The Applications of Loyalty Card Data for Social Science

Alyson. S. Lloyd

Thesis submitted in conformity with the requirements of

Doctor of Philosophy (Ph.D.)

Department of Geography

University College London

September 2018

Declaration

I, Alyson Lloyd, 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:

Thesis Outputs

Peer Reviewed Journal Publications

2018 Detecting Address Uncertainty in Loyalty Card Data, ***Applied Spatial Analysis and Policy***, A. Lloyd, J. Cheshire

Book Chapters

2018 The Provenance of Customer Loyalty Card Data, A. Lloyd, J. Cheshire in ***Consumer Data Analytics***, UCL Press, P. Longley, J. Cheshire and A. Singleton.

Peer Reviewed Conference Proceedings

2018 Classifying Spatio-temporal Consumption Patterns in Loyalty Card Data, ***GISRUK 2018***, Leicester, UK. A. Lloyd and J. Cheshire.

2017 Detecting Address Uncertainty in Loyalty Card Data, ***GISRUK 2017***, Manchester, UK. A. Lloyd and J. Cheshire

Other Conference Presentations

2017 Detecting Address Uncertainty in Loyalty Card Data, ***Association of American Geographers Annual Meeting 2017***, Boston, USA. A. Lloyd and J. Cheshire

2016 Challenges of Big Data for Social Science, ***Oxford Retail Futures Conference***, University of Oxford, UK. A. Lloyd and J. Cheshire

2016 Challenges of Big Data for Social Science, ***CDRC Forum***, University of Oxford, UK. A. Lloyd and J. Cheshire

Prizes and Awards

2017 Best Paper by a Young Career Researcher – **GISRUK 2017**

Acknowledgements

I would first and foremost like to thank my primary supervisor, Dr James Cheshire, for his invaluable insight and guidance throughout this completion of this thesis. His expertise, generosity with time and advice was very much appreciated. Thank you also to my secondary supervisor, Dr Helena Titheridge, for contributing her ideas and perspectives throughout. I am further grateful to Professor Paul Longley for his contributions of knowledge and giving me the initial opportunity to carry out this work. Extra special thanks go to my family and friends, for their encouragement and support. I would finally like to acknowledge the ESRC for providing the funding for this work.

Abstract

Large-scale consumer datasets have become increasingly abundant in recent years and many have turned their attention to harnessing these for insights within the social sciences. Whilst commercial organisations have been quick to recognise the benefits of these data as a source of competitive advantage, their emergence has been met with contention in research due to the epistemological, methodological and ethical challenges they present. These issues have seldom been addressed, primarily due to these data being hard to obtain outside of the commercial settings in which they are often generated. This thesis presents an exploration of a unique loyalty card dataset obtained from one of the most prominent UK high street retailers, and thus an opportunity to study the dynamics, potentialities and limitations when applying such data in a research context.

The predominant aims of this work were to firstly, address issues of uncertainty surrounding novel consumer datasets by quantifying their inherent representation and data quality issues and secondly, to explore the extent to which we may enrich our current knowledge of spatiotemporal population processes through the analysis of consumer activity patterns. Our current understanding of such dynamics has been limited by the data-scarce era, yet loyalty card data provide individual level, georeferenced population data that are high in velocity. This provided a framework for understanding more detailed interactions between people and places, and what these might indicate for both consumption behaviours and wider societal phenomena.

This work endeavoured to provide a substantive contribution to the integration of consumer datasets in social science research, by outlining pragmatic steps to ensure novel data sources can be fit for purpose, and to population geography research, by exploring the extent to which we may utilise spatiotemporal consumption activities to make broad inferences about the general population.

Table of Contents

List of Figures

List of Tables

1. Introduction

Throughout history, public bodies have sought to measure and record populations to provide infrastructure for societal decision-making. Historically, the leading producer of these data has been government statistical agencies engaged in collecting them through statistically representative surveys. In the current data era, large-scale digital datasets are capturing highly detailed records of people's daily lives, such as their patterns of consumption, work, travel, communication, leisure, interactions with organisations and preferences across both space and time. These 'Big Data' have shaped what has been coined the 'fourth paradigm of science', a fundamental shift towards data-driven research. Commercial organisations have been quick to realise that utilising these can offer a source of competitive advantage. Yet, their importance has been implicated far beyond the identification of customer tastes and preferences. These data are capturing the characteristics and movements of active citizens, and thus offer new opportunities to both enhance human geographical understanding and better comprehend the nature and functioning of societies.

Whilst the benefits and promises of such data have been numerous, their emergence has raised substantial epistemological, methodological and ethical questions and there has been concern over premature adoption to inform a broad spectrum of social, economic, political, and environmental processes. For example, so-called big data are often generated as by-products of alternative processes, leading to a substantial lack of quality control and an inherent bias towards self-selected populations. This has important implications for the data's content and coverage when they are reused for research purposes. Yet, these issues have seldom been addressed, primarily due to these data being relatively hard to obtain outside of the commercial settings they are often created.

This thesis presents the exploration of one form of big data – loyalty card data – for applications within the social sciences and humanities. Loyalty card data offer a typical example of a contemporary data source, allowing compilation of behaviours that inform consumption characteristics and long-term spatiotemporal activity patterns. Customer metadata such as age, gender and postcode, collected on application, also provide a valuable geodemographic dimension to these data that can be attributed to transactional behaviours. Access to a large UK high street retailer's (HSR) data was brokered through the Consumer Data Research Centre (CDRC), a big data initiative funded by the Economic and Social Research Council (ESRC). Such data have not previously been obtainable for academic endeavours on such a scale, therefore, this offered a unique opportunity to study the dynamics and applications of a commercial dataset for social science research.

The overarching aim of this thesis is to provide a substantive contribution to understanding how new forms of data may be employed in the study of populations. Traditionally, this has been achieved through census data, which essentially provides a static depiction of populations between extended time periods. This work seeks to establish to what extent a consumer dataset can function as an alternative, or supplement to, conventional sources. However, before endeavouring to repurpose these data to inform such phenomena, it is critically important that we begin by questioning the assumptions, values, and biases of this new wave of research. Largely due to access barriers, there is a substantial lack of understanding of how consumer datasets can be pragmatically applied to such causes. Therefore, the aims of this work were also to provide empirical evidence for the limitations and considerations necessary when attempting to do so.

It should be noted that these aims did not endeavour to develop specific and wide-ranging applications for implementing these types of data. For example, in the first instance, the extensive variety of existing consumer datasets means that each will be unique in its characteristics and applications outside of the context of this data would be inherently limited. However, in addition, given the current status of this underdeveloped area of research, it was considered important to make best use of access to this unique dataset to develop a framework upon which future researchers may analyse and interpret such data. For example, by outlining pragmatic steps that can be taken and the types of insight that might be obtainable from a novel, inherently uncertain consumer dataset.

1.1. Aims

As summarised above, the purpose of this research is to provide a thorough exploration of the provenance of loyalty card data, in order to understand its uses as a novel form of data in social science research. The more specific aims of the analysis can be understood across two main themes. Firstly, to understand the challenges encountered when applying a commercial data in a research context. This required investigation of the dynamics of these data and uncertainty/representation issues when inferring insights about the general population. This served to inform the inherent dynamics and limitations of these data to advise the proceeding analyses, however, also to provide a framework of necessary considerations for future research in this area.

Whilst this was an important step, it formed the preliminary stage of analyses on these data. The second key aim was then to understand how we may utilise a consumer dataset to enrich our understanding of population processes. The lack of incorporation of dynamic spatiotemporal activities (i.e. daily, weekly, seasonal patterns) has been recognised as one of the key limitations to our existing conceptualisations of geodemographic phenomena. Yet, a prominent advantage

of loyalty card data is their high temporal resolution and provision of both residential and consumption locations. Therefore, analyses in this thesis are focused particularly on the classification of spatiotemporal consumption habits and their relationship with the characteristics of both people and places. On this basis, the thesis has four broad aims:

1. To review current practices and perspectives in the study of populations and highlight opportunities for progression with a novel consumer dataset.
2. To assess and quantify data quality issues inherent in loyalty card data, and outline pragmatic ways of addressing them.
3. To explore the ability to extract spatiotemporal activity patterns from loyalty card data, and their relevance to making inferences about the general population.
4. To deliver recommendations about what loyalty card data, and consumer data more broadly, can contribute in terms of population insight and highlight prominent areas of future progression.

An underlying theme of all objectives in this work is evaluating the relevance of these data in terms of the general population and thus appraisal of their potential applications for matters of public and social good. This represents a sharp contrast to commercial endeavours, which are often focused on gaining an innate understanding of their specific consumer population in order to maximise profits. It was hypothesised that loyalty card data, both alone and in combination with other datasets, may advance our knowledge of the functional relationships between people and places.

The first substantive focus of this thesis (Aim 2) is reported in Chapters 3 and 4. These are concerned with quantifying uncertainty within loyalty card data and developing data-driven heuristics to address them. The second substantive focus (Aim 3) is reported in Chapters 5 to 7, which presents a three-fold analysis of extracting spatiotemporal activity patterns from loyalty card data. Together, they outline the types of insight we may derive from a novel consumer dataset in regards to complex interactions between people and places. A more detailed outline of this thesis structure is provided over the proceeding sections.

1.2. Thesis Structure

1.2.1. Chapter 2 - Literature Review

This chapter provides an overview of concepts and literature relevant to the explorations and analyses conducted throughout this thesis, demonstrating their value as an area of investigation. This includes, firstly, an overview of the current data landscape, the applications of novel forms of data in population research and key limitations that need addressing if research in this area is to progress. This is followed by an overview of traditional and current practices regarding the

study of people and populations, limitations of these practices and the potential uses of consumer data as population indicators. Finally, an overview of the provenance of loyalty card data is provided, discussing their potential uses as a social and spatial data source.

1.2.2. Chapter 3 - Data and Preliminary Analyses

Chapter 3 provides a detailed description of the attributes and characteristics of the loyalty card data utilised in this work. Following this, a preliminary analysis is presented of their various data quality issues, including data error, their representativeness in terms of the general population, and other data treatment processes that are necessary in order to extract meaningful insights. The purpose of this chapter was to inform interpretation of the analyses presented in Chapters 4 to 7, but also to provide data-driven evidence for many of the issues outlined in Chapter 2, such as the nature and dynamics of loyalty card data and pragmatic steps required when aiming to reliably integrate them into academic research.

1.2.3. Chapter 4 – Detecting Address Uncertainty in Loyalty Card Data

Chapter 4 provides an extension of the analyses aiming to quantify uncertainty in loyalty card data. A key requirement when harnessing these data for geographical insights is the ability to accurately link individuals to their residential location. Yet, the exploratory analyses from Chapter 3 indicated that the veracity of address information in loyalty card data is inherently uncertain. This chapter presents the development of data-driven heuristics that utilised customer transactions to estimate the credibility of their address information, by drawing on current knowledge and theory of spatial behaviour. Following this, results are contextualised through augmentation with census statistics and the effectiveness of the method discussed. These outputs informed data cleaning measures required for subsequent analyses, and also served to enforce considerations of, and solutions to, uncertainty in big data.

1.2.4. Chapter 5 - Temporal Profiling: Classifying Stores

Chapter 5 presents the first substantive analysis of spatiotemporal consumption dynamics using loyalty card data. Justified by the limitations of current population studies outlined in Chapter 2, this chapter aimed to provide data-driven evidence for the temporal rhythms of HSR store locations, and what these may indicate about the characteristics and functions of those locations. To achieve this, a cluster analysis of HSR store locations was conducted, using transaction frequencies over time. This endeavoured to understand how consumers interact with different location types, and if the distinctive characteristics of those places can be inferred from the temporal population flows they exhibit. Methodological steps are outlined, including considerations of data treatment and clustering methods. Following this, the outputs of a bespoke, temporal HSR store classification are presented and their locational attributes

explored. The implications of integrating temporal dynamics, and thus moving away from static classifications of places, are discussed.

1.2.5. Chapter 6 – Classifying HSR Customers

Following on from the previous analysis, Chapter 6 presents a segmentation of HSR customers based on their interactions with HSR location types that exhibit distinct temporal profiles, and explores what their spatiotemporal activities may indicate about their geodemographic identities. This analysis aimed to understand if focusing on everyday activities in the HSR population may reveal how the rhythms of people and retail spaces are ordered, and how these orderings may vary by social group. Methodological steps are outlined pertaining to the segmentation of individuals based on spatiotemporal habits, followed by an exploration of their characteristics in terms of geodemographic and consumption characteristics. The chapter concludes by discussing the integration of dynamic activity patterns to enriching representations of people and places.

1.2.6. Chapter 7 – HSR Areas and Activities

Chapter 7 intended to supplement outputs from the customer classification derived in Chapter 6 and demonstrate how loyalty card data may be utilised to quantify the distinct location-visiting patterns of customers who exhibit specific spatiotemporal rhythms. A secondary aim was then to explore how these characteristics varied over different geographical regions. The method presents an augmentation of customer activity patterns with the 2011 Census based COWZ classification and an analysis of behaviour over weekly and seasonal intervals and how these patterns vary regionally across England and Wales. Outputs demonstrate distinctions in the types of places that various social groups interact with, how these may vary based on the types of places that are accessible to regional populations, and serve to enforce the types of insight that may be extracted from loyalty card data.

1.2.7. Chapter 8 – Discussion, Applications and Research Prospects

Chapter 8 consolidates the principal findings from this work. Key methodological and knowledge contributions are highlighted in the context of loyalty card data, but also more widely for the integration of consumer data in population studies. Implications of incorporating dynamic spatiotemporal representations of activity patterns, as facilitated by novel consumer datasets, are discussed from the perspective of academics, public bodies and retailers alike. This discussion concludes by highlighting paths for future developments.

1.3. Note on Software and Code

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

1.4. Ethics

This research was approved by the UCL Research Ethics committee (Project ID: 9363/001).

2. Literature Review

This chapter provides an overview of concepts and literature relevant to the explorations and analyses conducted throughout this thesis. Section 2.1 provides background context to the work, including an overview of the current data landscape, changes this data era has incurred for scientific research and fundamental challenges that need addressing if we are to utilise novel data sources for population insight. Section 2.2 then provides an overview of traditional and current practices regarding the study of people and populations, in both the social sciences and commercial organisations. This includes the use of geodemographics to summarise complex population processes, considerations of their pitfalls, and how consumer data may enrich our understanding of such phenomena. In particular, how their inherent velocity may enhance our understanding of the temporal rhythms of both people and places. Section 2.3 then provides an overview of the provenance of loyalty card data and its potential as a source of social and spatial data.

2.1. Novel Sources of Population Data

2.1.1. The Current Data Landscape

Population data are a key resource, providing insights into societal phenomena that are utilised by governments, businesses and academia in order to monitor, regulate, profit from and make sense of the world. The production and collection of these data has traditionally been time-consuming and costly, resulting in static and often coarse representations of reality (Kitchin, 2014b). Yet, coinciding with both computational and technological advances, the data landscape has experienced a vast transformation in recent years, characterised by a colossal growth in the production and storage of data in a multitude of forms. A large proportion of these data contain both spatial and temporal references, describing when and where societal interactions occur and thus digitise a broad spectrum of social, economic, political, and environmental processes (Graham and Shelton, 2013). As a result, the variety of data producing systems that represent the interactions of everyday life – such as work, consumption, travel, communication and leisure - are now unprecedented.

The ‘Big Data’ being produced have received much attention, yet formal definitions remain contentious in literature. The most renowned has been Laney (2001)’s three Vs - Volume,

Variety and Velocity, yet various other 'V's have since been added to this list (most commonly, Value and Veracity). These can be summarised as follows:

- *Volume* – consisting of terabytes or petabytes of data, yet also refers to their size, scale and the (often large) number of dimensions.
- *Velocity* - data are often collected continuously and thus have a high temporal resolution (i.e. second-by-second).
- *Variety* – can refer to the diversity of information within a single big dataset (intra-data variety) or the diversity of datasets that fall under the big data umbrella (inter-data variety). These can be collected as structured, unstructured (which lack the structural organisation required by machines for analysis - i.e. text, images or audio), semi-structured, or mixed data.
- *Value* – refers to the value that the collected data can bring to the intended process.
- *Veracity* – refers to uncertainties surrounding data quality, which can be influenced by a number of factors including data origin, collection and processing methods.

Definitions have by no means been limited to these characterisations (see for example, Dutcher, 2014, for over 40 industry definitions or Press, 2014, for a review of 12 definitions). It is recognised, however, that size is not the primary defining factor of these data. For instance, government, academia and industry have long produced 'large' population datasets, such as national censuses. Yet, differentiations are apparent in that these have been produced manually and in tightly controlled ways, using sampling techniques that limit their coverage and size (Miller, 2010). In contrast, big data are being generated through automated, continuous systems. It is therefore velocity - and the additional 'V's that manifest as a result of their generation - that set big data apart from traditional data repositories and infrastructures (Kitchin, 2013).

These data can be further conceptualised by the nature of their production. This includes directed, automated and volunteered data:

- *Directed* - generated from digital forms of surveillance on a person or place by a human operator (such as CCTV or passport control collecting passenger details).
- *Automated* - generated as an automatic function of a device or system. Examples include traces from digital devices (i.e. smart phones), retail transactions, clickstream data (e.g. interactions on a website or app), sensor data (generated by sensors e.g. temperature or travel speed) or scanning of travel passes.

- *Volunteered* – generated by volunteered interactions, such as from social media (e.g. Twitter) or crowd-sourced data (where users contribute to a system), such as OpenStreetMap or Wikipedia.

These three forms of generation now mean that more data are being produced every 2 days than in all of history prior to 2003 (Strong, 2015). Large proportions of these data are georeferenced and therefore specify observations or facts about some location (Goodchild, 2013). Currently, the majority of georeferenced big data are being generated through location-based services such as mobile devices (Laurila et al., 2012). However, they exist in many forms including those collected with in-built locational attributes (i.e. georeferenced sensors or web content) or those that can be attributed to a spatial referencing system, such as residential postcodes.

Of particular relevance to this thesis is the vast volume of georeferenced, automated data that arise out of transactions between individuals and service organisations, known as consumer data. These data digitalise vast amounts of population consumption activities in both space and time, and now account for an increasing share of novel big datasets. Examples include (but are not limited to): store transactions, ecommerce/web interaction or online ordering, energy consumption and public transport usage. These data have long been recognised and utilised as a source of competitive advantage by the organisations producing them. For instance, to gain an innate understanding of when, where and what people consume, the frequency of consumption and the general nature of their habits, in order to maximise profits. However, from a non-commercial perspective, these data essentially document rich behavioural information pertaining to large numbers of individuals, and thus represent one of the many novel forms of population data that were not attainable prior to the data deluge.

2.1.1.1. Novel population data and their uses

The most prominent consequence of the emerging data landscape for the social sciences, and a key motivation of this research, is the potential to develop a greater understanding of population processes that can be used to benefit society as a whole. It is now commonly asserted that information derived from big data is likely to be one of the foundational elements for understanding future societies, for example by generating real time information about economic and social activity, or by generating new insights into human behaviour (Einav & Levin, 2014).

Data have long provided an infrastructure for individual and societal decision-making. Historically, the leading producer of these data has been statistical government agencies engaged in collecting them through large-scale statistically representative surveys. This has been a result of the scale of this process (generating large representative samples is expensive

and time consuming), but also public trust in the government to protect confidentiality through appropriate disclosure controls (Kitchin, 2014b). This administrative data comprises of information such as health, education, housing and tax records (UKDS, 2018a) and are utilised to improve the functioning of societies. For example, by evaluating economic performance and informing allocation of public services.

The public has clearly been served by the careful creation and dissemination of data from administrative sources (Lane et al., 2014). However, there is abundant evidence that access to big data can lead to even more profound social and economic benefits. Verhulst (2015) summarises their key contributions to public and social good from numerous cases studies, including: improving government, empowering citizens, creating opportunities and solving public problems. For example, open data projects are improving government, primarily by making government more accountable and efficient; empowering citizens, by facilitating more informed decision-making and enabling new forms of social mobilisation; creating new economic opportunities; and helping policymakers by providing solutions to public problems (e.g., related to public health or global warming). Specific examples of how big data can facilitate improved societal outcomes, as summarised by Verhulst, Young and Srinivasan (2018), include:

- *Improving situational awareness* – for example, Facebook's partnership with humanitarian organisations (such as UNICEF) in sharing locations, movement, and self-reported safety to help better understand demographic trends and the geographic distribution of various phenomena (for example, the spread of disease).
- *Increased knowledge creation and transfer* – for example, joining datasets to create a better understanding of correlations and causalities between societal issues.
- *Public service design and delivery* - private data sets often contain a wealth of information that can enable more-accurate modelling of public services and help guide service delivery in a targeted, evidence-based manner. An example comes from the open access of Transport for London's Application Programming Interface (API), and the huge subsequent economic benefits (TFL, 2017).
- *Optimised prediction and forecasting* – richer, more-complete information may enable new predictive capabilities for policy makers, allowing them to be more proactive.
- *Impact assessment and evaluation* – data collaborations can aid in monitoring, evaluation, and improvement, for example, public-interest bodies can rapidly assess the results of their actions (such as from social media).

These examples, amongst many others, have and continue to have vast implications across many domains, including for transportation systems, energy use, health, innovation, creation of jobs, increased efficiency, boosting of economies and understanding a wide range of population activities (Batty et al., 2012; Granickas, 2013). It is evident that access to these data can lead to the creation and continuous updating of new infrastructures that may ultimately contribute to a more effective governing of society.

2.1.2. The Fourth Paradigm of Science

Before further exploring the implications of these novel forms of data, it is first necessary to provide context surrounding their impacts on current research practices. Many have argued that big data has induced a scientific revolution, which is changing how knowledge is produced, business conducted, and governance enacted (Anderson, 2008; Bollier and Firestone, 2010; Floridi, 2012; Mayer-Schonberger and Cukier, 2013). To date, scientific research has been conducted in a data-scarce environment and methods have been developed to reflect this. Hey, Tansley and Tolle (2009) outline the three traditional paradigms of scientific enquiry, including *experimental science* (describing natural phenomena, also known as empiricism), *theoretical science* (modelling and generalisation) and *computational science* (simulation of complex phenomena). Theoretical science has been particularly prominent in the social sciences and humanities, where variables are defined based on pre-existing theory of what may provide insight about the topic of interest and data actively solicited through qualitative methods (i.e. surveys, interviews or case studies). As a result, conclusions are typically drawn from limited sample sizes and statistical methods have been developed to generalise findings to larger populations.

However, the emergence of big data has facilitated a paradigm shift towards what has been termed ‘data-driven science’. This seeks to generate hypotheses and insights ‘born from the data’ rather than ‘born from the theory’ (Kelling, Hochachka and Fink 2009), and has been coined the ‘fourth paradigm of science’ (Hey et al., 2009; Kitchin, 2014a). Whilst traditional methods have been underpinned by hypotheses to test theories, data-driven science challenges withstanding research epistemologies through its blending of abductive, inductive and deductive approaches (Kitchin, 2014a; Quan-Haase and Sloan, 2017). *Inductive reasoning* can be understood as taking specific instances known to be true and applying them to a generalised (often uncertain) conclusion. *Deductive reasoning* starts with generalised instances known to be true and applies them to a true and specific conclusion. *Abductive reasoning,* on the other hand, starts with an observation then seeks to find the simplest and most likely explanation.

Data-driven science, as Quan-Haase and Sloan (2017) summarise, uses a hybrid combination of these techniques. For instance, it incorporates an element of induction into research designs to guide knowledge discovery based on pre-existing theory. Yet, explanation through induction is not the intended end-point (as is with empiricist approaches). Instead, this provides the basis for the formulation of new hypotheses that can be deductively validated. Essentially, the fourth paradigm does not rely on hypothesis testing but rather the exploratory mining of data to extract patterns, which may be guided by, but is not driven by, pre-determined theories.

The emergence of georeferenced big data has also facilitated a shift from quantitative geographic research to what has been termed 'data-driven geography' (Miller and Goodchild, 2015). The proliferation of data that capture factual, quantified information about what is happening at particular places and times has allowed for a revolutionised way of studying the behaviours of individuals and populations, providing a foundation to robustly measure human interactions (such as counts, distance, cost, and time) on a large scale. Potential advances for quantitative geography, as Batty et al., (2012) point out, are that firstly, studies in the data-scarce era have focused largely on radical and massive changes to places over the long-term, with little ability to capture small spaces and local movements. Conversely, data-driven geography allows greater focus on the local and routine, providing deeper descriptions of what is happening and where (Miller and Goodchild, 2015). Secondly, it allows us to capture spatiotemporal dynamics at more granular intervals and at multiple scales. These data are collected continuously, meaning that both mundane and unplanned events can be captured.

However, this shift in scientific focus has not been without contention in literature. Mayer-Schonberger and Cukier (2013) summarise three fundamental challenges of data-driven science:

- *Populations, not samples* - Analysis techniques have traditionally been designed to extract insights from small, scientifically sampled data generated and analysed with a specific question in mind (Miller, 2010). However, data-driven methods advocate descriptive insights of voluminous populations. Whilst this may greatly increase the potential for knowledge, it also calls for novel methods of analysis and preliminary considerations that were not necessary when the samples being collected were under the control of the researcher.
- *Messy, not clean data* – Big data are created as a by-product of various processes and do not adhere to traditional collection practices. These datasets will be used for topics that are far removed from their original purpose (Goodchild and Longley, 1999) and their provenance therefore needs to be understood. Their inherently error-prone nature means that new data treatment methods are required to understand and clean them.

- *Correlations, not causality* - Data-driven science advocates correlation over causality, favouring identification of observed relationships rather than the causes of such phenomenon. Some have argued that 'correlation in enough' (Anderson, 2008; Prensky, 2009). However, causality may be important in many areas of social science (Walker, 2014) and correlations between variables can be random or spurious in nature (Kitchin, 2014a).

Despite these challenges, it is argued by some that data-driven science will become the new paradigm of scientific method because the epistemology is suited to extracting additional, valuable insights that traditional 'knowledge-driven science' would fail to generate (Kelling et al., 2009; Loukides, 2010; Miller, 2010; Kitchin 2014a). Some have postulated that this era represents 'the end of theory' with 'the data deluge making the scientific method obsolete' (i.e. Anderson, 2008). However, others argue that it simply presents a reconfigured version of traditional scientific method, providing a new way in which to build theory (Quan-Haase and Sloan, 2017).

2.1.3. Big Data Challenges

It is undisputed that the emergence of big data has brought with it the potential for greater information and knowledge. However, these data have only recently become established for repurposed academic use and present a multitude of issues to overcome if we are to reliably extract insights. The fundamental challenges facing the integration of these data in social science research are described throughout the proceeding sections, summarised across three key areas; acquisition, legal and ethical challenges, data uncertainty challenges and analysis challenges.

2.1.3.1. Acquisition, legal and ethical challenges

One of the main prospects of emerging big datasets is that they capture detailed interactions and transactions across space and time. However, as a result, these data are often personal in nature and bring substantial ethical and legal considerations to the fore that mean that access has to be limited. Such issues have informed the research throughout this thesis, particularly in the presentation of results.

Privacy is considered a basic human right and typically refers to acceptable practices with regards to accessing and disclosing personal and sensitive information (Elwood and Leszczynski, 2011). Privacy was protected in the UK, prior to May 2018, by the Data Protection Act (1998), and now the more stringent General Data Protection Regulation (GDPR, 2018) and there are various ways in which it can be breached (see Solove, 2006, for a review). Personally

sensitive data that come under legislation include anything that might be directly or indirectly embarrassing or detrimental to the identity of a person, such as: names, postcodes, email addresses, age, gender, preferences, political opinions and health/medical information, amongst many others.

The implications of the current data era on the concept of privacy has been the subject of much debate, with views varying largely between those with differing agendas (Kitchin, 2014b). It is a consensus, however, that this era has blurred ethical boundaries and facilitates circumstances in which individuals could be exploited. Prior to May 2018, privacy legislation was largely constructed around consent regarding the generation, use, and disclosure of personal data (Solove, 2013). However, as Kitchin (2014b) outlined, there were several cognitive and structural problems with existing legislation. For instance, whilst data holders were required to adhere to data protection laws, which is usually achieved through signing of a Privacy Policy (Milne and Culnan, 2004), it was demonstrated that people did not read or understand these policies (Nissenbaum, 2011) and did not anticipate that their data might be processed, packaged and sold on (Kitchin, 2014b). Thus, 'privacy policies often serve more as liability disclaimers for businesses than as assurances of privacy for consumers' (Tene and Polonetsky, 2012).

The recent implementation of GDPR has introduced more stringent and transparent rules on the storage and treatment of personal data by organisations, and introduces a wider definition on 'personal data' (i.e. see The General Data Protection Regulation & Social Science Research, CDRC, 2018). Yet, ethical debates surrounding the concept of privacy in the big data era are extensive and ongoing. The primary challenge facing the integration of these data into research is how a balance is achieved between protection of privacy and extraction of insights. Currently, when handing these data, legislation means that disclosure control is necessary to safeguard confidentiality. In general, the solution is to use anonymisation techniques such as deidentification (removing personally identifiable information), pseudonyms or aggregation, alongside encryption, secure storage and access limitations (Kitchin, 2014b). However, in the current era where locational information is routinely stored, disclosure is becoming increasingly challenging (Esayas, 2015). For example, removing sensitive information such as names and addresses is often not sufficient (Karr and Reiter and Lane, 2014) as linkage to other georeferenced data can facilitate the re-identification of individuals (Reiter, 2012). The most commonly applied disclosure technique in spatial terms is that of aggregation, for example, using coarser scale geographic units (Karr et al., 2014). However, these methods are of varying degrees of value as they often do not preserve relationships (Nowok, Raab and Dibben, 2015).

These issues have had a substantial influence of the ability to acquire data outside of commercial settings. Yet, on top of these restrictions, data have also been tightly held by organisations in order to maintain competitive advantage. There have been some developments in alleviating these barriers in recent years through data collaboratives. These emerging public-private partnerships are allowing participants to exchange data and combine analytical expertise to create new public value (Verhulst et al., 2018), which have largely taken the form of research partnerships (i.e. corporations sharing data with universities and researchers), trusted intermediaries, such as the ESRC funded CDRC, or direct access to corporate data streams through Application Programming Interfaces (APIs).

There has therefore been some progress in getting organisations to participate in causes that endeavour to access data for the public good and pursue research questions that contribute to understanding society. This was also encouraged by the Digital Economy Act (2017), which included provisions for businesses to assist in the compilation of National statistics. Nevertheless, the process of obtaining these data still requires highly sensitised considerations of the risks of disclosure and appropriate restrictions to access. Limited by these private ownership and access restrictions, the potential of big data for public good has so far gone largely untapped.

2.1.3.2. Data uncertainty challenges

Big data are created as a by-product of alternative processes, with no researcher control, which raises substantial methodological questions when applying them in research. The fundamental issues are that firstly, the quality of the data is unknown, and secondly, the sample population is unknown. As Longley et al., (2015) summarise, novel data sources have no obvious population reference points, and have not been created with any scientific sampling. Therefore, it is necessary to understand uncertainty in terms of data quality and fitness for purpose in order to extract and interpret meaningful insights.

Data quality is a broad concept that the International Organisation for Standardization (ISO) define as the "totality of characteristics of a product that bear on its ability to satisfy stated or implied needs". This encompasses a number of considerations including accuracy, completeness, vagueness, ambiguity and precision (Wang et al., 2005), consistency, scale, coverage, sample size and bias (Harris and Jarvis, 2014). Firmani et al., (2016) provide a detailed overview of these concepts, which can lead to various forms of the established scientific notion of data error – defined as the difference between reality and our representation of it (Heuvelink, 1999).

There has been a long history of controlling for these issues in quantitative research. However, where novel datasets are concerned, aspects of the production process are often unknown and preliminary data treatment is therefore required in order to understand their dynamics. This is impeded by the lack of reference data by which to validate against. Aspects of accuracy can be easily identifiable in some cases, for example spelling errors or syntax mistakes may be simple indicators. However, many will be more difficult to detect, such as where admissible but incorrect values are provided (Firmani et al., 2016). Similarly difficult to detect will be spatial errors, which have the potential to obscure, rather than reveal social and spatial processes (Graham and Shelton, 2013). Fisher (1999) outlines the various errors that can arise within georeferenced databases, including; entry errors (data are miscoded electronically), measurement errors (the property of a measure is erroneous), assignment errors (an entity is assigned to an incorrect class), class generalisation (generalisation that may be applied before digitalising, including spatial generalisation), processing errors (transformations that might occur due to rounding or algorithm error), but also temporal errors, where an object being represented changes character between the time of data collection and when the data are utilised. Further uncertainty can be introduced by incompleteness of data such as missing values or partial definition (Wang et al., 2005).

Solutions to issues of error in big data have been discussed in literature. Miller and Goodchild (2015) suggest that we can either restrict the assumptions and generalisations drawn from such analyses, or attempt to clean and verify the data. Goodchild and Li (2012) argue that traditional methods including, 1) the crowd solution, 2) the social solution and 3) the knowledge solution, will be particularly useful in the big data era. Most relevant to this work is that of the *knowledge solution*, which postulates that we may draw on existing theory to ascertain whether or not purported fact is false, or likely to be false. Consequently, we may attempt to create informed assumptions based on what has been termed 'logical consistency' in Geographical Information Science (GIS) literature (i.e. see Guptill and Morrison, 1995), and whether an observation is consistent with what is already known about the geographic world. However, as noted by Miller and Goodchild (2015), the development of explicit, formal, and computable representations of geographic knowledge can be a challenge in itself. Research has also not yet been able to apply these solutions in practice.

The second fundamental area concerning uncertainty in big data is representativeness, which refers to how well data capture the phenomena they seek to represent, and how well the sample of data represents the overall population (Kitchin, 2014b). Whilst traditional data have suffered with issues of sampling error (i.e. when a randomly chosen sample doesn't represent the underlying population by chance), big data suffer from sampling bias, where the sample isn't

randomly chosen at all. Due to the nature of their production, big data are self-selected rather than sampled populations and are inherently biased towards those who fall within the scope of the particular markets or activities that are being tracked. Thus, these data are still inherently a sample and representative of a set of people, even if that set is very large.

Similarly to many topics of uncertainty in big data, research has not yet been able to quantify representativeness dynamics due to lack of access. Some progress has been made in relation to more accessible contemporary datasets such as Twitter data (made available through their API), demonstrating how we can attempt to triangulate novel data with more conventional, administrative sources in order to ascertain representativeness. Longley et al., (2015) and Lansley (2014) illustrate examples of applying this, highlighting how efforts must be made to ascertain their relevance to the behaviours of the general population to avoid substantial generalisation pitfalls.

2.1.3.3. Analysis challenges

More practical challenges arise when attempting to analyse big data. The ways in which analysts can extract insight are also being revolutionised (Miller and Goodchild, 2015), for example, efficient methods are now necessary to process large volumes of diverse data into meaningful comprehensions (Gandomi and Haider, 2015). The potential of these novel data sources can only be realised if we are able to robustly extract these insights, yet, the characteristics of big data raise a number of foundational issues.

These challenges are twofold. Firstly, *data challenges* refer to the characteristics of these data (primarily their volume) leading to difficulties when applying traditional analysis techniques. For example, to date, methods have been developed to support small controlled samples, which are rooted in statistical inference. Inference has been required as collecting data from entire populations has been both prohibitively expensive and time consuming (Levy and Lemeshow, 2013). Therefore, these methods aim to quantify the probability that a measured characteristic in a sample is true of the population from which the sample was taken (Summerfield, 1983). Sampling and statistical inference therefore allow us to draw conclusions that are representative of the population from which the sample was drawn, without collecting data from every entity in the population. These methods have formed the very basis of statistical analysis for decades. In contrast, as some argue that big data represent ‘populations, not samples’ and are not designed to produce valid and reliable data amenable for scientific analysis (Lazer et al., 2015), traditional inferential statistics are no longer relevant tools. This calls for the development of novel methodologies that differ from established statistical data analysis.

Secondly, *process challenges* refer to identifying the correct tools and approaches to analyse these data effectively and proficiently, given the complexity of big data and the scalability of available algorithms (Candela, Castelli and Pagano, 2012). The management of massive datasets is substantially different from traditional data in terms of scale, and new methodological frameworks are required for data mining, analytics, visualization and modelling (Lane et al., 2014). An in depth review of these issues can be found in Sivarajah et al., (2017), Gandomi and Haider (2015) and Fan, Han and Liu (2014).

2.1.4. Summary

Our ability to produce, capture and store digital information has transformed our current data landscape and as a result, the variety of data producing systems is now unprecedented. Big data offers a significant opportunity to develop more sophisticated models of human interactions. However, they have only recently begun to become available for academic use and it is commonly observed in literature that there are a number of challenges to overcome, particularly in terms of data quality and representation. Important developments in this area can only arise from applying data-driven approaches to quantify uncertainty, through continual critique, truth propagation and contextualisation with contemporary social and geographical theory and administrative sources to understand their fitness for purpose. These steps are fundamental if we endeavour to explore their uses as novel sources of population data, yet access barriers have hindered developments.

2.2. Big data, Areas and Activities

Throughout history, much research has endeavoured to disentangle and summarise complex population processes due to their widespread value across government, academia and business. This section introduces traditional approaches to quantifying such processes and their applications, the growing importance of conceptualising societal dynamics through spatiotemporal population activities and the use of consumer data as indicators.

2.2.1. Understanding Society: Classification and Geodemographics

When aiming to understand human behaviour, we are faced with the issue of representing complex systems in a simplified format in order to comprehend them. This raises questions as to what to represent and how to represent it (Longley, 2005). Classification - the arrangement of entities into taxonomic groups according to observed similarities (Brenner, Staley and Kreig, 2005) – offers answers through reducing the complex dimensions of real life into more manageable chunks. From these we can infer dynamics without the noise created from individuals' interactions with the world. For the study of populations, geodemographic

classifications (or ‘geodemographics’) represent the analysis of people by where they live (Longley, 2017). They aim to encapsulate socio-spatial characteristics by quantifying how socioeconomic and behavioural factors vary across space.

The principle theory behind geodemographics is that similar people tend to cluster together thanks to the long-established presence of homophily: the pervasive fact that cultural, behavioural, genetic, and material networks tend to be localised and associations between similar people occur at a higher rate than among dissimilar people (McPherson, Smith-Lovin and Cook, 2001). Geodemographics represent the quantification of homophily, which manifests as neighbourhoods with similar demographic and psychological compositions, such as age, sex, race/ethnicity, education, intelligence, attitudes, and aspirations (e.g., Loomis, 1946, Richardson, 1940). The primary assumptions of geodemographics are that components of human identity arising out of social, economic, or demographic circumstances are likely to:

1) Cluster together in space (i.e. people living in the same neighbourhood are more likely to have similar characteristics than two people chosen at random) and,
2) Recur across multiple different locations (i.e. two neighbourhoods can be placed in the same category even though they are widely separated).

Geodemographic classifications are typically produced by means of cluster analysis, in order to identify a number of distinct socioeconomic groups. These have traditionally been derived from national census data, using variables such as age, household composition, income, occupation, education, ethnicity and religion, amongst many others. Based on clustering outcomes, a social profile is appended to an area, such as ‘Student Living’ for a student-populated neighbourhood or ‘Thriving Greys’ for an area dominated by affluent, older residents (Batey and Brown, 1995). In the UK, these profiles are produced at a small area level, such as Output Areas (OA; typically consisting of 5-10 postcodes), which provide stable and consistently sized areas over which to describe neighbourhood statistics whilst also preserving respondent confidentiality. These classifications exist both in open forms, such as the ONS's Output Area Classification (OAC; Gale et al., 2016), or proprietary such as CACI's Acorn (CACI, 2014) and Experian's Mosaic (Experian, 2018).

Such classifications have found uses across multiple industries (Longley, 2005). Commercial organisations have long recognised that consumer behaviour is partially driven by personal circumstance and neighbourhood influence, which plays an integral role in marketing based decisions. For instance, markets are typically conceptualised using ‘segmentation’ (the idea that any market can be broken down into different types) by measures of activity such as shopping

frequency, total number of purchases, spend and items purchased (Allaway, 2006), which is then augmented with geodemographic information to enrich customer profiles. In addition to census data, organisations often supplement these representations with other consumer data such as from lifestyle surveys or consumer registers (Harris, Slight and Webber, 2005).

The resulting data have applications for many business problems (Leventhal, 2016) such as understanding relationships between consumer purchasing behaviour and population characteristics (such as lifestyles and social attitudes) and subsequently identifying the best locations from which to serve their customer base (Longley et al., 2015; Clarke, 1999; Longley et al., 2005; Leventhal, 2016). In the public domain, the same classifications are used to facilitate resource allocation decisions of public goods, such as for policing (Ashby and Longley, 2005), education (e.g. Singleton and Longley, 2009), and health (e.g. Kandt, 2015b). There have also been widespread applications in academia, most notably within the domains of health and well-being, education, environmental/resource management and crime (Singleton and Spielman, 2014, provide a general overview of academic uses).

2.2.1.1. Limitations of geodemographic classifications

The use of geodemographics across academia, commercial and public sectors demonstrates their widespread value. However, current geodemographic practices also suffer from a number of limitations. Longley (2017) summarises these across two main themes, *substantive limitations* and *practical limitations*.

Substantive limitations have primarily arisen due to existing classifications being reliant on obtaining quality georeferenced population data. Within the UK, the census is still considered the most reliable source due to its high quality and coverage. However, the decennial nature of data collection, along with delays in making it publically available, mean that representations are essentially out of date before they are even published (Longley, 2017). In addition, census variables are predefined by government specifications and assumed to be suitable indicators of social, economic, and demographic factors, which has been criticised as an approach to summarising real world complexities (i.e. Openshaw, Blake and Wymer, 1995; Voas and Williamson, 2000). They can also lack consistency between regions, of which an example comes from the three regional Censuses across the UK where specific questions can be inconsistent, limiting the ease with which UK-wide analysis may be performed, in addition to Scotland and Northern Ireland employing different geographies for some data sources. Finally, the manual collection of census data means they are susceptible to self-reported survey limitations, such as non-response. This has seen a cumulative increase in recent years (Sax et al., 2003; Martin, 2006) and may exhibit social and spatial concentrations (Martin, 2010). The

expense and time involved in collection is also coming under increasing threat with economic constraint (Dugmore et al., 2011).

However, there are additional, more conceptual challenges imposed by the relatively scarce availability of population data compared with the present day. Established geodemographic practices are built on the assumption that residential neighbourhoods are the most relevant environmental exposures to individuals and that effects operate only through interactions among those in the same residential area. Although these measures are widely used, it is increasingly being highlighted that residential location only provides a partial understanding of human identity (Kwan, 2013). For example, geodemographic phenomena can be understood as a complex interaction between population characteristics and environmental attributes (Longley, 2017), where interaction with (or exposure to) contextual or environmental influences also shape their identities (termed 'neighbourhood effects' e.g., Kawachi and Berkman, 2003; Diez Roux and Mair, 2010). Yet, most individuals move around to perform routine activities and come under the influence of various neighbourhood contexts outside of their home (Matthews, 2008, 2011; Kwan 2009, 2012a, 2012b). Therefore, much of the contextual or environmental influences they experience, and the physical and social resources they utilise, might be located far from their area of residence (Matthews, Detwiler, and Burton, 2005).

This assumption that 'night-time' residence is key to understanding the relationship between human identity and the spatial organisation of society is becoming increasingly outdated. As Longley (2017) summarises, residential structure is only one indicator of social structure, that needs to be considered alongside indicators of activity patterns on a daily, weekly, seasonal, or longer-term basis. There have been some recent attempts to incorporate these dynamics, for example, the 2011 Census produced several workplace statistics with new small area geographies ('Workplace Zones'), based on 'daytime' characteristics (derived from social, economic, and environmental variables). Cockings, Martin and Harfoot (2015) subsequently created the 'Classification of Workplace Zones' or COWZ, describing the attributes of WZ populations. CACI (London, UK) also created the proprietary 'Workforce Acorn' (CACI, 2018), which moves worker's home classifications to their places of work. These classifications produced different cluster groups from those based on residential structure. However, although these provide important local economic indicators, and have found useful applications for convenience retail planning (Berry et al., 2016), it is inevitable that human identity will incorporate aspects beyond work and residence. For instance, many people do not undertake paid work, and for many, daily rhythms of activity will be substantially more complex (Longley, 2017).

On the other hand, geodemographic practices are faced with procedural limitations. Firstly, they lead to an implied assumption that the social profile assigned to an area represents the identity of all households. This gives rise to the well-recognised ecological fallacy (confounding the characteristics of areas with particular individuals who live within them), as in reality few areas are socially homogeneous (Dalton and Thatcher, 2015). O'Brien and Cheshire (2014) and Slingsby, Dykes and Wood (2011) visualise examples of these uncertainties across small areas. In addition to this, dynamics may vary with the characteristics of local areas. Singleton and Longley (2015) discuss the implications of performing local area analysis using global characteristics that may not be relevant due to the local considerations. Webber and Longley (2003) also demonstrate that locational context influences the impact of certain variables. The London Output Area Classification (LOAC; Singleton and Longley, 2015) and Lansley et al., (2015) demonstrate an attempt to account for these effects, producing separate classifications for the unique population of London. These revealed different clusters, highlighting the ecological fallacies that neighbourhood classifications can generate.

A final procedural limitation is that as the input units to geodemographic classifications are not naturally occurring (i.e. postcodes), the geographic scale and boundaries between areas can affect analytical results, also known as the modifiable areal unit problem (MAUP; i.e. see Openshaw, 1984). This can leads to two fundamental issues. Firstly, *scale effects*, where major analytical differences can occur depending on the size of units used (for example, correlations will generally be inflated the bigger the units) and secondly, *zonation effects*, where results vary if areal units are alternatively grouped at the same spatial scale (Openshaw and Taylor, 1979; Wong, 1996). These issues mean that caution is needed when conducting spatial analyses on aggregated data (Unwin 1996; Bailey and Gatrell, 1995).

2.2.1.2. Prospects of big data for geodemographics

Many substantive limitations in geodemographic practices have arisen from a lack of available population data. Georeferenced big data holds promise since its ‘velocity’ captures societal interactions far beyond what is possible through decennial census data. For example, these data offer continuous streams of various georeferenced population activities that are up-to-date and often national in scale. This may facilitate investigation of contextual and environmental influences that occur outside of residential neighbourhoods (i.e. see Longley, 2017; Kwan, 2013) and how this manifests in human identity.

Georeferenced big data also offer a means of creating more application-specific classifications, rather than attempting to summarise population dynamics from a limited set of pre-defined variables. This need was highlighted by Openshaw et al., (1995), who believed it doubtful that a

satisfactory general-purpose classification could be devised and Voas and Williamson (2000), who suggested that ad-hoc systems could be utilised to produce classifications with specific uses. Examples of more bespoke classifications are evident from the CIDER Migration Classification (Dennett and Stillwell, 2011), the Higher Education Classification (Singleton, 2008), the Internet User Classification (Riddlesden, 2014) and Health Milieu (Kandt, 2015a), which provide additional dimensions to the attributes of socioeconomic units. Yet, bespoke classifications that move focus away from residential geographies and instead focus on dynamic activity patterns are yet to be fully explored. Some commercial examples of this are evident, for example, instead of classifying neighbourhoods, Facebook and Twitter utilised online interactions (i.e. status updates and Tweets) in conjunction with location (Dalton and Thatcher, 2015). However, these endeavours have been commercially motivated, primarily to optimise advertisement, rather than publically available investigations motivated for public and social good.

Furthermore, big data facilitate a shift in focus from neighbourhoods or postcodes as a unit of measure, to the individual level, where personal locations, dispositions, attitudes, and socioeconomic characteristics are the object of analysis, rather than the homogenised, quantified areal units of geodemographics (Dalton and Thatcher, 2015). This may contribute to our understanding of procedural issues such as ecological fallacy and MAUP. Burns (2014) provides an exploration of producing classifications at an individual level using the UK's 2001 Small Area Microdata, demonstrating the potential for deeper profiling, classification validation and enrichment at this resolution.

The modelling of certain aspects of societal functions evidently requires data beyond the census and its categorical limits (Longley and Harris, 1999), which currently is only a prospect through the integration of novel big data sources. However, there are a number of barriers to the practical application of these data in geodemographic research. Primarily, issues of data provenance (i.e. data error and bias in coverage) must be explored if they are to make a significant contribution.

2.2.2. Enriching Geodemographics: Spatiotemporal Population Dynamics

Whilst geographers have long recognised the importance of time and activity patterns in understanding a wide range of human experiences, research has traditionally considered population dynamics primarily in static spatial terms. The proceeding section explores the concept of time geography, how shifting to a spatiotemporal focus can enrich population insights, limitations of current practices and how consumer data may contribute to understanding these phenomena.

2.2.2.1. Time geography

Research into human behaviour in space and time has been present for over five decades. Hägerstrand (1970) first introduced the notion of 'time geography', suggesting that time should be treated with equal importance to spatial factors in social analysis (Rainham et al., 2010). To broadly summarise this area of literature, time geography advocates that humans have goals - things we want to achieve, and projects - a series of tasks to complete in order to achieve those goals (Neutens et al., 2011). These projects take time to complete, and must be completed somewhere, therefore both space and time are relevant in the study of human activities. Projects and goals are also constrained by a number of factors, termed capability, coupling and authority constraints:

1) *Capability constraints* – physiological constraints, such as the need to sleep or eat, in addition to the ability to command tools such as transport (Thrift, 1977).
2) *Coupling constraints* – the fact that at certain times, people, activities and resources have to come together for given amounts of time. For example, going to work or meeting friends.
3) *Authority constraints* – refer to societal and institutional rules and norms. For example, people are not able to be in certain places at certain times, such as in a shop outside of its opening hours (Neutens et al., 2011).

These ideas were built on a range of early studies into human time-budgets and activity patterns (see Anderson, 1971) that have understandably evolved over time in line with technological advances and resulting societal changes. For example, the dissemination of the Internet/mobile devices means that people no longer have to physically be in a location at a certain time to complete goals (i.e. Internet shopping or banking) and consumption activities are no longer restricted by the opening hours of stores (Farag et al., 2007) - although there may be new constraints such as delivery slots and locations. Despite this, the underlying concept still remains that activities carried out by members of the population will be shaped by both individual and collective spatiotemporal constraints.

Motivated primarily by the need to model travel demand and investment, the main applications of time geography have been in the transport domain (e.g. Timmermans, Arentze and Joh, 2002; Chen, 2016). Yet, another area where time has provided insights not identifiable from static conceptualisations is accessibility (Kwan, 2013), referring to people's access to services and locations (such as shops or transport). Traditionally, this has been conceptualised in terms of locational proximity to amenities (for example, distances or travel costs between residential locations and facilities). However, these measures failed to account for constraints of everyday

life, such as people's needs to be at certain locations at certain times of day, the time taken to reach these locations and opening hours that may deem a location inaccessible (Kwan and Weber, 2008; Schwanen, 2007; Neutens, Versichele and Schwanen, 2010; Delafontaine et al., 2011). For example, a store is not necessarily accessible even if it is located right next to a person's residence if the person's space–time constraints (e.g. work schedule) make it difficult to visit during opening hours. These concepts are of enduring relevance to the study of population activity patterns and the organisation of societal flows, and have had applications across the domains of health (Widener et al., 2013), transport (Neutens et al., 2012), social interaction (Farber et al., 2013) and environmental exposure (Kwan, 2013).

2.2.2.2. Relationships between time, place, activities and identity

When examining spatiotemporal factors and constraints on daily life, much research has indicated the existence of 'temporal rhythms' in human behaviour. Lefebrve (2004) coined this concept 'rhythmanalysis' - the study of spatiotemporal rhythms at the individual, institutional, urban, regional, national, and even global scales. This theory asserts that everyday life is shaped by a multitude of habits, schedules and routines that are cyclic and predictable, due to the functioning of society requiring synchronisation of practices in order to achieve goals (Edensor, 2016).

It further postulates that differing routines of people in space are interlinked with the identities of those individuals and the subsequent formation and functions of places. For example, unique activity patterns can be derived from the daily flows of commuters, the multipurpose trips of working parents, the lifestyles of students or the slow pace of unemployment (Edensor, 2016) that are shaped by urban rhythms such as the schedules of public transport, the openings and closing of shops/workplaces, the flows of postal deliveries or even the rhythms of lunch/coffee breaks (Labelle, 2008), in addition to seasonal and annual cycles. Places can therefore be conceptualised as points of spatial and temporal intersection (Gren, 2001) for daily tasks, pleasures and rhythmic routines. This view argues that the daily, weekly, seasonal and annual rhythms of a place influence its on-going formation, and we can therefore identify the distinctive characteristics of a place according to its 'polyrhythmic ensemble' (Crang, 2001). Thus, places are "social constructs defined by the cumulative effects of highly distinctive interactions between population characteristics and environmental attributes over space and time" (Longley, 2017, p 10).

There is much evidence for the existence of temporal rhythms in human behaviour and recent years have seen an increased interest in their applications to understanding socioeconomic dynamics. For example, research has demonstrated a substantial amount of repetition and

predictability in individual activities (i.e. Simma and Axhausen, 2001; Buliun, Roorda and Remmel, 2008; Roorda and Ruiz, 2008; Song et al., 2010; Gonzalez et al., 2008) and also variations in temporal rhythms between social categories due to the differing space-time constraints imposed on them (Lefebrve, 2004; Farber et al., 2013). Examples include: the contrasting rhythms of different age groups (Lager, Van Hoven and Huigen, 2016), gender rhythms in place visiting (Schwanen, Kwan and Ren, 2008; Hanson, 2010; Schwanen, Banister and Anable, 2012), relationships with education, employment or family structure (Alhadeff-Jones, 2016; Fagan, 2001; Crouter and McHale, 1993) and travel methods, such as the rhythms of pedestrians and bicyclists' compared to cars (Hornsey, 2010). Lefebrve (2004) suggested that for the working population, aspects of every day (i.e. sleeping, eating, leisure and time at home) are subordinated around the need to work. Despite these suggestions, individual activities will obviously be largely variable in reality (Neutens et al., 2010). In literature, intra-person variability estimates have ranged between 20% and 80% depending on the metrics used and the days of analysis (Kang and Scott, 2010; Susilo and Axhausen, 2014), although, small qualitative studies have dominated these insights. In addition, various contextual influences can alter routines. Alheit (1994) highlighted how changes can be incurred by life events and also mark new phases of life - such as divorce, the birth of a child, illness or unemployment.

Research in this area indicates that aspects of human identity and place formation may manifest and be quantifiable from the study of population activity patterns. For instance, focusing on everyday activities may reveal how the rhythms of both places and people are ordered, and how these orderings may vary by social group (Lager et al., 2016). Recent work on human mobility and geographies of encounter (e.g., Sheller and Urry, 2006; Valentine, 2008; Adey, 2010) highlight the implications of these concepts for residential based geodemographics. For instance, if people's differing daily habitual obligations and behaviours shape their accessibility to various locations, people with specific personal and household attributes will have different temporal routines, accessibility options and thus environmental exposure influences (Neutens et al., 2010; Delafontaine et al., 2011). This research therefore highlights the need to advance our theoretical and empirical understanding of population rhythms over various scales and dimensions of the life course.

2.2.2.3. From data-scarce to data-rich activity patterns

Despite prolific evidence that incorporating elements of time and mobility could greatly enrich our understanding of societal phenomena, many notions in geography and social science research still continue, for the most part, to be conceptualised in static spatial terms. This has primarily been a result of space–time behavior studies being limited by data availability. For

example, in the data-scarce era, studies involving activity patterns and travel behaviour have largely used cross-sectional data collected through active solicitation (i.e. subjects and information on their travels are actively collected). This has included travel surveys where subjects are asked to self-report their activities and travels via paper, web, or phone interviews, and in some cases, Global Positioning System (GPS) loggers (e.g. Chen et al., 2010; Gong et al., 2011). The nature of these studies has meant that data collection is limited to small sample sizes and over limited time periods of say, a few days or months (Chen et al., 2016).

Big data is rapidly improving our ability to collect spatiotemporal activity data and there are now various sources from which we can attempt to infer the dynamics of time geography. Goodchild (2013) describes five domains of space-time in GIS, including tracking of continuous space-time paths, created from high-resolution data (i.e. separated by a few seconds, such as from GPS tracks) or lower resolution data where intervals are determined by the behaviour of the object, such as GPS tracking of when whales surface, or when a customer transacts. The latter, of particular relevance to this work, describes aspects of the *events and transactions* domain and the *change or snapshots* domain. The snapshots domain is concerned with capturing changes over time via a series of snapshots and is thus a sequence of cross-sectional continuous fields. Alternatively, events and transactions represent a single or series of events and locations in space and time. These data can still be useful for measurement at daily, weekly or seasonal scales.

Despite these emerging forms, access barriers have meant that studies endeavouring to capture population activities have so far only utilised more accessible datasets such as from social media (e.g. Mennis and Mason, 2011; Leak, 2017) or mobile phone data (e.g. González, Hidalgo and Barabasi, 2008). Much work remains to be done in terms of understanding the full spectrum of applications of various big data sources, such as consumer data. In addition, moving beyond the conventional focus of static residential spaces and toward temporally integrated perspectives poses challenges. For instance, there are currently few widely recognised methods for analysing complex relationships among human space–time trajectories, particularly in terms of reliable linkage to other relevant attributes such as socioeconomic context (Kwan, 2013). This links back to Longley's (2017) observation that there is a need to triangulate big data sources with traditional administrative datasets in an attempt to quantify their socioeconomic value and make sense of trends. Lansley and Adnan (2015) demonstrate an example of this, by identifying geotemporal demographics of Twitter flows in conjunction with the Census based OAC and Kamenjuk, Aasa and Sellin (2017) demonstrate an example of contextualising mobile phone data trends with Census migration statistics.

2.2.3. Consumer Data as Indicators

A current and growing area of research is the dissemination of consumer data as population indicators. As previously noted, these data capture consumption patterns continuously, represent entire consumer populations, convey information about actual purchasing behaviours, are high in temporal granularity, longitudinal in nature, and in some cases include georeferences of both customer residences and the locations where purchases occur. This has made them a particularly attractive source for generating societal insights.

Many analysis techniques have been applied within industry to extract insight about people's lives, activities and identities (see Hiziroglu, 2013, for a review of approaches) from consumer data. However, as these data are predominantly produced for insights about specific consumer populations, commercial aims have been focused on producing indicators that will benefit profits, rather than trying to better understand society by generalising trends to the wider population. Yet, for social science these data offer a framework for creating indicators relevant to wider societal phenomena. A well-known example comes from Tesco's Clubcard (Humby, Hunt and Phillips, 2004), where a correlation between increased consumption of nappies and beer could be attributed to the behaviour of new fathers, who showed an increase in drinking in the home rather than socialising. These kinds of trends demonstrate how our perception of distinct geodemographic groups can likely be enriched through the analysis of consumption patterns.

In addition to this, the spatiotemporal characteristics of consumer data offer a framework for exploring the concepts of time geography and spatiotemporal rhythms in a data-driven context, from a much larger population sample than has previously been obtainable, over granular temporal intervals and much more longitudinal periods. They further offer a means of creating bespoke indicators based on daytime consumption patterns, and thus creating representations that are not based solely on residential geographies. This would also be of relevance to commercial organisations, who similarly continue to utilise classifications based on residential based activity (Shearer et al., 2015). However, research has not yet been able to explore the implications of consumer data in this context. Only Sanford (2008) has attempted to use observed consumption patterns (as opposed to lifestyle surveys) in neighbourhood based classifications, demonstrating that they can be utilised as social indicators of community change and identity, such as race, education and income.

In addition to enriching knowledge of individual activities, these data have equally important implications for understanding the formation and functions of places. A prominent example comes from the challenges currently facing UK town centres and high streets, characterised by a

decline in economic health and vitality in recent years (Wrigley and Lambiri, 2014). It is commonly recognised that these places do not only represent shopping destinations, but are also valued as economic, social and community spaces by both public and commercial bodies (Carmona, 2015). Yet, notwithstanding the 2007 economic crisis (which brought many of these issues to the fore), changing consumer trends have threatened the resilience of these centres. This has included; the rise of the convenience shopping, where people prefer to make 'one stop', local, and often out-of-centre trips based on time constraints (Wrigley and Lambiri, 2014) versus destination trips or 'comparison shopping' (where consumers plan a trip to a retail centre in order to fulfil retail or leisure needs; see Guy, 1998); shifts in demographic composition such as age, ethnicity and household structure; the effects of accessibility and available transport; the rise of multi-channel retailing, such as online shopping; the impact of 'out-of-town' retail centres; and how factors vary with different regional economies, amongst many others (see Wrigley and Brookes, 2014, for an in depth review).

Whilst much work has endeavoured to understand these changes, solutions have been hindered by a lack of available data that captures complex consumer interactions with urban places. Wrigley and Lambiri (2014) highlight how exploitation of locally available, longitudinal, national scale time series data are needed in order to produce comparative, quantitative measures of high street performance and inform locally relevant decision making. The study of spatiotemporal rhythms in retail spaces, as facilitated by consumer data, could therefore be of enduring relevance to many of these issues. For example, understanding the functions that these spaces serve to consumers, their demographic compositions and the quantification of both short and long term impacts. This demonstrates an example of how consumer data may be applied to problems that concern the public, government and retailers alike.

Yet, for their potential in tackling issues of broader societal concern to be realised, the quality and provenance of consumer datasets needs to be fully understood. As outlined in Section 2.3.1, these types of data are not created for the edification of researchers and analysts, which generates a myriad of challenges. Data collaboratives such as the CDRC are beginning to acquire consumer data across multiple sectors (such as from retail, transport, energy consumption and footfall from Wi-Fi sensors - see Longley, Cheshire and Singleton, 2018), which endeavour to shed light of issues of uncertainty and investigate their use as alternative population indicators. However, such on-going efforts to repurpose these data represent the very beginnings of attempts to understand what different kinds of consumer data can provide in terms of population insight, and there remains a gap in research in realising their full potential.

2.2.4. Summary

Geodemographics help us summarise complex societal phenomena. However, data with finer temporal granularity that represent activity patterns and dynamics beyond residential location are lacking. These activity patterns have been implicated in helping us to better understand the socio-spatial organisation of society, for example, through the rhythms of distinct social groups and their relations to human identity and place formation. It is likely that the integration of consumer data will be a particularly important progression in this area of research, however, there is a need to understand the dynamics of these novel data sources before repurposing them.

2.3. The Provenance of Loyalty Card Data

Loyalty card schemes have been prominent since the early 1990s, when retailers began to recognise the cost benefits of retaining rather than obtaining customers (Kotler, 2002). They have since become extremely popular, with almost all major retail chains operating some form of scheme. In their most basic form, these schemes involve awarding points according to how much a customer spends, which can then be redeemed as discounts on future purchases. Data are collected firstly, through the process of an application (i.e. either online, or in a store) that typically asks for demographic (age, gender) and address information. Secondly, customers are provided with a membership card that records their purchasing habits at a point of sale. This essentially creates a system of marketing incentives that encourage customer loyalty by offering rewards for repeat patronage, as well as providing data that can be used to gain behavioural insights and encourage further spending. These schemes, facilitated by technological innovation, have placed retailers at the forefront of the big data revolution, since they now retain and interpret an immense body of data about their customers and their consumption patterns. Recent estimates suggest that approximately three-quarters (76%) of consumers carry between one and five cards with them at all times (YouGov, 2013) and collectively, almost 46.5 million people, or 92% of the UK adult population, are currently registered with at least one programme (Loyalife, 2017).

For retailers, loyalty schemes are primarily used for Customer Relationship Management (CRM), which aids understanding of customers on an individual level (Anderson and Kerr, 2001, provide a detailed overview of CRM). The ability to obtain a rich understanding of customers is possible due to the combination of both transactional data and customer metadata. For instance, variations in transactional behaviours are typically quantified using segmentation (see Section 2.2.1), which can then be augmented with geodemographic classifications through the provision of customer postcodes. Postcodes are also typically utilised for marketing strategies such as mail-based rewards or location-based targeting and also for GIS applications

such as location planning and catchment area mapping. Dorotic, Bijmolt and Verhoef (2012) provide a comprehensive literature review of loyalty programmes and Humby et al., (2004) provide an overview of their uses (for the Tesco Clubcard).

