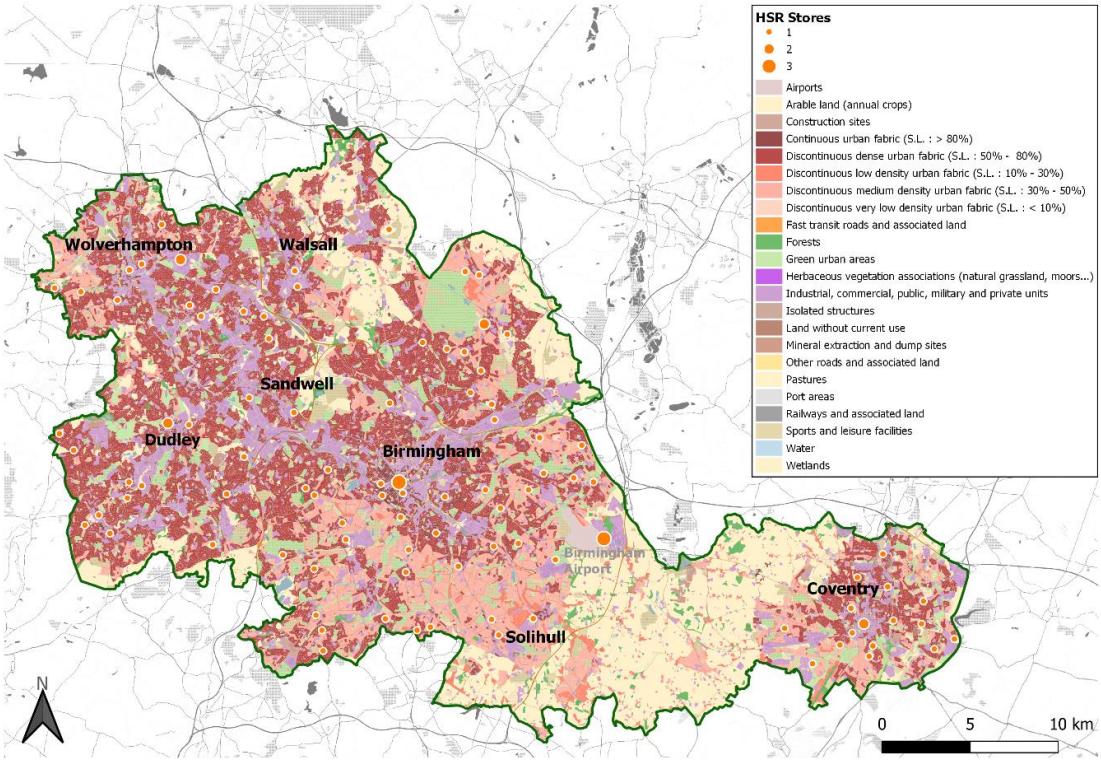


Figure 7.8 Store locations (shown as points proportional to the number of stores) at LSOA centroids and land use (European Environment Agency 2017)

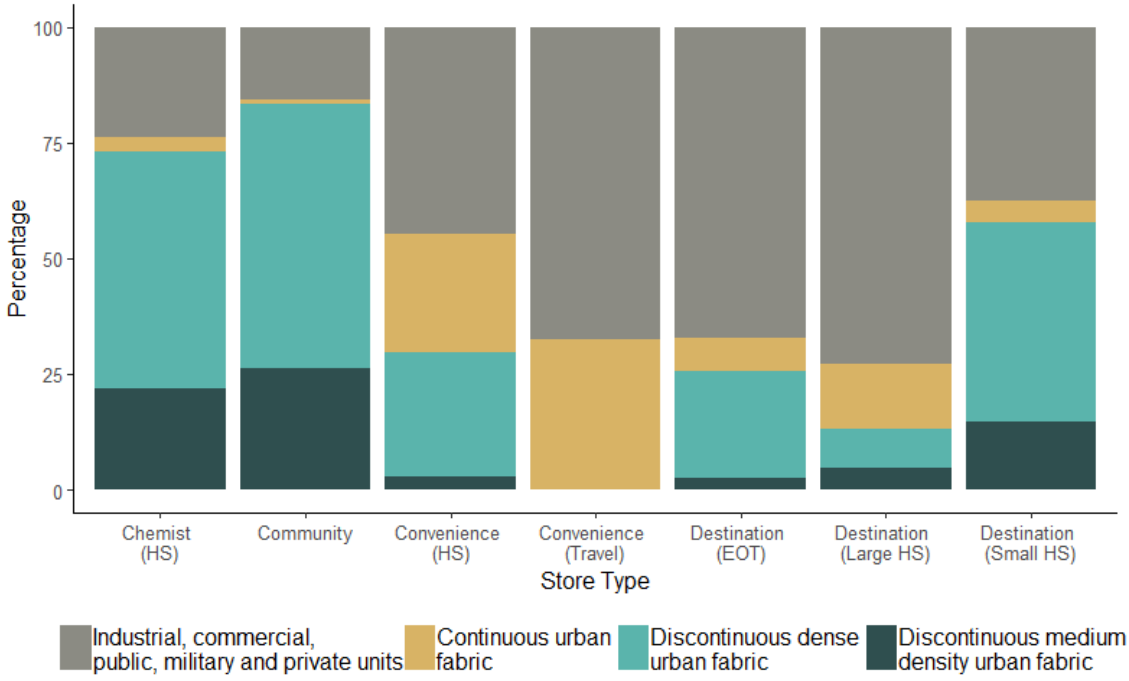


Figure 7.9 Percentage of each store type's service areas that intersected with the four most common land uses

Clear differences were evident in the land use surrounding each store type. 'Industrial, commercial, public, military and private units' refer to areas that are under industrial or commercial use, including large shopping centres as well as public and private facilities such as hospitals, schools and universities (European Environment Agency 2017). 'Convenience (Travel)', 'Destination (EOT)' and 'Destination (Large HS)' store types were primarily located in these areas. These store types were most likely located within shopping centres and retail parks, in particular 'Destination (EOT)' stores which were located in 'edge-of-town' retail areas. The high proportion of this land use within the service areas of 'Convenience (Travel)' store types was due to stores located within Birmingham New Street Station which is situated in close proximity to the Bullring shopping centre in central Birmingham. 'Convenience (HS)' and 'Destination (Small HS)' stores also contained a high proportion of 'industrial, commercial, public, military and private units' land use due to their location on high streets and retail areas.

'Continuous urban fabric' refer to areas where >80% of the land surface is covered by impermeable features like buildings, roads and artificially surfaced areas, i.e. urban structures and transport networks dominate the area. 'Discontinuous dense urban fabric' refer to areas where 50-80% of land surface is covered by impermeable features. 'Discontinuous medium density urban fabric' refer to areas where this is 30-50% (European Environment Agency 2017). In terms of 'continuous urban fabric', 'Convenience (HS)' and 'Convenience (Travel)' store types contained the highest percentages of this land use, followed by 'Destination (Large HS)'. These stores were therefore primarily located in built-up, urban areas and within central retail areas. 'Chemist (HS)', 'Community' and 'Destination (Small HS)' tended to be located in less built-up areas. The higher proportion of land classed as 'discontinuous dense urban fabric' suggests that these store types were located in suburban and smaller urban areas and were therefore more likely to serve local communities.

7.2.2.1.2 Population density

In addition to land use, understanding the location of HSR stores in terms of population density can give insight into the function of these stores and the types of individuals which may visit these stores. Figure 7.10 shows the location of HSR stores as proportional points at LSOA centroids and population density, i.e. the number of residents per one square kilometre at LSOA level. Figure 7.11 shows the average population density, for both the older and working age populations, within the service areas of each HSR store type.

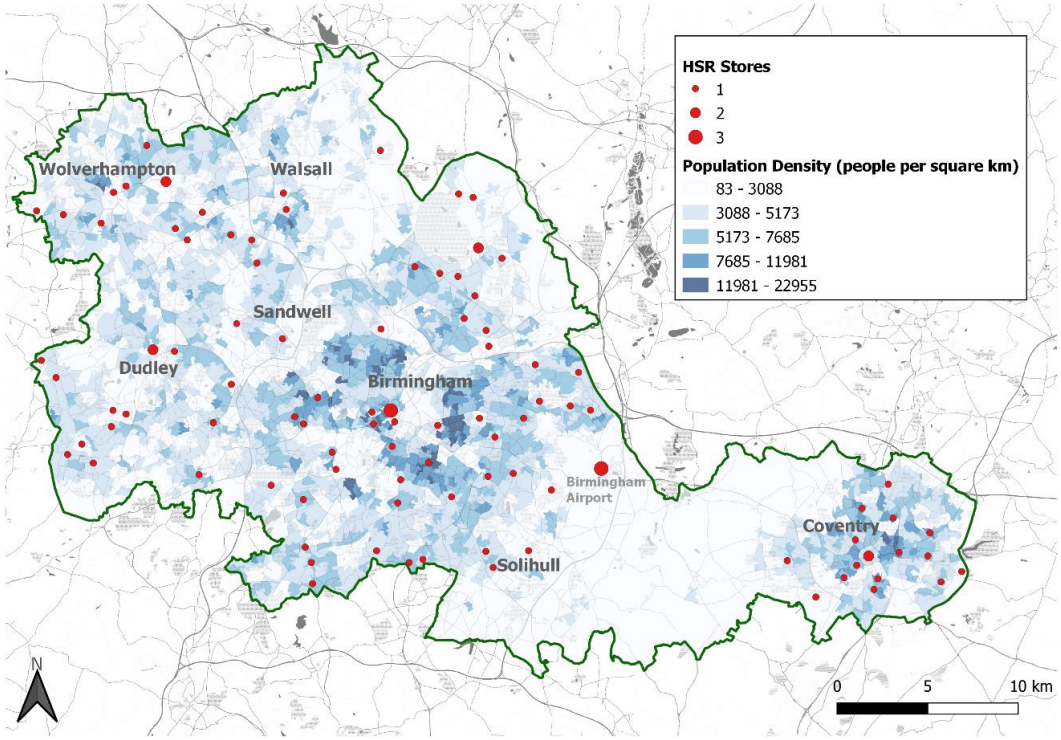


Figure 7.10 Store locations (shown as points proportional to the number of stores) at LSOA centroids and population density (residents per km^2)

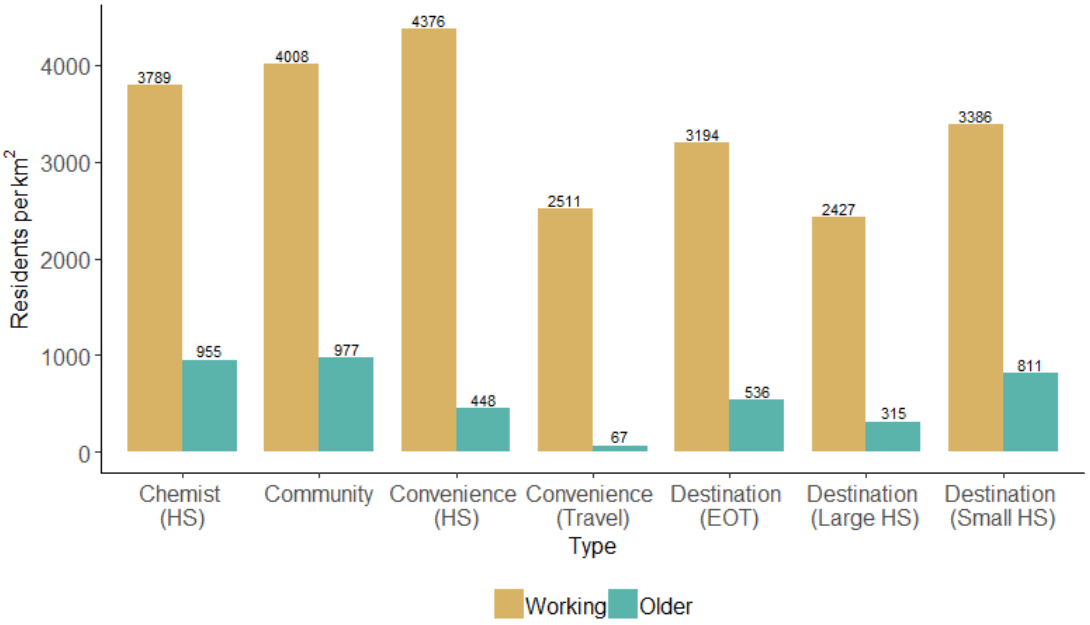


Figure 7.11 Average density of the older and working age populations in the service areas of each store type

Clear variations were evident between the population densities in the areas surrounding each store type, and between the older and working age populations. In terms of the working age population, 'Chemist (HS)', 'Community' and 'Convenience (HS)' stores were located in the most densely population areas, followed by 'Destination (Small HS)' and 'Destination (EOT)' store types. Comparing these to the land use percentages shown in Figure 7.9, these stores tended to be located in areas with a high percentage of 'discontinuous dense urban fabric' land use. This land use was less built-up than 'industrial' or 'continuous urban fabric' areas and therefore more likely to contain residential, rather than commercial, areas within their service areas.

There were clear differences evident between the population densities, and therefore distributions, of older and working age residents. The lowest density of older residents was found in the areas surrounding 'Convenience (Travel)' store types which were also located in the most central urban areas, with all service areas made up of 'industrial, commercial, public, military and private units' and 'continuous urban fabric' land uses. In comparison, 'Chemist (HS)', 'Community' and 'Destination (Small HS)' store types had the highest densities of older residents. These stores contained high proportions of 'discontinuous dense urban fabric' and 'discontinuous medium density urban fabric' land uses, and therefore were more likely to be located in less densely populated, suburban areas. Later in this chapter the store patronage patterns of the older and working age population are explored. Understanding where these populations reside in regard to different store types can be valuable in explaining differences in the store patronage patterns of different populations and their mobility levels. For example, cardholders that shop at local, community stores versus those that are more likely to travel to central urban retail areas.

7.2.2.1.3 IMD

The deprivation of the areas surrounding HSR stores can tell us about the distribution of stores throughout the study area, the accessibility of these stores to different population groups, and can also be used to explain the differences in shopping habits and store visiting patterns that are explored later in this chapter. Figure 7.12 shows the locations of HSR stores as proportional points at LSOA centroids and the IMD decile of each LSOA.

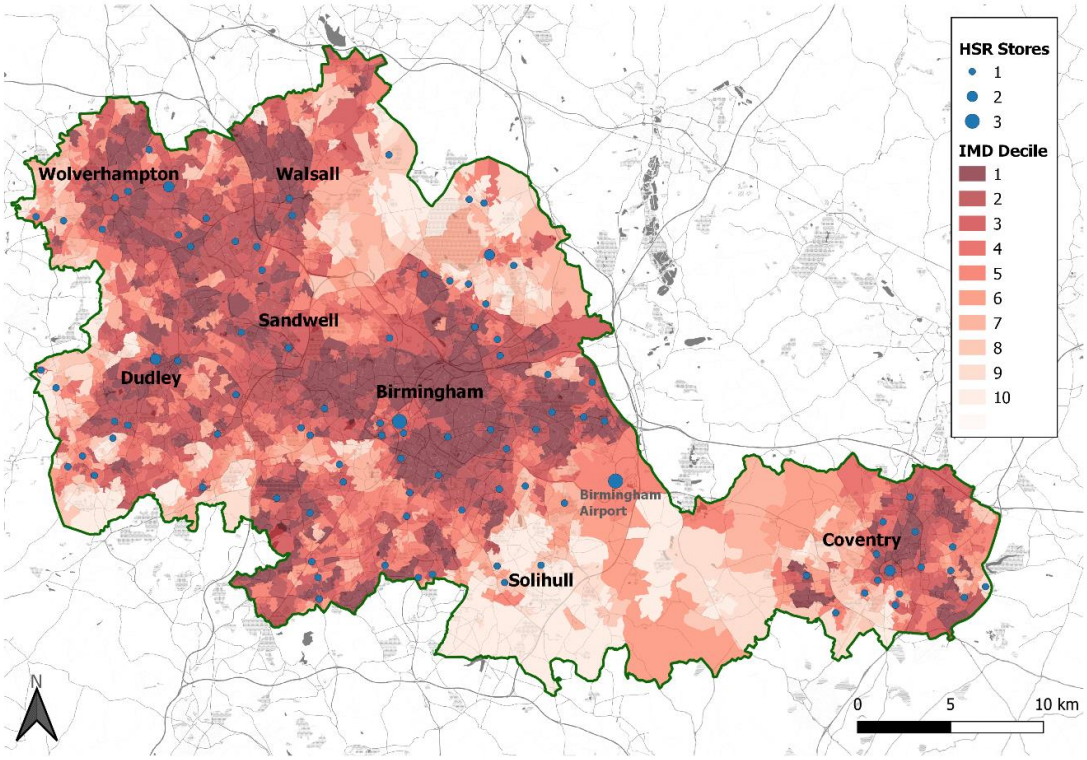


Figure 7.12 Store locations (shown as points proportional to the number of stores) at LSOA centroids and IMD deciles

Figure 7.12 shows that there did not appear to be a clear spatial pattern in the IMD deciles in which HSR stores were located. Central urban areas, in particular Birmingham, contained a large proportion of LSOAs in the most deprived IMD deciles and so it would be expected that a number of HSR stores were located within these most deprived areas. Nonetheless, a number of stores were also located in less deprived areas, such as those in Solihull and in the northern Birmingham area. Figure 7.13 explores this further in terms of store type; showing the IMD decile distributions of populations residing in the service areas of each store type.

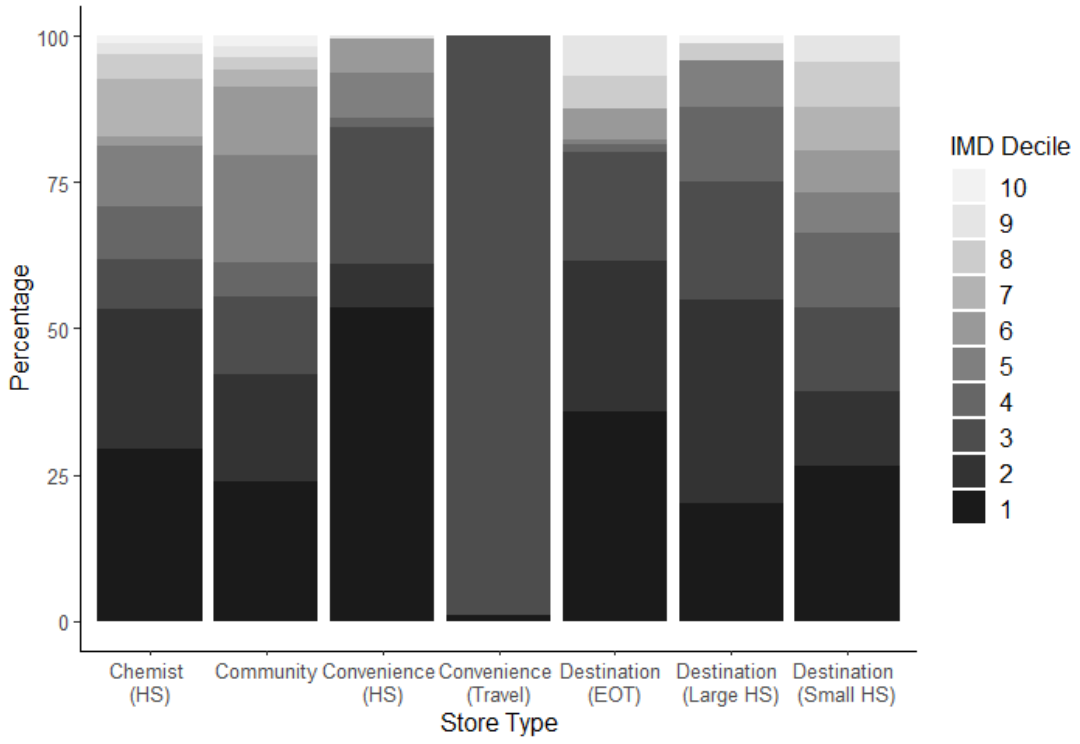


Figure 7.13 IMD decile distributions of populations residing in the service areas of each store type

Figure 7.13 shows that the service areas of each HSR store type differed significantly in terms of their IMD decile distributions. 'Convenience (Travel)' store types solely consisted of HSR stores located within Birmingham New Street station; explaining the large percentage of this service area located in Decile 3.

In terms of the other store types, 'Convenience (HS)' stores were likely to be located in the most deprived areas, with a large proportion of both the populations within these service areas residing in Decile 1 and almost negligible proportions in the least deprived LSOAs. These stores tended to be located in some of the most central urban areas, with high densities of working age residents. Similar IMD distributions could be observed between 'Chemist (HS)' and 'Community' store types, with the majority of service areas containing residents of Deciles 1 and 2. Similarly, 'Destination (EOT)' and 'Destination (Large HS)' also contained large proportions of residents of Deciles 1 and 2, and generally much fewer residents of less deprived deciles. 'Destination (Small HS)' stores exhibited a much more even distribution across IMD deciles, with similar proportions of residents residing in each decile.

The distributions shown in Figure 7.13 suggested that HSR stores were most likely to contain residents of the most deprived deciles. With the majority of LSOAs and residents in Deciles 1

and 2 (see Section 3.2.2), this was unsurprising and reflected the makeup of the study area. The differences between the distributions for each store type suggested that HSR stores served different purposes, such as local shopping or destination shopping, and were also likely to attract different types of customers. Store patronage by different population groups is explored further in Section 7.2.3.2; focusing on differences between the older and working age populations and residents of different IMD deciles.

7.2.3 Transaction data

The transaction data provided by the HSR contained a unique account ID for each cardholder, the store ID at which the transaction took place, the level 6 product ID of the item purchased, the value of the item, and the timestamp at which the transaction took place (DD/MM/YYY HH:MM). After filtering for the West Midlands metropolitan region, this resulted in a total of 44,969,650 records. However, as these transaction data were provided at product category level there were multiple records for each customer transaction; one for each product purchased. As the majority of analyses utilised transaction level data, these records were aggregated into single transactions. For products that were purchased as part of the same transaction, the account ID, store number and recorded timestamps were identical. This allowed for individual records to be grouped by these common variables. The sum of the product values and a count of the number of products bought were then calculated for each transaction. This resulted in a dataset containing 17,773,044 individual transactions made by 656,788 cardholders. Of these, there were 3,242,592 transactions made by 136,772 older cardholders and 14,530,452 transactions made by 520,016 cardholders of working age. Aggregate statistics of the transaction data are shown in Table 7.3 and age breakdowns shown in Figures 7.14 and 7.15.

Table 7.3 Transaction statistics

Average Statistics	**Working age population**		**Older population**	
Number of transactions	Average = 27.94	Min = 1 Max = 1538	Average = 23.67	Min = 1 Max = 645
Number of products per transaction	Average = 2.65		Average = 2.03	
Value (£)	Average = 19.19		Average = 16.46	
Stores visited	Average = 4.99	Min = 1 Max = 99	Average = 3.90	Min = 1 Max = 50

Slight differences were evident in the transaction statistics between the older population and the working age population. On average, the working age population tended to have made more transactions, purchased more products per transaction, had a higher basket value, and visited more distinct stores than the older population. These differences are explored in greater detail throughout this section.

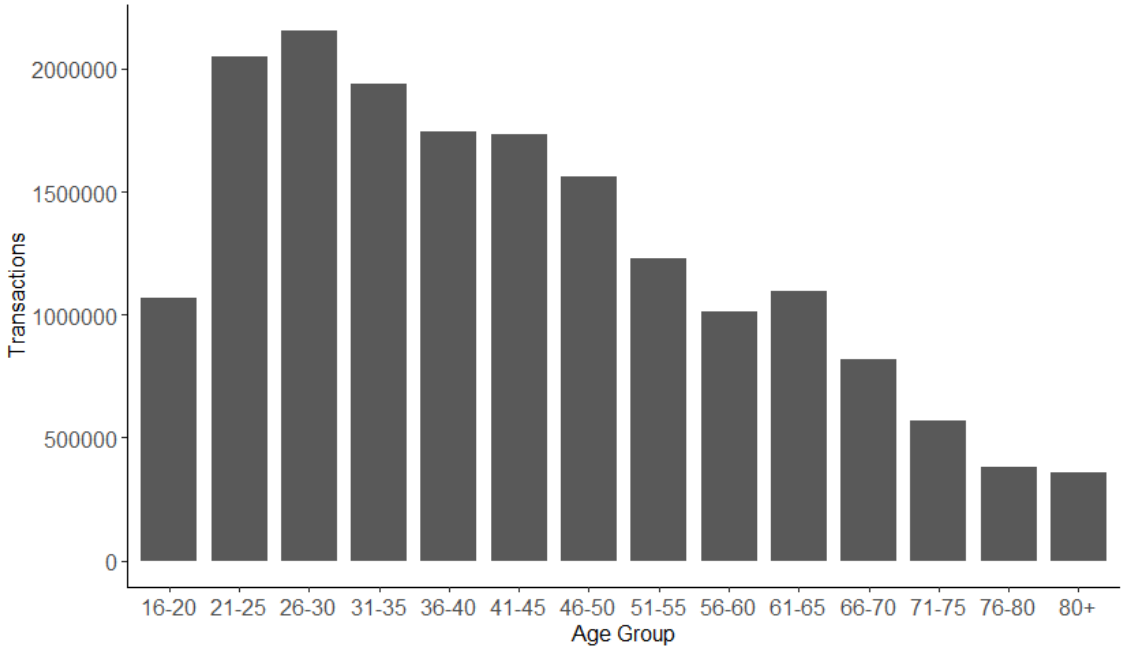


Figure 7.14 Total number of transactions per age group

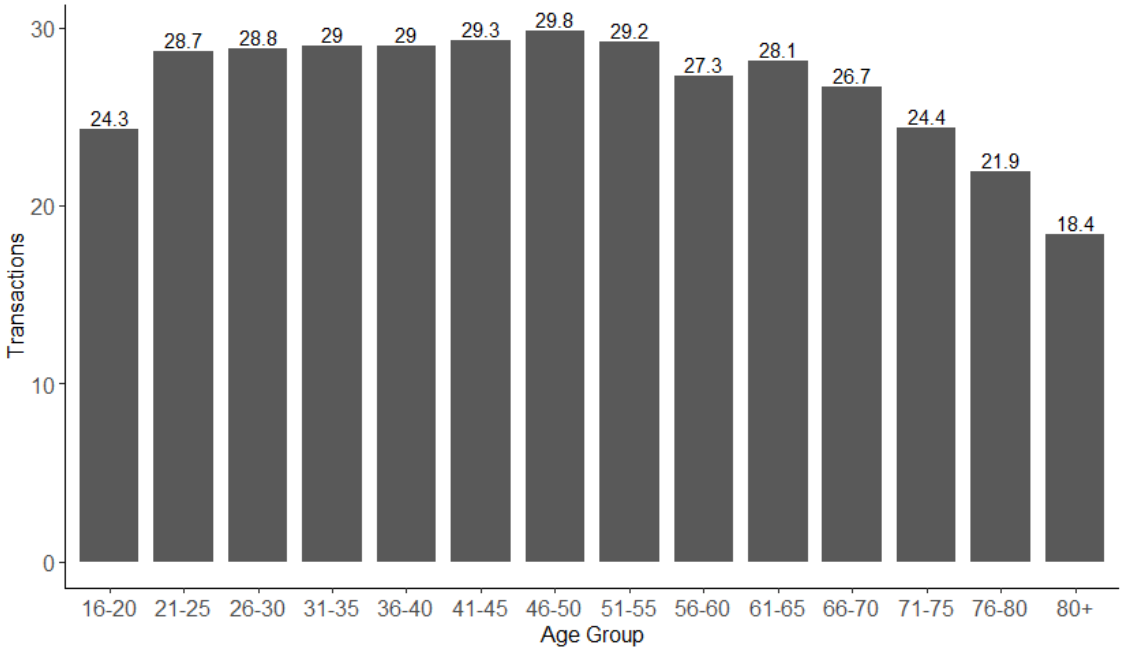


Figure 7.15 Average number of transactions per age group

Figures 7.14 and 7.15 show the total number of transactions undertaken by cardholders of each age group and the average number of transactions per cardholder undertaken across the 2.5 year study period respectively. Figure 7.14 shows the majority of transactions recorded in the data were undertaken by cardholders between the ages of 21-35, with the total number of transactions declining after this. This was largely reflective of the age distributions shown in Figure 7.5, with the number of cardholders peaking for age group 26-30 before declining with age. Although this did not provide insight into the activity of different age groups, as the underlying population sizes vary, it did show that the majority of transactions were undertaken by these younger age groups and therefore any aggregate results would likely be biased towards the behaviour of younger cardholders. To account for this bias, the majority of analyses undertaken in this chapter and Chapter 8 accounted for the age of cardholders; either grouping data into 'older' and 'working age' categories or including an 'age group' variable in the analyses.

To account for the underlying populations in each age group, Figure 7.15 shows the average number of transactions undertaken by cardholders over the entire study period for each age group, i.e. the total number of transactions undertaken by cardholders of each age group divided by the total number of cardholders in these groups. This resulted in a different trend to that shown in Figure 7.14, with the average number of transactions remaining relatively steady between the ages of 21-65 before declining with age after this. These age groups are the most likely to be economically active, with average retirement age around 62 to 65 years across the study period, and therefore the most likely to have disposable income. The decrease in the number of transactions undertaken by the older population could be reflective of their economic circumstances with retired populations generally having less disposable income that those of working age (Statista 2020).

7.2.3.1 Identifying 'active' users

As the majority of analyses throughout this chapter and Chapter 8 relied on the HSR transaction data to provide insight into activity patterns, it was vital 'active' users were identified to be included in these analyses. As the HSR transaction data likely only captured a proportion of cardholders' retail transactions, in order to maximise how representative these were of the activity of the population, these activity patterns should only be derived from those that actively made use of the loyalty card scheme. This prevented the aggregate results of any analyses being biased towards the large proportion of cardholders that had made little use of the loyalty card scheme.

Identifying ‘active’ cardholders considered two factors: the number of transactions that had been undertaken and the period over which these transactions were made. This method therefore considered the longitudinal activity of cardholders rather than relying solely on transaction frequency. In order to quantify this longitudinal activity, a new variable was created in the data representing the number of ‘active weeks’, i.e. the total number of weeks that each cardholder undertook at least one transaction. In order to maintain a suitable level of minimum transactions and a suitable number of cardholders for analysis, a threshold of 10 ‘active weeks’ was implemented. This meant that any cardholders who had not made at least 10 transactions and therefore been active over a period of at least 10 weeks were removed from future analyses. This resulted in 125,273 inactive cardholders being removed from the analysis, with 531,515 cardholders remaining.

7.2.3.2 Store patronage

Variations in card usage across store types were also explored, providing insights into the broad consumption patterns of the cardholder population. Differences in the patronage of stores by different population groups can tell us about the function of these stores and the retail behaviours of different populations. Figure 7.16 shows the percentage of total card transactions, product consumption and spend by store type, and can therefore provide insight into both transaction behaviour and card usage across stores.

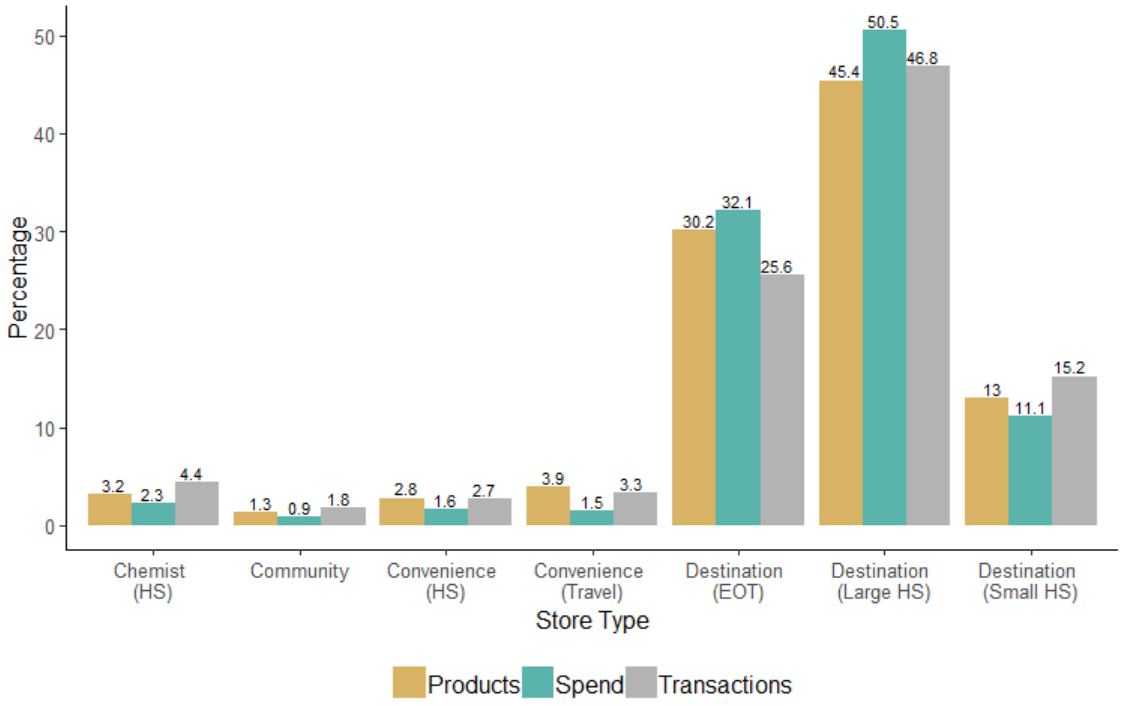


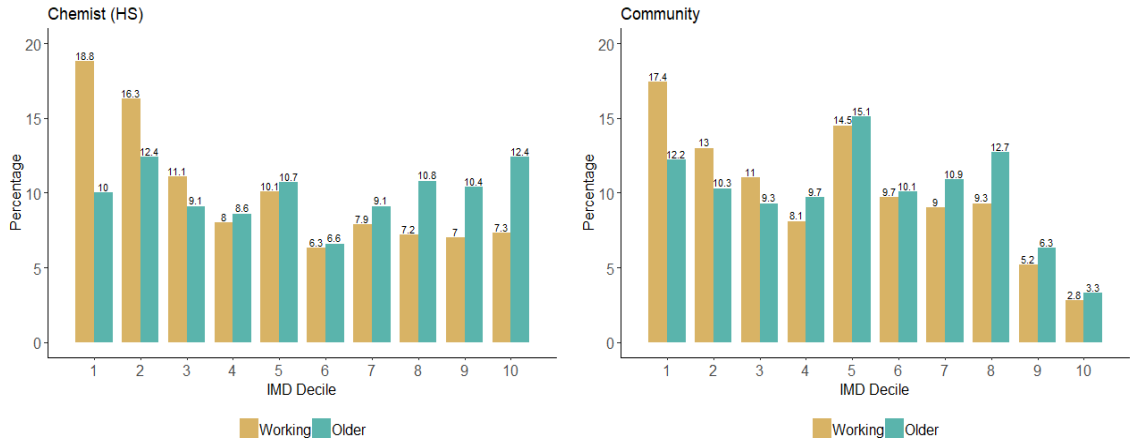
Figure 7.16 Percentage total card transactions, product consumption and spend by store type

Significantly high levels of card transactions, product consumption and spend were recorded at both ‘Destination (EOT)’ and ‘Destination (Large HS)’ stores, with ‘Destination (Small HS)’ stores representing the next highest volumes for all three variables. Lower levels of participation were observed in the remaining store types. These distributions were likely influenced by the HSR store network structure in the study area, which contained a higher number of ‘Destination’ store types than other store types, and thus would inevitably account for a higher percentage of overall data.

In addition, the number of products and the total spend differed between store types. This may be indicative of either a difference in purchasing behaviour between store types, or of the propensity of a cardholder to participate in the loyalty card scheme at different store types. This means there may have been a disproportionate representation of loyalty card transactions across different store types, which may not have been representative of actual transaction behaviour. Due to the lack of non-loyalty card data available, this difference in participation could not be explored further but was considered when interpreting the results of subsequent analyses by including information on store type.

7.2.3.2.1 Store patronage by IMD

As discussed in Section 7.2.2.1.2, stores types located in areas of different IMD deciles likely serve different functions and attract different groups of cardholders. Figure 7.17 shows the percentage of transactions made at each store type by residents of each IMD decile for both the older and working age populations.



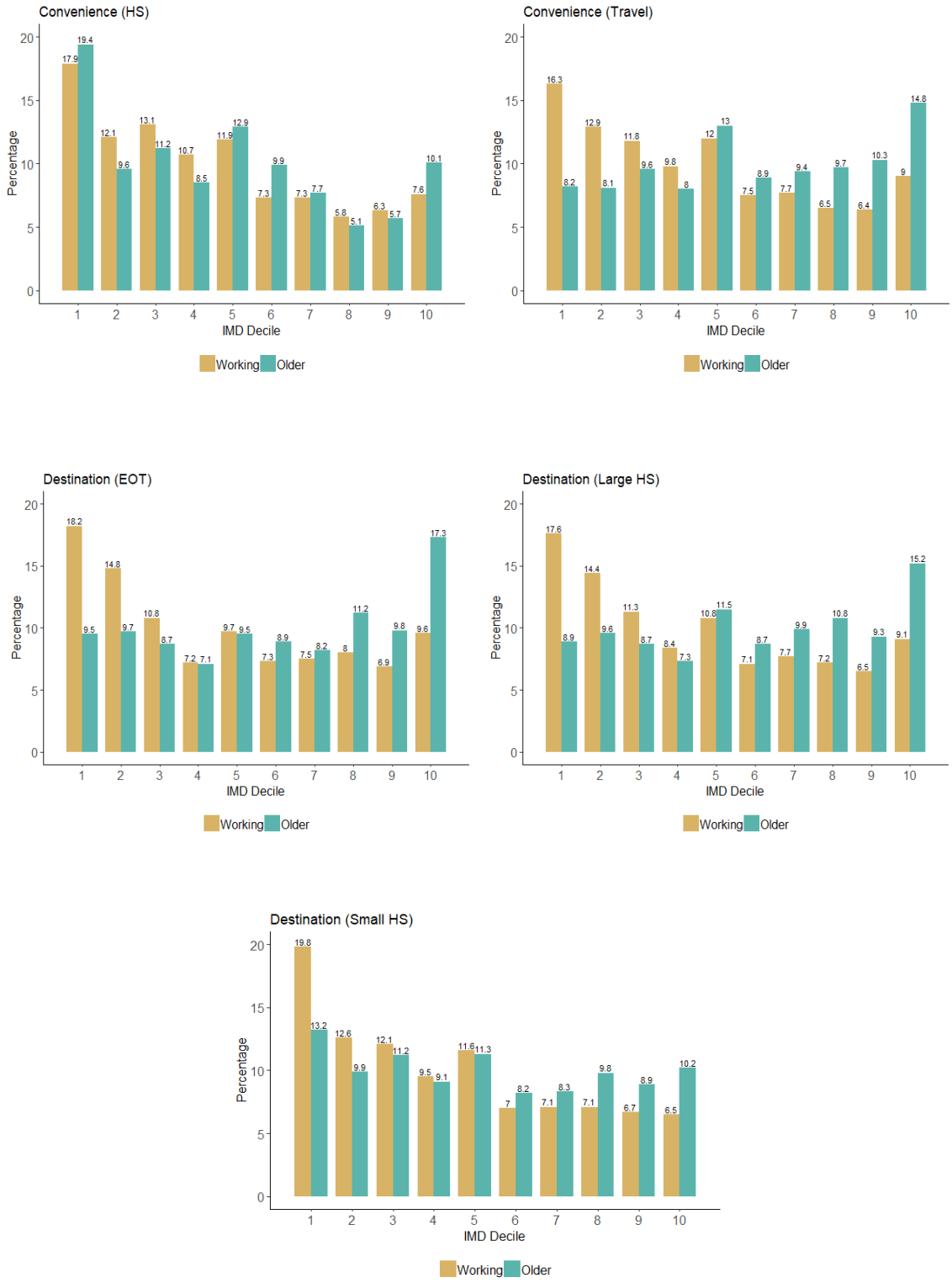


Figure 7.17 Percentage of transactions made at each store type by residents of each IMD decile

The IMD deciles of cardholders transacting at HSR stores showed differences between both the older and working age populations and the store type. As seen in Figure 7.3, there were more

cardholders residing in areas of Decile 1 than any other, and therefore it would be expected that the majority of transactions at each store type would have been undertaken by residents of these areas. However, this was found to not be the case for all types.