2.3.1. The Concept of Loyalty

Loyalty has received many definitions in the literature, yet is most commonly conceptualised as either behavioural loyalty – which focuses on the extent of repeat purchase patterns (Bridson, Evans and Hickman, 2008) – or attitudinal loyalty, which is concerned with preference or commitment to a particular store or brand (Meyer-Waarden, 2007). Whilst there has been a wealth of loyalty related literature in the academic community, the majority of this has focused on the concept of loyalty, influences on adoption rates (Demoulin and Zidda, 2009), scheme profitability and redemption dynamics (Smith and Sparks, 2009). A large proportion of this research has focused on whether or not these schemes are actually effective in maintaining loyal relationships, of which the findings are conflicting (e.g. Dowling and Uncles, 1997; Mauri, 2003; Gómez, Arranz and Cillán, 2012).

Behavioral research has focused primarily on the notion of segmentation, typically demonstrating that loyalty behaviour is extremely variable between individuals. Darden and Ashton (1974) provide one of the earliest examples, proposing seven segments ranging from the 'quality shopper' to the 'convenient location' shopper. Other segmentations have focused on measures such as repeat purchase patterns (Bridson et al., 2008) or shopping frequency over time (Demoulin and Zidda, 2009) in order to understand who are the most profitable customers. Most recently, Atkins, Kumar and Kin (2016) defined different types of shopper based on the lengths they may be willing to go to get the best deal. In short, there is great variability in consumers' strategies, preferences and behaviours.

Despite these research efforts, a fundamental drawback of loyalty research to date is that all have utilised qualitative methods (and consequently, small samples) rather than actual loyalty card data and these have also been mainly cross-sectional, rather than longitudinal. This, understandably, is partly due to the data's origins in privately owned businesses and their secure storage requirements since they provide information about consumer transactions, residential locations, movements and interactions. Yet, whilst qualitative studies have been useful for understanding the concept, behavioural and attitudinal dynamics of loyalty, there is a substantial lack of evidence from a data-driven perspective in academia.

2.3.2. Loyalty Cards as Social and Spatial Data

The wealth of information generated by loyalty cards offers huge potential for endeavours in data-driven science. In particular, the provision of customer postcodes and store locations provides a valuable geographic reference that can be regarded as the key to utilising these data for a broad range of social and spatial applications. The spatiotemporal analyses facilitated by these data are concerned with both the *events and transactions domains*, and the *changes and snapshots* domain. These data contain transactions (events) that are referenced in both space and time and linked to individual accounts, therefore creating space-time trajectories at an individual level. However, these trajectories are primitive in nature and limited in that they are reliant on an individual performing a transaction. It is reasonable to assume that the majority of loyalty scheme members will not transact more than once a day, or, even once a week, for example. Thus, they may not be particularly useful for analysis of detailed individual trajectories that require frequent intervals between data points (such as is possible from GPS trackers).

Nevertheless, as these data are particularly longitudinal in nature (i.e. data are collected over years), accumulation of events can provide insight into the general spatiotemporal trends of individuals, such as over daily, weekly or seasonal periods. This also means the snapshots domain is particularly relevant, as changes can be quantified between time periods. Combined, these data allow us to capture both short and long term dynamics of consumption patterns. They also offer a data-driven context in which to facilitate a more sophisticated view of spatiotemporal phenomenon, providing voluminous consumer data that are not compromised by uneven response rates, can be updated on a regular basis and permit consistent comparison between different behavioural datasets on a relatively granular scale (over 1.4 million postcode units across the UK). The spatial element in these data also means they can be appended to existing national statistics to infer relationships with existing understanding of population characteristics and neighbourhood types (Webber, Butler and Phillips, 2015).

Due to the additional georeferenced element of store locations, the data produced by loyalty cards allows us to investigate a broad number of variables relating to mobility, such as distances travelled, the size of store networks and the locations that individuals visit over time. By incorporating the temporal element of these movements, we can further utilise these data to understand more complex socio-spatial characteristics at both individual and aggregate levels. This evolving research may enable us to build bespoke classifications pertaining to specific phenomena, understand relationships between consumption characteristics with existing geodemographic representations, and summarise daily activity patterns in both time and space. These types of analyses could also greatly enrich our understanding of the formation and

function of places, which given the nature of these data, could have particularly important implications for issues of high street resilience.

2.3.3. Data Issues

Whilst these data offer substantial potential for informing population insights, they are also subject to the unique challenges imposed by big data (see Section 2.1.3). Whilst these data are adequate from a retailer's perspective, as variables are created and data interpreted with the primary focus of understanding and maximising the buying behaviours of their customer base, it would be impractical to assume that these data will meet the 'gold standards' of traditional population data in terms of both their quality and representativeness.

There are a number of data veracity considerations in the context of loyalty card data. In the first instance, preliminary explorations are necessary to identify basic forms of data error (i.e. incompleteness of records, identifiable spelling errors or syntax mistakes). However, more subtle uncertainties can also arise from the nature of their collection. For example, the accuracy of customer metadata is entirely dependent on human input when signing up to a scheme. Therefore, information pertaining to age, gender and address may be susceptible to entry errors, of which will only be identifiable if values are not admissible. Temporal errors may also be evident, for example, address attributes are dependent on customers updating this information if a residence changes. Issues of representation and bias are also inherent due to the effects of self-selection, where customers select themselves to participate and therefore represent a biased sample. Research into the representativeness of loyalty populations to date has been extremely limited in terms of longitudinal or data-driven studies, nevertheless, three important areas of consideration can be identified; the representation of these data in relation to the general population, the retailer population and the loyalty population.

General population. Retailers attract and target certain demographic groups and will only represent those who fall within the scope of the markets or activities that are being tracked. Numerous studies have indicated the presence of demographic biases in loyalty card data (i.e. Leenheer et al., 2007; Smith et al., 2003; Van Heerde and Bijmolt, 2005), such as an over-representation of middle-aged individuals (youths and over 65's may be the least likely to participate - Wright and Sparks, 1999) and a higher participation of larger or higher income households (Bell and Latin, 1998). This bias also varies with the characteristics of specific retailers, such as pharmacies being more representative of female populations, and petrol loyalty schemes more male dominated (Maritz Research, 2006). For high street retailers, bias in the distribution of customers may also be dictated by the physical locations of stores. For example,

Allaway, Berkowitz and D'Souza (2003) found that inconvenience of store locations plays a large role in determining card ownership.

Retailer population. Loyalty card populations represent a sample of a retailer's full customer base. Thus, it is similarly important to understand if these data are representative of all purchases taking place at a given outlet, or if behaviours can only be attributed to the loyalty card holding segment of customers. Numerous studies have indicated the existence of member and non-member differences in terms of purchasing behaviour (e.g. Heerde and Bijmolt, 2005), which can be attributed, in part, to individual/psychological dispositions such as receptiveness to offers.

Loyalty population. Individual differences in the behaviours of loyalty card holders can result in a disproportionate representation of customers across the database. For instance, data may not include all of the purchases made by card-holding accounts, due to not having the card or feeling that it is not worth using for small purchases (Wright and Sparks, 1999). In addition, many rarely or never use a card after signing up (Cortinas, Elorz, and Mugica, 2008) and will not be represented at all. The amount of data per individual may also be influenced by membership to competing loyalty programmes as well and those who shop only for price leaders and best deals (Allaway et al., 2006). These dynamics are quantified to some extent in commercial settings using segmentation, however, there is very little understanding of the effects of card usage on data quality in research. One exception is Allaway et al., (2006), who utilised approximately 1 million transactional records to demonstrate that only between 1% and 9% may actually exhibit consistently loyal behaviour (although definitions of what constitutes 'loyal' vary). There is also an obvious limited completeness when utilising the records of a single loyalty scheme.

Alternative considerations of representativeness arose from discussions with the HSR rather than insights that are evident in literature. Firstly, there are variations in card usage across different store locations. For example, lower levels of participation are observed in more transient locations, such as 'convenience' stores (i.e. smaller stores located in urban areas) in comparison to 'destination' based stores (i.e. city centre flagships), of which the latter typically sees higher basket sizes and thus higher participation due to the perceived benefits of points. This dynamic effects the distribution of behavioural data across different area types. Secondly, there is an over representation of product consumption within certain categories, due to cards being used more with high value items. These issues have further implications for the completeness of individual transactional histories, which may be influenced by these differing motivations to participate.

2.3.4. Summary and Research Potential

Loyalty card data offer an untapped opportunity for researchers to analyse societal and geographical questions in an entirely new way. They represent large numbers of people and allow analyses at a variety of spatiotemporal scales. However, there are a number of preliminary considerations and pragmatic steps required to ensure these data are fit for purpose in a research context. These cautions mirror those adopted in traditional methods of data handling in regards to data quality and sampling bias, however, efficient methods of revealing these inherent data issues requires exploration of which has been subject to relatively little appraisal from the academic community. This oversight is, in part, a symptom of disaggregate loyalty card data being hard to access outside commercial settings.

An important research direction is therefore to develop methods of handling and analysing these data. Traditional statistical methods have been focused on data-scarce science, where aims are to identify significant relationships from small, controlled sample sizes. Developments in big data research may involve applying data-driven approaches to quantify uncertainty within these data. Beyond this, there is a pressing need to develop a robust understanding of their applications to advancing our knowledge of population dynamics in respect to consumption behaviours, daytime activities, mobility patterns, spatiotemporal dynamics and the relationship of these patterns to geodemographic representations. For instance, how spatiotemporal routines, obligations and subsequent accessibility and environmental exposure dynamics may vary between distinct social groups. This would ultimately provide an enhanced description of what makes certain groups of people distinctive. It is critically important that analyses of this nature endeavour to achieve outputs that are both informative and safe, especially where data linkage is concerned. Nevertheless, the prospects of loyalty card data as a social and spatial data source presents promising applications for social science research.

3. Data and Preliminary Analyses

3.1. Introduction

This chapter provides an overview of the attributes and characteristics of the loyalty card data utilised throughout this thesis. Such data are rarely available outside of commercial settings, therefore this presented a unique opportunity to understand their inherent characteristics. The dataset was provided by one of the most prominent high street retailers in the UK, who have a national network of stores. Their loyalty card scheme exhibits one of the highest UK membership rates and comprises of (at the time of collection) over 18 million customers. The spatial extent, granularity and volume of these data is unparalleled in comparison to previous research regarding both the dynamics of loyalty card data and understanding the applications of commercially generated big data in a research context.

Access to these data was possible through the ESRC funded Consumer Data Research Centre: a government funded big data initiative that aims to facilitate the access of commercially generated consumer datasets to academic researchers. In order to secure the data, a number of strict procedures were necessary to minimise the risk of disclosing commercially or personally sensitive information about the retailer and its customers. These data are personal in nature (i.e. relate to identifiable living individuals – see Chapter 2, Section 2.1.3.1), describing residential locations, demographic characteristics and transactional behaviours at an individual customer level. These are classified as 'controlled data' under CDRC regulations – meaning data that need to be held under the most secure conditions with stringent access restrictions. This thesis represents one of the first investigations of a consumer dataset in this context and issues of access, data handling and presentation of results were important associated challenges. An overview of the processes required to conduct analyses on these data is provided in Figure 3.1.

Access was granted to these data via the CDRC's secure service - the Jill Dando Institute Research Laboratory (JDIRL) secure facility, UCL. In the first instance, this requires preliminary vetting and training procedures that ensure access is only granted to trusted researchers (see the CDRC User Guide, 2018). Following this, researchers must receive approval for proposed uses of the data and all analyses must be performed within the secure laboratory setting. To output data from the laboratory, the data must firstly conform to a number of statistical disclosure controls. This includes: aggregation to large geographical areas, suppression of disclosive cells, ensuring percentages do not allow deduction of disclosive units,

and where counts are concerned, a threshold rule of no less than 10 (see Appendix 1, for the full documentation).

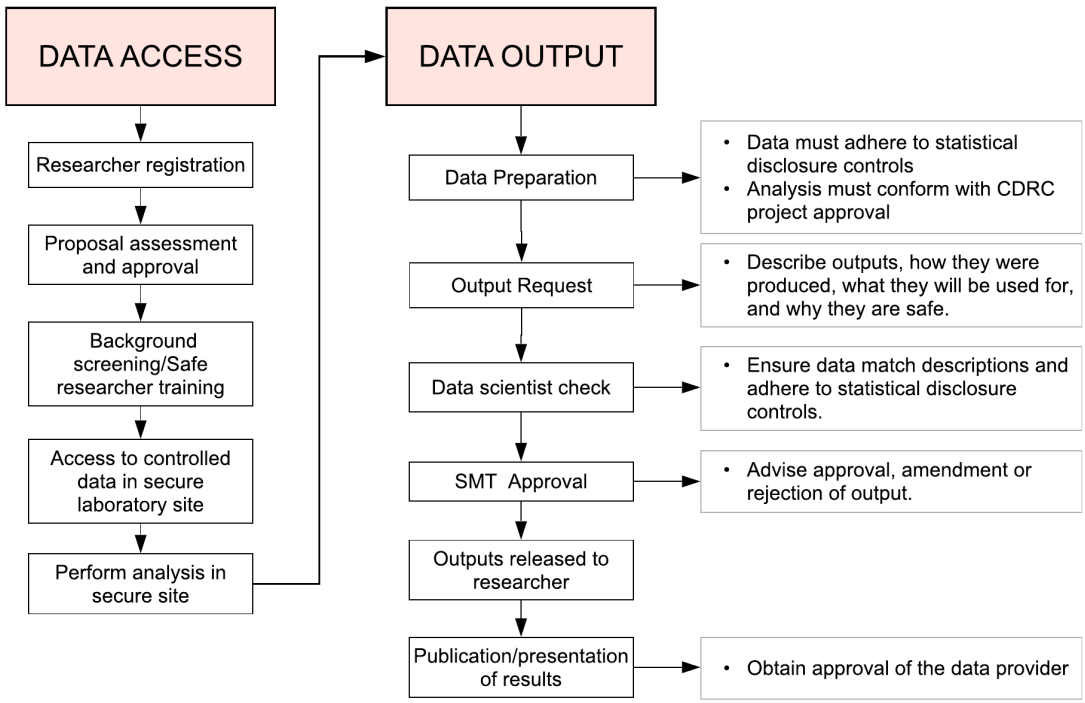


Figure 3.1: CDRC 'controlled data' procedures required to access, analyse, output and present HSR data.

These controls follow government specified rules and regulations on the handling of disclosive data (see the Government Statistical Service [GSS] guidelines, 2014). The second stage of data output then involves the assignment of two CDRC Data Scientists to carry out checks that ensure they match output request descriptions and adhere to statistical disclosure controls. Finally, two members of the CDRC Senior Management Team (SMT) review and advise the approval, amendment or rejection of these outputs. Once obtained, the presentation and publication of analyses must also be approved by the data provider for commercial disclosure purposes.

As a result of these procedures, the presentation of these data have been necessarily constrained in order to adhere to both statistical and commercial disclosure controls. This required data treatment measures, such as spatial aggregations, of which are described throughout the proceeding sections.

3.1.1. The High Street Retailer (HSR) Loyalty Scheme

The HSR loyalty scheme operates by awarding points to customers when they purchase products. The scheme exists in the form of issuing a physical loyalty card (i.e. like a bank card)

to a customer when they sign up. The process of signing up (either in store, or online) requires the customer to provide their date of birth, gender and postcode. To participate, customers are required to be over 12 years old. The points earned by customers are acquired by swiping the card when they transact in store, or purchasing online, of which differing products offer different rewards. Generally, the higher the value of the item, the more points are awarded. However, certain products can provide increased points during different time periods according to marketing strategies. The points earned by customers are accumulated and returned in monetary value, of which can be used to purchase further products with the HSR.

3.2. Data Overview

The data provided by the HSR included loyalty card transactions and non-card transactions between April 2012 and March 2014. These data represented every transaction recorded within a UK store within a 2.5 year period. In addition, metadata were provided for product categories (over 6 hierarchical levels), customer accounts (describing their gender, date of birth and postcode) and stores (containing locational/retail structure descriptions). These data represent a structured form of big data. For example, clear variables were provided with definitions of each and matching IDs between tables. Table 3.1 illustrates the structure of these tables and the variables utilised.

Table 3.1: Structure of the HSR data tables.

Transactions		**Metadata**		
Non-Card	***Card***	***Customers***	***Products***	***Stores***
Store ID	Account ID	Account ID	Level 1 ID	Store ID
Level 6 ID	Store ID	Gender	Level 1 name	Name
Value (£)	Level 6 ID	Date of Birth	Level 2 ID	Coordinates
Timestamp	Value (£)	Postcode	Level 2 name	Postcode
	Timestamp		Level 3 ID	Type
			Level 3 name	Format
			Level 4 ID	Opening date
			Level 4 name	Closing date
			Level 5 ID	
			Level 5 name	
			Level 6 ID	
			Level 6 name	

Despite the structured nature of these data, a number of disparities between datasets were identifiable, in addition to data quality issues such as incompleteness of records and uncertain attributes. This required preliminary data treatment procedures of which are described for each data table in the proceeding sections. In addition, the majority of customer data (95.7%) were generated within Great Britain (GB), with only 4.3% of customers and 0.1% of transactions from Northern Ireland (NI). Many areas within NI consisted of too few customers to enable safe presentation of results, therefore, analyses throughout this thesis focused on GB.

3.2.1. Transactional Data

Transaction variables included a customer account number, store number, product category, value (in GBP) and a timestamp given to the minute (in the format DD:MM:YYYY, HH:MM). Transactional data were provided, in their raw form, at a product category level. Therefore, for each customer transaction, multiple records existed (i.e. one for each product bought), resulting in 1,324,593,222 records. Data in this format could be utilised in analyses concerning product consumption. However, for many of the analyses conducted in this thesis, transaction level data was necessary. Within the card data, the timestamp for a single transaction was identical for each record. Therefore, transactions could be obtained by aggregating records by account number and timestamp (product data and value were aggregated to a count of products bought and sum of value for that transaction). This resulted in a total of 507,782,128 transactions. However, for the non-card data, transaction level data could not be obtained due to the absence of account numbers by which to link individual records. This eliminated any potential analysis regarding comparisons of card and non-card data at the transaction level. For this reason, in addition to the primary focus being on understanding the dynamics of loyalty card data, non-card data were not utilised.

The card transaction data covered a time period of the 1st of April 2012 to the 30th of September 2014. This included two full financial years (as defined by the HSR). Financial year 1 covered the 1st of April, 2012 to the 31st of March, 2013. Financial year 2 covered the 1st April, 2013 to the 31st of March, 2014. Therefore, there was an additional 6 months of data available proceeding financial year 2, providing a total time period of 912 days. Table 3.2 illustrates the volume of data available during each time period.

Table 3.2: Loyalty card records during each financial year, GB.

	Financial year 1 *365 days*	**Financial year 2** *365 days*	**Financial year 3** *182 days*
Product level	463,383,609	698,570,990	162,739,623
Transaction level	208,198,478	205,145,117	94,438,533

Information on payment methods, such as cash/card or whether loyalty points were used to make these purchases was not provided by the HSR.

3.2.2. Metadata Attributes

3.2.2.1. Products

Product metadata variables included unique codes and names for products over 6 hierarchical levels. Level 1 represented the most aggregate level, describing 2 categories (retail, or pharmacy). Level 6 represented the lowest level of this hierarchy, describing 329 categories. These still represented aggregate product groups (the HSR did not provide individual product

information). The references provided in the transactional data were at level 6, allowing linkage to this metadata. Table 3.3 provides a summary of each level of the product hierarchy, with examples of categorisations at each level.

Table 3.3: Example of the HSR product hierarchy structure.

Level	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 Mouthcare Children's wear Tissues
6	329	Deodorants Electrical hair Shampoo & conditioner Premium cosmetics

This information was used in analyses throughout this thesis to interpret consumption behaviours. However, some categories (at level 6) were not present in the transactional data and therefore were not utilised. In addition, some category meanings were not interpretable (i.e. 'Other' or 'Miscellaneous') and thus were also excluded when conducting product consumption patterns. In order to maintain disclosure of both customers and the HSR, outputs from these investigations are presented at more aggregate levels and with custom product category names.

3.2.2.2. Customers

Customer metadata variables included a unique account number, date of birth (DD/MM/YYYY), gender (M, F or U for undisclosed) and postcode. Postcodes were provided in 'postcode units' – the smallest geographical unit available, of which there are approximately 1.7 million across GB. These cover an average of 15 properties (although this ranges between 1 and 100). In total, there were 17,556,936 loyalty card account holders present in GB. However, not all accounts contained complete metadata records. For example, a number of customers either withheld or provided incorrect information (i.e. substituting '----' for a postcode). Table 3.4 provides a summary of data available for customer metadata attributes. Of total GB accounts, 16,797,398, or 95.6% of customers, provided full metadata records. Overall, there were 1,384,193 GB postcodes containing at least one customer.

Table 3.4: Customer metadata attribute completeness.

	Date of Birth	Gender	Postcode
Provided	96.3%	99.7%	99.7%
Withheld	3.7%	0.3%	0.3%

An initial data treatment process was to convert the date of birth field to age, performed in PostgreSQL, which required a timestamp from which to calculate. Two variables were created for this purpose. Firstly, for general analysis, ages were calculated from the start date of transactions (1st April 2012) in order to compare all customer ages equally and relative to the time period of the data. Secondly, for analyses involving census data, ages were calculated from the day of the Census (27th March 2011). This process revealed an abnormal range, with a minimum age of 2 and a maximum of 359. Approximately 8000 customers exhibited ages of over 100, for whom the majority demonstrated regular behavioural patterns (such as high frequency of transactions and spend). It was speculated that this could have been due to error in data processing or human input error. Due to these uncertainties, only customers between 16 and 85 years were considered in analyses that concerned age data. This was specified to remove the identifiable uncertainties and focus on the majority of the adult population. Still, errors falling inside of the normal human age range would not be identifiable in these data.

3.2.2.3. Store data

Store metadata included a unique store number, store name (i.e. the location name), store type (as defined by a HSR classification), store format (describing its primary retail function), opening date, closing date (if applicable), postcode, and location coordinates (latitude, longitude and eastings, northings). Throughout this thesis, the presentation of specific store locations is restricted in order to protect HSR anonymity. Where store locations are visualised, aggregations were performed from point locations to grid cell counts (5km or 1km depending on scale). In GB, there were a total of 2433 stores at the time of data collection. Store 'type' described a classification of these locations derived by the HSR across 9 groups. This was generated from a cluster analysis of locational, demographic and retail composition variables. Store 'format' described the primary retail focus of these stores across 4 groups. Table 3.5 provides an overview of the store types, formats, and the number of GB stores belonging to each class.

Stores were broadly separated into 'Chemists', 'Destinations', 'Convenience' and 'Community' types. High street chemists accounted for the largest proportion of HSR stores, followed by small high street destinations. Chemist 'health centres' represented a distinct type, which provided a pharmaceutical (i.e. prescription) service and were primarily located within GP surgeries. These did not represent a comparable retail dynamic to the other high street oriented store types and thus were not included for many analyses. 'Destination (EOT)' described 'Edge

of Town' locations, which were larger stores residing in retail parks and out-of-town shopping centres. 'Community' stores were typically smaller, pharmacy oriented, and served local communities. These were in a mix of location types, although predominantly small rural towns. The 'Convenience' types described smaller stores located in urban areas, and in the case of 'Convenience (Travel)', those located within or around transport hubs.

Stores of a 'Pharmacy' format were primarily healthcare and pharmaceuticals oriented (with limited offering of alternative products). 'Health and beauty' offered a wider range of products and were more cosmetics and beauty focused, but also offered pharmacy services (i.e. often comprised of a pharmacy in store). 'Flagships' were the key major urban stores (located in either shopping centres or city centres) and 'Airport' described those in airport locations.

Table 3.5: Overview of HSR store types and formats in GB.

Type	Count	Format	Count
Destination (Small high street)	542	Pharmacy	1254
Destination (Large high street)	238	Health and beauty	1078
Destination (EOT)	186	Flagship	63
Convenience (High street)	80	Airport	37
Convenience (Travel)	57	*Other*	1
Chemist (High Street)	742		
Chemist (Health Centre)	238		
Community	349		
Other	1		

Stores with less than a year of transactional data due to opening or closing within the time period were excluded from analyses. This eliminated 26 stores. Daily opening times for each store varied between location, store types and days, meaning differing temporal periods of data were available for each. Smaller, more rural stores typically demonstrated more conservative opening hours, whereas urban areas and transport hubs demonstrated trading both earlier in the morning and later at night. Sundays exhibited shorter opening periods across many stores (i.e. due to restricted trading hours).

3.2.3. Spatial Data Treatment and Aggregation

Whilst the majority of analyses presented in this thesis were conducted on non-aggregate data, spatial aggregations were necessary for the presentation of outputs. These were performed from the available postcode units to census derived geographies, primarily to facilitate the linkage of these data to existing national statistics for contextualising results.

3.2.3.1. Census geography

Census geographies describe the subdivision of geographical areas for the purposes of the census (Martin, 2002) and facilitate the reporting of sociodemographic population characteristics. In the UK, these geographies consist of a hierarchical subdivision of local

government areas, to sub-authority areas (such as wards), to lower levels created specifically for census purposes such as enumeration districts (EDs) in 1971, 1981 and 1991 or Output Areas (OAs) in the more recent 2001 and 2011 censuses. The EDs utilised prior to 2001 were designed to facilitate effective hand delivery of census questionnaires. As such, they were limited in their representation of social, economic and demographic distributions of populations (Openshaw, 1984), and exhibited large variations in size and social homogeneity (Martin, 2000). In response to this was the adaption of the Automated Zoning Procedure (AZP; Openshaw, 1977) by Martin (2002), which facilitated the grouping of adjacent postcode areas in England and Wales depending on set criteria to create the currently used OAs. The criteria followed to optimise these units included 1) population size controls to reduce inter-OA variance, 2) maximising social homogeneity, and 3) deriving shapes that were as compact and as circular as possible. In addition to these criteria, OAs were designed to be constrained by obvious boundaries, such as major roads, and to nest within administrative geographies. Thresholds were also placed on the minimum numbers of residents and households per OA to ensure confidentiality of the data. For a full overview of the methodology, see Martin (1998, 2000, 2002) and Martin, Nolan and Tranmer (2001).

OAs represent the base unit and the lowest geographical level at which census estimates are provided. Super Output Areas (SOA) represent more aggregate levels and are built up from groups of OAs. SOAs for England and Wales include lower layer super output areas (LSOA) and middle layer super output areas (MSOA). MSOAs represent the most aggregate level of census geography available. As the focus of this analysis was GB, equivalent data were also obtained from Scottish Census records. Scottish Census geographies include OAs, Data Zones (DZ), that are equivalent to LSOAs, and Intermediate Zones (IZ) that are equivalent to MSOAs. There are some disparities between the characteristics of these geographies; for example, DZs and IZs typically have smaller population sizes than their LSOA and MSOA counterparts in England and Wales. Table 3.6 provides population statistics for each unit. For simplicity, the combined usage of these units where utilised in this thesis are termed OA, LSOA or MSOA.

Table 3.6: Census geography threshold statistics, GB.

	England/Wales			**Scotland**		
	OA	**LSOA**	**MSOA**	**OA**	**DZ**	**IZ**
Total zones	181,408	34,753	7,201	46,351	6,796	1279
Minimum residents	100	1,000	5,000	50	500	2,500
Maximum residents	625	3,000	15,000	n/a	1,000	6,000

It is important to acknowledge that aggregating the HSR data to these units gives rise to issues of ecological fallacy and MAUP (see Chapter 2, Section 2.2.1.1). For instance, analyses conducted on HSR data post aggregation would be subject to scale effects, where statistical

outcomes may be more pronounced the larger the scale (such as in correlations/regressions), and zonation effects, where results may vary if divided up differently at the same scale. In this context, a prominent zonation related issue is that census geographies are derived from a different base population to that of the HSR data. OAs represent clusters based on the socioeconomic dynamics of the general population at the time of the census, yet, the naturally occurring boundaries the HSR population and their consumption characteristics would evidently be different to those of the socioeconomic characteristics of the general population in 2011. Therefore, any outcomes from analysis on the aggregated HSR data would be subject to these limitations. Having acknowledged this, it should be emphasised that for the majority of analyses presented here, data were utilised at the individual level and subsequently aggregated to census units to facilitate non-disclosive presentation of results. However, there were instances in which aggregations were necessary to draw comparisons with census statistics. For example, when aiming to contextualise a novel dataset in terms of general population characteristics, linkage to census data facilitated the only viable means for quantification in the absence of quality reference data. As census data are supplied at aggregate scales, HSR data were aggregated in order to facilitate these comparisons.

Table 3.7 gives an overview of the number of HSR customers present at each level of census geography. In order to maintain a suitable level of disclosure, the majority of outputs were required to be presented at the most aggregate level of MSOA. Figure 3.2 shows the Location Quotient (LQ), of cardholders per MSOA across GB. The LQ illustrates how well represented cardholders were per MSOA, in comparison to underlying population volumes as estimated by the 2011 Census, demonstrating over-representation in many rural areas (an LQ of 1 means a region has an identical share of the total population in comparison to the reference data, a negative LQ value indicates a lower share, and a positive value a higher share; see Miller, Gibson and Wright, 1991). Merged MSOA and IZ data were obtained from the UK Data Service online repository (UKDS, 2018b). This included MSOA population weighted centroids and boundary shapefiles (clipped to the coastline for mapping). Boundary data were simplified for visualisation purposes, with a tolerance of 1000m using the Simplify function (Douglas-Peucker algorithm) in ESRI ArcMap.

Table 3.7: Volume of HSR customers present at each geographic level.

Geography (GB)	Total with customers (GB)	Min	Median	Mean	Max
Postcode	1,384,193	1	10	13	975
OA	227,421	1	78	77	1377
LSOA	41,729	36	423	414	2750
MSOA	8,480	289	2035	2070	5772

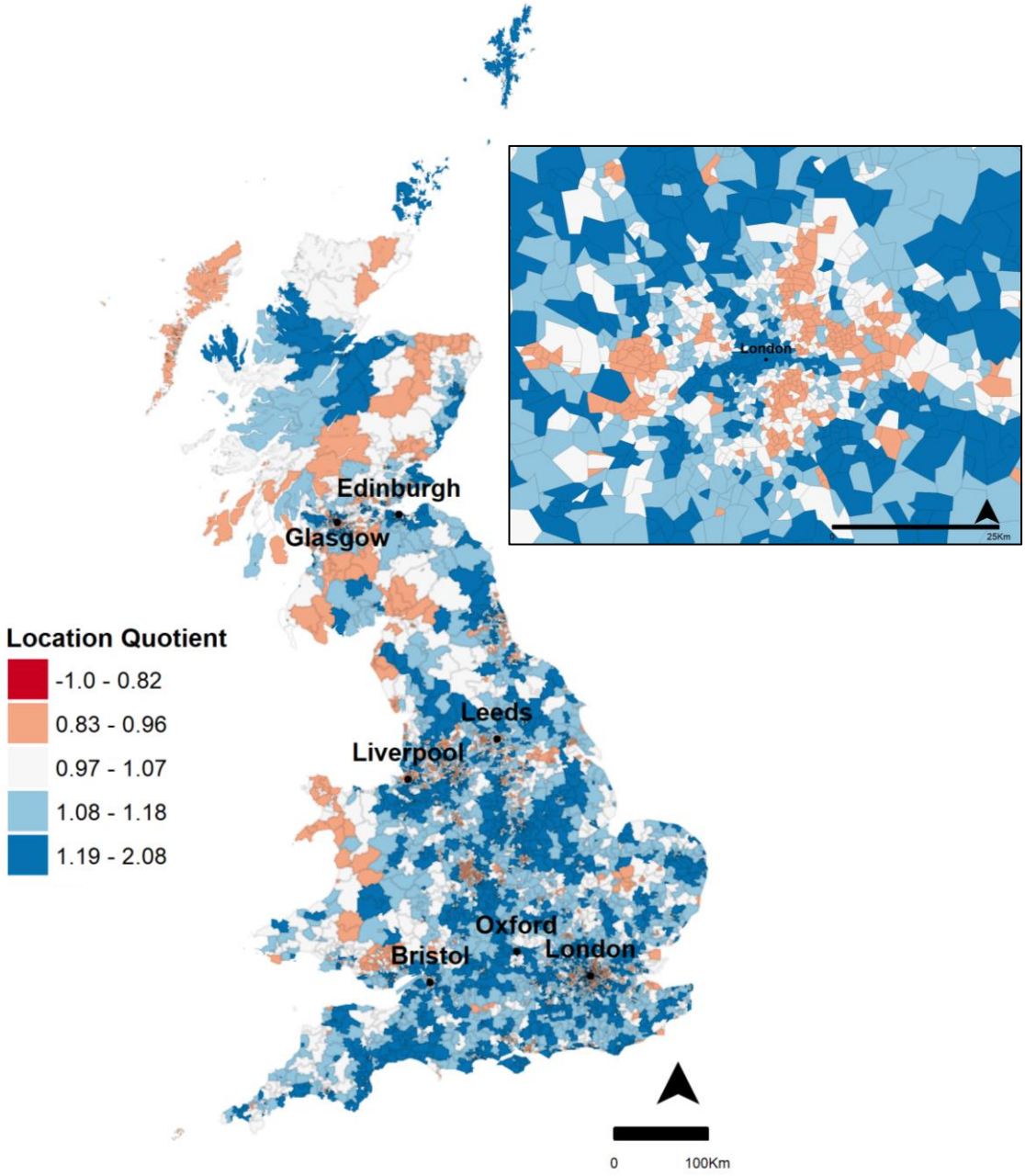


Figure 3.2: The LQ of cardholders per MSOA across GB and Greater London (inset).

Data were aggregated to the MSOA level to present the majority of outputs, as this was the lowest level that adhered to the disclosure controls outlined in Section 3.1 (and could therefore be extracted from the secure lab environment). However, other aggregation units (which align to OAs and SOA boundaries) included Local Authority Districts (LAD), Government Office Regions (GOR) and Workplace Zones (WZ). Larger units were utilised where MSOAs did not meet disclosure control requirements, or where the desired reference data were only available at a more aggregate scale. There are 348 LAD's in England and Wales (32 equivalent 'Council Areas' in Scotland) and 11 regions in GB (Scotland, Wales and 9 English regions).

Workplace zones (WZ) are a geography for England and Wales only that were produced using 2011 workplace data (information collected about workers and workplaces) by Cockings et al., (2015). These units were designed to supplement OAs and SOAs (which were created using residential population data) by providing statistics relevant to daytime workplace activities. Thus, this geography was relevant for analyses pertaining to the daytime activities of HSR customers rather than residential based, 'night-time' characteristics (OAs are designed to contain consistent numbers of people based on where they live, WZs are designed to contain consistent numbers of workers, based on where people work). WZs are constrained to MSOA boundaries to provide consistency between the OA and WZ geographies. These data were obtained from the Office for National Statistics (ONS) online data repository.

3.2.4. Supporting Data

In order to contextualise many of the findings presented in this thesis, a number of supporting datasets were obtained. These included both raw data and classifications derived from the 2011 Census. Tables 3.8 and 3.9 provide a description of each dataset.

Table 3.8: Supporting data – Census based classifications.

	Census Classifications	
Dataset	**Description**	**Reference**
Output Area Classification (OAC)	Describes geodemographic population characteristics across 8 Supergroups, 26 Groups and 76 Subgroups at the OA level, derived from Census variables (obtained from the ONS). Available at the OA level.	Gale et al., (2016)
Unified Rural Urban Classification (RUC)	Describes the characteristics of LAD's across 6 different classes based on population density, from the most rural 'Sparse/Remote Villages/Dwellings', to the most urban - 'Large Urban Areas' (obtained from the CDRC). Available at the LA level.	O'Brien (2016)
Local Authority Classification (LAC)	The LAC was created from Census variables and summarises the characteristics of LAD's across 8 Supergroups, 15 Groups and 29 Subgroups (such as 'Business and Education centres' and 'Rural England'). Obtained from the ONS. Available at the LA level.	ONS (2017)
Classification of Workplace Zones (COWZ)	Classifies the characteristics of WZ's based on Census data in order to differentiate different types of workers and workplaces. Characteristics are summarised across 7 Supergroups and 29 Groups (such 'Big City Life' and 'Market Squares'). Obtained from the ONS. Available at the WZ level.	Cockings, Martin and Harfoot (2015)

Table 3.9: Supporting Census data.

Census Data		
Dataset	**Description**	**Reference**
2011 Census Variables	Variables were obtained for analyses including those pertaining to demographic structure, housing and socioeconomic characteristics (i.e. employment, education). Obtained from the ONS for England/Wales and the National Records of Scotland (NRS). Available at the OA level.	ONS (2018a) NRS (2018)
Origin-destination statistics – Flow data	Includes the travel-to-work patterns of individuals. Obtained from the ONS for England/Wales. Available at the MSOA level.	ONS (2018b)
Origin-destination statistics - National internal migration	Describe moves that occurred from each area to elsewhere within the UK in the year preceding the Census. Obtained from the ONS for England/Wales. Available at the MSOA level.	ONS (2018b)
Origin-destination statistics - Student internal migration	Describes the migration patterns of those living at a student address in the year preceding the Census. Obtained from the ONS for England/Wales. Available at the LA level.	ONS (2018b)

These data were utilised in the analyses presented over the proceeding chapters, in order to draw broad comparisons between trends identified in the HSR data with those of the general population.

3.3. Preliminary Analyses – Representation and Uncertainty

Utilising data created in a commercial setting raises substantial methodological questions when attempting to apply them in research. For example, there is a complete absence of researcher control in the data collection process and it is therefore necessary to determine data quality, uncertainty and fitness for purpose to extract and interpret meaningful insights. As outlined in Chapter 2 (Section 2.3.1.2) important areas of consideration included identifying potential bias in the sample, determining the quantity, consistency and completeness of data and also assessing the plausibility of observed trends.

Access to this loyalty card dataset offered a unique opportunity to study its dynamics in an ethical and secure environment and therefore address substantial gaps in existing research. The proceeding section presents an exploratory analysis that aimed to understand the dynamics of loyalty card data and highlight a number of pragmatic measures that should be considered when implementing these data in research practice. This analysis aimed to firstly, identify potential data quality issues inherent in this novel form of data, and secondly, investigate to what extent we can generalise insights from loyalty card data to the wider population (and thus, quantify their representation). These questions had the wider aims of firstly, informing interpretation of the subsequent analyses presented in this thesis and secondly, understanding the potential limitations of applying big datasets such as these, in social science research.

3.3.1. Customer Attributes

An issue identified from loyalty card literature was that the loyalty population is likely subject to effects of self-selection, where customers select themselves to participate and therefore represent an inherently biased sample. In order to develop an understanding of the extent and dynamics of bias in these data, a data-driven approach was applied to investigate the demographic characteristics provided by HSR customer metadata.

3.3.1.1. Method

To understand the representativeness of these data in terms of the general population, firstly, age and gender attributes were compared to GB population estimates from the 2011 Census. Census data represent population attributes on the day of the census, therefore, customer ages were calculated (from their date of birth) on this date for comparability. Following this, each dataset was normalised by their total population, in order to account for underlying base populations (and thus compare proportional age distributions within each).

Secondly, facilitated by the provision of customer postcodes, comparisons were drawn with census based geodemographic indicators. This was achieved by firstly, performing an exploratory regression analysis between the numbers of cardholders per OA and census variables deemed indicative of socioeconomic status. Census variables are available at the OA level, therefore HSR customer counts were aggregated from postcodes to OA to facilitate analysis. A variety of variables were tested including those describing occupation, education, economic activity, estimated social grade, health and household structure. Regressions were conducted using the linear model (*lm*) function in R, using the number of cardholders as the dependent variable and census data as the independent variables. The simple linear regression model can be expressed as:

$$Y = a + bX \tag{3.1}$$

where X is the explanatory variable and Y is the dependent variable, b is the slope of the line, and a is the intercept (the value of $\boldsymbol{y}$ when $x = 0$). Finally, volumes of HSR customers were compared to geodemographic groups as derived by the OAC. For this, the frequency of individuals per classification group were obtained for both HSR customers and census statistics. Each were normalised by their relevant population denominators and proportions compared.

3.3.1.2. Results

Approximately 88.6% of HSR cardholders were female and only 10.9% male (0.3% undisclosed). In comparison to gender specific census population estimates, this translated to a representation of approximately 52.5% of the GB female population and only 6.2% of the male population. Figures 3.3 and 3.4 show the age and gender distributions of 16-85 year olds for the loyalty population (normalised by total customers) and the census estimates (normalised by total census population) and Figure 3.5 illustrates the relationships between the number of cardholders per OA and census variables: social grade, occupation and qualification.

It is evident from age distributions that the male population is substantially underrepresented by this dataset, although there are a higher proportion of younger male customers than other cohorts (i.e. 26-35). A peak can be observed across both genders at approximately 61-65 years, which is consistent with a growth in census population estimates for these age groups. These data are most representative of the older female population (ages 50-60), yet under representative of the youngest (16-20) and eldest (> 70). These data are also likely to be over representative of younger and middle-aged females (21- 50).

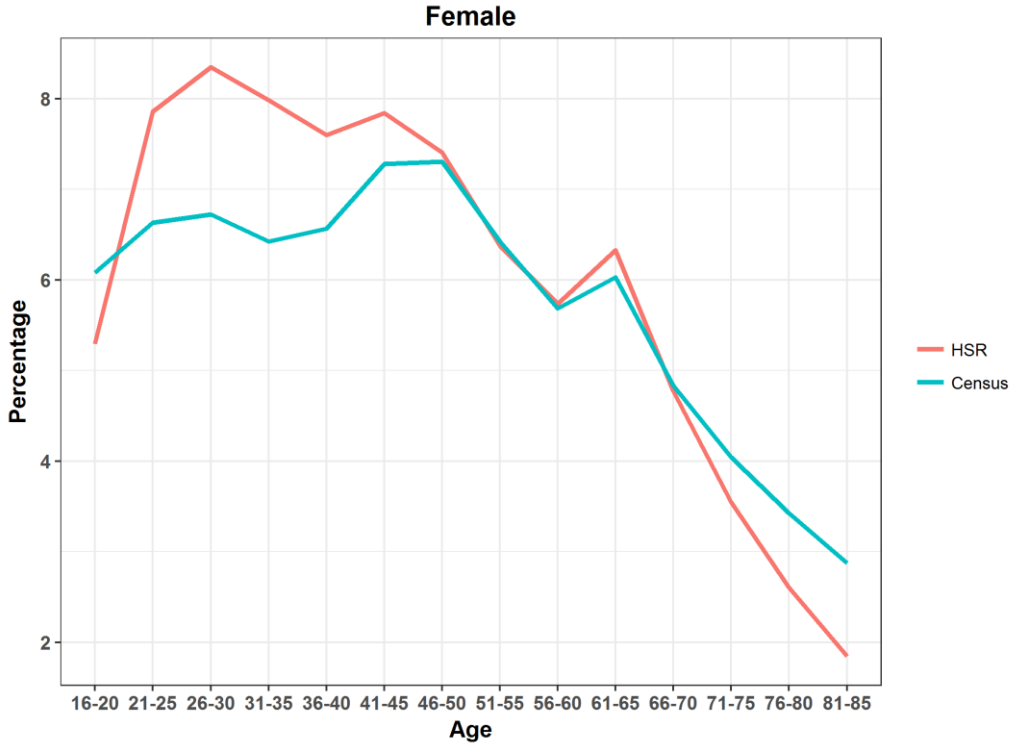


Figure 3.3: Female cardholder age distributions compared to census population estimates.

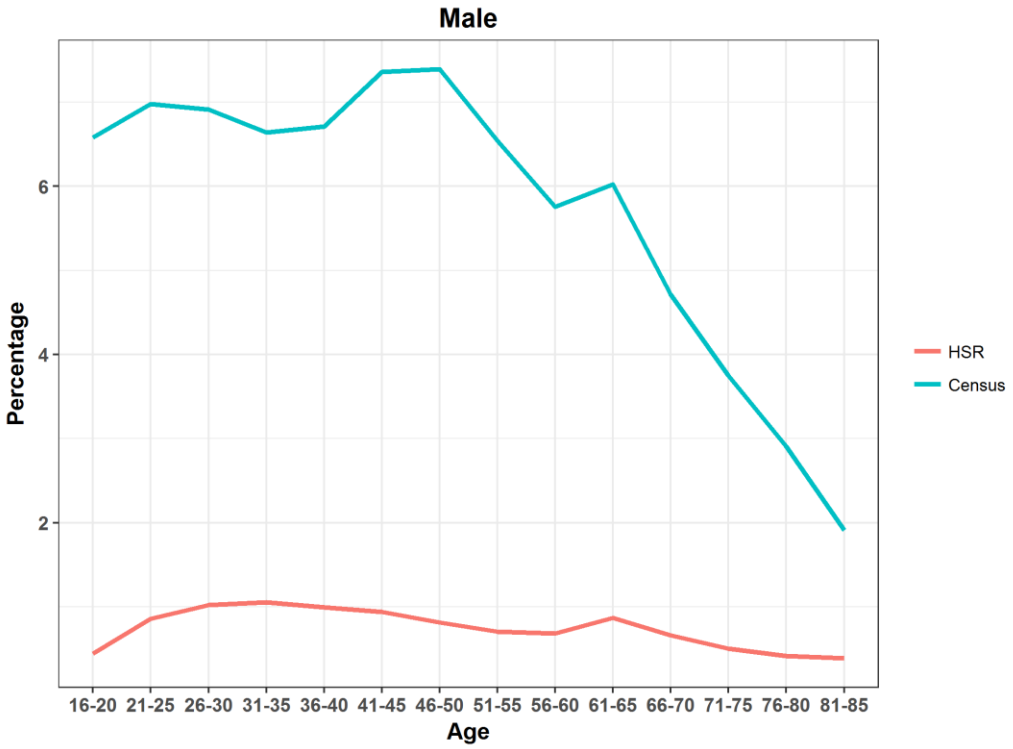
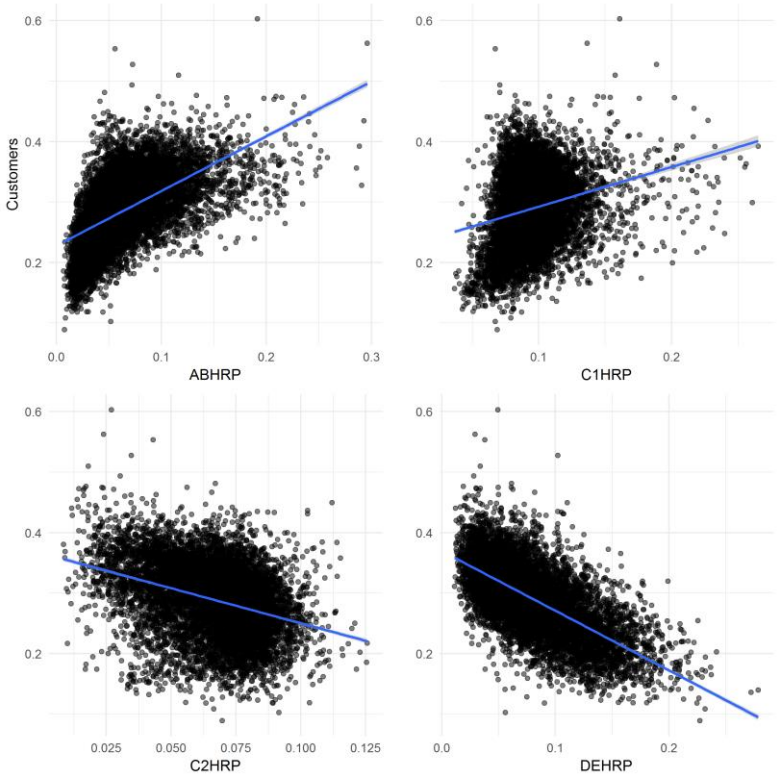


Figure 3.4: Male cardholder age distributions compared to census population estimates.

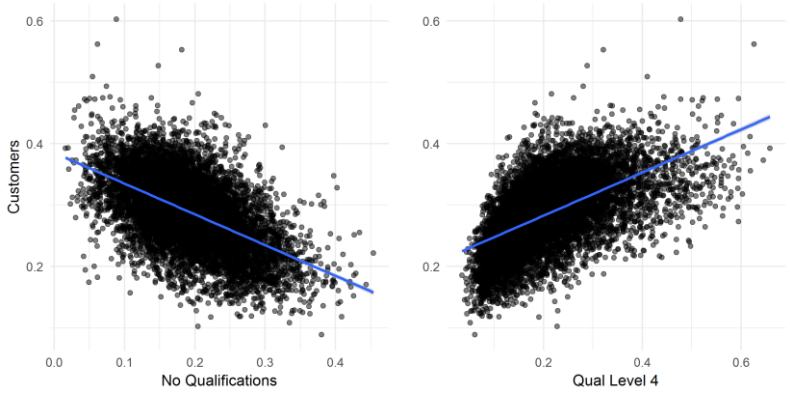
a) Social grade



AB *- Higher and intermediate professional occupations,* ***C1*** *– Supervisory, clerical and junior professional occupations,* ***C2*** *– Skilled manual occupations,* ***DE*** *– Semi-skilled/unskilled manual occupations, unemployed and lowest grade occupations.*

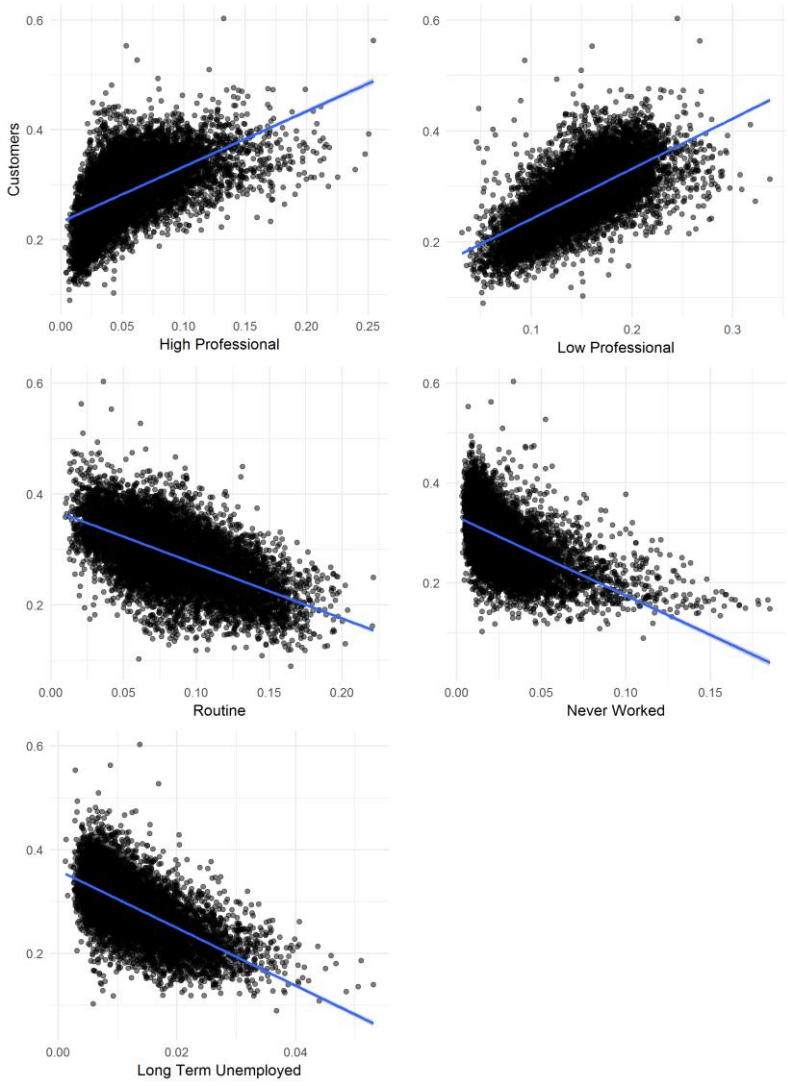
Adjusted R^2: 0.4735, $p < 2.2e\text{-}16$

b) Qualification



Adjusted R^2: 0.7922, $p < 2.2e\text{-}16$

c) Occupation



Adjusted R^2: 0.5382, $p < 2.2e\text{-}16$

Figure 3.5: Scatterplots demonstrating relationships between cardholders per OA and a) Social grade b) Qualification and c) Occupation.

The regression analysis indicated that cardholder volumes might be highly indicative of socioeconomic status and thus, these data may be under representative of less affluent segments of the general population. Rates of cardholders significantly declined with social grade, qualification levels and occupation type. There was a positive relationship between increased number of cardholders, high social grade, high tier qualifications and occupations. These relationships were consistent across many of the selected variables, showing positive relationships with older cohorts, very good health and smaller households and negative relationships with bad health, large households and economic inactivity. These relationships may help to explain patterns observed in the GB cardholder distributions (see Figure 3.6). For example, renowned affluent areas (such as Westminster and surrounding boroughs), and affluent suburban areas showed higher proportions of cardholders, whereas lower proportions were evident in less affluent areas such as Lambeth, South West London.

Figure 3.6: LQ of cardholders, Greater London.

Figure 3.7 demonstrates the volumes of HSR customers across OAC groups in comparison to census estimates. It's clear that geodemographic groups are disproportionately represented by these data, with more affluent groups likely being over-represented (particularly ageing suburban cohorts and young professionals), and deprived neighbourhoods/less affluent segments of the general population under-represented. These dynamics will likely bias the resulting spatial distribution of customers who are signed up to the scheme (for instance, there

may be an over-representation of suburban populations), with volumes of customers across areas reflecting underlying socioeconomic characteristics rather than the general population.

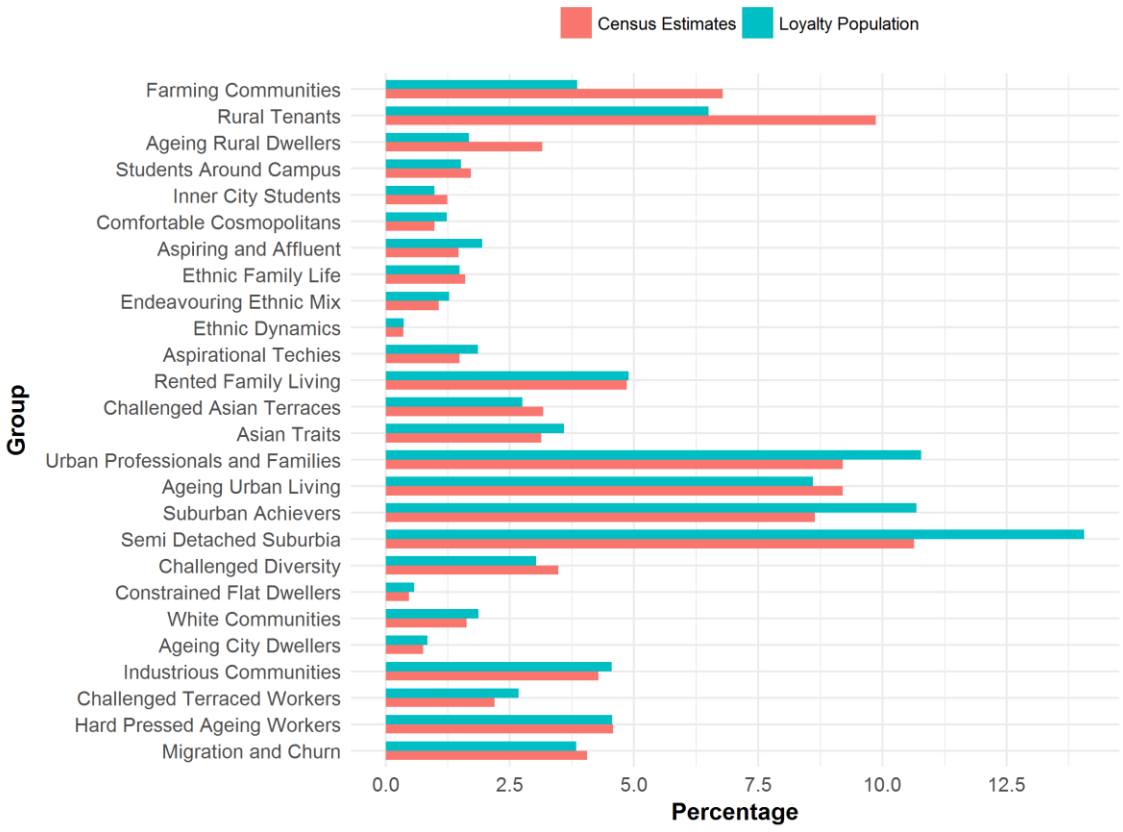


Figure 3.7: Proportion of customers per OAC Group in the loyalty versus census population.

A further observation from this analysis was an abnormality in the number of cardholders registered to some MSOAs. For example, in some instances there was a higher number of cardholders than residential census population estimates. This was evident, for example, in the City of London (Figure 3.6). Further investigation of these areas revealed that these could be attributed to postcodes comprising of either workplace locations or university campuses. This has important implications for the utilisation of postcode data provided in loyalty card data, indicating that the locational information provided by a customer may not always be representative of a place of household residence. These observations prompted the need for further investigation into the reliability of customer postcode attributes, of which are presented in Chapter 4.

3.3.2. Spatial Attributes

Further important bias considerations arise from the pre-defined HSR store location network. Research suggests that store locations can play a key role in loyalty card ownership, and that the distribution of cardholders may be influenced by the accessibility of these locations to

consumers. To understand potential bias in coverage as a result of the HSR network, store locations were augmented with the RUC. Following this, the number of customers per area type were augmented with this classification and compared to volumes within the general population (derived from total census population per area type). Figures 3.8 – 3.10 illustrate a) the percentage of stores per area type, b) the percentage of customers per area type and c) a comparison of cardholders and census population per area type.

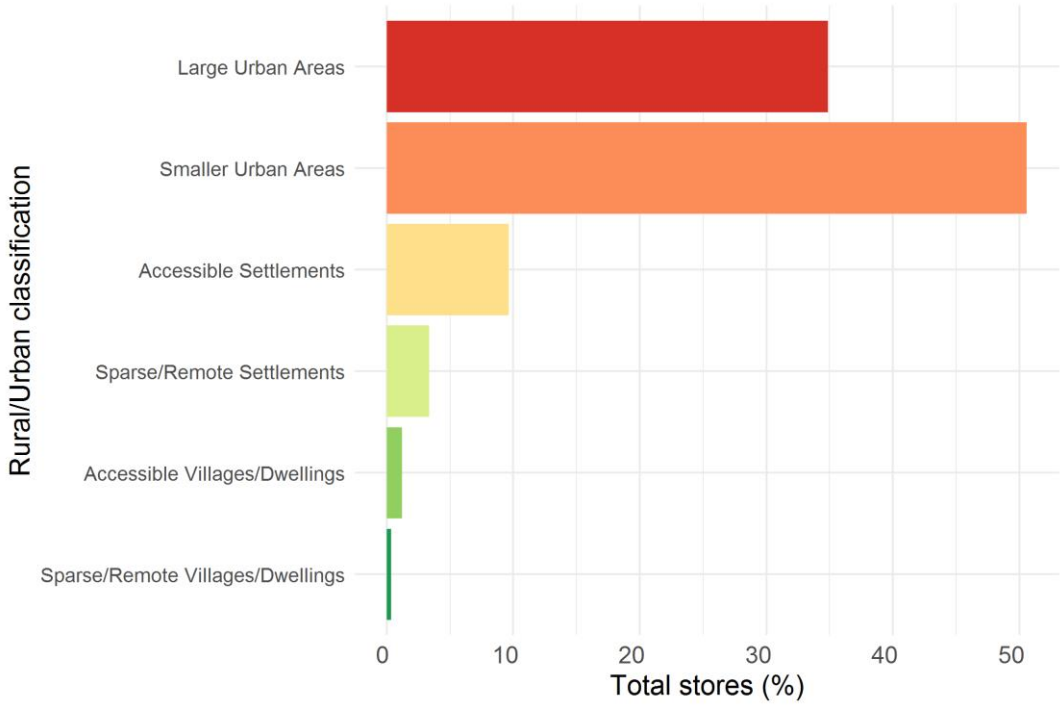


Figure 3.8: The percentage of HSR stores per RUC type.

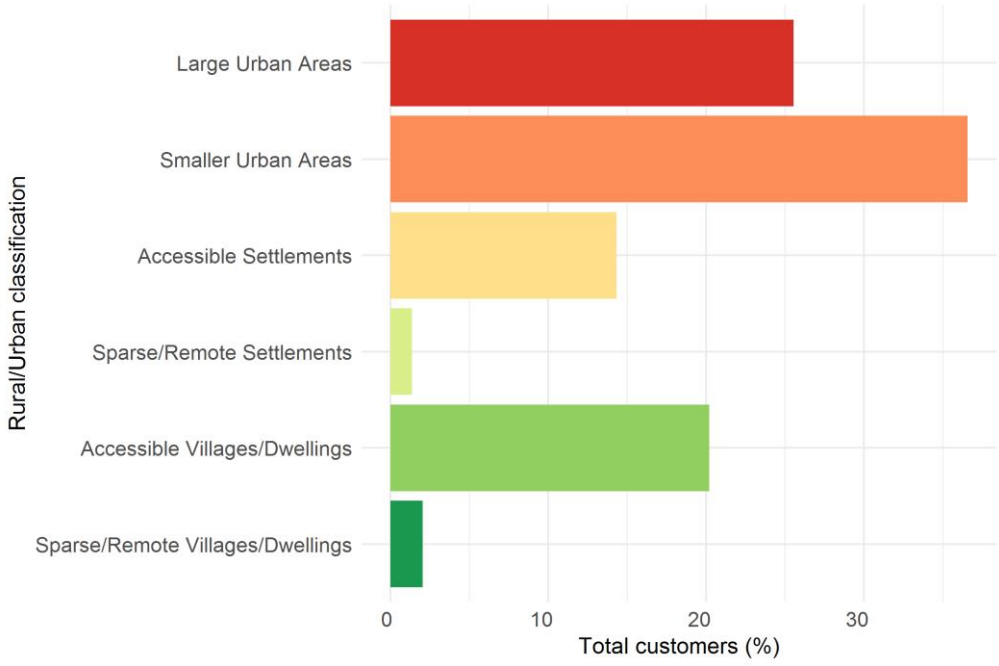


Figure 3.9: The percentage of HSR customers per RUC type.

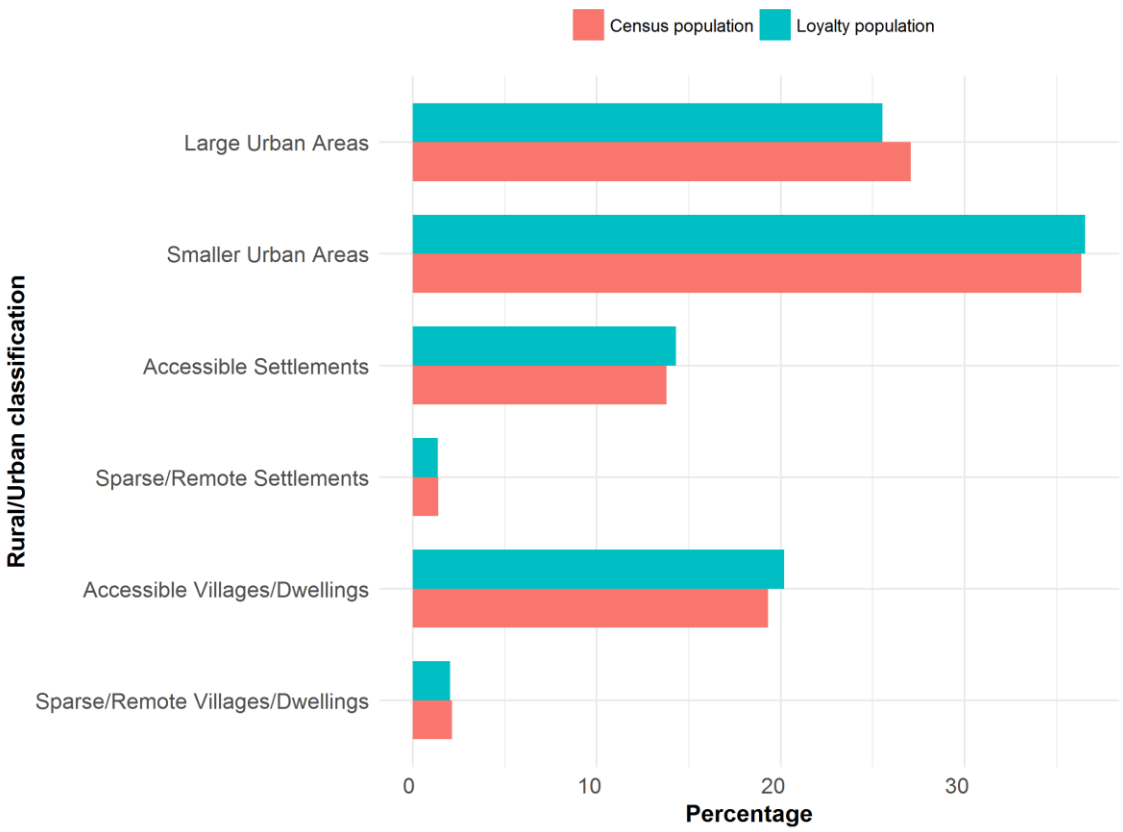


Figure 3.10: The percentage of HSR customers per RUC type compared to volumes in the general population (derived from 2011 Census estimates).

As is evident, the locations of HSR stores were considerably biased towards urban areas. Distributions of HSR customers per area also demonstrated this pattern, although with a larger proportion within 'Accessible Villages/Dwellings'. However, comparing proportions with census estimates revealed that volumes per area type are likely extremely representative of dynamics in the general population. Therefore, the biased location types of stores appeared to have minimal effect on the propensity to attract customers, for example, from more rural areas. Comparing the total census population per area type with the total customers (see Appendix 2) indicated that the loyalty card data represent approximately 30% of the population within each area. In contrast to the distribution of store locations, the most represented area type was 'Accessible Villages/Dwellings', and the lowest 'Large Urban Areas'.

3.3.3. Transactional Data

A final important area for investigation in loyalty card representativeness was the uncertainties that arise from individual differences in card usage. As outlined in Chapter 2 (Section 2.2.3), this could potentially lead to a disproportionate representation of behaviour across the database. For example, our ability to extract information about each individual may be influenced by

variations in transactional volumes and differing motivations to participate across location types, product categories and time periods.

3.3.3.1. Method

To investigate variation in individual card usage, descriptive statistics were obtained regarding the volume, frequency and duration of activity per customer over the 2.5 financial years. These are commonly utilised measures when aiming to quantify loyalty/consumption behaviours (i.e. see Allaway et al., 2006). Firstly, general activity volumes were quantified by investigating the total number of transactions per customer and secondly, frequency of activity was quantified by assessing the number of unique weeks that each customer was active throughout the database. Weeks were extracted and defined using the PostgreSQL *week* function. Average intervals were then calculated by dividing unique weeks active by the total duration of activity, in weeks (defined by the first minus the last day of transactions). These measures were applied as, firstly, obtaining only the number of days a customer was active would not provide insight into the longitudinal nature of their activities. Secondly, calculating intervals based only on duration divided by transaction volumes may bias activity intervals if a larger number of transactions were recorded within a short time period.

Variations in card usage across locations were investigated by comparing volumes of transactions, product consumption and spend by HSR store types. As these were derived from the clustering of locational characteristics, this was deemed appropriate for quantifying broad differences in participation (for example, within convenience versus destination type stores). Finally, variations in temporal consumption were investigated by obtaining frequencies of transactions during hourly, daily/weekly (Monday to Sunday) and monthly intervals over the 2.5 years of data.

3.3.3.2. Results

Transactional volumes varied substantially between customers. Overall, 0.66% of customers had never transacted, 2.6% had transacted only once and approximately a third of all customers (33%) less than 10 times over the 2.5 financial years. Approximately 23.4% of all customers were responsible for 60% of all transactions (see Figure 3.11). In terms of transaction frequencies, only 9% were active on a weekly basis, yet approximately 7 million, or 38%, of HSR customers exhibited monthly activity patterns over a two-year period. Therefore, in relation to the total sample size, these data still represented a significantly large and rich source of data in comparison to traditional studies of population activity over longitudinal periods. Variations in card usage were also evident across different retail locations (see Figure 3.12). These trends suggested that the representation of different store locations will vary in terms of

transactional, product and spend volumes and that certain locations will facilitate higher volumes of data than others.

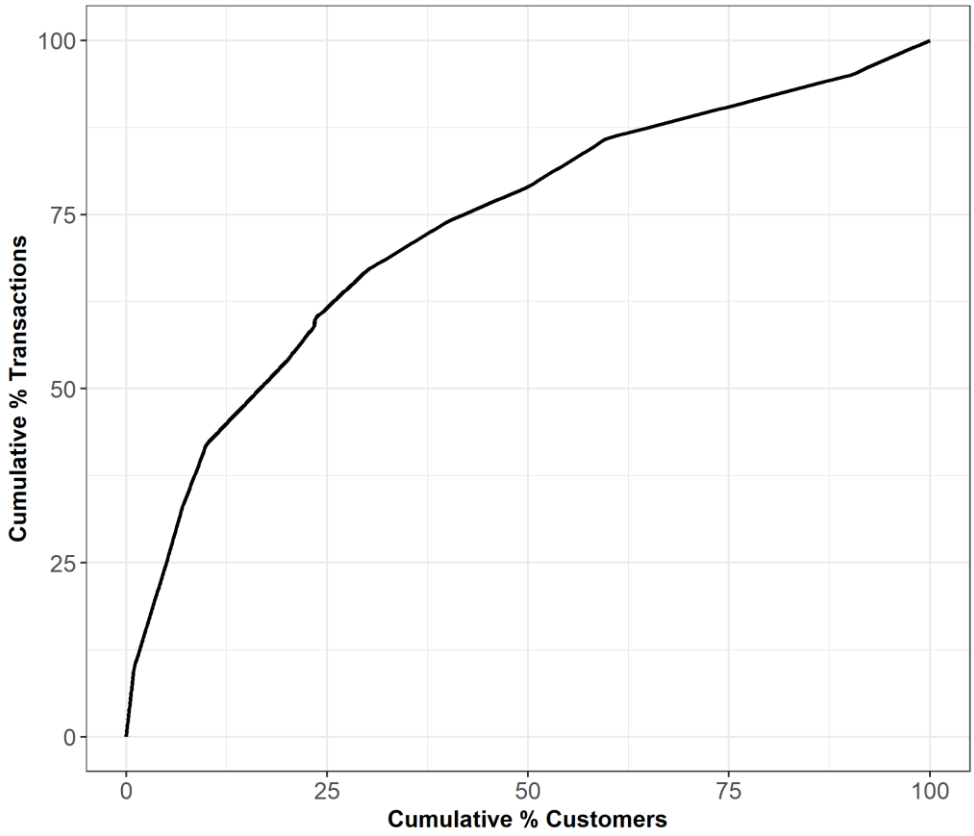


Figure 3.11. Cumulative percentage of transactions by percentage of customers.

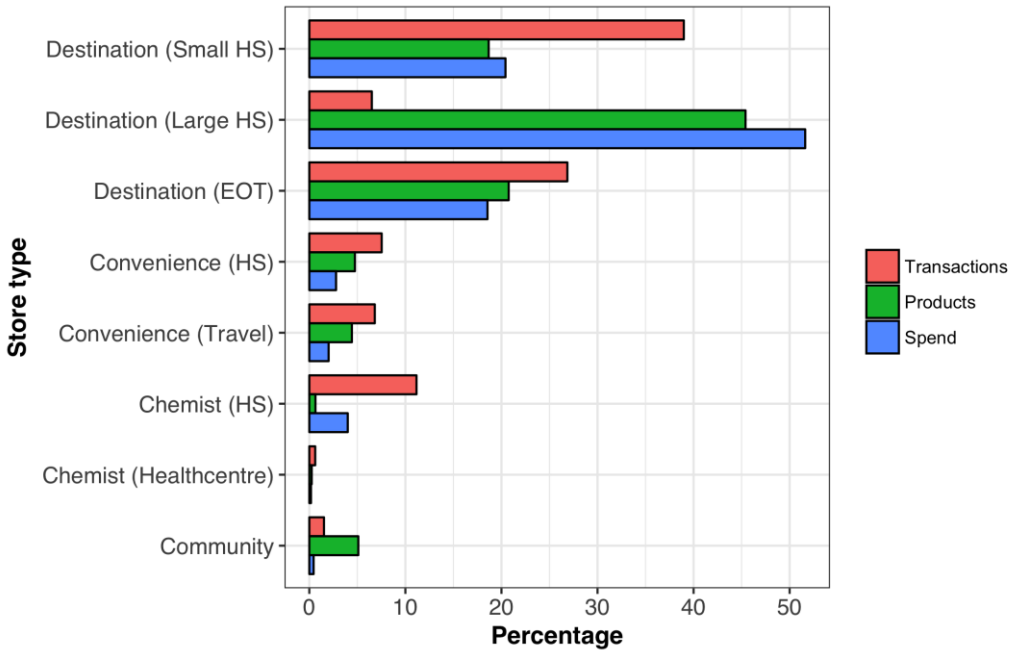


Figure 3.12: Percentage of total card transactions, product consumption and spend by store type.

Significantly lower levels of participation were observed in ‘Convenience’ and ‘Community’ type stores. Conversely, higher levels of participation are observed in ‘Destination’ locations, such as small high streets and retail parks. These trends will be influenced by the HSR store network structure, which contains higher proportions of certain store types (i.e. small high street stores) and thus will inevitably account for a higher percentage of overall data. However, biases may arise from the differing motivations to participate within these locations. For example, higher basket sizes (referring to the number of products bought in one transaction) are evident in some location types, as can be observed from transaction versus product volumes. This was apparent in ‘Destination’ type stores (primarily large high streets), which as highlighted in the literature, may produce higher loyalty participation due to the perceived benefits/rewards of more expensive purchases. Due to the inability to aggregate non-card data to a transaction level here, it was not possible to quantify if lower volumes of data in certain store types (i.e. ‘Convenience’, ‘Chemists’ or ‘Community’ stores) is a result of lower participation, or lower overall transactions. However, it is clear that there will be a disproportionate representation of behaviour across location types when utilising loyalty card data. The implications of these trends are that the distribution of behavioural data in space will be influenced by the characteristics of a store location. Ultimately, the completeness of individual trajectories may be influenced by these differing motivations to participate. Despite this, due to the volume of overall data, there is still a vast amount of data produced by loyalty cards available across all store locations. Finally, analysis of temporal trends revealed variations in transactional volumes during different time periods (see Figures 3.13 to 3.15). Higher volumes of activity are evident during lunchtimes, during weekday periods and during the summer and Christmas periods.

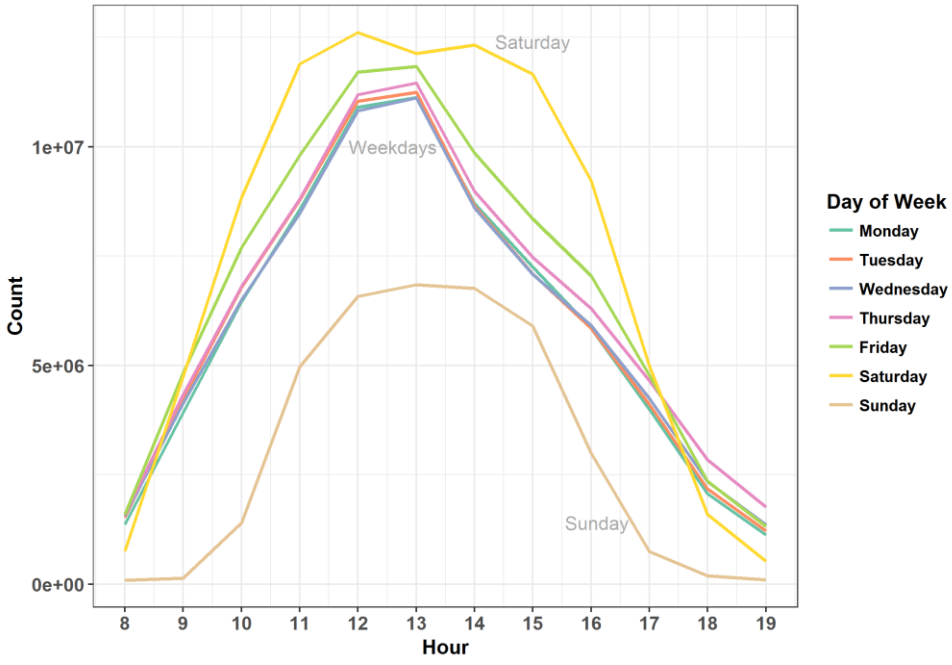


Figure 3.13: Transactions per hour, per day of week (over 2.5 financial years).

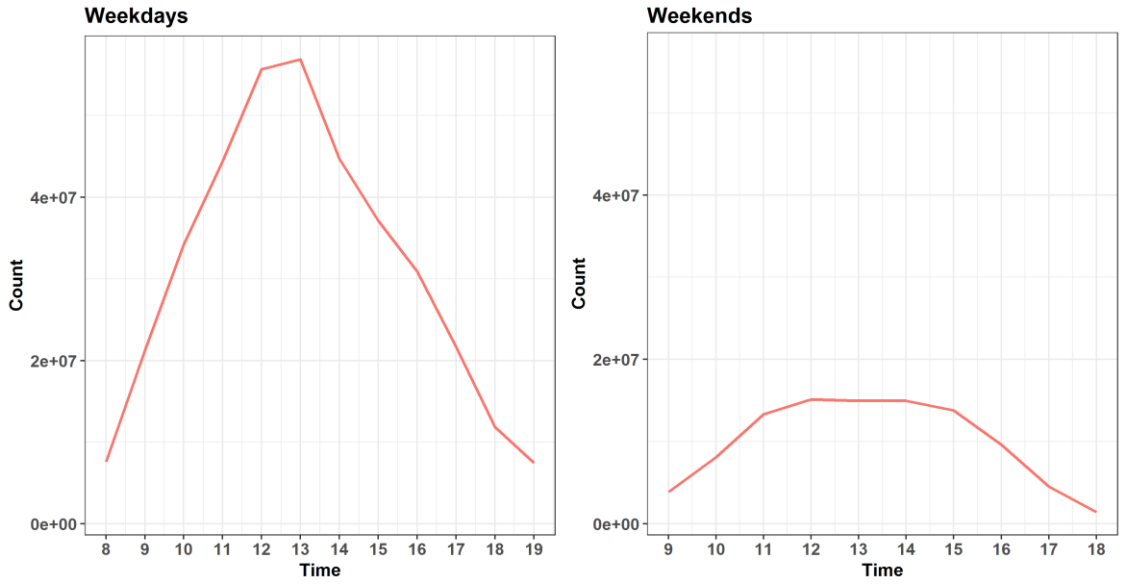


Figure 3.14: Total transaction for weekdays and weekends per hour (aggregated over 2.5 financial years).

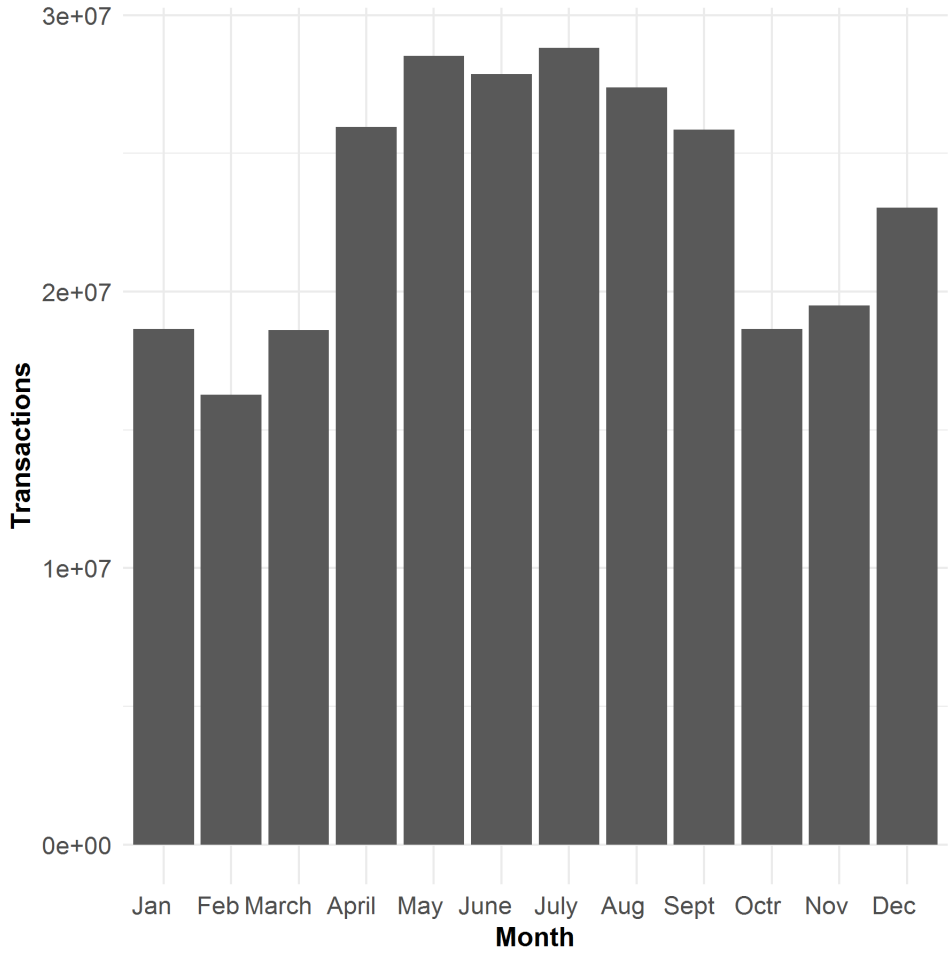


Figure 3.15: Total transactions (count) per month (aggregated over 2.5 financial years).

These trends demonstrated that certain time periods exhibit higher levels of behavioural data. This has important implications if conducting temporal analyses, as patterns will be biased by the underlying volume of transactional data per time period.

3.4. Summary and Conclusions

The exploratory analysis of customer attributes suggested that these data are biased towards certain segments of the general population, primarily middle-aged females of higher socioeconomic status. The male population, and those with lower socioeconomic status may be considerably underrepresented. Comparison to OAC Groups further implicated these findings, showing an uneven distribution of cardholders across geodemographic segments. These issues are fundamental to take into account if attempting to extrapolate the dynamics of loyalty card holders to that of the general population. Furthermore, analysis of the distribution of customers revealed that customer postcodes may not always represent a customer's place of residence (i.e. some may be workplaces), which has important implications when utilising these data to inform residential based geodemographic phenomena.

Important findings from the analysis of store distributions were that HSR locations are not evenly distributed across rural/urban place types, and predominantly reside in urban areas. However, comparing the distribution of HSR customers with Census estimates suggested these data are highly representative of the general population volumes residing in each area type (approximately 30% within each). This may be a result of the expansive HSR store network (i.e. with a presence on almost every GB high street) and therefore represents an advantage of these data. Analysis of card usage dynamics revealed uneven volumes in individual activity. These dynamics are particularly important when studying general consumption patterns as outputs will be biased by the amount of data available per customer. There may also be differing motivations to participate in certain store locations.

It is important to note that there are a number of areas in which uncertainty could not be quantified in these data. Firstly, whilst ages outside of normal ranges (i.e. 16-85) can be easily identified, errors falling within a normal age range will be unidentifiable in this context. Secondly, an area of uncertainty identified from loyalty card literature was potential differences between card member and non-member behaviour. Due to being unable to aggregate non-card data to a transaction level, direct comparisons could not be made in this context. Furthermore, direct comparisons of these data would be inherently problematic due to non-card data also comprising of instances where a cardholder did not use their card with a transaction. This creates particular difficulties when comparing product consumption between card and non-card customers and thus gauging if consumption of higher value products is over-indexed in loyalty card data. Thirdly, in relation to this, it is important to consider that it is unlikely that individual

records represent a person's entire transactional history due to non-consistent card usage. It is not possible to quantify, using data-driven methods, the extent to which this occurs.

These represent inherent limitations of utilising loyalty card data. However, access to this unique dataset allowed quantification of a number of representation and uncertainty/data quality issues that have not been previously obtainable via data-driven methods. These preliminary insights were central to the utilisation of these data to inform population dynamics, as understanding their applications is not possible without understanding of their limitations. These insights were used to inform proceeding analyses and interpret outputs, in order to understand the extent to which we can make broad inferences about the general population from a novel and inherently biased dataset.

4. Detecting Address Uncertainty in Loyalty Card Data

4.1. Introduction

One of the primary aims of this thesis was to investigate our ability to extract socio-spatial insights from a novel consumer dataset and understand their relevance as population indicators, in comparison to conventional census based measures. The customer postcodes provided in loyalty card data can be viewed as the key to achieving these aims. For example, they facilitate linkage of HSR customers to residential context and augmentation with census statistics (in order to understand bias and contextualise trends), which are both fundamental if we aim to enrich geodemographic representations through novel data sources. However, in this context, customer addresses are volunteered information and thus reliant on accurate human input, meaning that this information may be inherently uncertain. Given the importance of this element for applications in social science research, this raised substantial questions, and warranted preliminary investigations of the representativeness of these attributes. As summarised by Graham and Shelton (2013), when utilising novel forms of data it is crucial that we begin by assessing the accuracy of the spatial information provided to avoid obscuring important social and spatial processes. This represents one of the fundamental issues that arise due to the nature of these data being produced as a by-product of alternative commercial agendas, rather than conforming to the rigours of more traditional approaches to data collection.

Whilst many types of error in these data are to a certain extent, easily identifiable (for example, invalid postcodes), a more complex issue arises from temporal data errors – where an object being represented changes character between the time of data collection and when the data are utilised. Information regarding updated address information are not provided in these data, therefore, the information can only be assumed to be representative of a customer's current place of residence. This raises issues not only because the data here are historical in nature, but also because in the twenty-first century, places of residence may be transient (Van der Klis and Karsten, 2009; Sheller, 2011). Data pertaining to changes in residence are seldom able to be captured by traditional methods, however, the 2011 Census estimated that 7.5 million people changed address within the year prior to the Census (ONS, 2014). Recent research attempting to identify annual population change through novel forms

of data, such as consumer registers (Lansley, Li and Longley, 2017), also estimated a similar magnitude of migration.

A challenge in resolving this issue is that there is no guidance on suitable methods or heuristics to quantify the existence temporal address errors in this context. This is largely a result of lack of access outside of the commercial contexts in which they are created hindering both understanding of these dynamics and development of methods to address them. Since they are often hard to obtain for academic research, this investigation offered a unique opportunity to explore of the veracity of address attributes in the HSR loyalty card data. The primary objectives were too:

1) Develop a means of quantifying potentially inaccurate address information in the absence of reference data.
2) Explore and attempt to contextualise these findings in relation to existing population statistics.

To achieve these aims, firstly, data-driven heuristics were constructed that utilised customer transactions to estimate the credibility of their address information, by drawing on current knowledge and theory of spatial behaviour. Secondly, since the customer postcodes offer a basis by which each customer record could be linked to conventional statistical geographic units, results were compared to and augmented with existing national statistics in an attempt to provide a pragmatic means of validation. Finally, as an extension to these analyses, it is demonstrated how this information may be further utilised to some extent, to estimate the new locations of these individuals.

4.2. Exploratory Analysis

Exploratory analyses were conducted to identify potential uncertainty in the address data and inform method development. As discussed in Chapter 2 (Section 2.1.2) various methods of assessing veracity in novel data sources have been proposed. Of particular relevance for this work is that of the *knowledge solution* (Miller and Goodchild, 2015). This proposes that we can utilise existing theory to ascertain whether observable patterns are logically consistent with what is already known about the geographic world. This concept was applied, using abductive reasoning, to explore interactions between customer addresses and store visiting behaviours based on our existing knowledge of spatial behaviour and human mobility.

The theory was applied that assumptions can be made as to what constitutes uncertain travel patterns due to fundamental constraints imposed on daily human mobility. For example, a home location can be considered as one of the moorings that define spatial movement (i.e. journeys are likely to begin from and end at home) and the location in which one lives

therefore poses spatial and temporal constraints that affect the daily movement patterns and lifestyle of an individual (i.e. see Ellegård and Vilhelmson, 2004; Larsen and Urry, 2016). Notwithstanding increased ease of mobility due to available transportation (Sheller and Urry, 2006), daily movements surrounding this home mooring are still likely to be characterised by regularity (i.e. González et al., 2008; Song et al., 2010), as movement will be constrained by physical barriers of distance. For instance, there has been a long history of literature concerning the concept of distance decay (i.e. Tobler, 1970; Wilson, 1971; Taylor, 1971) and related gravity models (Tinbergen, 1962; Huff, 1963), postulating that the interaction between two locations declines with the increasing distance, time, and cost between them (although it may be positively associated with the amount of activity at each location; Isard, 1956). However, true interpretation of irregular behaviour in this context required understanding of complex travel patterns. For example, travel behaviours may not always fit with what appears geographically logical, due to dynamics such as incorporating store visits into daily routines or obligations (i.e. trip chaining; Adler and Ben-Akiva, 1979), which can vary according to purpose (i.e. work, leisure, tourism; Edensor, 2012). Therefore, a proposed model needed to take into account that relying on principles of geography alone was not enough to untangle complex trip dynamics.

Figures 4.1 and 4.2 demonstrate examples of observations from the exploratory analysis. Figure 4.1 illustrates a sample of travel flows from customers' home locations (shown here as population weighted MSOA centroids for disclosure purposes) to their most frequently visited store, for 'Community' stores (only one store type was selected due to the large amount of overall data masking intelligible flows). These patterns indicated potential instances of deviation from expectations based on our knowledge of spatial behaviour. For example, it is unlikely that customers frequently travel long distances (i.e. from Scotland to the South coast of England) to visit stores.

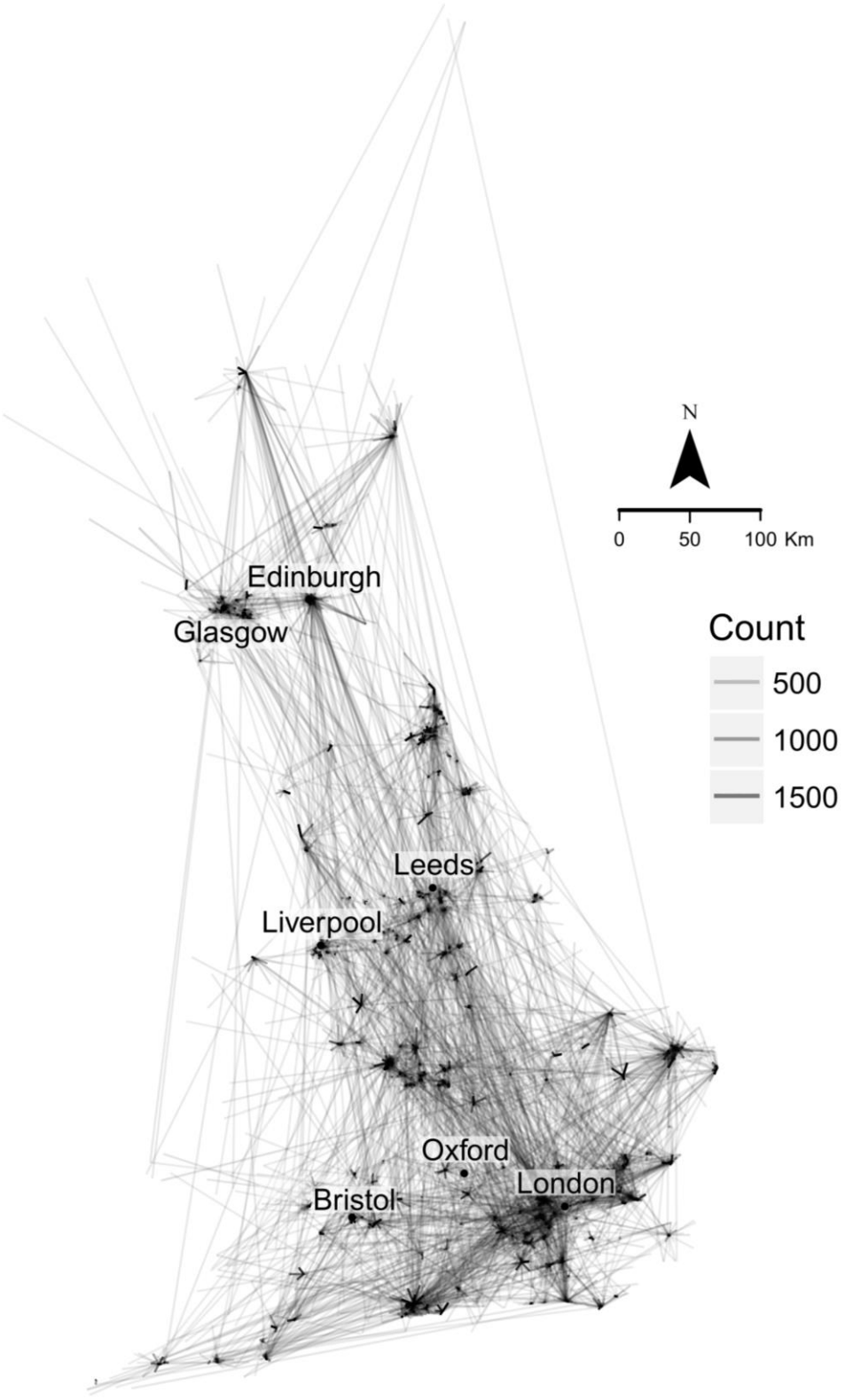


Figure 4.1: Flows from customers' origin MSOA to their most frequently visited store, for 'Community' type stores (showing a sample of 65,770 customers). Published in Lloyd and Cheshire (2018).

Figure 4.2: An example of ambiguous spatiotemporal transactional behaviour, recorded between 2012 and 2014. Published in Lloyd and Cheshire (2018).

Observing customers' transactional behaviours over time identified further ambiguities. Figure 4.2 shows an example for a customer registered to an MSOA in Northern England. In this instance, whilst their transactional behaviour appeared 'logically consistent' with their address at the beginning of records (i.e. within a local store network), they exhibited a permanent shift to inconsistent geographical areas after certain time periods (i.e. Oxfordshire in 2013, Bristol in 2014). This 'permanent shift' is defined in this context as an absence of further transactions within their initial network for the remainder of their recorded activity. These observations suggested a change in location that was not reflected by the postcode information provided in the data. Therefore, a method was needed to identify the extent to which these uncertain cases existed within the HSR database.

4.3. Detecting Address Uncertainty

4.3.1. Method

As outlined, the concept was applied that an individual will be anchored to their immediate geographical neighbourhood to some extent, and behaviour would be expected to occur (for

the majority), within a certain boundary of this location. In loyalty card data, an advantage is that spatial reference points are obtainable for both home locations (postcodes) and transactional behaviours (store locations). It was therefore possible to quantify the most prominent store locations for each residential area. This information could then be applied to interpret behaviours that were not consistent with residential areas, or when permanent changes in store networks occurred. This idea adopted early fundamental assumptions of human mobility, that a home boundary can be seen to represent an area in which the majority of time is spent and movement can be interpreted as when changes in the 'spatial points of reference' of a home mooring occur (Behr and Gober, 1982). Here, the address information provided home anchor points, and the method intended to define the importance of each store location (behavioural point of reference) to different home anchor points across Great Britain.

To achieve this, a data-driven method was constructed by drawing on knowledge and theory from multidisciplinary domains. Figure 4.3 gives an overview of the process applied.

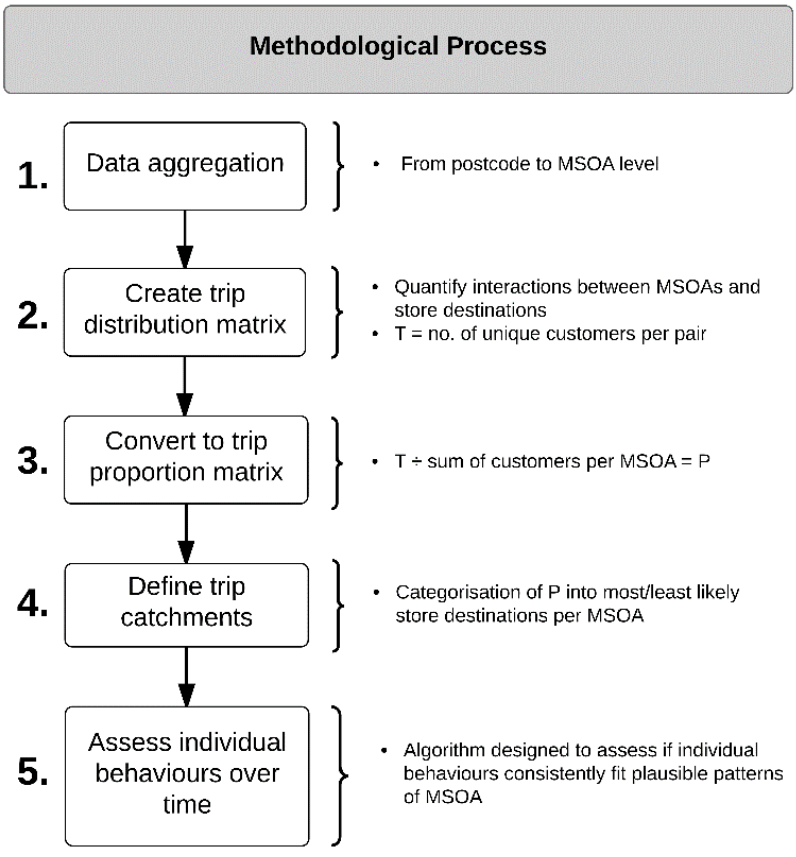


Figure 4.3: An overview of the methodological process. Published in Lloyd and Cheshire (2018).

In the first instance, quantification of interactions between each area of residence (origin) and store location (destination) was necessary. This was achieved by creating a trip

distribution (TD) matrix – a common technique applied in travel behaviour analysis that describes the frequency of trips recorded between each origin and destination. Typically, this represents the first stage in identifying and modelling interactions between places from a dataset, which can then used to predict future behaviours by applying gravity models.

There is a broad body of literature covering such retail spatial interaction models, which have been widely used both commercially and in academia (see Birkin et al., 2004, Clarke and Clarke, 2001; Birkin and Culf, 2001; Guy 1991) to forecast flows between origins and destinations (i.e. retail centres). The basic assumption of these models is that flows are a function of the attributes of origins and destinations (for example, their relative size or other measures of attractiveness), and the geographical proximity between them. Ultimately, these models are formulated to predict flows between places, and have been particularly popular because in many cases, detailed origin/destination data are not available or are incomplete (i.e. usually derived from small samples of survey data). Therefore, modelling flows between places have been valuable tools for informing transportation and land-use planning or retail analyses (see Birkin et al., 2004, for an overview).

Whilst spatial interaction models may conventionally be used for similar purposes, fitting a model to these data in this context was not considered an appropriate method for a number of reasons. Firstly, the aim of the analysis was not concerned with predicting or forecasting travel flows, but to quantify the retrospective flows that had occurred between areas of residence and HSR store locations over the two-year period. Secondly, the volume and longitudinal nature of these data provided a suitable basis for understanding flows between origins and destinations without the need for parameter calibration and model fitting. For example, the effects of distance and attractiveness between origins and destinations would be evident through the volumes of trips that had been captured.

Despite this, the principles of this area of research could be applied in this data-driven context. For example, firstly, identifying the volume of interactions between origins and destinations allowed analysis of what constituted the most likely travel patterns for a given area, based on observed events in the loyalty card data. Secondly, a method was then needed to categorise these patterns into the most/least likely locations of patronage. To achieve this, bespoke area ‘catchments’ were created by implementing percentage thresholds. This is a commonly utilised technique in retail centre catchment research, which typically involves the selection of one or more threshold values that represent the proportion of customers likely to patronise a certain store or retail centre (Dolega, Pavlis and Singleton, 2016). These categorisations often take the form of primary, secondary and tertiary catchments, of which definitions have varied across applications. Approaches adopted by some commercial consultancies define a primary catchment as the areal extent representing the flow of at least

50% of a particular centre's shoppers, the secondary retail catchment area typically between 25% and 50%, and the tertiary above 10% (i.e. Savills, 2005). Again, these methods are generally used to predict patronage extents from sampled data using gravity models (i.e. Huff, 1964) by utilising assumptions of distance, attractiveness and alternative retailer competition (Dramowicz, 2005). However, whilst the current method was concerned with description rather than prediction, these concepts could be adopted to create bespoke data-driven catchment areas.

Following these computations, an algorithm was designed that utilised this information to assess the frequency at which individual customers performed irregular travel patterns throughout their transactional histories. The specifics of these methodological stages are outlined in the next sections.

4.3.1.1. Data cleaning and pre-processing

A number of measures were taken to clean these data in preparation for analysis. Firstly, as noted in Chapter 3, transactional volumes varied substantially between customers. For the purpose of this analysis, active customers were defined as those that had transacted more than five times within the last financial year (April 2013-March 2014). This threshold was selected with the intention of eliminating inactive customers, whilst also retaining the maximum possible sample size. Secondly, records exhibiting missing or invalid postcodes (see Chapter 3, Section 3.2) were excluded from the analysis. These stages resulted in a sample of approximately 15.8 million customer accounts. Cleaning measures were also applied to customer metadata as these were utilised for interpreting characteristics post-analysis. Due to the metadata uncertainties identified (also see Section 3.2), customers were selected between the ages of 16–85, since this range captures the majority of the adult population. Those with withheld gender attributes were also removed. These stages removed 20.9% of the active customer database leaving a sample of approximately 12.5 million accounts with both sufficient volumes of transactional data and complete metadata attributes.

4.3.1.2. Deriving trip-distribution matrices

To create a TD matrix, customer origins were aggregated to the MSOA level. This aggregation was necessary in this context in order to produce large enough population groups to distinguish interpretable distributions. For example, some areas at the OA and LSOA levels contained few, if any, customers, which would be insufficient to summarise local patronage patterns. Alternatively, the MSOA level provided national coverage and a minimum of 289 customers per area, which was considered a suitable volume. On the other

hand, utilising larger scale units, such as LAs, reduced the sensitivity of the analyses to changes in store networks that were identifiable at the MSOA level.

The TD matrix was created by obtaining all MSOA to store journeys that had occurred within these data, resulting in 5,833,028 unique combinations. For each combination, the number of customers that had performed a journey (*T*) was obtained. This resulted in a matrix describing the frequency of customers that had performed each pair. Therefore, each unique trip a customer had performed was recorded. Table 4.1 shows an example of the matrix format. Subsequently, trip distributions were converted into trip proportions, by dividing *T* values by their *O* sum (total number of customers per MSOA). This was to interpret trips in relation to the differing volumes of customers per area.

Table 4.1: Example trip distribution matrix format. Published in Lloyd and Cheshire (2018).

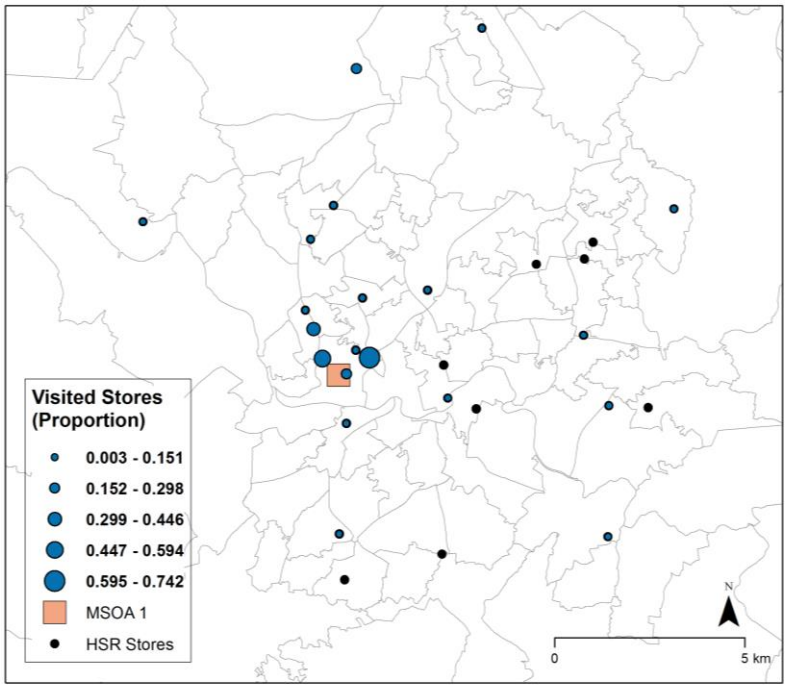
MSOA	**Store ID**					
	1	**2**	**3**	**4**	**5**	**Sum**
E02000001	T_1	T_2	T_3	T_4	T_5	O_1
E02000002	T_6	T_7	T_8	T_9	T_{10}	O_2
E02000003	T_{11}	T_{12}	T_{13}	T_{14}	T_{15}	O_3
E02000004	T_{16}	T_{17}	T_{18}	T_{19}	T_{20}	O_4
E02000005	T_{21}	T_{22}	T_{23}	T_{24}	T_{25}	O_5
Sum	D_1	D_2	D_3	D_4	D_5	

The TD proportion matrix allowed interpretation of the relative frequency with which the pairs were performed per MSOA. On average, 688 unique stores were visited per MSOA, with a minimum of 111 and a maximum of 1248. Individual customers visited an average of 11 different stores over the 2-year period. Figure 4.4 demonstrates the local trip distributions of 3 MSOA's within close proximity, in the area of Bristol, South West England. Data describing the behaviours of less than 10 people were removed from these visualisations for disclosure purposes.

Patterns illustrated that, unsurprisingly, there were many overlaps of likely store destinations between MSOAs within close proximity (for instance, a city centre flagship store that drew patronage from a large distribution of areas). However, importantly, it was possible to discern unique patterns of patronage for each area. In addition, whilst the full distribution of destinations per area was large (i.e. on a national scale), there were only a small number of destinations that received a high proportion of patronage, which were typically within close proximity to the origin MSOA (flows were evident to destinations far in proximity, but with

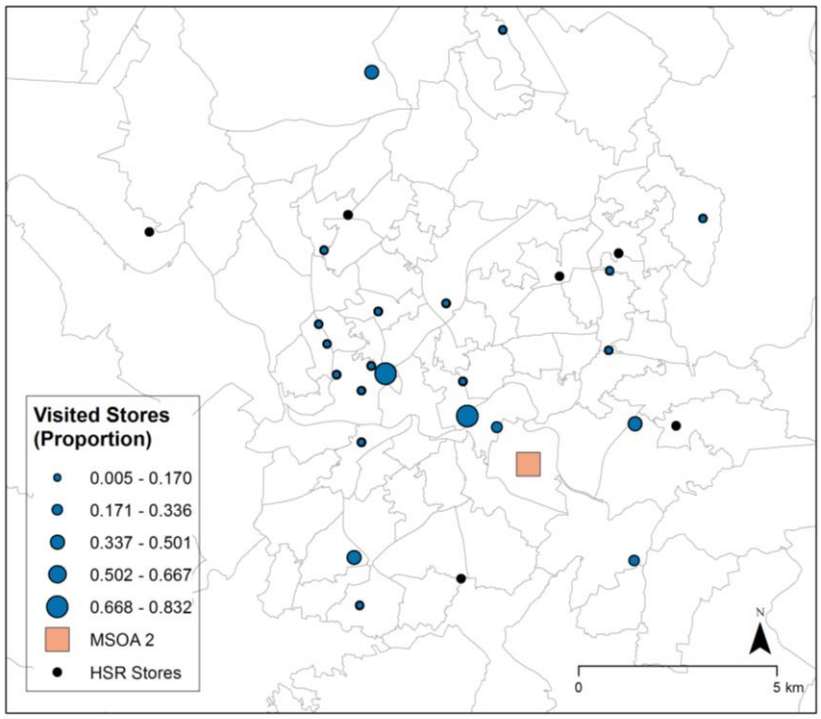
low proportional values). Trends also indicated that distance did not explain all observed patterns, for example, high proportions were also evident for town centre and retail park destinations, despite being further in proximity to some surrounding destinations. These observations are also consistent with established travel and catchment estimation knowledge, where factors such as trip chaining (i.e. workplace locations) or attractiveness of a destination (i.e. city centre or retail park versus local store) may prevail the effects of distance.

The TD information could also be utilised to examine variations in flow dynamics for different store location types, by aggregating flows by store, rather than MSOA. Figure 4.5 shows the proportion of flows to surrounding MSOA's for convenience high streets, large high streets, retail parks and community stores. TD's were normalised by the total trips per store, to account for differing volumes between store types.



a)

b)



c)

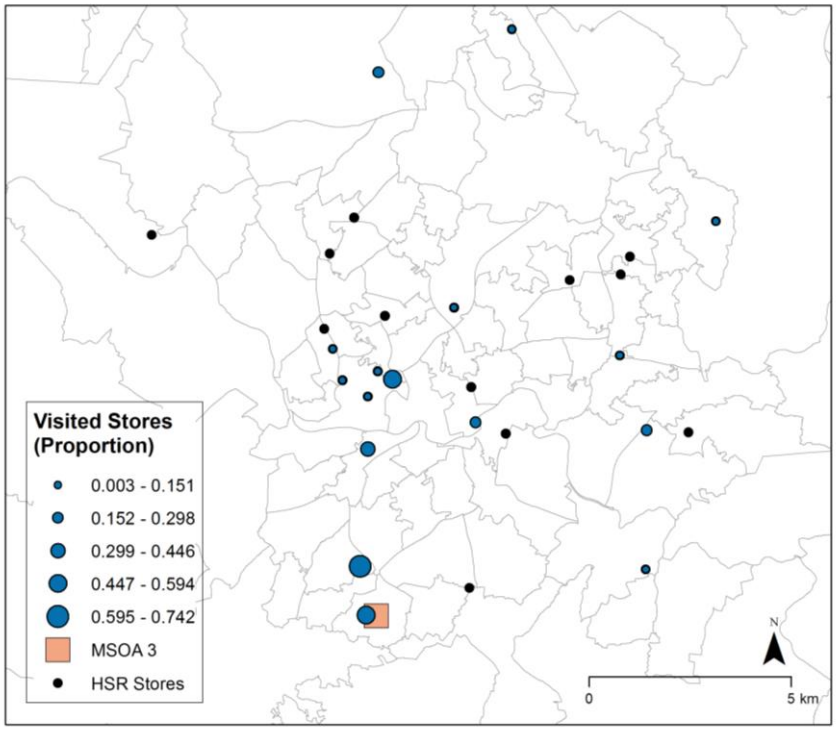
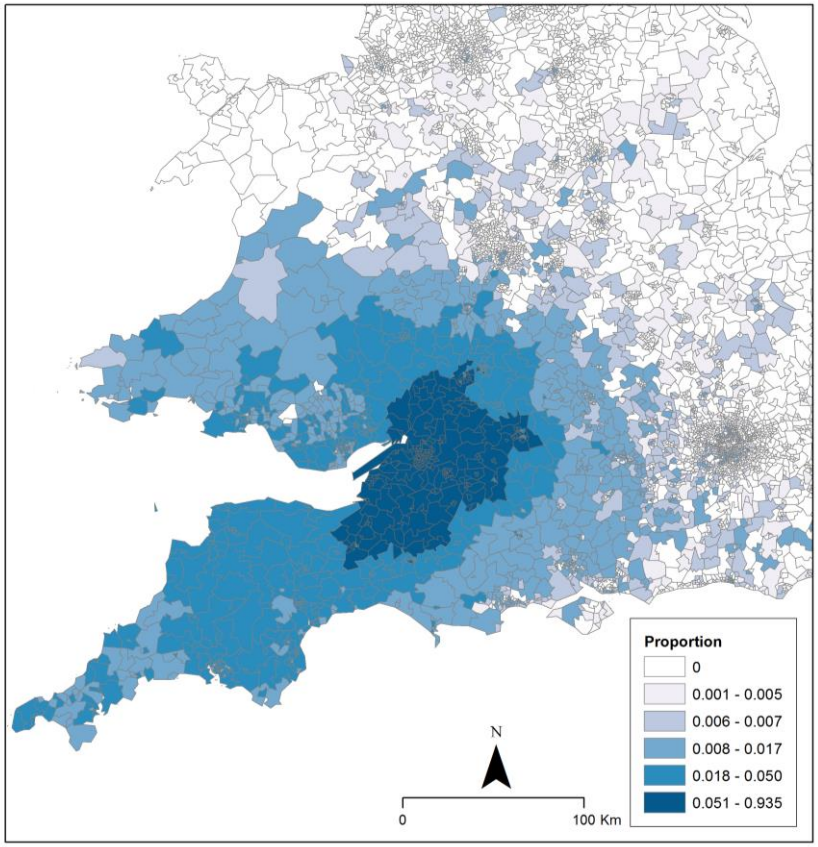


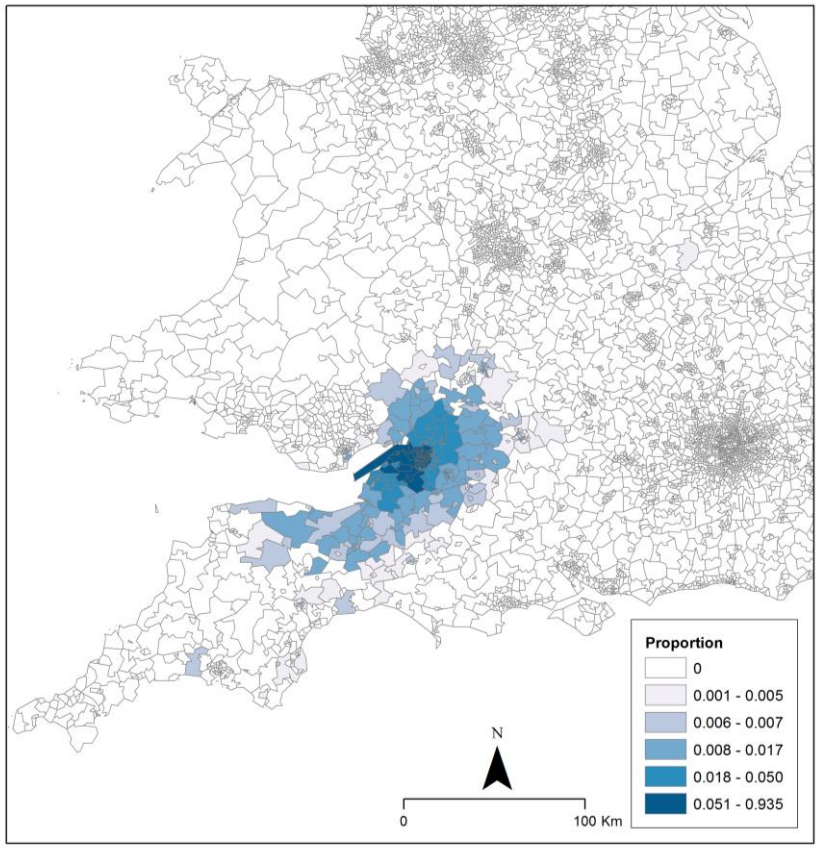
Figure 4.4: Local store destination distributions for MSOA's a) E02003043, b) E02003049 and c) E02003064

(normalised by total customers per MSOA, classified by equal intervals).

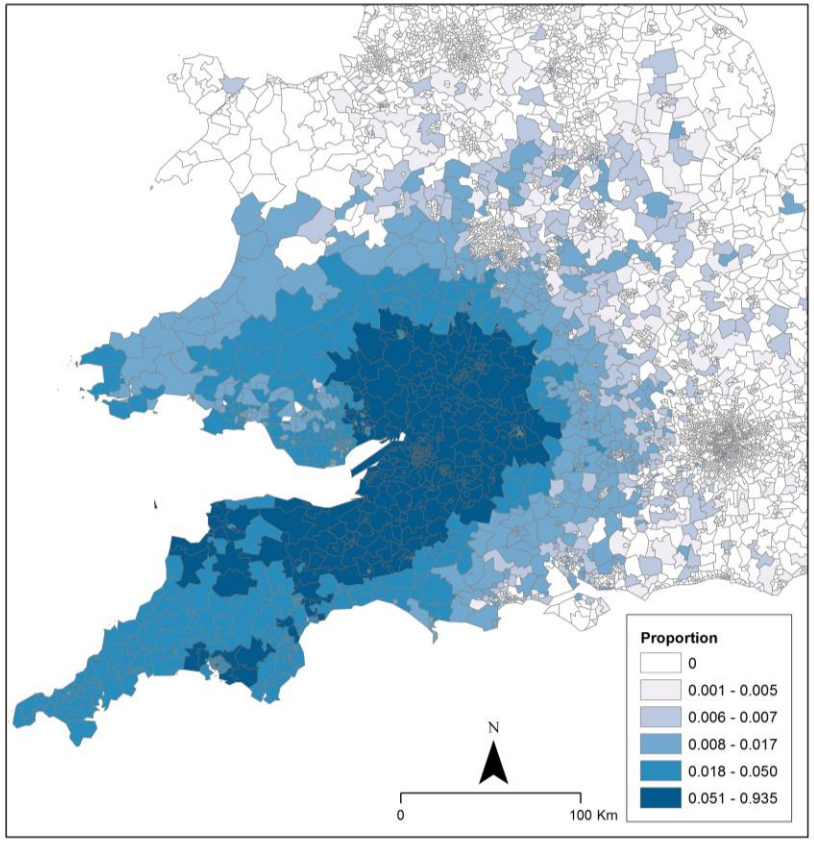
a) Destination (Large High Street)



b) Convenience (High Street)



c) Destination (EOT)



d) Community

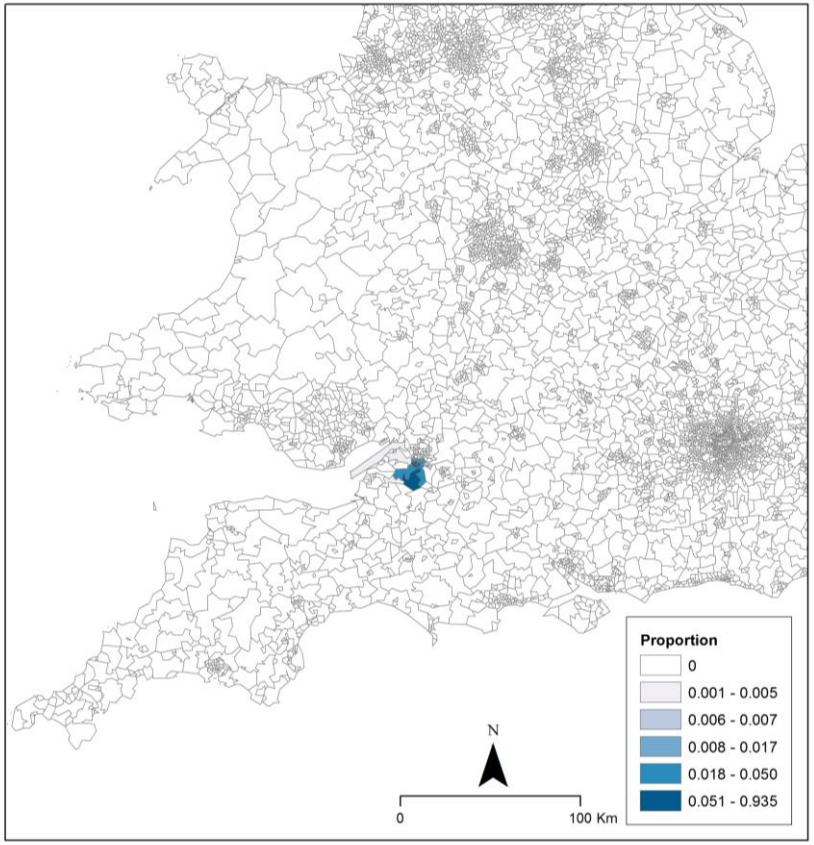
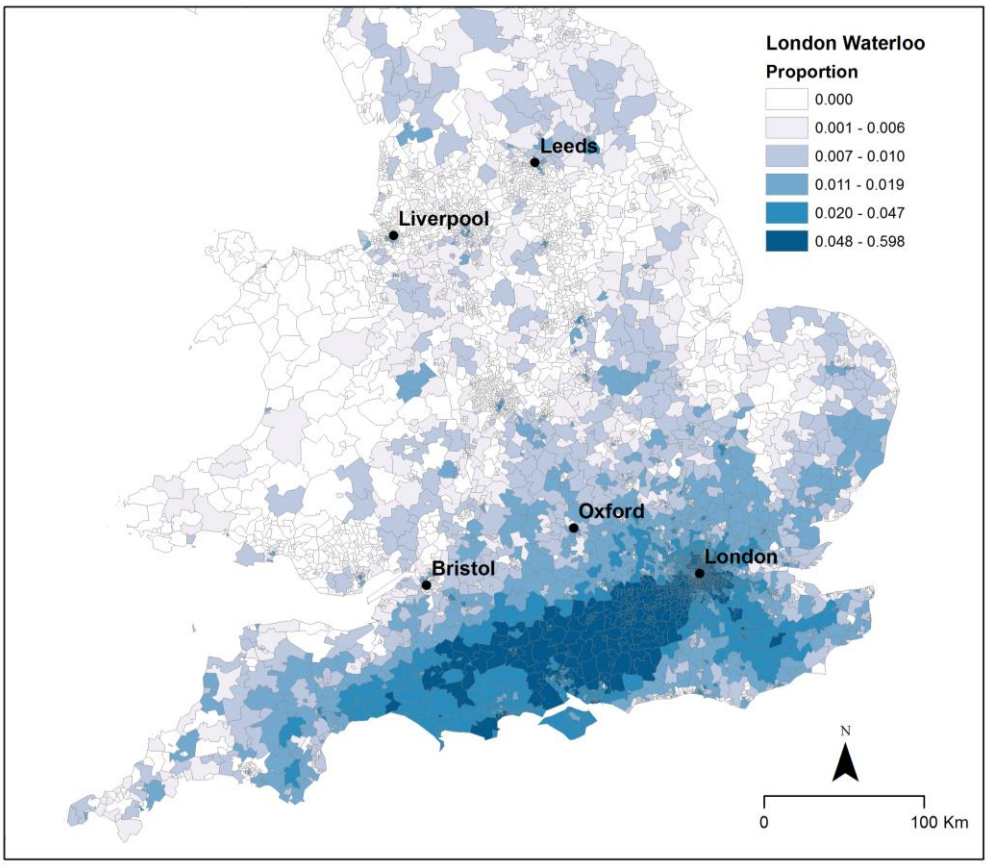
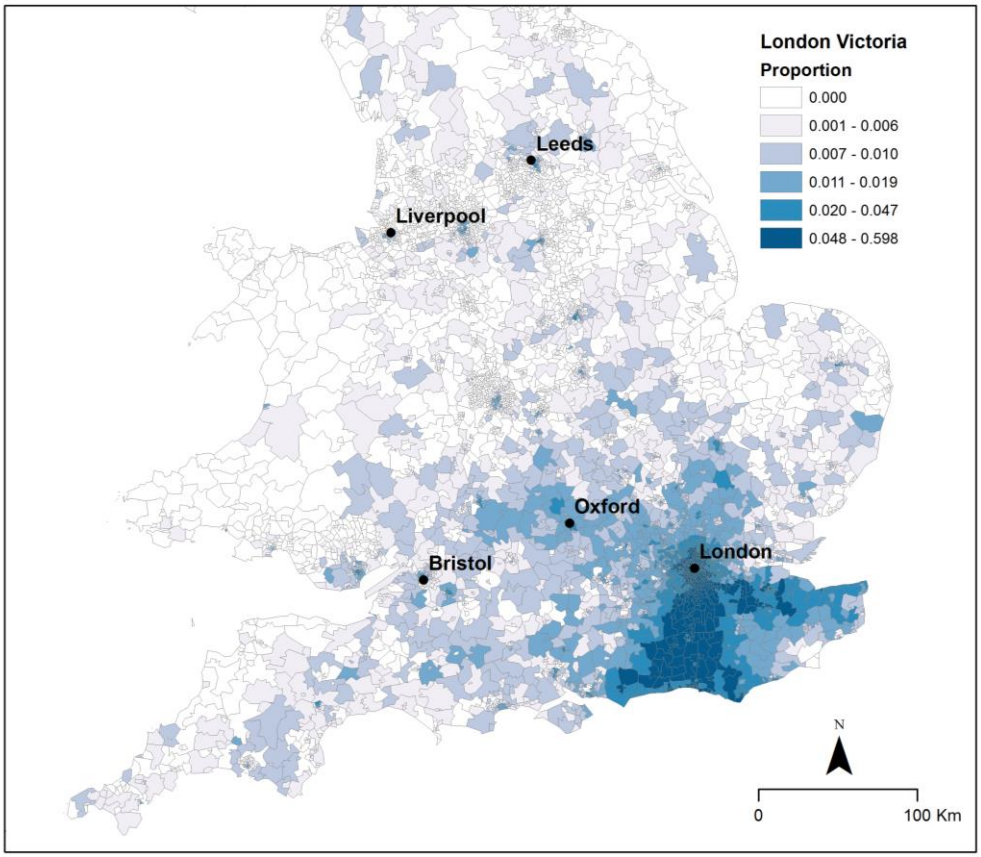


Figure 4.5: Trip distribution proportions to surrounding MSOA's from a) 'Destination (Large High Street)', b) 'Convenience (High Street)', c) 'Destination (EOT)' and d) 'Community' store.

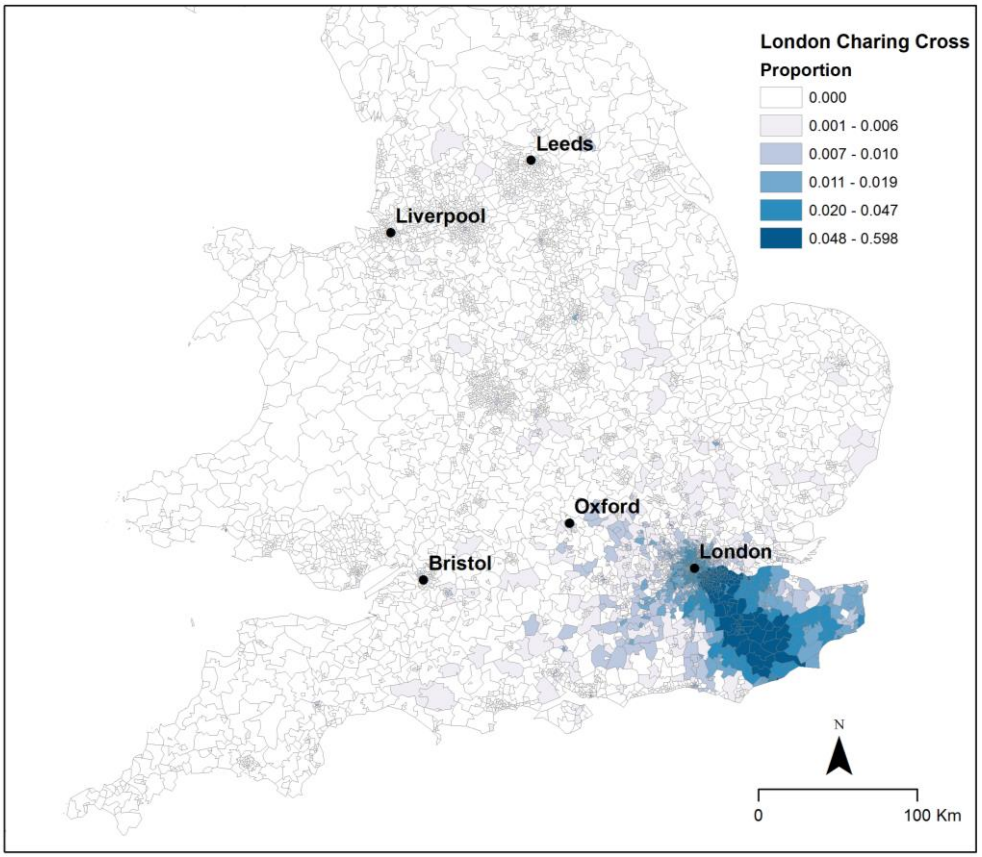
As may be expected, urban locations such as convenience and large high street destinations exhibited much wider distributions of patronage compared to community serving stores. Retail park flows showed moderately wide distributions, and patterns were skewed towards the nearest surrounding urban area/town. In the case of Bristol, this could be due to multiple retail park stores being situated around the outskirts of the city, each attracting their most proximal customers. It is also likely that transport accessibility of a location influences the resulting patronage flows. For example, Figure 4.6 illustrates distributions for stores located near major transport hubs in Central London. Clearly, the highest proportions of flows can be delineated by origin MSOAs with easy access to these railway lines.