Focusing on transactions made at 'Chemist (HS)' stores, differences between IMD deciles and between populations were evident. The majority of transactions conducted at these stores were by working age cardholders and residents of Deciles 1 and 2; the most deprived LSOAs. In comparison, transaction distributions were more evenly dispersed between deciles for the older population. This may reflect the health of the underlying populations. For those of working age, more deprived cardholders are more likely to suffer from poor health (Stafford and Marmot 2003) and may therefore purchase more healthcare products and services that 'Chemist (HS)' stores offer than less deprived populations. As health generally declines with age, deprivation may have less of a significant impact on the purchasing of healthcare products than age for those in the older population. Additionally, Figure 7.11 showed that 'Chemist (HS)' store types had the second highest density of older residents within their service areas. A lack of mobility and tendency to shop at local stores (Robinson et al. 2000) may therefore explain why 'Chemist (HS)' store types attracted older cardholders from all IMD deciles.

Similar patterns could be observed for 'Convenience (Travel)', 'Destination (EOT)' and 'Destination (Large HS)' stores, with the majority of transactions for the working age population undertaken by cardholders of Decile 1, in comparison to Decile 10 for the older population. While the transaction distributions for the working age population generally coincided with the population distributions shown in Figure 7.5, those for the older population showed the opposite pattern. The higher proportion of transactions made by older cardholders residing in Decile 10 LSOAs may be indicative of the economic wellbeing of this population. More deprived older cardholders are more likely to solely rely on the state pension and may therefore undertake fewer transactions than those that are less deprived.

Additionally, these trends can be also be indicative of mobility and spatial store visiting habits. Figure 7.11 shows that 'Convenience (Travel)', 'Destination (EOT)' and 'Destination (Large HS)' stores had relatively low older population densities when compared to 'Community', 'Chemist (HS)' and 'Destination (Small HS)' stores. The higher proportion of transactions undertaken at these store types by less deprived cardholders may therefore be indicative of high levels of mobility; with these cardholders more likely to be able to undertake trips to stores other than their local stores, for example those located in central urban areas and out of town retail parks. This may be due to higher access to private transportation and disposable income, which could be used for taxi fares, in these populations meaning that stores located at further distances from their residential location are more accessible.

Figure 7.9 shows that ‘Convenience (Travel)’, ‘Destination (EOT)’ and ‘Destination (Large HS)’ store types were the most likely to be located in ‘Industrial, commercial, public, military and private units’ land use areas, i.e. within large shopping centres, retail parks or industrial estates. In comparison, ‘Community’, ‘Convenience (Travel)’ and ‘Destination (Small HS)’ stores were more likely to be located in areas with ‘discontinuous dense urban fabric’ and ‘discontinuous medium density urban fabric’, i.e. in less urban and more suburban areas, and these stores were more likely to attract older cardholders residing in more deprived areas. This suggests that more deprived older cardholders may shop more locally and be less likely to visit large retail centres and central urban areas. These shopping patterns are likely due to lower levels of access to private transportation in these groups which coincides with greater reliance of public transport, meaning that longer distance trips may be more challenging to undertake.

7.2.3.2.2 Store patronage by age

Understanding how store patronage between age groups differs can tell us about the relationship between shopping behaviours and age. Figure 7.18 shows the percentage of transactions undertaken at each store type by cardholders of each age group.

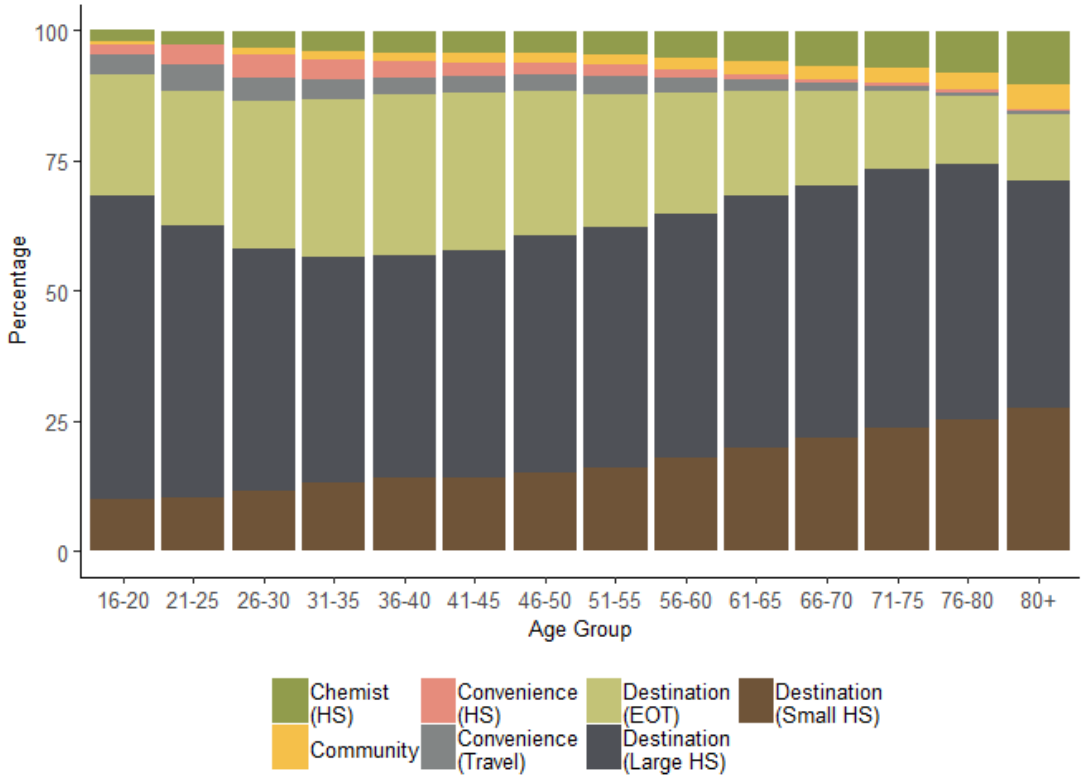


Figure 7.18 Percentage of transactions undertaken at each store type by 5-year age bands

The types of stores visited by cardholders of different ages showed distinct patterns. The percentage of transactions undertaken at 'Destination (Small HS)' stores clearly increased with age, with the smallest percentage found for the 16-20 population and the highest for the 80+ population. In contrast, visits to 'Destination (EOT)' store types showed the opposite trend with the percentage of visits made to these store types decreasing with age. 'Destination (Small HS)' and 'Destination (EOT)' stores are both 'destination' stores and therefore provided similar products and services. These differences in store visiting patterns were therefore likely due to the location of these stores rather than the store function. 'Small HS' stores are generally located in suburban communities and provide for local shopping purposes. These stores are also more accessible via public transport than 'EOT' stores which are located in out-of-town retail parks and would therefore be more easily accessible to the large proportion of the older population that lack access to private modes of transport (Mackett 2013).

The older population were also more likely to shop at 'Chemist (HS)' stores and less likely to visit 'Convenience (Travel)' and 'Convenience (HS)' store types than younger cardholders. This is likely a reflection of both the location of these stores and their function. Figure 7.11 shows that 'convenience' stores were more likely to be located in densely populated, urban areas whereas 'Chemist (HS)' stores were primarily location in 'discontinuous dense urban fabric areas'. Like previously discussed, the older population are more likely to shop in local, community stores rather than store located in central urban areas. 'Convenience' store types are primarily smaller stores located in and around transport hubs and in central retail areas. These stores are also geared towards meeting the needs of working and commuting populations, for example providing lunch and snacking options rather than the healthcare and medical services that are found in 'Chemist (HS)' stores. It is therefore logical that these stores would attract a greater percentage of the working age population, with the older population more likely to shop in local store types.

In terms of the younger population, these cardholders were more likely to shop at 'Destination (Large HS)' stores than any other age group, with the highest percentage of transactions found for the 16-20 age group. This likely reflects the spatial distribution of these groups, with younger age groups more likely to live and undertake recreational activities in central urban areas (Thomas et al. 2015). The highest percentage of transactions undertaken at 'Destination (EOT)' stores was found for cardholders between the ages of 31 and 45. As previously discussed, these 'edge-of-town' stores are generally not as easily reachable via public transport services and may therefore be inaccessible to younger populations that are not yet able to drive or may lack access to private transport and older populations who have lost the ability to drive. The differences in the store patronage patterns of different age groups are therefore a reflection of both the function that these stores provide and the areas in which they are located.

7.2.3.3 Spatial activity patterns

Previous sections have explored the more general store patronage patterns of cardholder. Spatial and temporal store visiting patterns can provide more granular insights; contributing to the understanding of both people and places by detecting where and when cardholders were active throughout the study area. The following sections explore the spatial and temporal activity patterns of the older and working age populations; benchmarking the typical store visiting patterns of these groups and identifying differences in terms of where and when these populations undertake retail activities. This section focuses on the store networks of cardholders, i.e. the spatial distribution of the stores they visit and the frequency of these visits.

Demand for both public transport services and key facilities, such as retail stores, are inherently linked, with demand for both also likely to vary between different groups of the population. For example, a large proportion of cardholders of working age would be expected to spend time at a work location for significant periods of the week. This may result in working age cardholders having a larger network of visited stores, as stores located both near a home and work location may be visited regularly. Understanding the differences between both groups of the population, and therefore being able to make inferences about the potential demand for services, can be valuable for both transport providers and retailers aiming to meet these needs.

These store networks needed to account for both the number of stores visited and the frequency of transactions at each store. Identifying cardholders' regular stores, i.e. the stores that are visited most often, can provide an understanding of the spatial distribution of a cardholder's regular network, which in turn can be used as a proxy for mobility. For example, cardholders with a high number of regularly visited stores over a large geographical area may be determined to be more mobile than those that have only transacted at their most local stores. Aggregate differences between the older and working age populations may also provide insight into how store visiting patterns change with age.

In addition to analysing cardholders' regular store visiting patterns, it was important that the spatial distribution of irregularly and rarely visited stores were also investigated. These stores can also be indicative of mobility levels, with a large range of visited stores, whether these were visited frequently or infrequently, suggesting a high level of mobility. Infrequent transactions conducted at stores a significant distance from a cardholder's registered address may be indicative of a cardholder being on holiday or visiting friends and family, all of which also suggest high mobility.

The method for constructing store visiting networks involved allocating each cardholder a regular, irregular and rare store network. For a cardholder to be allocated a store network, it was determined that they had to be an 'active' cardholder; preventing infrequent customers being

allocated a regular store network which would skew the aggregate results. These active cardholders were identified in Section 7.2.3.1 and consisted of 531,515 cardholders.

The method for identifying these store networks then consisted of determining what constituted a ‘regular’, ‘irregular’ and ‘rare’ store. A number of methods could have been implemented for identifying these stores, for example identifying the most visited *X* number of stores per cardholder or using a proportional threshold. As this analysis aimed to account for both the spatial distribution and the number of stores that a cardholder had transacted at, a proportion threshold was used. This proportional threshold allowed for cardholders to have different numbers of regular, irregular and rare stores. This was an important factor in this analysis and also allowed for the consideration of store-network density, i.e. some areas may have several stores in close proximity whereas others may be limited to just one local store.

Firstly, the number of times a cardholder transacted at each store was determined, and the proportion of a cardholder’s total transactions calculated. After analysis of the distribution of these proportions, the following thresholds were determined:

Regular Store: >20% of all transactions

Irregular Store: 10-20% of all transactions

Rare Store: <10% of all transactions

This resulted in a list of regular, irregular and rare stores for each ‘active’ cardholder. As these were conducted at individual level the results for each cardholder could not be presented due to disclosure controls. Figure 7.11 does, however, show an example of these store networks with simulated data so as not to compromise the anonymity of the cardholder or the data provider.

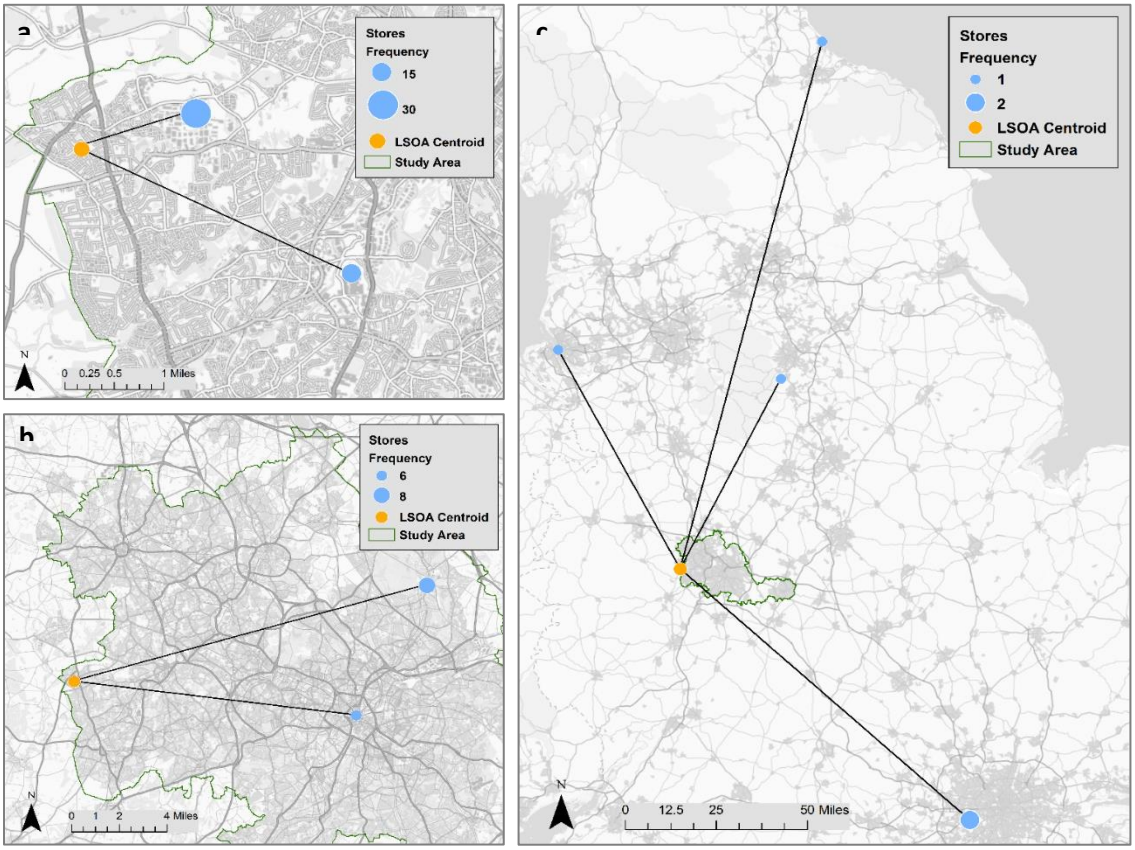


Figure 7.19 Simulated a) regular, b) irregular and c) rare store networks for a single cardholder

Figure 7.19 shows typical regular, irregular and rare store networks, with stores allocated as regular tending to be in closer proximity to a cardholder's registered LSOA than those that were irregularly or rarely visited. Differences in these networks are explored further throughout this chapter. Figure 7.20 shows the percentage of store types that were allocated as regular stores for the working age and older populations.

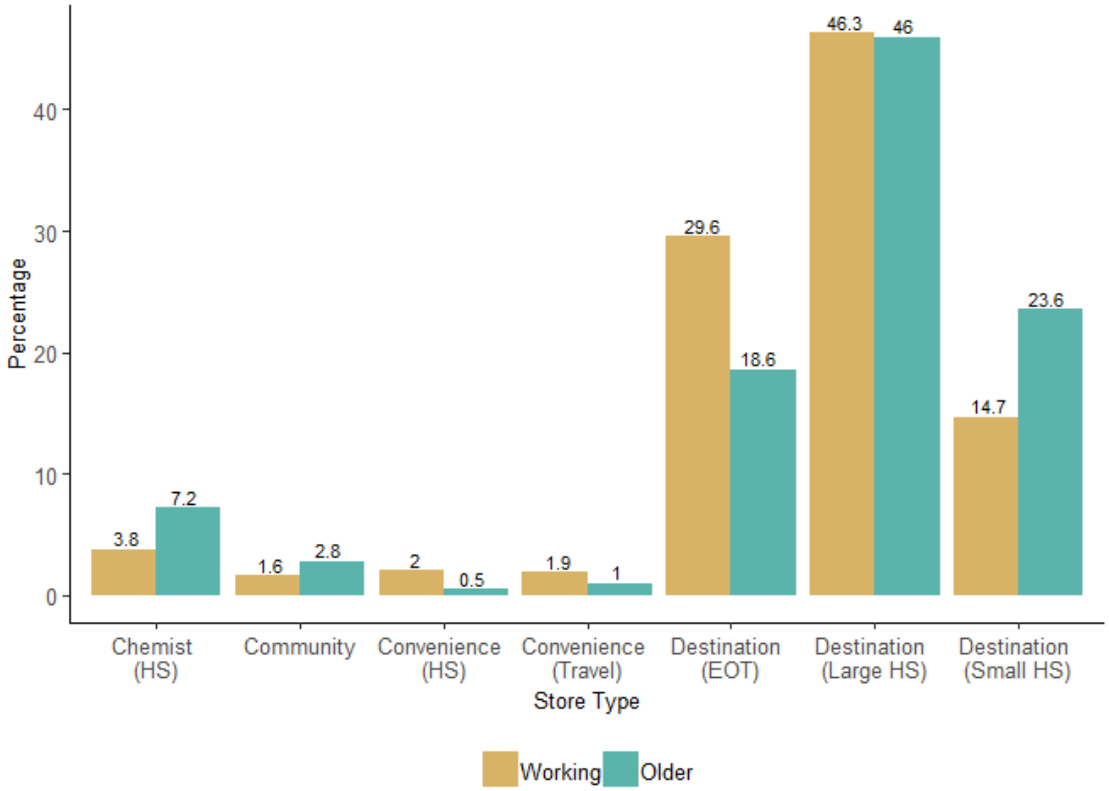


Figure 7.20 Percentage of store types allocated as 'regular' stores

These results suggested that 'Destination' store types were more likely to be allocated as a regular store than other store types. This is unsurprising, due to the high number of 'Destination' stores in the study area, as shown in Table 7.1. It was also found that the working age population were more likely to regularly visit 'Destination (EOT)' stores than the older population, with the older population more likely to regularly visit 'Chemist (HS)' and 'Destination (Small HS)' store types. 'EOT' stores tended to be located in out-of-town shopping centres and retail parks, which may have limited public transport access. As discussed in Section 7.2.3.2.2, this may limit the accessibility of these store types to groups of the population that lack access to private transportation. The propensity of the older population to regularly visit chemist store types is indicative of this population making use of prescription services and purchasing healthcare products.

Comparing the average number of stores allocated to each network between age groups can provide insight into how mobility level change with age. Although a low number of regularly visited stores may not necessarily be indicative of low mobility, due to differences in the store network, a high number of irregular and rarely visited stores located a significant distance from a cardholder's registered LSOA is likely indicative of a high level of mobility.

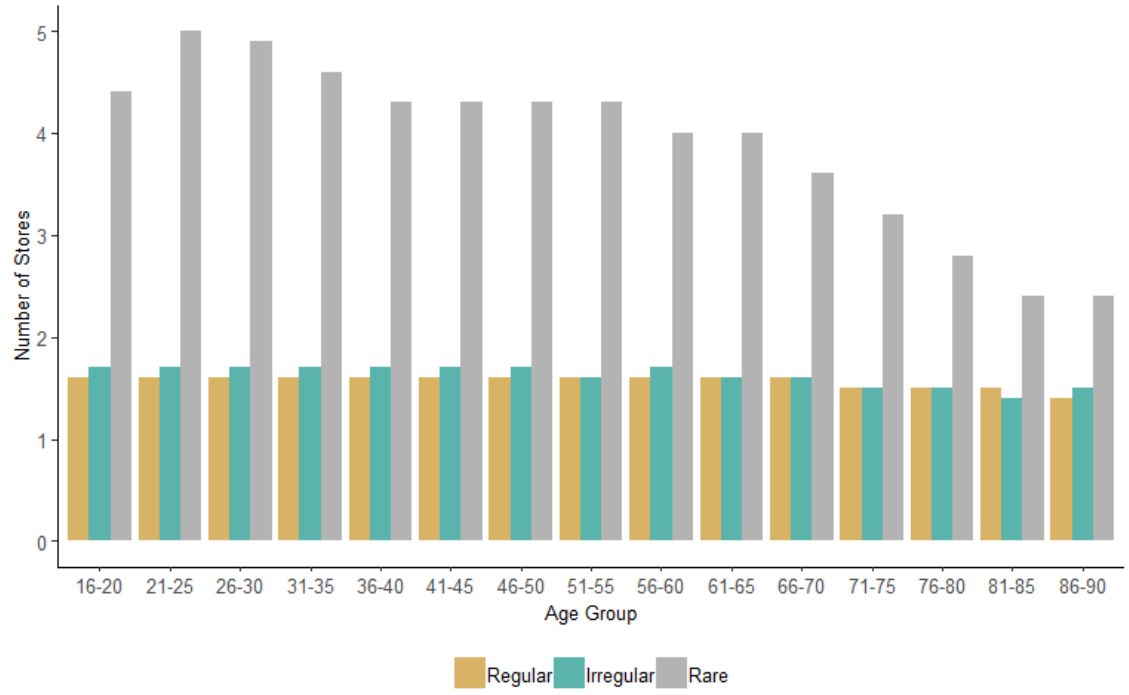


Figure 7.21 Average number of stores allocated as regular, irregular and rare (5-year age bands)

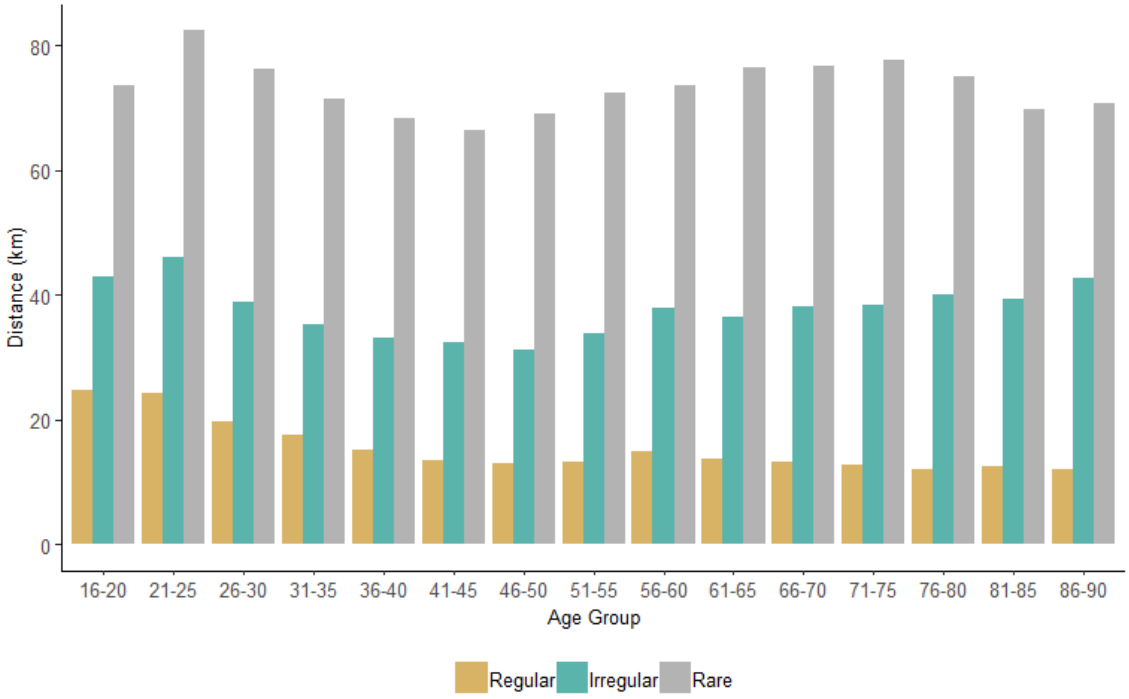


Figure 7.22 Average distance to regular, irregular and rare stores (5-year age bands)

Figures 7.21 and 7.22 show differences in store networks between different age groups. In terms of the average number of stores allocated to each network, rare store networks showed the largest differences. With the exception of the 16-20 age group, the average number of stores allocated to a cardholder's rare network decreased with age, from 5.03 for the 21-25 age group to 2.35 for the 86-90 age group. A slight decrease in the number of regular and irregular stores was also evident with increased age. The higher number of stores allocated to the store networks for the 21-25 age group was indicative of a more transient group with a greater risk of failing to update their registered address information (Lloyd and Cheshire 2019). The lower number of rare stores recorded for the older population suggested that these cardholders had a smaller store network than younger age groups, which could therefore reflect limited mobility in older populations.

Figure 7.22 shows less of a clear correlation between age and average distance to irregular and rare stores. Like Figure 7.21, the average distance travelled to each store type was greatest for the 21-25 population. The 41-45 and the 46-50 age groups exhibited the lowest average distance to each store type, with average distance then increasing up to the 71-75 age group. This suggests that although the older population may have smaller store networks in terms of the number of stores visited, this was not necessarily reflected in the spatial extent of these networks.

The exception to this was the average distance travelled to regularly visited stores, which clearly decreased with age. Again, this could reflect the more transient nature of the younger population; with the possibility of a registered address at a 'home' location with significant periods of time also spent at a second location, for example a university or a work location. Alternatively, this could suggest that the older population generally transact more locally than the working age population. This was discussed in Section 7.2.3.1, with older populations more likely to shop at more 'local' store types, such as 'Destination (Small HS)' and 'Community' stores.

In addition to age, IMD is closely related to activity, with those that are more deprived often at-risk of transport disadvantage and limited mobility (Hine 2012). Figures 7.23 and 7.24 show how these store networks differ by IMD decile for both the working age and the older populations.

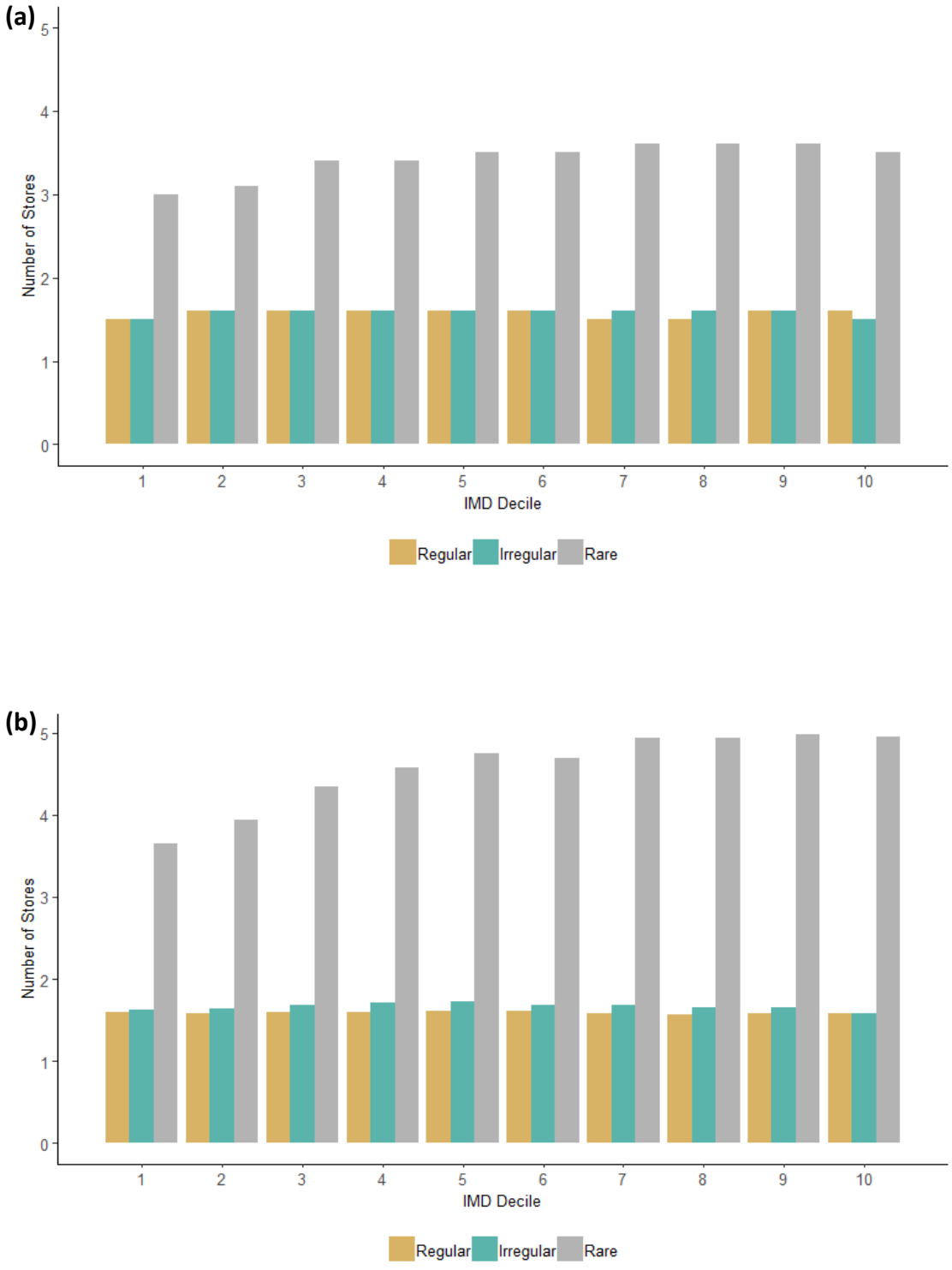


Figure 7.23 Average number of stores allocated as regular, irregular and rare per IMD Decile for a) the older and b) the working age populations

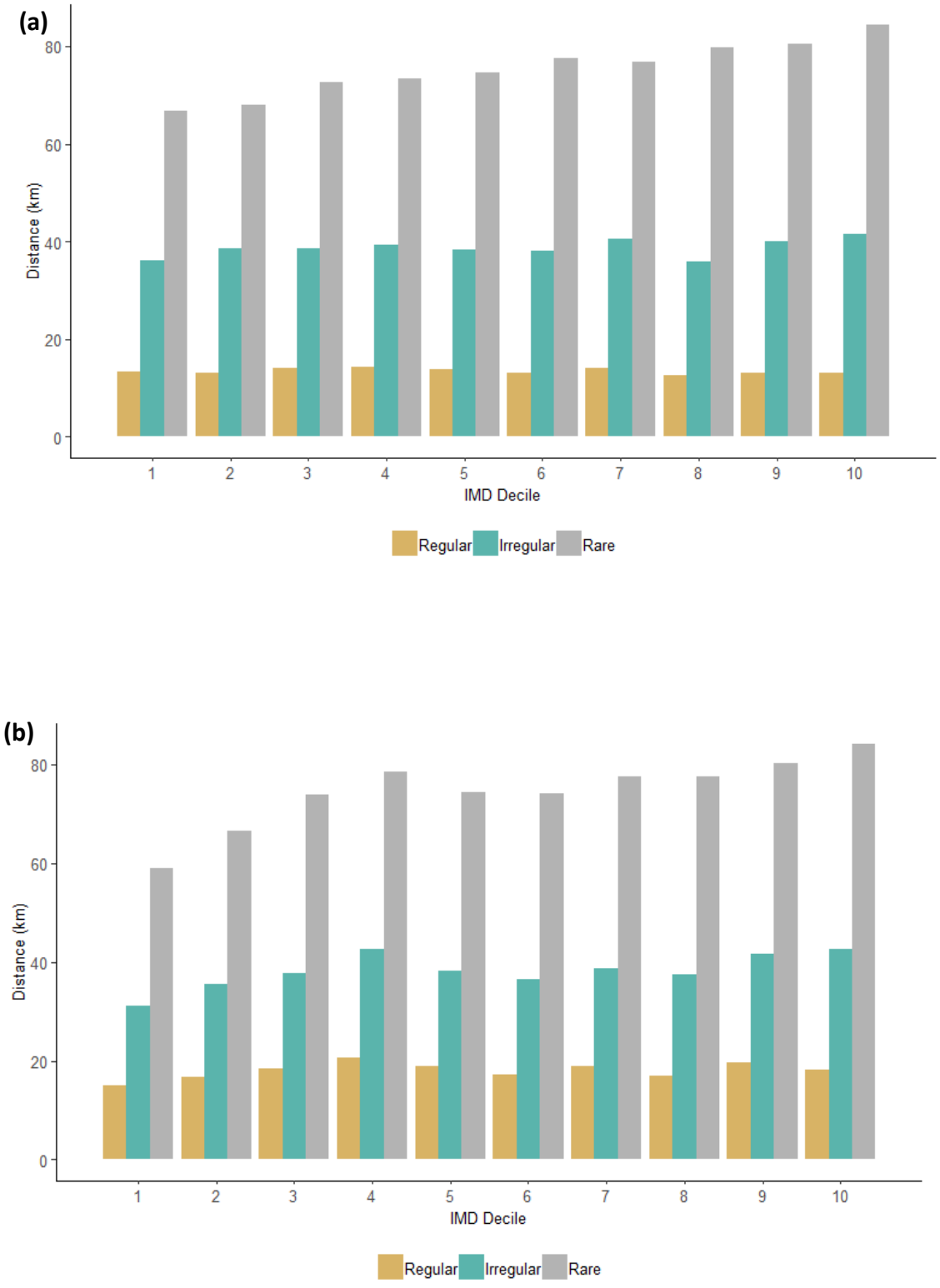


Figure 7.24 Average distance to regular, irregular and rare stores per IMD Decile for a) the older and b) the working age populations

Figures 7.23 and 7.24 suggest that there were differences in both the number of regularly, irregularly and rarely visited stores and the distance travelled to these stores between populations residing in different IMD deciles. In terms of the number of stores visited, the relationship was mostly evident when it came to rarely visited stores. For the working age population, those residing in the least deprived IMD deciles tended to have a higher number of rarely visited stores. This relationship, whilst still evident, was less pronounced for the older population; perhaps due to the lower average number of stores in their networks. There was little difference evident between the number of regularly and irregularly visited stores for each population.

In terms of the distance travelled to regularly, irregularly and rarely visited stores, there was less of a clear relationship evident. For the older population, the average distance travelled to rarely visited stores increased as deprivation decreased. Similarly to the number of stores visited, this relationship was not evident for regularly and irregularly visited stores. For the working age population, there was evidence that the average distance travelled for both irregularly and rarely visited stores increased as deprivation decreased, with a peaks at Deciles 10 and 4. This peak for those residing in Decile 4 may be attributed to the store network in the study area, with Figure 7.13 showing that few 'Community', 'Convenience (HS)' and 'Destination (EOT)' stores were located in LSOAs of these IMD deciles.

Although the size of store networks, both in terms of the number of stores and the distance to these stores showed correlations with both age group and IMD, the store network in the study area was also likely to have influenced these trends. A cardholder living within close proximity to an HSR store may be more likely to regularly visit that store, whereas those cardholders that live further from stores may have a wider network of stores visited regularly. Comparing these results to the IMD distributions shown in 7.13, HSR stores tended to mainly be located in areas of Deciles 1 and 2. This may explain the lack of relationship between distance and regularly visited stores. Rarely visited and irregular stores may be less influenced by the store network in the study area, and therefore be more indicative of how mobile a cardholder is.

7.2.3.4 Temporal activity patterns

In addition to the types of stores visited by different population groups, the temporal distributions of transactions made by populations can tell us when these groups are most active and how activity across store types, and therefore throughout the study area, differs. This can tell us when there is likely to be demand for services from different populations groups and can also give insight into store function. For example, stores that are most visited during lunch breaks and those that are visited most often at weekends may provide different services.

This section explores the temporal activity patterns of the older and working age populations. Firstly, the general daily and monthly activity patterns are presented and discussed. Secondly, the weekday and weekend activity patterns at each store type are explored; discussing how these differ between the working age and older populations and what this can tell us about demand for services.

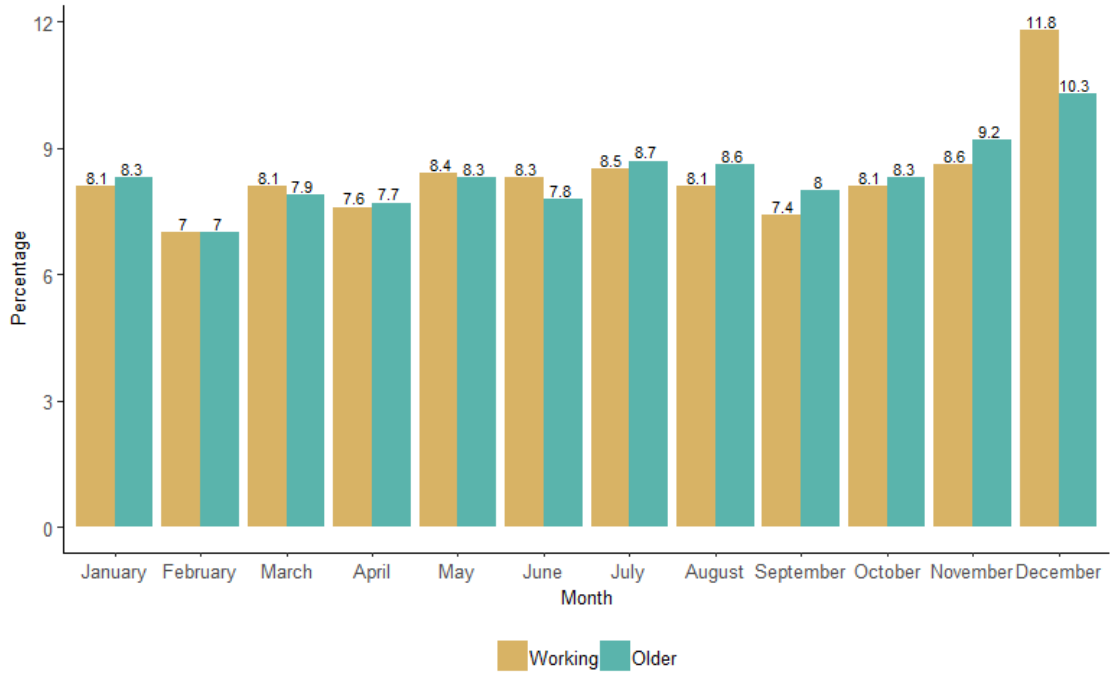


Figure 7.25 Average monthly transaction distributions for the working age and older populations

Figure 7.25 shows the percentage of total transactions made by the working age and older populations that were conducted during each month of the year. The distribution of transactions across months showed similar patterns between populations. Transaction frequencies tended to be lowest in February and September; reflecting post-Christmas and post-summer holidays slumps (ONS 2020). For both populations, transaction frequency peaked in December; likely due to increased shopping around Christmas, as well as the potential increase in healthcare products purchased which comes with an increase in colds and flus in the winter months.