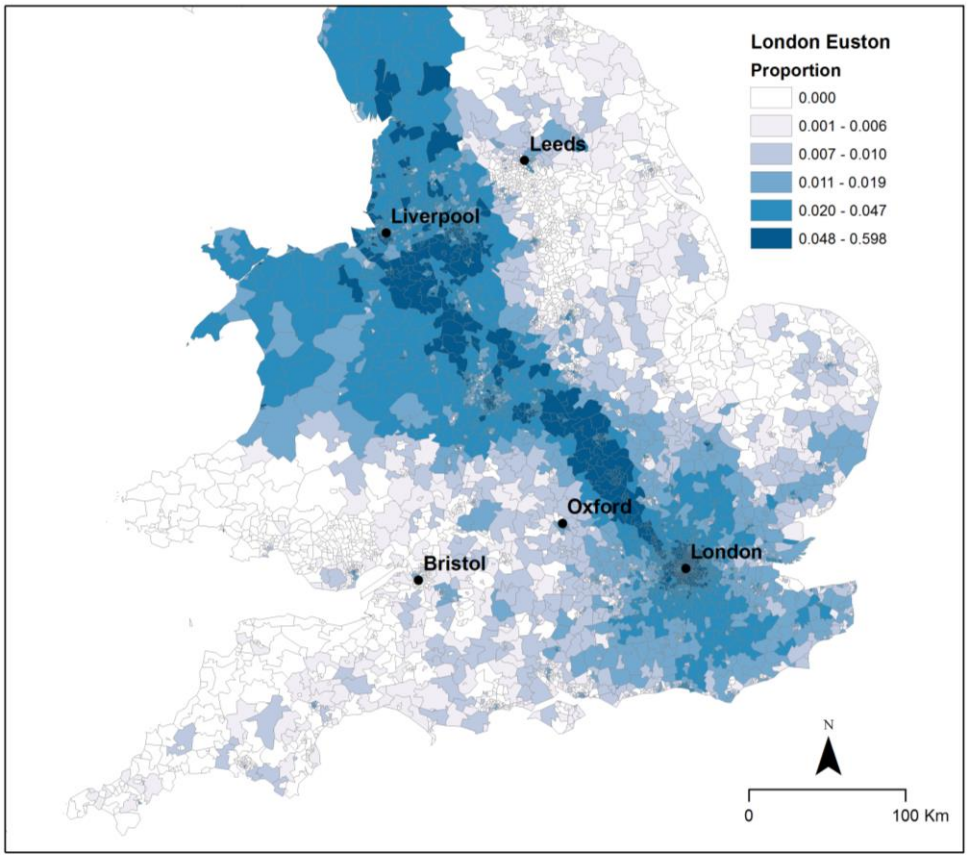
a)


London Victoria
Proportion
0.000
0.001 - 0.006
0.007 - 0.010
0.011 - 0.019
0.020 - 0.047
0.048 - 0.598
Leeds
Liverpool
Oxford
Bristol
London
N
0
100 Km

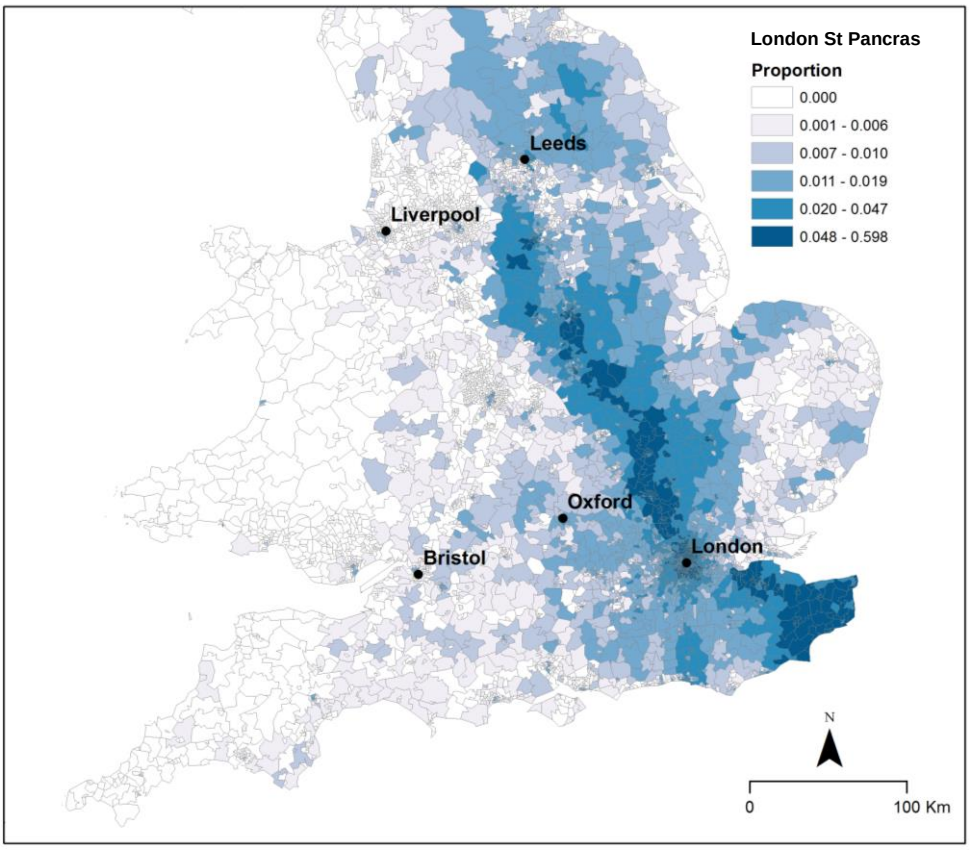
b)


London Charing Cross
Proportion
0.000
0.001 - 0.006
0.007 - 0.010
0.011 - 0.019
0.020 - 0.047
0.048 - 0.598
Leeds
Liverpool
Oxford
Bristol
London
N
0
100 Km

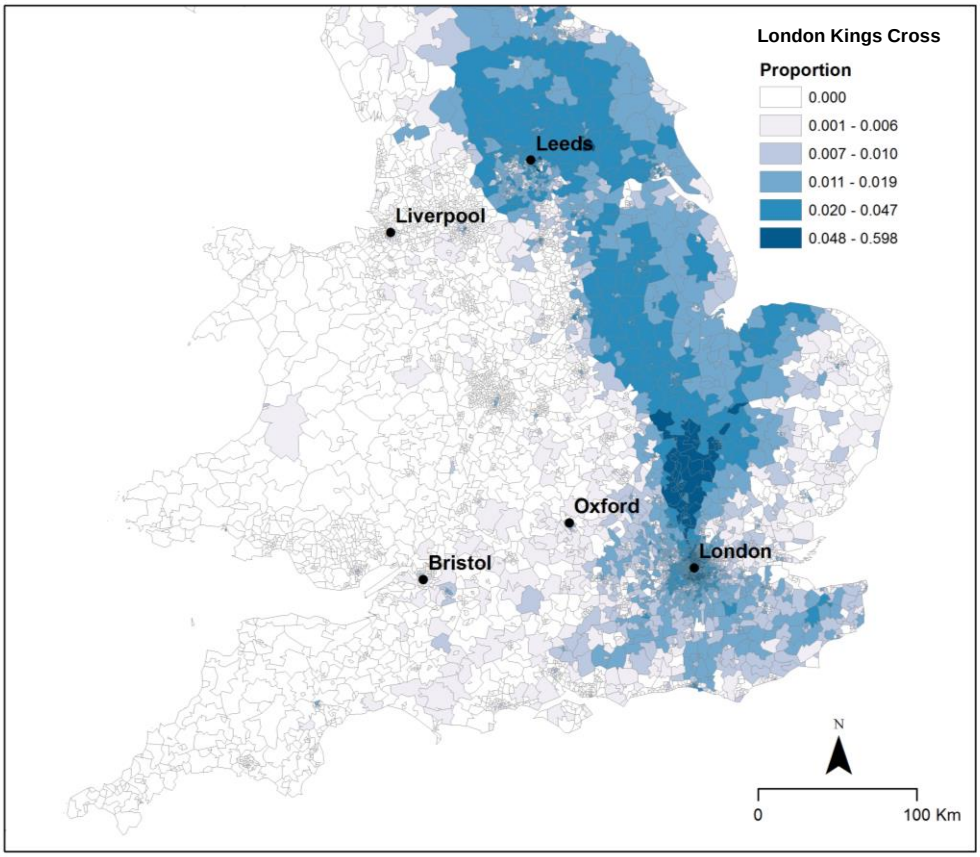
c)


London Euston
Proportion
0.000
0.001 - 0.006
0.007 - 0.010
0.011 - 0.019
0.020 - 0.047
0.048 - 0.598
Leeds
Liverpool
Oxford
London
Bristol
N
0
100 Km

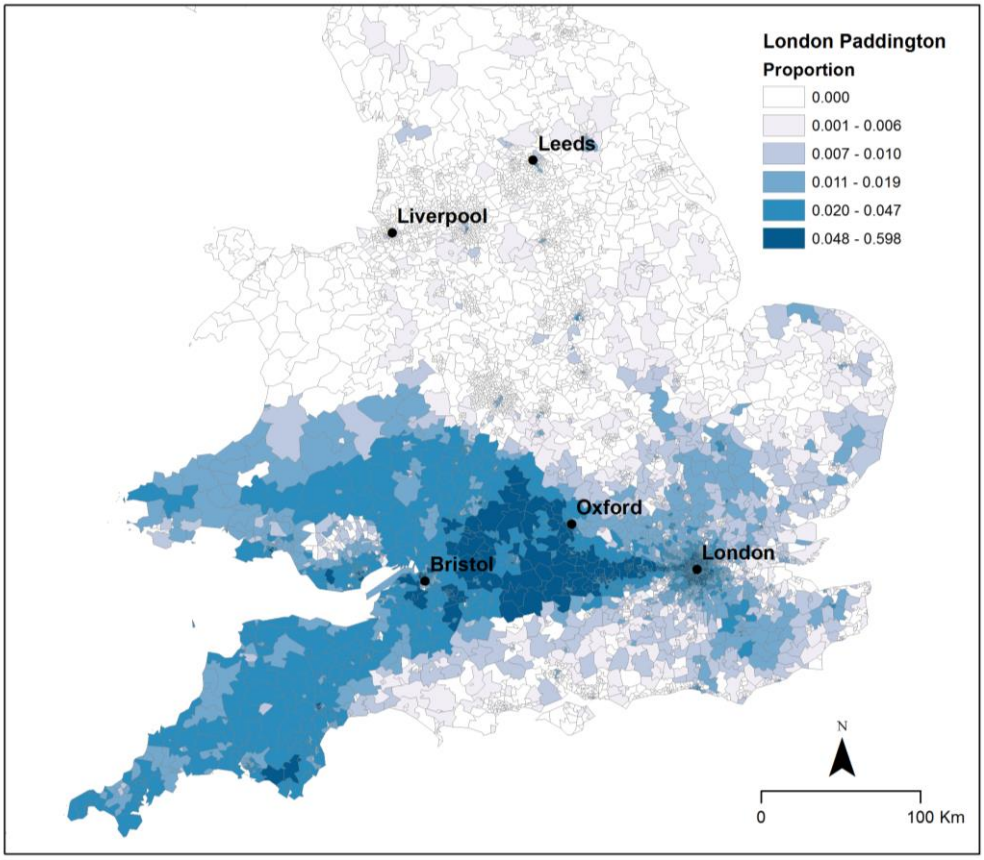
d)


London St Pancras
Proportion
0.000
0.001 - 0.006
0.007 - 0.010
0.011 - 0.019
0.020 - 0.047
0.048 - 0.598
Leeds
Liverpool
Oxford
London
Bristol
N
0
100 Km

e)



f)



g)

Figure 4.6. Trip distribution proportions to surrounding MSOA's for stores located near transport hubs in Central London (quantile breaks).

From these observations, it was concluded that analysis of TD's would allow sufficiently accurate delineation of the complex dynamics of customer flows. This also suggested that a large amount of customer addresses in the data were likely valid. For example, flows followed broadly expected trends of population movement, based on existing knowledge of travel behavior and distance decay dynamics. However, in order to isolate the uncertain cases identified in the exploratory analysis, this TD information was then utilised to classify irregular behaviour based on a customer's area of residence.

4.3.1.3. Threshold selection and distance constraints

Thresholds were defined to categorise TD's and select destinations that fell above or below these. This aimed to identify the point at which stores no longer constituted regularly patronised destinations for an MSOA. Stores above these thresholds were defined as 'primary' destinations (e.g. the highest 40% of visited locations) and those below as 'non-primary' destinations. Thresholds were calculated individually for each MSOA to reflect the unique dynamics of each.

Trip distribution tails for all MSOAs were positively skewed, a feature that can be explained by the behavioural dynamics of the data in context. For example, relatively few destinations are highly patronised by customers of a given area, largely due to effects of proximity. Figure 4.7 illustrates an example of trip distributions and (Euclidean) distances travelled, calculated from MSOA centroids to store locations. There is no consistent application of threshold values in retail catchment literature, therefore values were defined largely based on performing test trials and observing what values best described the trends in this dataset. Yet, due to distance decay dynamics, a plateau in patronage could be observed following the most highly patronised group of stores. This was reflected in the data by a large increase in variance between consecutive intervals in each MSOA's trip distribution tail. Therefore, it was deemed a practical solution to use this dynamic as a means to select threshold values.

Preliminary exploration was carried out to discern the efficiency of selecting thresholds in this way. This showed that utilising this natural trend in the data (which reflects the pervasive fact that trip distribution volumes will decline as distance increases) was able to sufficiently delineate what may constitute the most likely destinations for each area. Figure 4.8 illustrates this dynamic, showing the distributions tails and the threshold points for the 3 exemplar MSOAs illustrated in Figure 4.4. Thresholds were selected for each MSOA using an algorithm written in R, which essentially identified variance between distribution intervals and selected values based on the largest variance between consecutive intervals, proceeding the initial most patronised group of stores.

Across MSOAs, thresholds ranged between 31% and 55% with an average of 41%. Trips falling outside of these thresholds primarily described less patronised destinations, such as those far in

proximity. On average, there were 37 primary stores per MSOA, a minimum of 2 and a maximum of 64.

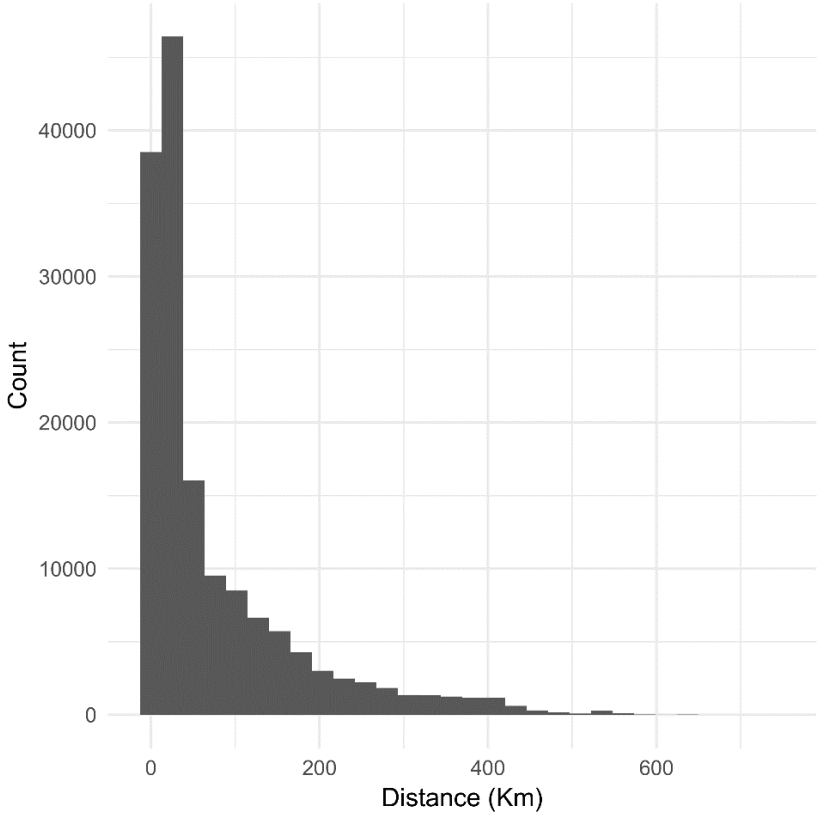
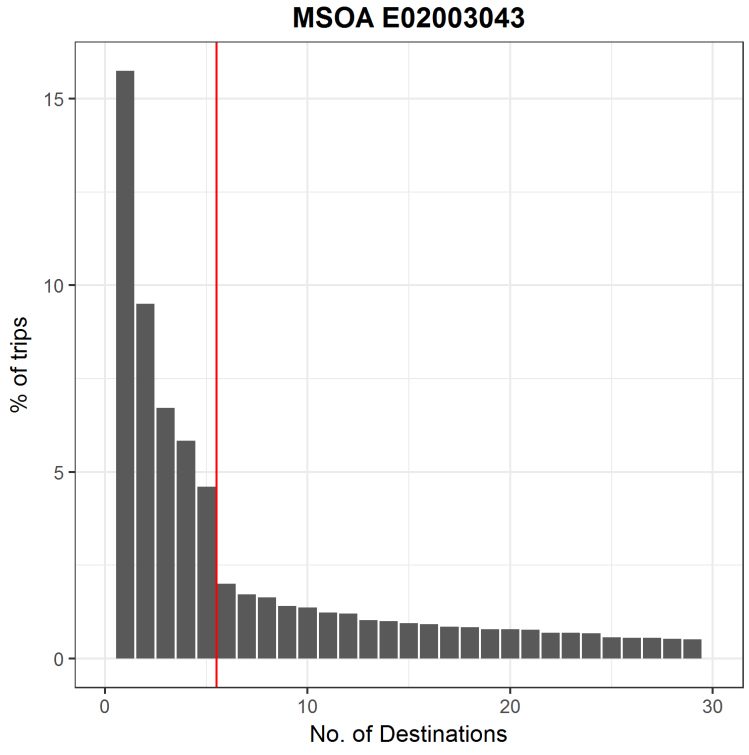
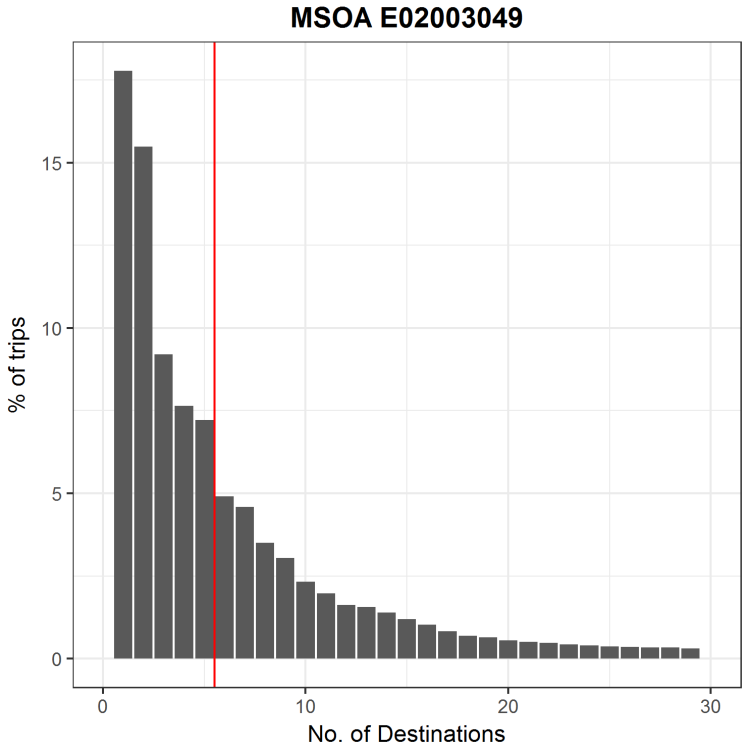


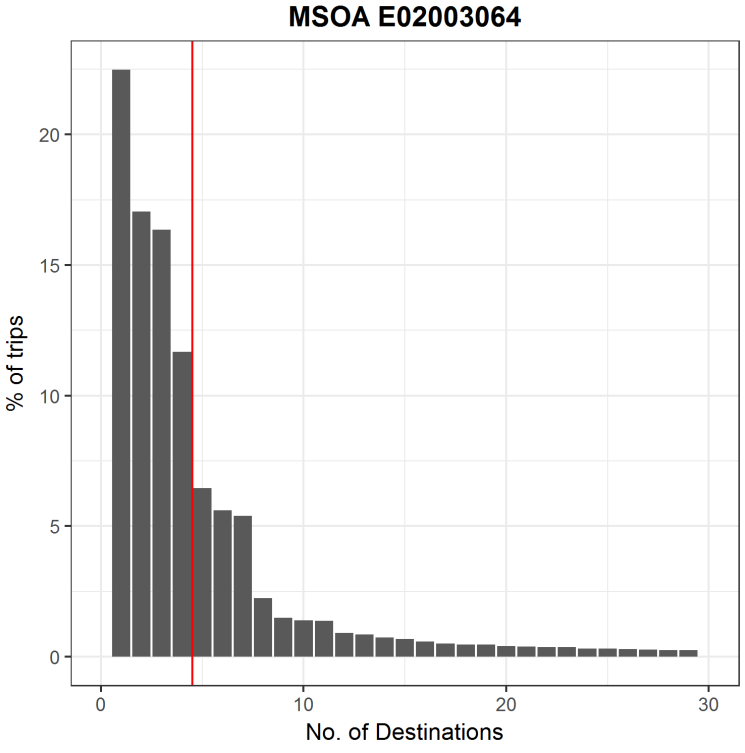
Figure 4.7: Example trip distribution tail and distance travelled. Published in Lloyd and Cheshire (2018).



a)



b)



c)

Figure 4.8: Example trip distribution tails (percentage of trips by number of destinations) and threshold points (in red) for MSOAs a) E02003043, b) E02003049 and c) E02003064.

A final methodological stage was the introduction of distance constraints as a result of issues identified in the data. For example, distances were a key indicator of whether or not behaviours could be deemed 'logically consistent' with an area. Yet, in some instances, destinations within relatively close proximity to a residential area were categorised as 'non-primary' and those very far as 'primary'. Exploratory analysis suggested that this was likely due to two factors; firstly, the limited time span of the data producing small counts for fairly local stores (therefore a reflection of lack of data rather than true customer behaviour) and secondly, destinations that exhibited high counts across many areas, but were unlikely to be a routinely patronised store (such as Oxford Street in central London).

To create a suitable constraint, the average Euclidean distance travelled from MSOA centroids to primary stores was calculated for each area. It is acknowledged that alternative (i.e. network) measures would provide more accurate portrayals of travel distances, however, the aim of including this constraint was to isolate cases that were significantly above or below normative behaviour in relation to overall Euclidean distances per area (i.e. given that this may vary considerably between rural and urban areas), rather than to quantify precise travel behaviours. These irregular instances were minimal, however customers who were identified in the succeeding analysis based on these re-categorised stores were flagged in the output for further investigation.

4.3.1.4. Implementation

The resulting output from this process was a list of primary stores per MSOA. The final methodological stage was then to design an algorithm that could implement this information and flag irregular patterns of behaviour in the database. Similarly to previous stages, identification of the 'optimum' method required exploratory analysis and trial-and-error of different techniques. To understand how an algorithm may achieve this, abductive reasoning was applied to investigate variations in TD behaviour on an individual level. This suggested that, generally, 'normal' behaviour took the form of consistent primary trips throughout transactional histories and non-primary trips on an intermittent basis. However, two fundamental patterns of potential uncertainty could be identified. Firstly, (for the purpose of this analysis), *address errors* were defined as customers who had never transacted at a primary store location. Secondly, *address changes* were defined as customers who demonstrated a change in patronage behaviour within the time span of the data. These could typically be identified as a permanent shift to a unique network of stores that was outside of their registered area's primary destinations. Figure 4.9 gives an overview of the algorithm designed to detect these cases.

Firstly, account numbers were individually selected alongside their time-ordered transactional histories. Catchment thresholds were then obtained for a customer's MSOA and their

transactions categorised. If an *address error* was not identified at this stage, accounts were assessed for an *address change* by analysis of their time-ordered transactions to detect a change in store network – defined as no further occurrences of primary store transactions recorded after certain timestamps in their transactional history. If a change point was detected and patterns exhibited a new network of stores, account numbers were appended as an address change and a timestamp (as per their last primary transaction) of this change recorded. However, if at least one store had been previously visited, the account was appended as a *change risk*. It was speculated that *change risk* instances may either indicate a customer who had not changed location (but had not visited a primary store in a substantial amount of time), yet, could alternatively indicate a location change that was close enough in proximity to warrant continued patronage of certain stores (for example, a city centre flagship store). Only analysis of more recent transactional data would be able to clarify these cases, however, these instances were still recorded as demonstrating abnormal transactional behaviour.

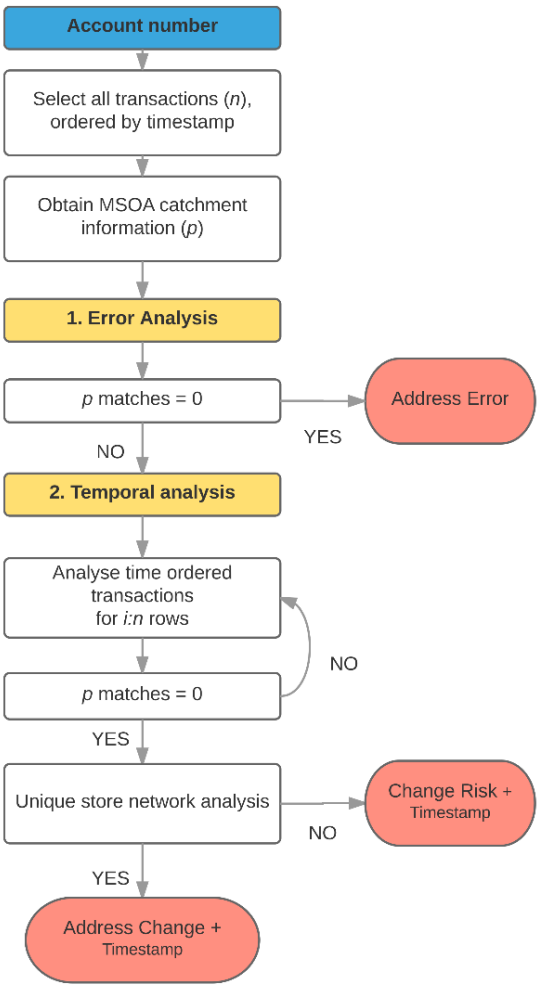


Figure 4.9: Overview of the algorithm process for detecting uncertain address information. Published in Lloyd and Cheshire (2018).

Finally, a constraint was also necessary to avoid returning customers who may have exhibited deviating behaviours based only on their most recent transactions (i.e. due to the data only capturing a two-year period, this would not provide enough information to classify a location change). All identified cases were therefore subject to a time constraint of 1 month to ascertain if the person had spent sufficient time transacting in the new area to be defined as a permanent change. This time period was chosen based on a trade-off between accounting for the relatively high average transaction interval of the active customer sample (12 days, yet this was positively skewed) and avoiding omitting identification of those with more frequent transactions. This algorithm was implemented in R, using the *RPostgreSQL* package (Conway et al., 2008) to obtain the relevant data stored within the database.

4.3.1.5. Contextualising outputs

To contextualise the characteristics of flagged customers, comparative analyses were conducted with a number of census outputs. Due to availability of these statistics, only England and Wales data were utilised for these comparisons. In the first instance, accounts were compared across 2011 OAC groups. This classification, derived from census variables, describes geodemographic characteristics (i.e. derived from demographic and socioeconomic variables) across 8 Supergroups, 26 groups and 76 subgroups at the OA level. Customer postcodes were thus aggregated to the OA level and frequency of estimated moves compared across the groups. Counts were normalised by total customers per group in the database, to account for underlying variation in volumes.

To contextualize the estimated moves in terms of migration patterns, results were also compared to census migration statistics, which describe moves that occurred between MSOAs in England and Wales between 2010 and 2011. Events captured using the loyalty card data were selected between 2013 and 2014, as this were the only available full year of data comparable to census dates. However, at this stage, the aim was to establish whether the moves identified within the loyalty card data followed broadly expected flows, rather than use them to make broader inferences across the population. Card estimates were adjusted to reflect census population volumes per MSOA by creating a coefficient (total census population per MSOA divided by total card population per MSOA). Card migration counts were then multiplied by this value. Relationships between migration estimates were measured using Spearman's rank correlation.

4.3.2. Results

Implementation of this algorithm returned a total of 447,141 accounts – approximately 3.6% of the analysis sample. This comprised of 213,395 estimated *address errors* and 233,748 *address changes*. Whilst it is unlikely that the largest proportion of customers provided incorrect address information at sign up, it is possible that an *address change* occurred before the time period of

the available data. In addition, a large proportion of the *address change* customers could only be categorised as *change risk* due to lack of available transactional data (45% of risk customers exhibited less than 10 transactions in comparison to 6% of those conclusively categorised), resulting in final analysis samples of 213,395 address errors and 169,943 address changes. Analysis of spend characteristics suggested that these customers had not deserted the card scheme, with an average spend of £344, 37 transactions and 64 products over the 2.5 financial years. Figure 4.10 illustrates the travel flows presented in Figure 4.1 (using the raw data), in comparison flows after removing accounts flagged in the analysis.

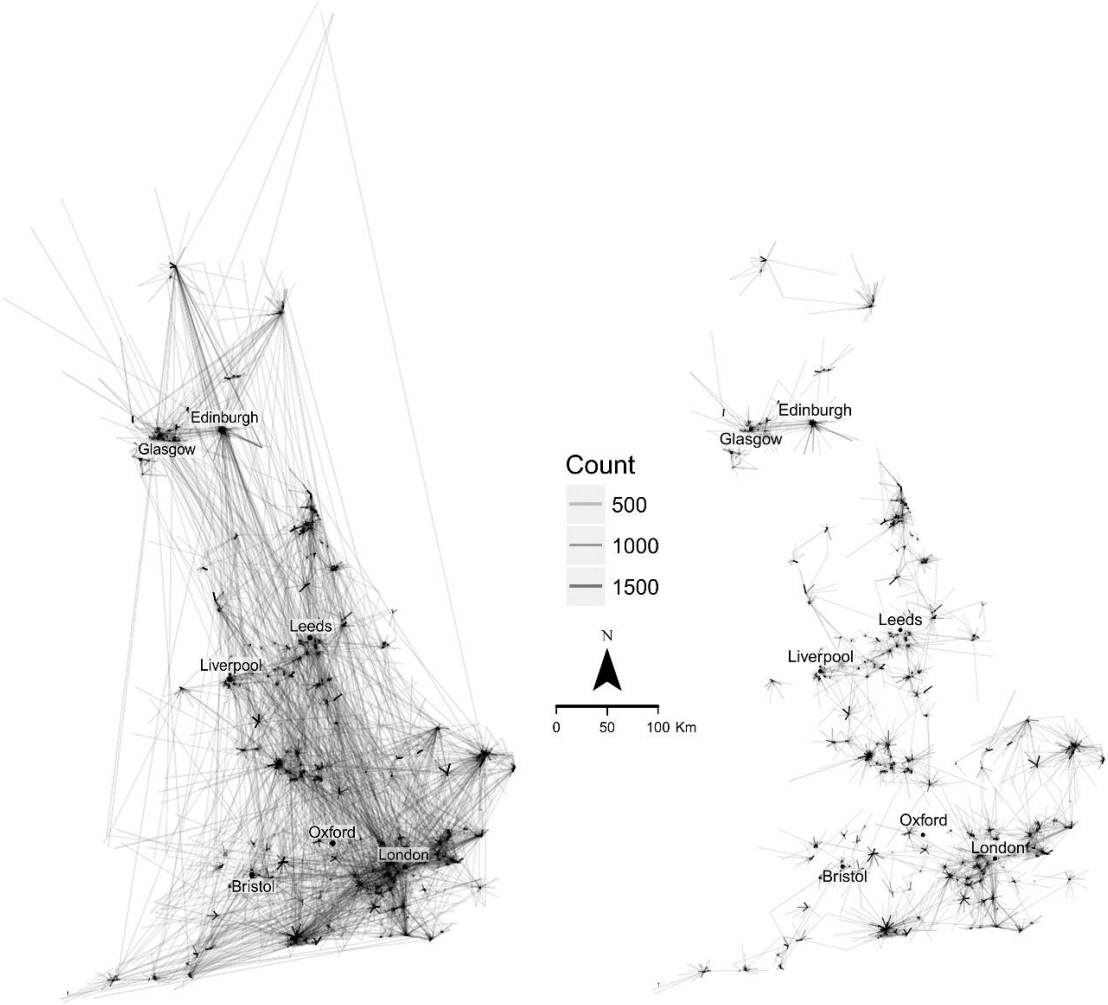


Figure 4.10: Raw versus cleaned flows from customers' origin MSOA to their most frequently visited store, for 'Community' type stores (showing sample sizes of 65,770 customers before cleaning, 53,141 remaining after). Published in Lloyd and Cheshire (2018).

Applying this cleaning method produced flows that were consistent with expectations for this store type, which primarily serves local surrounding communities. In comparison to the raw data, it was speculated that the majority of patterns that were inconsistent with our existing knowledge of spatial behaviour could be identified. Analysis of the demographic attributes of

these customers was possible using the metadata provided at sign up. Figure 4.11 shows the age distributions of customers recorded at the time of their change point, normalised by total customers per age group. This suggested that customers flagged as exhibiting a location change were considerably skewed towards younger cohorts, particularly between the ages of 18-20. This could be indicative of a more transient group with a greater risk of failing to update their address information.

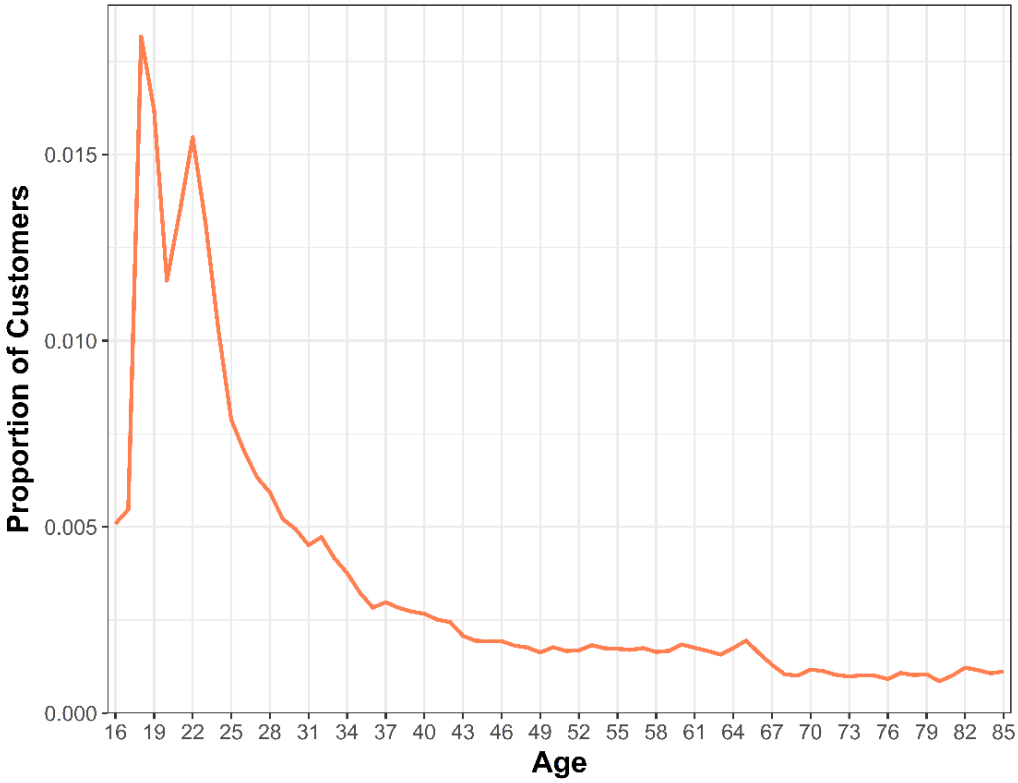


Figure 4.11: Ages recorded at time of estimated change point, normalised by total customers per year of age. Published in Lloyd and Cheshire (2018).

Furthermore, Figure 4.12 shows the comparison of these customers to the OAC at the Supergroup and Group levels (counts were normalised by the total customers per area group in the database). This suggested that the largest proportion of the flagged customers were likely registered to cosmopolitan areas and in particular, primarily student populated neighbourhoods. Analysis of the Subgroups indicated that the highest proportions were registered to: 'Student Digs', 'Student Communal Living', 'Students and Commuters' and 'Multi-cultural Student Neighbourhood' Groups. Higher proportions amongst other subgroups indicated those also less likely to have a long-term stable location, such as 'Young Families and Students' (Ethnicity Central Supergroup) and 'Private Renting New Arrivals' (Multi-cultural Metropolitans Supergroup). Supergroups with the lowest proportions of flagged customers included 'Suburbanites' and 'Hard-pressed Living'. These trends suggested that the method was able to

highlight areas that we expect to have more transient residents, whilst also providing insight into customer segments who may be more likely to exhibit inaccurate address attributes.

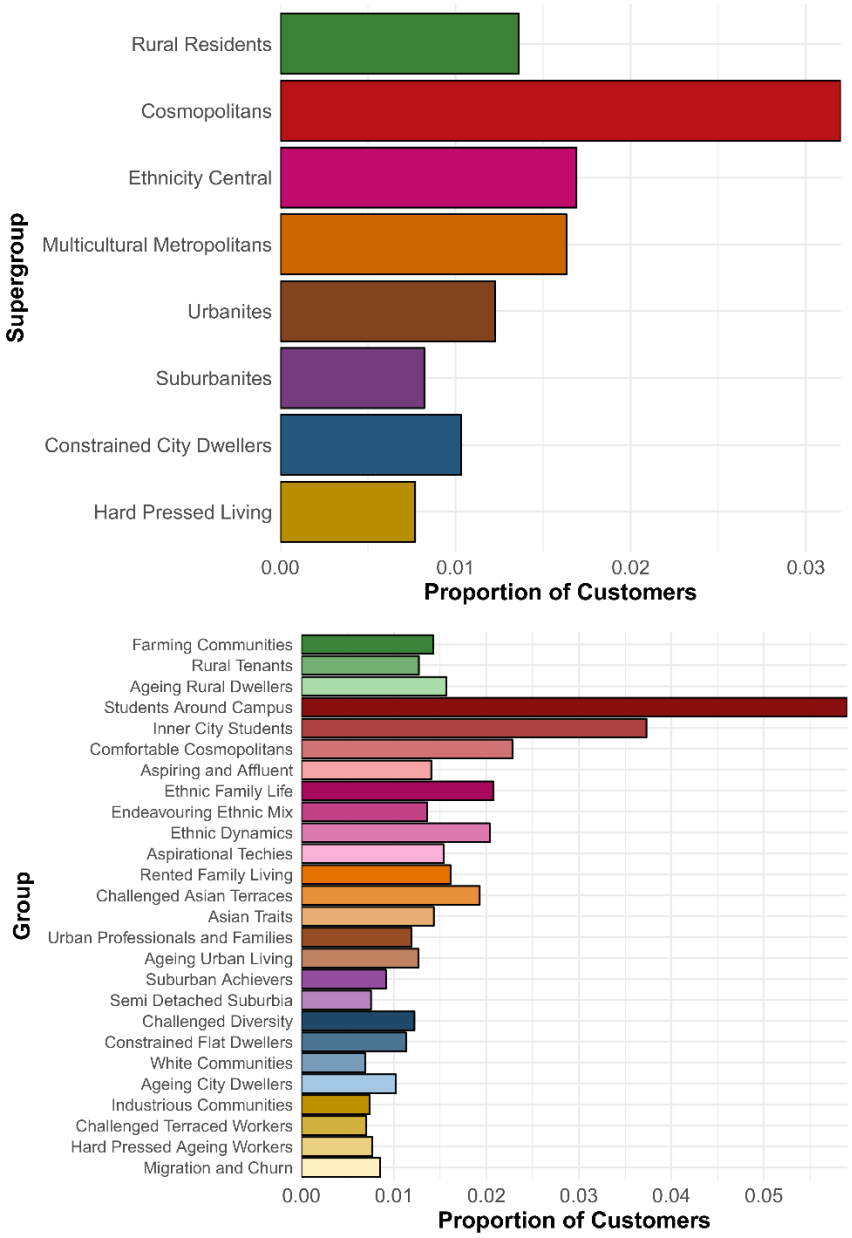


Figure 4.12: Migration counts across OAC a) Supergroups and, b) Groups. Published in Lloyd and Cheshire (2018).

Finally, correlation of migration events between the datasets showed a moderately strong positive relationship (rho = 0.53, $p < 2.2e\text{-}16$). However, in light of the previous observations, it was likely that the card migration was skewed by the amount of student migration captured.

Correlation with census student migration estimates at the local authority level indicated a strong positive relationship of 0.87, $p < 2.2\text{-}e6$ (see Figure 4.13).

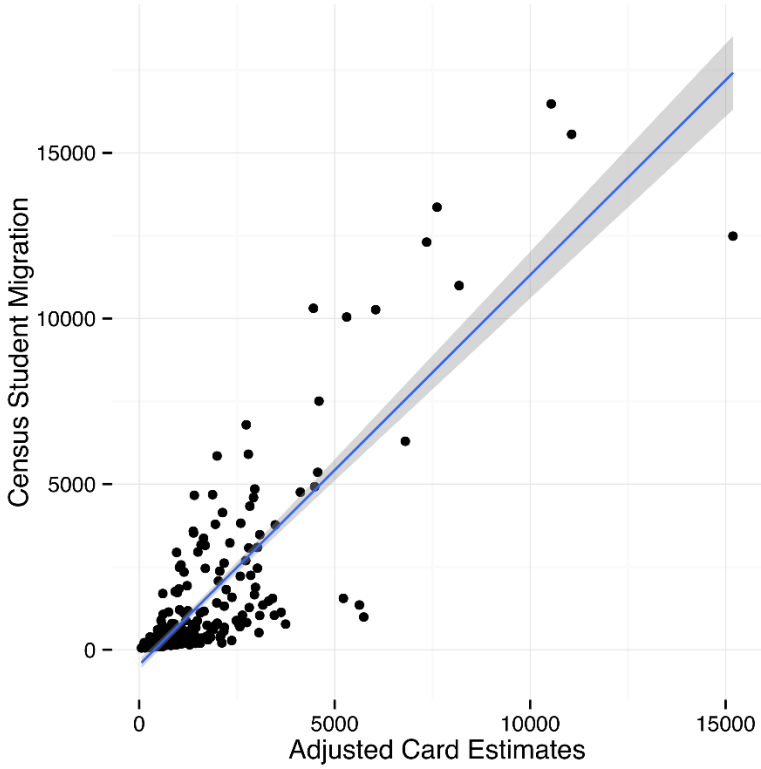


Figure 4.13: Relationship between card migration and census student migration estimates. Published in Lloyd and Cheshire (2018).

We speculate that the ability of the method to flag high levels of student migration could be due to the highly transient residential nature of this demographic group. However, it is also possible that the method is better equipped at flagging long distance moves, as these cases will fundamentally exhibit a more discernible change in store network. It follows that an unavoidable limitation is that this method is not able to detect location changes that do not cause a modification in store visiting behaviours. This may limit the extent to which we are able to detect close proximity moves, and will also be constrained by the geography of the store locations, as these are the only spatial point of reference for observing irregular behaviours. This may affect the method where store networks are less dense, for example, more rural areas.

4.4. Estimating Relocations

Having quantified store-visiting behaviour at a small area level, it was possible to extend these analyses to estimate potential areas of relocation (i.e. their new area of residence) for flagged customers. For example, providing that enough transactional data were available, the localities that their new store visiting behaviours were consistent with could be identified. The proceeding

sections of this chapter therefore present a brief extension to the customer address analysis, which aimed to understand the extent to which we could extract further insight from these outputs. Results were then contextualised using census migration statistics.

4.4.1. Method

In order to estimate relocations, a simple pattern-matching algorithm was designed to match new customer store networks to the primary store networks of different small areas. This was conducted at both the MSOA and also LA level to assess the granularity at which such analyses could be conducted. The algorithm was implemented in R, which firstly selected transactional histories for flagged customers, using all transactions for *address error* customers and transactions succeeding a change point for *address change* customers. Following this, stores visited by these individuals were matched with primary stores identified per MSOA or LA. Primary stores per LA were computed by aggregating the MSOA level information created from the computations in Section 4.3. Per LA, there was an average of 83 primary stores and a maximum of 262. Outputs included the number of total area matches and the number of primary stores that were matched from each area.

4.4.1.1. Contextualising outputs

Results were firstly contextualised with existing migration statistics captured by the 2011 Census Origin-Destination data (describing the origins and destinations of moves that occurred in England and Wales between 2010 and 2011). As per the previous analysis, card estimates were adjusted to reflect census population volumes (see Section 4.3.1.5). These analyses similarly aimed to investigate whether the moves identified within the loyalty card data followed broadly expected patterns of migration, rather than use them to make inferences across the population. Secondly, data were appended to census derived LAC Groups to investigate potential social mobility characteristics (i.e. movement between area types). This was conducted using both the card and census data to compare characteristics observed within the general population. The LAC was created from census variables (including demographic structure, housing, socioeconomic characteristics and employment) and summarises the characteristics of LA's across 8 broad Supergroups, 15 groups and 29 subgroups (such as 'Business and Education centres' and 'Rural England'). LAC migration patterns were analysed firstly using all flagged customers, and secondly, by different life stages. This was in order to assess potential differences between demographic attributes. For the purpose of this analysis, life stages were defined using the following age bands; Young Adults (16-24), Early Adulthood (25-34), Midlife (35-50) and Mature Adulthood (50-80), and differences in loyalty card migration characteristics compared.

4.4.2. Results

Due to this segment of customers being an actively transacting group, 97.45% could be accurately matched to one LA. Analyses at the MSOA level were able to match 6.1% of *address errors* and 2.5% of *address change* customers to one small area. These customers had an average of 121 transactions and 24 unique stores, indicating that customers with a larger overall store network could be matched with finer granularity. Due to these relatively low match rates, only LA level data were considered for further analysis. Table 4.2 shows the average transactions and available store data per customer required to suitably match individuals to their relocated areas.

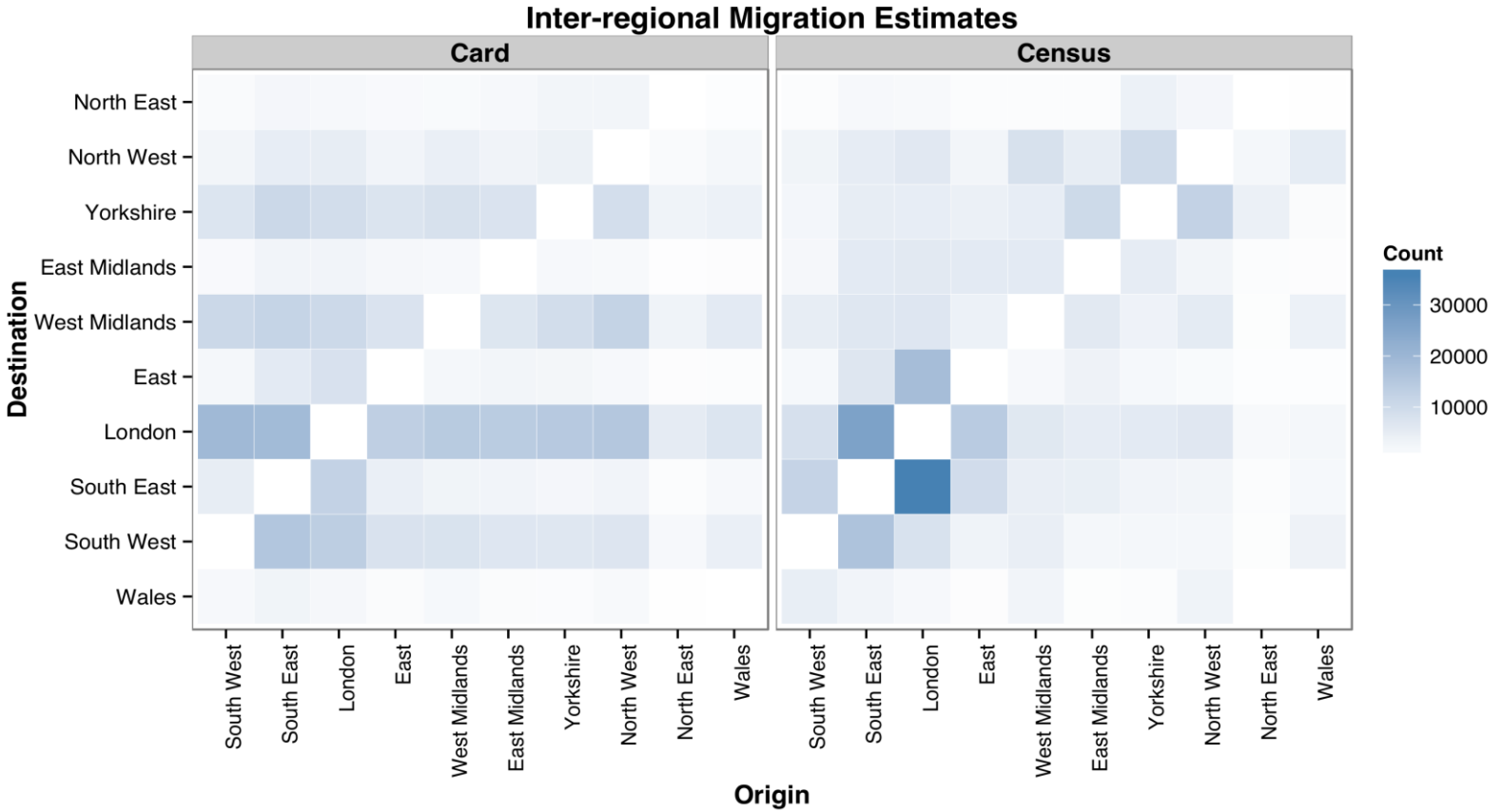
Table 4.2: Average data required for relocation estimation accuracy.

LA Matches	Average Transactions	Average Stores
1	180	23
2	56	13
3	23	2
4	6	1

Comparison with census origin-destination statistics indicated that whilst there may be a relationship with inter-regional flows, the card data substantially underestimated intra-regional migration from what we may expect. This is likely due to the limitations acknowledged in Section 4.3, such as students dominating the sample or the inability of the method to highlight local migration patterns. Therefore, only inter-regional observations were considered for further analysis. Figure 4.14 demonstrates a comparison of inter-regional flows using the card and census data. Spearman's correlation indicated a moderately strong positive relationship between inter-regional flows (rho = 0.69, $p < 2.2e\text{-}16$). Comparable proportions of movement between regions could be observed, although the card data overestimated moves between areas in some cases (i.e. London to South East). This could be due to the card data covering a different and longer temporal period. However, overall the card estimates did follow broadly expected patterns of population migration.

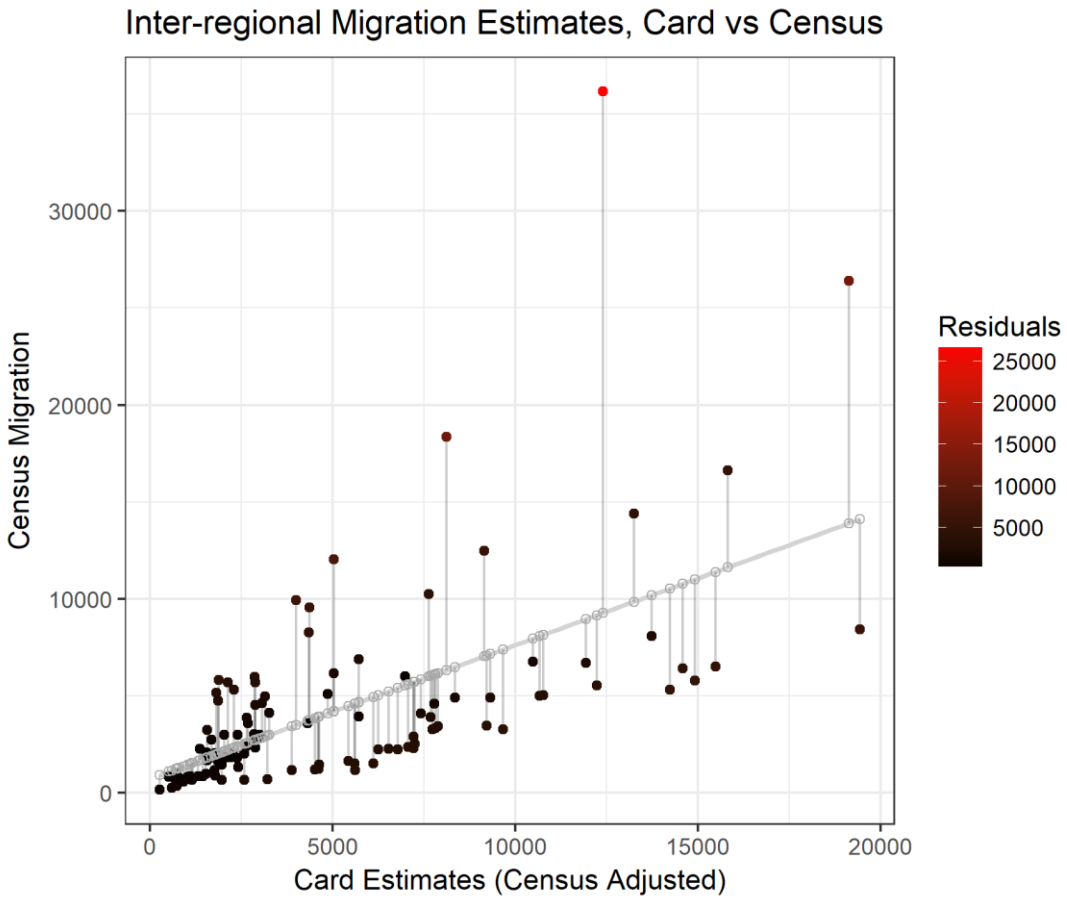
Figure 4.15 shows flows between LA characteristic groups using the card and census data between a) Supergroups and b) card flows between groups. Figure 4.16 shows the segmentation of these migration patterns by life stage. Similar patterns of social mobility could be observed between the card and census estimates at the Supergroup level, such as a large proportion of flows from numerous characteristic groups to 'Business and Education Centres', which typically describe migration to larger cities or cosmopolitan areas. Outflows from these areas were also comparable across the two datasets, such as high proportions to the 'London Cosmopolitan' Group, 'Prosperous England' and the 'Suburban Traits'. Analysis at the Group level showed further expected patterns such as large interactions between 'Business and Education Centres' and 'Rural Coastal and Amenity' areas. Segmenting these migration

patterns by life stage showed how the highest proportion of flows for different age demographics varied substantially. For example, 'Young Adults' were most likely to migrate to 'Business and Education centres' (this sample most likely included the majority of the student population). Similar trends could be observed for 'Early Adulthood'. Conversely, the largest proportion of both the 'Midlife' and 'Mature Adults' groups could be seen to migrate towards more rural and suburban areas or 'Prosperous England'.


Inter-regional Migration Estimates
Card
Census
Destination
North East
North West
Yorkshire
East Midlands
West Midlands
East
London
South East
South West
Wales
Origin
Count
30000
20000
10000

a)

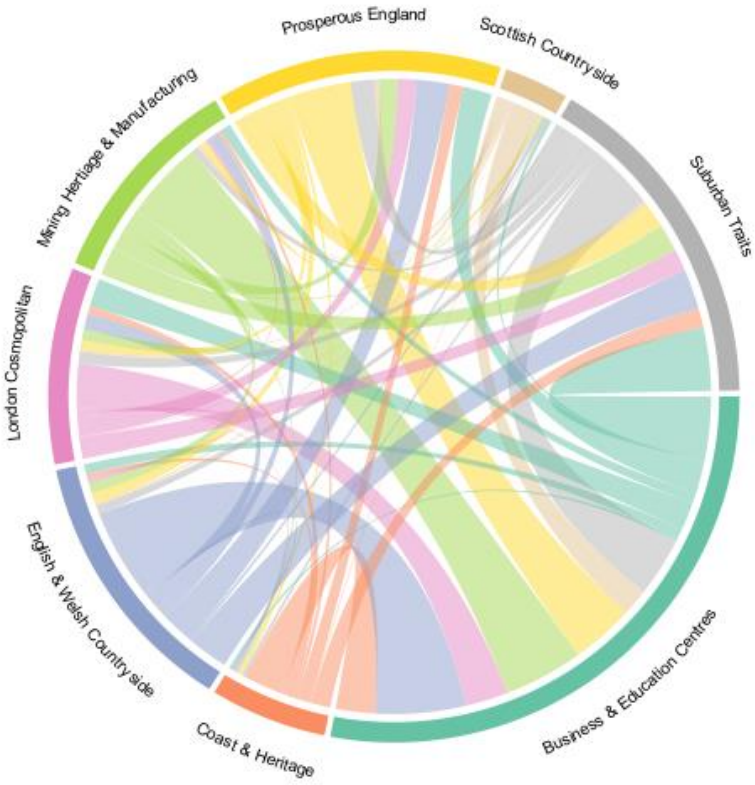
b)



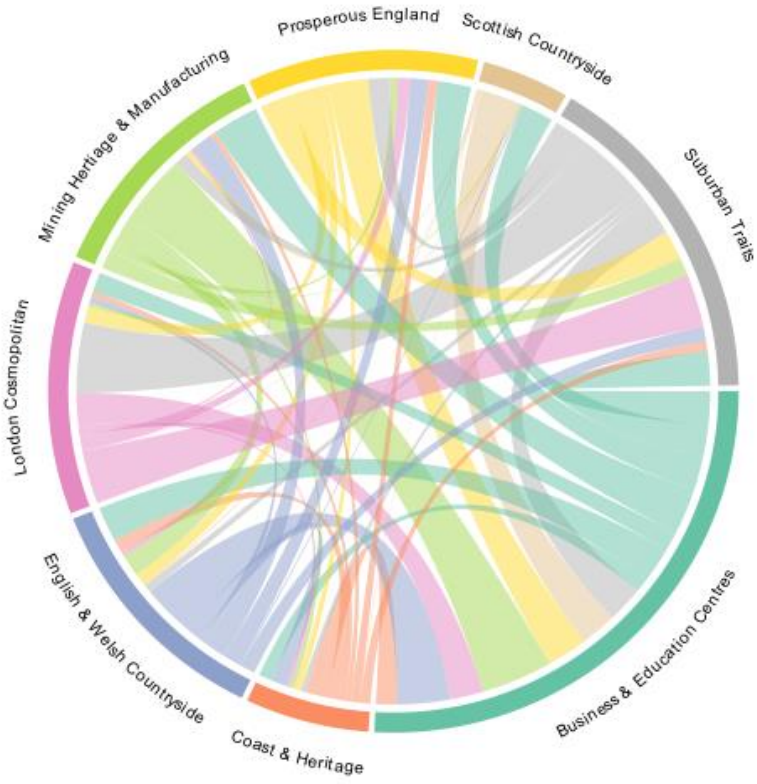
Coefficient = 0.69, p < 2.2e-16

Figure 4.14: a) Inter-regional migration estimates using loyalty card data and census origin-destination statistics (published in Lloyd and Cheshire, 2018) and b) inter-regional census vs card estimates with residuals and fitted values (method = '*lm*').

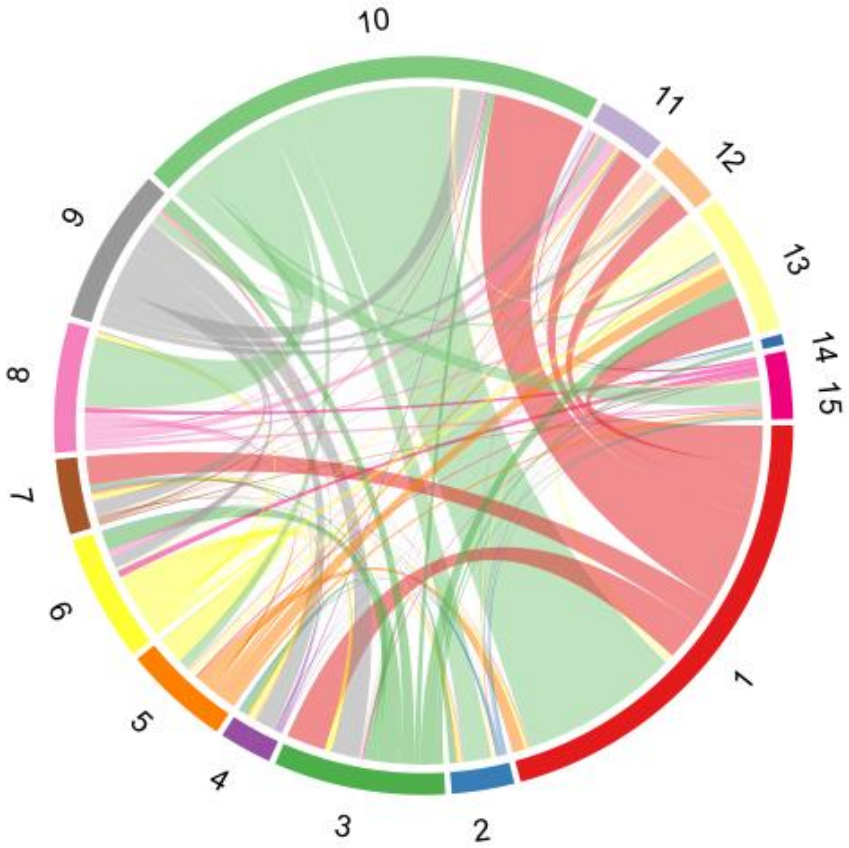
a) Card Estimates



Census Estimates



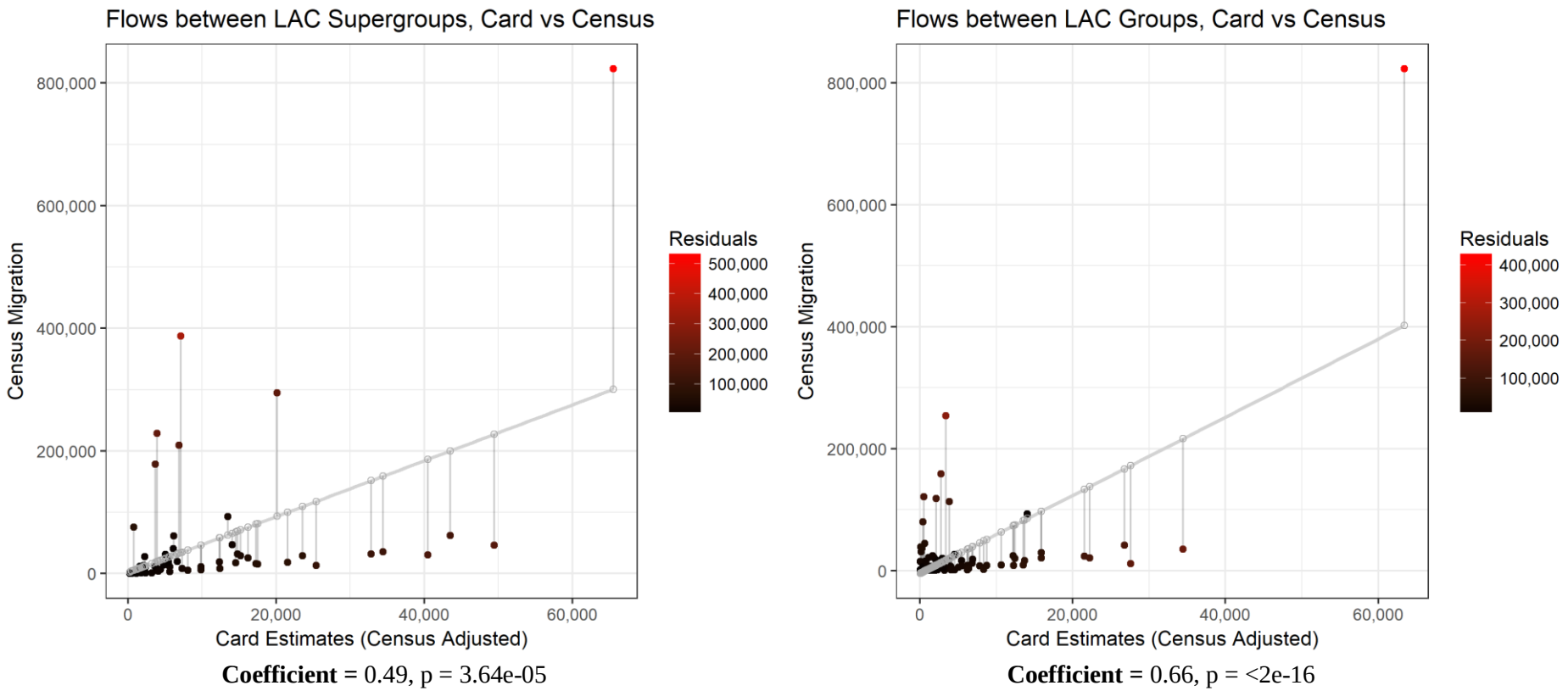
b) Card flows between Groups



Key

1. Rural Coastal and Amenity
2. Rural Hinterland
3. Rural England
4. Rural Scotland
5. Remoter Scotland and Glasgow Suburbs
6. London Cosmopolitan Suburbia
7. London Cosmopolitan Central
8. Growth Areas and Cities
9. Multicultural Suburbs
10. Business and Education Centres
11. Coastal Resorts and Services
12. Heritage Centres
13. Prosperous England
14. Manufacturing Traits
15. Mining Heritage

c)

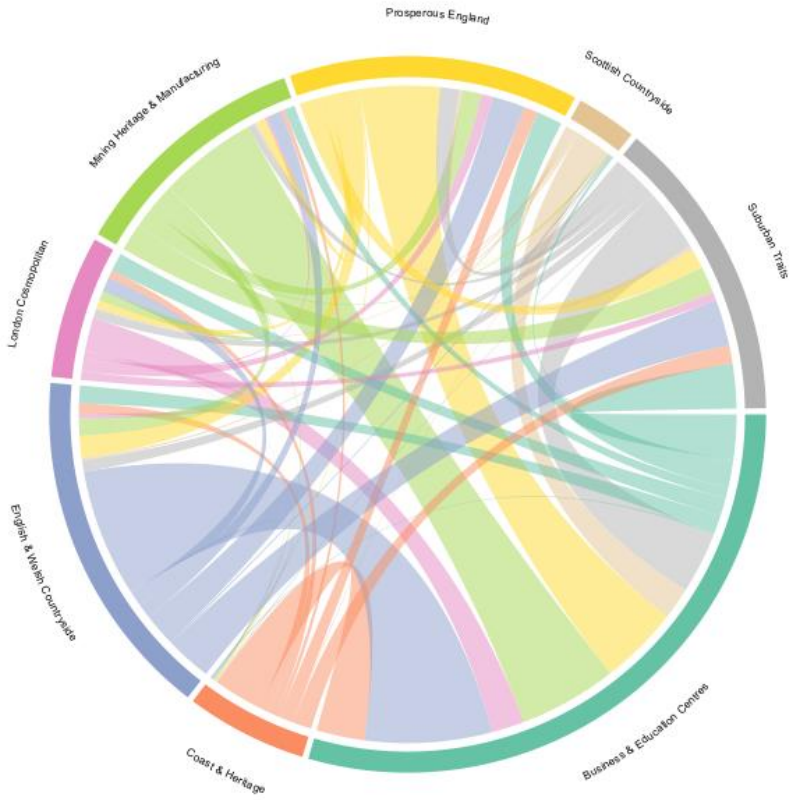


Coefficient = 0.49, p = 3.64e-05

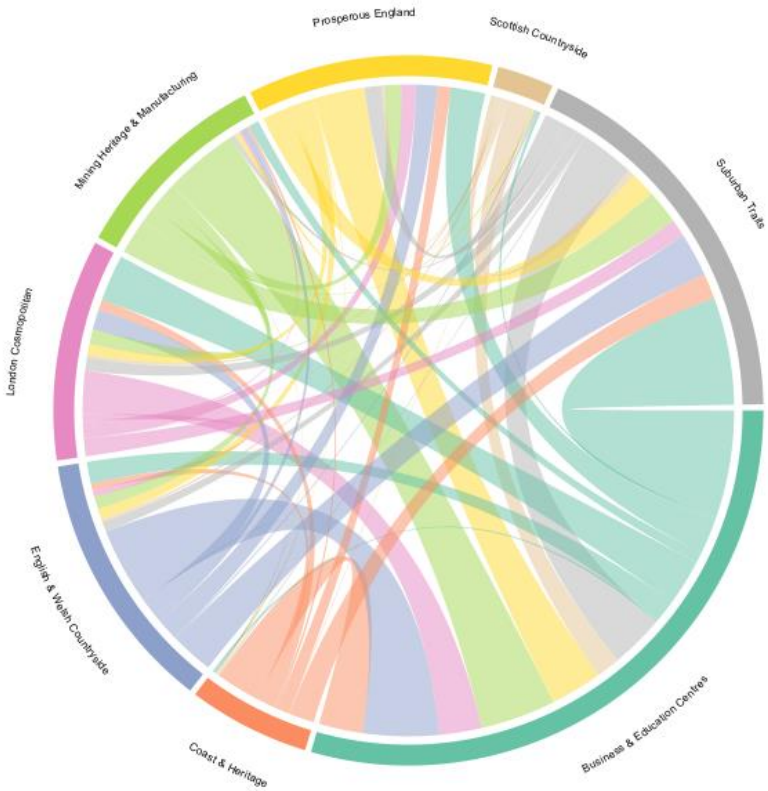
Coefficient = 0.66, p = <2e-16

Figure 4.15: Origin and relocation characteristics for a) LAC Supergroups using card and census data, b) card flows between Groups and c) census vs card flows between LAC Supergroups and Groups with residuals and fitted values (method = '*lm*').