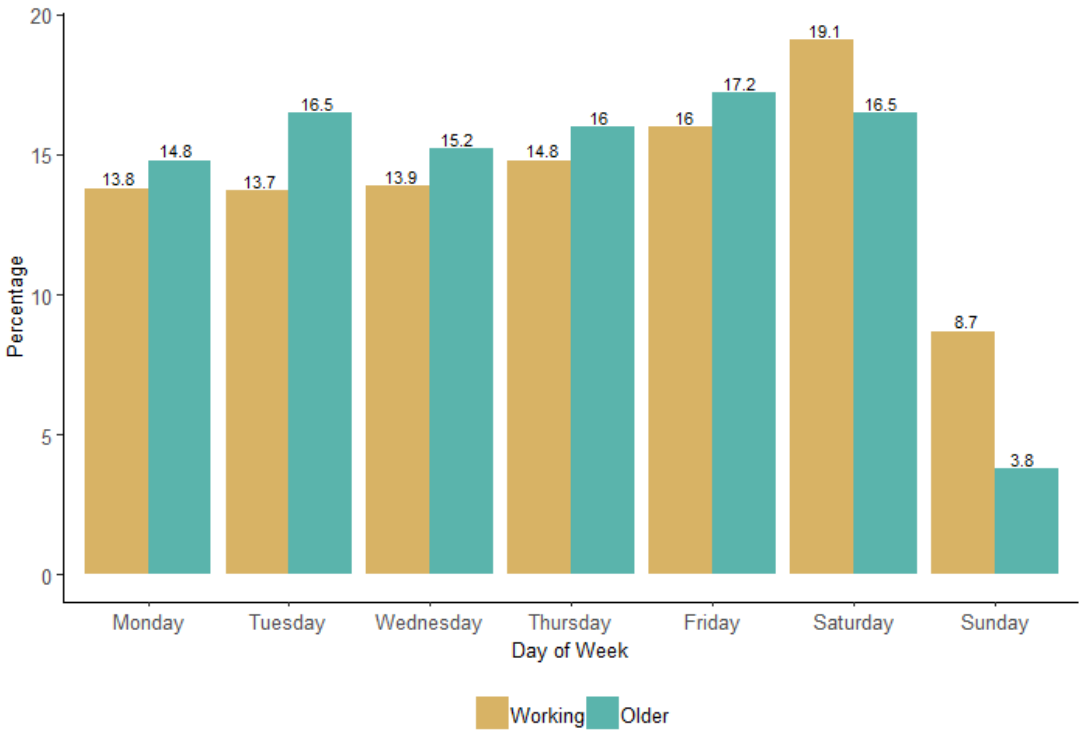


Figure 7.26 Average daily transaction distributions for the working age and older populations

Figure 7.26 shows the percentage of total transactions made by the working age and older populations that were conducted during each day of the week. The distribution of transactions across days of the week showed that the working age population made a higher proportion of transactions on weekends than the older population. This likely reflects the work week, with typical Monday to Friday working days meaning that a large number of shopping trips are likely to take place during weekends. Despite this, high proportions of transactions were still recorded throughout the week; likely indicative of the lunch and snacking options, as well as the essential items that HSR stores provide. The older population exhibited similar transaction levels throughout the week, with the exception of Sundays. This is likely a reflection of both reduced opening hours during Sundays, as well as limited public transport services which may restrict accessibility for those cardholders that do not have access to private transportation.

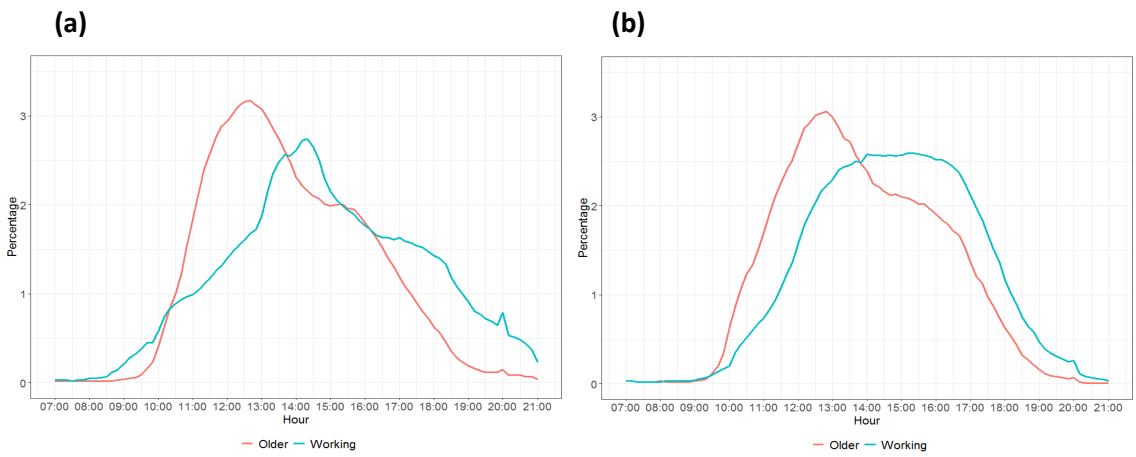


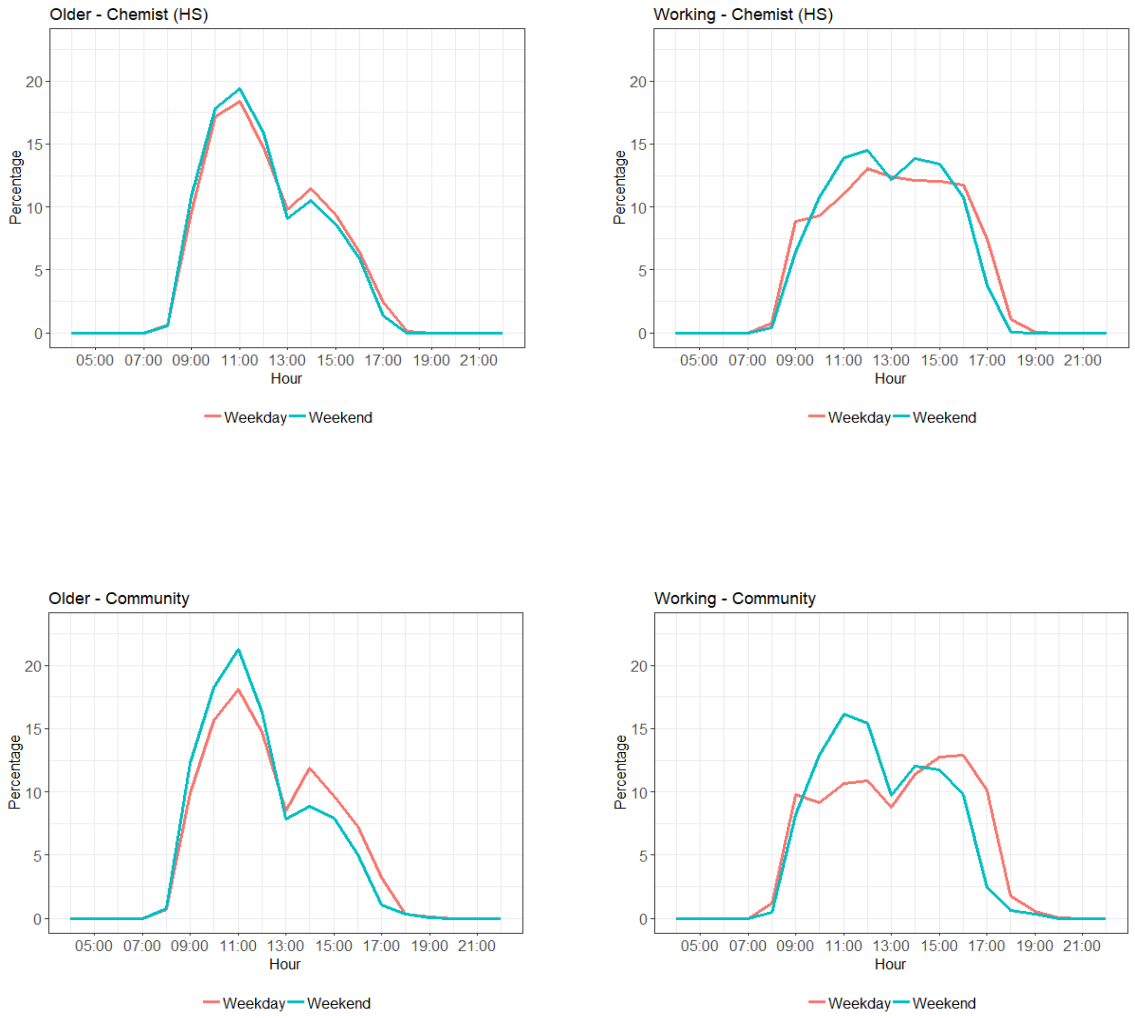
Figure 7.27 Aggregate 10-minute temporal profiles for a) weekdays and b) weekends

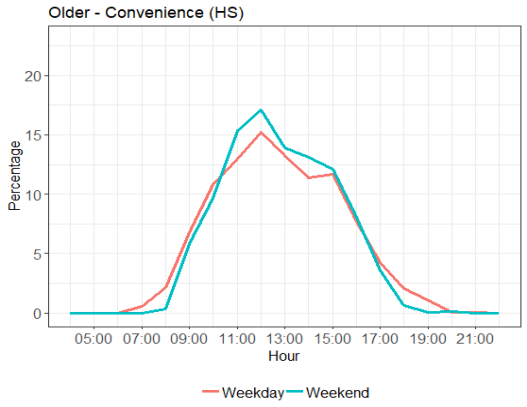
Figure 7.27 shows the aggregate temporal profiles of the older and working age populations for weekdays and weekends at 10-minute intervals, i.e. the percentage of total transactions for each population conducted during these 10-minute intervals. These can give an indication of the general temporal activity patterns of these populations. The temporal profiles of the older population showed little difference in the distribution of transactions on weekdays and weekends, with activity peaking at around 11am – 12pm and again at around 3pm, as well as showing a low number of transactions around peak times. These profiles suggested that the aggregate temporal activity of the older population differed significantly from the working age population. Public transport services are often most regular around peak times in the morning and afternoon; however, these may not be the times at which the older population have the highest demand for services. These profiles can therefore provide insights into when the older population are likely to be 'out and about' and therefore demand for services by this population group are likely to be at its highest.

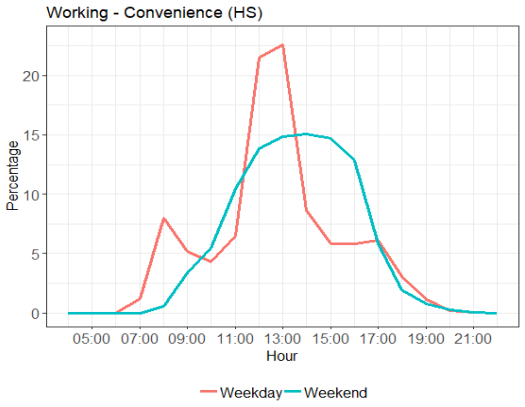
In contrast, the working age population exhibited distinct weekday and weekend temporal profiles. The weekday profiles showed a significant peak in activity at around 1pm, in contrast to a greater distribution of activity throughout the day on weekends. This weekday peak appeared to be indicative of a typical lunch hour and may be due to cardholders purchasing food and drinks during this time. Slight peaks at around 9am – 10am and 4pm – 6pm were also evident, again likely due to the beginning and end of a typical working day.

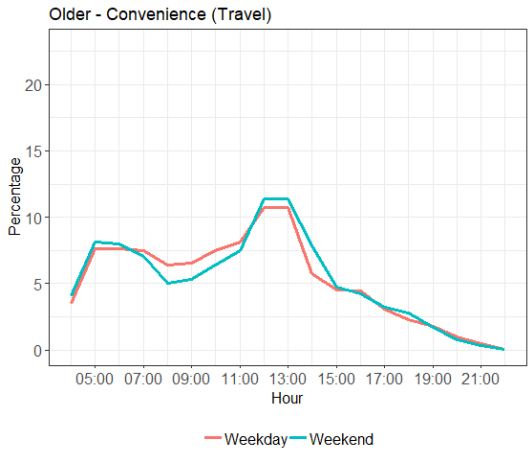
7.2.3.4.1 Store type

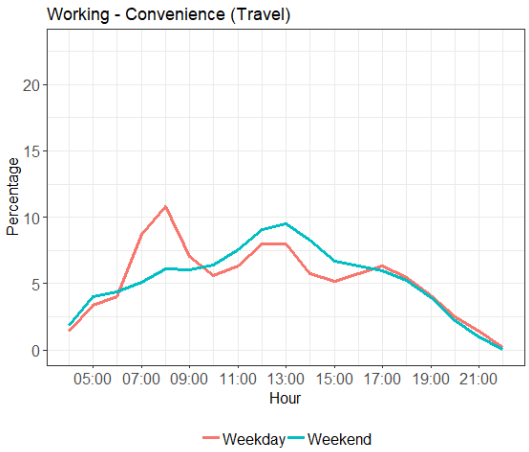
Distinguishing between store types gives insight into the demand for different stores and the function of these stores, for example if they are primarily used for shopping during lunch breaks or more likely to be visited during weekends. These temporal profiles were constructed as the percentage of total weekday and weekend transactions that were made during each 1-hour interval. Figures 7.28 shows the average temporal store visiting patterns of the working age and older populations for each store type on weekdays and weekends.

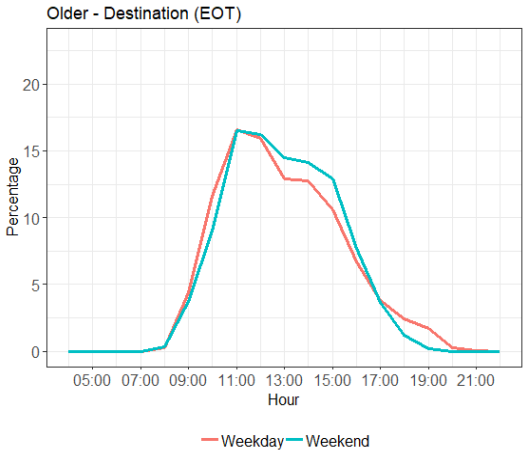


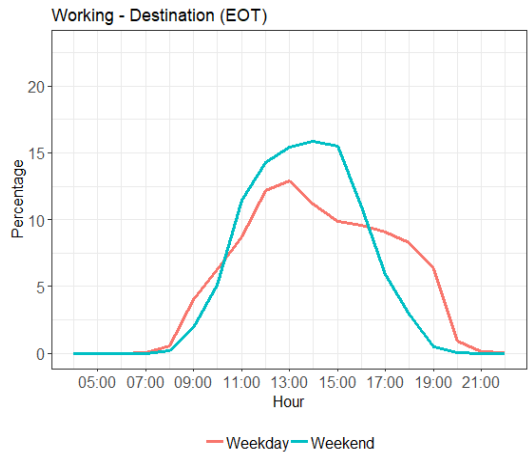

Older - Convenience (HS)
Percentage
0
5
10
15
20
05:00 07:00 09:00 11:00 13:00 15:00 17:00 19:00 21:00
Hour
Weekday
Weekend


Working - Convenience (HS)
Percentage
0
5
10
15
20
05:00 07:00 09:00 11:00 13:00 15:00 17:00 19:00 21:00
Hour
Weekday
Weekend


Older - Convenience (Travel)
Percentage
0
5
10
15
20
05:00 07:00 09:00 11:00 13:00 15:00 17:00 19:00 21:00
Hour
Weekday
Weekend


Working - Convenience (Travel)
Percentage
0
5
10
15
20
05:00 07:00 09:00 11:00 13:00 15:00 17:00 19:00 21:00
Hour
Weekday
Weekend


Older - Destination (EOT)
Percentage
0
5
10
15
20
05:00 07:00 09:00 11:00 13:00 15:00 17:00 19:00 21:00
Hour
Weekday
Weekend


Working - Destination (EOT)
Percentage
0
5
10
15
20
05:00 07:00 09:00 11:00 13:00 15:00 17:00 19:00 21:00
Hour
Weekday
Weekend

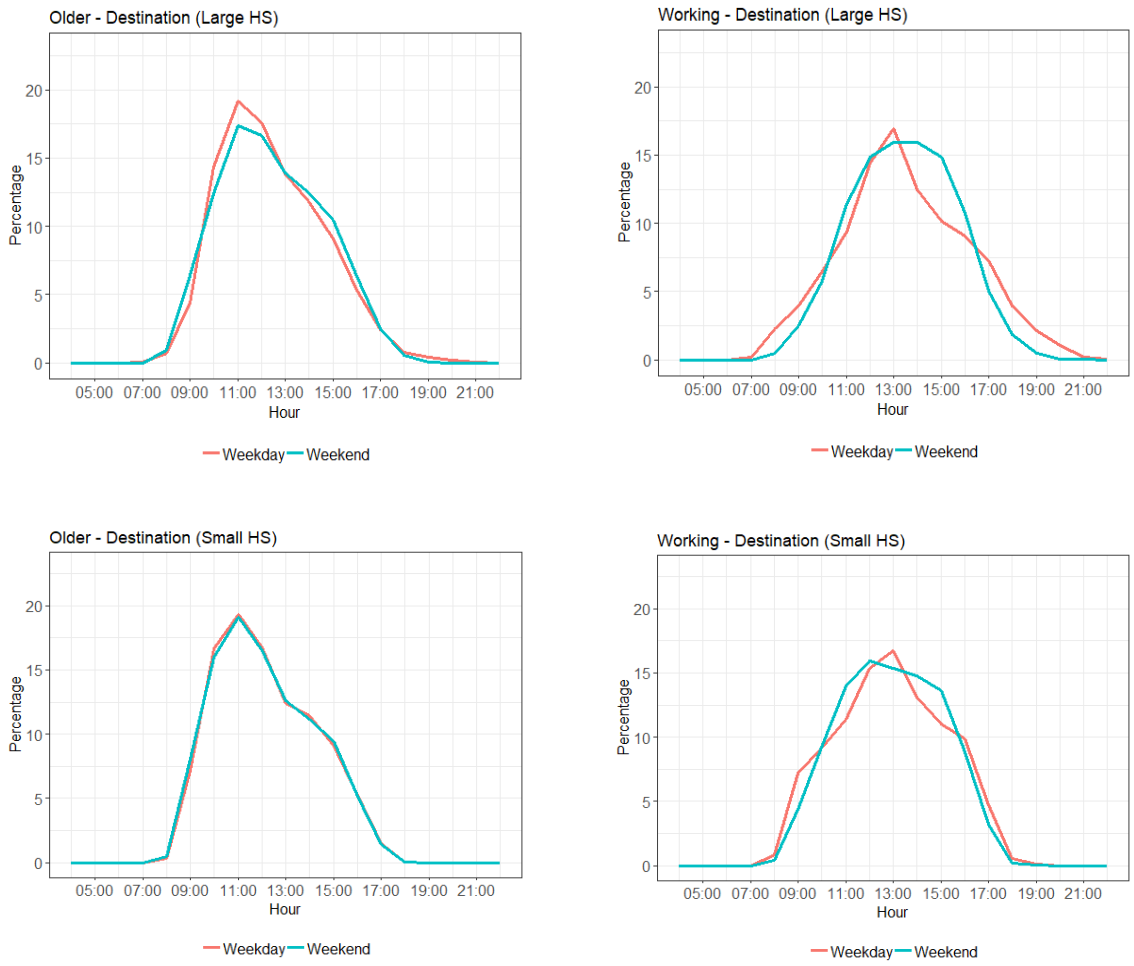


Figure 7.28 Aggregate 1-hour weekday and weekend temporal profiles of the older and working age populations for each store type

The average temporal profiles between store types and populations showed differences in how stores were used and the activity patterns of population groups. The temporal profiles for the older population showed much less variation between store types and between weekends and weekdays than the working age population, reflecting the fact that this population was less likely to be constrained by working hours. Still, differences between the activity of the older population at different store types were evident. 'Chemist (HS)' and 'Community' store types showed the most significant peaks at 11am and 2pm. All other store types, with the exception of 'Convenience (Travel)' stores, displayed a less significant peak at 2pm. 'Convenience (Travel)' stores showed a peak around the morning rush hour and typical lunch hours. This reflection of typical working hours may be indicative that those older cardholders that transacted at 'Convenience (Travel)' store types were more likely to still be in work and therefore transacting at these stores during commuting trips and during lunch breaks. Very little variation was found

between weekend and weekday temporal patterns for the older population, which was likely a reflection of a large proportion of cardholders who have retired and therefore were not constrained by the typical working week.

In contrast, the temporal activity patterns for the working age population differed significantly between store types and between weekdays and weekends. The store types that showed the most significant variations were both the 'Convenience (HS)' and 'Convenience (Travel)' stores. These stores showed peaks around the morning and evening rush hour periods and during typical lunch times on weekdays. These stores, therefore, were likely used by commuting populations. On weekends, the activity at these store types peaked around 1pm; highlighting the difference in the activity, and therefore demands, of the working age population on weekdays and weekends.

The differences between the aggregate temporal activity patterns of these populations highlighted the different mobility needs and demands of groups of the population. With current transport services often tailored towards meeting the needs of commuting populations (Wang 2007), understanding the demands of different population groups can improve the effectiveness of the transport network for meeting mobility needs. Non-transport related data, such as the HSR data used in this study, have the potential to provide insights into demand beyond the scope of the transport network. Public transport users are constrained by the routes and scheduling of the current transport system, and therefore non-transport related data can provide valuable insight into whether the current network is meeting the demands of different populations. As discussed in Section 2.2.3, much research and policy are focused around encouraging a shift from private transportation to public transportation for both environmental and social reasons. For this to be achieved, public transport services must be able to meet the demands and mobility needs of different population groups. Although this section focused on aggregate spatial and temporal patterns, future research utilising consumer data has the potential to provide detailed, granular insights that can expand the understanding of the demands of these populations and ensure future public transport services meet their mobility needs.

7.2.4 Product data

The product data contained unique codes and product names over 6 hierarchical levels. Level 1 represented the most aggregate level, containing just two categories: retail and pharmacy. Level 6 was the most detailed level of the hierarchy, describing 329 product categories. The transaction data contained product data at level 6, meaning data could be linked at the most detailed level and then aggregated to higher levels for analysis where necessary. Table 7.4

provides a summary of this product hierarchy, with examples of the product categories included in each level.

Table 7.4 Example of the HSR product hierarchy structure

Level	No. of categories	Example categories
1	2	Retail Pharmacy
2	7	Healthcare Beauty Pharmacy dispensing Pharmacy services
3	160	Baby Electrical NHS dispensing Private dispensing
4	221	Baby consumables Lunch and snacking Winter medicines Summer medicines
5	287	Skincare Mouth care Children's wear Tissues
6	329	Deodorants Electrical hair Shampoo and conditioner Premium cosmetics

Product information was utilised throughout this chapter to explore consumption patterns and differences between population groups. For ease of interpretation and to maintain the anonymity of both customers and the HSR, the general product consumption patterns of cardholders shown in Figure 7.29 were conducted at Level 2 of the hierarchy. As loyalty cards were rarely used for pharmacy services, with just 0.02% of all recorded transactions, these were disregarded for future analyses. Level 2 of the product hierarchy therefore consisted of 5 categories: 'toiletries', 'retail healthcare', 'beauty', 'lifestyle' and 'other'. Figure 7.29 shows the percentage of products purchased by older and working age cardholders of each category.

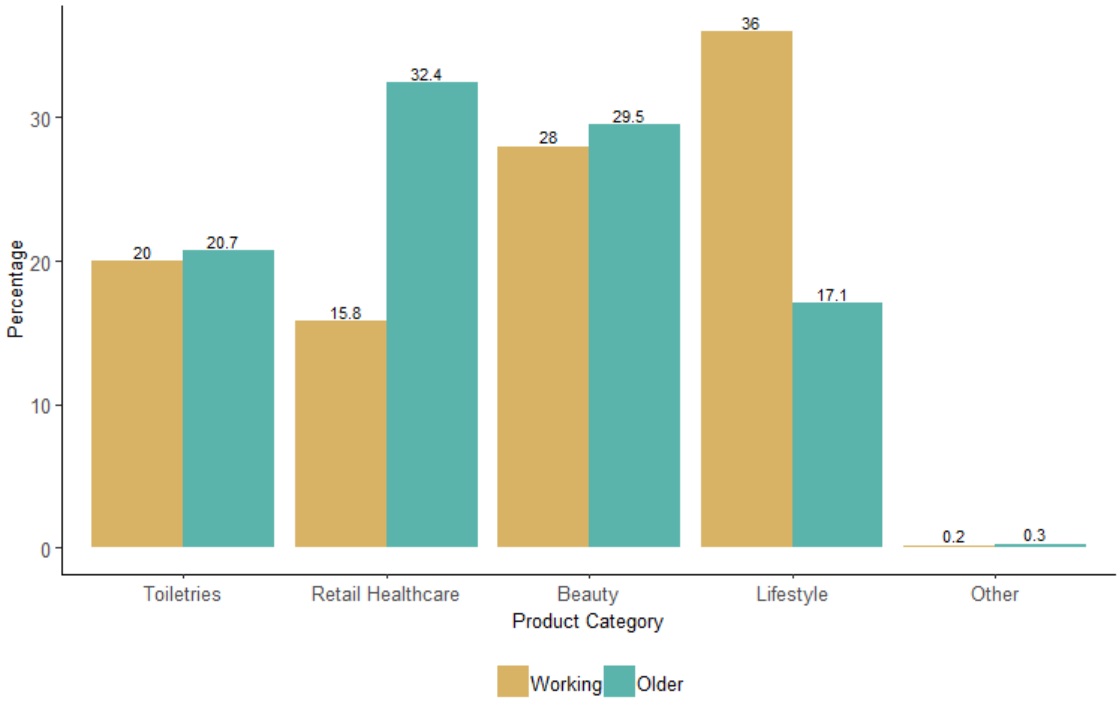


Figure 7.29 Level 2 product distributions for the working age and older populations

There were clear differences evident in the product distributions of both populations, with the older population purchasing a significantly higher proportion of 'retail healthcare' products and a lower proportion of 'lifestyle' products than the working age population. Lifestyle products included products such as baby consumables, lunch and snacking, and gifts. These may be expected to be more commonly purchased by those buying for families or those that are in work purchasing lunch and snacking products during a lunch break. The higher proportion of healthcare products by the older population was unsurprising, with health generally declining with age.

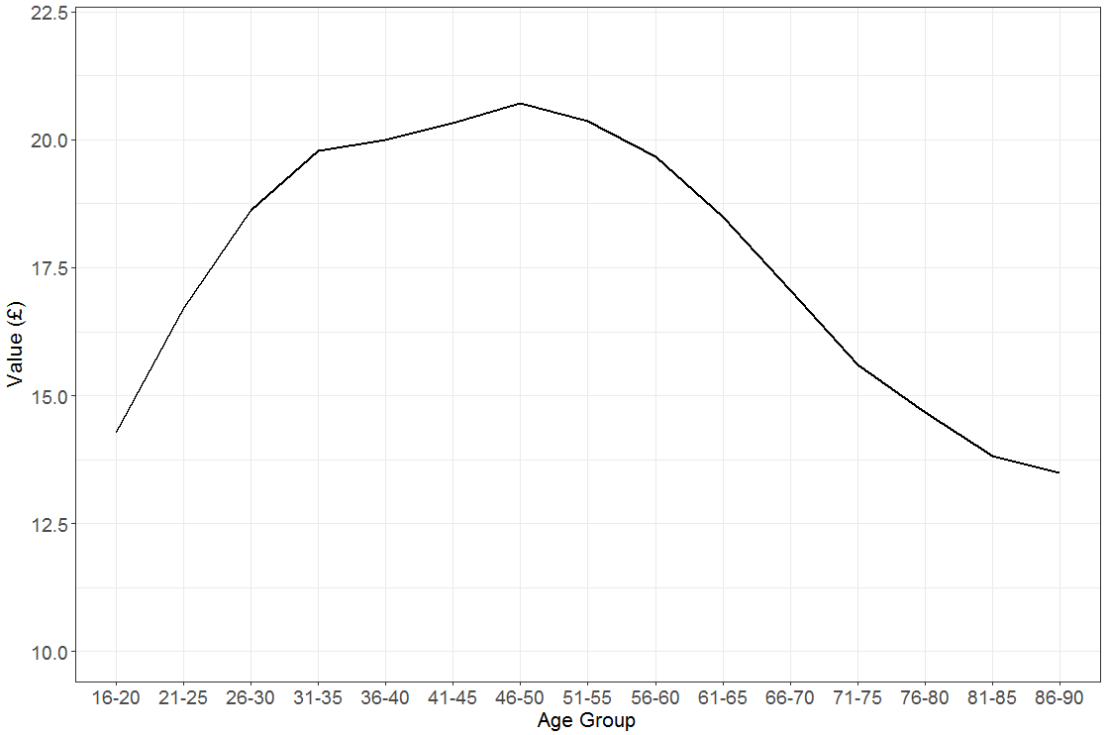


Figure 7.30 Average basket value (£) by age (5-year age bands)

Average basket value correlated with age; peaking for the 46-50 age group before declining to its lowest for the 86-90 age group. This corresponds with those population groups that are most financially stable, and therefore suggested that retail spending recorded by loyalty cards may be a good indicator of economic wellbeing. This higher level of spending for the 30-50 population may also be indicative of single loyalty cards used for purchasing for entire families, with this population group more likely to have a number of dependants and therefore may be more likely to be purchasing products for multiple individuals.

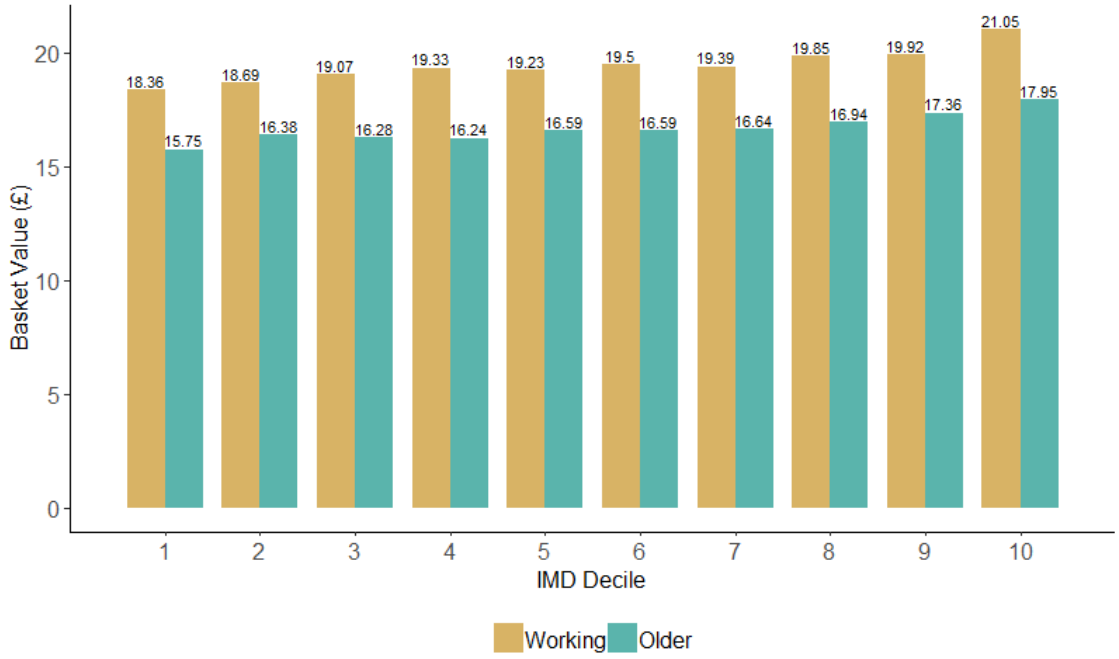


Figure 7.31 Average basket value by IMD

Figure 7.31 shows that average basket value was lower for the older population than the working age population, and for both populations this increased as relative deprivation decreased. For the working age population, average basket value for those residing in LSOAs of Decile 10 was £2.69 more than those residing in the most deprived LSOAs. This difference was less for the older population, with an average difference of £2.20 per transaction. Comparing this with Figure 7.30 therefore suggested that economic wellbeing may not have such a strong influence on the types of products purchased but may instead dictate the amount of money spent on these products.

7.2.4.1 Classification of 'essential' products

In addition to general product distributions, this section also aimed to understand differences in the purchasing of essential and discretionary products. Discretionary purchasing can be defined as "consumer spending on things they want to buy rather than what they need" (Danziger 2004) and is an important indicator of economic wellbeing and deprivation (Sarial-Abi et al. 2016). Prior research has primarily been descriptive and largely limited in empirical testing of who is more or less likely to make discretionary purchases. The loyalty card data allow for detailed analysis of this.

As there is no standard classification of which products are considered essential or discretionary, the classification developed for this analysis was primarily subjective but also relied on guidance from the government's 'Discretionary Housing Payment' scheme. This payment is an additional payment that councils can make to an individual if it is determined that extra help is needed to meet housing costs and cover essential expenses (Department for Work and Pensions 2016). Some of the products deemed to be essential by this scheme include toiletries, baby items and NHS health expenses. This classification was conducted at Level 3 of the product hierarchy, consisting of 160 product categories. Examples of products classified as essential and concessionary are shown in Table 7.5.

Table 7.5 Example products classified as 'essential' and 'discretionary'

Essential	**Discretionary**
Walk in prescriptions	Premium cosmetics
Summer medicines	Fashion brands
Heart health and dental	Fragrance
Baby consumables	Accessories
Screening and testing	Photography services
Personal care	Gifting

Essential products mainly consisted of healthcare and medicines, baby products and personal hygiene products, whereas discretionary products primarily included beauty and cosmetic products, as well as snacks, drinks and gifts. Figure 7.32 shows the percentage of products purchased that were classified as either essential or discretionary in 5-year age bands.

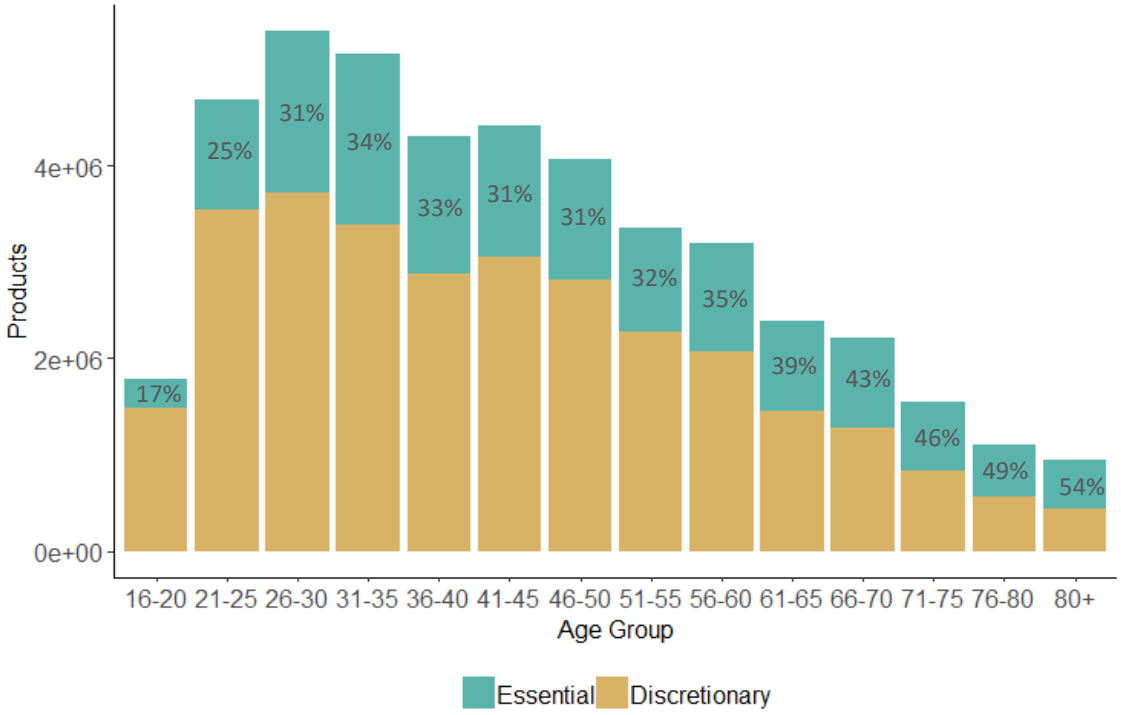


Figure 7.32 Number of essential and discretionary purchases by age group (percentage of essential products purchased is labelled)

The proportion of products purchased that were classified as 'essential' increased with age, with products purchased by those in the highest age group (80+) 3.18 times more likely to be essential than those in the youngest age group (16-20). This increase was likely due to increased purchasing of healthcare products by the older population and a decrease in disposable income meaning fewer discretionary products are purchased. The increase in essential purchasing for those in the 31-35 age group was likely due to the inclusion of baby items as essential products and may have also been due to one cardholder purchasing essential products for dependent family members, for example toiletries or medicine.

Comparing the number of essential and discretionary products purchased by residents of each IMD decile can give insight into how essential purchasing relates to economic wellbeing. Figures 7.33 and 7.34 show the total number of essential and discretionary products purchased by cardholders of each IMD decile, with the percentage of these products that were classified as essential labelled.