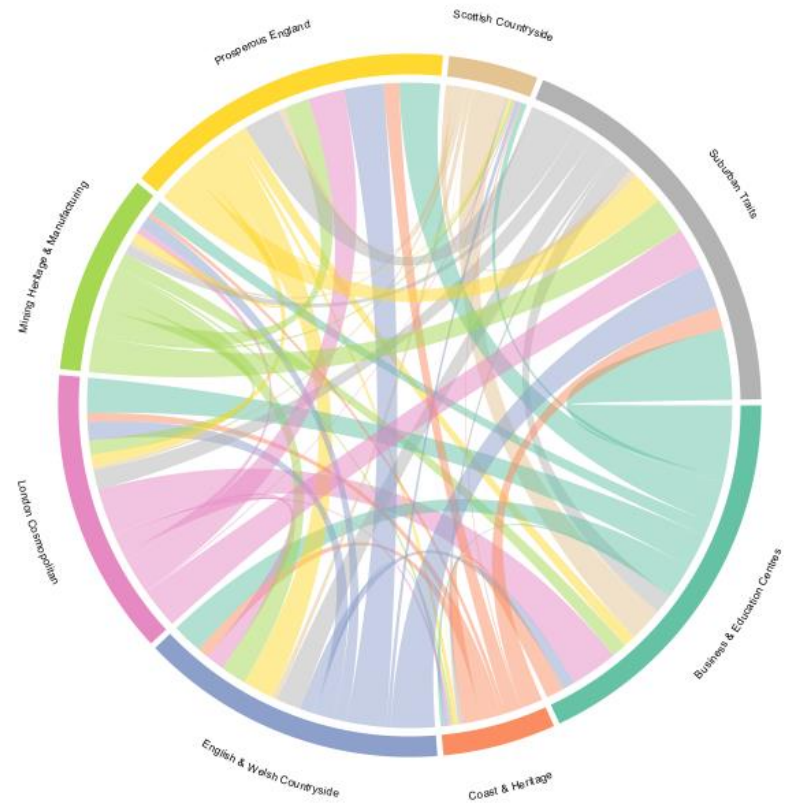
a) Young Adults



b) Early Adulthood



c) Midlife



d) Mature Adults

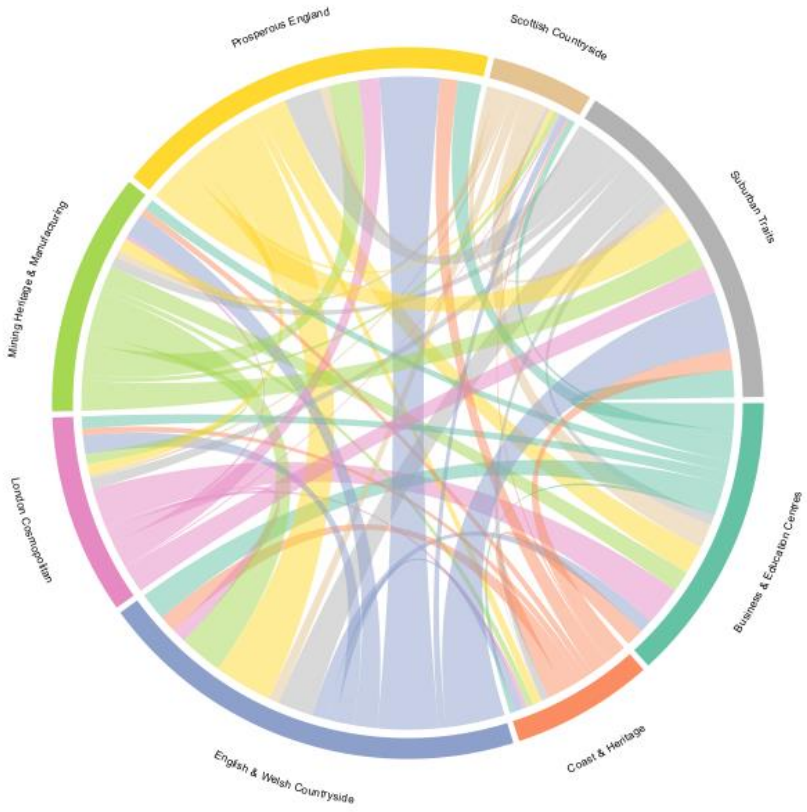


Figure 4.16: Social mobility by life stage using loyalty card data

These results suggested that the method was able to firstly, follow broadly expected trends of inter-regional migration patterns in the general population and secondly, indicate expected patterns of social mobility. For example, we may expect youths to migrate towards business and education areas and older cohorts away from such areas, due to differing stages in the economic cycle. Whilst the customers flagged may be substantially biased towards a young student population, it appeared that we may still be able to utilise these data to understand social mobility between groups within the general population. Nevertheless, the card sample largely underestimated moves between local areas, which will in turn limit the overall representation of social migration patterns.

4.4.3. Method Limitations

It is important to acknowledge a number of limitations with the approach adopted here. Firstly, as the method fundamentally utilises transactional behaviours to determine plausible geographical information, it is unable to detect changes in location that do not cause a modification in store visiting behaviour. For example, a customer who relocates to an area within close proximity may be unlikely to change patronage behaviours, especially if the store competition in a particular area is low (i.e. more rural locations). These relocations are essentially unidentifiable error in the data. Therefore, results will be constrained by the pre-defined spatial distribution of stores as anchor points for behaviour, which creates bias in the amount of data available across different areas. This study therefore highlights important considerations for the adoption of such data as indicators of population and social statistics; primarily that analyses are heavily dependent on the data available, which limits the scope of insights than can be derived. In this case, the dynamics and time-span of these data predispose the heuristics to be best suited to extracting moves over longer distances. Secondly, due to these data being static and historical in nature, there were a number of accounts flagged as showing deviation from normal behaviour, yet were unable to be conclusively classified due to a lack of transactional data. This is likely due to the fact that the data only cover a two-year window, and many customers transact less frequently than others.

More longitudinal records would be required to understand these patterns further. For example, despite being large in volume, many limitations arose from the limited time period of this sample, and the lack of complete data pertaining to both individuals and stores as a result. This inherently restricted the extent to which we could infer dynamics such as changes in transactional behaviour over time. Implementation on more longitudinal and frequently updated data would undoubtedly improve estimations of the small area trip distributions derived here. This may allow analysis at a finer spatial granularity, which would be more sensitive to network changes. Despite these limitations, access to a unique dataset facilitated the construction of a

method to detect these uncertainties, and there is scope for this technique to be adapted for implementation on real-time big data.

It is finally important to recognise that the method presented here is bespoke to this particular dataset. For example, the loyalty card data consisted of considerably voluminous behavioural data that had both residential and behavioural spatial reference points. Due to the large variation between content, structures and attributes present in the current big data landscape, much further work is needed to understand how we might address veracity issues in various sources. It is likely that bespoke methods will be necessary dependent on the specific nature of the dataset in question. However, it is hoped that these analyses demonstrate how we can potentially apply data-driven methods through the utilisation of pre-existing theory to mitigate these uncertainties to some extent, in order to support the reliable adoption of large consumer datasets in social research and identify insights that would not be practically obtainable using traditional methods.

4.5. Discussion

These analyses presented insights regarding the types of uncertainties that may arise due to novel forms of data being produced as a by-product of alternative commercial agendas, rather than conforming to more traditional approaches to data collection. As these data have been particularly hard to obtain for academic research, these results offer unique insights into the dynamics of and inaccuracies within a commercial dataset. The development of a data-driven method to address these uncertainties utilised knowledge and theory from multi-disciplinary domains to identify inaccurate addresses based on customers' stated home location and their transactional histories. Outputs from the implementation of this method suggested that firstly, the majority of addresses in these data are likely correct. However, secondly, a segment of the population within loyalty card data might be unrepresentative of a current place of residence, and these errors are likely not random. Despite being unable to unequivocally verify these findings due to the absence of reference data, comparisons with existing national statistics suggested that the method was able highlight customers that we may expect to have more transient residential locations and demonstrated comparable trends to inter-regional migration. In addition, despite a bias towards a student population, results were still able to reveal expected trends of social mobility between area characteristics, demonstrating that whilst further investigations are necessary, big data veracity issues of this kind may also be of value to understanding certain dynamics within the general population.

These insights have a number of implications for the use of large consumer datasets in social science research. Firstly, it exposes veracity issues inherent in consumer data, of which have been poorly understood to date due to a lack of access outside of the commercial settings in which they are created. Secondly, it presents the development of heuristics by which we can

attempt to address these issues, highlighting the potential to identify uncertainties using the data alone, where linkage to reference data is not possible. Such methods may be adopted in other relevant systems or settings in an attempt to clean spurious patterns, for example, in any dataset where both residential and behavioural points of reference are accessible. Thirdly, whilst further investigations are needed, the observed relationships with census data present a promising example of the potential to use alternative datasets as a means of creating more frequent indicators of migration. For example, census data and population registers offer information on acts of migration but do not create an overview of daily patterns. Surveys that focus on migration use small samples and are retrospective. Other monitoring methods over short periods of time, such as travel diaries, can deliver information on daily patterns but rarely capture the occurrence of a change in residence. This investigation shows an example of employing alternative approaches to overcoming these obstacles by enabling a more detailed approach in considering temporal patterns for a change of residence, and therefore allowing more precise focus to be placed on when a change of residence occurs.

Objectives for the Census 2021 and beyond suggest the integration of more address-level intelligence from administrative, commercial and open-data sources to help estimate non-response rates and to ultimately move away from the 10-yearly census approach (Stillwell 2016). Whilst it is acknowledged that further development is needed, these types of analysis have positive implications for utilising novel data sources to supplement these conventional data sources. The implications from this analysis for this cause are twofold. Firstly, key to this integration will be the ability to link data efficiently and accurately, yet we highlight here how preliminary data treatment is necessary to ensure the veracity of the commercial data being integrated. However, secondly, the methods presented here show an example of how we can attempt to mediate big data veracity effects, whilst also demonstrating a means of highlighting addresses, areas and specific population characteristics that may be more transient in nature. In terms of census objectives, this information could also facilitate targeting of non-responding households.

The potential applications of these insights may not only be of interest for uses in research, but also for retailers and other consumer data collectors and users who are operating reliant on consumers keeping up to date address records. For instance, results provide insight into the extent of customers who no longer live at their stated address and will therefore no longer receive their mail-based rewards, or be correctly identified for location-based targeting efforts. This could have negative impacts on proceeding loyalty behaviour and if aiming to distribute a limited number of offers, these errors could hinder the impact of such campaigns. In addition, understanding the demographic and geodemographic attributes of the customers most at risk of these uncertainties could help to mitigate these negative effects.

5. Temporal Profiling: Classifying Stores

5.1. Introduction

The inherent spatiotemporal nature of loyalty card data offers a framework by which to investigate the concepts of time geography and spatiotemporal population rhythms in a data-driven context. As summarised in Chapter 2, there is a need to explore how the velocity of such data may contribute to our understanding of both people and places. Research in this area has suggested that firstly, the temporal rhythms of a place (i.e. daily, weekly, seasonal or annual fluctuations) may influence its on-going formation and secondly, we can potentially identify the distinctive characteristics of a place according to its rhythmic ensemble (i.e. Crang, 2001; Edensor, 2016). This research further indicates that the formation and function of places may be inherently interlinked with the spatiotemporal rhythms of distinct social groups of people. This view would suggest that we can identify the characteristics of a place according to its temporal rhythm, which may also indicate the characteristics of individuals who patronise it.

Quantifying such trends requires data that capture detailed population activity patterns over both space and time, and as a result these notions have been largely founded on qualitative methods and theory. The following chapters present a series of analyses that aimed to make use of access to this unique dataset to explore these notions from a data-driven perspective. It should be acknowledged that the concept of 'place' here refers only to the specific context of retail centres in which HSR stores reside. The aim of this analysis was therefore to investigate trends in this specific context, and to explore the implications of these dynamics for retail centres, the concepts of time geography and spatiotemporal population rhythms more generally. This chapter presents the first stage of this analysis, which focused on classifying HSR store locations based on their temporal rhythms. Following this, Chapter 6 presents an investigation of the individuals who patronise those locations, and what their spatiotemporal profiles may indicate about their geodemographic characteristics.

To examine the temporal rhythms of HSR stores, a cluster analysis was conducted on transactional frequencies over time. Stores with similar temporal profiles were then explored in terms of their locational and retail composition characteristics. This aimed to understand if we could identify the distinctive characteristics of different retail locations based solely on the temporal fluctuations they exhibit. It should be noted that this investigation, in addition to those

presented in the succeeding chapters are largely exploratory in nature, but rationales are given for the methodological steps taken where appropriate.

5.2. Method

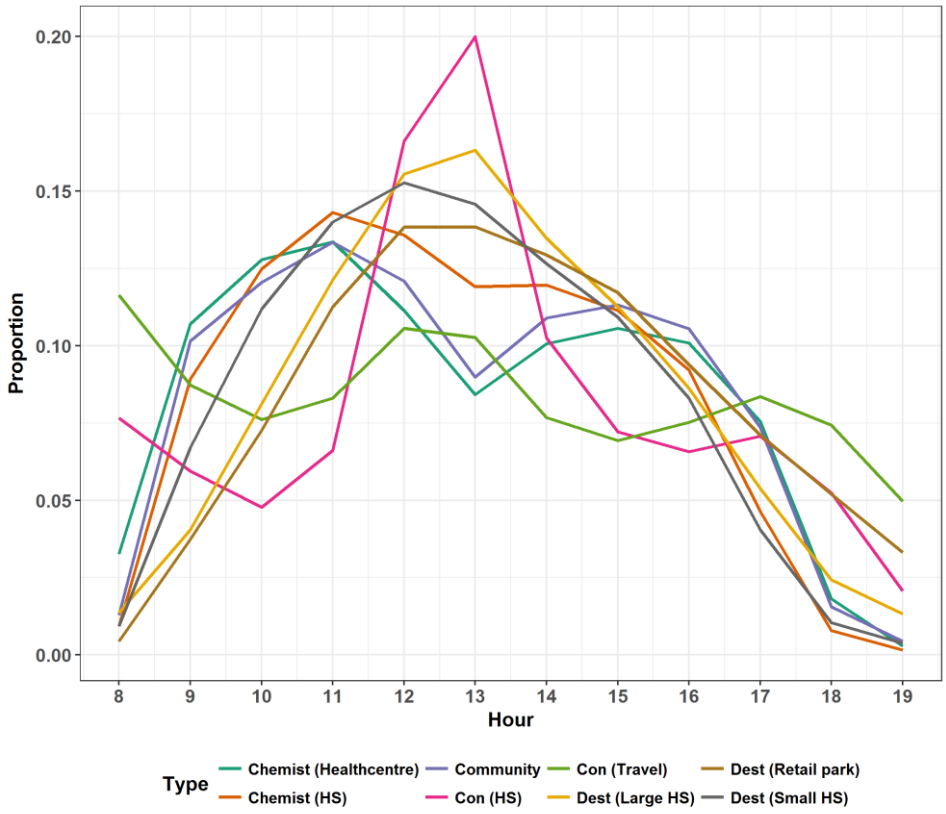
Clustering is, for the most part, an unsupervised process where the composition and number of clusters in the data are not defined a priori, and implementation requires consideration of data manipulation and cluster analysis methods in order to produce an optimum solution. This is often an iterative process that requires applying different techniques and observing the effects on cluster outcomes using a mixture of both quantitative and qualitative assessments (making it 'as much art as it is science'; Harris et al., 2005). Generally, an optimum solution may be viewed in terms of 1) compact clusters, with the objects in each group being as similar in characteristics as possible, and 2) having the highest possible separation in characteristics between different clusters (Berry and Linoff, 1996). However, it is similarly important that outputs provide a valuable representation of the real world that is both useful and easy to interpret.

The following sections outline the processes adopted for the creation of temporal HSR store profiles. This details the selection of variables, data treatment techniques, such as rate calculation and transformation, and implementation of the final clustering solution.

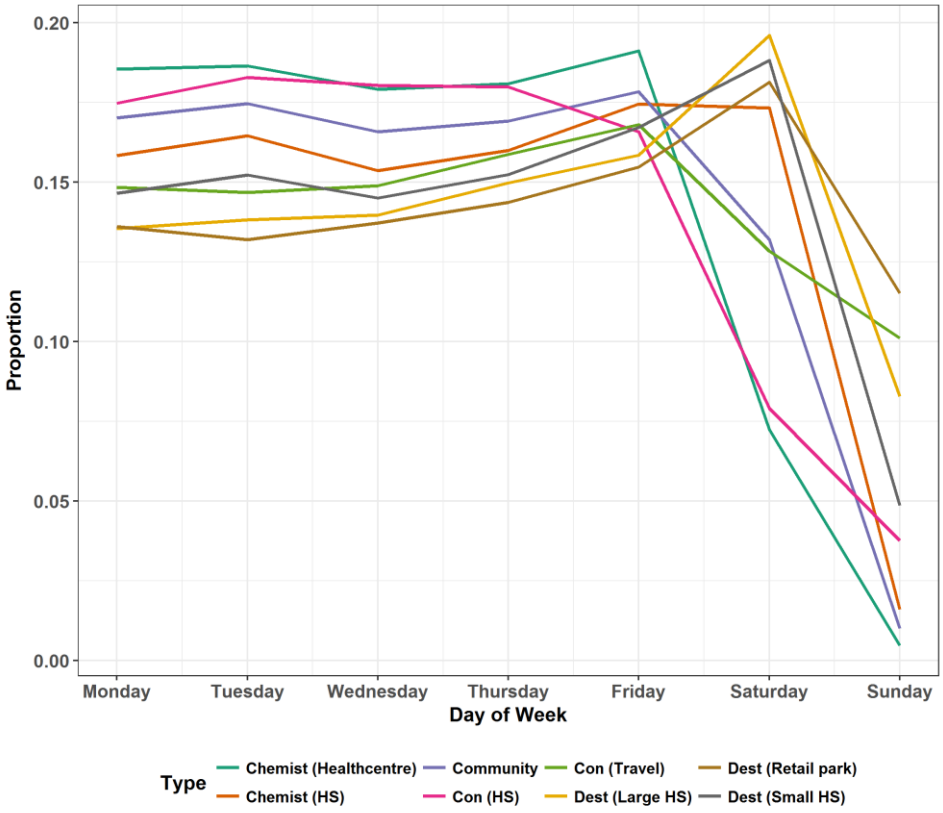
5.2.1. Data Preparation

5.2.1.1. Exploratory analysis

To inform the appropriate selection and treatment of data, exploratory analyses were conducted to understand the broad temporal trends of HSR stores. Data were obtained firstly, for each HSR store type (see Chapter 3, Section 3.2.2.3 for an overview of store types defined by the HSR) over various temporal intervals. Figure 5.1 shows transactional frequencies by a) hour and b) day of week for each store type, demonstrating how each exhibited distinct temporal patterns. These trends suggested that firstly, it may be important to consider temporal variations separately across weekday and weekend periods, as this may be a significant distinguishing factor between store types. Secondly, these patterns provided preliminary insight of potential relationships between temporal fluctuations and place types. For instance, convenience high street stores, which are primarily located in urban areas, exhibited evident periodic fluctuations (i.e. 8am, 12pm, post 5pm) and low activity during weekend periods, likely delineating a workplace population.



a)



b)

Figure 5.1: Transactional frequencies per store type, by a) hour, and b) day of week (normalised by total transactions per store type).

In contrast, chemists exhibited off-peak activity (i.e. weekday mid-morning) and increased weekend morning consumption, and large high streets and retail park locations exhibited high weekend volumes from late afternoon to evening. These trends suggested that underlying socio-spatial relationships with consumption patterns might be discernible through the temporal analysis of store locations.

5.2.1.2. Store selection

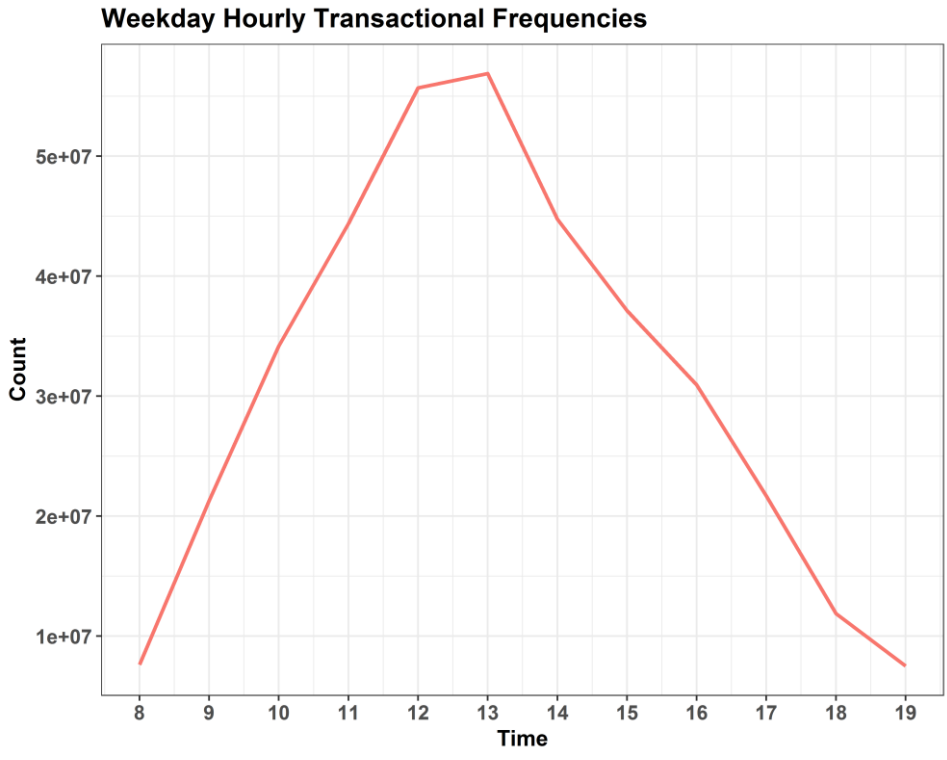
As outlined in Chapter 3, there were 2433 HSR stores distributed across GB. However, the removal of some stores was necessary prior to this analysis. Firstly, stores were excluded that were not considered representative of the general high-street oriented consumption patterns of the majority of locations. This included 'health centres' - a distinct type that represented small prescription dispensaries located in GP surgeries and 'Airport' format stores. Further exclusions were stores that did not exhibit at least 1 year of transactions, due to opening or closing during the time period of the data. Finally, there were a number of stores for which locational attributes were supplied, but were absent in the description data (likely due to disparities between the datasets provided). As store attributes were used to describe clusters post-classification, these were also excluded from the analysis. In all 182 stores were removed with 2251 stores going forward for analysis.

5.2.1.3. Temporal aggregation

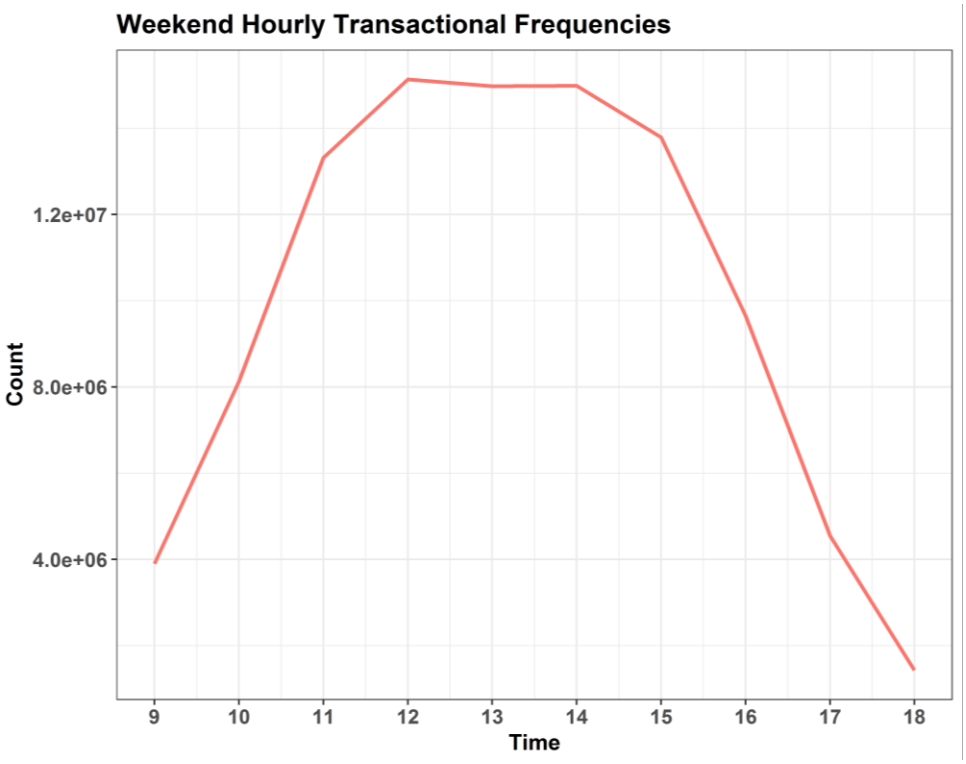
In order to obtain variables for clustering, a certain level of temporal aggregation was necessary. This refers to the process of summing or averaging the values of time series data over regular intervals. Whilst it is acknowledged that this can lead to a disregard of potentially important information (i.e. data smoothing or pattern alteration, see Steffensmeier et al., 2014), this was a necessary measure in order to comprehend trends from the large amount of data available. Thus, definition of appropriate time intervals involved a trade-off between reducing the dimensions of the raw data to a manageable size, whilst also minimising the loss of temporal patterns. The scale of interest for this analysis was that of daily consumption patterns, rather than over longer-term periods (i.e. months, seasons). However, based on the previous observations that weekday and weekend fluctuations were an important distinguishing factor between store types, it was also considered important to retain separation of these periods. Initial consideration was given to incorporating each day of the week, however, previous observations of consumption during these periods (see Chapter 3, Figure 3.13) indicated that whilst the magnitude of transactions varied between days (i.e. higher on Saturdays than Sundays), daily fluctuations were relatively uniform between Monday-Friday, and Saturday-Sunday. Therefore, data were aggregated to represent weekdays and weekends only.

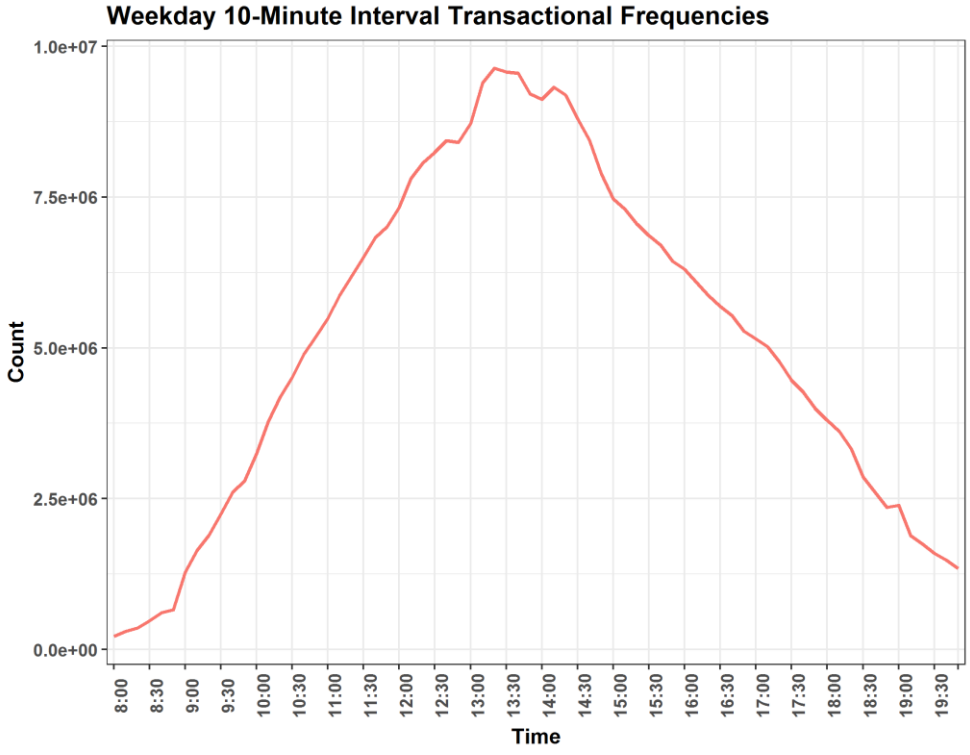
After consideration of appropriate representation of daily intervals, it was decided that initial clustering would be performed on hourly transaction volumes. This was in order to provide a general description of variation between stores that could likely be summarised by a reasonable number of clusters. Aggregating by more granular time periods at this stage would have created too complex outputs to describe basic temporal trends. Following this, a separate dataset was created, where transaction volumes were aggregated by 10-minute intervals. This aimed to provide a more in-depth analysis of fluctuations over daily periods and potentially facilitate further segmentation of stores, in addition to ensuring that important patterns were not disregarded by hourly aggregations.

Figure 5.2 depicts the resulting data (across all stores) utilised for cluster analysis, which pertained to transactions recorded per hour/10-minutes across the 2.5 financial years, during weekdays and weekends. Aggregating these data resulted in a total of 19 hours in which transactions were recorded. Abnormally early/late retailing hours were not reflective of the majority of stores in the sample, and typically described a small subset located in transport hubs. Data were therefore reduced to represent the most normative trading hours during weekdays (8am to 7pm) and weekends (9am to 6pm).



a)


Weekend Hourly Transactional Frequencies
Count
1.2e+07
8.0e+06
4.0e+06
9
10
11
12
13
14
15
16
17
18
Time


Weekday 10-Minute Interval Transactional Frequencies
Count
1.0e+07
7.5e+06
5.0e+06
2.5e+06
0.0e+00
8:00
8:30
9:00
9:30
10:00
10:30
11:00
11:30
12:00
12:30
13:00
13:30
14:00
14:30
15:00
15:30
16:00
16:30
17:00
17:30
18:00
18:30
19:00
19:30
Time

b)

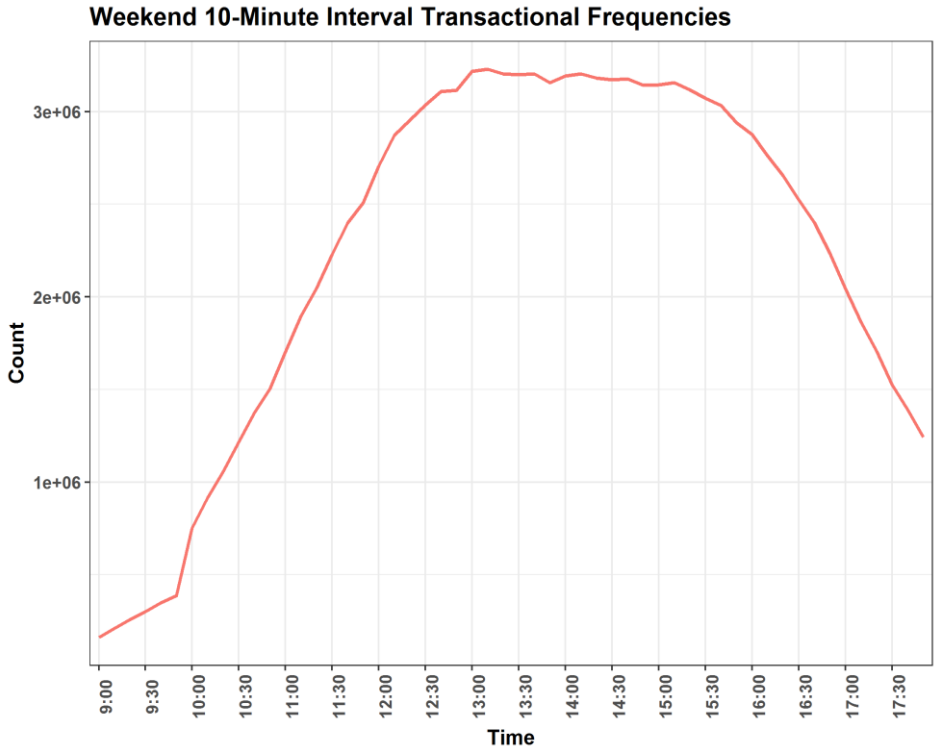


Figure 5.2: Total transactions during weekdays and weekends per a) hour and b) 10-minute interval.

5.2.1.1. Rate calculation

Before clustering, it is vital to ensure that all variables are measured on comparable scales, ensuring equal weighting and control of underlying population structures (Gale et al., 2016). This often requires a processes of rate calculation, transformation and standardisation to place the variables onto a single scale. The data here exhibited inconsistent underlying population structures in their raw form, dependent on the total number of transactions that occurred across different store locations. Clustering of this raw count data would have failed to identify unique transactional peaks independent of base populations, therefore rate calculation was necessary. This typically involves turning variables into percentages or ratios by dividing values by their relevant population denominator.

However, further considerations arose from the temporal nature of these data. Time-series analysis presumes that univariate and multivariate social systems are composed of short, medium, and long-term processes of dependencies between variables (Steffensmeier et al., 2014). These characteristics mean that time-series data are composed of trends and cycles that are governed by relationships that evolve over time, and thus tools must be applied to correctly identify underlying patterns. All time-series data have three basic parts (Wu and Wei, 1989):

1) *A trend component*
    a. Non-stationarity (long term change in mean)

2) *A seasonal/cyclic component*
    a. Seasonality (i.e. annual variation)
    b. Cyclicity (variation fixed in period. i.e. diurnal)
3) *An irregular component*
    a. Signal after removal of trend and seasonal/cyclic variations

Prior to the temporal aggregation outlined in Section 5.2.1.3, these data exhibited both non-stationarity and seasonal components, such as increased sales during the Christmas period (a common retail trend). However, aggregating data across the 2.5 years to represent only a single daily/weekly period alleviated further consideration of these attributes, as transactions occurring in each time period were considered as a whole sum, rather than over longitudinal periods that may exhibit these moving averages. Nevertheless, a second rate transformation technique was still required to account for a trend component that was evident across days. For example, as depicted in Figure 5.2, transaction counts demonstrated moving averages across intervals based on underlying variation in the total number of transactions recorded during different times of the day and week. This is a similarly common dynamic of shopping behaviour, for example, where there is higher demand across all stores at certain time periods. This could be delineated here by a midday peak on weekdays, and late morning to early afternoon on weekends.

Clustering of these data therefore required accounting for both the base population of stores and the trend component of the relative time period. To achieve this, data were firstly normalised by the total transactions that occurred across the full sample of data, in the relevant time period (e.g. hour, or 10-minute interval) and secondly, by total store transactions. Resulting values pertained to the percentage of (weighted) transactions that occurred within each time interval, for each store. This facilitated identification of unique consumption variations across time whilst also accounting for differences in capacity among store locations.

An important subsequent consideration of this process was that these data were now compositional in nature, also referred to as ‘closed data’ (Aitchison & Greenacre, 2002), with each value describing relational parts (i.e. the proportion of transactions per time period, per store). Values therefore conveyed only exclusively relative information, which represented a unique data type requiring specific treatment for further analysis.

5.2.1.2. Transformation of Compositional Data

The dynamics of compositional data were originally explored by Aitchison (1986), who outlined how data of this type have a number of consequences for analysis, such as an inability to perform well with standard statistical procedures. A transformation-based methodology to deal with such data was proposed in the early 1980s (Aitchison, 1986), due to the recognition that compositions provided information only about the relative magnitudes of their components,

and hence a need for data analysis focused on the ratios between components (Palarea-Albaladejo, Martín-Fernández & Soto, 2012). In brief, a composition multiplied by any positive constant will contain the same information as the former (Pawlowsky-Glahn and Buccianti, 2011), and this invalidates many statistical approaches.

As outlined by much research (see Aitchison, 1986; Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015, for an in-depth review), these data can be adjusted in order to give meaningful inferences by means of log ratio transformation. This is based on the rationale that if compositional data carry only relative information about components, we can utilise the logs of ratios to make them appropriate for analysis (Tsagris, Preston & Wood, 2011). This transformation essentially converts the data from its compositional sample space (the simplex) to a Euclidean sample space (three dimensional space), therefore removing the problem of a constrained sample and opening up analyses to all standard multivariate techniques. The process of log ratio analysis for compositional data, as outlined by Aitchison (2008), can be stated as follows:

1) Formulate the compositional problem in terms of the components of the composition.
2) Translate this formulation into terms of the log ratio vector of the composition.
3) Transform the compositional data into log ratio vectors.
4) Analyse the log ratio data by an appropriate standard multivariate statistical method.
5) Translate back into terms of the compositions the inference obtained at Step 4.

Several log ratio transformations have been proposed for compositional data. These include the centred log ratio transformation (*clr*), isometric log ratio transformation (*ilr*) and the additive log ratio transformation (*alr*), of which *clr* is the most often used (Aitchison, 2008). It is beyond the scope of this analysis to detail the workings of each of these transformations (an in depth review can be found in van den Boogaart & Tolosana-Delgado, 2013), however, there has been considerable discussion of the advantages and disadvantages of each. Whilst some research suggests that these transformations yield similar results (Godichon-Baggioni, Maugis-Rabusseau and Rau, 2017), selection of the most appropriate technique is often dependent on the data in question.

The *clr* was selected for this analysis for a number of reasons. Firstly, the *alr* is typically no longer used, due to its inability to preserve distance and variance in the data. It also requires a specified denominator as a divisor, the choice of which can largely affect outcomes. Secondly, the *ilr* has the disadvantage of the loss of a variable; meaning data are no longer directly interpretable in terms of the original composition. Alternatively, the *clr* works by dividing each feature by the geometric mean then taking the logarithm (Fernandes et al., 2014), meaning

distances are preserved and data are directly interpretable in terms of the original structure. The *clr* can be expressed as:

$$\mathrm{clr}(x) = (\log(x1/g(x)), \ldots, \log(x_d/g(x))) \tag{5.1}$$

where x represents the composition vector, $g(x)$ is the geometric mean of the composition x, and x_d is Euclidean distances between the individual variables. This was applied to the data using the *clr* function from the R package *compositions* (van de Boogaart, Tolosana & Bren, 2015). A final common implementation before clustering is to standardise the data to ensure the variables are on a comparable scale. However, these procedures are relatively non-informative for compositional data, as the arithmetic mean and the variance or standard deviation of individual components do not fit with the Aitchison geometry as measures of central tendency and dispersion (Pawlowsky-Glahn, Egozcue and Tolosana Delgado, 2007). In short, applying traditional standardisation techniques can destroy relationships between compositional parts. Whilst it is possible to perform standardisation of compositional data, (see Palawsky-Glahn et al., 2007, for a review of methods), this is typically deemed necessary when variables are not derived from the same scale (i.e. comparison of multiple compositions or inclusion of non-compositional variables) and thus was not implemented for this analysis.

5.2.2. Clustering Method Selection

Many clustering methods exist in literature, yet selecting an optimum method is often a subjective process (Singleton and Longley, 2009). Methods can primarily be divided into two classes: model-based methods, such as mixture models (McLachlan and Peel, 2004), and methods based on dissimilarity distances, such as hierarchical clustering (Ward Jr, 1963), K-means (MacQueen, 1967), or Kmedians (Cardot, Cenac and Monnez, 2012). However, despite the large number of existing clustering methods, there has been relatively little attention paid to the most appropriate strategy for clustering compositional data (Godichon-Baggioni et al., 2017; Tauber, 1999; Zhou, Chen and Lou, 1991; Martin-Fernandez, Barcelo-Vidal and Pawlowsky-Glahn, 1998). The methodological approach adopted for this analysis drew heavily on previous research demonstrating that the K-means algorithm can be particularly effective in delineating trends in this context (see Godichon-Baggioni et al., 2017, who demonstrate the suitability of this algorithm with data of a comparable structure to the sample here). In addition to this, K-means is the most commonly used technique in geodemographics (Lansley, Wei and Rains, 2015) and this approach therefore follows in the path of existing literature surrounding conventional geodemographics (Harris et al., 2005).

5.2.2.1. Implementing K-means

K-means is a top-down, iterative relocation algorithm based on an error sum of squares (SSE – the sum of squared differences between each observation and its group's mean) measure approach where the number of cluster groups is defined by the user (Harris et al., 2005). This can be expressed as:

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} || x_i^{(j)} - c_j ||^2 \quad (5.2)$$

The algorithm seeks to reduce the sum distance between each data point $x_i^{(j)}$ and their respective cluster centre c_j. It firstly initialises with k 'seeds' randomly placed within the multidimensional space of the input data. Data are then assigned to their closest seed, creating an initial cluster assignment. Cluster centroids are then recalculated as the average of the attribute values for all data points assigned to each cluster. The data points are then reassigned if they become closer to new cluster centroid and this process repeats iteratively until the centroid locations cannot be moved as an optimum solution has been reached (Harris et al., 2005). The R default algorithm of 'Hartigan-Wong' (Hartigan and Wong, 1979) was selected, as this focuses on maximising similarities rather than minimising differences between groups. Details of this algorithm in comparison to alternate options that aim to reduce differences (i.e. Lloyd/Forgy, MacQueen) are outlined further in Morissette and Chartier (2013).

The simplistic nature of K-means makes it one of the most commonly used methods, however, given the random allocation of initial seeds, the algorithm is stochastic in nature and a single iteration will not achieve an optimum solution. Singleton and Longley (2009) suggest it is appropriate to iteratively implement this algorithm a minimum of 10,000 times. In addition, it requires a user-specified number of partitions of which choice can largely affect clustering outcomes.

5.2.2.2. Cluster number selection and classification structure

Selecting the number of clusters when implementing K-means is often a subjective decision based on the context of each individual system (Vickers, Rees and Birkin, 2005; Singleton and Longley, 2009). Typically, the higher the number of clusters, the smaller the mean distances between each data point and its nearest cluster centroid (Singleton and Longley, 2009), yet, it is also commonly recognised that a classification with too many groups ceases to become useful. For instance, this makes the model harder to interpret, and often groups can be difficult to distinguish (Lansley, Wei and Rains, 2015). Watson and Callingham (2003) suggested that for

ease of interpretability, the highest level of a classification should describe no more than 6 clusters and a second tier around 20 clusters, to enable good visualisation and description. This is evident in many existing widely used classifications.

The decision on the final number of clusters was based on a trade-off between cluster homogeneity, classification complexity and assessing if cluster characteristics demonstrated realistic representations of the data. The within cluster sum of squares (WCSS) and total between cluster sum of squares (BCSS) are calculated as part of the algorithm, which indicate how tightly clustered a particular dataset is. The WCSS describes how close objects within each cluster are to their centroids, providing a measure of cluster homogeneity. The BCSS measures the distances between the clusters, and therefore how similar they are to each other. The construction of clusters here was guided by the WCSS value, as emphasis was on ensuring clusters were as homogeneous as possible rather than as dissimilar as possible.

The number of layers in a classification is also subject to critical analysis. A two-tiered classification was implemented here, with the highest tier utilising the hourly interval data and the lower tier 10-minute intervals. It was decided that analysis at a more temporally disaggregate level, or creating a subgroup level would produce unnecessarily complex outputs rather than providing further meaningful insights, given the relative simplicity of input variables. Figure 5.3 provides an outline of the classification structure, which sought to describe broad temporal trends at a more aggregate ‘Supergroup’ level. Subsequently, data were segmented by these initial groupings and more granular variations clustered using the 10-minute interval data, which created the ‘Group’ level.

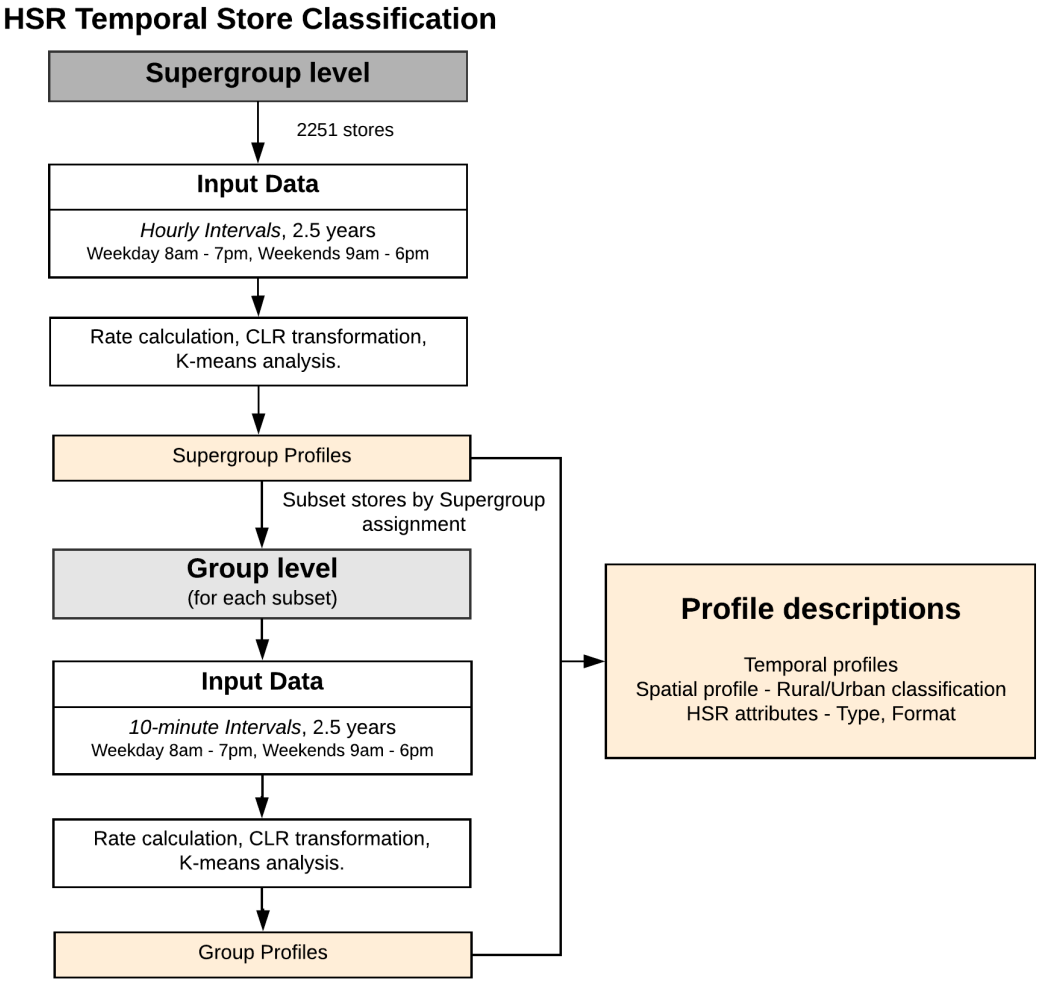


Figure 5.3: Overview of the temporal store classification process and structure.

Plotting WCSS values by number of clusters can indicate how many groupings may be necessary in order to best summarise the data by illustrating the proportion of total variation that is explained by each number of groups. This was repeated multiple times, suggesting a bend at 3-5 clusters (indicating a reduction in the amount of variance explained by additional groups - see Figure 5.4). Iteratively running the K-means algorithm using various K values indicated that the optimum number of clusters to both improve WCSS values, yet also create representative characteristics was 5. After this had been delineated, the algorithm was run over 10,000 iterations to identify the best fitting solution by iteratively running the *kmeans* in R, and recording the resulting WCSS and BCSS values of each iteration. From observing multiple outcomes, a prominent finding was that clusters were extremely similar each time, with only a small number of stores being reallocated across groups. This is typically observed when K-means is able to efficiently segment a dataset, and therefore supported the selection of this method. Total WCSS values ranged between 26.6 and 34.5.

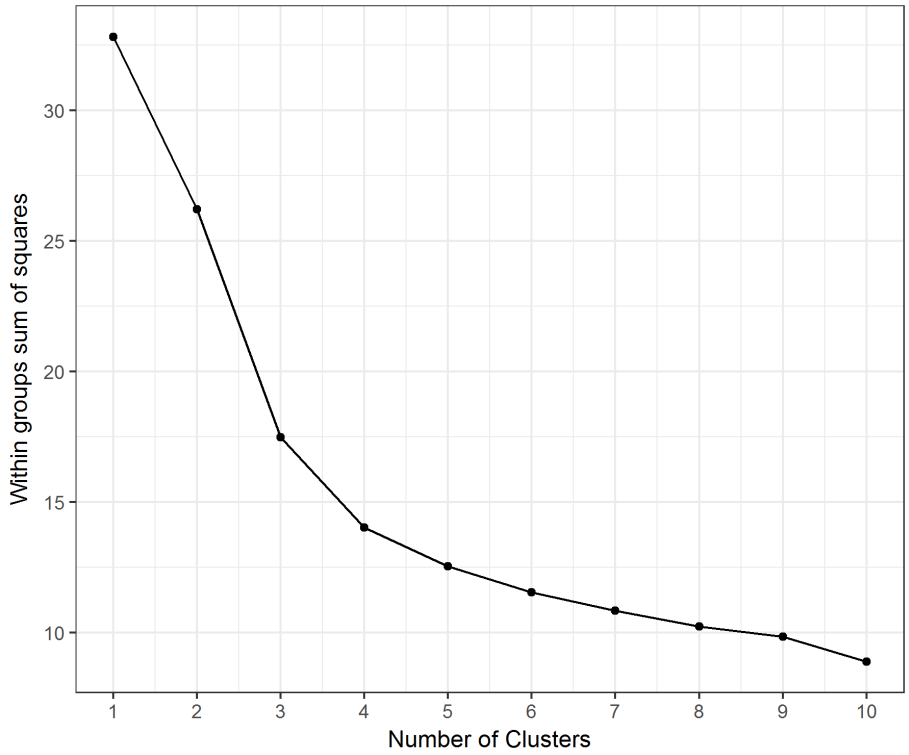


Figure 5.4: Plot of WCSS values by number of clusters, demonstrating the proportion of total variation explained by each number of groups.

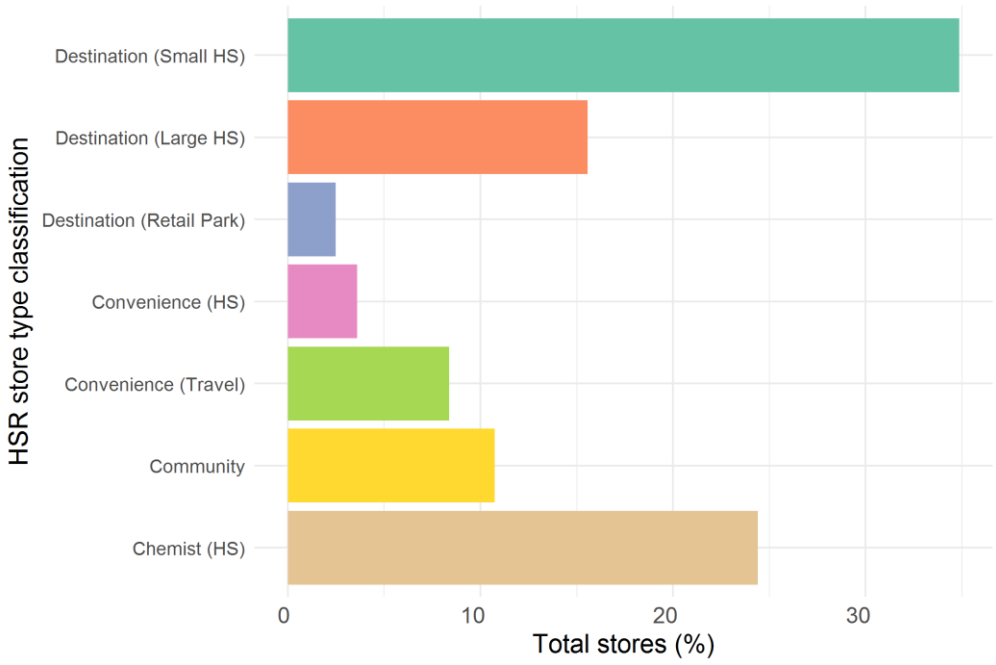
Once Supergroups had been created, data were segmented by their assignment and 10-minute interval data selected for each subset. These subsets were treated with the same methodology as outlined so far including rate calculation, transformation and implementation of K-means. WCSS plots indicated that between 2-4 clusters might be most suitable for each subset. Finally, evaluation of the resulting cluster outputs was necessary. Whilst the WCSS value was used as a metric in the creation of clusters, this value is often not directly comparable across clusters with different numbers of observations. Therefore, to evaluate within-cluster variability, distance from centroid measures were calculated (referring to squared Euclidean distance or SED), including the maximum, minimum, mean and the number of stores that fell above or below the mean. This provided a measure of the extent to which stores classified into any particular cluster varied from the 'average' characteristics of that cluster and thus quantification of those closer to, or further away from, the cluster centroid.

As noted in Chapter 3, to visualise results whilst also preserving HSR anonymity, store locations were aggregated to 5km grid cells for national level visualisations and 1km grid cells for more granular visualisations. This was achieved by truncating coordinates in R and representing locations as cell centres. Cells may represent more than one store location, therefore counts per cell were also obtained.

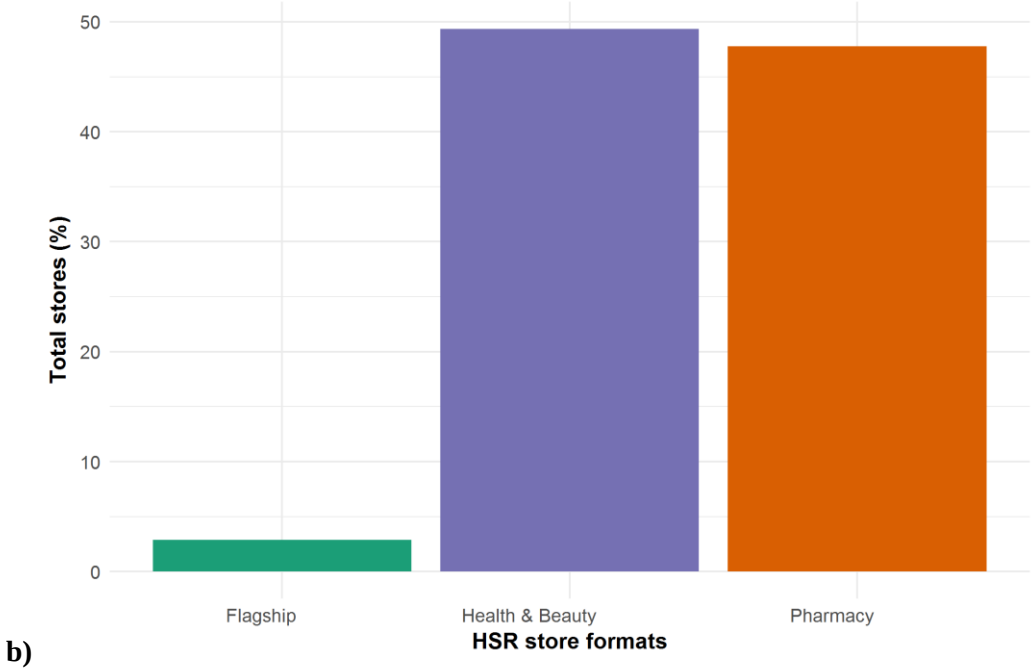
5.2.2.3. Cluster descriptions and pen portraits

To describe resulting clusters, names and 'pen portraits' were developed to aid understanding of their characteristics. This was achieved through observation of their temporal profiles in addition to store attributes (such as type and format) and their geographical distributions (i.e. rural/urbanity). Store attributes were investigated by comparing cluster compositions to the pre-existing HSR store classification, and rural/urban interactions were contextualised using the unified rural urban area classification (RUC; O'Brien, 2016). As outlined in Chapter 3 (Section 3.3.2), this describes the characteristics of small areas across 6 different classes, from the most rural 'Sparse/Remote Villages/Dwellings', to the most urban - 'Large Urban Areas'. Cluster names were decided based on the combination of these attributes.

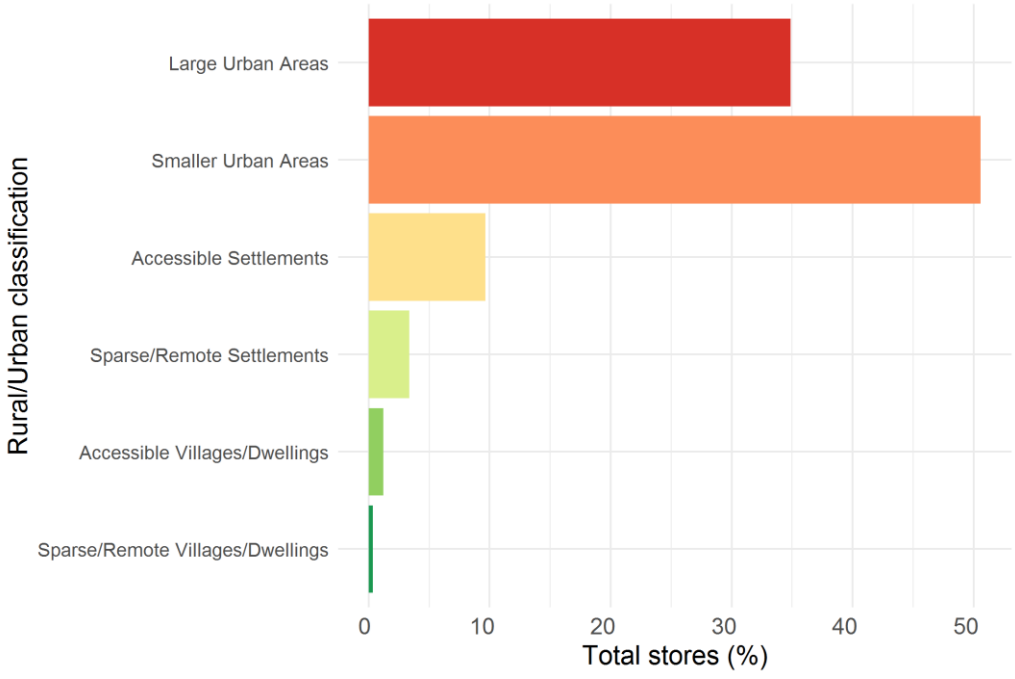
When interpreting characteristics, rate calculation was required to account for underlying variations in both the volume of stores per type and per rural/urban area. To illustrate, Figure 5.5 shows a) the total number of stores in the HSR network per type, b) total store formats and c) the total number of stores per rural/urban classification group.



a)



b)



c)

Figure 5.5: The total number of stores per a) HSR store type, b) HSR store format and c) RUC type.

The HSR network consisted of substantially more small high street and chemist high street stores in urban areas. Therefore, when interpreting the characteristics *between* clusters, values were normalised by the total frequency of types overall. Similarly, for interpreting rural/urban characteristics, values were normalised by the total existing stores per classification group. This was in order to understand which area types/store types may be most prominent in each group, independent of underlying volumes.

5.3. Results

5.3.1. Temporal Profiles – Supergroups

Table 5.1 outlines the overall distance to centroid measures. Figures 5.6- 5.10 show a) radial plots, describing the mean profile (cluster centroids) for each time point and, b) boxplots, demonstrating the distribution of stores from their cluster centroid, illustrating the extent of variation within clusters.

Table 5.1: Distance to centroid measures per Supergroup.