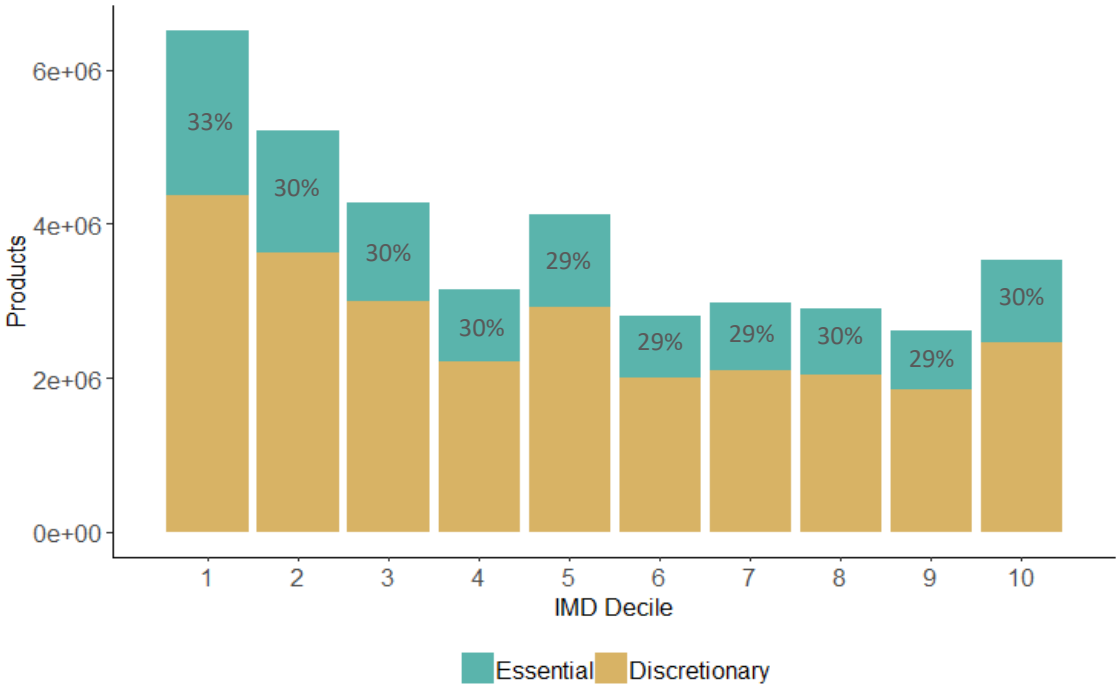


Figure 7.33 Number of essential and discretionary products purchased by the working age population per IMD decile (percentage of essential products purchased is labelled)

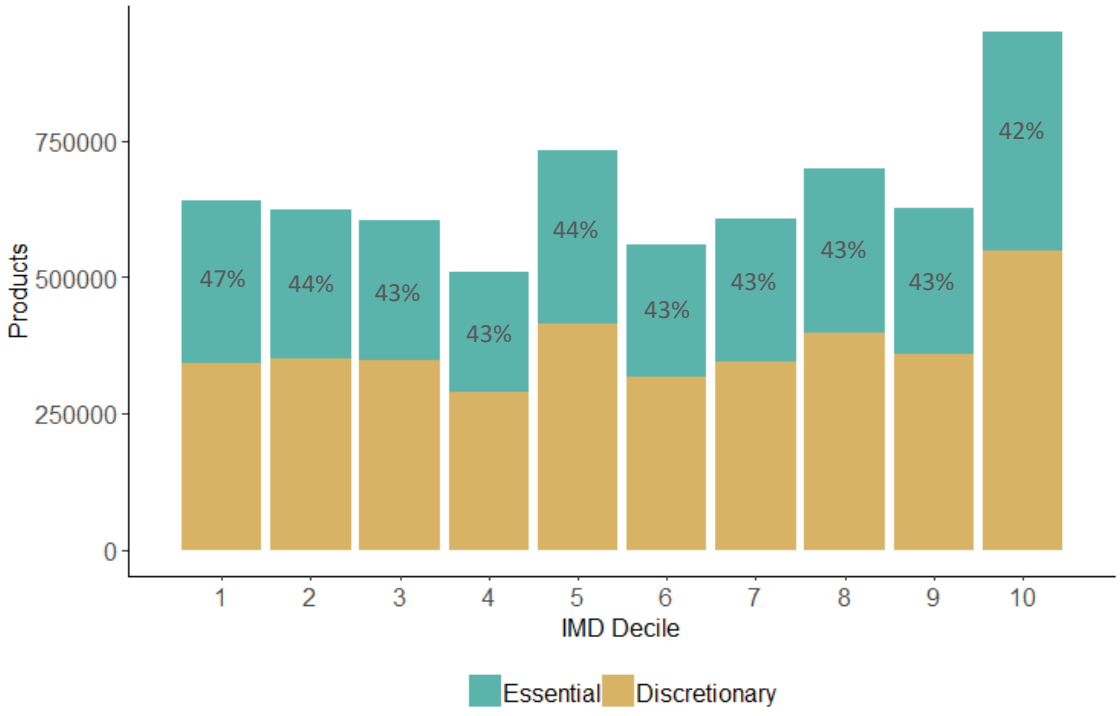


Figure 7.34 Number of essential and discretionary products purchased by the older population per IMD decile (percentage of essential products purchased is labelled)

From Figures 7.33 and 7.34 it was clear that the numbers of products purchased by cardholders residing in each IMD decile differed significantly. For the working age population, the total number of products purchased followed a similar pattern to the population distributions shown in Figure 7.3, with both significantly more products purchased and cardholders residing in LSOAs of Decile 1. This would be expected, with a larger number of cardholders in an IMD decile coinciding with a larger number of products purchased. In contrast, Figure 7.34 showed the opposite trend to that shown in Figure 7.3. Despite the largest proportion of cardholders residing in LSOAs of Decile 1, the greatest number of products purchased were by cardholders in Decile 10. This, again, highlights the economic wellbeing and greater disposable income of less deprived populations. Future analyses could normalise the number of products purchased by the number of cardholders residing within each IMD decile to further explore these distributions.

It may have been expected that those residing in more deprived areas would purchase a higher proportion of essential products due to lower levels of disposable income. However, Figures 7.33 and 7.34 show that this was not the case. Essential purchasing consistently made up between 29% and 33% of total purchases on average across all IMD deciles for the working age population, and between 42% and 47% for the older population. This may reflect differences in deprivation within LSOAs, with IMD recorded at LSOA level rather than individual level. Therefore, levels of deprivation likely differ between individuals in these areas. It may also reflect the fact that, as mentioned in Section 7.2.4, economic wellbeing appears to have a larger influence on the value of products purchased rather than the types of products purchased. It could therefore be expected that populations of different IMD deciles may purchase a similar number of essential products, but the money spent on these products may be higher for those that are less deprived.

7.3 Discussion and Conclusions

The aims of this chapter were two-fold; to firstly give an overview of the HSR loyalty card data, including their representativeness and potential bias, and secondly, to explore the retail behaviours of cardholders with a particular focus on comparing the older and working age populations. This chapter also demonstrated the potential of loyalty card transaction data as an indicator of activity, mobility and wellbeing; concepts which are explored further in Chapter 8.

7.3.1 The potential of loyalty card data

Results presented in Section 7.2.1 suggested that while loyalty card transactions provide large, longitudinal data, in order for this to be utilised for academic insight the representativeness of the data must be considered. The data utilised in this thesis appeared to be representative of the female population in the study area in terms of the age groups represented however less so in terms of deprivation. These data also contained a significantly lower proportion of male cardholders than the total study area population. The representativeness of these loyalty card data therefore limited the insights that could be deduced into the retail behaviours of the male population. This meant that the demographics and socioeconomics of cardholders had to be included in subsequent analyses to enable results to be generalised beyond the loyalty card population.

The insights into store networks and visiting patterns demonstrated the potential of retail transaction data as an indicator of activity and mobility. Spatial and temporal store visiting patterns can provide insight into who is 'out and about' and active throughout the study area, as well as differences in the demand for both retail and transport services. As the older population's needs, both in terms of retail and transport differ from those of working age, insight into the differences in demand can allow for transport authorities and operators to ensure that the needs of different population groups are being met. The spatial and temporal references contained within consumer data can also be important indicators of activity (Graham and Shelton 2013). Current research argues that indicators of activity patterns on a daily, weekly, seasonal and long-term basis are needed to be used alongside residential locations to gain a more comprehensive view of social structure (Longley 2017). These analyses therefore added to the growing research into how consumer data can be utilised to further the understanding of activity patterns and be used as indicators of the spatial organisation of society.

Exploration of the store patronage uncovered clear differences between both the older and working age populations and between IMD deciles. The differences between the patronage of each store type shown in Figure 7.17 suggested that the most deprived deciles of the older population were constrained in their store visiting patterns; with low numbers of visits to less 'local' stores, including 'Convenience (Travel)', 'Destination (EOT)' and 'Destination (Large HS)' stores. These insights were also reflected in the results of the spatial store visiting patterns and the analysis of cardholders' store networks. Many previous studies have looked at how deprivation influences retail behaviour and the accessibility of stores (Williams and Hubbard 2011; Adeniyi et al. 2020), however little research has utilised big consumer data. These consumer data can allow for a more comprehensive view of the differences in retail behaviour and store visiting by different population groups.

The spatial store visiting patterns presented in Section 7.2.3.3 provided insight into both mobility and demand. Comparisons of the regular store visiting patterns of the older and working age populations presented in Figure 7.21 showed differences between populations, in particular between edge-of-town and small high street stores. These edge-of-town stores can often lack regular public transport services (Williams and Hubbard 2011) which may influence the propensity of the older population to regularly visit these stores. The older population were also found to be more likely to regularly visit 'Chemist' and 'Community' store types, and less likely to regularly visit 'Convenience' store types, which corresponds with the purchasing behaviours presented in Figure 7.29.

As well as differences in the types of stores visited by populations, there were clear differences in the store networks of cardholders. The size of an individual's store network, both in terms of the number of stores and the distance travelled to these stores can provide insights into how mobile they are. Figures 7.21 to 7.24 suggested that age, rather than IMD, had a significant relationship with the size of a cardholder's store network, however only for rarely visited stores. The high number of rarely visited stores by the younger population is representative of the more transient nature of these groups in comparison to the more settled older populations (Sheller 2011). Age also had a clear relationship with the distance travelled to stores, with a particular correlation with regularly visited stores. For older age groups, the average distance between a cardholder's registered LSOA and their regularly visited stores was less than half that of the youngest age groups. Again, this may reflect the store network in the study area but may also be indicative of the working age population regularly visiting stores that may be located near a work location rather than the registered home address.

IMD had less of a clear relationship with patterns of store visiting than age, with little difference evident in both the number of stores and the distance travelled to these stores between IMD deciles. This is likely a reflection of the store network, with Figure 7.13 showing that the majority of stores in the study area are located in and around the most deprived LSOAs, which may explain the marginally lower distances travelled to stores for the most deprived deciles shown in Figure 7.24. Previous research has found differences in the locations of retail areas between IMD deciles (Adeniyi 2020) and these results further previous insights by looking at how this may influence retail behaviour. The idea that the store network has a strong influence on the store visiting patterns of cardholders is explored further in Chapter 8.

The temporal transaction patterns presented in Section 7.2.3.4 suggests that those of working age are more constrained by working hours, with the majority of transactions taking place during typical lunch breaks and afternoon peak hours during weekdays. The older population's propensity to undertake the majority of transactions during the morning, after-peak hours

suggests that demand for both transport services and retail services likely differ significantly between these groups of the population. With transport services often designed around meeting the needs of commuting populations, understanding differences in demand can aid transport authorities in increasing the utility of the transport network.

In addition to store visiting patterns, the purchasing behaviours of cardholders provided insight into economic wellbeing. Economic activity is strongly related to life satisfaction, with household spending found to be more strongly related to how life satisfaction is rated than household income (ONS 2019). Insights into both basket value, discretionary spending and product purchasing can therefore be indicative of both economic wellbeing and life satisfaction. The higher spending on healthcare products by the older population was reflected in the higher proportion of essential products purchased by these age groups. This means that as well as lower overall spending by older population groups, these groups are also buying significantly fewer discretionary products, which may have an impact on wellbeing and life satisfaction.

7.3.2 Links to smart card insights

Generally, insights into mobility have relied upon transport-related data and specifically designed surveys and questionnaires. To an extent, this chapter has demonstrated how other sources of data can be indicative of mobility and how retail activity can provide insights into which groups of the population are active in the study area, both spatially and temporally.

Insights generated from the smart card data in Chapter 5 suggested that the decrease that has been recorded in concessionary bus usage was primarily due to those that resided in and around urban centres decreasing their usage of bus services. These urban areas also tended to be more deprived, with significant proportions of smart card holders exhibiting decreasing bus usage residing in LSOAs of Deciles 1 and 2. The comparisons between spatial store visiting patterns and IMD in this chapter suggested that those that resided in more deprived areas were likely to have smaller store visiting networks, in particular in terms of rarely visited stores, and were more likely to transact at ‘local’ store types. As both the number of stores and the distance travelled to those stores are likely highly influenced by the store network in the study area for regular store visiting patterns, in this instance differences in individual’s rarely visited store networks can provide greater insight into mobility. The smaller ‘rarely visited’ store networks for more deprived groups of the population could therefore indicate constrained mobility. Furthering the understanding of mobility within deprived populations can add to ongoing research into social segregation and exclusion (Lathia et al. 2012). Chapter 8 explores the relationship between mobility and both the store and transport network further.

The spatial and temporal transaction patterns also coincided with insights from Chapter 6. In terms of trip purpose, the majority of trips to HSR stores would be classed as shopping trips, with the exception of some trips to 'Chemist (Health centre)' store types which may be classed as personal business trips. The temporal transaction patterns shown in Figure 7.27, and the start and end times of inferred shopping trips shown in Figures 6.7 and 6.8 showed strong correlations. Similar profiles were found in both the smart card data and the loyalty card data; peaking around 10 am – 11am with another peak later in the afternoon. In terms of spatial store visiting patterns, the high number of transactions made at both large and small high street stores corresponded with the shopping hotspots shown in Figure 6.13, with the main activity areas concentrated around these high streets.

This chapter demonstrated the utility of novel consumer data to provide insight into activity and demand beyond transport-specific data and the public transport network. The similarities in the results produced by the smart card data and the HSR data are promising for the triangulation of multiple consumer datasets to explore social phenomena. If big consumer data are increasingly going to be used to substitute traditional data, the agglomeration of multiple consumer datasets will be necessary to ensure the validity and reliability of these results.

7.3.3 Future research potential

The utility of loyalty card data for exploring mobility and activity patterns is analysed further in Chapter 8, however the temporal element of these data is not explored. This chapter presented a preliminary analysis of the temporal dimension of the data, however future research could explore this concept in greater depth. Transaction volumes, and therefore retail activity, can be used as indicators of demand for both retail facilities and public transport services. Further research into the differences in temporal activity between population groups could therefore aid both transport providers and retailers in tailoring services towards those that are most likely to be using facilities at different times of the day and over longer temporal periods.

The data used in this thesis spans a period of 2.5 years, from April 2012 to September 2014. However, had data been available over a longer period there would have been potential to explore the impact of concessionary travel on retail behaviours. Identifying those that had become eligible for concessionary travel during the period for which loyalty card transactions were recorded would allow for changes in activity patterns once becoming eligible for concessionary travel to be detected. This could have provided a useful case study into the utility of concessionary smart cards for the older populations, and whether there is an impact on mobility levels once travels becomes free of charge. There has been much research conducted on the impacts of the ENCTS and how free bus use influences travel behaviour (Laverty and

Millet 2015; Mah and Mitra 2017). Future research utilising consumer data could further this by exploring the impacts of the ENCTS on retail activity and spatial travel behaviour. Due to both the limited study period, and the fact that only a year of birth was supplied by the HSR, this study was unable to explore this further.

Chapter 8 expands upon many of the concepts of this chapter, with a specific focus on the bus network and the relationship with retail activity. The results presented in this chapter provided a promising starting point for the triangulation of multiple sources of data for the exploration of mobility and activity patterns. The increase in the availability of novel, big datasets over recent years will only produce more opportunities to utilise diverse data to explore these concepts.

Chapter 8 – Activity patterns and the relationship with public transportation

8.1 Introduction

For transport authorities and operators, understanding how public transport services are used by different groups of the population, and the impact that access to these services can have on the mobility of individuals is imperative to be able to provide services that meet mobility needs. The activity patterns emanating from neighbourhoods can provide insight into these mobility needs. Accurate study of these activity patterns can help us to understand how they are shaped by the demographic and socioeconomic characteristics of residents, the availability of public transportation and the accessibility of key facilities. This chapter utilises loyalty card transactions as a proxy for activity; extracting both consumption patterns and spatial activity patterns from the HSR loyalty card data and then analysing the factors that impact these.

Having previously provided a comparison of the retail behaviours of the working age and older populations, this chapter focuses solely on the older population to allow for comparability to the patterns found in the smart card data. In a study into mobility and social participation, Farber and Paez (2009: 217) stated that "activity analyses require large quantities of disaggregated travel and activity behaviour data". The availability of both detailed public transport data and large retail transaction data therefore presents a unique opportunity for a 'real-life' case study into the relationship between public transport provision, accessibility and activity. With many of the HSR stores located in central urban areas and on local high streets, these can also represent the more general shopping and socialising opportunities that have been linked to wellbeing and social inclusion (Farber et al. 2011).

This study used HSR transaction behaviour as a proxy for activity, primarily focusing on the consumption and spatial store visiting patterns of cardholders. For this analysis, spatial store visiting patterns were used to describe a cardholder's 'activity space'. Activity spaces are a measure of spatial behaviour used to summarise mobility patterns (Hasanzadeh et al. 2019) and have been defined as "a set of geographically distributed locations which are physically contacted by individuals" (Reynolds and Horton 1971: 37). Activity spaces have been applied in a wide range of research, including transportation research, geography and public health. More recently, there has been a growing body of research on the links between the characteristics of activity spaces and wellbeing, with these being used as an alternative measure of social exclusion (Schonfelder and Axhausen 2003; Manaugh and El-Geneidy 2012).

Activity spaces have been so widely used in research as they go beyond residential neighbourhood to give insight into how people use and move about space (Leverentz 2020). Previous research has found that adults spend a high proportion of their physical activity time outside of their residential neighbourhood, and that activity spaces and residential buffers have poor spatial overlap (Holliday et al. 2017). Activity spaces can therefore give a better indication of the areas that populations visit and spend time in than residential information and have been used widely in health research due to their relationship with health behaviours and wellbeing (Kestens et al. 2018; Holliday et al. 2017). Using consumer data to understand the factors that impact the characteristics of activity spaces can therefore provide insight into the wellbeing and social inclusion of the older population in the study area. Additionally, it can help identify the factors that can constrain activity spaces and provide insight into the impact of public transport provision and the accessibility of key facilities.

In addition to activity spaces, consumption patterns can provide insight into the utilisation of both the public transport and HSR store networks. Consumption patterns, such as the types of products purchased and the value of these purchases, can be used to determine the types of trips being undertaken, such as recreational shopping trips or essential trips undertaken to purchase necessities. Comparing these patterns to public transport provision and the HSR store network can therefore be used the assess the current use of the transport network and to identify how public transport services are utilised by different groups of the population.

To start, this study analysed the bus timetable data provided by TfWM to calculate bus service provision and quantify the accessibility of HSR stores for each LSOA in the study area. In this context, ‘accessibility’ referred to the time that it would take to access HSR stores via the bus network. The concept of accessibility was discussed in Section 2.2.3.3 and has been widely linked to social inclusion and wellbeing; with an increase in the accessibility of key services and facilities related to increased social interaction and a decrease in social exclusion (SEU 2003). Spatial regression analyses were then conducted between these public transport variables and a number of retail activity and consumption variables, whilst also considering the socioeconomics and demographics of cardholders. These analyses aimed to quantify the relationship between the activity patterns emanating from neighbourhoods, the provision of bus services, and the accessibility of key facilities via public transportation. Understanding this influence through a large-scale, ‘real-life’ case study, made possible through the availability of the HSR data, can provide novel evidence of the influence of public transport on the mobility of the older population, in addition to an evaluation of the utility of the current transport system.

8.2 Method

This section presents the methods used to assess the impact of public transport provision and accessibility on the activity patterns of loyalty card holders. Firstly, the research design is presented; outlining and justifying the processes and analyses undertaken throughout this chapter. The subsequent sections then describe the methods used to undertake these analyses and detail how these methods were applied to the data used in this study.

8.2.1 Research design

Figure 8.1 outlines the research design of this chapter.

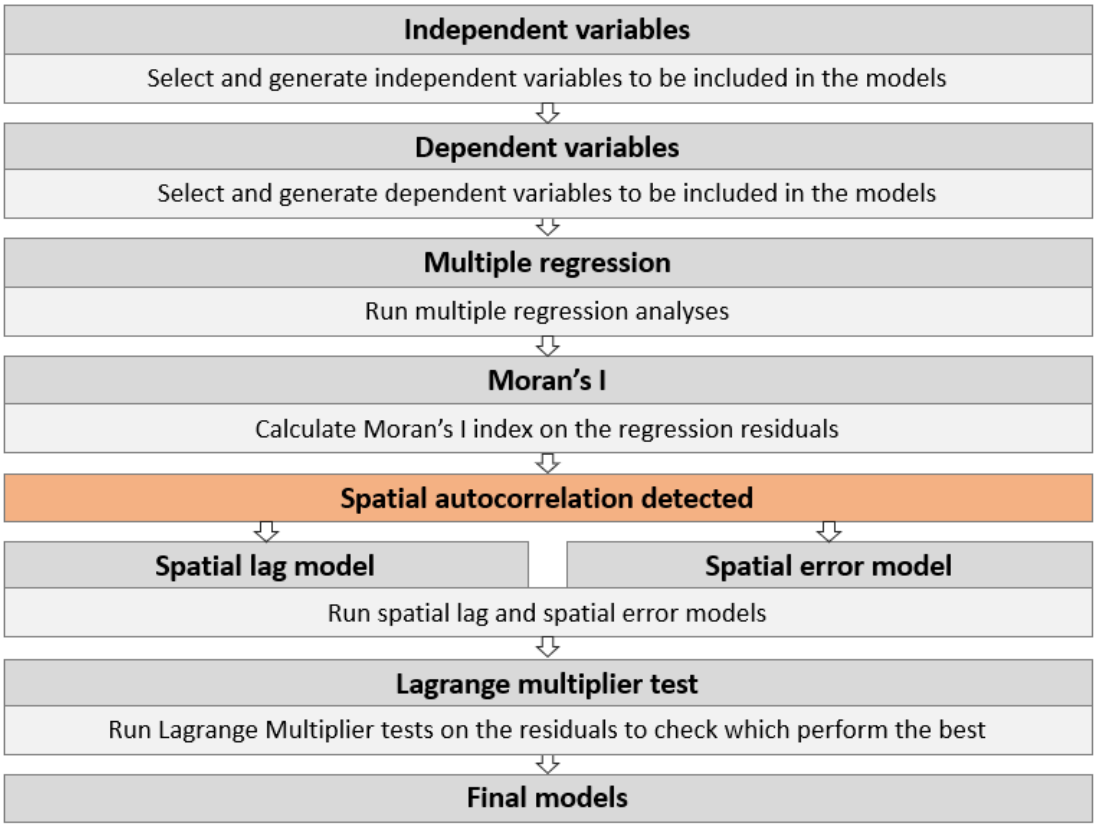


Figure 8.1 Research design

To assess the relationship between public transport provision, accessibility and the activity patterns of cardholders, this study employed regression analyses. Regression analysis is a statistical method that allows for the relationship between a set of independent variables and a dependent variable to be examined (Seber and Lee 2012). As this study aimed to determine the

linear relationship between continuous variables, linear regression models were implemented. Linear regression is a linear approach to modelling the relationship between a dependent variable and one or more explanatory variables (Montgomery et al. 2012). For this study, where the impact that a number of factors have on the activity patterns of cardholders was analysed, multiple regression models were implemented as these allowed for more than one independent variable to be included.

Multiple regression relies on several assumptions. The first of these is that there must be a linear relationship between the dependent and independent variables and that the residuals are normally distributed. These models also assume that there is no multicollinearity between independent variables, i.e. they are not highly correlated with each other (Daod 2017). To test the correlation between independent variables, one can compute a matrix of Pearson's bivariate correlations among all independent variables. Pearson's correlation coefficient measures the statistical relationship between two continuous variables, with values ranging from -1, indicating a perfect negative relationship, to +1, indicating a perfect positive relationship (Benesty et al. 2009). As has been implemented in numerous previous studies (SAGE 2019), this analysis takes values of ±0.8 to indicate multicollinearity between variables.

Once the variables to be included in the multiple regression analyses were selected, the models could be run; with each model using all the selected independent variables and one dependent variable. The results of each of these models contained the residuals, the coefficient estimate for each independent variable, the coefficient standard error, the coefficient t-value and p-value, and the adjusted R^2 and p-value for each model. The model residuals indicate the difference between each point and the regression line, i.e. the difference between the observed and predicted values (Zelterman 2010). The coefficient estimates are the expected change in the dependent variable with a one unit increase in the independent variable, whilst holding all other independent variables constant (Altman and Krzywinski 2015). The coefficient standard error is the standard deviation of each estimate, and therefore measures how precisely the model estimates the coefficient's unknown value (Altman and Krzywinski 2015). The coefficient t-value is the estimate divided by the standard error and therefore measures the size of the difference relative to the variation in the data (Montgomery et al. 2012). The p-value tests the null hypothesis that the coefficient estimate is equal to zero and therefore has no effect on the dependent variable. P-values of <0.05 are significant and therefore the null hypothesis can be rejected (Weisberg 2005). The adjusted R^2 is the percentage variation in the response variable that is explained by the variation in the independent variables and is therefore used to measure how well the model explains this variation. The adjusted R^2 considers the number of independent variables included in the model and only considers variables that are significant, and therefore have a p-value of <0.05. The p-value of the model tells us whether the overall

model is statistically significant (Weisberg 2005). A significant p-value means that there is enough evidence of a relationship between the independent variables and the dependent variable to reject the null hypothesis of no correlation.

One of the issues with basic multiple regression is the potential presence of spatial autocorrelation. Spatial autocorrelation describes the presence of systematic spatial variation in a variable, for example areas of high or low values may cluster together rather than being randomly distributed throughout the study area (Larch and Walde 2007). This causes an issue when conducting regression analyses as linear regression models rely on the assumption that observations are independent from one another. The data must therefore be examined for the presence of spatial autocorrelation before the results of the multiple regression analyses can be accepted.

The most commonly used test of spatial autocorrelation is the Moran's index (Moran's I), which measures spatial autocorrelation based on both feature locations and feature values simultaneously (Chen 2013). The Moran's I is expressed as:

$$I = \frac{N \sum_i \sum_j \omega_{ij} (x_i - \overline{x})(x_j - \overline{x})}{\left(\sum_i \sum_j \omega_{ij}\right) \sum_i (x_i - \overline{x})^2}$$

where I is the Moran's I, N is the number of evaluation units in the study area, w_{ij} is the spatial weight, x_i and x_j are the respective element attribute values of the evaluation units, and $\overline{x}$ is the average element attribute value of the evaluation units. The Moran's I has a value ranging from -1 to +1, where -1 indicates a strong negative spatial autocorrelation, 0 indicates a random distribution and +1 indicates a strong positive correlation (Chen 2013). After computing the index value, the expected value of each polygon is computed. The expected value and observed values are then compared and a z-score and p-value, indicating whether the difference between these values is significant, is computed.

The Moran's I can be calculated using either the regression variables themselves or the residuals produced after a multiple regression model has been run on the data. In this study, the Moran's I tests were run on the regression residuals rather than the individual variables in order to assess whether the models worked equally well across the study area or whether there were areas that the models were better or worse at predicting the dependent variable.

If spatial autocorrelation is present in the model residuals, a spatial regression model is conducted. These are similar to multiple regression models but also account for spatial

dependence. A number of spatial regression models exist, including spatial error models and spatial lag models; the two most commonly used. Spatial error models correlate the error terms across different spatial units, and therefore account for spatial dependence in the error terms rather than the dependent variable (LeSage 1997). Spatial lag models reflect the effect of spatial units on other nearby units in the whole study area, and therefore consider spatial lag in the dependent variable rather than in the error terms.

The outputs of spatial error models can be interpreted similarly to multiple regression models, however, also include the lambda (ʎ) value, likelihood-ratio, and Wald statistic (LeSage 1997). The lambda value is the spatial autoregressive coefficient and tells us whether there is spatial dependence in the errors, the likelihood-ratio tell us whether the inclusion of spatial dependence improves the model, and the Wald statistic tells us whether the explanatory variables included in the model are significant.

Unlike spatial error models, spatial lag models include a spatial lag term and therefore the coefficient estimate cannot be interpreted directly, as a change in one area cascades throughout the entire area (Anselin 2017). To generate variables that can be interpreted similarly to multiple regression coefficients, an additional analysis to calculate the direct, indirect and total impact of each independent variable on the dependent variable can be undertaken. The direct impact is the local impact and can be interpreted similarly to multiple regression coefficients, the indirect impact refers to the spill-over effect and is the average impact of one's neighbours on one's outcome, with the total impact being the total of the direct and indirect impacts of the predictor (Anselin 2017). Spatial lag models also include a rho *(*ρ) value which is the spatial autoregressive parameter and can be interpreted in the same way as the lambda value in the spatial error models, as well as the likelihood-ratio and Wald statistic which, like for spatial error models, refer to the whether the inclusion of spatial dependence improves the model and whether the explanatory variables included in the model are significant (LeSage 1997).

To test whether the spatial error or spatial lag models produce the best fit, Lagrange Multiplier tests can be conducted. These tests are conducted on the residuals of a spatial regression model and are employed to determine whether, and in which form, spatial correlation is present i.e. whether there is spatial dependence in the variable and/ or spatial dependence in the error terms (Larch and Walde 2007). These tests result in 'LM lag' and 'LM error' values, referring to the spatial lag and spatial error models respectively, and a p-value for each. A high LM value and a significant (<0.05) p-value indicates that that spatial regression model produced the best fit. If both the 'LM lag' and 'LM error' values are significant, their 'robust' counterparts can be analysed. These tests are robust to missing error or missing lag information and to the presence of other types of autocorrelation (Breusch and Pagan 1980). The outputs from these tests are

'RLM lag' and 'RLM error' values, referring to the robust spatial lag and robust spatial error models, in addition to the p-values. These outputs can be assessed in the same way as 'LM lag' and 'LM error', with a high value and significant p-value indicating either the spatial lag or spatial error model produced the best fit. This method is outlined in Figure 8.2.

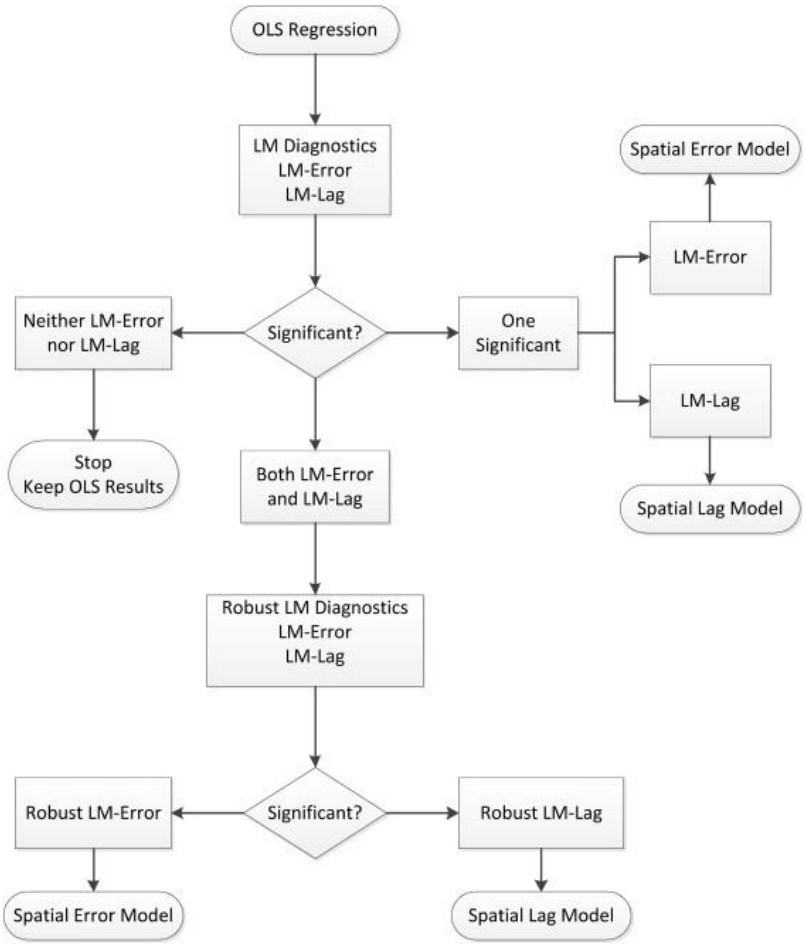


Figure 8.2 Method for interpreting the results of a Lagrange Multiplier test (Anselin 2005)

Once it has been determined whether a spatial error or spatial lag model produces the best fit, this can be implemented as the 'final model' and run on the data. For this study, the independent variables in each model were the demographic, public transport provision and HSR store accessibility variables, with the dependent variables consisting of the product consumption and retail activity variables. As only one dependent variable could be included in each model, the previously outlined methods were run multiple times; once for each dependent variable.

8.2.2 Independent variables

8.2.2.1 Public transport provision

Throughout this thesis, access to public transport services has been discussed extensively in relation to quality of life and social inclusion. By analysing variances in the provision of public transport throughout the study area, it is possible to gain an understanding of how these are related to activity behaviour, and in turn, how this could impact inclusion and wellbeing.

For this analysis, service provision was defined as the average number of buses servicing an area per hour. Weekday service provision was selected rather than weekend service provision. This was due to the large variation between Saturday and Sunday services that would have skewed any mean service provision calculations, as well as the propensity of the older population to make shopping trips on weekdays. This was evidenced in the travel day distributions shown in Figure 6.6 and the HSR daily transaction distributions shown in Figure 7.8. Similarly, off-peak service provision was chosen due to two inter-related factors – the propensity of the older population to undertake journeys during off-peak hours and the fact that the concessionary travel pass does not allow for travel during the morning peak period. It is acknowledged that this latter factor probably determined the former.

For this analysis, bus timetable data were utilised; an outline of which can be found in Section 4.2.3. The timetable data contained information on the stop number along the route ('n'), the route and direction that each bus runs ('O' for outbound and 'I' for inbound), the NaPTAN code for each scheduled stop, the scheduled arrival and departure times, and the day of the week for which the service is scheduled. Table 8.1 shows an extract of this dataset.

Table 8.1 Extract of the bus timetable data

n	Route	Journey scheduled	NaPTAN code	Arrive	Depart	Day of week	Direction
1	48	09:35:00	43000234101	09:35:00	09:35:00	1111100	O
2	48	09:35:00	43000230601	09:38:00	09:38:00	1111100	O
3	48	09:35:00	43002230102	09:39:00	09:39:00	1111100	O
			.....				
1	6A	17:29:00	43001056101	17:29:00	17:29:00	0000001	I
2	6A	17:29:00	43000075302	17:31:00	17:31:00	0000001	I
3	6A	17:29:00	43001056002	17:34:00	17:34:00	0000001	I

To develop a measure of the level of bus service provision in an area, the average hourly service provision across each LSOA was calculated. These aggregate measures of service provision were conducted at LSOA level to allow for comparability to previous analyses conducted with the smart card data and also linkage to additional census data.

Firstly, the timetable data were filtered to contain only the journeys that were scheduled to arrive at each stop on a weekday and during off-peak hours, using the '*Day of week'* and '*Arrive*' variables. The data could then be grouped by NaPTAN code, and the number of buses scheduled to visit each stop calculated. This resulted in a value corresponding to the total number of buses that were scheduled to visit each stop during off-peak, weekday hours. The average hourly service provision was then calculated as the mean number of buses that visit each stop per hour. These bus stops were then aggregated to LSOA, with the total number of buses that visit each LSOA per hour calculated.

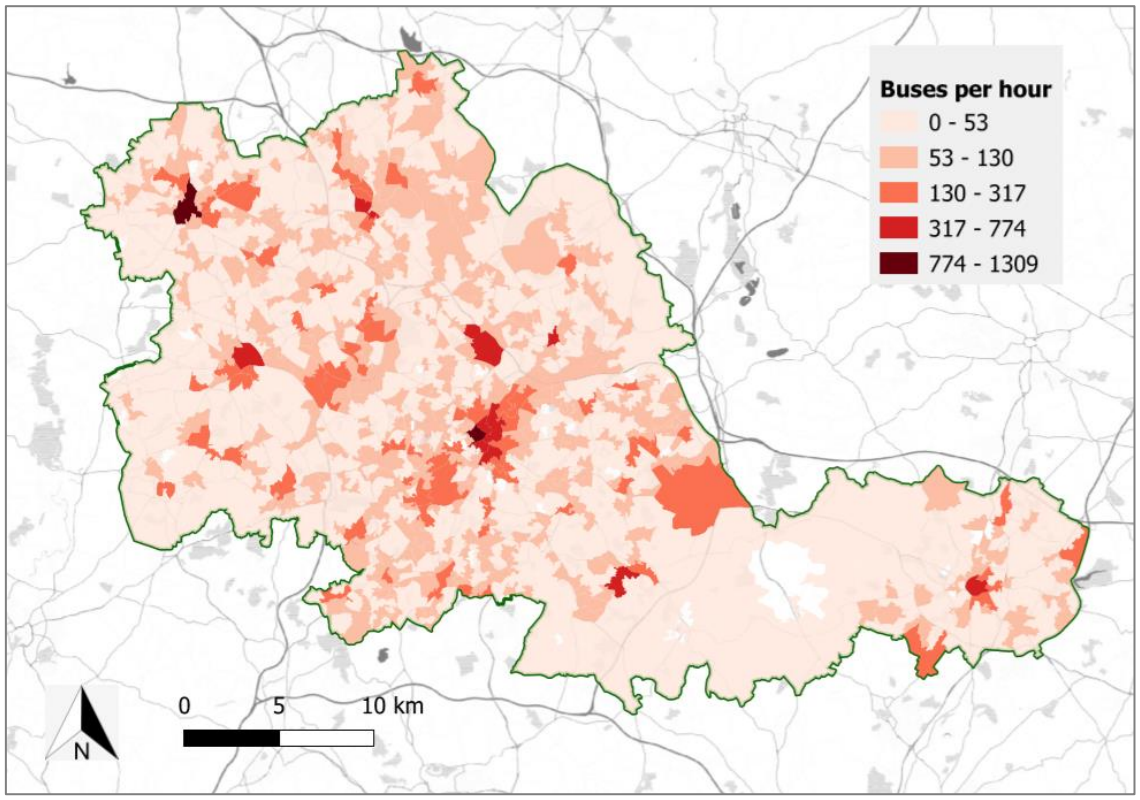


Figure 8.3 Hourly, weekday, off-peak service provision (LSOA level)

Figure 8.3 shows the total hourly service provision within each LSOA throughout the study area. Unsurprisingly, areas with the highest service provision were located in central urban areas and around main transport hubs, including Birmingham Airport. More rural areas, including

large areas of Solihull, had significantly fewer bus services per hour. This is due to both demand and population; as areas where many people live or that attract many visitors would be expected to have greater service provision. These service provision values were included in the final spatial regression models to gain an understanding of the impact that service provision has on activity behaviour, and to account for varying levels of service provision when assessing the impact of the other variables included in the models.

8.2.2.2 HSR accessibility

In addition to aggregate service provision, travel times via bus services to HSR stores were also analysed. It has been discussed throughout this thesis that for the full benefit of the ENCTS to be realised it is not only vital that there is an adequate service in terms of the number of buses supplying an area, but also that these buses are servicing the right areas, i.e. those for which there is a high demand for travel. Including both bus service provision and travel time via the bus network to HSR stores allowed for this concept to be explored further.

To calculate average travel times to HSR stores, bus timetable data were once again utilised. For consistency, the same off-peak, weekday time periods that were used for the service provision analysis were utilised. 'Chemist (Health centre)' store types were not included in this analysis due to the low number of transactions conducted at these stores using an HSR loyalty card, as shown in Figure 7.12. Additionally, because of unique patterns of use, HSR stores located within Birmingham Airport were excluded.

The first step in calculating travel times was to identify the bus stops that service each HSR store. To achieve this, the bus stop service areas that were generated in Section 6.2.7.2 were utilised. These service areas consisted of 400m polygons built along the road network; corresponding to an average 5-minute walk time. For this analysis, if an HSR store was located within the service area of a bus stop then this stop was determined to service that store. To identify these stops, a spatial join was conducted between the HSR store locations and the bus stop service areas. This resulted in a table containing all bus stops that service at least one HSR stores, as well as the departure and arrival times of all buses that service these stops. Figure 8.4 shows a simplified example of this analysis.

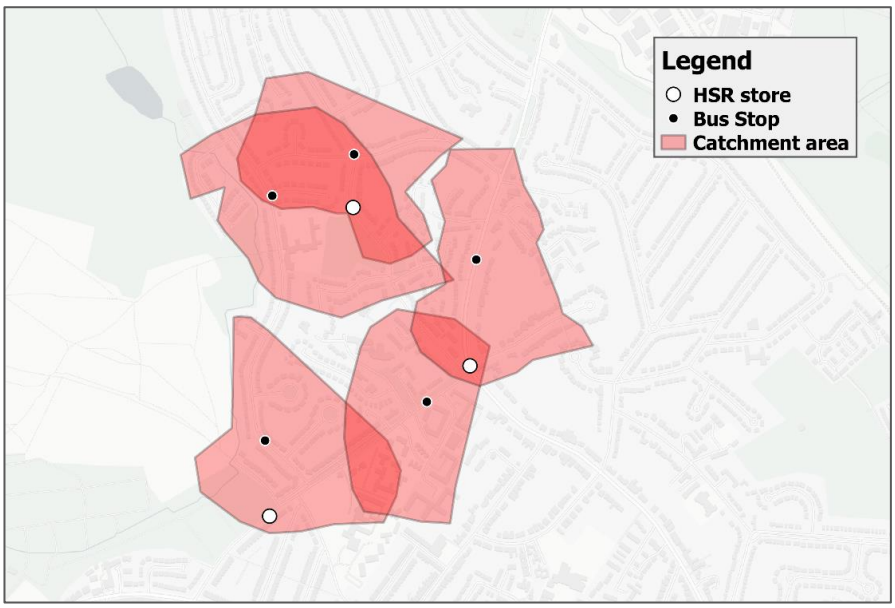


Figure 8.4 Example service area analysis

Joining this dataset to the original timetable data by the ‘route’ variable then resulted in a dataset that contained the departure time of each bus service from every bus stop along each route, and the subsequent arrival time at the stops that service HSR stores along the same route. The time difference between these departure and arrival times was then calculated and the minimum time retained. This final dataset therefore contained the minimum time it would take to reach a bus stop that services an HSR store from every other bus stop along the same route, i.e. the ‘direct travel time’.