Cluster	**Distance measures (SED)**				
	Mean	Maximum	Minimum	No. stores above mean	No. store below mean
1	6.3	12.4	1.5	47.3%	52.7%
2	6.4	12.8	0.8	47.3%	52.7%
3	6.2	13.7	1.3	42.2%	57.8%
4	4.7	10.9	0.7	48.8%	51.2%
5	3.1	8.1	0.7	47.8%	52.3%

As can be observed in Figures 5.6- 5.10, Supergroups demonstrated unique fluctuations in weekday and weekend consumption patterns. Supergroup 1 exhibited increased activity at off-peak periods (i.e. mid-morning and afternoon) on weekdays and weekend morning peaks. Supergroup 2 demonstrated similar patterns during weekdays and weekends (primarily midday to afternoon), yet much higher overall weekend consumption (primarily midday to afternoon). Supergroup 3 demonstrated similar patterns to Supergroup 1, however could be differentiated by opposing magnitudes of consumption on weekdays and weekends (Supergroup 3 showed highest transactional activity on weekdays and low weekend activity).

Supergroup 4 demonstrated patterns suggestive of a working population on weekdays (i.e. peaks in the morning, lunch time and evening) and significantly decreased consumption on weekends. Finally, Supergroup 5 could be differentiated from other clusters by high levels of activity during both weekdays and weekends, although showed higher overall consumption on the weekend with peak times in the afternoon and evenings. Whilst Supergroup 1 and Supergroup 3 both exhibited similar temporal patterns, the differentiation of magnitude between weekday and weekend consumption was considered an important an aspect to retain when testing the outcomes of different cluster solutions. In addition, alternative solutions (i.e. testing of 3-6 K) did not reduce these to a single cluster, suggesting that the number of groups specified did not influence the separation of these Supergroups.

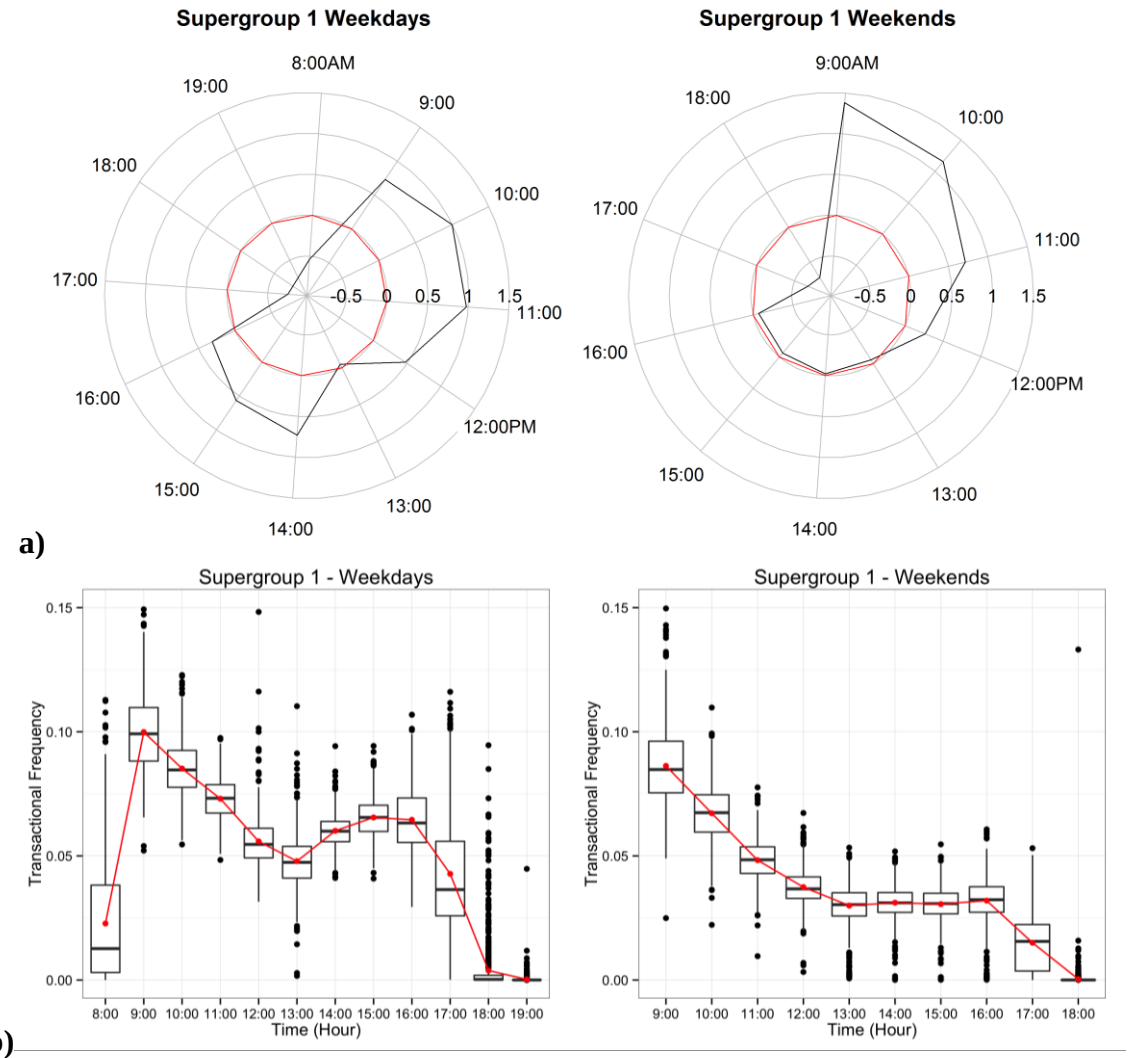


Figure 5.6: Radial plots and boxplots (showing cluster centroids and the distribution of stores from this centre) for *Supergroup 1*.

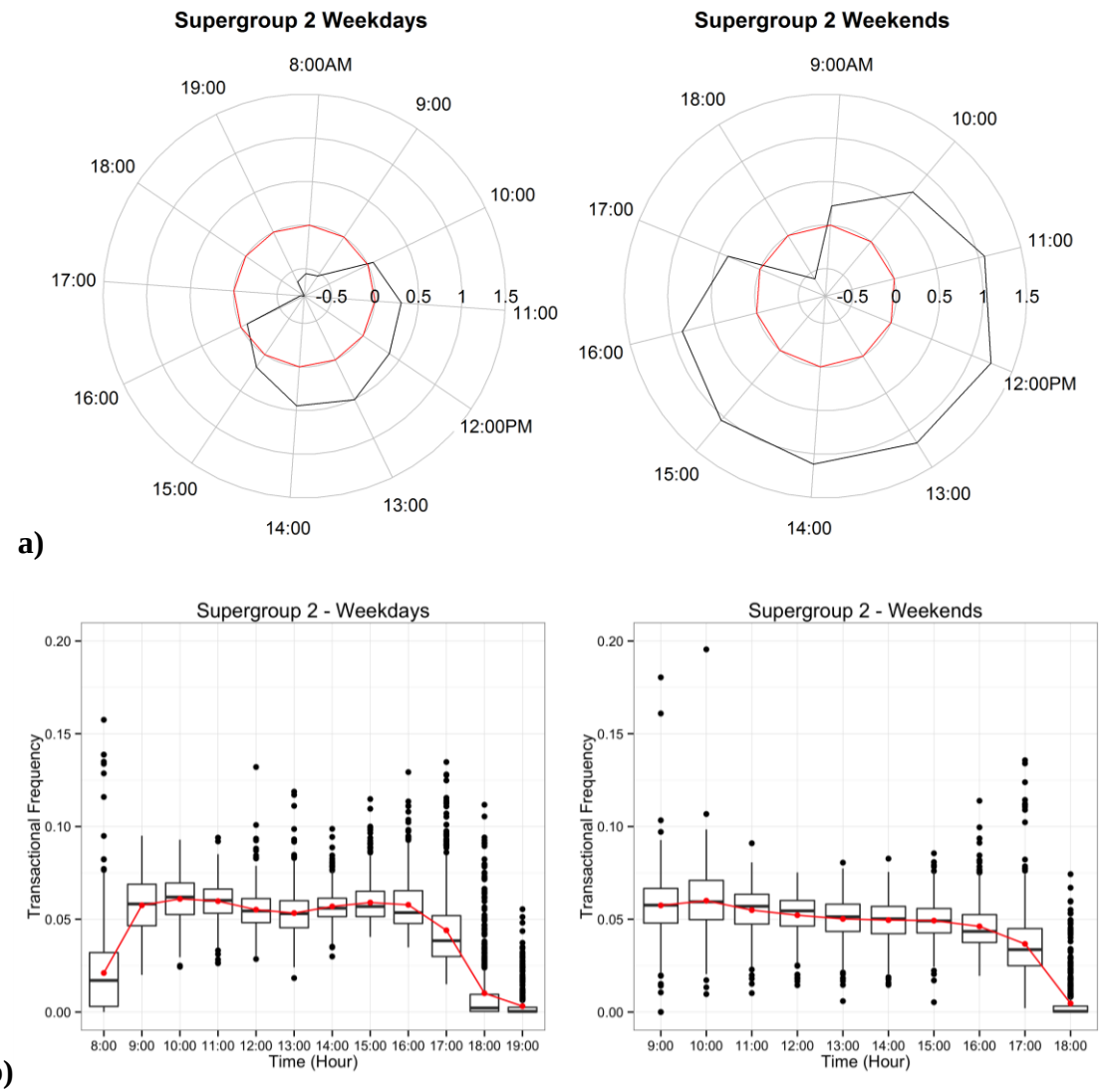


Figure 5.7: Radial plots and boxplots (showing cluster centroids and the distribution of stores from this centre) for *Supergroup 2*.

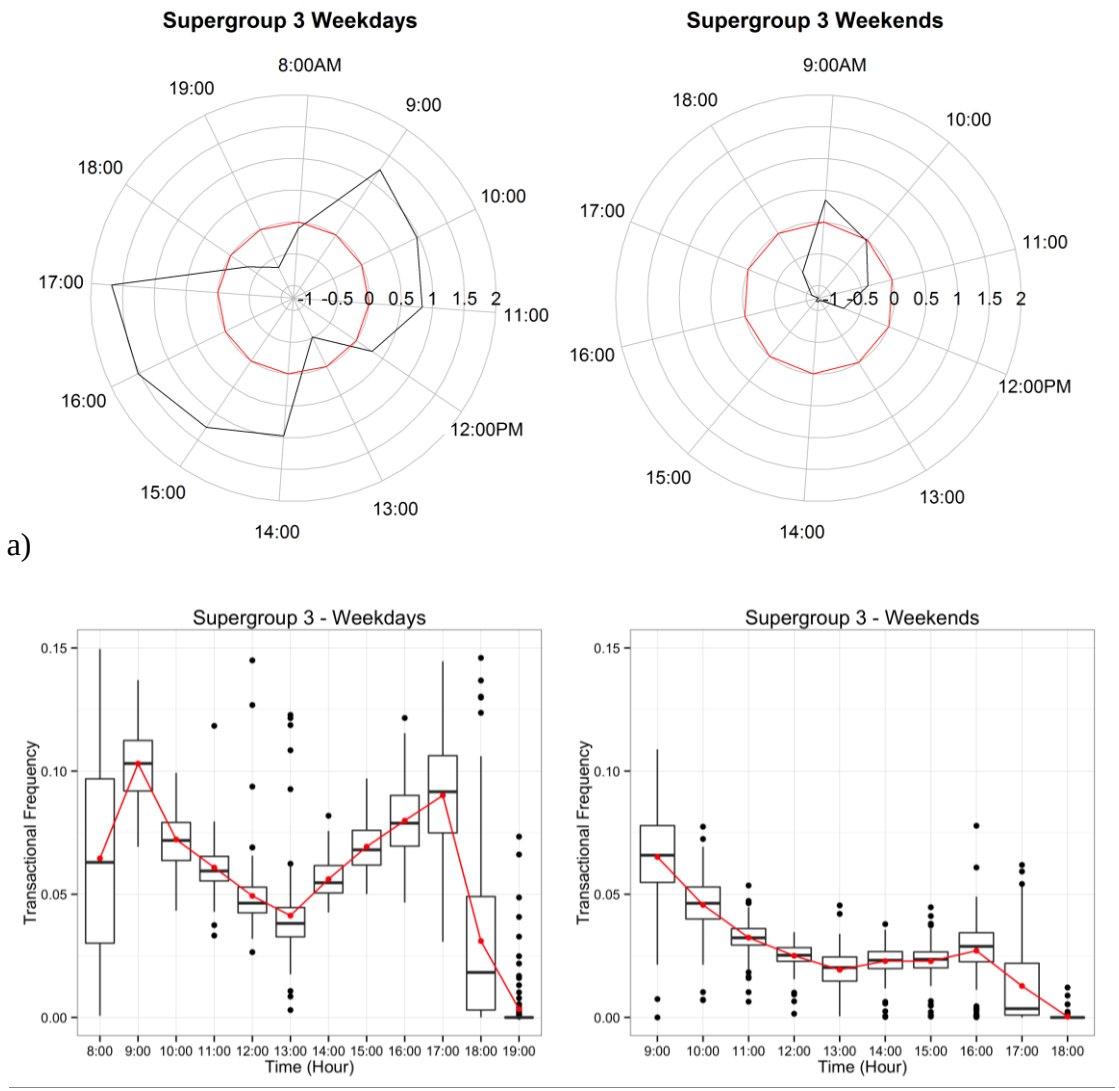
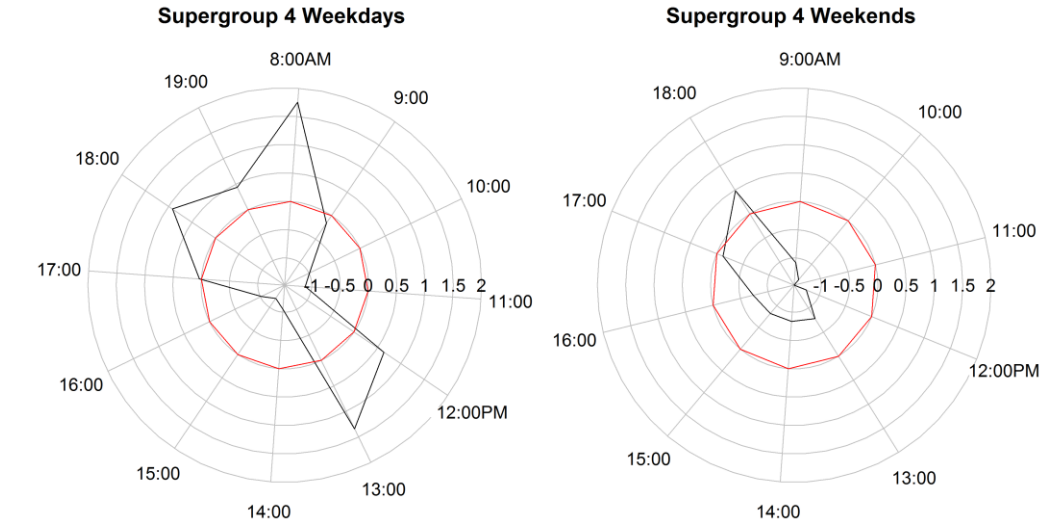
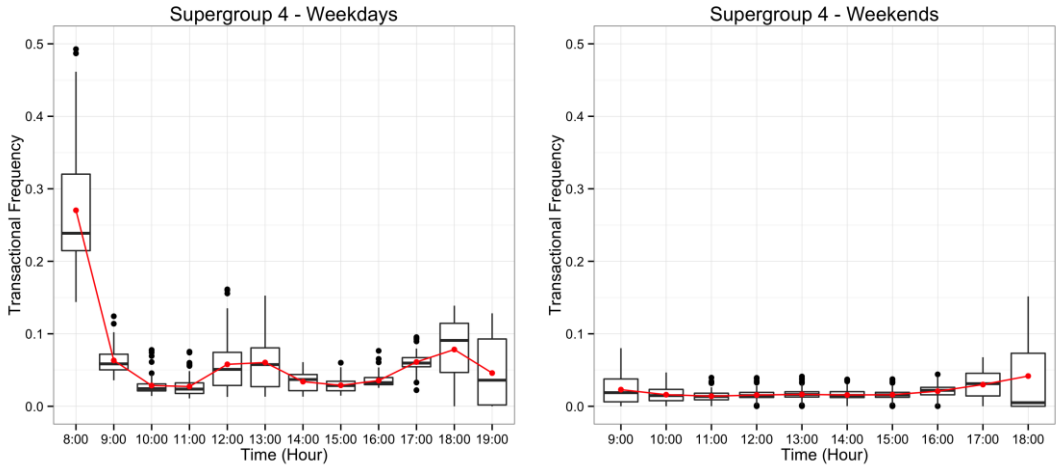


Figure 5.8: Radial plots and boxplots (showing cluster centroids and the distribution of stores from this centre) for *Supergroup 3.*

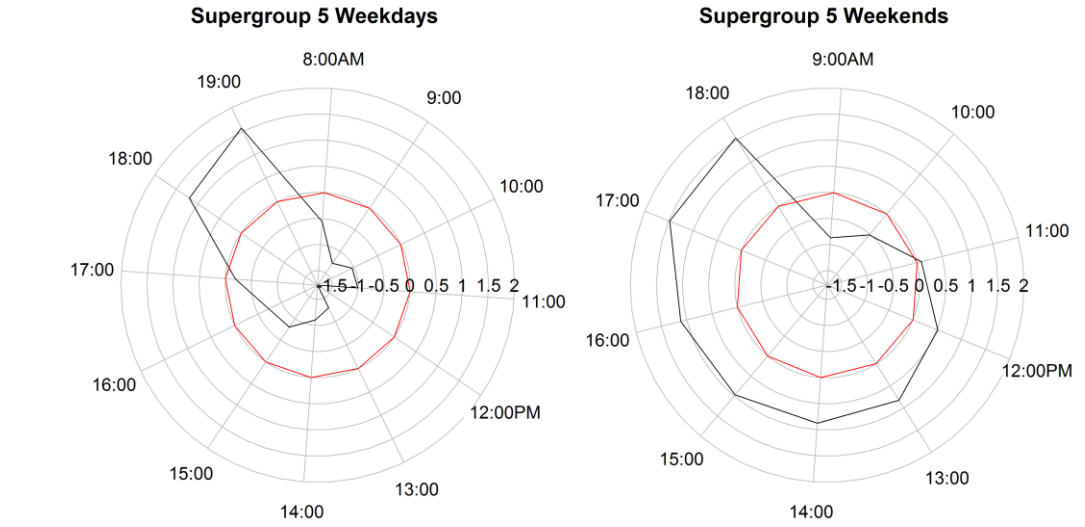


a)

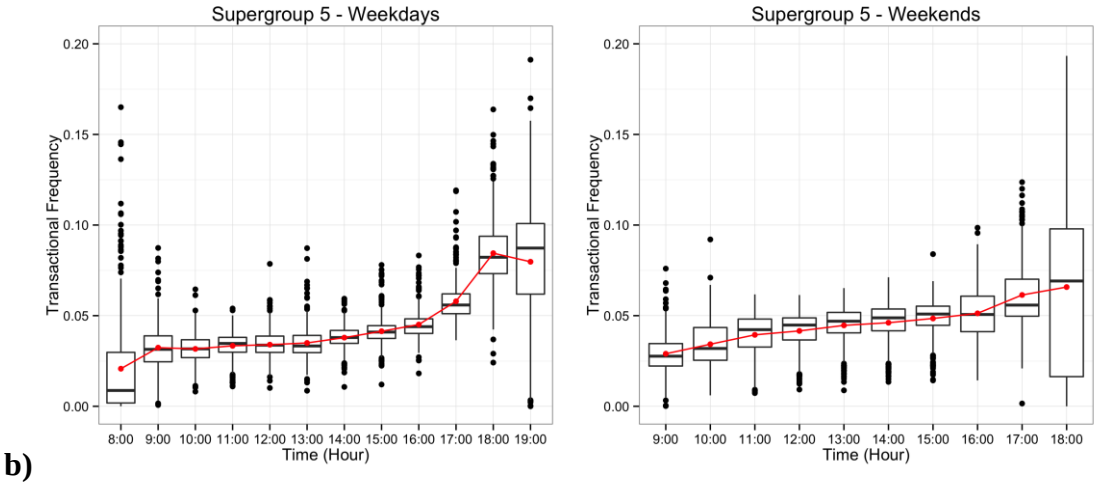


b)

Figure 5.9: Radial plots and boxplots (showing cluster centroids and the distribution of stores from this centre) for *Supergroup 4*.



a)



b)

Figure 5.10: Radial plots and boxplots (showing cluster centroids and the distribution of stores from this centre) for *Supergroup 5.*

5.3.2. Supergroup Descriptions

Combining temporal profiles with store metadata and analysing their spatial attributes provided useful context for profile interpretations, demonstrating that temporal consumption patterns were able to segment stores into groups that exhibited similar types, sizes, formats and locational attributes. Descriptions of Supergroup attributes are outlined in Table 5.2. These were derived from the temporal profiles, spatial attributes and retail compositions of stores, of which are illustrated for each Supergroup in Figure 5.11 and Figures 5.12-5.21. The resulting Supergroups can be summarised as follows:

- **Supergroup 1 ('General Off-peak Shopping')** - contained the highest proportion of chemist high street and small high street stores. These stores were primarily pharmacy format and accounted for the majority of rurally located HSR stores. This was the largest Supergroup, which is likely reflective of the HSR's store structure, as these store types make up the largest proportion of their market. This Supergroup also contained a relatively large segment of community stores, however, only those located in predominantly rural locations.
- **Supergroup 2 ('Weekend Peak Destinations')** - typically described larger stores including 72.5% of all HSR large high streets. These were 80% health and beauty format. Locations were of a rural/urban mix, in town centres of varying size.
- **Supergroup 3 ('Weekday Off-peak Shopping')** - contained the largest proportion of community stores of all Supergroups, however, only those located in predominantly suburban/urban locations. These were 98% pharmacy format and locations were predominantly within 'Accessible Settlements', 'Smaller Urban Areas' and 'Large Urban Areas'.
- **Supergroup 4 ('Weekday Convenience')** - contained 60% of all convenience high street stores and 90% of convenience travel stores. These were primarily of health and beauty format, and were located almost exclusively in 'Large Urban Areas'.
- **Supergroup 5 ('Stable Destinations')** - contained the largest stores including the highest proportion of HSR flagships and 94% of all HSR out-of-town retail parks. These were primarily located in 'Large Urban Areas' and 'Smaller Urban Areas', however, this Supergroup also included a segment of more rural retail park stores (in 'Accessible Dwellings').

Table 5.2: Store Supergroup descriptions.

Supergroup	No. Stores	Description
1 **General Off-peak Shopping** The most rural Supergroup, serving a pharmacy/health and beauty mix, primarily off-peak hours during the week and early risers on weekend.	902	**Store attributes** • Predominantly small high streets, chemists and rural community stores. • 66.5% pharmacy focused, 33.5% health and beauty. **Temporal profile** • Contrasting weekday-weekend patterns. • *Weekdays* - late morning and afternoon peaks. • *Weekends* – early morning peak. **Spatial profile** • Rural/urban mix, but contained the largest proportion of rural stores.
2 **Weekend Peak Destinations** Town centre destinations of varying size, serving health and beauty needs primarily on weekends.	572	**Store attributes** • Large and small high street stores. • 80.8% health and beauty format. • Contained 72.5% of all large high street stores. **Temporal profile** • Similar weekday and weekend fluctuations, but much higher volumes on weekends. • Both periods characterised by mid-day activity peaks, but a steady flow of transactions from late mornings to early evenings. **Spatial profile** • Rural/urban mix. • Town centres of varying size.
3 **Weekday Off-peak Shopping**	337	**Store attributes** • Contained the highest proportion of community stores (60.1%), yet predominantly those in more suburban/urban community locations. • 98% pharmacy format.

Stores serving pharmaceutical needs, primarily in off-peak hours during weekdays to local surrounding communities.		**Temporal profile** • Exhibited a similar temporal profile to Supergroup 1, however, contrasting weekday versus weekend patterns (higher volumes on weekdays). • *Weekdays* – late morning/afternoon peaks. • *Weekend*s – early morning peaks. **Spatial profile** • Urban/suburban community areas.
4 **Weekday Convenience** Urban convenience stores primarily serving the weekday working population.	90	**Store attributes** • Contained 60% of all convenience high street stores and 90% of convenience travel stores. • 78% health and beauty format. **Temporal profile** • Contrasting weekday and weekend activity. • *Weekday* – peaks around 8am, lunchtime and evening. • *Weekend* – minimal activity overall, but evening peaks. **Spatial profile** • Primarily urban centres and transport hubs.
5 **Stable Destinations** Large flagship stores characterised by similar demand on both weekdays and weekends.	350	**Store attributes** • Primarily retail parks (94% of all retail park stores), but also a mix of all other store types. • Contained the highest percentage of all flagship stores (63.1%). **Temporal profile** • Differentiated from other Supergroups by similar magnitudes of activity during both weekdays and weekends. • Afternoon to evening peaks during both weekdays and weekends. **Spatial profile** • Predominantly urban, yet also contained a segment of retail parks in 'accessible' rural locations.

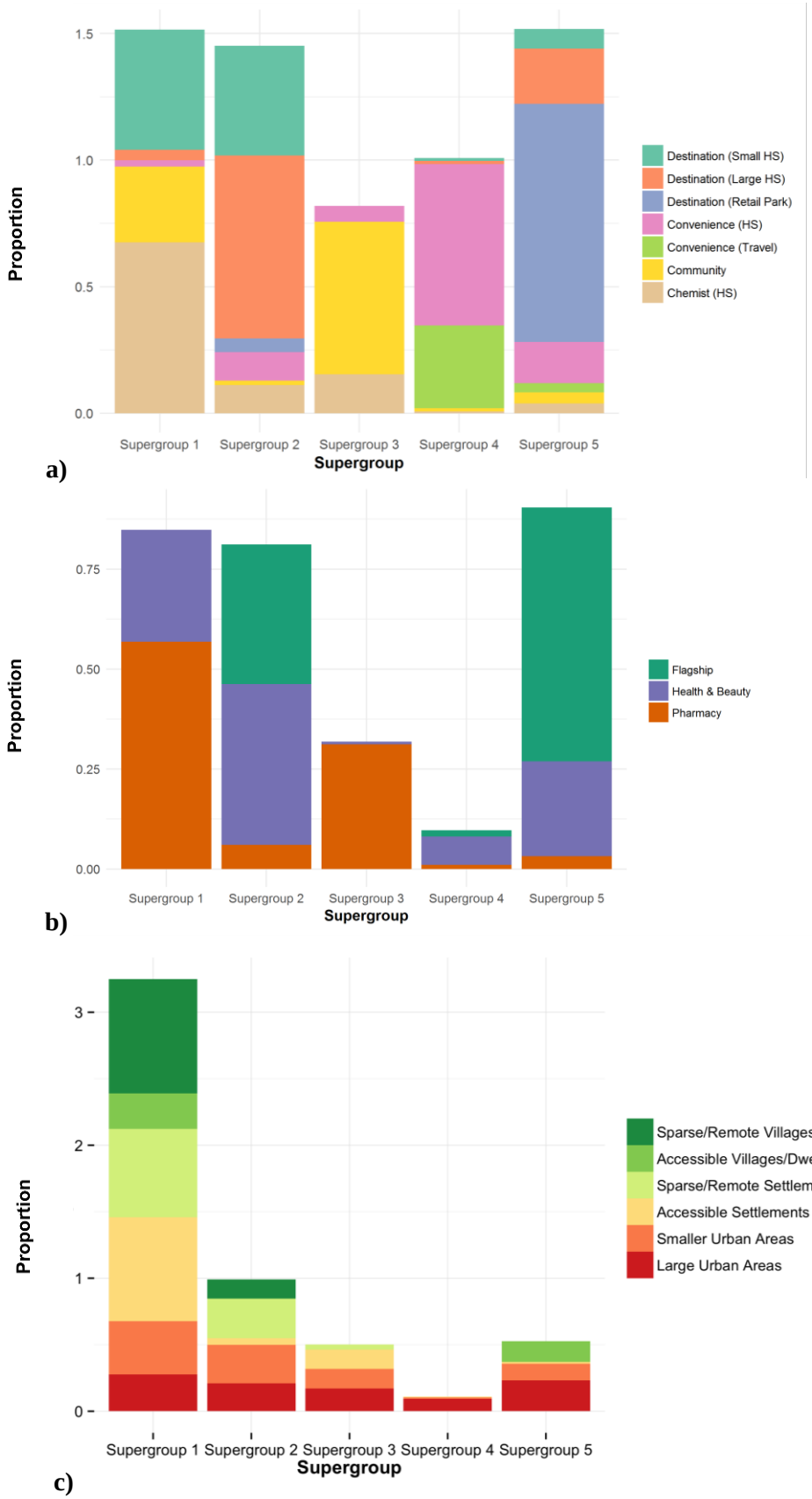


Figure 5.11: Proportion of a) HSR store types, b) HSR store formats and c) RUC location types, per Supergroup.

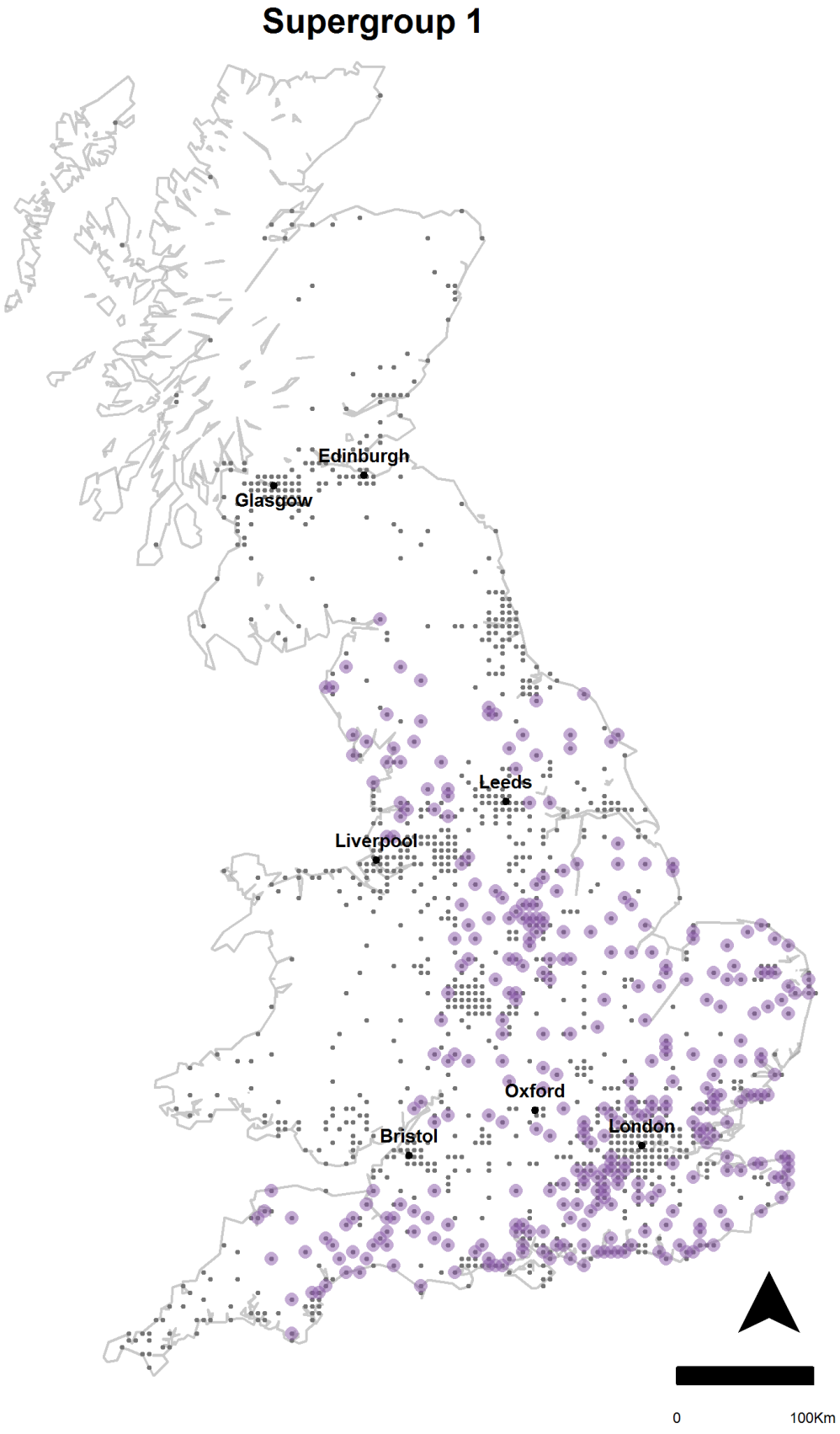


Figure 5.12: Distribution of all HSR stores (shown in grey) and Supergroup 1 ('General Off-peak Shopping', highlighted in purple) across Great Britain (represented by 5km grid cell centres).

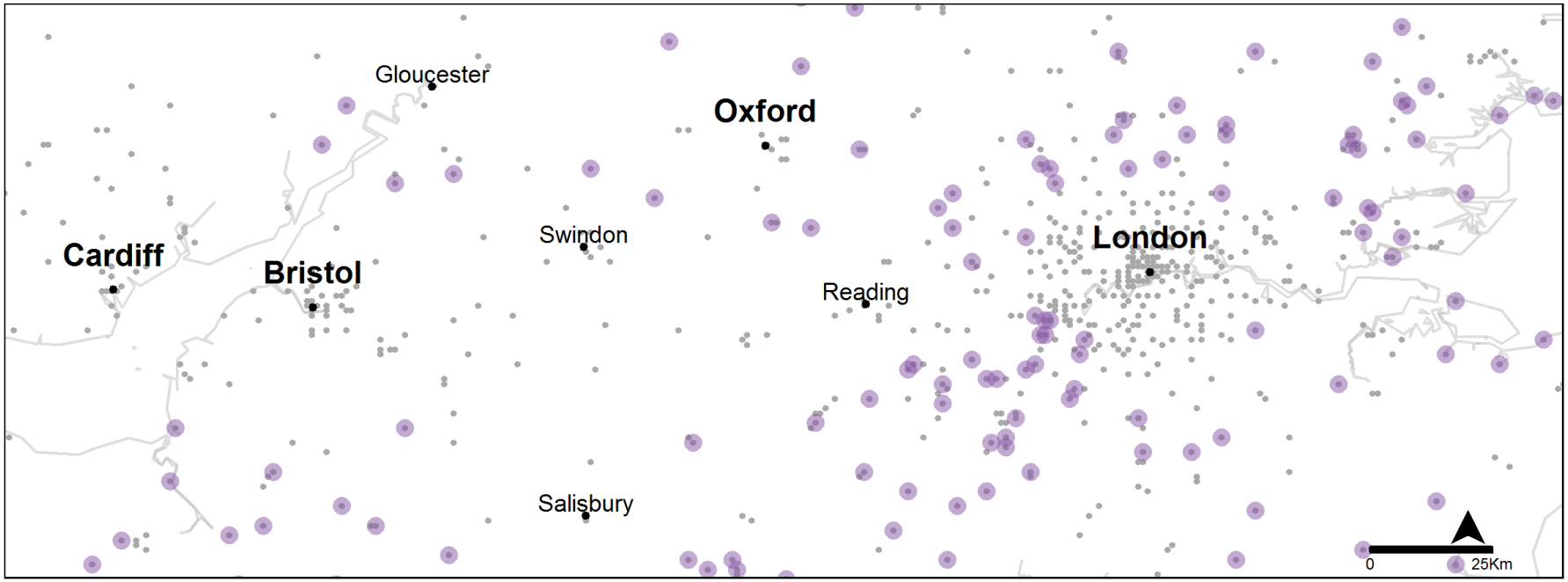


Figure 5.13: Distribution of all HSR stores (shown in grey) and Supergroup 1 ('General Off-peak Shopping', highlighted in purple) across Southern England (represented by 1km grid cell centres).

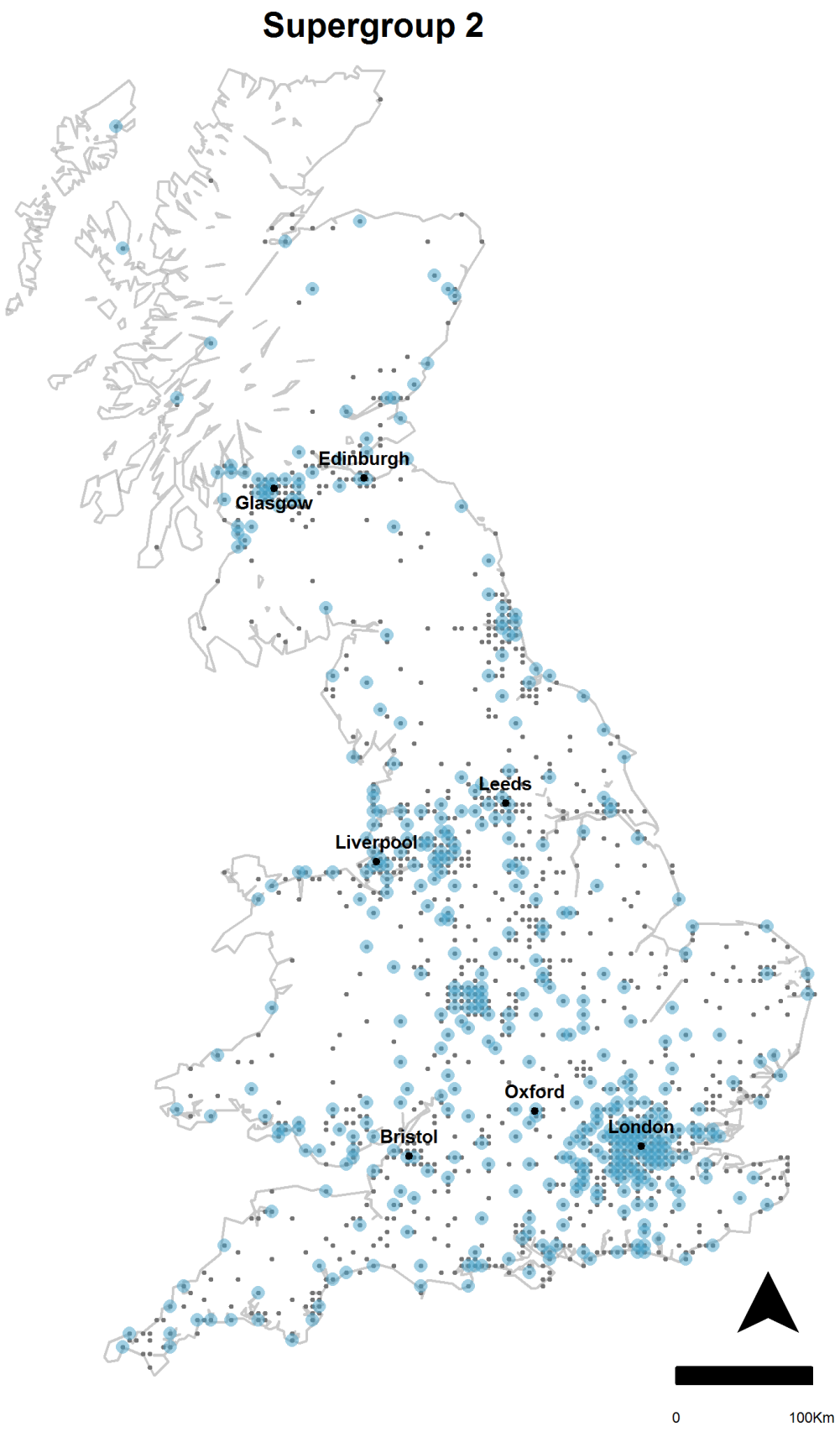


Figure 5.14: Distribution of all HSR stores (shown in grey) and Supergroup 2 ('Weekend Peak Destinations', highlighted in blue) across Great Britain (represented by 5km grid cell centres).

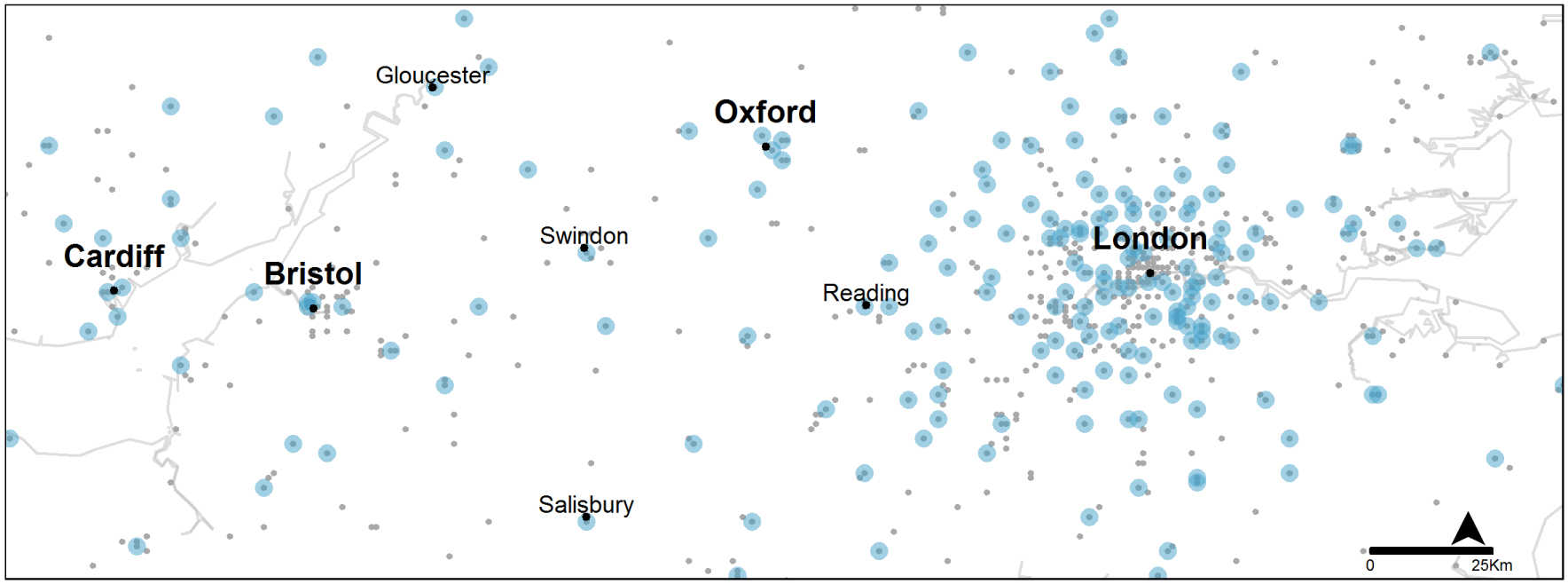


Figure 5.15: Distribution of all HSR stores (shown in grey) and Supergroup 2 ('Weekend Peak Destinations', highlighted in blue) across Southern England (represented by 1km grid cell centres).

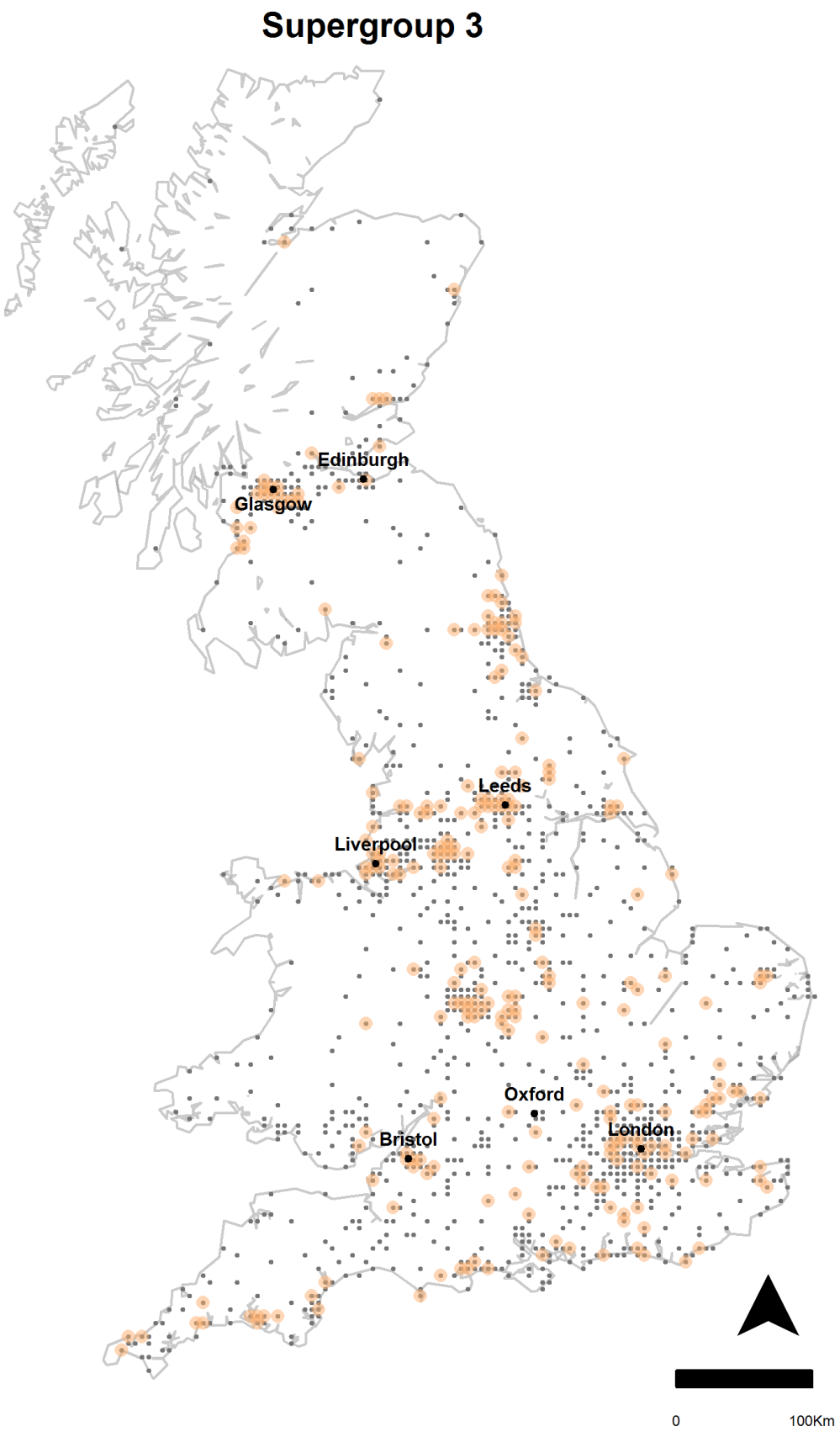


Figure 5.16: Distribution of all HSR stores (shown in grey) and Supergroup 3 ('Weekday Off-peak Shopping', highlighted in orange) across Great Britain (represented by 5km grid cell centres).

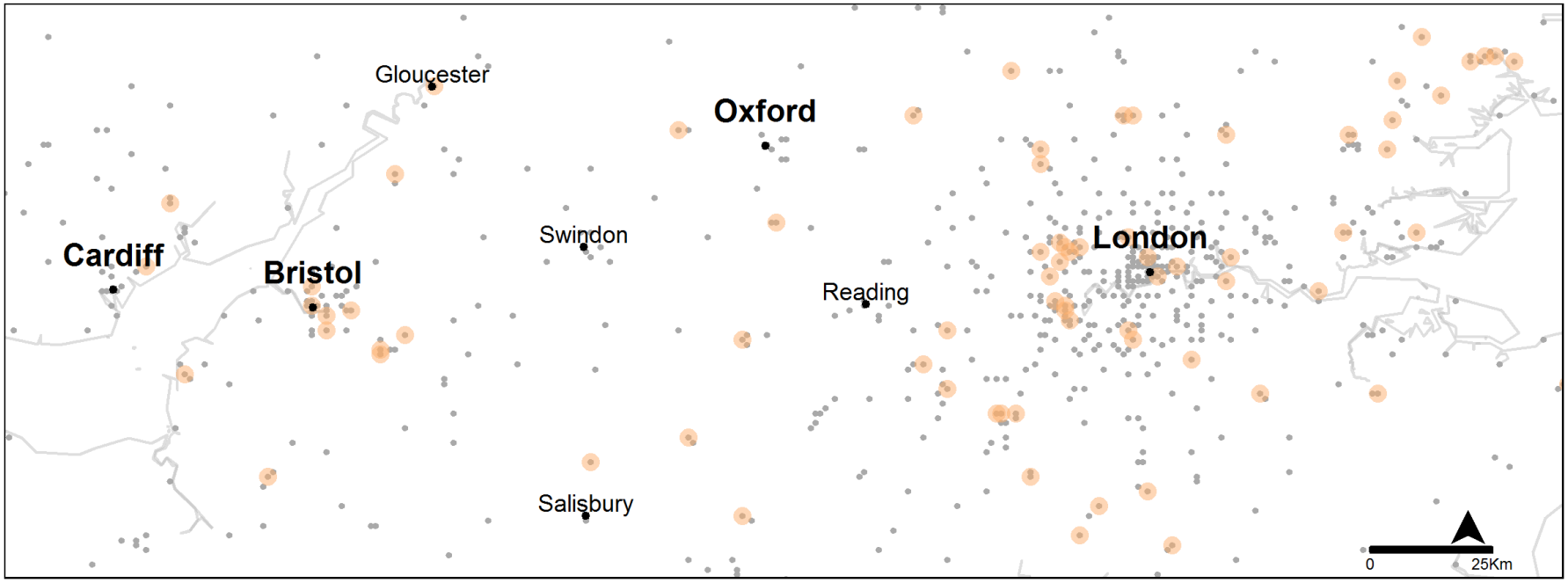


Figure 5.17: Distribution of all HSR stores (shown in grey) and Supergroup 3 ('Weekday Off-peak Shopping', highlighted in orange) across Southern England (represented by 1km grid cell centres).

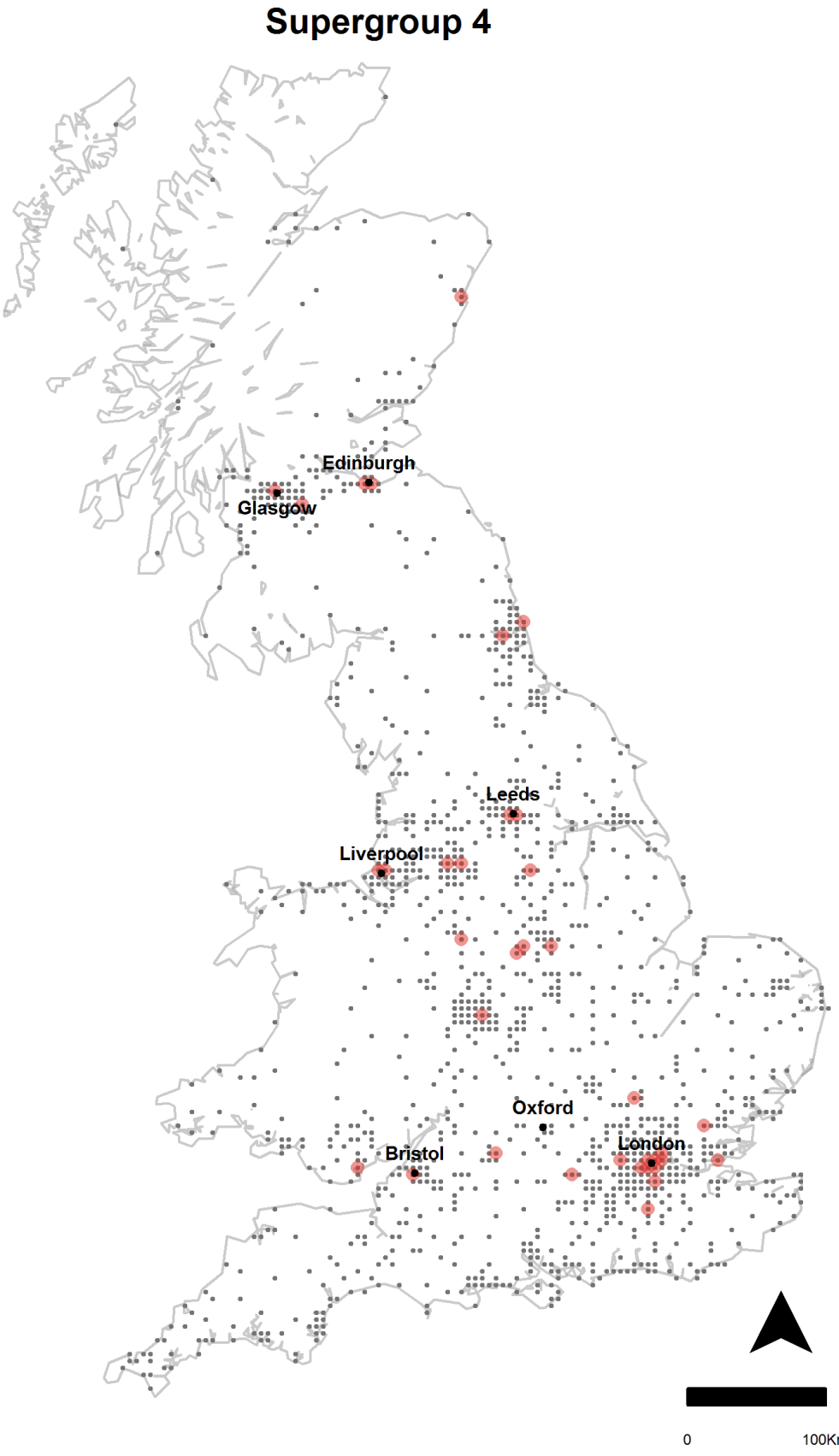


Figure 5.18: Distribution of all HSR stores (shown in grey) and Supergroup 4 ('Weekday Convenience', highlighted in red) across Great Britain (represented by 5km grid cell centres).

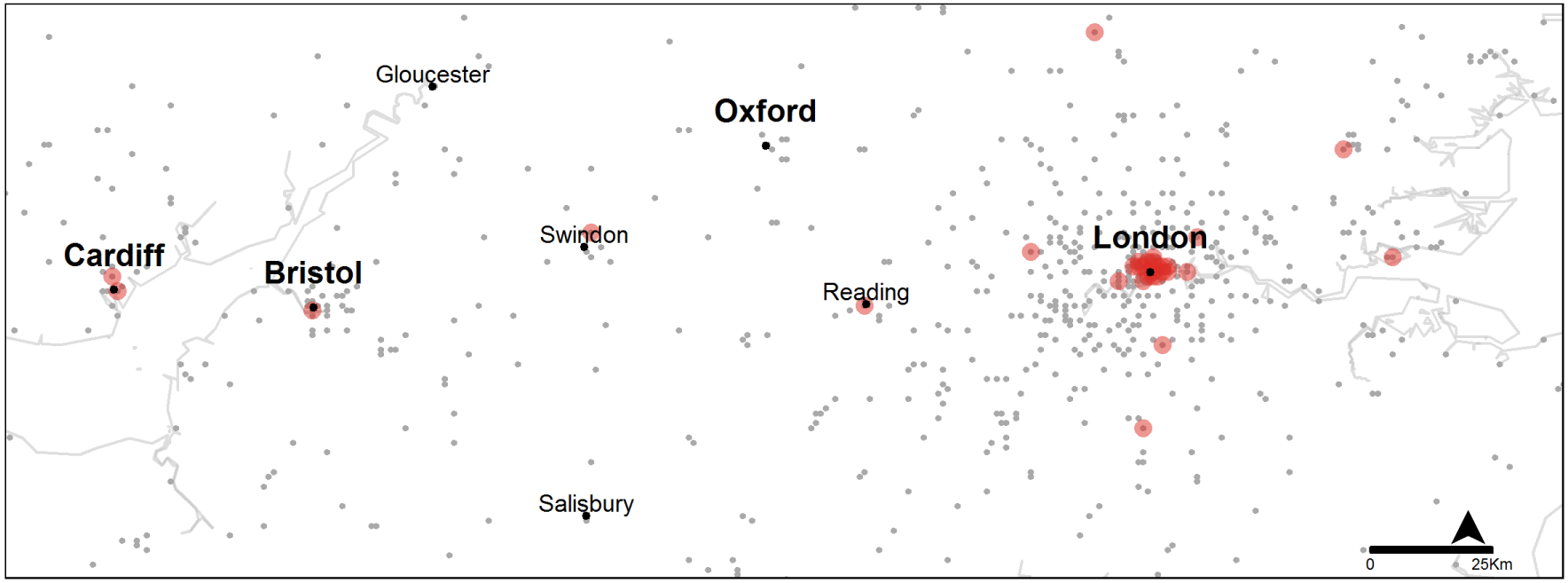


Figure 5.19: Distribution of all HSR stores (shown in grey) and Supergroup 4 ('Weekday Convenience', highlighted in red) across Southern England (represented by 1km grid cell centres).

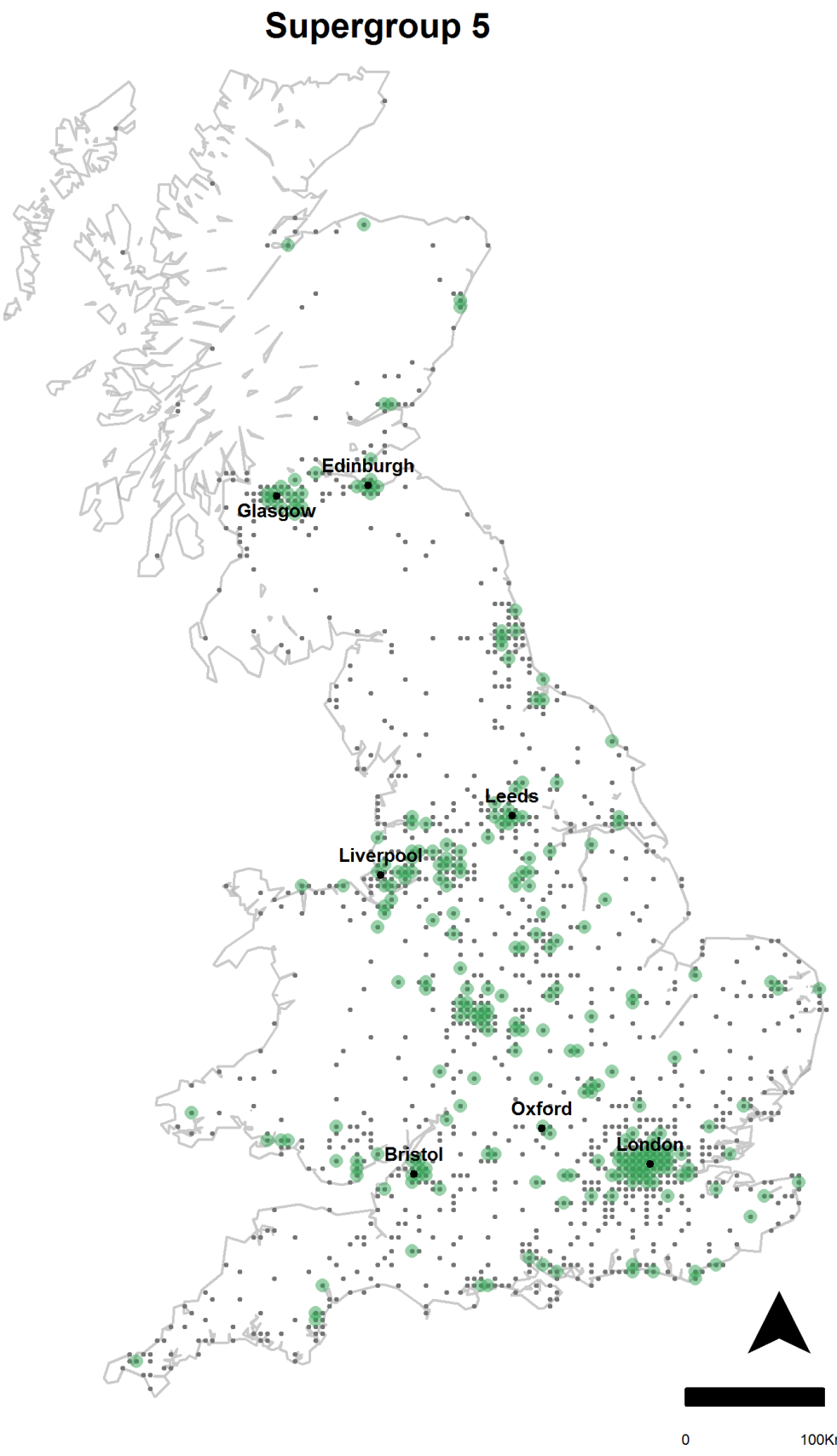


Figure 5.20: Distribution of all HSR stores (shown in grey) and Supergroup 5 ('Stable Destinations', highlighted in green) across Great Britain (represented by 5km grid cell centres).

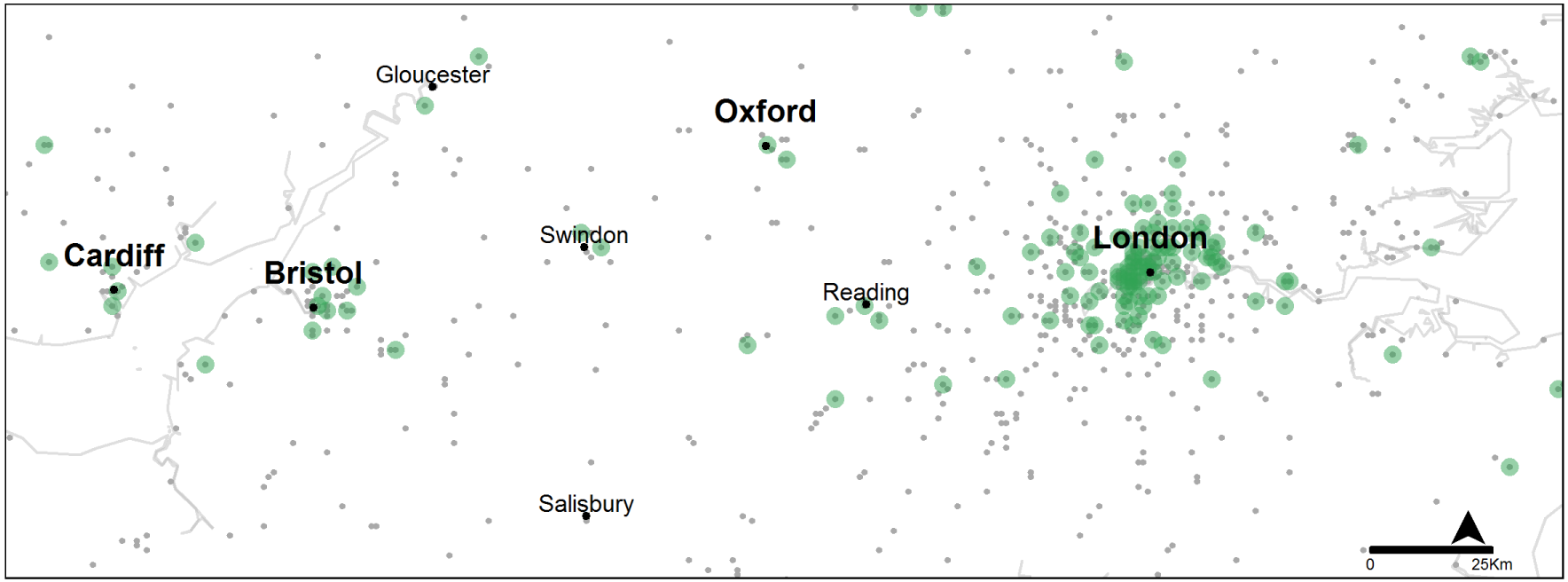


Figure 5.21: Distribution of all HSR stores (shown in grey) and Supergroup 5 ('Stable Destinations', highlighted in green) across Southern England (represented by 1km grid cell centres).

Results suggested that temporal consumption patterns may be indicative of population activities and thus provide insight into the attributes and functions of store locations. For example, Supergroup 4 ('Weekday Convenience') demonstrated a temporal profile typical of a working population (i.e. peaks outside of business hours and minimal weekend activity). The store attributes of this Group were extremely consistent with this observation, consisting predominantly of convenience high street and convenience travel stores, in urban locations where a working population are likely to reside during the day. Supergroup 5 ('Stable Destinations') described stores that demonstrated a similar magnitude of consumption on both weekdays and weekends. This included locations that are workplace oriented yet also prominent retail destinations (such as Oxford Circus, London) and retail park locations that demonstrate destination shopping during both periods. Conversely, Supergroup 2 ('Weekend Peak Destinations') showed the most geographically distributed stores, with a mix of rural and urban locations. However, these stores were all located in town centres and exhibited the same store attributes (primarily larger, health and beauty oriented) with increased weekend consumption. This indicated that these stores may fulfil a particular function to consumers, such as weekend 'destination' shopping trips to local town centres by the surrounding population.

These temporal variations were also, to some extent, able to segment stores by their spatial attributes. For example, notwithstanding the inherent bias of store locations towards urban areas, the vast majority of rurally located stores were assigned to Supergroup 1 (and to a lesser extent, Supergroup 2). This suggests a strong correlation between the temporal rhythms and the characteristics of locations. A further example is evident from the assignment of rural 'Community' type stores exclusively to Supergroup 1, and the more urban 'Community' stores to Supergroup 3. These observations suggest that we may be able to differentiate places, and identify their characteristics, by the temporal flows of consumption they exhibit. They also suggest that we may be able to enrich descriptions of the functions of different locations to consumers, in comparison to utilising only locational and physical store characteristics. For example, despite HSR defined store types showing obvious uniformities in their assignment to clusters, many stores were differentially classified by this analysis in comparison to the HSR groupings. For example, a mix of store types were assigned to the convenience cluster here (Supergroup 4), suggesting that utilising locational characteristics alone may not be sufficient in delineating whether a store fulfils a weekday convenience function to consumers. These instances are explored in more detail in the following Group level analysis.

5.3.3. Temporal Profiles - Groups

Based on WCSS observations and iteratively testing the outcomes of 2-4 clusters, the optimum number of Group clusters obtained was between 2 and 3 for each Supergroup. Table 5.3 illustrates distance to centroid measures. The most homogeneous clusters belonged to

Supergroups 4 and 5. The least homogeneous were the Groups derived from Supergroup 3, in particular, Group 3a, again reflecting the higher variation within this Supergroup. However, the stores in this Supergroup still demonstrated unique temporal patterns.

Table 5.3: Distance to centroid measures per Group.

Supergroup	Group	Distance Measures (SED)				
		Mean	Maximum	Minimum	No. stores above mean	No. store below mean
1	**a**	9.4	18.3	3.2	46.4%	54.6%
	b	8.4	18.2	4.6	41.5%	58.5%
	c	6.8	11.4	3.8	46.3%	54.7%
2	**a**	7.4	18.4	1.2	50.0%	50.0%
	b	10.6	19.4	3.1	48.6%	51.4%
	c	7.0	13.9	2.2	44.0%	56.0%
3	**a**	13.4	32.1	6.8	29.5%	70.5%
	b	9.2	23.6	4.3	52.2%	47.8%
4	**a**	7.3	20.0	2.9	44.1%	56.9%
	b	6.8	16.2	2.4	47.4%	52.6%
5	**a**	7.7	14.7	4.1	56.8%	44.2%
	b	4.7	10.2	1.8	46.1%	54.9%
	c	7.1	10.5	4.1	46.3%	53.7%

Combining Groups with store metadata and observing spatial attributes was similarly useful for interpretation of these cluster profiles, of which are summarised in Table 5.4. Figures 5.22 to 5.36 illustrate the radial plots for each Group (describing the mean profile, or cluster centroid, for each time point), their spatial attributes and retail compositions. This demonstrated that the 10-minute interval data were able to further segment stores based on their temporal variations, characteristics/functions and location types. For ease, the attributes of each Group are summarised as follows:

- **Supergroup 1 ('General Off-peak Shopping')**
    - **Group 1a ('Off-peak Late Risers')** - These stores showed mid-day to afternoon peak consumption patterns similarly across weekdays and weekends. This Group contained the majority of health and beauty oriented small high street stores of this Supergroup, but also a mix of HSR store types including convenience high street and large high street stores located in a mix of urban and rural areas.
    - **Group 1b ('Off-peak Early Risers')** – exhibited morning peaks across both weekdays and weekends and contained predominantly pharmacy oriented (92%) 'Community' and 'Chemist' stores. This Group contained the highest proportion of rural store locations.
    - **Group 1c ('General Off-peak Activity')** - exhibited almost identical store types to Group 1b (97% pharmacy format), but were differentiated by both

weekday morning and afternoon peaks (low mid-day trade) and a higher proportion of urban store locations.

- **Supergroup 2 ('Weekend Peak Destinations')**
  - **Group 2a ('Weekend Destinations – Late Risers') -** characterised by peak consumption patterns on weekend afternoons. Contained the majority of large high street stores and the highest proportion of retail park stores in this Supergroup, located in medium-large sized towns. These were predominantly health and beauty format stores (86.7%).
  - **Group 2b ('Weekend Destinations – Early Risers')** – characterised by peak consumption patterns on weekend mornings. These were primarily small high street, health and beauty focused stores (92.8%) and represented the largest Group (reflective of the higher number of overall HSR small high street stores).
  - **Group 2c ('Weekend Destinations - General Activity')** – The most rural segment of this Supergroup, that demonstrated peak consumption patterns on weekend mornings, yet also weekday late mornings. These stores were primarily within medium-sized town centres and of chemist and community types, serving a mix of pharmacy (54.3%) and health and beauty (44.8%) needs.
- **Supergroup 3 ('Weekday Off-peak Shopping')**
  - **Group 3a ('Weekday Early Risers')** – characterised by weekday morning peaks and low consumption on weekends. Primarily community stores and chemists (94.8% of all stores were 'Pharmacy' format) in suburban/urban locations.
  - **Group 3b ('General Weekday Activity')** – Characterised by consistent morning to evening activity weekdays. Primarily urban community chemists (100% 'Pharmacy' format).
- **Supergroup 4 ('Weekday Convenience')**
  - **Group 4a ('Commuter Convenience') -** Characterised by peak consumption between 8-9am on weekdays, yet also increased lunchtime and evening consumption. Comprised of the highest proportion of convenience travel stores located in 'Accessible Settlements', urban areas and transport hubs.
  - **Group 4b ('General Convenience') -** These stores demonstrated general convenience usage during weekdays (morning, lunchtime, and evening peaks). Primarily convenience high street stores yet contained a mix of store types. These were located exclusively in 'Smaller Urban Areas' and 'Large Urban Areas'.

- **Supergroup 5 (‘Stable Destinations’)**
  - **Group 5a (‘Stable Destinations - Late Risers’)** – characterised by afternoon/evening peaks both on weekdays and weekends. These were primarily retail park stores (yet contained a mix of store types), located in small and large urban areas but also ‘Accessible Settlements’.
  - **Group 5b (‘Stable Destinations – Early Risers’)** – characterised by late morning peaks both on weekdays and weekends. A large proportion of this group consisted of retail park stores in ‘accessible’ rural locations.
  - **Group 5c (‘Stable Urban Destinations’)** – These stores showed convenience usage during weekdays (morning, lunchtime, and evening peaks) and a similar magnitude of activity on weekends, primarily mid-morning. Stores were a mix of all types and formats, however, primarily convenience types and large high streets. These stores are likely used for convenience on weekdays and destination shopping on weekends. They were also exclusively located in large urban areas.

Table 5.4: Store Group descriptions.

Group	No. Stores	Description
1a **Off-peak Late Risers**	458	**Store attributes** • Predominantly chemists and small high street stores (contained almost all small high street stores in Supergroup 1). These were predominantly health and beauty oriented (60.7%). **Temporal profile** • Transactional volumes were similar across weekdays and weekends. • *Weekdays* – midday-afternoon peaks. • *Weekends* – midday-afternoon-evening peaks. **Spatial profile** • Rural/urban mix.
1b **Off-peak Early Risers**	212	**Store attributes** • Predominantly chemists (75% chemists, 18.9% community stores). • 92% pharmacy format. **Temporal profile** • Similar transactional volumes across weekdays and weekends. • Morning peaks during both periods. **Spatial profile** • Highest proportion of rural stores in Supergroup 1.
1c **General Off-peak Activity**	232	**Store attributes** • Predominantly chemists (71.1% chemists, 27.2% community stores). • 97% pharmacy format. **Temporal profile** • Morning (9 and 10 am) and evening peaks, during both weekdays and weekends. • Characterised by low volumes midday.

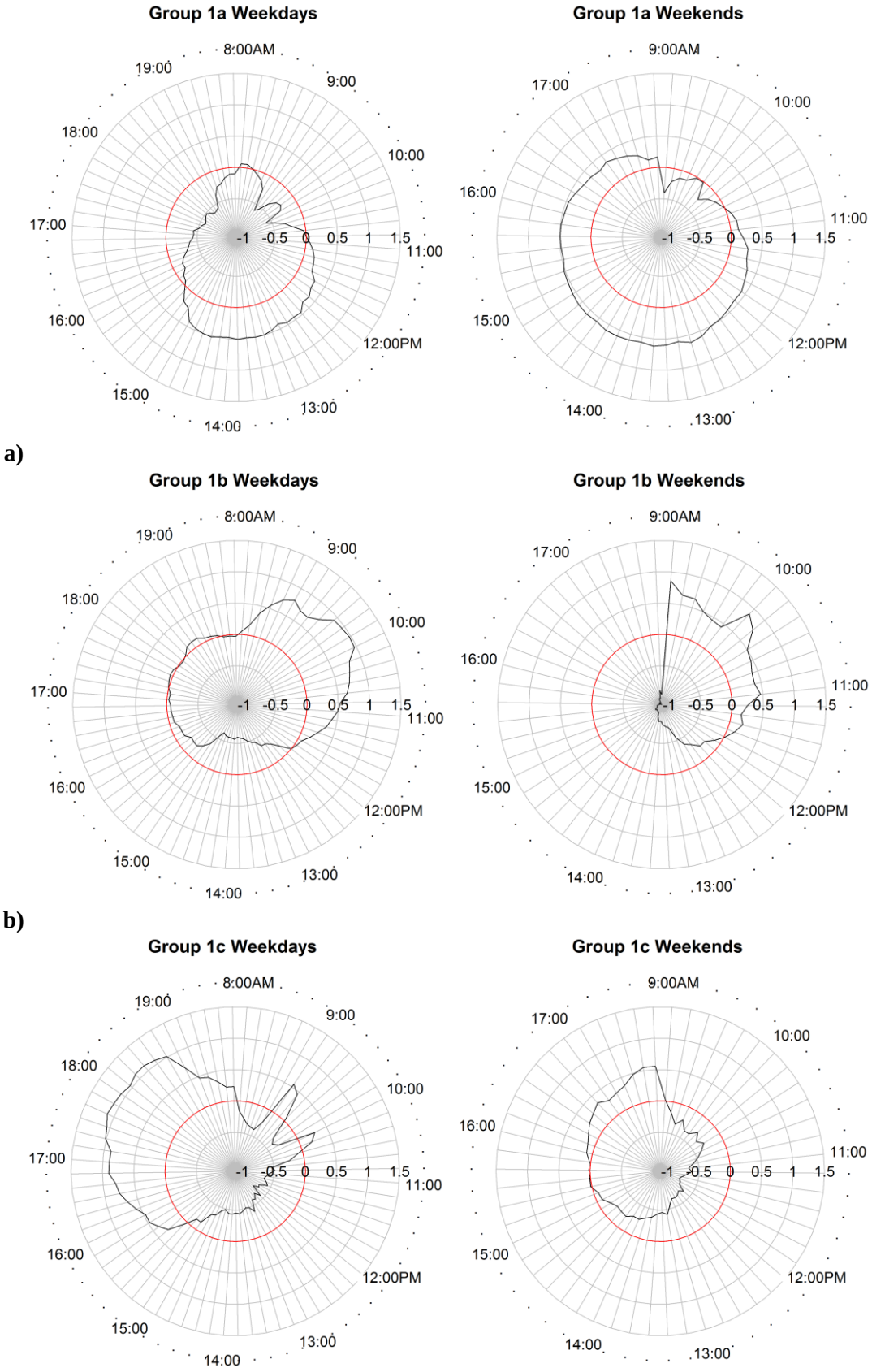
		Spatial profile • Contained a higher proportion of urban locations types than others in this Supergroup.
2a **Weekend Destinations – Late Risers**	137	**Store attributes** • Primarily large high streets (51.1%). • Health and beauty focused (85.7%). **Temporal profile** • Higher volumes weekend. • *Weekdays* – some evidence of convenience usage (morning, lunch, evening peaks). • *Weekends* – afternoon/evening peaks. **Spatial profile** • Medium-large sized towns.
2b **Weekend Destinations – Early Risers**	319	**Store attributes** • Primarily small high streets (54.5%), and large high streets. • Health and beauty focused (92.8%). **Temporal profile** • Higher volumes weekends. • *Weekday* – morning peak (9 and 10am) and midday. • *Weekend* – morning peak. **Spatial profile** • Small-medium sized towns.
2c **Weekend destination - general activity**	116	**Store attributes** • The most pharmaceutically oriented of this Supergroup (57.8% chemists). • Health and beauty (44.8%) and pharmacy mix (54.3%). **Temporal profile** • Higher volumes weekends. • Characterised by low volumes midday. • *Weekday* – morning peaks (9 and 10am), afternoon and evening.

		• *Weekend* – morning and evening peaks. **Spatial profile** • The most rural Group of Supergroup 2 (i.e. largest proportion located in 'Sparse/Remote Villages/Dwellings').
3a **Weekday Early Risers**	152	**Store attributes** • Primarily chemists (49%) and community stores (47.7%). • 94.8% pharmacy format. **Temporal profile** • Higher weekday activity/ low weekend activity. • *Weekdays* – morning peaks (9, 10am) substantially higher than all other periods. **Spatial profile** • Suburban/urban locations.
3b **General Weekday Activity**	185	**Store attributes** • 74% community stores, 100% pharmacy format. • Highest percentage of community stores of all Groups. **Temporal profile** • Higher weekday activity. • *Weekday* - Steady flow of transactions – mid-morning to evenings. • *Weekend* – morning to midday peak. **Spatial profile** • Urban locations.
4a **Commuter Convenience**	44	**Store attributes** • Primarily convenience high street stores (61.4%) and travel stores (31.8%). • Contained 70% of all travel stores. • Predominantly health and beauty focused (93.2%). **Temporal profile**

		• Higher volumes weekday, minimal activity weekends. • Weekday – morning peaks (around 8am and 9am). **Spatial profile** • 'Accessible Settlements', urban areas and transport hubs.
4b **General Convenience**	46	**Store attributes** • Primarily convenience high street stores. However, also comprised of chemists, community, and small/large high street stores (that likely function as convenience stores during weekdays). **Temporal profile** • Higher volumes weekdays. • Morning peak 9-930.am. **Spatial profile** • Exclusively in urban areas.
5a **Stable Destinations – late risers**	156	**Store attributes** • Primarily retail park stores. • 80% health and beauty format. • Contained the highest percentage of flagships across all Groups. **Temporal profile** • Slightly higher weekend activity volumes. • Afternoon/evening peaks during both weekdays and weekends. **Spatial profile** • Urban areas but also 'Accessible Settlements'.
5b **Stable Destinations – Early Risers**	153	**Store attributes** • Predominantly retail park stores. **Temporal profile** • Similar transactional volumes both during weekdays and weekends. • Late morning peaks during both periods.

		Spatial profile • Urban and ‘accessible’ rural locations.
5c **Stable Urban Destinations**	41	**Store attributes** • Locations likely used for convenience on weekdays and destination shopping on weekends. • Mix of all store types and formats, but predominantly convenience high streets and large high streets. • 65% health and beauty format. **Temporal profile** • Weekday convenience trends (morning, lunchtime and evening peaks). • General activity weekends (peak mid-morning). **Spatial profile** • Large urban areas.

Supergroup 1



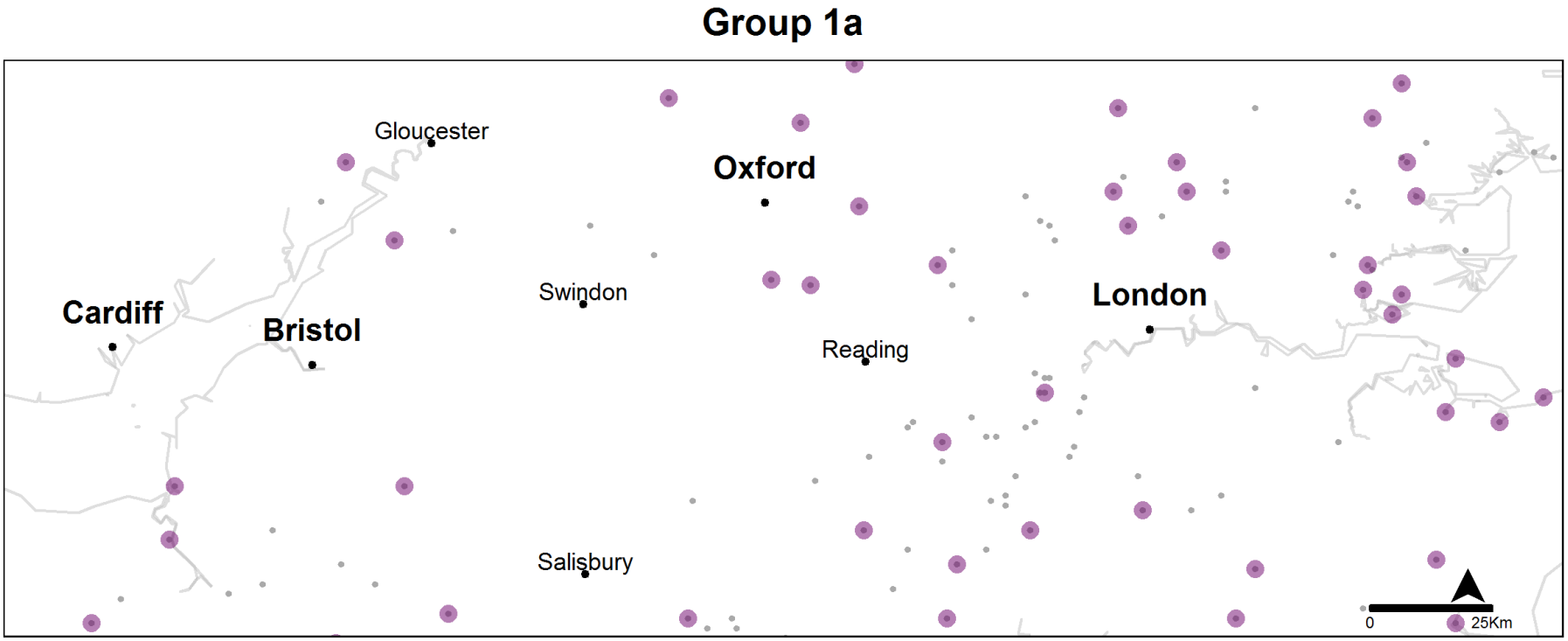
a)

b)

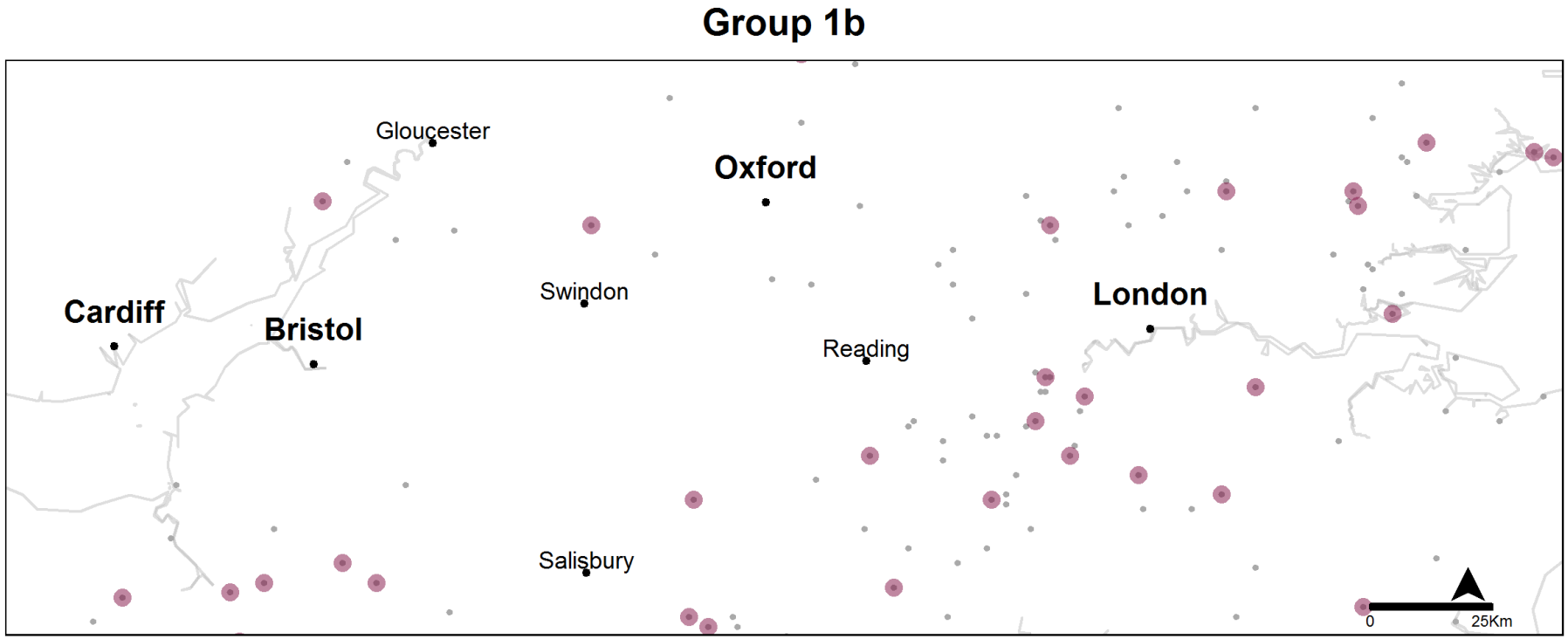
c)

Figure 5.22: Radial plots for a) *Group 1a*, b) *Group 1b* and, c) *Group 1c*.

a)


Group 1a
Gloucester
Oxford
Cardiff
Bristol
Swindon
London
Reading
Salisbury
0
25Km

b)


Group 1b
Gloucester
Oxford
Swindon
Cardiff
Bristol
London
Reading
Salisbury
0
25Km

c)

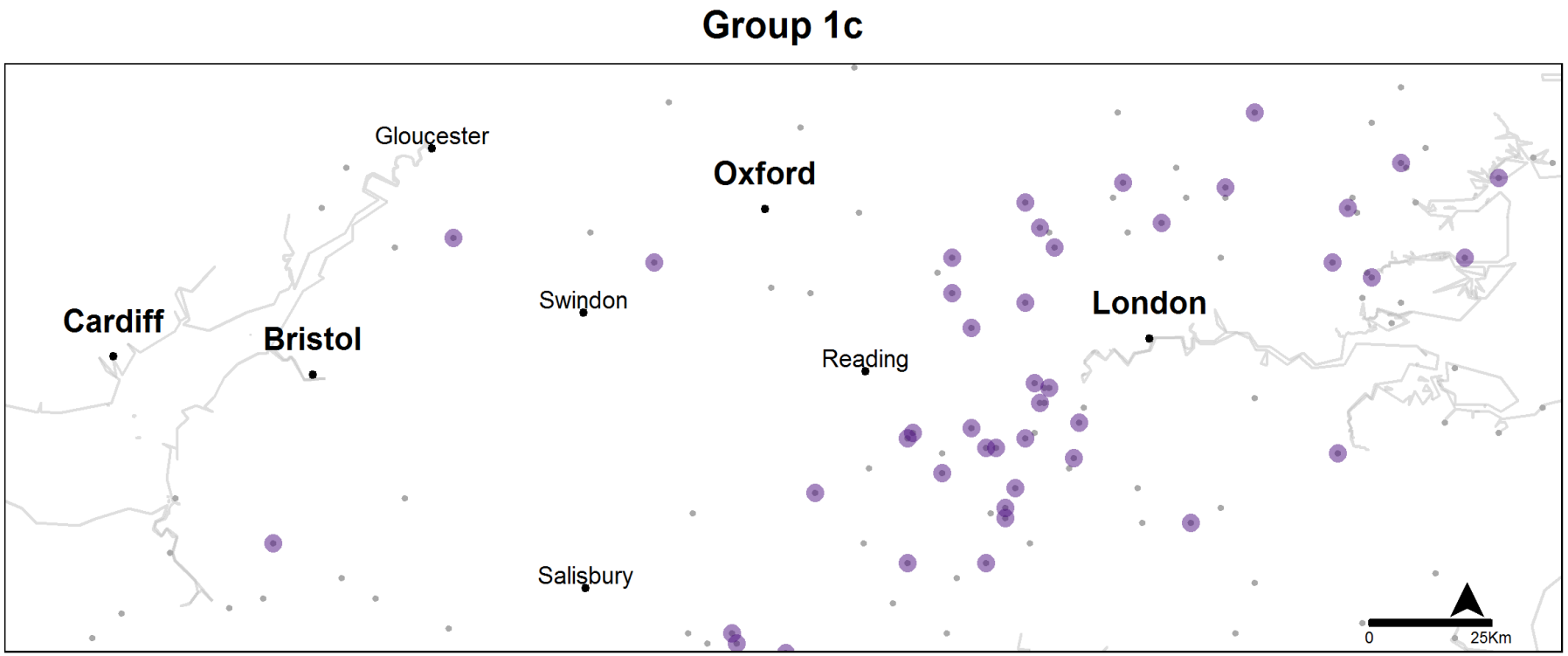
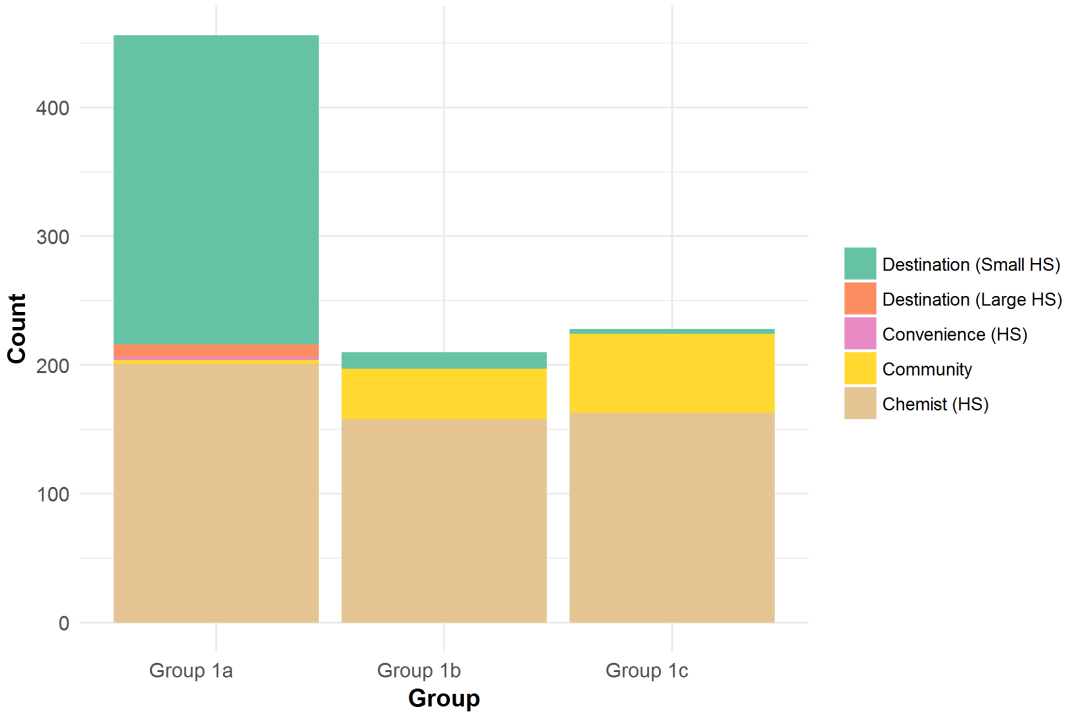
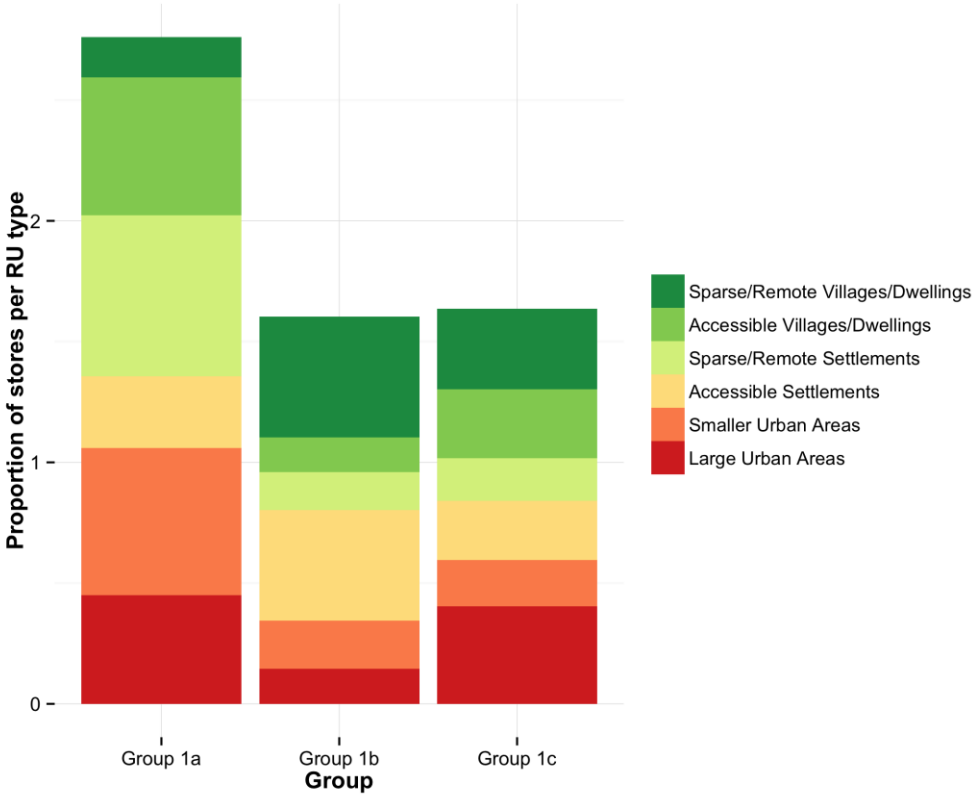


Figure 5.23: Distribution of all Supergroup 1 stores (shown in grey) and a) Group 1 - 'Off-peak Late Risers', b) Group 1b – 'Off-peak Early Risers', and c) Group 1c - 'General Off-peak Activity' across Southern England (represented by 1km grid cell centres).



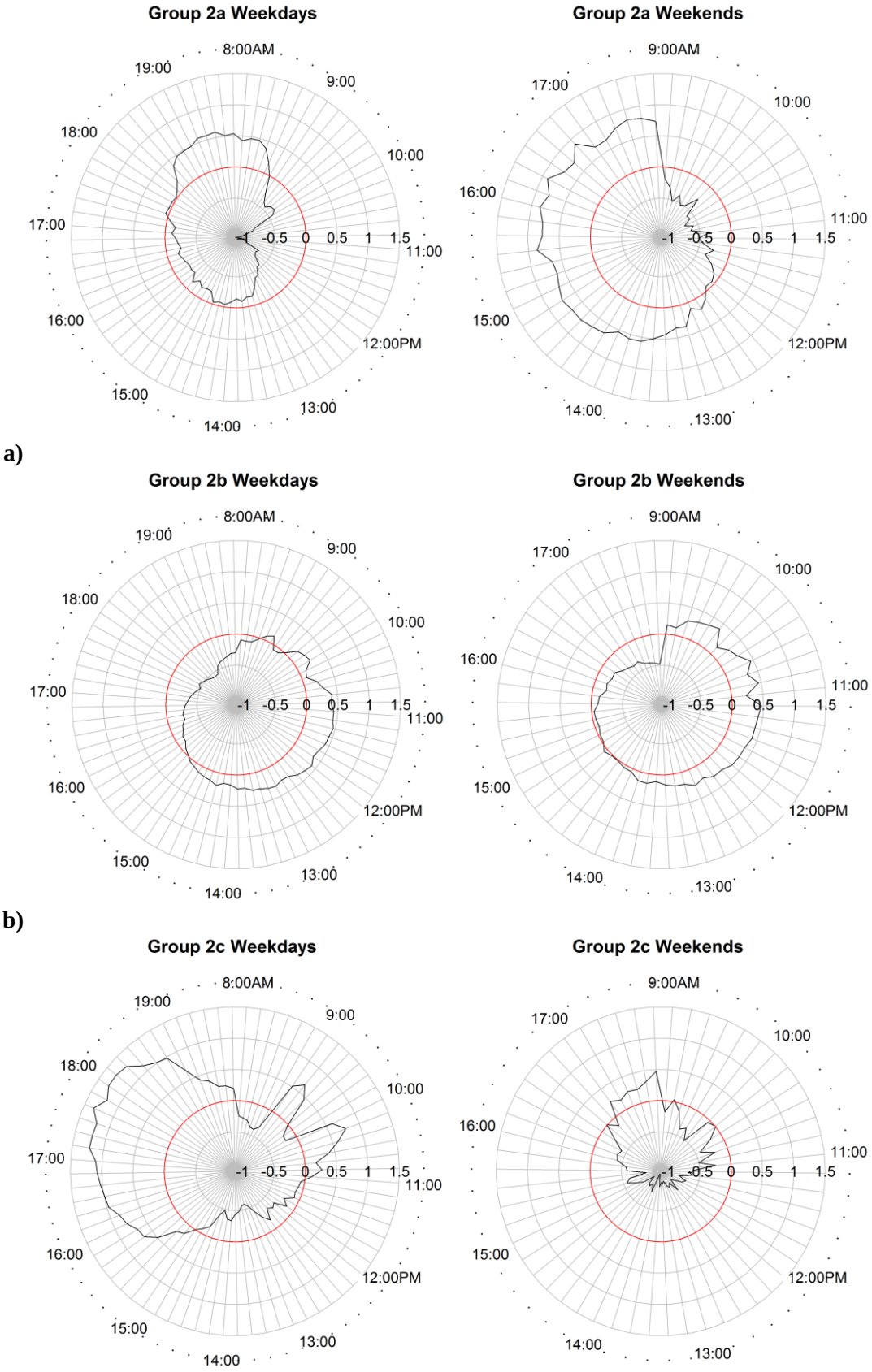
a)



b)

Figure 5.24: Supergroup 1, a) store type counts per Group and b) proportion of rural/urban store locations per Group (normalised by total stores per RUC type in Supergroup 1).

Supergroup 2

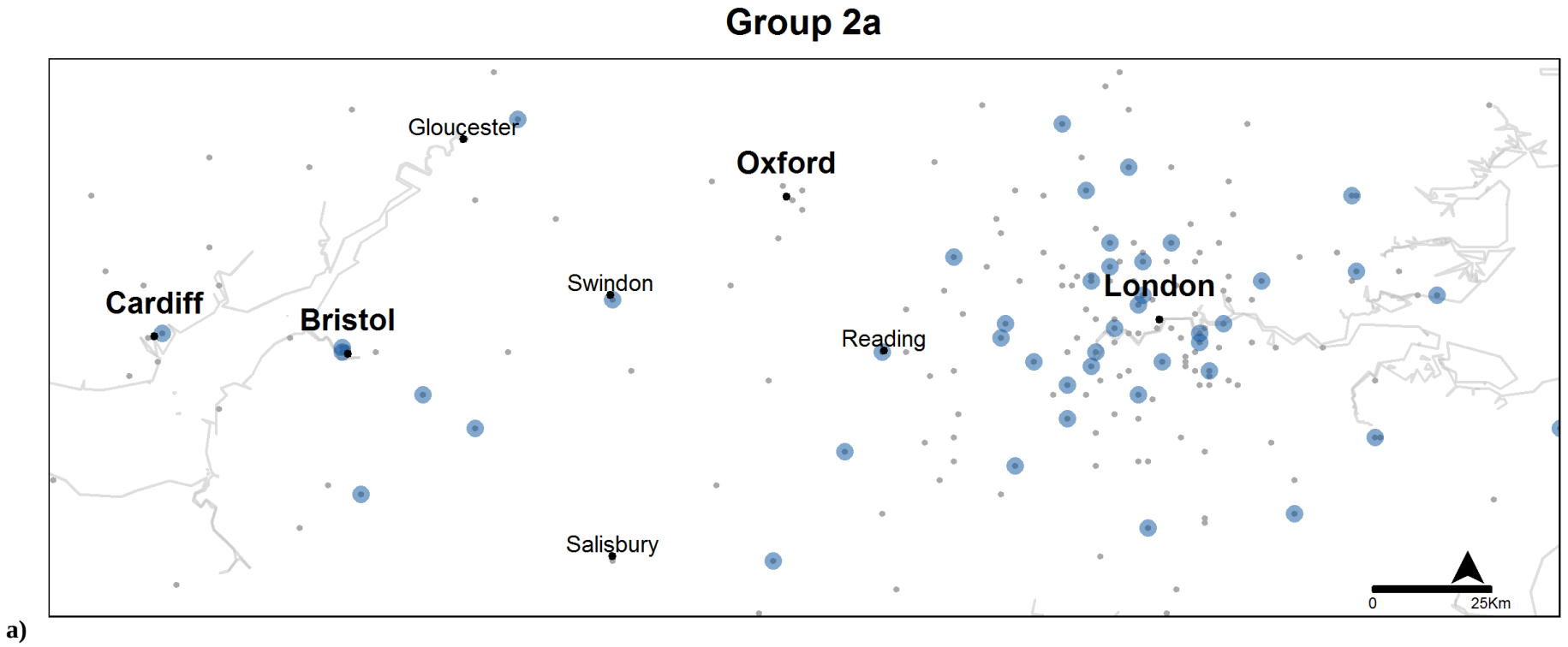


a)

b)

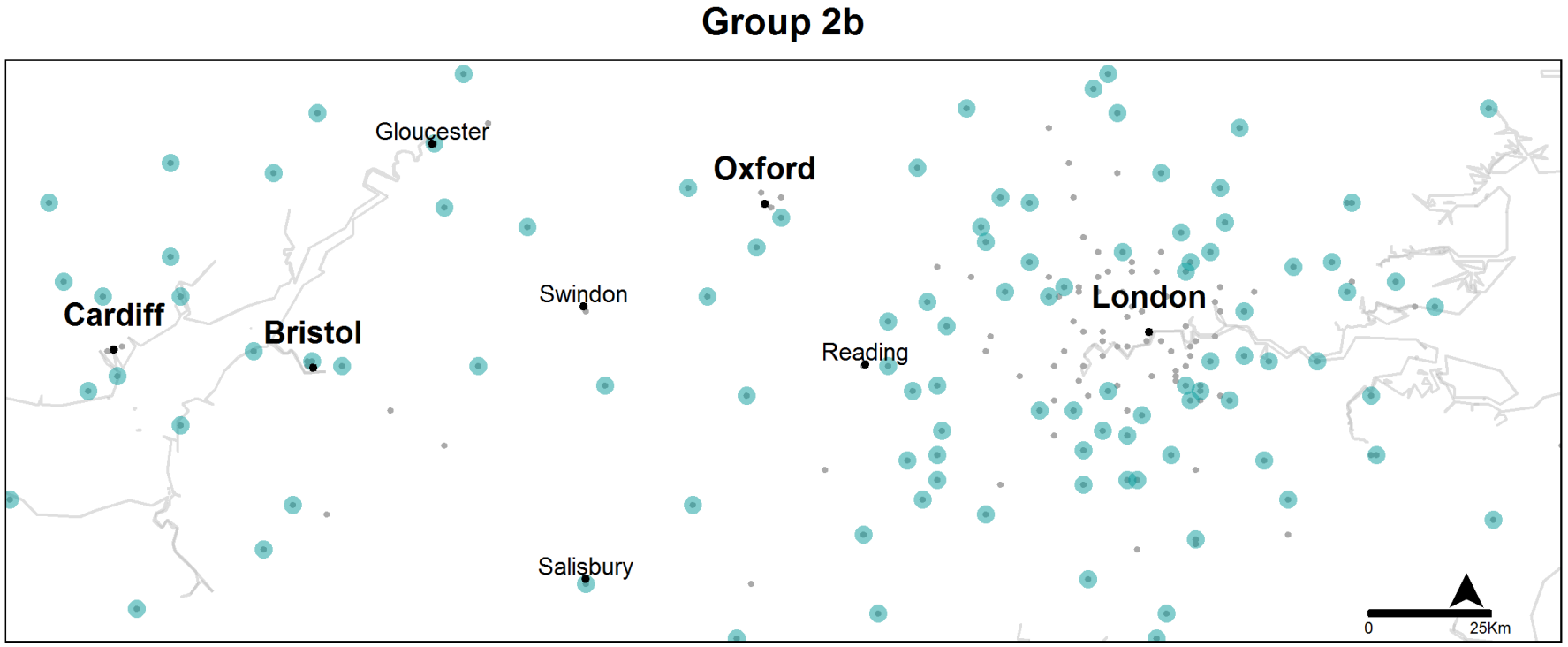
c)

Figure 5.25: Radial plots for a) *Group 2a*, b) *Group 2b* and, c) *Group 2c*.


Group 2a
Gloucester
Oxford
Swindon
Cardiff
Bristol
Reading
London
Salisbury
0
25Km

a)

b)


Group 2b
Gloucester
Oxford
Swindon
Cardiff
Bristol
Reading
London
Salisbury
0
25Km

c)

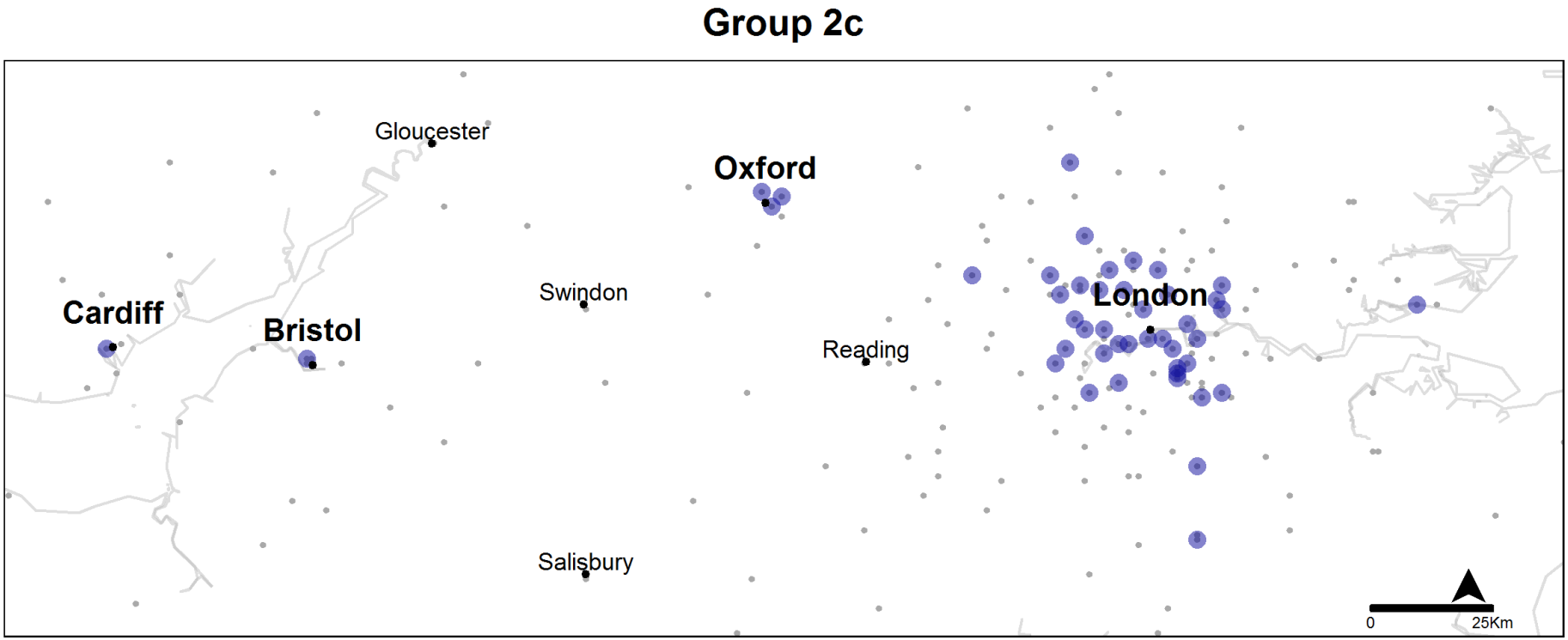
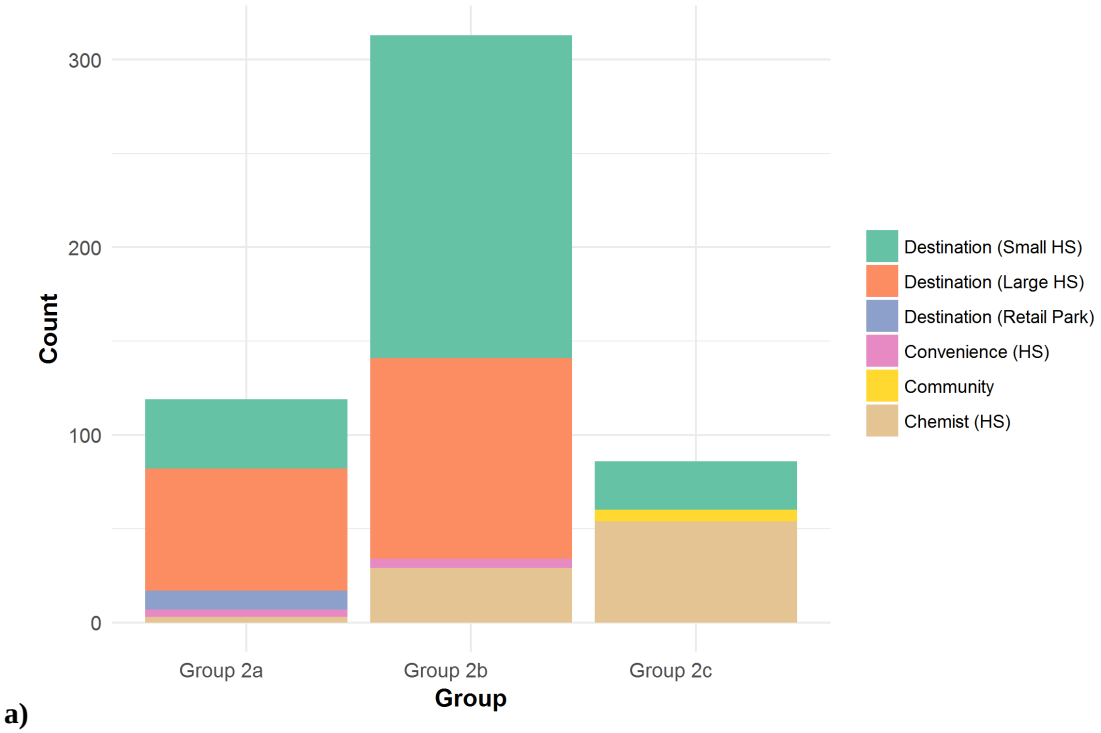
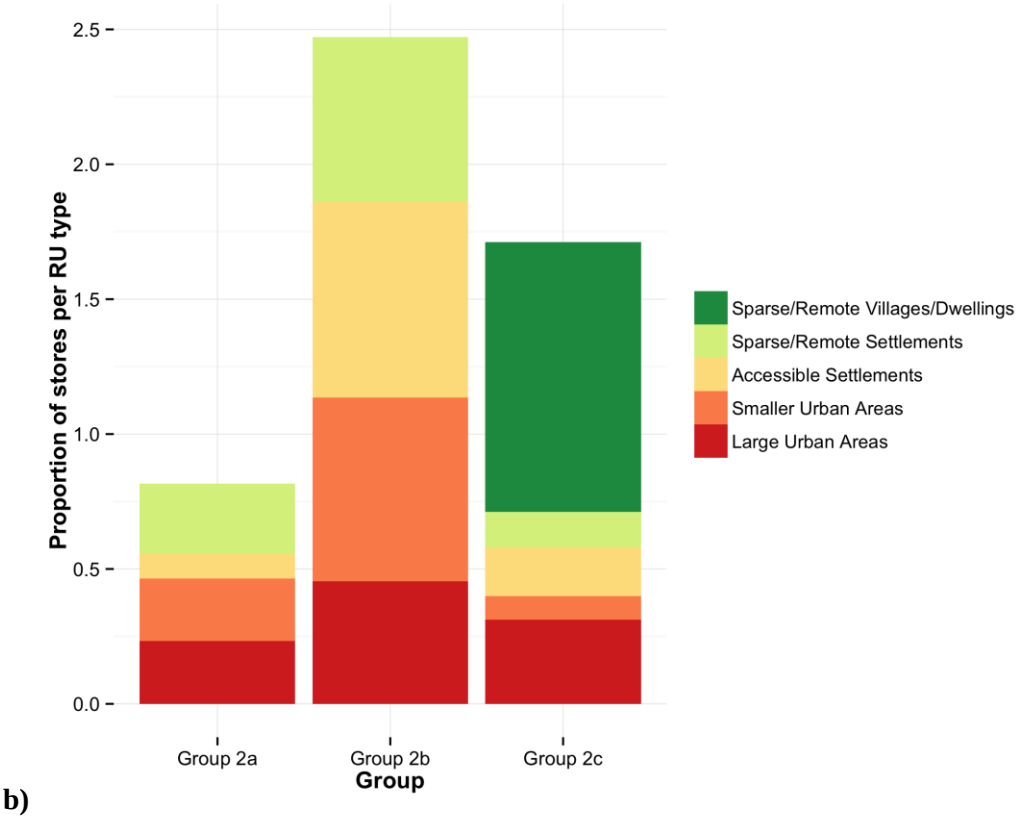


Figure 5.26: Distribution of all Supergroup 2 stores (shown in grey) and a) Group 2a - 'Weekend Destinations – Late Risers', b) Group 2b - 'Weekend Destinations – Early Risers', c) Group 2c - 'Weekend Destinations - General Activity', across Southern England (represented by 1km grid cell centres).



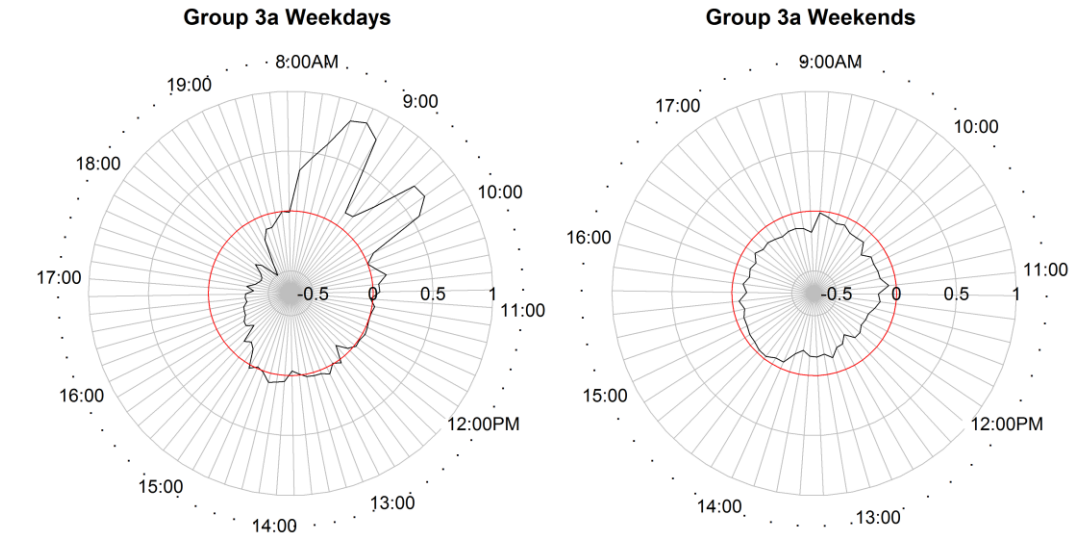
a)



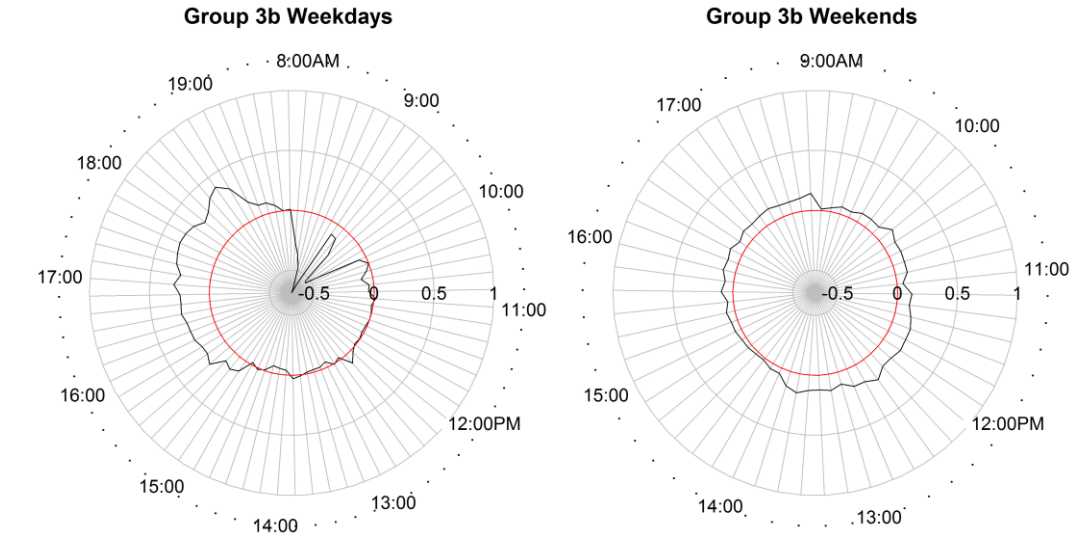
b)

Figure 5.27: Supergroup 2, a) store type counts per Group and b) proportion of rural/urban store locations per Group (normalised by total stores per RUC type in Supergroup 2).

Supergroup 3

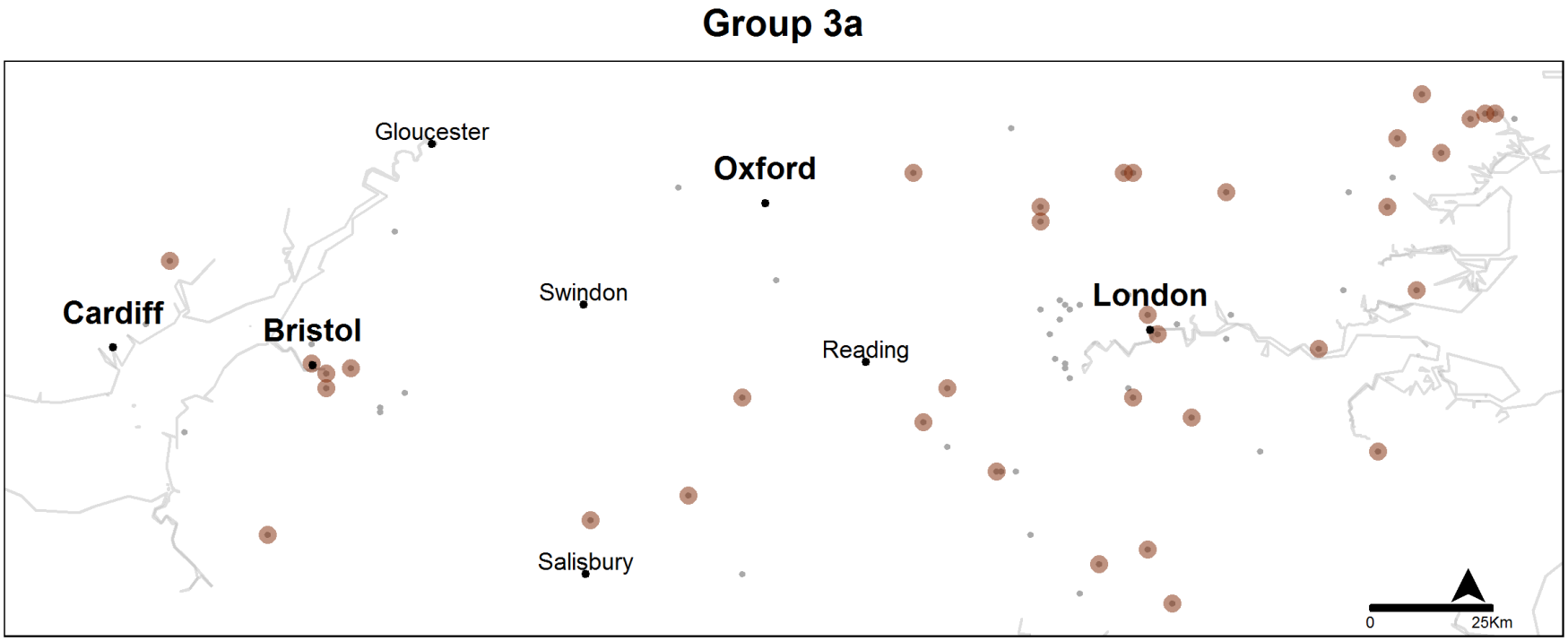


a)

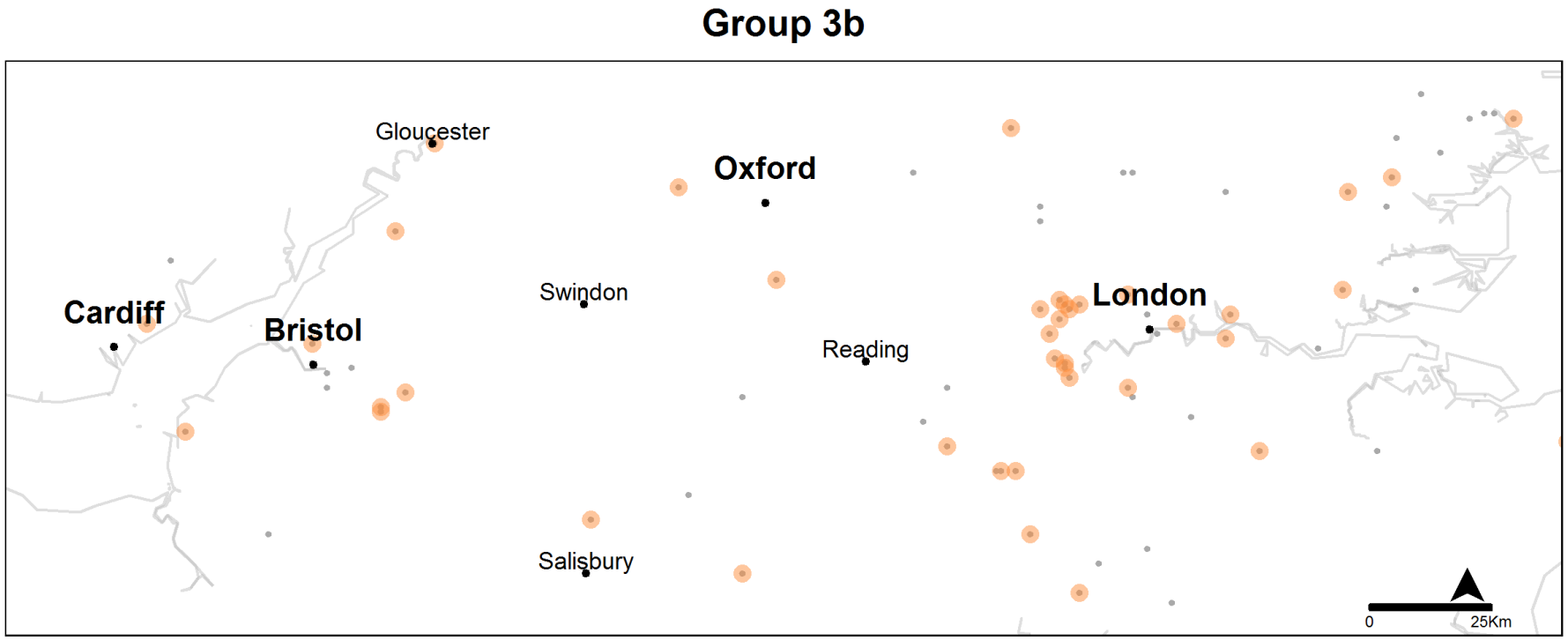


b)

Figure 5.28: Radial plots for a) *Group 3a* and b) *Group 3b*.

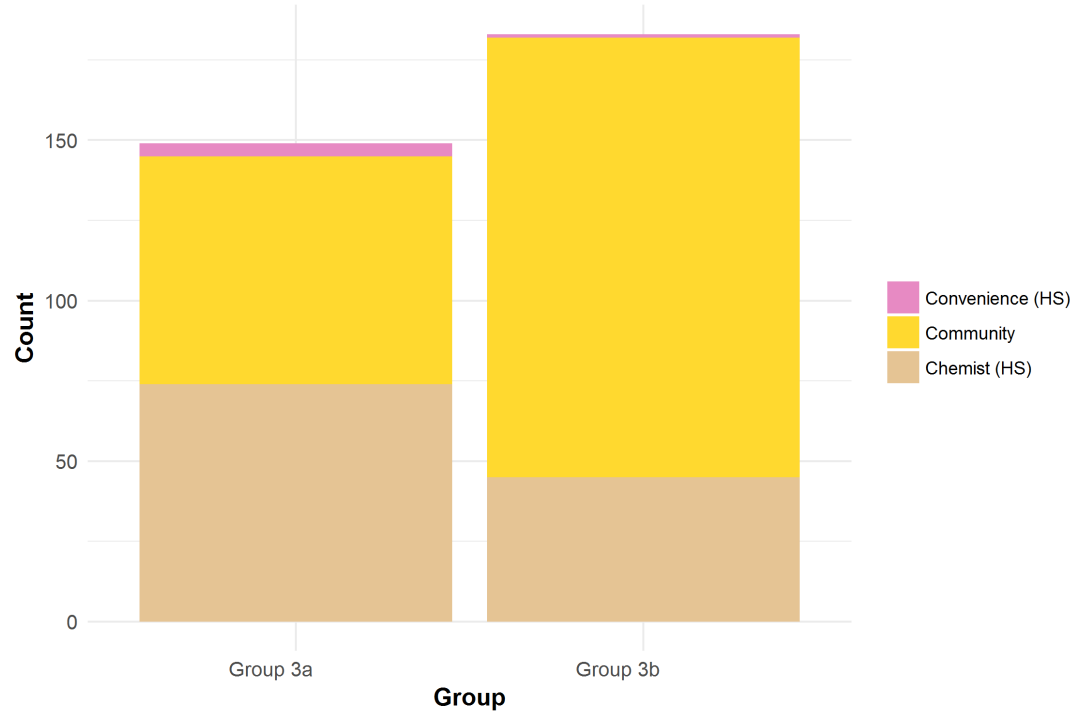

Group 3a
Gloucester
Oxford
Cardiff
Bristol
Swindon
Reading
London
Salisbury
0
25Km

a)

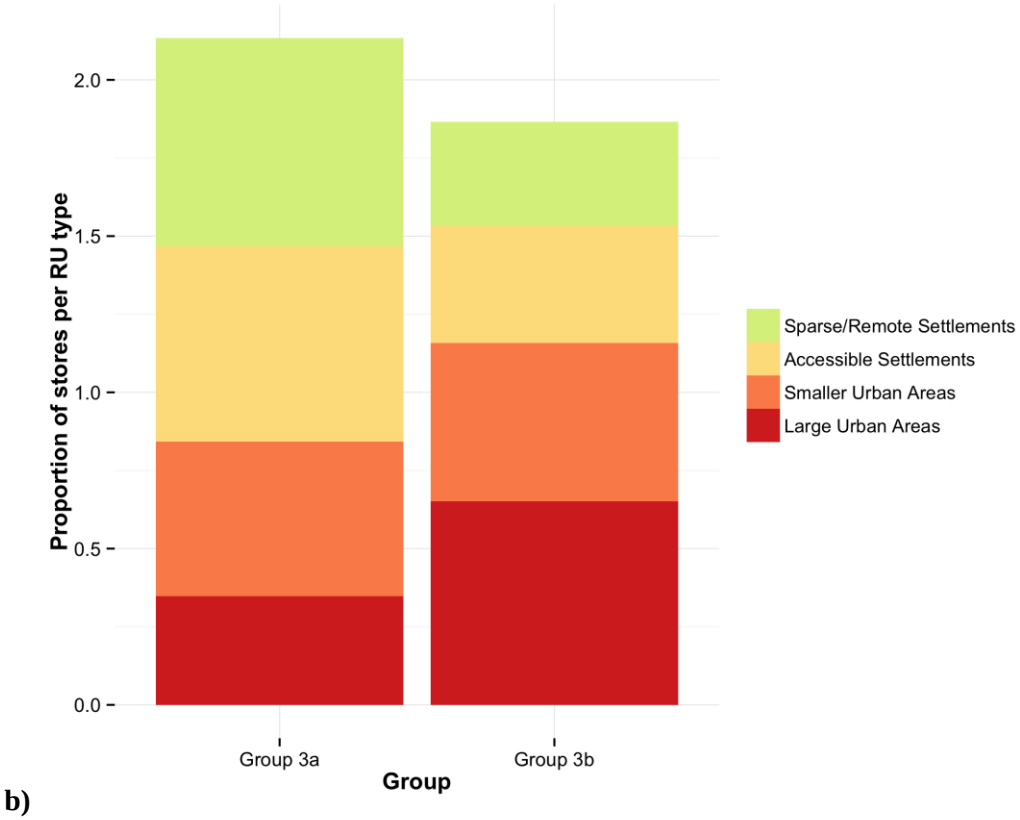


b)

Figure 5.29: Distribution of all Supergroup 3 stores (shown in grey) and a) Group 3a - 'Weekday Early Risers' and b) Group 3b - 'General Weekday Activity', across Southern England (represented by 1km grid cell centres).