HSR service stops

NaPTAN code	Route	Journey scheduled	Arrive	Depart
43000234202	48	09:35:00	09:46:00	09:46:00
43000230407	6A	10:05:00	10:17:00	10:17:00
43002230118	13	10:34:00	11:02:00	11:02:00

1. Join

2. Calculate difference

Original timetable data

n	Route	Journey scheduled	NaPTAN code	Arrive	Depart	Day of week
1	48	09:35:00	43000234101	09:39:00	09:39:00	1111100
2	48	09:35:00	43000230601	10:02:00	10:02:00	1111100
1	6A	10:05:00	43002230102	10:12:00	10:12:00	1111100
2	6A	10:05:00	43001056101	10:30:00	10:30:00	1111100
1	13	10:34:00	43000075302	10:47:00	10:47:00	1111100
2	13	10:34:00	43001056002	11:08:00	11:08:00	1111100

Figure 8.5 Method for joining timetable data for HSR service stops to the original timetable data and calculating the 'direct travel time'

As only 302 of the 314 routes in the study area directly serviced an HSR store, an additional analysis was conducted to account for transfers between services. The same method used for direct routes was applied but instead of bus stops that service HSR stores, 'transfer' bus stops, i.e. stops that service more than one route, were used. For bus stops without a direct route to an HSR store, the minimum time it takes to travel to a transfer stop and the subsequent minimum time it would take to then travel to an HSR store from this transfer stop were calculated. This method also considered wait times at these transfer bus stops by calculating the average time between the arrival of the first service and the departure of the next service.

Once the direct travel times and the travel times including one transfer had been calculated, the minimum for each stop was retained as the 'optimum travel time'. The average optimum travel time to each bus stop's three closest HSR stores was then calculated. Rather than just using the lowest optimum travel time, calculating the average travel time to the three closest stores acknowledges that individuals may not always wish to transact at their closest store. The reason for the journey, to buy or access a variety of products and services, will determine which of the

different store types – and the retail areas in which they are located – were visited. The results of this analysis were then aggregated to LSOA level and are shown in Figure 8.6.

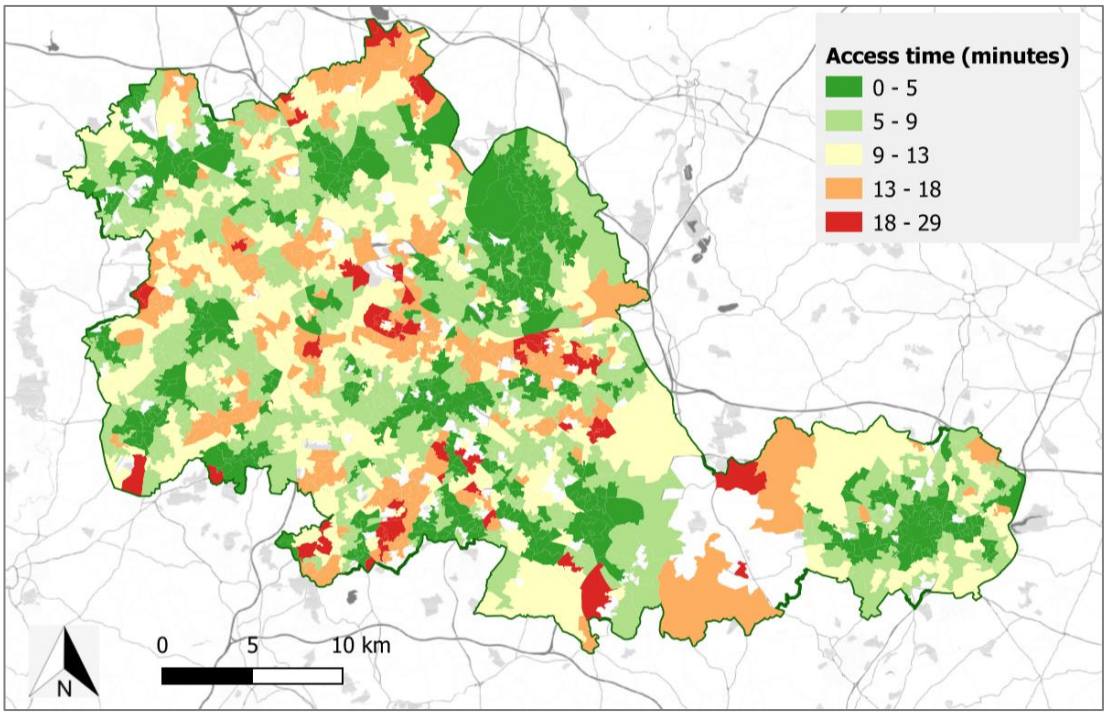


Figure 8.6 Travel times to HSR stores (average to 3 closest stores, LSOA level)

Figure 8.6 shows the average travel time to the closest 3 HSR stores for each LSOA. To protect the anonymity of the data provider, the exact locations of these stores could not be displayed. Like Figure 8.3, central urban areas generally had lower average travel times due to the propensity of HSR stores to be located in these areas. The areas to the north of Birmingham, located around Sutton Coldfield, showed low travel times despite lower service provision than some more urban areas due to the location of multiple HSR stores in this area. For those areas that do not have a travel time value, it was found that 3 HSR stores were not able to be accessed with only one transfer between bus services. These areas were primarily located in the most rural parts of Solihull, as well as isolated LSOAs throughout the study area.

Like the bus service provision analysis, the ‘travel time’ variable was also included in the regression models. By including this variable in the final models, and accounting for other variables such as distance to HSR stores and the availability of private transport, insights could be generated into the impact that bus service efficiency has on the activity of the older population.

8.2.2.3 HSR store network

As discussed in Chapter 7, the structure of the HSR store network was likely to have a significant influence on the retail behaviours of cardholders. Those living in close proximity to an HSR store might be expected to transact more frequently and at fewer stores than those living further away. To get a clear view of the impact of public transport provision and accessibility on retail activity, it was therefore important to account for the structure of the store network in the regression models. This also relates to the idea of choice, i.e. a cardholder might choose to transact locally if there is a good provision of local retail services but might be forced to travel further if these retail options are inadequate or not suitable.

To gain a general view of the structure of the store network in terms of the location of HSR stores in relation to a cardholder's home residence, the average road network distance to the nearest 3 stores was calculated for each LSOA centroid in the study area. Again, for reasons detailed earlier, this analysis did not include 'Chemist (Health centre)' store types or those stores located within Birmingham Airport. The road network distance from each LSOA centroid to each HSR store was determined using the *Closest Facility* tool in ArcMap, and the average value of the 3 nearest stores calculated.

8.2.2.4 Demographic and socioeconomic characteristics

In addition to the HSR store network, the demographic and socioeconomic characteristics of cardholders have also been shown to influence retail activity and transaction behaviour (see Chapter 7) and therefore had to be accounted for in the final regression models.

The demographic and socioeconomic variables included in the models were chosen for both their comparability to previous chapters and their relationship with mobility and activity. Table 5.3 gave a detailed description of each of these variables and an outline of their relationships with mobility and activity. Table 8.2 gives another brief description of these variables and an overview of the measures used for the following regression analyses. Table 8.3 outlines some descriptive statistics of each variable for both the West Midlands study area and England and Wales for comparison.

Table 8.2 Demographic and socioeconomic variables and their measures

Variable	Measure
IMD score *(ONS 2015)*	Description: o Official measure of relative deprivation for small areas in England o The higher the score, the more deprived an area Measure: o The raw deprivation score for each LSOA
NS-SeC (%) *(2011 Census, Source: ONS 2018b)*	Description: o The official socioeconomic classification in the UK o For those that are retired, NS-SeC is recorded as their former occupation Measure: o The percentage of the over 65 LSOA population recorded as belonging to the 'higher managerial, administrative and professional occupations', 'lower managerial, administrative and professional occupations' and 'intermediate occupations' classes
Poor health (%) *(2011 Census, Source: ONS 2018b)*	Description: o A self-assessment of health from the 2011 Census o Classes include: 'very good', 'good', 'fair', 'bad' and 'very bad' Measure: o The percentage of the over 65 LSOA population recorded as having 'bad' or 'very bad' health
Car or van availability (%) *(2011 Census, Source: ONS 2018b)*	Description: o The number of cars or vans that are available to each household Measure: o The percentage of the over 65 LSOA population with access to at least one private car or van
White ethnic group (%) *(2011 Census, Source: ONS 2018b)*	Description: o Ethnic group recorded in the 2011 Census o Groups include: 'white', 'mixed', 'Asian', 'black', 'other' Measure: o The percentage of the over 65 LSOA population recorded as belong to the 'white' ethnic group
Age *(2016, Source: HSR data)*	Description: o Included in the HSR customer data o Estimated as single year age bands in Section 7.2.1 Measure: o Average age of 'older' cardholders in each LSOA

Table 8.3 Descriptive statistics of demographic and socioeconomic variables

Variable		Mean	Standard deviation	Minimum	Maximum
IMD score	*England*	17.4	15.6	0.5	92.6
	West Midlands	31.3	17.7	1.0	85.1
NS-SeC (%)	*England & Wales*	49.0	17.2	0.0	100.0
	West Midlands	42.7	18.1	6.7	88.2
Poor health (%)	*England & Wales*	15.0	7.2	0.0	71.7
	West Midlands	21.0	7.6	0.0	51.6
Car or van availability (%)	*England & Wales*	70.4	17.4	3.6	100.0
	West Midlands	60.6	15.9	5.9	96.2
White ethnic group (%)	*England & Wales*	98.7	13.6	1.1	100.0
	West Midlands	91.9	16.1	4.0	100.0
Age	*England & Wales*	71.6	1.4	65.4	79.4
	West Midlands	69.7	1.7	60.5	78.0

Table 8.3 shows that each variable varied significantly throughout the study area, with generally large differences between the minimum and maximum. Additionally, whilst these variables were generally comparable between the study area and England and Wales, on average the West Midlands tended to be more deprived, have fewer residents belonging to 'higher managerial', 'lower managerial' and 'intermediate' NS-SeC classes, and have higher proportions of residents with 'bad' or 'very bad' health. Private transport availability was also generally lower, with an average of 60.6% of LSOA residents with access to a private car or van compared to an average of 70.4% in England and Wales. The West Midlands had a generally higher proportion of residents from a non-white ethnic background and a younger population on average. Detailed demographic and socioeconomic analyses of the study area were presented in Chapter 3.

Section 8.2.1 stated that regression models rely on the assumption that there is no multicollinearity between variables. Multicollinearity refers to a situation where the independent variables in a regression model are highly correlated. To check for this, the Pearson's correlation coefficients between all pairs of independent variables were examined, as shown in Table 8.4.

Table 8.4 Pearson's correlation coefficient matrix of demographic and socioeconomic variables

Independent variables	*IMD score*	*NS-SeC (%)*	*Poor health (%)*	*Car or van availability (%)*	*White ethnic group (%)*
IMD score	---	-0.40	0.74	-0.76	-0.44
NS-SeC (%)	-0.40	---	-0.80	0.83	0.41
Poor health (%)	0.74	-0.80	---	-0.74	-0.52
Car or van availability (%)	-0.76	0.83	-0.74	---	0.34
White ethnic group (%)	-0.44	0.41	-0.52	0.34	---

These results showed that 'NS-SeC' was highly correlated with both 'poor health' and 'car or van availability', with Pearson's correlation coefficient values of more than ±0.8. As this was indicative of multicollinearity, the 'NS-SeC' variable was not included in the regression models. The remaining demographic and socioeconomic variables did not have Pearson's correlating coefficients of greater than ±0.8 and were therefore included in each model as the independent variables, in addition to the previously outlined public transport and HSR network variables.

8.2.3 Dependent variables

The dependent variables included in each model were selected to be indicative of different retail consumption behaviours and activity patterns. Table 8.5 firstly gives a description of each dependent variable and the remainder of this section then outlines the methods used to generate these variables. As the aim of this section was to generate variables that were indicative of cardholders' general consumption and activity patterns, only those users deemed to regularly transact at HSR stores were included. These were identified in Section 7.2.3.1 and consisted of those users that had been active over at least 10 weeks of the study period.

Table 8.5 Dependent variable descriptions

Variable	Description
Transaction frequency	Average frequency of loyalty card transactions conducted by cardholders in each LSOA
Basket value (£)	Average basket value (£) for cardholders in each LSOA
Essential products (%)	Average percentage of items purchased by cardholders in each LSOA that were classed as 'essential' by the classification outlined in Section 7.2.4.1
Activity space area (km^2)	Average area (km^2) of the activity spaces of cardholders in each LSOA
Activity space compactness	Average compactness value of the activity spaces of cardholders in each LSOA
Distance travelled (km)	Average distance travelled (km) per transaction by cardholders in each LSOA
Stores visited	Average number of stores visited by cardholders in each LSOA

This section outlines the methods used to generate the retail activity variables that were explored in relation to public transport service provision and HSR store access. These variables were used as the dependent variables in the subsequent regression models. This chapter utilised a number of the retail variables generated in Chapter 7 in addition to variables describing cardholders' activity spaces. The variables employed could broadly be split into two aspects of retail behaviour; (a) consumption behaviours and (b) activity space characteristics.

8.2.3.1 Consumption behaviours

8.2.3.1.1 Transaction frequency

Transaction frequency was defined as the total number of transactions that each cardholder had made over the study period. Section 7.2.3 outlined some of the main statistics in regard to the transaction frequency of cardholders in this study. Analysing the impact of public transport provision and accessibility on the frequency of transactions made at HSR stores, whilst

accounting for socioeconomic and demographic characteristics, can provide insight into the importance of public transport for accessing retail opportunities (Palacin et al. 2016).

Transaction frequency for each cardholder was calculated by grouping transactions by their unique account number and counting the total number of transactions for each individual account. This resulted in a transaction frequency value for each cardholder which was implemented in the regression analyses outlined in Section 8.2.1.

8.2.3.1.2 Basket value

The basket values of cardholders can provide insight into the economic wellbeing of individuals and the consumption behaviours of customers (ONS 2019). In addition, comparing basket value with the public transport and HSR network can evidence how the older population make use of public transport for shopping purposes, for example how basket values vary between local retail sites and those located further from a cardholder's residence. Average basket value was calculated as the average value of a cardholder's total transactions (£) in Section 7.2.4 and these same values were used in this chapter.

8.2.3.1.3 Essential products

The percentage of products bought by each cardholder deemed to be essential can provide insight into the shopping behaviours of cardholders (Sarial-Abi et al. 2016). Comparing this with public transport variables could help transport authorities determine how services are utilised for different types of trips, for example transactions containing a large proportion of essential products may be deemed as 'necessary' trips. The relationship between public transport and essential purchasing may therefore help to distinguish between cardholders that rely on public transport services for every day, essential shopping trips, and those that primarily utilise public transport services for recreational trips. Similarly to basket value, the percentage of essential products purchased by each cardholder was calculated in Section 7.2.4.1, using a classification of essential and discretionary products based upon the Government's 'Discretionary Housing Payment' scheme (Department for Work and Pensions 2016).

8.2.3.2 Activity space characteristics

As outlined in Section 8.1, activity spaces have been linked to wellbeing and social inclusion. There are many measures of activity spaces, including the size, the intensity of interactions and

the number of distinct locations visited. This section outlines the different measures of activity spaces used in this study and how these were generated using the HSR transaction data.

8.2.3.2.1 Activity space area

The area of an individual's activity space is related to how mobile they are and the number of opportunities that they are able to access. An individual with a limited activity space may be determined to be disadvantaged, however it must be acknowledged that there is a difference between those that choose to travel locally due to an amenable local environment, and those that travel locally due to a lack of choice (Manaugh & El-Geneidy 2012). The previous analyses into accessibility and the HSR store network were conducted to account for this.

Many studies have implemented different methods to measure the size of an individual's activity space. The most commonly used methods are to measure the area of a minimum convex polygon (MCP) covering all locations visited by an individual (Buliung and Kanaroglou 2006; Fan and Khattack 2008; Manaugh and El-Geneidy 2012), or to generate a standard deviational ellipse (SDE) calculated using location coordinates (Jarv et al. 2015; Miranda-Moreno et al. 2012). For this analysis the area of each individual's activity space was calculated as the size of an MCP covering all HSR stores visited by each individual. This was chosen over the SDE method as it resulted in a more accurate measure of activity space area; only covering those locations that a cardholder was known to have visited.

To construct the MCPs, firstly the latitude and longitude of each HSR store visited by each cardholder were extracted, along with the latitude and longitude of the centroid of each cardholders' registered LSOA. These data were then imported into ArcMap and an MCP was generated for each cardholder using the *Minimum Bounding Geometry* toolbox. The area of the resulting MCP, in km^2, was then calculated for each cardholder.

8.2.3.2.2 Distance travelled

In addition to the size of an individual's activity space, many studies have also considered the average distance travelled to each visited location (Fan and Khattack 2008; Hasanzadeh et al. 2018). The MCPs covered all locations visited by an individual, however, did not account for the number of times that a location was visited or where these were located in relation to a cardholder's residence. The calculation of the average distance travelled included the distance travelled to HSR stores for all transactions rather than all distinct stores visited. This resulted in

a more accurate measure of distance travelled by considering the intensity of visits to each location.

To calculate the average distance travelled for each transaction undertaken by a cardholder, the network distance between the LSOA centroid of each cardholder's residence and the location of the HSR store was calculated. The mean distance travelled, in kilometres, was then calculated for each cardholder.

8.2.3.2.3 Activity space compactness

Solely relying on measures of activity space area and distance travelled can be misleading when analysing local travel behaviours. For example, distance travelled does not consider the direction of travel or the resulting space. The area of an activity space can be misrepresentative, as a small area does not necessarily equate to local travel behaviour (Manaugh and El-Geneidy 2012).

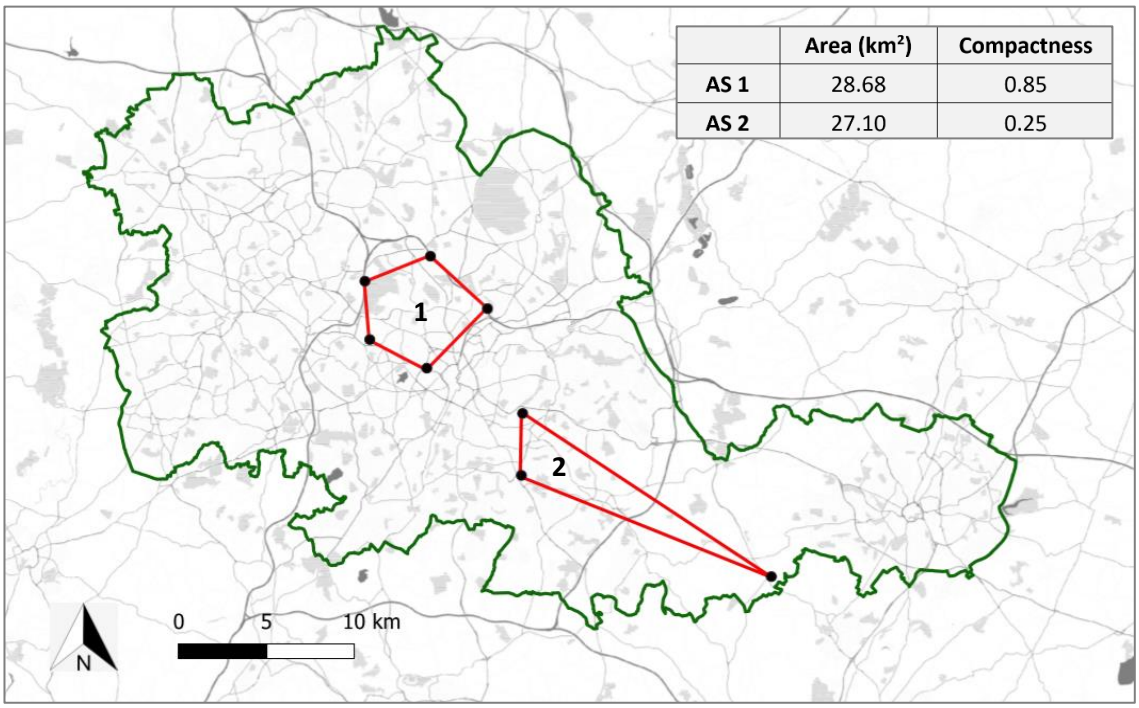


Figure 8.7 Example activity spaces, showing the area and compactness values

Comparing the two activity spaces shown in Figure 8.7, although clearly representing different types of spatial behaviour, both had similar activity space areas. To differentiate between these,

it was therefore necessary to include another variable that describes the shape of the activity space. The measure used for this was the 'compactness' of the activity space (Hirsch et al. 2014). The calculation for this measure was the ratio between the area of the activity space and the area of a circle with the same perimeter; known as the circularity ratio. This calculation can be expressed as:

$$Comp = \frac{A_r}{p^2} \times 4\pi$$

where *Comp* is the compactness of the activity space, A_r is the area of the activity space, and *p* is the perimeter. To ensure this value lies between 0 and 1, the resulting fraction was multiplied by 4π. The more compact an activity space, the closer to 1 the compactness value. In the case of the two example activity spaces shown in Figure 8.7, this measure separated 'Activity Space 1', which showed a compact activity space, from 'Activity Space 2', which showed a more elongated activity space.

8.2.3.2.4 Stores visited

The final measure of spatial store visiting behaviour included in the analysis was the number of stores that were visited by each cardholder. After accounting for the HSR store network in relation to cardholders' home locations, the number of stores visited can give insight into the mobility of cardholders. Comparing this with the public transport variables can also provide insight into how varying levels of provision and accessibility impact store visiting patterns.

To calculate this, the HSR transactions made by each cardholder were grouped by 'Account number' and 'Store ID', with the number of distinct store IDs transacted at by each unique account then calculated.

8.2.4 Data standardisation

Before running any models, the data were rescaled to ensure that the results were easily comparable. This refers to the process of rescaling data measured in different units or on different scales to a notionally common scale. One of the most common forms of rescaling is standardisation (also called z-score normalisation), whereby features are rescaled to z-scores such that μ=0 and σ=1, through the following formula:

$$z = \frac{x - \mu}{\sigma}$$

where μ *is the mean and* σ is the standard deviation from the mean (Kaur and Kumar 2016). This study utilised this standardisation method; converting both the dependent and independent variables to the same scale and therefore allowing for the coefficient estimates to be easily compared between variables.

8.3 Results

This section presents the results of the models following the methods outlined in Figure 8.1. Firstly, multiple regression models were run on each of the dependent variables in turn. The relationships between the dependent variable and the independent variables were then discussed in terms of the strength and direction of the parameters and how these relate to the hypotheses generated for each model. As discussed in Section 8.2.1, spatial autocorrelation was likely to be present in many, if not all, models due to the spatial nature of both the dependent and independent variables. Moran's I and Lagrange Multiplier tests were then conducted to detect spatial autocorrelation in the residuals of the multiple regression models and if present, in what form this existed, i.e. whether there was spatial dependence in the variable and/ or spatial dependence in the error terms (Larch and Walde 2007). If necessary, spatial regression models were then conducted to account for this spatial autocorrelation and the resulting relationships discussed in terms of the hypotheses for each model and the differences from the results generated by the multiple regression models.

8.3.1 Transaction frequency

'Transaction frequency' refers to the total number of transactions made over the study period by each cardholder, aggregated to average frequency at LSOA level. This was hypothesised to be related to a number of the independent variables. In terms of demographics and socioeconomics, 'transaction frequency' was hypothesised to decrease as deprivation increases owing to the generally lower levels of disposable income available to more deprived populations (Clark et al. 2014). Those living in areas with generally poorer health and older cardholders would also be expected to make fewer transactions, as these can limit the ability of cardholders to travel to stores. In terms of the transport network, those with access to private transportation would be likely to undertake more transactions due to the ease of access that private modes of transport afford. Cardholders that reside in areas with high public transport service provision and those

with shorter travel times to HSR stores were also hypothesised to have a higher transaction frequency due to the increased accessibility of HSR stores for frequent visits. To test these hypotheses, a multiple regression model was run, with 'transaction frequency' as the dependent variable and those variables outlined in Section 8.2.2.4 as the independent variables. This model tells us whether there was a statistically significant relationship between 'transaction frequency' and any of the independent variables.

8.3.1.1 Multiple regression

Table 8.6 shows the results of the multiple regression model. The results included the estimate, standard error, t-value and p-value for each independent variable, and the adjusted R^2 and p-value for the whole model; the definitions of which were outlined in Section 8.2.1.

Table 8.6 'Transaction frequency' – multiple regression results

Independent variable	**Estimate**	**Standard error**	**t-value**	**p-value**	**Significance**
IMD score	-0.12	0.02	-2.69	0.01	.
Poor health	-0.22	0.04	-5.89	<0.0001	***
Car or van availability	1.63	0.04	4.00	<0.0001	***
White ethnic group	0.04	0.04	1.64	0.10	-
Age	0.11	0.02	5.32	<0.0001	***
Network distance	-0.21	0.02	-9.00	<0.0001	***
Service provision	0.02	0.02	0.80	0.42	-
Travel time	-0.13	0.02	-5.72	<0.0001	***
Adjusted R^2	0.30				
p-value	< 2.20e-16				

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

This model showed that all independent variables except 'white ethnic group' and 'service provision' were found to have a significant relationship with 'transaction frequency'. Although hypothesised that areas with a higher level of bus service provision may also have a higher

average transaction frequency, this was not found to be statistically significant. The lack of a significant relationship between 'transaction frequency' and 'service provision', as well as the negative relationship with 'travel time' suggested that regular bus services may only benefit residents if these services run efficiently to the areas that public transport users wish to visit. Understanding the demands of transport users in terms of the areas they wish to visit, an aspect of which was explored in Chapter 6, is therefore vital to enable transport authorities to ensure that public transport services are meeting the mobility needs of different groups of users, for example both commuting populations and the older population.

'IMD score', 'poor health', 'network distance' and 'travel time' were all found to have a negative impact on the average frequency of transactions. Focusing on the impacts of each of these variables, 'poor health' was found to have the greatest impact, followed by 'network distance'. Due to the healthcare and medical products and services that the HSR provides, it may be expected that those with poor health would want or need to transact frequently at stores. The negative impact that 'poor health' had on 'transaction frequency' instead meant that in areas with higher proportions of residents with 'bad' or 'very bad' health, fewer average HSR transactions were made. This suggests that people residing in these areas were actually physically unable to take advantage of these necessary services and supplies. Similarly, the negative relationship with 'IMD score' meant that those that resided in more deprived areas transacted less frequently on average, which again suggested that these populations were unable to take advantage of retail opportunities.

'Network distance' and 'travel time' referred to different aspects of accessibility and ease of access. 'Network distance' related to the distance from each LSOA centroid to an HSR store, whereas 'travel time' referred to the average time it would take to travel to an HSR store via the bus network from each LSOA. The negative impact that these variables had on the frequency of transactions made in HSR stores highlights the importance of real accessibility to retail opportunities to allow for regular store visiting.

'Age' and 'car or van availability' were the only independent variables found to have significant positive impacts on 'transaction frequency'; with an increase in the average age of cardholders and an increase in the proportion of the population with access to private transportation found to result in an increase in the average frequency of transactions. Again, this positive relationship with 'age' may be due to the healthcare and medical services that these stores provide, with Section 7.2.4.1 discussing how older cardholders were more likely to purchase these healthcare products. Additionally, this points to the fact that a large proportion of this population were likely to be retired and therefore potentially have more time to shop. Additionally, the older population would be expected to generally be less 'tech-savvy' and therefore less likely to shop

online for these goods and services. E-commerce transactions are not included in the HSR data and so the increase of transaction frequency with age may be due to the omission of online purchases in these data. The positive relationship with 'car or van availability' means that in areas where a high proportion of the population have access to private transportation, cardholders were likely to transact at HSR stores more often.

The overall R^2 of the model meant that 30% of the variation seen in average transaction frequency could be explained by the independent variables, and the p-value of <2.20e-16 meant that the model was statistically significant. Although this R^2 value may be viewed as relatively low, this can be explained by the spatial aggregation of the data to LSOA level and the inferences that were made regarding the demographic and socioeconomic characteristics of cardholders. This generalisation of both the transaction frequency variable and the cardholder characteristics likely resulted in a decreased R^2 value as variances within LSOAs were not captured.

8.3.1.2 Spatial autocorrelation

As discussed in Section 8.2.1, there was likely to be spatial autocorrelation present in the residuals of the model. To test for this, a Moran's I test was run on the residuals of the multiple regression model. The results of this are shown in Table 8.7.

Table 8.7 'Transaction frequency' – Moran's Index

Dependent variable	Moran's I	Expectation	Variance	z-value	p-value
Transaction frequency	0.26	-0.0024	0.0003	16.01	<2.2e-16

The results of the Moran's I test presented in Table 8.7 showed positive spatial clustering, i.e. similar residual values clustered together, with a positive Moran's I index value of 0.26. The positive z-value and a statistically significant p-value meant that the residuals were spatially clustered and that this clustering was statistically significant, and therefore the null hypothesis of no spatial clustering could be rejected. This evidence of spatial autocorrelation meant that the results of the multiple regression model were likely to be unreliable. This was because this observed spatial autocorrelation violated the assumption that the outcomes were independent of

each other which would affect the coefficient estimates (Kuhn and Dormann 2012). Therefore, to gain reliable results, regression models that account for spatial autocorrelation were implemented.

To determine whether a spatial lag or a spatial error model would be the most appropriate to implement in order to account for this spatial clustering, a Lagrange Multiplier test was conducted. The Lagrange Multiplier tests this by analysing spatial lag in the dependent variable and the presence of spatial lag specifically in the error terms. Spatial lag in the dependent variable would mean a spatial lag model would produce the best results, whereas spatial lag in the error terms means a spatial error model should be implemented. A detailed overview of the Lagrange Multiplier test is given in Section 8.2.1. Table 8.8 shows the results of the Lagrange Multiplier test, including the results of the robust spatial error and spatial lag tests.

Table 8.8 'Transaction frequency' – Lagrange Multiplier test results

	LMerr		**LMlag**		**RLMerr**		**RLMlag**	
Dependent variable	***Value***	***p-value***	***Value***	***p-value***	***Value***	***p-value***	***Value***	***p-value***
Transaction frequency	249.47	<2.2e-16	290.19	<2.2e-16	1.60	0.21	42.33	7.7e-11

A significant LMerr and LMlag meant that the robust counterparts had to be analysed. These 'robust' tests were robust to missing error or missing lag information and also to the presence of other types of autocorrelation (Breusch and Pagan 1980) and could therefore give a clearer indication of which spatial regression model would be the most appropriate to conduct. Analysing the RLMerr and RLMlag, the value of the RLMlag value was greater and statistically significant. This meant that a spatial lag model would produce the best fit.

8.3.1.3 Spatial lag model

Similarly to the multiple regression model, the spatial lag model was run using 'transaction frequency' as the dependent variable. As spatial autocorrelation in the residuals of a multiple regression model can often lead to an overstatement of the magnitude of the coefficient estimates it was hypothesised that the magnitudes of the estimates found by the spatial lag model would be smaller than those found by multiple regression, but the direction of these relationships would be the same. The results of this spatial lag model are shown in Table 8.9.

Table 8.9 'Transaction frequency' – spatial lag model

Independent variable	Transaction frequency					
	SLM			Impacts		
	Estimate	*z-value*	*p-value*	*Direct*	*Indirect*	*Total*
IMD score	-0.05	-1.30	0.19	-0.05	-0.03	-0.09
Poor health	-0.17	-5.05	<0.001 **	-0.18	-0.11	-0.29
Car or van availability	0.14	3.72	<0.001 **	0.14	0.09	0.23
White ethnic group	0.03	1.17	0.24	0.03	0.02	0.05
Age	0.06	3.28	<0.001 **	0.06	0.04	0.10
Network distance	-0.14	-6.46	<0.001 **	-0.14	-0.09	-0.23
Service provision	0.03	1.30	0.19	0.03	0.02	0.05
Travel time	-0.08	-3.81	<0.001 **	-0.08	-0.05	-0.13
Rho (ρ)	0.10 (***)					
Likelihood-ratio	238.26 (**)					
Wald statistic	249.85 (***)					

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

The relationships between the independent variables and 'transaction frequency' shown in Table 8.9 differed from those found by the multiple regression analysis. Although found to have a significant relationship in the multiple regression analysis, 'IMD score' was not found to significantly impact the frequency of transactions when accounting for spatial autocorrelation. It was hypothesised that increased deprivation would result in decreased transaction frequency due to lower levels of disposable income. However, this model suggested that when accounting for other factors such as health and the accessibility of HSR stores, higher levels of deprivation did not significantly decrease the frequency of transactions. This may be due to the number of essential products that HSR stores provide, which is explored through a further model in Section 8.3.3.

The remaining results showed similar relationships between the dependent and independent variables. 'Poor health', 'network distance' and 'travel time' were still found to negatively impact 'transaction frequency' however the direct impacts of these were found to not be as strong as by the multiple regression analysis. This was hypothesised due to the impact that spatial autocorrelation can have on the magnitude of coefficient estimates. The total impact,

which included the spill-over from one's neighbours, was greater for each 'network distance' and 'poor health' and the same as was found via multiple regression for 'travel time'.

Like with multiple regression, 'car or van availability' and 'age' were found to have significant positive impacts on 'transaction frequency'. Again, these were not as strong as found with the multiple regression model, likely due to the impact of spatial autocorrelation. In particular, the relationship with 'car or van availability' was found to have a much smaller impact than shown in Table 8.6. This suggests that the coefficient estimates generated by the multiple regression model were greatly influenced by the presence of spatial autocorrelation.

The rho, likelihood-ratio and Wald statistic shown in Table 8.9 were all significant, meaning that the inclusion of the lagged values improved the model and, at least some, of the explanatory variables were significant. This means that the results of this spatial lag model should be considered rather than the results of the multiple regression model which were impacted by the presence of spatial autocorrelation in the residuals.

8.3.2 Basket value

'Basket value' refers to average value of transactions conducted by each cardholder. This can be indicative of economic wellbeing. It can also indicate how the bus network is used by the older population in terms of how the value of transactions varies in terms of the accessibility of HSR stores. The value of transactions was hypothesised to be related to socioeconomic variables such as 'IMD score', 'poor health' and access to private transport. As disposable income is generally lower for those that are more deprived or of poorer health (Clark et al. 2014), 'basket value' was hypothesised to have a negative relationship with these variables. In terms of access to private transport, it was predicted that an increase in private transport availability would lead to an increase in 'basket value', both due to the increased economic wellbeing that is associated with private transport access and the ability to purchase more products when travelling by car over public transport or walking. In terms of 'age', 'basket value' was expected to decrease as average age increases, with the older population likely to rely primarily on pensions and therefore generally have less disposable income than working age populations. Focusing on the public transport and HSR store network variables, it would be expected that 'basket value' would increase with the distance or time taken to travel to HSR stores. This is because the less accessible HSR stores are, the less likely cardholders are to undertake regular, lower value transactions, but may rather purchase more items during each visit.

8.3.2.1 Multiple regression

The results of the multiple regression analyses are shown in Table 8.10, with 'basket value' as the dependent variable.

Table 8.10 'Basket value' – multiple regression results

Independent variable	Estimate	Standard error	t-value	p-value	Significance
IMD score	0.02	0.05	0.47	0.64	-
Poor health	-0.05	0.04	-1.12	0.26	-
Car or van availability	0.10	0.05	2.16	0.03	.
White ethnic group	-0.10	0.03	-3.44	<0.001	**
Age	-0.20	0.02	-8.79	<0.0001	***
Network distance	0.17	0.03	6.74	<0.0001	***
Service provision	0.07	0.02	2.96	<0.001	**
Travel time	0.11	0.03	4.11	<0.0001	***
Adjusted R^2	0.12				
p-value	< 2.20e-16				

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

Although hypothesised that 'IMD score' and 'poor health' would have a negative relationship with the value of transactions these relationships were not found to be significant. These results were unexpected, in particular the lack of a relationship with 'IMD score', as those living in more deprived areas would be expected to have less disposable income and therefore make lower value purchases. The essential nature of many of the products sold by the HSR could explain the lack of relationship between deprivation and transaction value. The value to the individual of such essential purchases is highly likely to be less dependent on disposable income and economic wellbeing.

This model found that 'white ethnic group' and 'age' negatively impacted the average value of transactions. Of these, 'age' had the greatest impact, of -0.20. In comparison, 'white ethnic group' had a negative impact of -0.10. The significant negative impact of 'age' on the value of

transactions made by cardholders provided insight into economic wellbeing and suggested that those that are younger may have more disposable income whilst the older age group are dependent on their fixed pension. The smaller, but still significant, negative impact of 'white ethnic group' meant that the average value of products purchased was lower in areas with a higher proportion of residents from white ethnic backgrounds.

'Car or van availability', 'network distance', 'service provision' and 'travel time' were all shown to have a positive impact on the average value of transactions made by cardholders. An increase in the distance to HSR stores and the time taken to access these stores via public transport compared with the decrease in transaction frequency found in the first spatial lag model suggested that the more expensive a retail trip was, whether this be financially or the time taken to undertake the trip, the less often these trips were made. The increase in the transaction value suggested that the less easily accessible retail opportunities were, the more infrequent, high value transactions that were undertaken. The positive impact of 'car or van availability' meant that those residing in areas with higher access to private transportation generally made higher value purchases. This may partly be due to the constraints of relying on public transport services for shopping trips, for example limits on the number of goods that can be easily carried on services meaning that those without access to private transportation may purchase fewer products, and therefore have lower basket values.

The p-value of the model meant that the overall model was statistically significant, and the R^2 value of 0.12 suggested that 12% of the variation in 'basket value' could be explained by variations in the independent variables. This R^2 value was relatively small, meaning that only a small amount of the differences in average basket value in each LSOA could be explained by the independent variables included in this model. This may have been due to the impact of additional variables that were not captured in the model due the unavailability of data, for example income. 'Basket value' was also likely to be related to the regularity of transactions and the types of users that were making transactions. Active users were identified in Section 7.2.3.1; however, this model did not account for different types of active users. For example, 'basket value' would be likely to differ between those that made regular purchases of essential products and those that used HSR stores for purchasing lunch and snacking products throughout the week. The value of these transactions may therefore differ due to factors not included in the model. Future research accounting for this could improve the outcomes of these models.

8.3.2.2 Spatial autocorrelation

Like the previous 'transaction frequency' model, the spatial nature of the data meant that the presence of spatial autocorrelation in the residuals of the multiple regression model was likely. To test for the presence of spatial autocorrelation, a Moran's I test was conducted on these residuals. The results of this are shown in Table 8.11.

Table 8.11 'Basket value' **–** Moran's Index

Dependent variable	Moran's I	Expectation	Variance	z-value	p-value
Basket value (£)	0.13	-0.0024	0.0003	7.83	2.45e-15

The positive z-value and a statistically significant p-value shown in Table 8.12 meant that data were spatially clustered, with areas of high and low values grouped together, and that this clustering was statistically significant. The presence of spatial autocorrelation in the residuals of the model meant that a regression model that accounts for spatial autocorrelation had to be conducted. To determine whether this should be a spatial lag model or a spatial error model a Lagrange Multiplier test was run.

Table 8.12 'Basket value' – Lagrange Multiplier test results

Dependent variable	LMerr		LMlag		RLMerr		RLMlag	
	Value	***p-value***	***Value***	***p-value***	***Value***	***p-value***	***Value***	***p-value***
Basket value (£)	58.47	2.07e-14	61.76	3.89e-15	0.06	0.81	3.35	0.04

The results of the Lagrange Multiplier test showed that both the LMerr and LMlag were significant, and therefore the robust counterparts were analysed. Comparing RLMerr and RLMlag it was clear that a spatial lag model should be conducted as this had a higher value and a significant p-value.

8.3.2.3 Spatial lag model

Due to the presence of spatial autocorrelation in the residuals of the multiple regression model, the coefficient estimates were likely to be biased. It was hypothesised that conducting a spatial lag model instead of a multiple regression model would improve the fit and account for this bias. This would result in coefficient estimates that were similar to those found in the multiple regression model but of a smaller magnitude as the impact of neighbourhood effects were accounted for. Table 8.13 shows the results of the spatial lag model.

Table 8.13 'Basket value' – spatial lag model

Independent variable	Value					
	SLM			Impacts		
	Estimate	*z-value*	*p-value*	*Direct*	*Indirect*	*Total*
IMD score	0.03	0.63	0.53	0.03	0.01	0.04
Poor health	-0.05	-1.29	0.20	-0.05	-0.02	-0.07
Car or van availability	0.10	2.30	0.02 .	0.10	0.03	0.13
White ethnic group	-0.09	-3.32	<0.001 **	-0.09	-0.03	-0.12
Age	-0.19	-8.37	<0.0001***	-0.19	-0.06	-0.25
Network distance	0.14	5.29	<0.0001 ***	0.14	0.04	0.18
Service provision	0.06	2.34	0.02 .	0.06	0.02	0.08
Travel time	0.09	3.57	0.00 **	0.10	0.03	0.13
ρ	0.06 (***)					
Likelihood-ratio	54.55 (***)					
Wald statistic	55.32 (**)					

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

Similarly to the multiple regression model, all independent variables except 'IMD score' and 'poor health' were found to have significant relationships with 'basket value'. In terms of the direct impact of each of these variables, these were similar to those that were found with the multiple regression model. 'White ethnic group' and 'age' were found to have significant negative impacts on 'basket value', whereas 'car or van availability', 'network distance', 'service provision' and 'travel time' had significant positive impacts. These relationships were

hypothesised, with increased age generally corresponding with decreased disposable income. The negative relationship with 'white ethnic group' may be due to areas with high proportions of residents from non-white ethnic groups often containing specialist stores where these groups conduct a large proportion of their transactions (Hamlett et al. 2008). As predicted, 'basket value' decreased as the accessibility of stores increased. This is likely due to an increased cost of making a shopping trip resulting in more items being purchased so that these trips can be undertaken less regularly.

The significance of the rho, likelihood-ratio and Wald statistic meant that the inclusion of the lagged values improved the model and, at least some, of the explanatory variables were significant. Again, this meant that the results of the spatial lag model were more reliable than those produced by the multiple regression model, as the spatial autocorrelation present in the residuals produced bias in the coefficient estimates.

8.3.3 Essential products

The percentage of products purchased that were classed as essential can also tell us about the retail behaviours of cardholders, with essential products more likely to be purchased locally than discretionary products. In terms of the independent variables, it was hypothesised that 'IMD score', 'poor health' and 'age' would increase the percentage of products purchased that were classified as 'essential'. Higher levels of deprivation often mean lower levels of disposable income which can result in fewer discretionary products being purchased (Clark et al. 2014). Areas with a high proportion of cardholders of poor health or older age may be expected to generally purchase a greater proportion of essential products due to the classification of healthcare and medical products as essential in Section 7.2.4.1. In terms of the transport and HSR store network variables, it was hypothesised that the further that one has to travel to an HSR store and the longer it takes to travel to these stores, the fewer essential products would be purchased as essential purchasing would be expected to be higher at local stores that are visited regularly.

Like previous models, it was predicted that there would be spatial autocorrelation present in the residuals. Again, this was tested using Moran's I and Lagrange Multiplier tests to determine whether an additional spatial regression model should be implemented to improve the fit of the model.

8.3.3.1 Multiple regression

Table 8.14 shows the results of the multiple regression model. The dependent variable was 'essential products' which refers to the percentage of total products purchased that were classified as essential.

Table 8.14 'Essential products' – multiple regression results

Independent variable	Estimate	Standard error	t-value	p-value	Significance
IMD score	-0.04	0.05	-0.92	0.36	-
Poor health	-0.18	0.04	-4.64	<0.001	**
Car or van availability	0.00	0.04	0.05	0.96	-
White ethnic group	-0.11	0.03	-4.47	<0.001	**
Age	0.23	0.02	10.12	<0.0001	***
Network distance	-0.27	0.02	-11.24	<0.0001	***
Service provision	-0.01	0.02	-0.27	0.79	-
Travel time	-0.14	0.02	-5.72	<0.0001	***
Adjusted R^2	0.21				
p-value	< 2.20e-16				

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

Of the independent variables included in the model, 'poor health', 'white ethnic group', and 'travel time' all showed negative relationships with essential purchasing. This negative relationship with health meant that in areas with a high proportion of residents with poor health, essential purchasing was lower than areas with better general health. This could be seen as contradictory, with those of poor health likely to be in greater need of essential products, such as healthcare products and medication, but could be indicative of a relationship between good health and the purchasing of healthcare products. It should also be noted that prescriptions would be free of charge for a large number of these older cardholders, and therefore a loyalty card would not be used for these transactions, and consequently, the transactions would not be recorded in the HSR loyalty card data.

The significant negative relationship with 'white ethnic group' meant that areas with a higher proportion of residents from a white ethnic background generally had lower proportions of essential products purchased. As discussed in the previous section, this may be due to greater purchasing at specialist stores often located in these areas (Hamlett et al. 2008).

Focusing on the 'travel time' variable, the negative relationship with 'essential products' suggested that those residing in areas for which HSR stores were less accessible via the bus network were more likely to conduct discretionary, rather than essential, purchases. Considering the negative relationship with 'transaction frequency' and the positive relationship with 'average value', these results indicated that those that resided further from HSR stores were less likely to use these stores as their local stores for frequent, necessary purchases.

'Age' was the only variable found to have a significant positive impact on essential purchasing. This relationship was hypothesised as health generally declines with age, and therefore an increase in the purchasing of essential items such as healthcare products would be expected in areas with a high proportion of older residents. Additionally, generally lower levels of disposable income in older populations would be expected to result in decreased levels of discretionary purchasing (Clark et al. 2014).

The R^2 value of 0.21 meant that 21% of the variation in the percentage of essential products purchased could be explained through the independent variables, and the p-value of < 2.2e-16 meant that the overall model was significant. Although 0.21 was relatively low, this was likely explained by the aggregation of variables to LSOA level. Variations between cardholders within these LSOAs would result in a lower R^2 value than may have been found had this analysis been conducted at the individual level.

8.3.3.2 Spatial autocorrelation

To test for the presence of spatial autocorrelation in the residuals of the multiple regression model, a Moran's I test was run. Table 8.15 shows the results of this test.

Table 8.15 'Essential products' – Moran's Index

Dependent variable	Moran's I index	Expectation	Variance	z-value	p-value
Essential products (%)	0.25	-0.0024	0.0003	15.33	<2.2e-16

The positive z-value and significant p-value showed that spatial autocorrelation was present in the residuals of the model, with high values clustering around high values and low around low. To test in what form this spatial autocorrelation was present, a Lagrange Multiplier test was run on these data.

Table 8.16 'Essential products' – Lagrange Multiplier test results

Dependent variable	LMerr		LMlag		RLMerr		RLMlag	
	Value	***p-value***	***Value***	***p-value***	***Value***	***p-value***	***Value***	***p-value***
Essential products (%)	228.33	<2.2e-16	287.68	<2.2e-16	3.15	0.08	62.50	2.67e-15

The results of the Lagrange Multiplier test showed that both LMerr and LMlag were significant, however when the robust counterparts of these were tested, only the RLMlag value was significant. Therefore, a spatial lag model was run to account for spatial autocorrelation in the residuals of the multiple regression model.

8.3.3.3 Spatial lag model

Like previous models, the presence of spatial autocorrelation was hypothesised to have increased the value of the correlation coefficients of the independent variables. It was therefore predicted that the coefficients found through a spatial lag model would have smaller magnitudes, but the directions of these impacts would remain the same. Table 8.17 shows the results of the spatial lag model, including the estimate, z-value, p-value and direct, indirect and total impacts of each independent variable on 'essential products', as well as the rho, likelihood ration and Wald statistic of the overall model.

Table 8.17 'Essential products' – spatial lag model

Independent variable	Essential products					
	SLM			Impacts		
	Estimate	*z-value*	*p-value*	*Direct*	*Indirect*	*Total*
IMD score	-0.01	-0.20	0.84	-0.01	-0.01	-0.02
Poor health	-0.11	-3.15	<0.001 **	-0.12	-0.08	-0.20
Car or van availability	0.02	0.42	0.68	0.02	0.01	0.03
White ethnic group	-0.07	-3.06	<0.001 **	-0.07	-0.05	-0.12
Age	0.17	8.52	<0.0001 ***	0.18	0.12	0.30
Network distance	-0.16	-7.12	<0.0001 ***	-0.17	-0.12	-0.29
Service provision	0.00	0.05	0.96	0.00	0.00	0.00
Travel time	-0.10	-4.58	<0.0001 ***	-0.11	-0.07	-0.18
ρ	0.11 (***)					
Likelihood-ratio	238.67 (***)					
Wald statistic	261.38 (***)					

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

Like the multiple regression model, the spatial lag model found that 'IMD score' and 'car or van availability' did not exhibit statistically significant relationships with 'essential products'. However, a statistically significant negative relationship was found between 'essential products' and 'network distance' that was not found in the multiple regression model. This negative relationship meant that as the distance to HSR stores increased, the percentage of products purchased that were classed as essential decreased. This relationship was hypothesised, with essential products more likely to be purchased locally than discretionary products.

The remaining independent variables showed similar relationships as found in the multiple regression model. However, as hypothesised, the direct impacts of these variables on 'essential products' were not found to be as strong. This was due to the presence of spatial autocorrelation resulting in bias in the coefficient estimates. Like the previous spatial lag models, the significant rho, likelihood-ratio and Wald statistic meant that the inclusion of lagged values improved the fit of the model.

8.3.4 Activity space area

'Activity space area' refers to the size of a cardholder's activity space, derived from the network of stores visited. The method for calculating this was outlined in Section 8.2.3.2.1. In terms of how the independent variables would be expected to impact the area of cardholders' activity spaces, those related to economic wellbeing and health ('IMD score', 'poor health' and 'age') were hypothesised the negatively impact 'activity space area'. The variables related to transport and the HSR store network ('car or van availability', 'network distance', 'service provision' and 'travel time') were expected to have positive relationships with 'activity space area'. Areas with high private transport availability and bus service provision would likely result in residents with larger activity spaces due to the higher levels of mobility afforded by transportation access. Higher 'network distance' and 'travel time' values were hypothesised to have the opposite effect, with the greater cost of travelling to HSR stores decreasing the area of cardholders' activity spaces.

8.3.4.1 Multiple regression

Table 8.18 'Activity space area' – multiple regression results

Independent variable	Estimate	Standard error	t-value	p-value	Significance
IMD score	-0.18	0.05	-4.10	<0.0001	***
Poor health	-0.14	0.04	-3.59	<0.001	**
Car or van availability	0.12	0.04	2.89	<0.001	**
White ethnic group	0.11	0.03	4.18	<0.0001	***
Age	-0.06	0.02	-2.91	<0.001	**
Network distance	0.05	0.02	1.92	0.06	-
Service provision	0.06	0.02	2.78	0.01	.
Travel time	0.01	0.02	0.53	0.59	-
Adjusted R^2	0.22				
p-value	< 2.20e-16				

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

‘IMD score’ was shown to have the greatest impact on ‘activity space area’. An increase in score, and therefore the deprivation of an LSOA, resulted in a decrease in the average area for cardholders residing in that LSOA. Similarly, ‘poor health’ also resulted in a decrease in area. ‘Poor health’ refers to the proportion of LSOA population classified as having bad or very bad health. This model therefore suggested that an increase in the population classed as having poor health in an LSOA resulted in limited average activity space areas. The average age of cardholders residing in each LSOA was also found to correspond with a decrease in the average size of activity spaces, suggesting that older cardholders may have constrained activity spaces.

‘White ethnic group’, ‘car or van availability’ and ‘service provision’ all had significant positive coefficient estimates, meaning that as these variables increased, so did ‘activity space area’. ‘White ethnic group’ referred to the proportion of the population that were of a white ethnic background and was found to have the largest positive impact on activity space area. This therefore suggested that a higher proportion of white residents in an LSOA resulted in a larger average activity space. ‘Car or van availability’ referred to the proportion of the LSOA population that had access to a private car or van, and therefore the positive coefficient for this variable suggested that the higher the proportion of the population with access to private transportation, the larger the average activity spaces of the cardholders residing in these areas.

In terms of the public transport variables, only ‘service provision’ was shown to have a significant impact on ‘activity space area’. An increase in the hourly provision of bus services in an area was found to result in a slight increase in the average activity space of cardholders, which suggested that reliable public transport services allowed for greater travel.

Both ‘travel time’ and ‘network distance’ were not found to have a significant impact on the size of a cardholder’s activity space. These variables are linked, however do not represent the same concept. ‘Network distance’ refers to the average distance along the road network to the closest 3 HSR stores from each LSOA, whereas ‘travel time’ refers to the average time it would take via the bus network to travel to these stores. The lack of a significant relationship between these variables and the average size of a cardholder’s activity space area therefore meant that the average time taken to visit HSR stores did not significantly impact the averagesize of cardholders’ activity spaces.

The R^2 value of 0.22 meant that 22% of the variation in ‘activity space area’ could be explained by variations in the independent variables, and the p-value of the overall model meant that this was statistically significant. Again, this relatively low R^2 value was likely the result of the aggregation of variables to LSOA level.

8.3.4.2 Spatial autocorrelation

Like previous models, to check for the presence of spatial autocorrelation in the residuals, a Moran's I test was conducted. The results of this are shown in Table 8.19.

Table 8.19 'Activity space area' – Moran's Index

Dependent variable	Moran's I index	Expectation	Variance	z-value	p-value
Activity space area (km^2)	0.03	-0.0024	0.0003	2.19	0.01

The positive z-value and significant p-value suggested the presence of spatial autocorrelation. The z-value was smaller and the p-value less significant than previous models, suggesting that the spatial autocorrelation in the residuals was not as great. However, as this was still found to be significant, a spatial regression model was conducted. Again, a Lagrange Multiplier test was run to determine whether to select a spatial error or spatial lag regression to account for this spatial autocorrelation.

Table 8.20 'Activity space area' – Lagrange Multiplier test results

Dependent variable	LMerr		LMlag		RLMerr		RLMlag	
	Value	*p-value*	*Value*	*p-value*	*Value*	*p-value*	*Value*	*p-value*
Activity space area (km^2)	4.12	0.04	3.58	0.06	0.54	0.46	0.00	0.98

The values and p-values for the LMerr and LMlag test showed that the LMerr produced the highest value and the only significant p-value. This meant that spatial autocorrelation was present in the error terms and therefore a spatial error model would produce the best fit. A detailed description of spatial error models is given in Section 8.2.1.

8.3.4.3 Spatial error model

Like previous models, it was hypothesised that similar coefficient estimates would be found by the spatial lag model but that these may be of a lower magnitude. Unlike spatial lag models, the estimates produced by spatial error models could be interpreted in the same way as those produced by multiple regression analyses. Table 8.21 shows the results of the spatial error model, with 'activity space area' as the dependent variable.

Table 8.21 'Activity space area' – spatial error model

	Activity space area		
Independent variable	***Estimate***	***z-value***	***p-value***
IMD score	-0.18	-3.85	<0.001 **
Poor health	-0.14	-3.53	<0.001 **
Car or van availability	0.13	3.03	<0.001 **
White ethnic group	0.11	4.06	<0.0001 ***
Age	-0.06	-2.96	<0.001 **
Network distance	0.05	1.81	0.07
Service provision	0.07	2.88	<0.001 **
Travel time	0.01	0.60	0.55
Lambda (λ)	0.02 (***)		
Likelihood-ratio	3.86 (.)		
Wald statistic	3.74 (***)		

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

Like the multiple regression model, all independent variables with the exception of 'network distance' and 'travel time' were found to be statistically significant. The estimates of the spatial error model showed similar relationships as found through multiple regression, with 'IMD score', 'poor health' and 'age' the only variables to have a negative relationship with 'activity space area'. These variables are all related to both disposable income and physical mobility and therefore suggested that these factors, rather than those relating to the accessibility of HSR stores, constrained the activity and mobility of cardholders. The significant positive relationship with 'service provision' also highlighted the importance of access to public transport services in allowing for greater travel.

The significance of the ʎ value and the likelihood-ratio suggested that the inclusion of spatial error dependence improved the models when compared to a standard multiple regression model. The Wald statistic and the p-values also showed that a number of the explanatory variables were significant, and therefore had a significant impact on the average area of cardholders' activity spaces.

8.3.5 Distance travelled

'Distance travelled' referred to the average distance travelled to access an HSR store. Again, it was expected that demographic and socioeconomic variables that can constrain mobility, e.g. 'IMD score' and 'age', would have negative relationships with the average distance travelled to HSR stores. Additionally, it was expected that transportation variables would have positive relationships with 'distance travelled', with greater provision of public transport services and access to private transportation meaning that stores located a further distance from cardholders' registered addresses could be easily accessed.

8.3.5.1 Multiple regression

A multiple regression analysis was conducted, with 'distance travelled' as the dependent variable in the model. Table 8.22 shows the results of this analysis.

Table 8.22 'Distance travelled' – multiple regression results

Independent variable	Estimate	Standard error	t-value	p-value	Significance
IMD score	-0.13	0.05	-2.67	0.01	.
Poor health	-0.22	0.04	-5.27	<0.001	**
Car or van availability	-0.19	0.04	-4.20	<0.001	**
White ethnic group	0.05	0.03	1.99	0.05	.
Age	-0.07	0.02	-2.88	<0.001	**
Network distance	0.24	0.03	9.37	<0.0001	***
Service provision	0.10	0.02	4.09	<0.001	**
Travel time	0.14	0.03	5.67	<0.001	**
Adjusted R^2	0.15				
p-value	< 2.20e-16				

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

In terms of the average distance travelled to HSR stores, all independent variables were significant. Those variables related to public transport and the HSR store network; 'network distance', 'service provision' and 'travel time', were all found to have a positive relationship with the average distance travelled. 'IMD score', 'poor health', 'car or van availability' and 'age' were all found to have a significant negative relationship.

The negative relationships of 'IMD score', 'poor health' and 'age' with 'distance travelled' evidenced that those that lived in more deprived areas, areas with a higher proportion of residents with poor health orolder age were more likely to be constrained in terms of their ability to undertake long-distance trips. In terms of the impacts that each of these variables had on the average distance travelled, 'poor health' was found to have had the biggest impact, of -0.22. 'IMD score' had the next largest impact, of -0.13, and 'age' was found to have had a relatively small negative impact of -0.07. This suggested that whilst an increase in the average age of cardholders may have constrained mobility, this was not as significant as the impact of deprivation and health.

Additionally, 'car or van availability' was found to have a negative relationship with 'distance travelled'. This was unexpected, with access to private transportation generally leading to

greater accessibility of services, a greater choice of destinations, and often a lower cost of travel (Robinson et al. 2000). This negative relationship may instead be revealing of the limitations of the bus network. Access to private transportation allows cardholders more freedom and choice in the destinations that are visited and could therefore result in a lower average distance travelled if local shopping opportunities are not easily accessible via the bus network. This would result in public transport users being forced to seek retail opportunities further afield if these opportunities are more easily accessible via the bus network than local retail opportunities; leading to longer average trip distances.

Unsurprisingly, a significant positive relationship was found between the average network distance to HSR stores and the average distance travelled to these stores. For this analysis, the 'network distance' variable was primarily included to account for road network distance when analysing the remaining variables. After 'network distance', 'travel time' had the greatest impact on the average distance travelled to HSR stores. As 'travel time' is linked to distance, the 'network distance' variable was included to account for this. However, 'network distance' was calculated using the shortest available route, and thus not necessarily the route that bus services would take. This positive relationship between 'travel time' and 'distance travelled' may therefore instead be indicative of the proportion of this relationship that was not accounted for. An increase in the provision of bus services to an area also corresponded with an increase in the average distance travelled, and therefore indicated that a higher level of service allowed for greater freedom of travel and higher mobility levels.

8.3.5.2 Spatial autocorrelation

To check for spatial autocorrelation in the residuals of the multiple regression model a Moran's I test was conducted, with the results shown in Table 8.23.

Table 8.23 'Distance travelled' – Moran's Index

Dependent variable	Moran's I index	Expectation	Variance	z-value	p-value
Distance travelled (km)	0.09	-0.0024	0.0003	5.51	1.82e-08

The positive z-value suggested that residuals of similar values clustered together and the p-value meant that this clustering was statistically significant and therefore a spatial regression model should be implemented. To determine whether a spatial error or spatial lag model would produce the best fit, Lagrange Multiplier tests were conducted with the result shown in Table 8.24.

Table 8.24 'Distance travelled' – Lagrange Multiplier test results

Dependent variable	**LMerr**		**LMlag**		**RLMerr**		**RLMlag**	
	Value	***p-value***	***Value***	***p-value***	***Value***	***p-value***	***Value***	***p-value***
Distance travelled (km)	28.46	9.57e-08	34.61	4.03e-09	1.03	0.31	7.18	0.01

The Lagrange Multiplier tests suggested that a spatial lag model would produce the best fit, with analysis of the robust counterparts resulting in a higher value and a significant p-value compared to RLMerr results.

8.3.5.3 Spatial lag model

Like previous models, accounting for spatial autocorrelation in the residuals was hypothesised to reduce the magnitude of the relationships found between 'distance travelled' and the independent variables. The results of the spatial lag model are shown in Table 8.25, with 'distance travelled' implemented as the dependent variable.

Table 8.25 'Distanced travelled' – spatial lag model

Independent variable	Distance travelled					
	SLM			Impacts		
	Estimate	*z-value*	*p-value*	*Direct*	*Indirect*	*Total*
IMD score	-0.11	-2.44	0.01 .	-0.14	-0.02	-0.16
Poor health	-0.19	-4.79	<0.0001 ***	-0.19	-0.04	-0.23
Car or van availability	-0.16	-3.67	<0.0001 ***	-0.16	-0.03	-0.19
White ethnic group	0.04	1.64	0.10	0.04	0.01	0.05
Age	-0.06	-2.69	0.01 .	-0.06	-0.01	-0.07
Network distance	0.21	7.98	<0.0001 ***	0.21	0.04	0.25
Service provision	0.10	4.03	<0.0001 ***	0.10	0.02	0.12
Travel time	0.12	4.94	<0.0001 ***	0.12	0.03	0.15
ρ	0.05 (***)					
Likelihood-ratio	30.82 (***)					
Wald statistic	30.15 (***)					

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

Similar results were found for the spatial lag model as were found by the multiple regression model, with the exception of 'white ethnic group' which was no longer statistically significant. The strength and direction of the relationship between each independent variable and 'distance travelled' were similar to those found in the multiple regression model. The demographic and socioeconomic variables ('IMD score', 'poor health' and 'age') had significant negative relationships with 'distance travelled'. This suggested that populations residing in more deprived areas and areas with a higher proportion of residents with poor health were generally more constrained in how far they were able to travel. Unsurprisingly, 'network distance' had the largest impact on 'distance travelled' with the further cardholders reside from HSR stores resulting in greater average distances travelled to access these stores. 'Car or van availability' had the largest negative impact, with a decrease in the number of residents with access to private transportation resulting in an increase in the average distance travelled to stores. Like all previous models, the rho, likelihood-ratio and Wald statistic were significant.

8.3.6 Activity space compactness

'Activity space compactness' referred to the distribution of each cardholder's activity space, i.e. how dispersed the HSR stores were that were visited by each cardholder. This can give an indication of retail behaviour, with compact activity spaces a possible indicator of constrained mobility. Conversely it can also be indicative of a good provision of local retail opportunities. It was therefore hypothesised that variables related to deprivation and health ('IMD score', 'poor health' and 'age') would have a positive relationship with compactness; indicating more compact activity spaces for cardholders residing in more deprived areas and areas with a high proportion of people with poor health. Those variables related to the accessibility of the HSR store network ('network distance' and 'travel time') were also hypothesised to have a positive relationship with compactness, with the greater time or distance to access HSR stores resulting in fewer stores being visited and therefore more compact activity spaces.

8.3.6.1 Multiple regression

Table 8.26 shows the results of the multiple regression model.

Table 8.26 'Activity space compactness' – multiple regression results

Independent variable	Estimate	Standard error	t-value	p-value	Significance
IMD score	-0.04	0.05	-0.75	0.45	-
Poor health	-0.06	0.04	-1.54	0.12	-
Car or van availability	0.06	0.05	1.32	0.19	-
White ethnic group	0.10	0.03	3.67	<0.001	**
Age	-0.05	0.02	-2.00	0.06	-
Network distance	0.06	0.03	2.18	0.03	.
Service provision	-0.12	0.02	-4.81	<0.001	**
Travel time	0.15	0.03	5.73	<0.001	**
Adjusted R^2	0.11				
p-value	< 2.20e-16				

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

Of the independent variables included in the model, 'IMD score', 'poor health', 'age' and 'car or van availability' were not found to have significant relationships with the compactness of cardholders' activity spaces. 'IMD score', 'poor health' and 'age' were predicted to have a positive relationship with compactness, with previous analyses suggesting that increased deprivation, age and poor health can constrain activity spaces. This lack of a significant relationship suggested that the makeup of cardholders' local retail options had more of an influence on how compact their activity spaces were.

'White ethnic group', 'network distance' and 'travel time' were shown to have a positive relationship with compactness. This meant that as the proportion of the LSOA population from a white ethnic background increased, so did the compactness of their activity spaces. In terms of 'network distance' and 'travel time', an increase in the distance or time taken to travel to HSR stores resulted in an increase in the average compactness of cardholders' activity spaces. This meant that the longer it took to access retail opportunities, the more 'local' and compact a cardholder's activity space was likely to be.

The negative relationship between 'service provision' and 'activity space compactness' meant that an increase in the number of buses servicing an area resulted in a more dispersed average activity space. This may be the result of areas with a high level of service provision benefitting from public transport that service a number of different destinations, and therefore providing more choice in terms of the opportunities that can be easily accessed.

The adjusted R^2 value of 0.11 meant that 11% of the variation in 'activity space compactness' could be attributed to variations in the independent variables, and the p-value meant that the model was statistically significant. This R^2 value was relatively low, which again could be down to the aggregation of variables to LSOA level. The hypothesised relationships between compactness and the independent variables were less clear than previous models due to the impact that local shopping facilities in the areas in which cardholders reside can have on the dispersion of shopping trips, and therefore, as this was not accounted for in the model, this may also explain the low R^2 value.

8.3.6.2 Spatial autocorrelation

A Moran's I test was conducted on the residuals of the multiple regression model to check for the presence of spatial autocorrelation and determine whether a spatial regression model should be applied.

Table 8.27 'Activity space compactness' – Moran's Index

Dependent variable	Moran's I index	Expectation	Variance	z-value	p-value
Activity space compactness	0.39	-0.0024	0.0003	23.80	<2.2e-16

The positive z-value and significant p-value shown in Table 8.27 meant that the residuals of the multiple regression model exhibited spatial clustering and therefore a regression model that accounts for spatial autocorrelation should be conducted. Table 8.28 shows the results of a Lagrange Multiplier test to determine whether a spatial lag or spatial error model would produce the best fit.

Table 8.28 'Activity space compactness' – Lagrange Multiplier test results

	LMerr		LMlag		RLMerr		RLMlag	
Dependent variable	***Value***	***p-value***	***Value***	***p-value***	***Value***	***p-value***	***Value***	***p-value***
Activity space compactness	554.33	<2.2e-16	563.76	<2.2e-16	4.44	0.04	13.87	0.00

As both the LMerr and LMlag were significant, the robust counterparts were analysed. The RLMlag had a higher value and a more significant p-value and therefore it was determined that a spatial lag model should be conducted.

8.3.6.3 Spatial lag model

Table 8.29 shows the results of the spatial lag model.

Table 8.29 'Activity space compactness' – spatial lag model

Independent variable	Activity space compactness					
	SLM			Impacts		
	Estimate	*z-value*	*p-value*	*Direct*	*Indirect*	*Total*
IMD score	-0.03	-0.82	0.41	-0.04	-0.04	-0.08
Poor health	-0.04	-1.12	0.26	-0.04	-0.05	-0.09
Car or van availability	0.04	0.94	0.35	0.04	0.04	0.08
White ethnic group	0.05	2.13	0.03 .	0.05	0.06	0.11
Age	-0.07	-3.48	0.07	-0.07	-0.08	-0.15
Network distance	0.02	0.81	0.42	0.02	0.02	0.04
Service provision	-0.08	-3.73	<0.001 **	-0.08	-0.09	-0.17
Travel time	0.09	4.19	<0.0001 ***	0.10	0.11	0.21
ρ	0.14 (***)					
Likelihood-ratio	450.39 (*)					
Wald statistic	573.03 (*)					

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

Unlike the multiple regression model, only 'white ethnic group', 'service provision' and 'travel time' were found to have significant relationships with the compactness of cardholders' activity spaces. Areas with higher proportions of white residents were found to have more compact activity spaces. Higher service provision was again found to decrease compactness and lead to more dispersed activity spaces; however, the strength of the direct impact was less than found through multiple regression. This highlighted how provision of public transport can allow for greater travel and more diverse destinations to be reached. 'Travel time' was found to have the greatest, both direct and total, impact on 'activity space compactness'. This suggested that areas with more efficient public transport services allowed for more dispersed activity areas and therefore more choice in destinations that could be accessed.

Like the previous models, the significant rho, likelihood-ratio and Wald statistic meant that at least one of the independent variables included in the model was significant and the inclusion of the lagged values improved the model.

8.3.7 Stores visited

'Stores visited' referred to the average number of stores visited by cardholders. This variable differed from the transaction frequency variable as it referred to the number of different stores visited by cardholders. It can therefore be seen as an indicator of retail behaviour in terms of the capacity of individuals to access a number of retail opportunities. Like previous models, socioeconomic and demographic variables that have been shown to constrain mobility, such as 'IMD score' and 'age', were hypothesised to result in fewer stores being visited by cardholders. Those related to the accessibility of the HSR network ('network distance' and 'travel time') were hypothesised to have a negative impact on the number of stores visited. This was due to the greater cost of undertaking shopping trips predicted to result in fewer trips undertaken, and therefore stores visited.

8.3.7.1 Multiple regression

Table 8.30 shows the results of a multiple regression analysis implemented with 'stores visited' as the dependent variable.

Table 8.30 'Stores visited' – multiple regression results

Independent variable	Estimate	Standard error	t-value	p-value	Significance
IMD score	-0.15	0.04	-3.70	<0.001	**
Poor health	-0.21	0.04	-5.86	<0.001	**
Car or van availability	0.18	0.04	4.64	<0.001	**
White ethnic group	0.17	0.02	7.25	<0.001	**
Age	0.03	0.02	1.15	0.25	-
Network distance	-0.12	0.02	-5.31	<0.001	**
Service provision	-0.14	0.02	-0.67	0.50	-
Travel time	0.05	0.02	2.35	0.02	.
Adjusted R^2	0.35				
p-value	< 2.20e-16				

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

All independent variables, with the exception of 'age' and 'service provision', were found to have a significant impact on the number of stores visited by cardholders. Similarly to 'activity space area', increases in 'IMD score' and 'poor health' were found to result in a decrease in the average number of stores visited. In contrast, an increase in 'car or van availability' was found to significantly increase stores visited. As activity space area was calculated using the locations of stores visited, it is unsurprising that similar patterns were found. The coefficient estimates for these variables suggested that living in areas with high access to private transportation had the greatest impact on the number of stores that were visited, with the level of deprivation in an area also having a significant impact.

Unlike the 'activity space area' model, 'network distance' to HSR stores had a significant impact on the number of stores visited, with an increase in the average distance resulting in a decrease in the average number of different stores visited. This was expected, with increased distance related to increased cost of travel; highlighting how the store network influences the retail behaviour of cardholders.

'White ethnic group' was shown to result in an increase in the number of stores visited, meaning an increase in the proportion of the population from a white ethnic background resulted in an increase in the average number of stores visited by cardholders. This suggested that those from white ethnic backgrounds may have more diverse retail behaviours, transacting at larger number of stores, than those of other ethnic groups. Additionally, this was indicative of the tendency of individuals from other ethnic backgrounds to shop in specialist stores which spring up in areas where such groups put down roots (Hamlett et al. 2008).

The R^2 value of 0.35 was the highest of all models and meant that 35% of the variation in the average number of stores visited by cardholders could be explained by variations in the independent variables. The p-value of < 2.20e-16 also meant that the overall model was statistically significant.

8.3.7.2 Spatial autocorrelation

A Moran's I test was ran on the residuals of the multiple regression model to check for the presence of spatial autocorrelation. The results of this test are shown in Table 8.31.

Table 8.31 'Stores visited' – Moran's Index

Dependent variable	Moran's I index	Expectation	Variance	z-value	p-value
Stores visited	0.47	-0.0024	0.0003	29.10	<2.2e-16

The results shown in Table 8.31 suggested that spatial autocorrelation was present in the residuals and therefore a spatial regression model would produce the best fit. To determine which spatial regression model should be implemented, Lagrange Multiplier tests were conducted.

Table 8.32 'Stores visited' – Lagrange Multiplier test results

Dependent variable	LMerr		LMlag		RLMerr		RLMlag	
	Value	***p-value***	***Value***	***p-value***	***Value***	***p-value***	***Value***	***p-value***
Stores visited	830.57	<2.2e-16	644.69	<2.2e-16	186.01	<2.2e-16	0.13	0.71

Table 8.32 shows the results of these Lagrange Multiplier tests. As both of the LMerr and LMlag values were statistically significant, the robust counterparts were analysed with the RLMerr producing a statistically significant p-value. This meant that a spatial error model should be conducted rather than a spatial lag model.

8.3.7.3 Spatial error model

Like previous models, the presence of spatial autocorrelation in the residuals of the multiple regression model likely increased the magnitude of the coefficient estimates. The estimates found through the spatial error model were therefore predicted to be of smaller magnitude, but the overall model was predicated to produce a better fit. The results of this spatial error model are shown in Table 8.33.

Table 8.33 'Stores visited' – spatial error model

Independent variable	Stores visited		
	Estimate	***z-value***	***p-value***
IMD score	-0.17	-4.62	<0.0001 ***
Poor health	-0.10	-3.48	<0.0001 ***
Car or van availability	0.25	7.45	<0.0001 ***
White ethnic group	0.19	7.43	<0.001 **
Age	-0.04	-2.36	0.02 .
Network distance	-0.12	-4.21	<0.0001 ***
Service provision	0.01	0.49	0.63
Travel time	0.03	1.38	0.17
Lambda (λ)	0.16 (***)		
Likelihood-ratio	634.43 (***)		
Wald statistic	827.47 (***)		

Significance Coefficients: *** <0.0001, ** < 0.001, * < 0.01, . < 0.05

Unlike the multiple regression model shown in Table 8.30, all independent variables were found to be significant with the exception of 'service provision' and 'travel time'. 'IMD score', 'car or van availability', 'white ethnic group' and 'age' were found to have stronger relationships with 'stores visited' when considering spatial autocorrelation. 'Car or van availability' had the largest impact of 0.25, suggesting that high private transportation availability in an area greatly influenced the number of stores that were accessible to cardholders. 'White ethnic group' was also found to have a significant positive relationship with stores visited, meaning that areas with a higher proportion of residents from a white ethnic background generally had a higher average number of stores visited. Again, this may be due to the retail behaviours of different ethnic groups, with ethnic minority groups more likely to shop at specialist stores and therefore visit fewer HSR stores. 'IMD score' was found to have the largest negative impact on 'stores visited'. This evidences the idea that deprivation can lead to constrained mobility and restricted activity spaces. Unlike the results of the multiple regression model, 'age' was found to have a significant negative impact on 'stores visited'; suggesting that those of older age may be restricted in the retail opportunities that they are able to access.

Like the previous model, the lambda value and the likelihood-ratio meant that the inclusion of spatial error dependence improved the models, and the Wald statistic and p-values also showed that several of the explanatory variables were significant.

8.4 Discussion and conclusions

This section firstly discusses the results presented in Section 8.3 in relation to the mobility and activity of the older population, as well as the previous insights generated throughout this thesis. Secondly, the methods employed in this chapter are critically examined in terms of their limitations, and potential future research directions are outlined.

8.4.1 Transport provision and accessibility

As discussed in Chapter 2, inequalities in the provision of public transport and the accessibility of both public transport services and key facilities can impact wellbeing and reduce social cohesion (Lupton et al. 2015). Public transport provision plays a key role in the sustainability of urban areas and the quality of life of residents. However, even with high service provision, poor links to opportunity destinations can contribute to social isolation and exclusion and prevent full participation in life-enhancing opportunities (Lucas et al. 2019).

The distribution of both service provision and travel times to retail opportunities, shown in Figures 8.3 and 8.7, varied significantly throughout the study area. This would be expected in any large conurbation, with areas having different transport needs due to varying populations and their specific needs. Including these variables in each of the models allowed for a 'real-life' case study of how public transportation impacts the activity of individuals. By accounting for demographics, socioeconomics and the HSR store network, the influence of both the provision of bus services and the efficiency of these services on the retail behaviours of cardholders could be explored.

8.4.1.1 The impact of public transportation on retail activity

The role of public transport in the activity patterns of individuals and population groups has been widely researched, however is often either focused solely on commuting populations or hindered by the lack of information relating to trip purpose. Traditional travel behaviour data collection methodologies have therefore often not captured the *why* of demand (McCray and Brais 2007). By conducting these analyses on behavioural variables and spatial activity patterns

derived from the HSR transaction data rather than the smart card data, this *why* of demand was captured with all records the result of a retail transaction and therefore a shopping trip. Additionally, the product information contained within the HSR data allowed for deductions about the kinds of trips that were being undertaken, for example necessary trips to purchase essential products or recreational shopping trips.

The results presented in this chapter found that both public transport service provision and the accessibility of HSR stores via the bus network were found to have statistically significant relationships with a number of consumption and activity variables. In terms of the impact that service provision was found to have on the spatial activity patterns of cardholders, higher public transport service provision was found to correspond with larger and more dispersed activity spaces (Tables 8.21 and 8.29) and a higher average distance travelled to HSR stores (Table 8.25). Accounting for private transport availability, through the 'car or van availability' variable, and the HSR store network, through the 'network distance' variable, meant that the impact of service provision could be largely isolated from other influencing factors. These results therefore substantiated the hypothesis that access to reliable public transport services can increase the mobility of the older population and allow for a wider range of opportunities to be accessed.

The large variations in service provision shown in Figure 8.3 may therefore be of concern to transport authorities, with low levels of service in many suburban and rural areas potentially leading to the constrained activity spaces of older residents in these areas. Quantifying the relationship between service provision and activity behaviour allowed for the impact of changes to bus scheduling and routing to be understood. Additionally, with questions surrounding the suitability of the ENCTS scheme (see Section 2.2.3.1), the impact of bus service provision evidenced the benefits of this scheme for maintaining high levels of mobility within the older population, with knock-on implications for social inclusion and wellbeing (Spinney et al. 2009).

Similarly, the accessibility of HSR stores via the bus network was also found to have a significant impact on a number of retail activity variables. The less accessible the HSR store network, and therefore the higher the travel time to these stores, the less frequently transactions were made (Table 8.9) and the more compact activity areas were (Table 8.28). These results highlighted the impact that public transport services can have on the activity spaces of individuals and the role these services can play in determining the opportunities that individuals can reach. The transport-disadvantage caused by a lack of access to efficient public transport services can lead to increased social exclusion and decreased wellbeing, in particular for populations that do not have access to key services and facilities in the local areas where they live (Lucas et al. 2019).

In addition to spatial activity patterns, consumption patterns were also found to be impacted by the accessibility of the HSR store network. Differences in the types of products purchased and the value of transactions provided insights into the types of trips that were being undertaken, i.e. if these trips were essential or instead formed part of a recreational trip. The average value of transactions had a significant positive relationship with the time taken to travel to stores via the bus network, and the proportion of essential products purchased had a significant negative relationship. This meant that the more time taken to access HSR stores, the more discretionary, higher value purchases that were made. This gives insight into the retail behaviours of the older population, suggesting that higher bus travel times lead to a reduction in regular, essential shopping trips.

Disadvantage arising from this is contingent on the retail mix in the areas in which cardholders reside. Those that have access to a wide range of local retail opportunities are less likely to be negatively impacted by the inaccessibility of other retail opportunities. The relationships between retail activity and travel times do, however, suggest that those that reside in areas with high travel times to retail opportunities may be forced to purchase essential products within their local areas, which can often be more costly and the range of products available can be limited (Williams and Hubbard 2011). In the case of this study, areas that suffer from high travel times to 'higher order' retail opportunities, such as out-of-town retail parks and shopping centres, in combination with poor local retail provision are likely to be disadvantaged in terms of the range of retail opportunities that are able to be accessed.

This chapter explored the potential for non-transport related consumer data to provide insight into the mobility of the older population and demand for services. The results of these analyses were promising; demonstrating how differences in retail behaviour can give insight into the efficiency of public transport services and the ability of the network to meet mobility needs and demands. Quantifying the impact of public transport on activity patterns is important for local transport authorities aiming to understand the influence that changes to current services and routes, or policy changes may have on the activity of different population groups. Insights into consumption demonstrated how the use of both retail data and public transport data can help both transport authorities and retailers to understand the utility of current services and the function that different facilities serve.

8.4.1.2 Private transport availability

Section 2.2.2.1.1 outlined the impact that access to private transportation can have on mobility, and also discussed how losing the ability to drive with age can impact wellbeing and social inclusion (Musselwhite et al. 2015; Mackett 2013). Focusing on the 'car or van availability'

variable, it was clear that access to private transportation had a significant impact on retail activity. An increase in the proportion of the population with access to private transportation was found to correspond with an increase in 'activity space area', 'stores visited', 'transaction frequency' and 'basket value', and a decrease in 'distance travelled'. The positive relationship with the average area of cardholders' activity spaces, the number of different stores visited, and the frequency of transactions highlighted the importance of access to private transportation in allowing for high levels of mobility. In turn, this also emphasised how those that rely on public transport services may be constrained in terms of their spatial activity patterns and the retail opportunities that they can access.

Current transport policies are often focused on encouraging individuals to move away from private transportation towards public transport services, both for environmental reasons and to ease congestion on roads. The significant relationship between private transport access and a number of retail activity variables, such as 'activity space area' and 'stores visited', suggested that current public transport services may not be able to provide the levels of mobility that private transport can. The negative relationship between private transport access and the average distance travelled to HSR stores suggested that public transport services may not be providing adequate access the most 'local' and convenient stores, resulting in public transport users being constrained by the current, predetermined transport network. Encouraging a move towards public transport would therefore require transport authorities to ensure that mobility needs are met and that individuals are able to maintain the high levels of mobility currently afforded by private transportation.

8.4.2 Demographic and socioeconomic factors

In addition to those variables related to public and private transportation access, the demographic and socioeconomic characteristics of cardholders were also found to have statistically significant relationships with retail activity.

Throughout this thesis, the concept of deprivation and the impact that this can have on mobility has been widely discussed. To explore this further, the spatial regression analyses undertaken included 'IMD score'; a raw score given to each LSOA based on their level of deprivation. Contrary to what may have been expected, 'IMD score' was not found to have a significant impact on the majority of dependent variables, with only the relationships with 'activity space area', 'stores visited' and 'distance travelled' found to be significant. These variables were all found to have a negative relationship with deprivation, suggesting that those that reside in deprived areas may be constrained in terms of the retail opportunities that they are able to access. The idea that poorer groups of the population lack access to a wide range of retail

opportunities has been widely researched (Williams and Hubbard 2011). Quantifying this relationship and showing that those residing in more deprived areas suffer from disadvantage in terms of their spatial activity patterns further evidences these insights. This disadvantage can lead to a reduction in quality of life, with less opportunity for social interaction and increased social exclusion (Williams and Hubbard 2011).

Contrary to what was hypothesised, the consumption variables ('transaction frequency', 'basket value' and 'essential products') were not found to have a significant relationship with deprivation. This suggested that those residing in deprived areas were constrained in terms of their spatial mobility patterns, however, did not exhibit significantly different consumption patterns to those that were less deprived. One reason behind this may be the function of the HSR, with a large proportion of products classified as essential meaning that these may not be subject to the differences in consumption expected between those with more or less disposable income. The availability of data may have hindered this analysis, with prescriptions purchased at HSR stores not recorded in the data. As these would also have been classified as 'essential', the percentage of essential products purchased by cardholders may have been higher if these had been taken into account.

The 'poor health' variable included in the models referred to the proportion of the population recorded as having 'bad' or 'very bad' health. The negative relationships found with 'activity space area', 'stores visited', 'transaction frequency' and 'distance travelled' therefore suggested that those with poor health exhibited spatially constrained patterns of shopping behaviour. Although this is unsurprising, with the relationship between health and mobility widely researched (Boniface et al. 2015), the significant relationship found between a number of spatial activity variables and 'poor health' highlighted the role that health can play in the mobility of individuals. This emphasised the need for transport services to have the capacity to meet the needs of different groups of the population, in addition to the working age, commuting populations that public transport systems are often designed around (Wang 2007).

The impact of increasing age on the mobility of the older population was widely discussed in Section 2.2.2.2 and was found to limit mobility and increase the risk of social exclusion. Including the 'age' variable in the regression models allowed for this to be explored in terms of how an increase in age impacts mobility and activity patterns. 'Age' was found to have a significant impact on the majority of dependent variables, with an increase in the average age of cardholders found to have a positive impact on 'transaction frequency' and 'essential products' and a negative impact on 'activity space area', 'stores visited', 'basket value' and 'distance travelled'. The majority of these impacts were relatively small, with the exception of the impacts on 'basket value' and 'essential products', with total impacts of -0.25 and 0.30

respectively. These impacts meant that an increase in age influenced the consumption behaviour of cardholders, with transactions more likely to be of lower value and consist of a larger proportion of essential products. Like with deprivation and health, the negative relationship with 'activity space area', 'stores visited' and 'distance travelled' suggested that as cardholders age, their spatial mobility may become restricted. Section 2.2.3 discussed the suitability of current bus services and the ENCTS scheme, and the potential of alternative transport services, such as demand responsive transport, to better meet the needs of those that are unable to maintain high levels of mobility through traditional services. The evidence provided in this chapter of the impact on spatial mobility that deprivation, health and age can have suggested that alternative transport services tailored to meet the needs of those that are most at risk of transport disadvantage may be more effective in meeting the unique mobility needs of these populations.

8.4.3 Method limitations

Although this chapter presented promising insights into the utility of alternative consumer data for transport-related research, there were a number of limitations with the methods that were employed and the assumptions that were made throughout. Many of these limitations were related to the quality and completeness of the HSR data, which was assessed in detail in Section 7.2. This section therefore discusses the specific limitations of the methods employed in this chapter.

The primary limitation of the methods employed in this chapter was the aggregation of each variable to LSOA level. Although this was conducted to allow for the linkage of LSOA-level census data to the HSR data and to protect the anonymity of cardholders, issues of ecological fallacy can arise. Ecological fallacy occurs when inferences about the nature of individuals are deduced from inferences about the group to which those individuals belong (Lavrakas 2008). In this case, this could arise if deductions were made about individual cardholders based on the LSOAs to which they were registered. The conclusions deduced throughout this chapter did not make any inferences about individuals, but rather about the general mobility and activity patterns emanating from neighbourhoods and the factors that may impact these aggregate patterns. Future research may wish to explore these individual activity patterns, and therefore alternative statistical methods, such as multilevel models or methods using qualitative data, could be employed.

This chapter utilised HSR store locations as proxies for retail opportunities. Although the HSR stores in the study area were located on many high streets and in all urban centres, these locations were unlikely to encapsulate all retail opportunities available throughout the study area. Similarly, HSR transactions only captured one dimension of a cardholder's transaction

records and were unlikely to be fully representative of their transaction behaviour across all retail stores. Individual choice and the role that HSR stores play in an individual's full spending habits could not be accounted for. This restriction to only one retailer is one of the disadvantages of novel consumer data when compared to traditional data collection methods, such as surveys and questionnaires, which have the ability to capture an individual's full retail behaviour. This is a trade-off that is often encountered when working with large consumer data; capturing only a dimension of all cardholders' retail behaviour rather than gaining a full view of a select number of individuals' behaviour. To account for this, only 'active' cardholders were included in the analysis (see Section 7.2.3.1), however a more thorough analysis of the types of cardholders present in the data, for example those that undertake regular, essential transactions may have improved this analysis.

When calculating travel times to HSR stores, there were a number of assumptions that were made. The same 400 meter bus stop service areas used in Chapter 6 were utilised for this analysis, however the average distance that one would be willing to walk to access a public transport stop may not be equivalent to the distance one would walk to an HSR store. Although this may have resulted in the number of bus stops that service an HSR store being underestimated, there was likely to be minimal difference in the travel times between subsequent bus stops. Therefore, potentially not including all bus stops that may be used to access an HSR store was unlikely to have had a significant impact on the aggregate travel times.

Additionally, this analysis could have been expanded to include more than one transfer between services. However, due to the computational cost associated with this, it was deemed unnecessary for the scope of this study. As this analysis focused on minimum travel times, transferring between two or more services whilst considering wait times at stops, in many cases would not have resulted in shorter travel times. The aggregation of these travel times to LSOA level would also have minimised the impact that any outliers had on the overall results of the regression models. The unavailability of data from other modes of transport may also have limited the accuracy of the estimated travel times, with some stores that have high bus travel times being easily accessible by metro or train services. Inclusion of these data in future analyses would undoubtedly improve the overall accuracy of the public transport travel time distributions shown in Figure 8.6.

In regard to the regression analyses implemented, the spatial autocorrelation detected in the residuals of all multiple regression models meant that the initial results produced by these models were likely to be biased as the assumption that the outcomes were independent of one another was violated. This meant that the initial coefficient estimates were not valid and further steps had to be undertaken to detect and address spatial autocorrelation. Although the presence

of spatial autocorrelation in the residuals of the multiple regression models meant that the results of these models were likely to be bias, detecting this through a Moran's I test and addressing it by conducting spatial regression models meant that the results of these models were more robust and replicable.

8.4.4 Future research potential

This chapter highlighted the benefits of consumer data; demonstrating the potential of non-transport consumer data to provide insights into the current utility of the public transport network and the mobility and activity patterns of the older population in the West Midlands. The ongoing growth in the collection of these data and their increasing availability for research purposes present scope to expand upon the analyses and insights presented in this chapter.

As outlined in Section 8.4.3, one of the limitations of these analyses was the availability of data from only one retailer. The inclusion of data from additional retailers or facilities that service a different need, such as supermarkets or leisure facilities, could improve the analyses and widen the scope of the research. Additionally, as discussed in Chapter 7, HSR transaction data available over a longer temporal scale would have opened up the possibility of conducting analyses into the impact that changes in public transport provision and travel times have on the retail activity of cardholders.

This chapter highlighted the importance of not only service provision, but also the public transport network providing efficient services to the locations that individuals wish to travel to. Future research could expand upon this by combining demand extracted from the smart card data with demand obtained from the HSR data. This could highlight areas of apparent high demand from both a public transport and a retail perspective and identify the areas where these demands do not align, and therefore where there may be a lack of, or oversupply of, public transport provision.

For the analyses presented in this chapter, public transport provision and travel times were aggregated to LSOA level to allow for the inclusion of census data. Future research could forgo this aggregation; analysing service provision and travel times at a more granular scale, such as bus stop or service area level. This could also allow for the inclusion of more detailed travel time and accessibility measures at the bus stop level, for example delays on the bus network, the availability of shelters and seating at the stop, and the accessibility of the stop itself.

Although this chapter presented a thorough analysis, and highlighted strong causal relationships between retail activity and public transport, there is scope to expand this further. Much of this future research relies on the availability of additional and complementary data, whether this is

more granular transport network data or additional consumer data. The increasing collection and availability of these data are promising for future research in this area.

Chapter 9 – Discussion, applications and research prospects

9.1 Introduction

Current research into transport, mobility and the relationship with social exclusion and disadvantage primarily relies upon traditional data collection methods, such as questionnaires and travel surveys (Long and Thill 2015). These 'traditional' data are often hindered by limited sample sizes and the lengthy time taken for data to be collected and subsequently released for research purposes. With the possibility that the 2021 Census will be the last (BBC News 2020), there is heightened emphasis on the importance of research into the potential for consumer data to supplement, and eventually replace, these traditional data sources.

This thesis aimed to explore the potential utility of large consumer data for supplementing traditional data and furthering the understanding of the mobility and activity patterns of the older population. Through the use of both smart card travel data and loyalty card transaction data, insights were generated into the use of the bus network, changes in the mobility patterns of the older population over the study period, and the effectiveness of the current network to meet mobility needs and travel demands. The utility of novel consumer data and non-transport related data to gain insight into the broader dimensions of activity, which are often unable to be captured by traditional transport and survey data (Kieu et al. 2014), underscores the importance of research into this area and the benefits of data triangulation; exploring the same phenomenon through multiple data sources.

This chapter seeks to consolidate the findings of this thesis. The applications of this research and the potential implications the results could have are discussed in terms of the practical implementation of consumer data for this type of research. Additionally, the insights generated into the mobility and activity patterns of the older population and the effectiveness and efficiency of the current transport network in the West Midlands are summarised. Following this, reflections are made on the data and methods that were applied throughout the thesis, and the challenges and limitations that were encountered when working with these data and implementing these methods. Finally, paths for future research are discussed.

9.2 Applications and implications

9.2.1 Data cleaning and preparation processes

This thesis provided many valuable insights into the utility of the current transport system, the mobility of the older population, and the potential value of alternative consumer data sources.

The first of these applications related to the inherent data quality and representativeness issues of consumer data. There has been a fundamental lack of understanding of these issues due to the novelty of these data, with no standard practices for dealing with data quality issues (Kitchin 2014). This research highlighted issues of uncertainty, bias and errors through analysis of two consumer datasets. These included missing data, discrepancies between variables, for example mismatched 'card type' and 'age group' records in the smart card data, and issues of representativeness. This thesis presented methods of identifying missing data and errors and justified the choice of omitting these from subsequent analyses rather than inferring any missing records. It is hoped that this work can aid future researchers aiming to implement these data and provide insight into the types of issues, and the methods of dealing with these, that may be encountered.

In terms of the representativeness of data, although the smart card data was relatively representative of the older population due to the high uptake of the scheme, the loyalty card data represented only a proportion of the wider population in the study area. Discussions of this in Chapters 4 and 7 highlighted the importance of ascertaining representativeness prior to undertaking analyses to determine to what extent insights can be applied beyond the study population. In the case of this thesis, the focus was on including the demographic and socioeconomic characteristics of the populations in the analyses. This was achieved by including both the known characteristics that were recorded in the data, as well as inferring characteristics through comparisons with census data. This method was implemented to avoid ecological fallacy and ensure that results from both the smart card and loyalty card data were comparable. The research undertaken in this thesis evidenced the increased importance of accounting for different representations in the data when utilising multiple consumer datasets.

In addition to data quality issues, methods had to be devised to process and prepare the data for subsequent analyses. There has been little research conducted on smart card data that spans the time period that this study does (7 years), with the majority of research conducted on data spanning only days, weeks or months. This thesis presented methods for dealing with data that covers longer periods; principally, adjusting the data to account for population changes over the study period and highlighting the need to consider the longitudinal nature of the data. As consumer data becomes more widely collected and used in research, the time periods for which data is available will increase (Morency et al. 2015). Accounting for wider changes in the population, for example neighbourhood composition, will become an increasingly important first step in the analysis of these data.

These insights emphasise the need to develop formal frameworks that outline and benchmark the applications of various forms of consumer data and the processes required to prepare these

data for analysis. The implementation of standard practises for the application of consumer data could provide both academia and industry with a consistent point of reference, where potential sources of data error and bias can be identified and addressed. The creation of these standard practices, for example the inference of cardholder characteristics to ascertain representativeness, would widen the scope for consumer data to be implemented for research, ensure that methodologies are consistent, and increase the feasibility of these data supplementing or replacing traditional data sources.

9.2.2 Insights into mobility and transport

The move away from the sole use of traditional data sources for mobility and transport insights has brought about a need to understand the specific role that consumer data will play in this (Stillwell 2016). This thesis implemented smart card data and loyalty card data to develop a new perspective on the current usage of the bus network, changes in travel behaviour, and the wider mobility and activity patterns of the older population. One the aims of this thesis was to understand how this smart card data and loyalty card data can provide insights into mobility and be used to supplement traditional transport data. This section outlines the successes in achieving this aim, as well as how this research could be expanded upon in the future.

This thesis firstly aimed to adjust the trends in bus boardings by changes in the eligible population; due to natural population change, migration and the increase in the age of eligibility for concessionary travel. The increase in the age of eligibility had not been considered by the local transport authority as having a significant impact on the number of boardings made by older concessionary travellers. However, the majority of LSOAs experienced a decline in the number of residents eligible for concessionary travel over the study period. Adjusting for these changes demonstrated that the decline in bus usage by older concessionary travellers was less substantial than that shown by non-concessionary bus users. Although both populations exhibited a decline in boardings, evidencing that this was not a behavioural change unique to the older population was a significant insight.

One of the motivations behind this study was the need to understand the recorded decline in the use of bus services by the older population. Limitations in the available data, for example granular boarding data for other modes of transport, restricted the potential for this research to explore the reasons behind this decline. However, this thesis was able to identify the common characteristics of the cardholders that had decreased their use of the bus network and identify groups that may be at risk of transport disadvantage and exclusion. The cardholder clustering presented in Chapter 5 showed that those that had decreased bus usage were most likely to reside in areas of high deprivation and the areas surrounding urban centres. Individuals and

households that are deprived are at an increased risk of transport disadvantage as they often lack access to private modes of transport and therefore are more likely to rely on the public transport network (Lucas 2012). These inner city and urban areas were also shown to have a high proportion of cardholders who were most active on the bus network; found in Clusters 3 and 4. Analyses presented in Chapter 8 found that these areas, whilst benefitting from high service provision, also tended to have less efficient bus services with higher travel times to retail opportunities after adjusting for distance. Cardholders residing in suburban and rural areas, for which bus services were found to be more efficient, were found to have been more likely to have increased their bus usage. This relationship between bus usage and the efficiency of the bus network, whilst not a novel concept, is important for transport authorities and planners to consider alongside demand metrics to ensure that mobility needs and demands are met (Mohamed et al. 2014).

Comparing these results to the trip purpose inferences suggested that at least a proportion of this decline may have been the result of a decline in the number of retail trips undertaken using bus services. The higher travel times in areas with high proportions of cardholders exhibiting declining bus usage also suggested that this may be due to a decrease in the efficiency of the transport network for reaching these opportunities. The propensity of transport networks to be designed around meeting the needs of commuting populations (Battellino 2007), and the shift of retail opportunities away from high streets and town centres to out-of-town shopping centres and retail parks (Scheiner 2006) may have contributed to this change in behaviour. The hotspots generated for shopping trips support this idea of a changing retail landscape. Shopping hotspots were found to have moved away from main urban centres and high streets, towards smaller high streets and out-of-town retail areas. Similar results were not found for leisure trips, which remained largely within urban centres. This suggested a change in behaviour unique to shopping trips and therefore evidenced a change in the retail landscape in the study area.

Additionally, the rise of e-commerce may have resulted in a behavioural change among some cardholders with shopping transactions increasingly being conducted online and therefore a decreasing need for these trips to be undertaken via the public transport network. The propensity to partake in e-commerce varies significantly within the older population (ONS 2019), which may result in groups of the population that do not engage with online retailing being unable to access retail opportunities if the trend of retailers moving away from local high streets and town centres continues. The unique mobility needs of the older population; with often no need to access job opportunities and differing temporal constraints to the working age population, mean that current public transport services often struggle to meet these demands. The insights generated in this thesis suggest that the current transport network may not have adapted to adequately service the changing retail landscape, and subsequently the changing demands of

different users, and therefore may leave certain groups of the older population vulnerable to transport disadvantage.

Spatial activity patterns are explored throughout this thesis, with activity hotspots in Chapter 5, trip purpose hotspots in Chapter 6, and spatial store visiting patterns in Chapters 7 and 8. Differences in the activity hotspots of each bus usage cluster showed that those that were most active on the bus network tended to undertake activities concentrated around central Birmingham, with no significant hotspots found elsewhere. Those that were less active tended to have more dispersed hotspots, concentrated around central Birmingham, Coventry and Solihull, with additional, less significant hotspots throughout the study area. Additionally, the most active cardholders were also found to reside in the most deprived areas. It was found that whilst both the number of stores visited and the distance travelled to regularly visited stores were relatively equal between IMD deciles, those that were more deprived tended to have fewer rarely visited stores and would not to travel as far to visit these stores. These results suggested that cardholders residing in the most deprived LSOAs exhibited spatially constrained patterns of activity. Constrained activity areas in deprived populations have been found in previous research (Adeniyi et al. 2020; Hasanzadeh et al. 2018) and the results produced in this thesis add to the evidence base around this issue, from both a transportation and retail perspective. Paired with the evidence that those residing in more deprived areas were more likely to have experienced an increase in bus travel times and a decline in the use of the bus network, this suggested that these cardholders may be at-risk of transport disadvantage.

Transport disadvantage and the associations with social exclusion and the wellbeing of individuals is a multi-faceted issue. Throughout this thesis, consumer data has been shown to offer valuable insight into these complex issues. Clear differences in the use of the public transport network were found within the older population, highlighting the fact that this population and their travel behaviours are not homogenous and should not be treated as such by transport authorities and policy makers. With the changing retail environment, it is imperative that transport services are able to adapt to these changes whilst accounting for the unique needs and demands of the older population. It is hoped that the insights produced throughout this thesis can shed light on the decline in the use of the ENCTS scheme in the West Midlands, and highlight the areas in which the current transport network may not be meeting the mobility needs and demands of the older population. One of the aims of this thesis was to identify changes in the use of the bus network and the potential reasons for these changes. This section has highlighted how this thesis was able achieve this through the implementation of both smart card data and loyalty card data.

9.2.3 Wider applications and policy recommendations

Previous research has found that those residing in urban areas make regular use of public transport and those that reside in more rural areas tend to use private transportation due to the irregularity of public transport services (Laverty 2015). Chapter 5 of this research evidenced that this is the case in the West Midlands, with cardholders that regularly use the transport network more likely to reside in central urban areas. However, evidence presented in Section 5.3.2 suggested that there may be a shift in this trend in the future. Those residing in more rural areas exhibited increasing use of the bus network and those residing in central urban areas were found to be more likely to have reduced their use of bus services. Currently, the ENCTS allows for free public transport use for those that are eligible. However, analysis of both service provision and travel times showed that both the provision of bus services and the accessibility of key facilities via the bus network differed significantly throughout the study area. Rural areas were particularly disadvantaged, with generally higher travel times and significantly lower service provision than central urban areas (Palacin et al. 2016). With increasing bus usage, and therefore increasing demand for bus services in these areas, it could be argued that improving rural bus services may bring about more benefits for the older population than the ENCTS. There have been many arguments that the ENCTS scheme is too widely targeted, with the majority of cardholders making little use of their concessionary travel cards (Mackett 2014). It has also been found that rural populations are often more vulnerable to social exclusion, in particular those that lack access to private transportation (Gray et al. 2006). Targeting future schemes towards these populations, for example improving rural bus services or funding demand-responsive transport in rural areas, may therefore help to improve the mobility and reduce the social exclusion of these most at-risk populations.

Transport in the West Midlands is currently undergoing a shift towards ‘smart’ technologies and transport solutions. With the West Midlands selected as the UK’s first ‘future mobility zone’ there has been increased investment in new transport solutions such as electric bikes, car sharing, and MaaS (TfWM 2018). A shift in the transport system towards smart technologies and new modes of transport, along with a change in local highstreets and retail areas largely driven by an increase in e-commerce, risks exacerbating transport disadvantage and leaving certain groups who are unable to take advantage of these new technologies vulnerable to social exclusion (Nimrod 2018). Improving the current transport system in order to meet the needs and demands of these at-risk populations can help to ensure that the implementation of new technologies and modes of transport does not exacerbate social exclusion and widen the gap between those that are able to take advantage of new technologies and those that are not.

9.2.3.1 COVID-19 impacts on the future of public transportation

At the time of writing, COVID-19 has had wide ranging impacts on society, mobility and the use of public transportation. Between March to June 2020, demand for public transport services in the West Midlands dropped by 86% (TfWM 2020), and whilst this is expected to begin to increase again over the short to medium term, the way public transport services are used is unlikely to return to normal in the near future.

Changes are expected in both how and why people travel. With an increase in the population working from home, commuting trips would be expected to decline (DfT 2020). Additionally, with social distancing measures in place and fears surrounding the safety of public transport, an increase in the use of the network to previous levels may take some time. This may be especially true for the older population who are generally the most at risk and therefore may be more reluctant to return to using public transport services. Throughout this thesis, issues of transport disadvantage and social exclusion have been widely discussed. If the older population are unable or unwilling to make use of public transport services, the risk of these populations becoming excluded from society may worsen. Ensuring an inclusive, safe, efficient and effective public transport system is now even more vital to allow for participation in society and to minimise these negative social impacts.

Smart technologies and novel data will be vital for understanding the impacts of COVID-19 and assessing the future of public transport services (Gutierrez et al. 2020). Consumer data, including those used throughout this thesis, can be implemented to understand demand and changes in travel behaviour. Additionally, smart card data can provide insights into the use of different routes and services and potentially be used to identify overcrowding on the network. Chapter 5 of this thesis identified groups of the population that make use of public transport services, and also identified those that have exhibited declining usage of services and therefore may be particularly vulnerable to transport disadvantage and social exclusion. With social exclusion likely to worsen due to social distancing measures, identifying these groups would allow for specific targeting of measures and policies to minimise the potential negative impacts, for example increased demand responsive transport or ‘mobility credits’ for use on taxi services.

With a modal shift towards private transportation, walking and cycling likely, there is potential for a large proportion of the older population to be left behind. Chapter 6 of this thesis inferred the purpose of trips made on the bus network. This method could be applied to more recent data to understand changes in the use of bus services with more people working from home and a reduction in non-essential travel (DfT 2020). Insights from these inferences could then be used to better plan services around the needs of bus users, for example increased services to areas that attract a large number of essential trips. Likewise, Chapters 7 and 8 explored demand from

a retail perspective and the impact of public transport provision and efficiency on the activity of the older population. Insights into the relationship between public transport and activity could be used to predict how changes in public transport scheduling could impact the activity patterns of bus users.

Additionally, the decline in bus usage and the predicted long-term impacts of COVID-19 on public transport usage bring about questions as to whether the ENCTS is best suited to meet the mobility needs of the older population in the future. As seen in Chapter 5, the majority of ENCTS cardholders make little use of the bus network, and with the probable reduction in public transport usage in coming years there is an argument that this money may be better invested in other schemes to ensure social inclusion, such as demand-responsive transport or credits for use on taxi services where issues of safety may not be as concerning. Novel consumer data provides a unique opportunity to gain an in-depth view of public transport usage before, after and during the COVID-19 pandemic and have the potential to be utilised to minimise the negative impacts that this may have on the social exclusion of the older population.

9.3 Reflections on data

9.3.1 The utility of consumer data

This thesis presented a broad range of insights, including the representativeness of the smart card and loyalty card data, and the types of insight that can be extracted to explore mobility and activity patterns. Throughout this thesis, the limitations of specific analyses are outlined in each chapter. This section discusses the utility of consumer data and the more general limitations related to the use of these data for mobility insights that were encountered throughout this research.

Errors and uncertainties in data are encountered in both traditional data and novel consumer data. The main sources of errors in traditional data are human errors in the data collection and processing and recollection errors from research subjects (Lyberg and Kasprzyk 2011). The main source of errors encountered when working with consumer data throughout this thesis was the uncertainty around how the data were collected and processed by the data partners. Clear errors were detected during preliminary analysis of the smart card data, for example cardholders with an age group of less than 60 recorded as having a ‘60 Plus’ smart card. Whether these errors occurred during the collection of the data or when the data were processed was unknown. This raises concern around errors in the data that were unable to be detected, for example incorrect age group, sex or registered LSOA. Although these issues are also present in

traditional data, the uncertainty around how data are collected and processed before being provided for research means that potential sources of error are unknown and can therefore not easily be accounted for (Firmani et al. 2016).

The provenance of unknown errors and missing data in consumer data may limit their application for granular insights into individual behaviour. However, the 'total-capture' nature of these data means they can often be advantageous for the analysis of aggregate insights. Chapter 5 of this thesis aimed to group cardholders based on their activity on the bus network. Over 250,000 cardholders were found to have rarely used the network and therefore very few records were available for these users, however, they were still captured in the data. Often, travel surveys and questionnaires are undertaken either on the bus network or at stops and stations (Long and Thill 2015). This means that the likelihood of capturing regular users is higher than those that rarely use the network, and therefore the results of analyses may be more representative of the behaviours of these regular travellers. Consumer data can therefore be beneficial for the analysis of total populations, in particular for those whose data may have previously been difficult to acquire.

One of the primary limitations of consumer data is the issue of representativeness. The majority of consumer data are collected from specific populations, for example cardholders of a specific loyalty card scheme. For insights to be extracted and applied in a wider research context the representativeness of these data must be ascertained (Kitchin 2014). Due to the anonymisation of data and the aggregation of address data to higher spatial scales, such as LSOA level, inferences had to made about many of the demographic and socioeconomic characteristics of the study population. This meant that the specific representativeness of the data in terms of demographic and socioeconomic characteristics could not be determined but instead had to be inferred based on known characteristics (age, sex and registered LSOA). This is generally less of an issue with traditional data collection methods where specific characteristics of interest can often be collected directly from participants.

Similarly, this anonymisation of data can present issues when utilising multiple datasets. This thesis utilised both HSR data and smart card data collected on the bus network. Due to the anonymisation of these data, these could only be linked at LSOA level rather than individual level, which would have been possible had data not been previously anonymised. This is likely to always be an issue when working with consumer data, with issues of privacy and ethics meaning it is rare to have access to personal data with no anonymisation. Spatial linkage and profile matching, i.e. linking subjects through similarities in their characteristics, are generally the current solutions used to overcome these issues and were shown throughout this thesis to produce robust aggregate results and insights into general trends. This thesis aimed to ascertain

the extent to which consumer data can be used for transport and mobility research. The need to infer the characteristics of data subjects and aggregate the data to larger spatial scales limited the extent to which these data could provide granular, individual insights. Due to the anonymisation of consumer data, results were therefore largely limited to aggregate population insights. It is therefore argued that, in their current state, these data would not be suitable as a replacement for traditional transport data but would be best used to supplement findings.

9.3.1.1 Data triangulation

Throughout this thesis, the idea of the triangulation of multiple data to provide a more comprehensive view of mobility has been discussed. This research explored the idea of utilising non-transport related consumer data to provide insight into mobility and activity patterns that would traditionally be collected from travel surveys. Implementing multiple datasets and methods to investigate the same phenomenon can be beneficial for a number of reasons that are outlined in this section.

The implementation of both smart card and loyalty card data in this thesis allowed for the concepts of mobility, activity and disadvantage to be explored from different perspectives. This can be particularly beneficial when conducting research with consumer data, due to differences in the representativeness of each dataset (Liu et al. 2016). This also highlighted the advantage of spatial data, with the lack of personal information contained in the smart card and loyalty card data meaning that comparisons between these data, and the insights generated from each, relied upon spatial linkage; primarily at LSOA level. With the smart card data only capturing users of the bus network and the loyalty card data only those that participate in the HSR scheme, combining the insights from these data provided a more comprehensive understanding of the phenomena being investigated.

Chapter 6, which inferred the purpose of trips made on the bus network, was largely methodological and demonstrated how the combination of both novel and traditional data can enhance insights generated from both. Although there has been much discussion around replacing traditional data with alternative, novel data (Milne and Watling 2019), this transition will likely take time and consist of at least some period of cross-over between both; with traditional data still needed for validation and contextualisation purposes. This chapter demonstrated how this can be implemented in practice. Like the data cleaning and processing methods, standard practices for the triangulation and amalgamation of novel and traditional data could open this up to wider research and accelerate the process of supplementing, and eventually replacing, these traditional surveys.

Although these chapters presented promising results, these were often hindered by data availability. The potential of consumer data for mobility insights has been widely discussed, however the inclusion of additional consumer data sources will only improve these insights. This thesis presented processes and methods for the cleaning and analysis of records generated by retail transactions. Many of these are transferable between data, and therefore could be implemented on additional data to widen the scope of this investigation or to provide guidance for further research utilising multiple sources of consumer data.

9.4 Summary and conclusions

In many respects, the use of consumer data for transport and mobility research is in its infancy. In reference to the applications and current limitations outlined earlier in the chapter, future research should continue to develop our understanding of the processes necessary to undertake in order to utilise consumer data for this type of research. These processes include data linkage between consumer datasets and with other traditional data, standard processes for data cleaning, considerations of the representativeness and bias of the data, and the creation of metadata to increase the accessibility of consumer data for research. Investigation of these concepts in alternative settings could increase the understanding of these uncertainties and the development of standard processes for consumer data research.

Additionally, this thesis developed methods for the analysis of both smart card data and loyalty card data for mobility insights that can be applied in a wider research context. For example, the trip chaining method utilised in Chapter 6 to link individual smart card transactions into full trips was conducted on transaction data built on the ITSO standard. This method could therefore be applied to all smart card data collected in this format. Chapters 7 and 8 developed methods for extracting spatial behaviours from records of loyalty card transactions; identifying regular, irregular and rare store networks and quantifying the activity spaces of cardholders. Applications of these methods on alternative consumer datasets could further improve these methods and facilitate the development of standardised methods for such research.

In terms of specific insights into the transport system and mobility of the older population in the West Midlands, this thesis produced insights into the recorded decline in bus boardings by concessionary travellers and the effectiveness of the current transport network. Implementing these results alongside insights that are available to the local transport authority could support policy and transport planning. One primary finding of this thesis was that the current transport network may not be suited to meeting the specific mobility needs of the older population, in particular where these differ from those of working age. The continuing change of the retail

environment may exacerbate this in the future, with potentially significant impacts on those that are unable to take advantage of these new forms of retail (Hughes and Jackson 2015; Hubbard 2017). Analyses of loyalty card data highlighted the role that non-transport related consumer data could play in understanding these impacts. Future implementation of further consumer datasets could widen the scope of research beyond retail activity to gain a more comprehensive view of the impact of the public transport network on the activity of populations.

9.5 Closing remarks

In closing, this thesis has largely addressed the aims outlined in Chapter 1 and produced two primary outputs; the development of processes and methods for the analysis of consumer data, and specific insights into the mobility and bus usage of the older population. Whilst this thesis has presented processes and methods through which insights can be extracted from consumer data, discussion of these methods has highlighted the limitations of these processes and how these can be improved and expanded upon by continued research. The range of insights into the mobility patterns of the older population extracted from both smart card and loyalty card data are promising in terms of their role in future research. Whilst these data are by no means a viable substitution for conventional transport data in isolation, such research agendas highlight their utility within this context and present new ways in which mobility and activity patterns can be investigated.

References

Adeniyi O, Brown A, Whysall P. (2020). *Retail location preferences: A comparative analysis.* Journal of Retailing and Consumer Services, 55, 102-146.

Agard B, Morency C, Trepanier M. (2006). *Mining public transport user behaviour from smart card data*. IFAC Proceedings, 39(3), 399-404.

Age UK. (2012). *600,000 are 'prisoners in own home'*. Available at: https://bit.ly/2VgPKaR. (Accessed: 24th February 2020).

Allaway A, Gooner R, Berkowitz D, Davis L. (2006). *Deriving and exploring behaviour segments within a retail loyalty card program*. European Journal of Marketing, 40(11/12), 1317-1339.

Alsnih R, Hensher D. (2003). *The mobility and accessibility expectations of seniors in an ageing population*. Transportation Research Part A: Policy and Practice, 37(1), 903-916.

Altman N, Krzywinski M. (2015). *Simple linear regression*. Nature Methods, 12(11), 999-1000.

Andrews G, Parkhurst G, Susilo Y, Shaw J. (2012). *The grey escape: investigating older people's use of the free bus pass*. Transport Planning and Technology, 35(1), 3-15.

Anselin L. (1995). *Local indicators of spatial association – LISA*. Geographical Analysis, 27(2), 93-115.

Anselin L. (2005). *Exploring spatial data with GeoData: A workshop*. Spatial Analysis Laboratory, University of Illinois, Chicago.

Anselin L. (2017). *Spatial regression: specification of spatial dependence*. [Online]. Available at: https://bit.ly/306MIa0. [Accessed: 14/07/2020].

Ashby D, Longley P. (2005). *Geocomputation, geodemographics and resource allocations for local policing*. Transactions in GIS, 9(1), 53-72.

Baernholdt M, Yan G, Hinton I, Rose K, Mattos M. (2012). *Quality of life in rural and urban adults 65 years and older: Findings from the National Health and Nutrition Examination Survey*. Journey of Rural Health, 28(4), 339-347.

Bagchi M, White P. (2005). *The potential of public transport smart card data*. Transport Policy, 12(5), 464-474.

Banister D, Bowling A. (2004). *Quality of life for the elderly: the transport dimension.* Transport Policy, 11, 105-115.

Barry B. (1998). *Social exclusion, social isolation and the distribution of income*. Centre for Analysis of Social Exclusion, London School of Economics, London, UK.

Battellino H. (2007). *In search of an effective service delivery model to provide services for the transport disadvantaged*. Thredbo 10 – Thredbo International Conference Series on Competition and Ownership in Land Passenger Transport, Sydney, Australia, 2007.

BBC News. (2020). *UK's 2021 Census could be the last, statistics chief reveals*. [Online]. Available at: www.bbc.co.uk/news/uk-51468919. (Accessed: 13th February 2020).

Ben-Hur A, Weston J. (2009). *A user's support guide to support vector machines*. Data Mining Techniques for the Life Sciences, 609, 223-239.

Benesty J, Chen J, Huang Y, Cohen I. (2009). *Pearson correlation coefficient*. Noise Reduction in Speech Processing. Springer Topics in Signal Processing, 2, 1-4.

Berk R. (2008). *Statistical learning from a regression perspective*. Springer, New York.

Binder J, Matern A. (2019). *Mobility and social exclusion in peripheral regions*. European Planning Studies, 28(6), 1049-1067.

Birkin M. (2019). *Spatial data analytics of mobility with consumer data*. Journal of Transport Geography, 76, 245-253.

Birkin M, Clarke G. (1998). *GIS, geodemographics and spatial modelling in the UK financial service industry*. Journal of Housing Research, 9(1), 87-111.

Bishop C. (2006). *Pattern recognition and machine learning.* Springer, New York.

Boarnet M, Crane R. (2001). *The influence of land use on travel behaviour: specification and estimation strategies*. Transportation Research Part A, 35, 823-845.

Bocarejo J, Oviedo D. (2012). *Transport accessibility and social inequalities: a tool for identification of mobility needs and evaluation of transport investments*. Journal of Transport Geography, 24, 142-154.

Bollier D. (2010). *The promise and peril of big data.* The Aspen Institute: Communications and Society Program, Washing DC, USA.

Boniface S, Scantlebury R, Watkins S, Mindell J. (2015). *Health implications of transport: Evidence of the effects of transport on social interactions*. Journal of Transport and Health, 2, 441-446.

Bora D, Gupta A. (2014). *A comparative study between fuzzy clustering algorithm and hard clustering algorithm*. International Journal of Computer Trends and Technology, 10(2), 108-113.

Breiman L, Friedman J, Olshen R, Stone C. (1984). *Classification and Regression Trees*. Chapman and Hall, New York.

Breiman L. (2001). *Random Forests*. Machine Learning, 45, 5-32.

Breusch T, Pagan A. (1980). *The Lagrange Multiplier test and its applications to model specification in econometrics*. The Review of Economic Studies, 47(1), 239-253.

Buliung R, Kanaroglou P. (2006). *Urban form and household activity-travel behaviour*. Growth and Change: A Journal of Urban and Regional Policy, 37(2), 172-199.

Burke M, Brown A. (2007). *Distances people walk for transport*. Road and Transport Research, 16(3), 16-29.

Busari A, Oluwafemi D, Ojo S, Oyedepo J, Ogbiye A, Ajayi S, Adegoke D, Daramola K. (2019). *Mobility dynamics of the elderly: Review of literatures.* IOP Conference Series: Materials Science and Engineering, 640.

Butcher L. (2020). *Concessionary bus travel*. [Online]. Available at: https://bit.ly/3kPzl7f. (Accessed: 01/11/2020).

Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville A. (2014). *The use of triangulation in qualitative research*. Methods and Meaning: credibility and trustworthiness of qualitative research, 41(1), 89-91.

Cass N, Shove E, Urry J. (2005). *Social exclusion, mobility and access.* The Sociological Review, 53(3), 539-555.

Chakirov A, Erath A. (2012). *Activity identification and primary location modelling based on smart card payment data for public transport.* Eidgenössische Technische Hochschule Zürich, IVT. Institute for Transport Planning and Systems.

Chen Y. (2013). *New approaches for calculating Moran's Index of Spatial Autocorrelation*. PLoS ONE, 8(7), 1-14.

Chien S, Qin Z. (2004). *Optimisation of bus stop locations for improving transit accessibility*. Transportation Planning and Technology, 27, (3), 211-227.

Church A, Frost M, Sullivan K. (2010). *Transport and social exclusion in London*. Transport Policy, 7(3), 195-205.

Clark W, Ham M, Coulter R. (2014). *Spatial mobility and social outcomes*. Journal of Housing and the Built Environment, 29, 699-727.

Coutinho F, van Oort N, Christoforou Z, Alonso-Gonzalez M, Cats O, Hoogendoorn S. (2020). *Impacts of replacing a fixed public transport line by a demand responsive transport system: case study of a rural area in Amsterdam*. Research in Transportation Economics, 83.

Crawford K. (2013). *The hidden biases in big data*. Available at: https://hbr.org/2013/04/the-hidden-biases-in-big-data. (Accessed: 21st February 2020).

CrowdTwist. (2018). *Loyalty Research Report: Engaging consumers in an age of increasing brand distrust*. [Online]. Available at: https://bit.ly/3i2cKDv. (Accessed: 26/06/2020).

Cui J, Loo B, Lin D. (2016). *Travel behaviour and mobility needs of older adults in an ageing and car-dependent society*. International Journal of Urban Sciences, 21(2), 109-128.

Dalton C, Thatcher J. (2015). *Inflated granularity: Spatial 'big data' and geodemographics*. Big Data and Society, 2(2), 1-15.

Daniels R, Mulley C. (2013). *Explaining walking distance to public transport: The dominance of public transport supply*. The Journal of Transport and Land Use, 6(2), 5-20.

Danziger P. (2004). *Why do people buy things they don't need?* Paramount Market Publishing, New York, USA.

Daod J. (2017). *Multicollinearity and regression analysis.* Journal of Physics: Conference Series, 949, 1-6.

Davenport A, Zaranko B. (2020). *Levelling up: where and how?* [Online]. Available at: https://www.ifs.org.uk/publications/15055. (Accessed: 30/10/2020).

Davison L, Enoch M, Ryley T, Quddus M, Wang C. (2014). *A survey of demand responsive transport in Great Britain*. Transport Policy, 31, 47-54.

De Mauro A, Greco M, Grimald M. (2015). *What is big data? A consensual definition and review of key research tropics.* AIP Conference Proceedings, 1644, 97-104.

Deng Z, Ji M. (2010). *Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach*. 7th Annual Conference on Traffic and Transportation Studies, Kunming, China.

Department for Transport (DfT). (2020). *Transport use during the coronavirus (COVID-19) pandemic*. [Online]. Available at: https://bit.ly/2Or3nQn. (Accessed: 15/07/2020).

Department for Transport (DfT). (2020). *Capability, health and travel behaviour of older people*. [Online]. Available at: https://bit.ly/35BpeN2. (Accessed: 28/10/2020).

Department for Transport (DfT). (2018). *National Travel Survey*. [Online]. Available at: https://bit.ly/1g5x1Ak. (Accessed: 20/05/2018).

Department for Transport (DfT). (2017). *Commuting trends in England 1988 – 2015*. [Online]. Available at: https://bit.ly/2XeBmyY. (Accessed: 12/06/2019).

Department for Transport (DfT). (2016). *National Travel Survey 2016: notes and definitions*. [Online]. Available at: https://bit.ly/33TFQ0g. (Accessed: 12/06/2019).

Department for Transport (DfT). (2016). *GB Driving License Data*. [Online]. Available at: https://bit.ly/3fExZsw. (Accessed: 04/09/2018).

Department for Transport (DfT). (2015). *National Public Transport Data Repository*. [Online]. Available at: https://bit.ly/2ZoVvUw. (Accessed: 30/06/2020).

Department for Transport. (2014). *National public transport access nodes*. [Online]. Available at: https://bit.ly/2M6qQpj. (Accessed: 20/05/2018).

Department for Work and Pensions. (2015). *Employment statistics for workers aged 50 and over, by 5-year age bands and gender*. [Online]. Available at: https://bit.ly/3dYzTUf. (Accessed: 03/07/2020).

Department for Work and Pensions. (2016). *Claiming discretionary housing payments*. [Online]. Available at: https://bit.ly/2Nm32NI. (Accessed: 12/12/2019).

Devillaine F, Munizaga M, Trepanier M. (2012). *Detection of activities of public transport users by analysing smart card data.* Transportation Research Record: Journal of the Transportation Research Board, 2276(1), 48-55.

European Environment Agency. (2017). *Urban Atlas.* [Online]. Available at: https://bit.ly/2CKyHn9. (Accessed: 16/07/2019).

Everitt B, Landau S, Leese M, Stahl D. (2011). *Cluster Analysis.* Fifth Edition. Arnold, London.

Falk M, Hagsten E. (2015). *E-commerce trends and impacts across Europe*. International Journal of Production Economics, 170, 357-369.

Fan Y, Khattack A. (2008). *Urban form, individual spatial footprints, and travel: Examination of space-use behaviour*. Transportation Research Record: Journal of the Transportation Research Board, 2082(1), 98-106.

Farber S, Paez A. (2009). *My car, my friends and me: a preliminary analysis of automobility and social activity participation*. Journal of Transport Geography, 17(3), 216-225.

Farber S, Paez A, Mercado R, Roorda M, Morency C. (2011). *A time-use investigation of shopping participation in three Canadian cities: is there evidence of social exclusion*. Transportation, 38, 17-44.

Farrington J. (2007). *The new narrative of accessibility: its potential contribution to discourses in (transport) geography*. Journal of Transport Geography, 15(5), 319-330.

Felgate M, Fearne A, Di Falco S, Martinez M. (2012). *Using supermarket loyalty card data to analyse the impact of promotions*. International Journal of Market Research, 54(2), 221-240.

Ferrari L, Berlingerio M, Calabrese F, Curtis-Davidson B. (2013). *Measuring public transport accessibility using pervasive mobility data*. IEEE Pervasive Computing, 12(1), 26-33.

Ferriera A, Beukers E. (2012) *Accessibility is gold, mobility is not: a proposal for the improvement of Dutch transport-related cost-benefit analysis*. Environment and Planning B: Planning and Design, 39, 683-697.

Firmani D, Mecella M, Scannapieco M, Batini C. (2016). *On the meaningfulness of "big data quality"*. Data Science and Engineering, 1, 6-20.

Fleiss J. (1981). *Statistical methods for rates and proportions.* John Wiley and Sons, London.

Forkenbrock D, Benshoff S, Weisbrod G. (2001). *Assessing the social and economic effects of transportation projects*. University of Iowa, Economic Development Research Group.

Gabriel Z, Bowling A. (2004). *Quality of life from the perspective of older people*. Ageing and Society, 24, 675-691.

General Data Protection Regulation (GDPR). (2018). *General Data Protection Regulation (GDPR)*. [Online]. Available at: https://gdpr-info.eu. (Accessed: 20/05/2018).

Gong H, Chen C, Bialostozky E, Lawson C. (2012). *A GPS/ GIS method for travel mode detection in New York City*. Computers, Environment and Urban Systems, 36, 131-139.

Goodwin P. (1985). *Evidence on car and public transport demand elasticities 1980 – 1988*. Oxford University Transport Studies Unit.

Gordon J, Koutsopoulos H, Wilson N, Attanucci J. (2013). *Automated inference of linked transit journeys in London using fare-transaction and vehicle location data*. Transportation Research Record: Journal of the Transportation Research Board, 2343, 17-24.

Gou J, Du L, Zhang Y, Xiong T. (2012). *A new distance-weighted k-nearest neighbour classifier*. Journal of Information and Computational Science, 9(6), 1429-1436.

Graham M, Shelton T. (2013). *Geography and the future of big data, big data and the future of geography*. Dialogues of Human Geography, 3, 255-261.

Gray D, Shaw J, Farrington J. (2006). *Community transport, social capital and social exclusion in rural areas*. Area, 38(1), 89-98.

Greengauge 21. (2009). *Fast Forward: A High-Speed Rail Strategy for Britain.* Greengauge 21. London, UK.

Griffiths S, Vaughan L, Haklay M, Jones C. (2008). *The sustainable suburban high street: A review of themes and approaches*. Geography Compass, 2(4), 1155-1188.

Guardian, The. (2009). *UK transport mapped: Every bus stop, train station, ferry port and taxi rank in Britain. Welcome to the ultimate transport data.* [Online]. Available at: https://bit.ly/2oHRQmc. (Accessed: 30/06/2020).

Guo-Xun Y, Chia-Hua H, Chih-Jen L. (2012). *Recent advances of large-scale linear classification.* Proceedings of the Institute of Electrical and Electronics Engineers, 100(9), 2584-2603.

Gutierrez A, Miravet D, Domenech A. (2020). *COVID-19 and urban public transport services: emerging challenges and research agenda*. Cities and Health, Special Issue: COVID-19.

Hamlett J, Bailey A, Alexander A, Shaw G. (2008). *Ethnicity and consumption: south Asian food shopping patterns in Britain, 1947-1975.* Journal of Consumer Culture, 8(1), 91-116.

Han G, Sohn K. (2016). *Activity imputation for trip-chains elicited from smart card data using a continuous hidden Markov model*. Transportation Research Part B: Methodological, 83, 121-135.

Handy S. (2002). *Accessibility vs. mobility-enhancing strategies for addressing automobile dependence in the US*. US Department of Environmental Science and Policy, European Conference of Ministers of Transport, Davis, USA.

Hanley J, McNeil B. (1982). *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 143(1), 29-36.

Hartigan J, Wong M. (1979). *Algorithm AS 136: A k-means clustering algorithm*. Journal of the Royal Statistical Society, Series C (Applied Statistics), 28, 100-108.

Hasanzadeh K, Laatikainen T, Kytta M. (2018). *A place-based model of local activity spaces: individual place exposure and characteristics*. Journal of Geographical Systems, 20, 227-252.

Hasanzadeh K, Czepkiewicz M, Heinonen J, Kytta M, Ala-Mantila S, Ottelin J. (2019). *Beyond geometries of activity spaces: A holistic study of daily travel patterns, individual characteristics, and perceived wellbeing in Helsinki metropolitan area*. Journal of Transport and Land Use, 12(1), 149-177.

Haustein S. (2011). *Mobility behaviour of the elderly: an attitude-based segmentation approach for a heterogeneous target group*. Transportation, 39(6), 1079-1103.

Hayes T, Usami S, Jacobucci R, McArdle J. (2015). *Using classification and regression trees (CART) and random forests to analyse attrition: results from two simulations.* Psychology and Aging, 30(4), 911-929.

Heckman J. (1990). *Selection bias and self-selection*. In: Eatwell J, Milgate M, Newman P. Econometrics. The New Palgrave. Palgrave Macmillan, London.

Herwangi Y, Pradono P, Syabri I, Kustiwan I. (2013). *Assessing transport disadvantage and transport-related social exclusion for inclusive transportation planning: a review of the methodology*. The Second Planocosmo Conference, ITB, Bandung.

Hess D. (2012). *Walking to the bus: Perceived versus actual walking distance to bus stops for older adults*. Transportation, 39, 247-266.

Hildebrand E. (2003). *Dimensions in elderly travel behaviour: a simplified activity-based model using lifestyle clusters*. Transportation, 30(3), 285-306.

Hine J. (2012). *Mobility and Transport Disadvantage.* Mobilities: New perspectives on transport and society. Ashgate Publishing Ltd., Farnham, UK.

Hirsch J, Winters M, Clarke P, McKay H. (2014). *Generating GPS activity spaces that shed light upon the mobility habits of older adults: a descriptive analysis*. International Journal of Health Geographics, 13(1), 51.

Hobbs A, Hanley S. (2014). *Big and open data in transport*. Number 472. Parliamentary Office of Science and Technology.

Holliday K, Howard A, Emch M, Rodriguez D, Evenson K. (2017). *Are buffers around home representative of physical activity spaces among adults?* Health and Place, 45, 181-188.

Huang Q, Wong D. (2016). *Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us?* International Journal of Geographical Information Science, 30(9), 1873-1898.

Hubbard P. (2017). *The 'death' of the high street.* The Battle for the High Street. Palgrave Macmillan, London, UK.

Hughes C, Jackson C. (2015). *Death of the high street: identification, prevention and reinvention*. Regional Studies, Regional Science, 2(1), 237-256.

Hughes M, Moreno J. (2013). *Further analysis of data from the household electricity usage study: Consumer archetypes.* Final Report for the Department of Energy and Climate Change and the Department for the Environment, Food and Rural Affairs.

International Transport Forum (ITF) (2015). *Big Data and Transport: Understanding and assessing the options.* Organisation for Economic Co-operation and Development (OECD).

Jackson S, Firth J, Firth J, Veronese N, Gorely T, Grabovac I, Yang L, Smith L. (2019). *Social isolation and physical activity mediate associations between free bus travel and wellbeing among older adults in England.* Journal of Transport and Health, 13, 274-284.

Jacobs-Crisioni C, Rietveld P, Koomen E, Tranos E. (2014). *Evaluating the impact of land-use density and mix on spatiotemporal urban activity patterns: An exploratory study using mobile phone data*. Environment and Planning A, 46, 2769-2785.

Jain A. (2010). *Data Clustering: 50 years beyond K-means.* Pattern Recognition Letters, 31(8), 651-666.

Jain A, Dubes R. (1988). *Algorithms for clustering data*. New Jersey, USA. Prentice Hall.

Jarv O, Muurisepp K, Ahas R, Derudder B, Witlox F. (2015). *Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia*. Urban Studies, 52(14), 2680-2698.

Jiang S, Ferreira J, Gonzales M. (2012). *Clustering daily patterns of human activities in the city*. Data Mining and Knowledge Discovery, 25(3), 478-510.

Johnson J. (1986). *Ecological fallacy*. The Dictionary of Human Geography, 2nd Edition, Blackwell, Oxford, UK.

Joyce J. (2003). *Bayes' Theorem*. The Stanford Encyclopedia of Philosophy, Stanford University, USA.

Jun C, Dongyuan Y. (2013). *Estimating smart card commuters' origin-destination distribution based on APTS data.* Journal of Transportation Systems Engineering and Information Technology, 13(4), 47-53.

Kaisler S, Armour F, Espinosa J, Money W. (2013). *Big Data: Issues and challenges moving forward*. 46th Hawaii International Conference on System Science, Hawaii, USA.

Kaur N, Kumar K. (2016). *Normalisation based k-means data analysis algorithm*. International Journal of Advanced Research in Computer Science and Software Engineering, 6(6), 455-457.

Kenyon S, Lyons G, Rafferty J. (2002). *Transport and social exclusion: Investigating the possibility of promoting inclusion through virtual mobility*. Journal of Transport Geography, 10, 207-219.

Kestens Y, Thierry B. Shareck M, Steinmetz-Wood M, Chaix B. (2018). *Integrating activity spaces in health research: Comparing the VERITAS activity space questionnaire with 7-day GPS tracking and prompted recall*. Spatial and Spatio-temporal Epidemiology, 25, 1-9.

Key W, Culliney M. (2016). *The oldest old and the risk of social exclusion*. Social Policy and Society, 17(1), 47-63.

Khan K, Rehman S, Aziz K, Fong S, Sarasvady S. (2014). *DBSCAN: Past, present and future.* The Fifth International Conference on the Applications of Digital Information and Web Technologies, Bangalore, 2014, 232-238.

Kieu L, Bhaskar A, Chung E. (2014). *Passenger segmentation using smart card data*. IEEE Transactions on Intelligent Transport Systems, 16, 1537-1548.

Kilby K, Smith N. (2012). *Accessibility Planning Policy: Evaluation and future directions – final report.* Centre for Research in Social Policy.

Kim S, Ulfarsson G. (2004). *Travel mode choice of the elderly: effects of personal, household, neighbourhood and trip characteristics*. Transportation Research Record: Journal of Transportation Research Board, 1894, 117-126.

Kimpel T, Dueker K, El-Geneidy A. (2007). *Using GIS to measure the effect of overlapping services areas on passenger boardings at bus stops.* Journal of the Urban and Regional Information Systems Association, 19(1), 5-12.

Kitchin R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, London, UK.

Knox P. (1981). *Retail Geography and Social Well-Being: A note on the changing distribution of pharmacies in Scotland*. Geoforum, 12(3), 255-264.

Kotler P. (2002). *Marketing Places*. Simon and Schuster, New York, USA.

Kuhn M. (2019). *The caret package*. [Online]. Available at: https://bit.ly/33Sb4ov. (Accessed: 08/03/2020).

Kuhn I, Dormann C. (2012). *Less than eight (and a half) misconceptions of spatial analysis*. Journal of Biogeography, 39(5), 995-998.

Kusakabe T, Asakura Y. (2014). *Behavioural data mining of smartcard data: a data fusion approach*. Transportation Research Part C: Emerging Technologies, 46, 179-191.

Landis R, Koch G. (1977). *The measurement of observer agreement for categorical data.* Biometrics, 33(1), 159-174.

Langlois G, Koutsopoulos H, Zhao J. (2016). *Inferring patterns in the multi-week activity sequences of public transport users*. Transportation Research Part C: Emerging Technologies, 64, 1-16.

Larch M, Walde J. (2007). *Lag or error? Detecting the nature of spatial correlation*. Data Analysis, Machine Learning and Applications, 301-308.

Larose D. (2005). *Hierarchical and k-means clustering. Discovering Knowledge in Data: An Introduction to Data Mining.* Second Edition. Hoboken, New Jersey. John Wiley and Sons, Inc.147-162.

Last A. (2010). *Smart card data on use of free concessionary travel by older and disabled bus passengers*. European Transport Conference, Glasgow, UK.

Laverty A, Millet C. (2015). *Potential impacts of subsidised bus travel for older people.* Journal of Transport and Health, 2(1), 32-34.

Lavrakas P. (2008). *Ecological fallacy*. Encyclopedia of Survey Research Methods. Sage Publications, Thousand Oaks, California.

LeSage J. (1997). *Regression analysis of spatial data*. The Journal of Regional Analysis and Policy, 27(2), 83-94.

Leverentz A. (2020). *Beyond neighbourhoods: activity spaces of returning prisoners*. Social Problems, 67(1), 150-170.

Li H, Raeside R, Chen T, McQuaid R. (2012). *Population ageing, gender and the transportation system.* Research in Transportation Economics, 34(1), 39-47.

Li J, Xu L, Yao D, Mao Y. (2019). *Impacts of symbolic value and passenger satisfaction on bus use*. Transportation Research Part D: Transport and Environment, 72, 98-113.

Liao L, Fox D, Kautz H. (2007). *Extracting places and activities from GPS traces using hierarchical conditional random fields*. International Journal of Robotics Research, 26(1), 119-134.

Liu L, Andris C, Ratti C. (2010). *Uncovering cabdrivers' behaviour patterns from their digital traces*. Computers, Environment and Urban Systems, 34(6), 541-548.

Liu L, Hou A, Biderman A, Ratti C, Chen J. (2006). *Understanding individuals and collective mobility patterns from smartcard records: a case study in Shenzhen*. Proceedings of the 12th International IEEE Conference on Intelligent Transport Systems, St. Louis, MO, USA.

Liu Z, Li Q, Zhaung Y, Xiong J, Li S. (2018). *Inferring travel purposes for transit smart card data.* In: Qui M. Smart Computing and Communication. SmartCom 2017. Lecture Notes in Computer Science, 10699, Springer, Cham.

Liu J, Li J, Li W, Wu J. (2016). *Rethinking big data: A review on the data quality and usage issues*. ISPRS Journal of Photogrammetry and Remote Sensing, 115, 134-142.

Luiu C, Tight M, Burrow M. (2016). *The unmet travel needs of the older population: a review of the literature*. Transport Reviews, 37(4), 488-506.

Lloyd A, Cheshire J. (2019). *Detecting address uncertainty in loyalty card data*. Applied Spatial Analysis and Policy, 12(2), 445-465.

Loader C, Stanley J. (2009). *Growing bus patronage and addressing transport disadvantage – the Melbourne experience*. Transport Policy, 16(3), 106-114.

Long Y, Shen Z. (2013). *Disaggregating heterogeneous agent attributes and location from aggregated data, small-scale surveys and empirical researches*. Computers, Environment and Urban Systems, 42, 14-25.

Long Y, Thill J. (2015). *Combining smart card data and household travel survey to analyse jobs-housing relationships in Beijing*. Computers, Environment and Urban Systems, 53, 19-35.

Longley P. (2005). *Geographical Information Systems: a public renaissance of geodemographics for public service delivery*. Progress in Human Geography, 29(1), 57-63.

Longley P. (2017). *Geodemographic Profiling*. The International Encyclopedia of Geography.

Lu Y, Zhu S, Zhang L. (2012). *A machine learning approach to trip purpose imputation in GPS-based travel surveys.* 4th Annual Conference on Innovations in Travel Modelling, Tampa, USA.

Lucas K. (2019). *A new evolution for transport-related social exclusion research?* Journal of Transport Geography, 81, 1-4.

Lucas K. (2012). *Transport and social exclusion: Where are we now?* Transport Policy, 20, 105-113.

Lucas K. (2003). *Making the Connections: Final report on transport and social exclusion.* Social Exclusion Unit, Office of the Deputy Prime Minister.

Lucas K, Stokes G, Bastiaanssen J, Burkinshaw J. (2019). *Inequalities in mobility and access in the UK Transport System.* Future of Mobility: Evidence Review, Government Office for Science.

Lupton R. (2015). *Deprived neighbourhoods and spatial inequalities in England*. Geographical Association Conference, University of Manchester.

Lyberg L, Kasprzyk D. (2011). *Data collection methods and measurement error: an overview*. Measurement Errors in Surveys, 235-257.

Ma L, Kent J, Mulley C. (2018). *Transport disadvantage, social exclusion, and subjective wellbeing: The role of neighbourhood environment – evidence from Sydney, Australia.* The Journal of Transport and Land Use, 11(1), 31-47.

Mackett R. (2014). *Has the policy of concessionary bus travel for older people in Britain been successful?* Case Studies on Transport Policy, 2(2), 81-88.

Mackett R. (2013). *Impact of concessionary bus travel on the wellbeing of older and disabled people*. Transportation Research Record: Journal of the Transportation Research Board, 2352, 114-199.

Mageean J, Nelson J. (2003). *The evaluation of demand responsive transport services in Europe*. Journal of Transport Geography, 11(4), 255-270.

Mah S, Mitra R. (2017). *The effects of a free bus program on older adults' travel behaviour: A case study of a Canadian suburban municipality*. Case Studies on Transport Policy, 5(3), 460-466.

Manaugh K, El-Geneidy A. (2012). *What makes travel 'local'?: Defining and understanding local travel behaviour*. The Journal of Transport and Land Use, 5(3), 15-27.

Marshall S. (1999). *Restraining mobility while maintaining accessibility: an impression of the 'City of Sustainable Growth'.* Built Environment, 25(2), 168-179.

Mayer-Schonberger V, Cukier K. (2013). *Big data: a revolution that will transport how we live, work and think*. Houghton Mifflin Harcourt, Boston, New York, USA.

McCray T, Brais N. (2007). *Exploring the role of transportation in fostering social exclusion: The use of GIS to support qualitative data*. Networks and Spatial Economics, 7(4), 397-412.

Milne D, Watling D. (2019). *Big data and understanding change in the context of planning transport systems*. Journal of Transport Geography, 76, 235-244.

Miranda-Moreno L, Eluru N, Lee-Gosselin M, Kreider T. (2012). *Impact of ICT access on personal activity space and greenhouse has production: Evidence from Quebec City, Canada*. Transportation, 39(5), 895-918.

Mohamed K, Come E, Baro J, Oukhellou L. (2014). *Understanding passenger patterns in public transit through smartcard and socioeconomic data*. UrbComp' 2014, New York, USA.

Mohamed K, Come E, Oukehllo L, Verleysen M. (2017). *Clustering smart card data for urban mobility analysis*. IEEE Transactions on Intelligent Transportation Systems, 18(3), 712-728.

Montgomery D, Peck E, Vining G. (2012). *Introduction to linear regression analysis*. John Wiley and Sons, New Jersey.

Morency C, Munizaga M, Greaves S, Chu K. (2015). *Two-year worth of smartcard transaction data – extracting longitudinal observations for the understanding of travel behaviour*. Transportation Research Proceedia, 11, 365-380.

Morissette L, Chartier S. (2013). *The k-means clustering technique: General considerations and implementation in Mathematica.* Tutorials in Quantitative Methods for Psychology, 9(1), 15-24.

Musselwhite C, Haddad H. (2010). *Mobility, accessibility and quality of later life. Quality in Ageing and Older Adults*, 11(1), 25-37.

Musselwhite C, Holland C, Walker I. (2015). *The role of transport and mobility in the health of older people*. Journal of Transport and Health, 2(1), 1-4.

Musselwhite C. (2018). *Mobility in later life and wellbeing*. Quality of Life and Daily Travel. Springer, New York, USA.

Newhaus F. (2009). *Urban diary – A tracking project*. UCL Working Paper Series, 151.

Newman M. (2004). *Power laws, Pareto distributions and Zipf's law.* Contemporary Physics, 46(5), 323-351.

Nimrod G. (2018). *Technophobia among older internet users*. Educational Gerontology, 44(2-3), 148-162.

Office for National Statistics (ONS). (2020). *Retail Sales Index*. [Online]. Available at: https://bit.ly/2BFjQx1. (Accessed: 26/06/2020).

Office for National Statistics (ONS). (2019). *Internet access – households and individuals*. [Online]. Available at: https://bit.ly/2SH39rc. (Accessed: 13/02/2020).

Office for National Statistics (ONS). (2019). *Personal and economic well-being: what matters most to our life satisfaction?* [Online]. Available at: https://bit.ly/2WlWqDh. (Accessed: 24/02/2020).

Office for National Statistics (ONS). (2019). *Labour market statistics time series.* [Online]. Available at: https://bit.ly/2OdIPKE.

Office for National Statistics (ONS). (2018a). *Population estimates.* [Online]. Available at: https://bit.ly/28IITPF. (Accessed: 20/05/2018).

Office for National Statistics (ONS). (2018b). *2011 Census Data.* [Online]. Available at: https://bit.ly/2fsjDkI. (Accessed: 20/05/2018).

Palacin R, Vigar G, Peacock S. (2016). *Transport poverty and urban mobility.* Justice and Fairness in the City: A multidisciplinary approach to 'ordinary' cities. Policy Press, Bristol University Press, Bristol, UK.

Peace R. (2001). *Social exclusion: a concept in need of definition?* Social Policy Journal of New Zealand, 16, 17-36.

Pelletier M, Trepanier M, Morency C. (2011). *Smartcard data use in public transit: A literature review.* Transport Research Part C, 19, 557-568.

Petersen J, Gibin M, Longley P, Mateos P, Atkinson P, Ashby D. (2010). *Geodemographics as a tool for targeting neighbourhoods in public health campaigns*. Journal of Geographical Systems, 13(2), 173-192.

Pooley C. (2016). *Mobility, transport and social inclusion: Lessons from history.* Transport Policy and Social Inclusion, 4(3), 100-109.

Preston J, Raje F. (2007). *Accessibility, mobility and transport-related social exclusion*. Journal of Transport Geography, 15(3), 151-160.

Preston J. (2010) *The Case for High Speed Rail: An Update.* RAC Foundation, University of Southampton.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ratti C, Pulselli R, Williams S, Frenchman D. (2006). *Mobility landscapes: Using location data from cell phones for urban analysis.* Environment and Planning B: Planning and Design, 33(5), 727-748.

Ray S, Turi R. (1999). *Determination of the number of clusters in K-means clustering and application in colour segmentation.* The 4th International Conference on Advances in Pattern Recognition and Digital Techniques.

Reynolds D, Horton F. (1971). *Effects of urban spatial structure on individual behaviour.* Economic Geography, 47, 36-48.

Richardson A, Ampt E, Meyburg A. (1995). *Survey methods for transport planning*. Eucalyptus Press, Melbourne, Australia.

Rish I. (2001). *An empirical study of the naïve Bayes classifier*. IJCAI Workshop on Empirical Methods in Artificial Intelligence, 3, 41-46.

Robinson N, Caraher M, Lang T. (2000). *Access to Shops: The views of low-income shoppers.* Health and Education Journal, 59(2), 121-136.

Rye T, Scotney D. (2004). *The factors influencing future concessionary bus patronage in Scotland and their implications for elsewhere.* Transport Policy, 11, 133-140.

SAGE. 2019. *Learn to test for multicollinearity in R with data from the English Health Survey*. 2019 SAGE Publications Ltd.

Sarial-Abi G, Gurhan- Canli Z, Kumkale T, Yoon Y. (2016). *The effect of self-concept clarity on discretionary spending tendency*. International Journal of Research in Marketing, 33(3), 612-623.

Sasirekha K, Baby P. (2013). *Agglomerative hierarchical clustering algorithm – a review.* International Journal of Scientific and Research Publications, 3(3), 2250-3153.

Sen A. (1981). *Poverty and famines: An essay on entitlement and deprivation*. Clarendon Press, Oxford, UK.

Scheiner J. (2006). *Does the car make elderly people happy and mobile? Settlement structures, car availability and leisure mobility of the elderly.* European Journal of Transport and Infrastructure Research, 6(2), 151-172.

Schluchter M. (2005). *Mean square error*. Encyclopedia of Biostatistics, John Wiley and Sons Ltd., London.

Schmocker J, Quddus M, Noland R, Bell M. (2004). *Estimating trip generation of the elderly and disabled: an analysis of London data.* Transportation Research Record, 1924(2), 9-18.

Schonfelder S, Axhausen K. (2003). *Activity spaces: measures of social exclusion?* Transport Policy, 10(4), 273-286.

Seber G, Lee A. (2012). *Linear Regression Analysis, Second Edition*. John Wiley and Sons, Hoboken, New Jersey, USA.

Sheller M. (2011). *Mobility*. Sociopedia. International Sociological Association. [Online]. Available at: https://bit.ly/1JzNAWX. (Accessed: 12/01/2019).

Singleton A. (2008). *Comparing Classifications: Some preliminary speculations on an appropriate scale for neighbourhood analysis with reference to Geodemographics Information Systems*. UCL Centre for Advanced Spatial Analytics, Working Paper Series, 127.

Singleton A, Longley P. (2009). *Creating open source geodemographics: Refining a national classification of census output areas for applications in Higher Education*. Regional Science, 88(3), 643-666.

Singleton A, Spielman S. (2014). *The past, present and future of geodemographic research in the United States and the United Kingdom*. The Professional Geographer, 66(4), 558-567.

Sinha K, Labi S. (2007). *Social and cultural impacts*. Transportation Decision Making: Principles and Project Evaluation and Programming. John Wiley and Sons Inc., Hoboken, New Jersey, USA.

Smith C, Quercia D, Capra L. (2013). *Finger on the pulse: Identifying deprivation using transit flow analysis*. Local is Where It's At, San Antonio, Texas, USA.

Social Exclusion Unit (SEU). (2003). *Making the Connections: Final report on transport and social exclusion.*

Song J, Eom J, Lee K, Min J, Yang K. (2015). *Public transportation service evaluations utilising Seoul transportation card data*. Procedia Computer Science, 52, 178-185.

Songchitruksa P, Zeng X. (2010). *Getis-Ord spatial statistics to identify hot spots by using incident management data*. Transport Research Record: Journal of Transportation Research Board, 2165(1), 42-51.

Spinney J, Scott D, Newbold B. (2009). *Transport mobility benefits and quality of life: A time-use perspective of elderly Canadians.* Transport Policy, 16(1), 1-11.

Stafford M, Marmot M. (2003). *Neighbourhood deprivation and health: does it affect us all equally?* International Journal of Epidemiology, 32(3), 357-366.

Statista. (2020a). *Mean disposable income in the UK by age, 2017/18*. [Online]. Available at: https://bit.ly/2Z5XZHa. (Accessed: 26/06/2020).

Statista. (2020b). *Retail e-commerce sales worldwide from 2014 to 2023*. [Online]. Available at: https://bit.ly/3f2oN1j. (Accessed: 08/06/2020).

Statista. (2020c). *Share of individuals who made purchases online in Great Britain from 2012 to 2019, by age*. [Online]. Available at: https://bit.ly/2XHKL3N. (Accessed: 08/06/2020).

Steinberg D. (2009). *CART: classification and regression trees.* The Top Ten Algorithms in Data Mining. Chapman and Hall, Bacon Raton, USA.

Stillwell J. (2016). *The Routledge handbook of census resources, methods and applications*. Unlocking the UK 2021 Census. Routledge, London, UK.

Tharwat A, Gaber T, Ibrahim A, Hassanien A. (2017). *Linear discriminant analysis: a detailed tutorial.* AI Communications, 30(2), 169-190.

Thomas E, Serwicka I, Swinney P. (2015). *Urban demographics: why people live where they do.* [Online]. Available at: https://bit.ly/3g1Anu9. (Accessed: 6/06/2020).

Titheridge H, Achuthan K, Mackett R, Soloman J. (2009). *Assessing the extent of transport social exclusion among the elderly*. Journal of Transport and Land Use, 2, 31-48.

Tomaney J, Pike A. (2020). *Levelling up?* The Political Quarterly, 91(1), 43-48.

Tran T, Drab K, Daszykowski M. (2013). *Revised DBSCAN algorithm to cluster data with dense adjacent clusters*. Chemometrics and Intelligent Laboratory Systems, 120, 92-96.

Transport for West Midlands (TfWM). (2020). *Public thanked for 86% drop in bus, train and tram journeys following coronavirus outbreak*. [Online]. Available at: https://bit.ly/2B4GSxu. (Accessed: 15/07/2020).

Transport for West Midlands (TfWM). (2018). *West Midlands chosen as UK's first Future Mobility Zone.* [Online]. Available at: https://bit.ly/390d8hW. (Accessed: 15/07/2020).

Transport for West Midlands (TfWM). (2017). *West Midlands Travel Trends 2017*. [Online]. Available at: https://bit.ly/3kkDYGB. (Accessed: 07/08/2020).

Transport for West Midlands (TfWM). (2016). *Bus boarding data*. Data Insight. [Online]. Available at: https://bit.ly/3kWJQGl. (Accessed: 07/08/2017).

Urban Transport Group. (2019). *What's driving bus patronage change? An analysis of the evidence base.* [Online]. Available at: https://bit.ly/38iC8k4. (Accessed: 29/06/2020).

Urry J. (2007). *Mobilities*. Polity Press, Cambridge, UK.

Vickers D, Rees P, Birkin M. (2005). *Creating the National Classification of Census Output Areas: Data, Methods and Results*. Working Paper 05/2. School of Geography, University of Leeds.

Viggiano C, Koutsopoulos H, Attanucci J, Wilson N. (2016). *Inferring public transport access distance from smart card registration and transaction data*. Transportation Research Record: Journal of the Transportation Research Board, 2544, 55-62.

Wagstaff K, Cardie C, Rogers S, Schroedl S. (2001). *Constrained k-means clustering with background knowledge.* Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, USA. 577-584.

Waller L, Gotway C. (2004). *Applied spatial statistics for public health data*. John Wiley and Sons, Hoboken, New Jersey, USA.

Walsh K, Scharf T, Keating N. (2017). *Social exclusion of older persons: a scoping review and conceptual framework*. European Journal of Ageing, 14, 81-98.

Wang F. (2007). *Meeting older persons' mobility and access needs – a re-think on public transport*. Institute of Transport and Logistics Studies, The University of Sydney, Sydney, Australia.

Weisberg S. (2005). *Applied linear regression*. John Wiley and Sons, Hoboken, New Jersey, USA.

West Midlands Combined Authority (WMCA). (2019). *Proposals drawn up to provide free travel to women worst effected by pension age changes*. [Online]. Available at: https://bit.ly/2ZmwOIh. (Accessed: 30/06/2020).

White P. (2009). *Factors behind recent bus patronage trend in Britain and their implications for future policy*. International Journal of Transport Economics, 36(1), 13-31.

Whitley E, Craig P, Popham F. (2020). *Impact of the statutory concessionary travel scheme on bus travel among older people: a natural experiment from England*. Ageing and Society, 40(11), 2480-2494.

Williams P, Hubbard P. (2011). *Who is disadvantaged? Retail change and social exclusion*. The International Review of Retail, Distribution and Consumer Research, 11(3), 267-286.

YouGov. (2018). *A nation of loyalty members: Three quarters of Brits belong to a loyalty scheme.* [Online]. Available at: https://bit.ly/2Uq7oIe. (Accessed: 09/06/2020).

Zahedi S, Shafahi Y. (2018). *Estimating activity patterns using spatio-temporal data of cell phone networks*. International Journal of Urban Science, 22(2), 162-179.

Zelterman D. (2010). *Applied linear models with SAS*. Cambridge University Press, Cambridge, UK.

Zhou J, Murphy E, Long Y. (2014). *Commuting efficiency in the Beijing metropolitan area: an exploration combining smart card and travel survey data*. Journal of Transport Geography, 41, 175-183.

Ziegler F, Schwanen T. (2011). *'I like to go out and be energised by different people': An exploratory analysis of mobility and wellbeing in later life*. Ageing and Society, 31(5), 758-781.