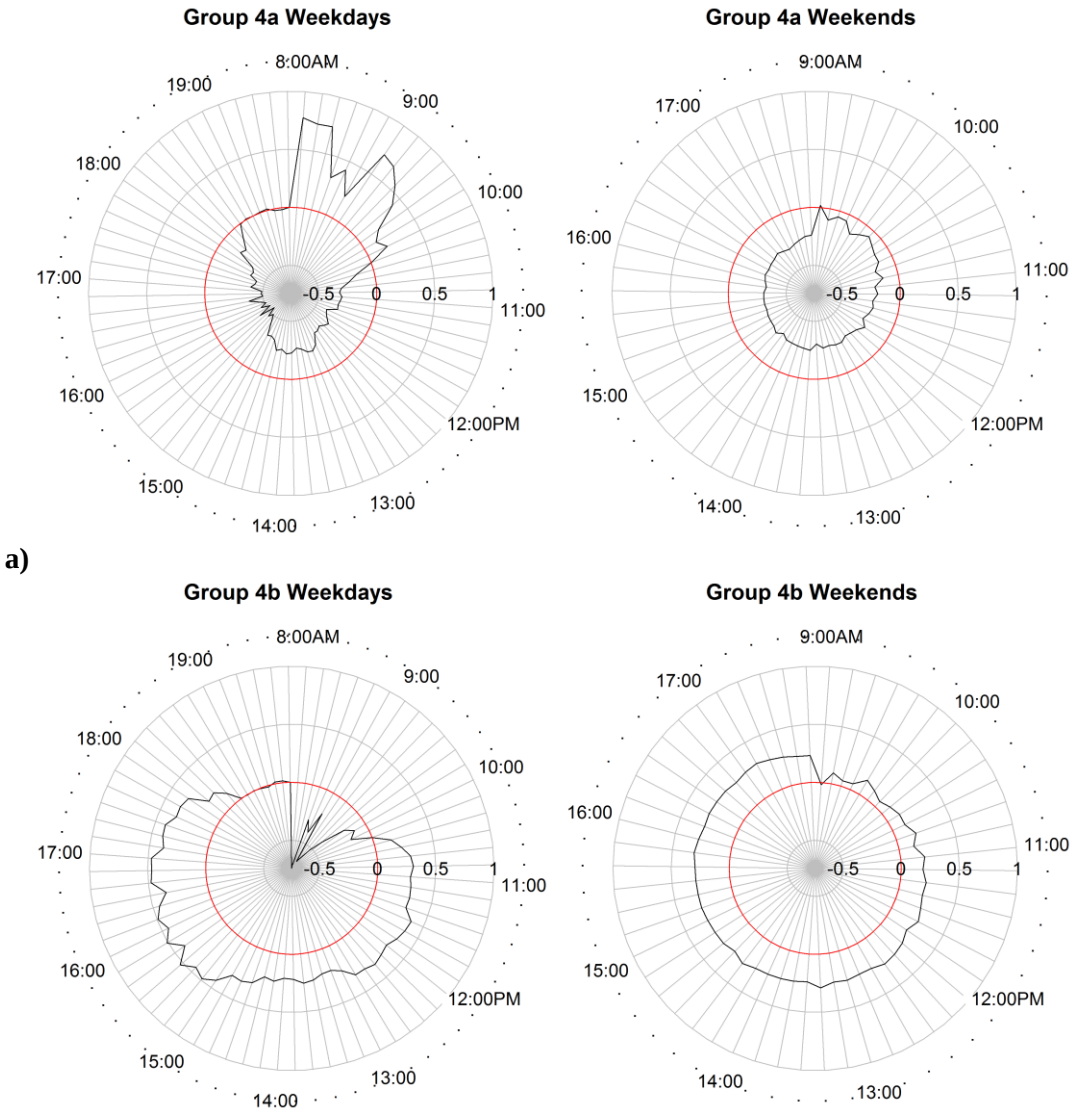
a)



b)

Figure 5.30: Supergroup 3, a) store type counts per Group and b) proportion of rural/urban store locations per Group (normalised by total stores per RUC type in Supergroup 3).

Supergroup 4

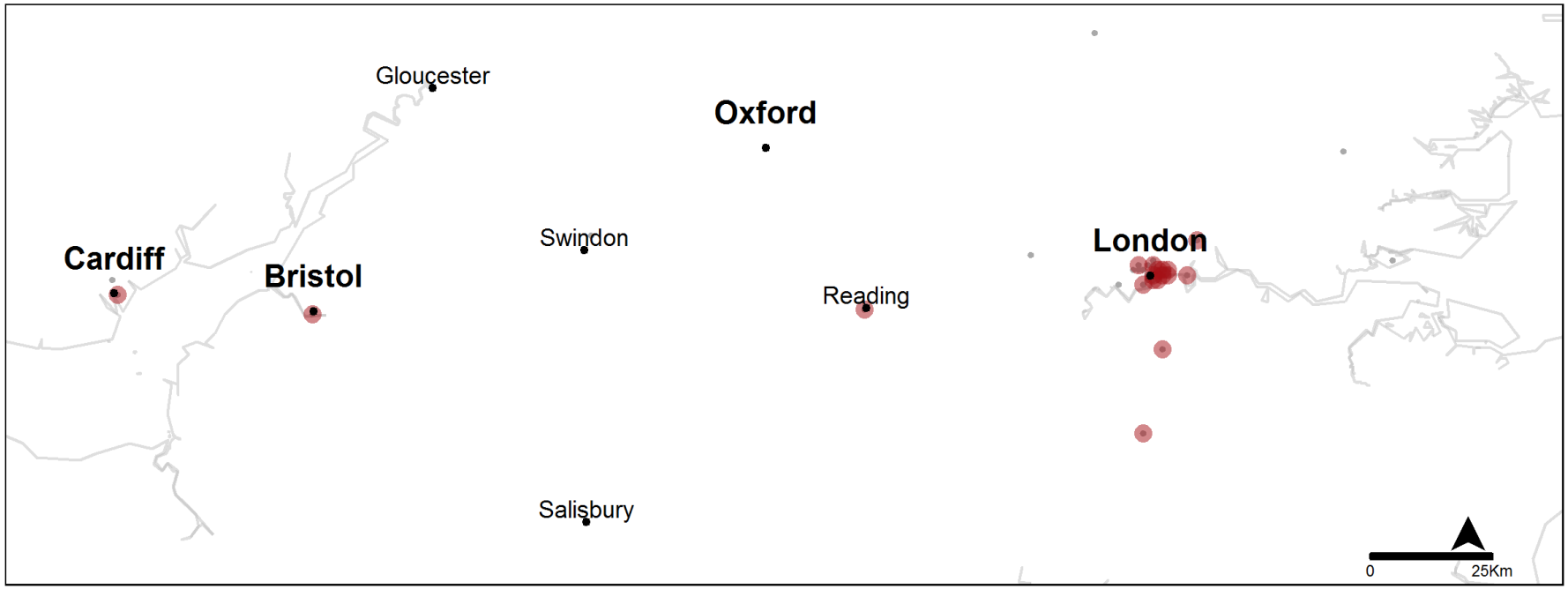


a)

b)

Figure 5.31: Radial plots for a) *Group 4a* and b) *Group 4b*.

Group 4a


Gloucester
Oxford
Swindon
Cardiff
Bristol
Reading
London
Salisbury
0
25Km

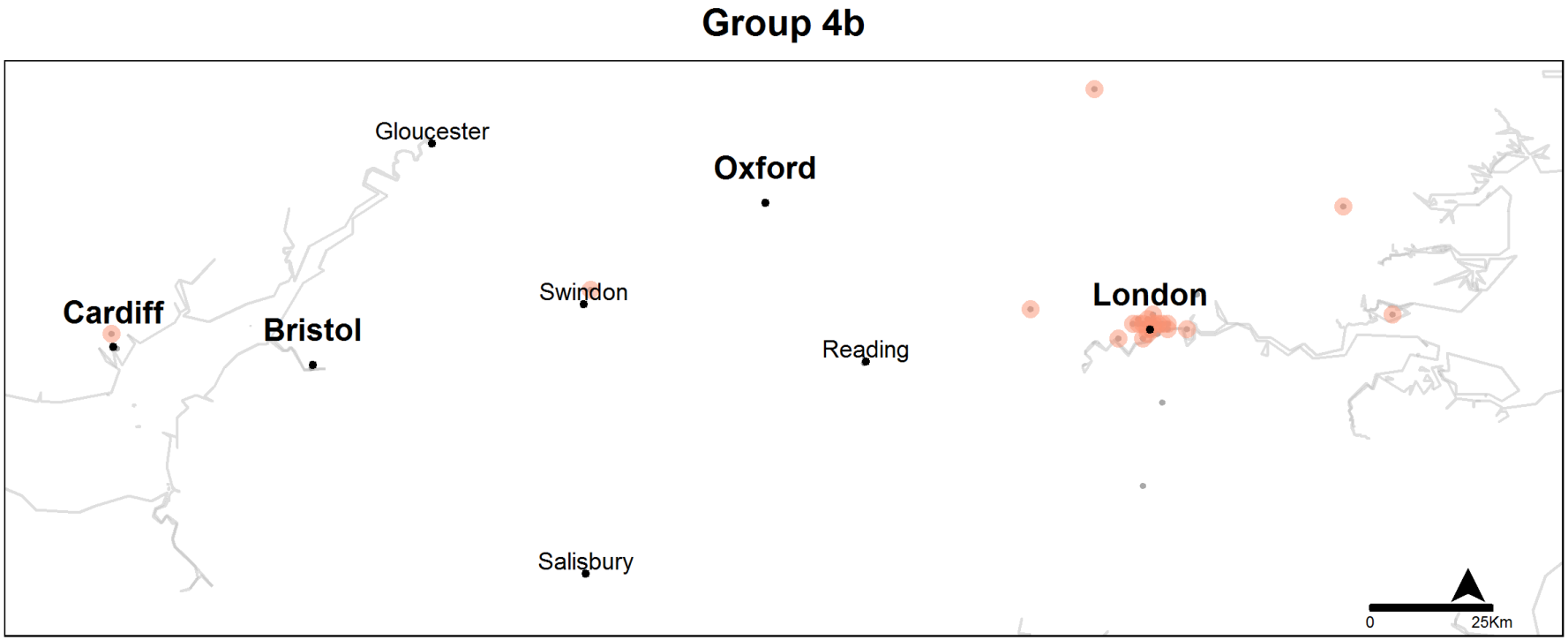
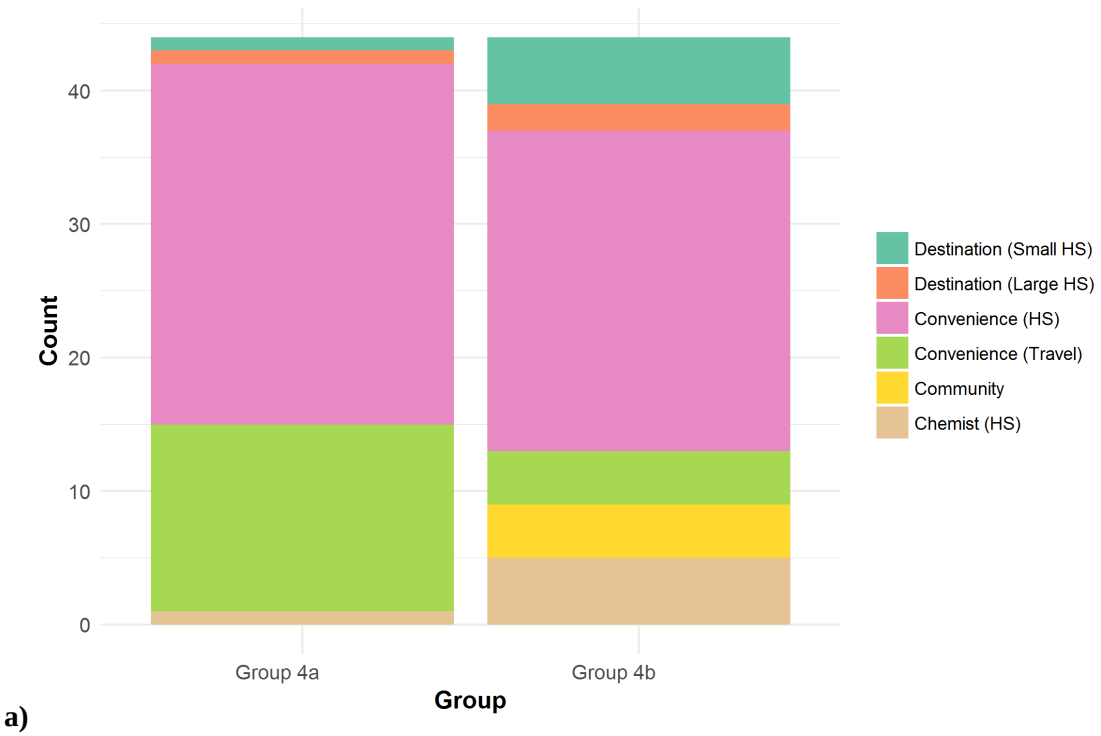
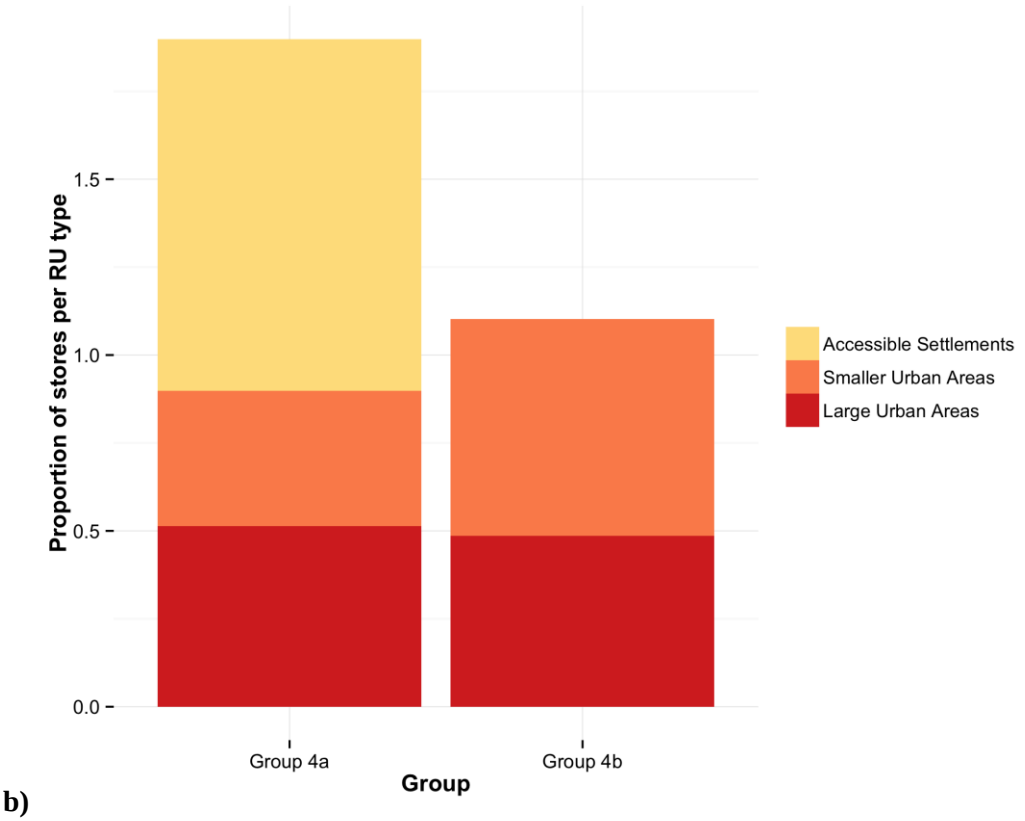


Figure 5.32: Distribution of all Supergroup 4 stores (shown in grey) and a) Group 4a - 'Commuter Convenience' and b) Group 4b - 'General Convenience', across Southern England (represented by 1km grid cell centres).



a)



b)

Figure 5.33: Supergroup 4, a) store type counts per Group and b) proportion of rural/urban store locations per Group (normalised by total stores per RUC type in Supergroup 4).

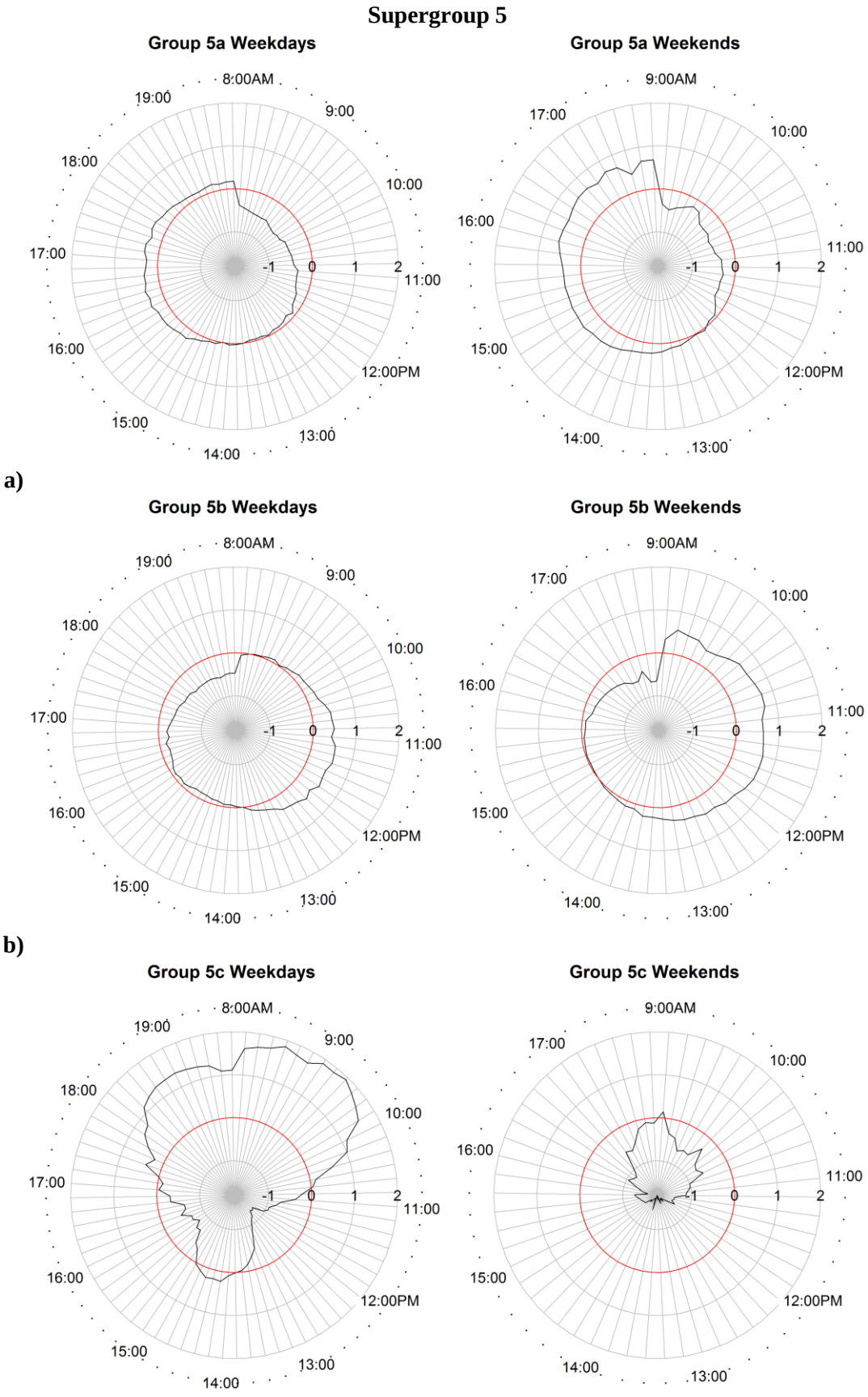
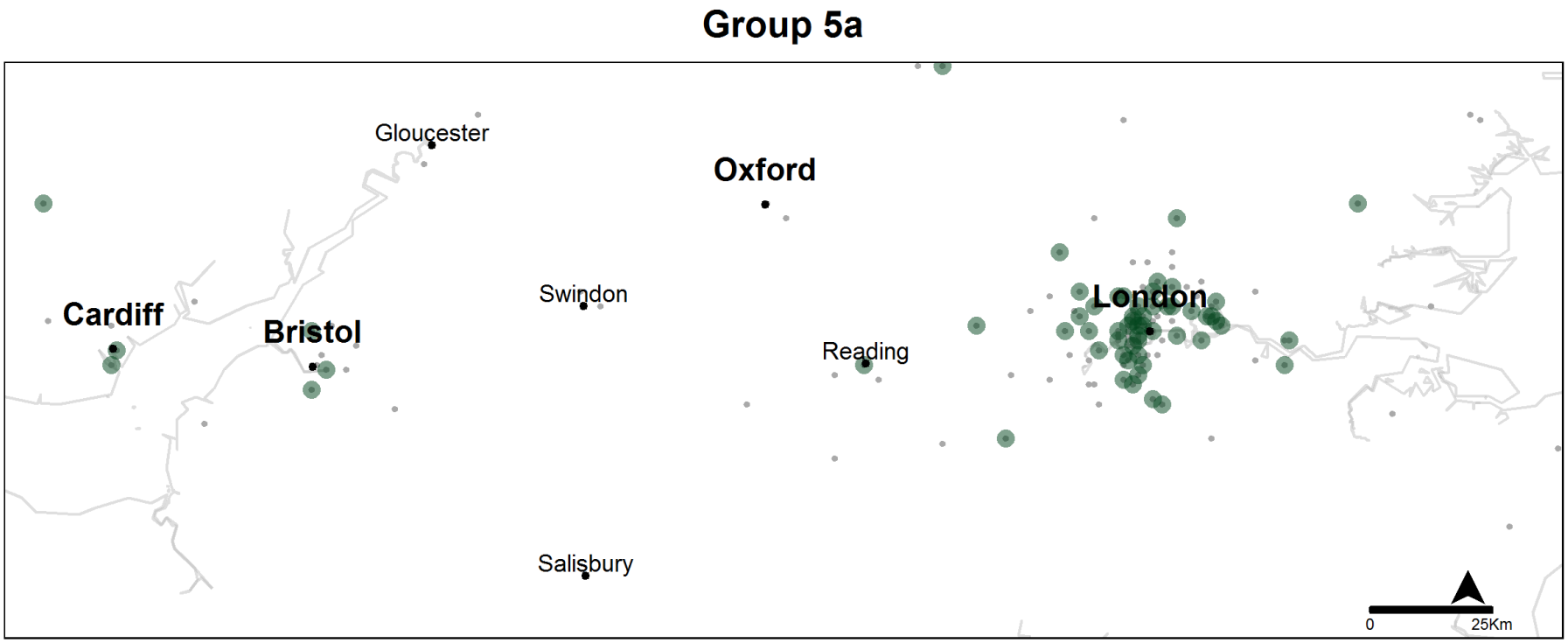
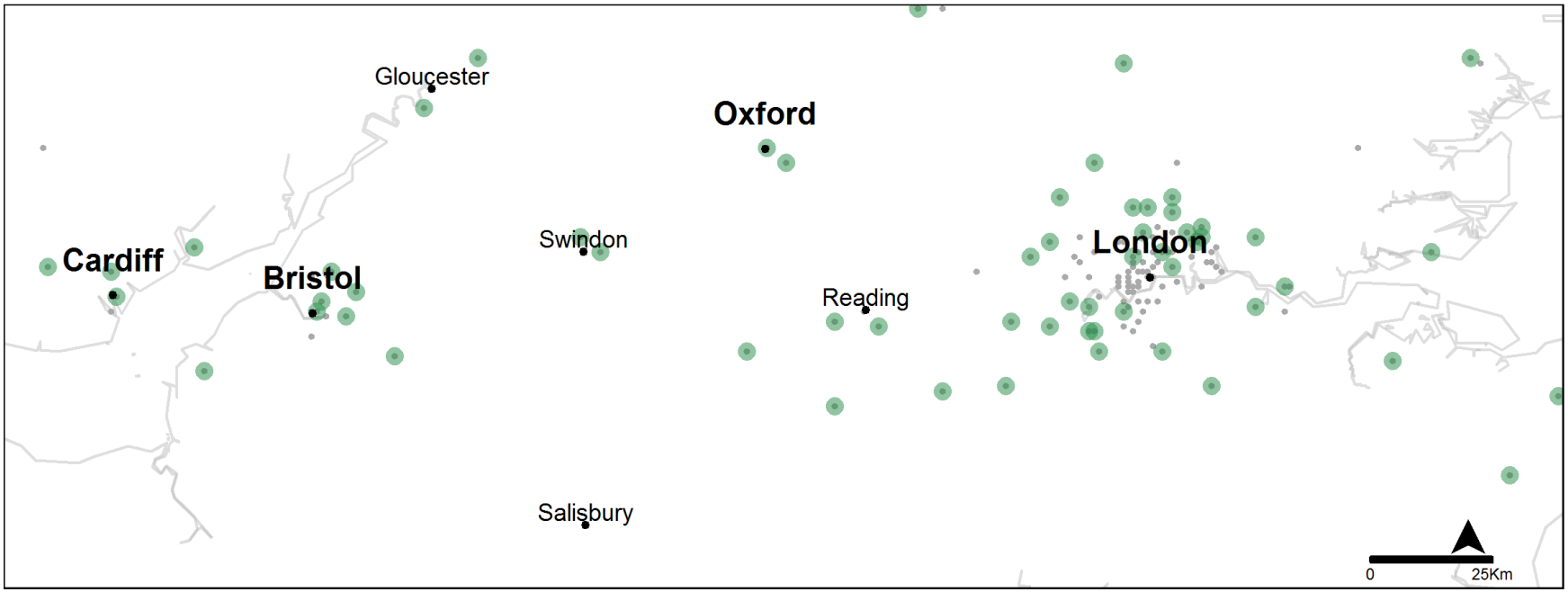


Figure 5.34: Radial plots for a) *Group 5a*, b) *Group 5b* and, c) *Group 5c*.

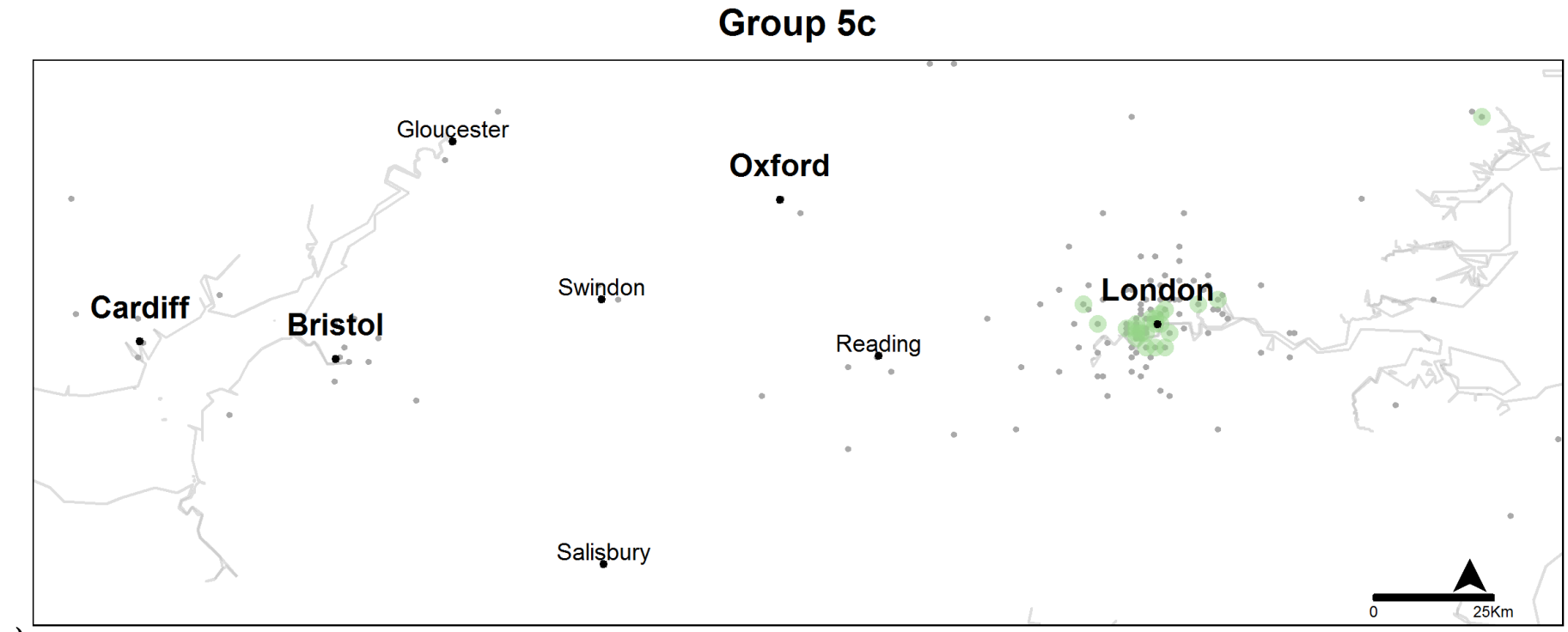

Group 5a
Gloucester
Oxford
Swindon
Cardiff
Bristol
Reading
London
Salisbury
0
25Km

a)

Group 5b

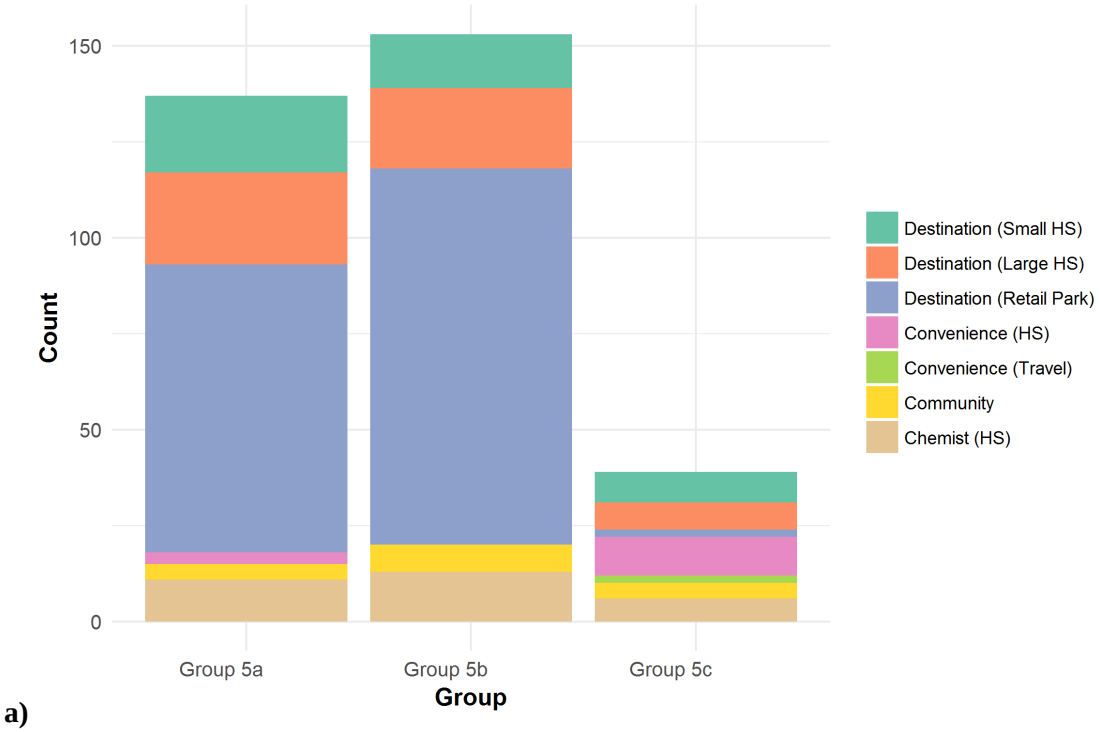

Gloucester
Oxford
Swindon
Cardiff
Bristol
Reading
London
Salisbury
0
25Km

b)

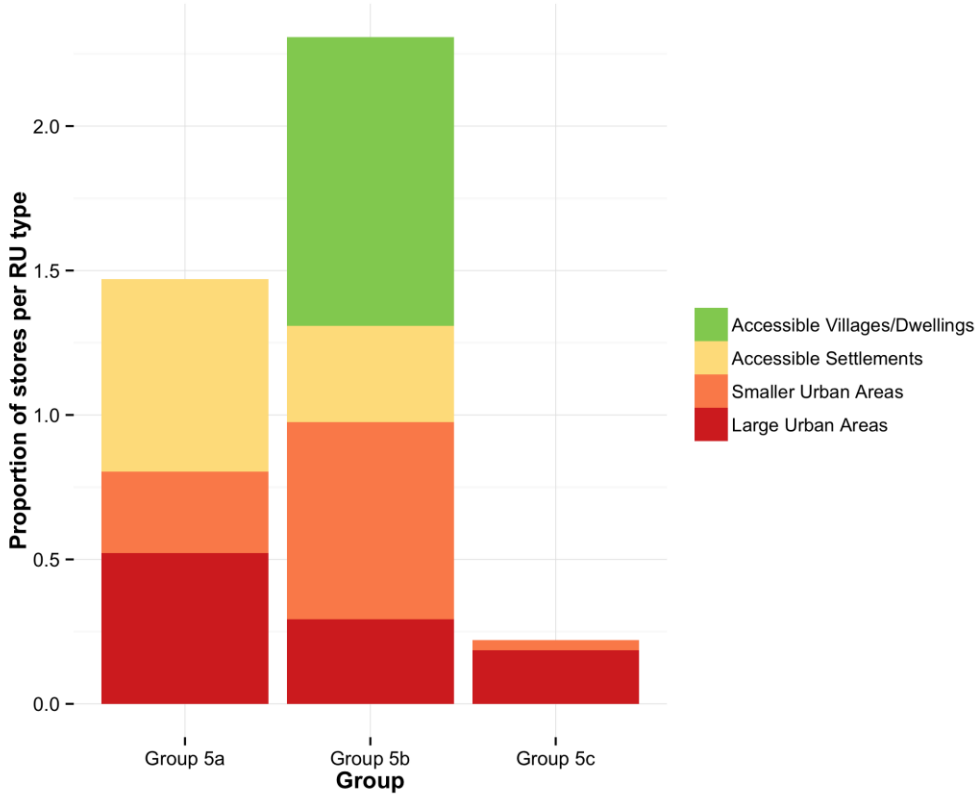


c)

Figure 5.35: Distribution of all Supergroup 5 stores (shown in grey) and a) Group 5a - 'Stable Destinations - Late Risers', b) Group 5b - 'Stable Destinations – Early Risers', c) Group 5c - 'Stable Urban Destinations', across Southern England (represented by 1km grid cell centres).



a)



b)

Figure 5.36: Supergroup 5, a) store type counts per Group and b) proportion of rural/urban store locations per Group (normalised by total stores per RUC type in Supergroup 5).

These results extended the findings of Section 5.3.1 and demonstrated how stores could be further segmented by utilising more granular time intervals. Similarly to the previous Supergroup analysis, temporal consumption patterns were able to differentiate stores into Groups that exhibited similar types, sizes, formats and locational attributes. Stores in Supergroup 1 ('General Off-peak Shopping') could be further separated by differences in morning/afternoon consumption peaks, which subsequently classified this initially mixed Supergroup into Groups with more uniform formats and location types. Supergroup 2 ('Weekend Peak Destinations') were further segmented into medium-large town stores (Group 1a), smaller town stores (Group 1b) and more rural fringe town centres (Group 1c). Supergroups with initially similar attributes (Supergroup 3 – 'Weekday Off-peak Shopping') could also be further differentiated by their temporal attributes.

Group 3b ('General Weekday Activity') showed the least discernible temporal pattern of all Groups and was the least homogeneous cluster. Yet, this group still exhibited identical store attributes, which suggests that variable consumption patterns may be a defining characteristic of these stores. Supergroup 4 ('Weekday Convenience') was further differentiated into those predominantly located in transport hubs (Group 4a) and other stores that exhibited general convenience trends (Group 4b). These were not exclusively 'Convenience' types as defined by the HSR. Both exhibited temporal patterns consistent with a working population, however Group 4b demonstrated less extreme peaks that began later in the morning. Supergroup 5 ('Stable Destinations') could be further segmented into two Groups predominantly located in retail parks (differentiated by their morning and evening consumption peaks) and a third group that exhibited weekday convenience and weekend destination mixed profiles. These stores were primarily located in London, which is consistent with the urban structure of this area (i.e. it has many both workplace and retail destination oriented locations). Other locations in this Group included city centre flagship stores such as Manchester Old Trafford Centre and Birmingham Bullring Centre. This suggests that temporal variations were able to accurately identify areas that may exhibit this weekday workplace convenience dynamic in addition to weekend destination shopping.

On the whole, each temporal cluster was dominated by a single HSR store type, of which was consistent with what we might expect. For example, Supergroup 4, 'Weekday Convenience', consisted predominantly of convenience high street and travel stores, yet no retail park stores. HSR defined 'Convenience' stores were also present in the convenience-oriented Group of Supergroup 5 (Group 5c). There are similar examples of these trends across all Supergroups and Groups. This suggests that locational factors may be highly correlated with temporal rhythms. However, there were also many instances that this was not the case. For instance, 6 'Chemists' and 6 small high street stores exhibited the temporal attributes of Supergroup 4 ('Weekday

Convenience') and 9 convenience high street stores demonstrated the temporal attributes of Supergroup 2 ('Weekend Peak Destinations'). The multiple HSR store types present in Groups 4a and 4b suggested that many location types may serve 'Weekday Convenience' functions.

The implications of these findings are that, firstly, they support the notion that we can identify the distinctive characteristics of a place according to its rhythmic ensemble. For instance, these trends suggested that temporal consumption patterns may be indicative of the activities of the populations who patronise these locations, which in turn may be predictive of locational attributes and the functions of stores. Secondly, they suggest that incorporating temporal attributes may provide enriched descriptions of stores compared to classifications that are constructed using only locational and store characteristics. For example, Group 5a, 'Stable Urban Destinations', exhibited convenience trends during the week, yet destination trends during the weekends and included a variety of HSR store types (including those classed as either destination or convenience). Thus, whilst the HSR descriptions may accurately represent them in terms of their physical and locational attributes, they do not acknowledge that stores may serve both convenience and destination functions, depending on the time of week. These instances suggest that classifications based only on physical and locational attributes may not represent the full complexity of functions that a location serves to consumers.

These trends would suggest that we can potentially use these insights to infer the activities of the consumers who patronise these different locations. For example, as is the case for Groups 5a and 5b, it may be of interest to investigate why retail park stores can be segmented into either those with afternoon/evening peaks or those with late morning peaks. From observing the geographical distribution of these stores it could be identified that Group 5a described more urban/suburban fringe retail park locations and Group 5b more rural fringe retail parks. Similarly, Groups 1b and 1c demonstrated uniform store characteristics, yet showed differing daily consumption trends. It is speculated that the patterns observed may demonstrate relationships with the characteristics of the population that patronise these different locations. Chapter 6 investigates this further.

5.3.4. Method Limitations

Before outlining the implications of these findings, there are limitations of this method to acknowledge. Firstly, not all cluster solutions demonstrated high levels of homogeneity. However, it is speculated that this may be due to the underlying characteristics of consumption rather than methodological issues. For example, the most homogeneous group was Supergroup 4 ('Weekday Convenience'). Less homogeneous clusters included Supergroup 3 ('Weekday Off-peak Shopping') and Supergroup 2 ('Weekend Peak Destinations'). Supergroup 4 demonstrated clear consumption peaks outside of business hours (early morning, lunch time and

evenings), which is indicative of trip chaining behaviour motivated by visits between business hours. Conversely, stores that might typically facilitate less time-constrained journeys are likely to exhibit more varied consumption patterns. This may include weekend destination or leisure trips (i.e. Supergroup 2) and off-peak local pharmacy visits (Supergroup 3). These groups exhibited much more gradual increases to peak consumption times than Supergroup 4. Therefore, it is possible that higher variation within some clusters may be a reflection of the function that some stores fulfil in consumer journeys, with leisure-based trips showing less specific peaks and thus adding more noise to a cluster.

A further limitation is that this analysis presents a relatively simple classification of temporally aggregated data. Whilst this provided sufficient insight into general variations in temporal consumption patterns for the purpose of this analysis, extensions to such analyses may benefit from investigating more granular temporal trends in specific contexts (i.e. of specific local areas, rather than the global approach applied here) or as previously mentioned, how trends vary over more longitudinal periods, to provide a more enriched description of variations between locations.

5.4. Discussion

The aim of this analysis was to understand if the distinctive characteristics of various retail centres were identifiable through their temporal rhythms. The results demonstrated that the temporal rhythms of consumption exhibited by HSR stores might be highly indicative of the attributes and functions of those locations, as clusters of stores exhibited distinct geographies, locational and retail characteristics. Findings implied that we may be able to enrich our understanding of retail centre characteristics by incorporating time as a dimension and moving away from traditional, static conceptualisations. For example, conceptualisations that describe stores based only on their locational and physical attributes only may overlook the function that a location serves to people during different time periods. Therefore, temporal trends may enrich our understanding of how, why and when people interact with different centres.

From the perspective of time geography, these findings indicate the potential strength of relationships between temporal consumption activities and the characteristics of retail locations. Research into the socio-spatial characteristics of both people and places over time have, to date, been largely inferred through qualitative methods, small sample sizes and limited time periods. This analysis provides data-driven evidence for how incorporating a temporal dimension could enrich our perception of retail spaces and their functions/formations, and how this may be achieved through the integration of novel consumer datasets. However, it should be noted that these findings are limited in their generalisability outside of this specific context, as HSR stores were the only proxy for a ‘place’ available in these data. In addition, these data are generated by

an inherently biased sample of the population (see Chapter 3, Section 3.4, for a summary) and therefore analysis of alternative data, with differing representations, are necessary to understand the full extent to which these patterns can be extrapolated to both retail centres as a whole, and the behaviours of the general population. However, the implication still holds that loyalty card data, and other spatiotemporal consumer datasets, have the potential to provide a more dynamic depiction of the characteristics of such places. As outlined above, future analyses could also aim to apply such data over various time periods (i.e. months, seasons or years) to assess how the spatiotemporal dynamics may change over more longitudinal periods.

These insights also have implications for retailers and high streets more widely. For example, as outlined in Chapter 2 (Section 2.2.3), there are a number of issues hindering our current understanding of high street resilience. Most notably, there is a considerable gap in knowledge regarding how consumers interact with UK high streets and town centres in the face of various contextual changes (for example, the rise of the 'convenience culture', multi-channel retailing and the impact of 'out-of-town' retail centres). In addition, there is a prominent lack of quantitative, coherent and consistent measures between locations. The insights presented here provide an example of how novel consumer datasets may aid in providing solutions to many of these core issues. For example, the spatiotemporal analysis of retail centres (as facilitated by loyalty card transactions) may provide insight into volumes of interactions with various high street locations in addition to how and why consumers are using them. These data are locally available, longitudinal, and national in scale, which could be of great use to creating the comparative, quantitative measures of high street performance highlighted by Wrigley and Lambiri (2011), to aid locally relevant decision making.

Specific implications of this analysis, in the first instance, are that the identification of distinct convenience trends across various HSR locations may allow quantification of the emerging convenience market and to what extent various high streets or town centres are utilised for this purpose. This analysis demonstrates how distinct temporal rhythms are associated with 'convenience' versus 'destination' oriented locations, indicating how consumer data may be utilised to quantify trip motivations, which could therefore aid modelling of locally relevant solutions to meet the needs of evolving consumer trends. These place functions were not discernible from the static HSR conceptualisation of stores in many cases, therefore, incorporating time may be crucial in helping us understand the specific function and trip motivations that retail centres attract, during various time periods.

In addition, the provision of both store locations and customer postcodes in loyalty card data provides a means of quantifying flows between consumers and high streets. This could aid understanding of a number of both short and long term impacts, such as the effects of accessibility and transport, the impact of changes such as multi-channel retailing and 'out-of-

town' retail centres on consumer flows and how these factors vary nationally. Whilst there will evidently be limitations to utilising insights from one retailers loyalty card for such endeavours, this work highlights the vast potential benefits that may be possible if public bodies were able to access and integrate various forms of consumer data (in particular, high street retailer data) to inform the resilience of UK high streets. This may provide a viable means of creating consistent quantitative measures of these important economic, social and community spaces that have been unobtainable to date. Thus, it is hoped that this analysis demonstrates how loyalty card data, and other consumer data, may be relevant to problems that concern the public, government and retailers alike.

6. Classifying HSR Customers

6.1. Introduction

As outlined in Chapter 2, research has indicated that the temporal rhythms of places may be inherently interlinked with the spatiotemporal rhythms of distinct social groups of people (Lager et al., 2016; Longley 2017). For example, the rhythms of certain places may be reflective of the differing daily obligations and trip chaining activities of the individuals who patronise them. This would suggest that focusing on everyday activities may reveal how the rhythms of both places and people are ordered, and how these orderings may vary by social group. Based on the analyses conducted in Chapter 5, it was hypothesised that the temporal consumption patterns exhibited by HSR stores may provide a means of extracting the spatiotemporal rhythms and related social characteristics of HSR customers. For example, characteristics regarding the identities of individuals may be identifiable through the analysis of those who patronise urban 'Weekday Convenience' stores, versus those who exhibit rural, off-peak shopping trends. This chapter therefore aimed to investigate the extent to which customer characteristics can be extracted from their spatiotemporal store interactions. Thus, the aims of this analysis were:

1) To classify HSR customers based on their store visiting behaviours during different time periods.
2) To explore relationships between customer profiles and identity factors, as measured by their demographic, geographic and consumption characteristics.

The following sections present the development of an appropriate methodology to address these aims, the segmentation of individuals based on their spatiotemporal consumption habits and an exploration of the resulting groups' characteristics. The chapter concludes by discussing the implications of integrating dynamic population activity patterns into traditional representations of people and places.

6.2. Method

In order to investigate relationships between the temporal profiles of HSR locations and the customers who patronise them, an appropriate methodology was required to, 1) identify store profile interactions for each customer, and 2) group customers who exhibited similar behavioural patterns. To achieve this, a simple pattern matching technique was developed that firstly, assigned a profile to each customer based on the store types they interacted most prominently with during either weekdays or weekends, and secondly, segmented customers by

those who exhibited matching profile assignments. This allowed exploration of the characteristics of segments that exhibited distinct spatiotemporal consumption habits. The proceeding sections present the development and implementation of this method.

6.2.1. Exploratory Analysis and Segmentation Method

The first consideration in constructing this method was the selection of suitable time periods over which to analyse consumption patterns. Observations from the store level profiling indicated that the analysis of weekday versus weekend behaviour was the most prominent distinguishing factor. Therefore, individual store visiting behaviours were similarly assessed over these periods. This allowed sufficient identification of variation in customer activities, whilst also avoiding creating substantially complex outputs (i.e. in comparison to utilising daily patterns).

Exploratory analyses were conducted to inform a method of assigning customers to temporal store profiles. A practical solution here would have evidently been to identify to the stores that individuals most prominently interacted with and append their associated profiles (i.e. the Supergroups or Groups derived from Chapter 5). However, it was recognised that this could potentially lead to a loss of important variation in individual behaviour. For example, this would assume patronage to only one store type, when in reality, customers could exhibit high or equal patronage across multiple types. To understand the suitability of assigning customers based on the store types that facilitated their most transactions, the number of store profiles that each individual interacted with were quantified for each time period. This indicated that customers patronised an average of 3 different store profiles during weekdays and 2 during weekends (see Figure 6.1). On weekdays, 31.4% exhibited activity across two different store profiles and 47.9% across three. On weekends, a higher percentage demonstrated activity within only one profile (23.7%), yet the majority across two profiles (38.3%). This suggested higher overall variation during weekday periods. In both cases, a minority of customers demonstrated patronage to more than 3 profiles during either period. Yet, those belonging exclusively to one store profile were also a minority.

Despite this, an investigation of transactional frequencies indicated that whilst customers may exhibit varying behaviour overall, a substantial proportion of their transactions could be attributed to one store type. For example, on average, a customer's primary profile (i.e. the store type that facilitated the most transactions) accounted for 77.5% of all weekday transactions, and 78.6% of weekend transactions. Only a small proportion (7.6%) exhibited less than 50% of their transactions within their highest-ranking weekday profile, and 4.7% on weekends. Therefore, it was concluded that grouping customers based on the store profiles for which they exhibited the

highest patronage to during weekday and weekend periods would be representative of the majority of their activities.

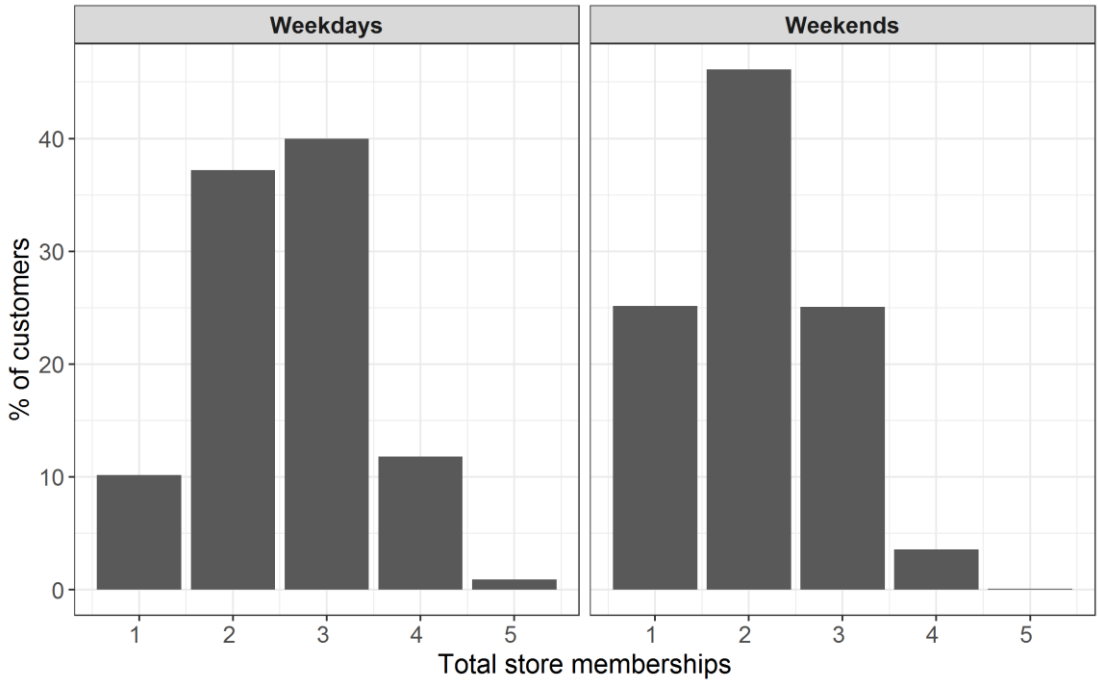


Figure 6.1: Total individual interactions with store profiles (across the 5 store Supergroups) evident over the 2.5 years, during weekdays and weekends.

However, to account for the loss of potential variation through the implementation of this method, a number of measures were obtained for interpreting results. This included how many profiles customers belonged to overall during each time period, differences in transactional frequencies between rankings and also analysis of the alternative store types customers were likely to patronise. Thus, whilst customers were grouped based on their most patronised store profile, variation in behaviour was also quantified.

6.2.2. Method Implementation

6.2.2.1. Active customer selection

Utilising the individual level data required consideration of the large variances in behavioural data pertaining to each customer (see Chapter 2). Selecting customers who had generated a sufficient volume of data for analysis required consideration of both the frequency and duration of their activity. As the time period of interest was activity patterns over weekdays and weekends, customers were assessed based on the number of unique weeks they transacted over the 2.5 years. In order to maintain a suitable level of minimal transactions, the threshold was set at least 2 transactions a month for a minimum of 1 year (thus a minimum of 24 transactions). This resulted in a sample of 4,834,282 customers. Transactional levels varied largely within this

active sample, with an average of 69 and a maximum of 2102 transactions per customer. Figure 6.2 shows the distribution of transactions. Approximately 24% of customers fell into the lower quartile.

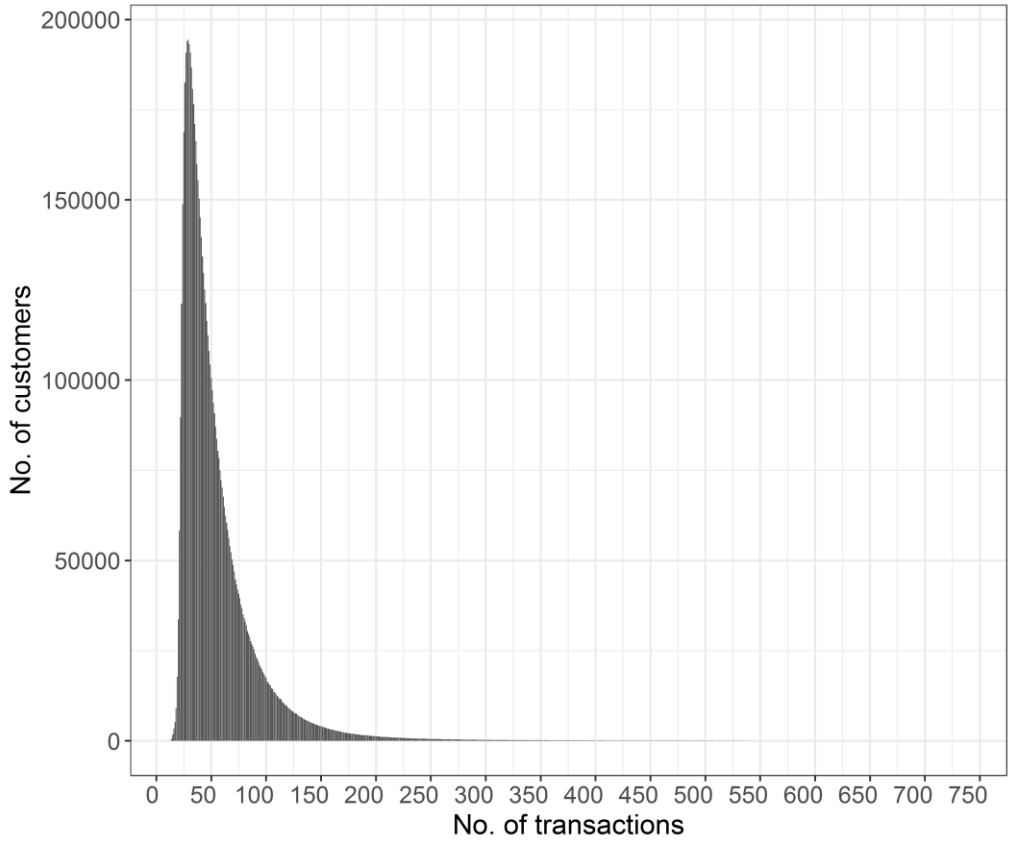


Figure 6.2: Distribution of total transactions for active customers.

Total transactions generated by active customers varied between store types, with the most active store types being Supergroup 2 ('Weekend Peak Destinations') and 5 ('Stable Destinations'). Supergroup 3 ('Weekday Off-peak Shopping') exhibited the lowest activity overall. These dynamics may be related to the characteristics of different store types. For example, Supergroup 3 described a relatively specific group of small, pharmacy-oriented stores. The highest percentage of data was lost through active customer selection for these stores, suggesting that the most active of HSR customers may be least likely to patronise this store type. Table 6.1 illustrates the number of transactions recorded in each store Supergroup.

Table 6.1: Transactions recorded within each store Supergroup.

Store Supergroup	**Total loyalty card transactions** *(2.5 years)*
1 – 'General Off-peak Shopping'	3,844,299

2 – ‘Weekend Peak Destinations’	7,942,408
3 – ‘Weekday Off-peak Shopping’	481,170
4 – ‘Weekday Convenience’	2,339,921
5 – ‘Stable Destinations’	7,577,902

However, due to the nature of loyalty card usage, this could also be a reflection of less loyalty card participation (i.e. high volumes of non–card transactions may still occur). These uncertainties represent a disadvantage of loyalty card data analysis. However, attempts can still be made to identify trends from the large volume of remaining data.

6.2.2.2. Data preparation

To obtain data for analysis, all transactions generated by active customers were selected across the 2.5 years, store ID’s and transaction counts were obtained for each customer during weekdays or weekends with temporal store types appended (at both the store Supergroup and Group levels), and transactional counts aggregated per type. Data therefore pertained to the number of transactions performed within each store profile during either weekday or weekend periods. The stores included in this analysis were those also utilised in the store profiling analysis (see Chapter 6, Section 6.2), and thus 182 were excluded. In addition, due to the utilisation of customer addresses post analysis, all customers identified as showing uncertain postcode information (see Chapter 4) were excluded. This removed 141,659 accounts from the sample.

This segmentation method was reliant on identifying customers who exhibited matching weekday-weekend patronage behaviours. However, it is improbable to assume that all customers would demonstrate activity during both of these periods. Therefore, customers were separated into those who transacted exclusively on weekdays, exclusively on weekends, and those that showed mixed profiles. Overall, 88,741 individual’s demonstrated weekday-only behaviours, and 621 were only active on weekends. These relatively small figures are likely due to the selection of only substantially active customers, the longitudinal nature of these data and also that HSR customers are generally more active during weekdays. As the aim of this analysis was to segment customers based on weekday-weekend activity patterns, only mixed profile customers were utilised.

To assign weekday and weekend profiles for each customer, interactions with store types were ranked based on transactional volumes within each. This required consideration of rank-ties, where in some cases, customers exhibited equal transactional frequencies across one or more profiles. Ranks were calculated using the *ranks* function in R with the ‘average’ ties method, which created split rankings for tied instances. At the store Supergroup level, there were 518,390 (10.9% of customers) split ranks (113,356 during weekdays, 405,034 during weekends)

and at the store Group level, 711,610 (14.9%) split ranks (162,012 weekdays, 549,598 weekends).

For the remaining customers, highest-ranking store profiles were selected. This resulted in a dataset consisting of account numbers, a weekday profile and a weekend profile, which could be utilised to identify customers who exhibited matching behaviour. It should be noted that the process of ranking weekday and weekend transactional frequencies separately subsequently ignored consideration of whether an individual transacted more overall during weekdays or weekends. However, it was hypothesised that the most useful information could be derived from the interaction of store visiting behaviours rather than the magnitude of transactions that occurred during each period. The inclusion of this information would have also lead to a large number of possible combinations, which would create an unnecessarily complex number of outputs.

Using the ranked data, the number of customers exhibiting primary memberships to store profiles could be examined (see Table 6.2). As can be observed, there were large variations in the number of customers per store profile. These volumes were likely influenced by firstly, the nature of the HSR store network (for example, there are many more of Supergroup 2 stores than 3) and secondly, the overall levels of loyalty card activity that store types exhibit (as depicted in Table 6.1). However, it was not considered important in this context (although typical of many classification outputs) for segments to be of equal sizes, as these frequencies reflected the differing volumes of customers exhibiting particular patronage patterns, of which was in itself informative about the HSR loyalty population.

Outputs revealed variation in primary membership frequencies across weekday and weekend periods. For example, store Supergroups 3 and 4 experienced low primary memberships during weekend periods, consistent with observations from the previous analysis that they serve predominantly weekday-based consumption. Similarly, store Supergroups 2 and 5 exhibited higher volumes of primary memberships during weekend periods. Supergroup 3 showed the lowest overall memberships, as expected based on the amount of available data for these stores. Exploring the distribution of memberships per store Supergroup provided further insight into the dynamics of store type usage. Figure 6.3 shows the frequency at which a store Supergroup was assigned each possible rank, demonstrating that in the case of Supergroup 3 (‘Weekday Off-peak Shopping’), a higher proportion of customers utilised them as second or third choice locations. This suggested that this store type did not function as a primary location to the majority of the HSR population, potentially due to their specific retail offering and locational characteristics.

Table 6.2: The total frequency of customers who exhibited primary membership to each store type, during weekdays and weekends.

Supergroup	Weekday *(Count)*	Weekend *(Count)*
1 – 'General Off-peak Shopping'	636,143	564,448
2 – 'Weekend Peak Destinations'	1,962,000	2,113,974
3 – 'Weekday Off-peak Shopping'	31,421	15,790
4 – 'Weekday Convenience'	308,575	81,784
5 – 'Stable Destinations'	1,665,122	1,827,265

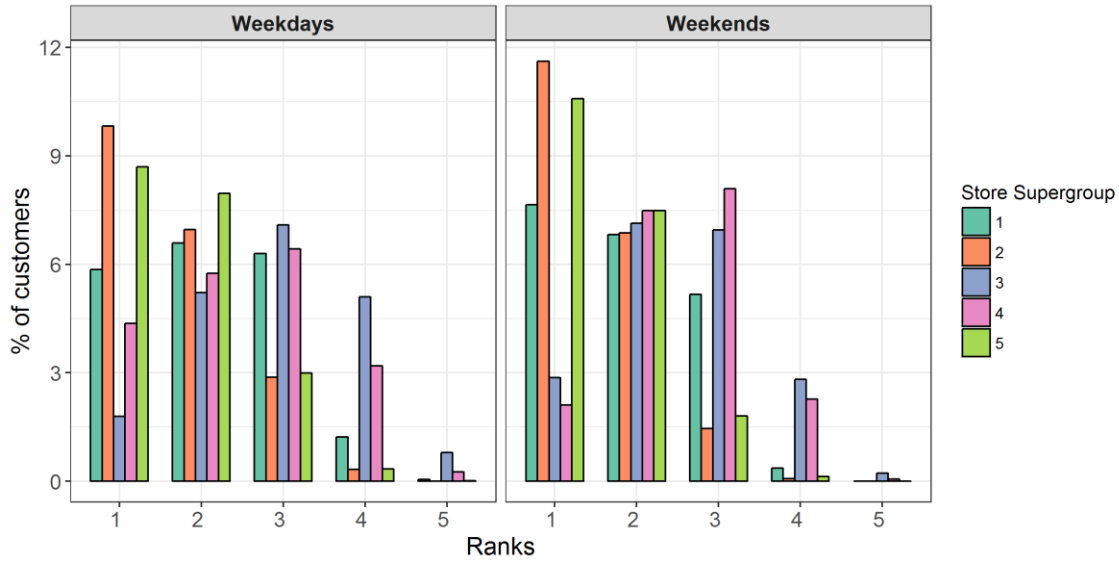


Figure 6.3: Customer ranks assigned to each store Supergroup, during weekdays and weekends.

Supergroup 4 ('Weekday Convenience') stores were also most likely to be ranked as second (29%) or third choices (32%) than first (22%). Store Supergroup 1 ('General Off-peak Shopping') was most likely to be the second highest ranked (32%) during the week, but first (38%) during the weekend and Supergroups 2 ('Weekend Peak Destinations') and 5 ('Stable Urban Destinations') both showed the highest volumes of primary memberships during both periods. These trends may explain the low primary membership to Supergroup 3, as customers may typically patronise alternative store types more frequently.

6.2.2.3. Classification structure

The next methodological stage was to derive a meaningful segmentation structure from customers' weekday and weekend profiles. This was achieved by firstly, grouping customers based on their top weekday store type membership at the store Supergroup level (5 types). The weekday profile was allocated higher importance in the initial segmentation, as this was the period that exhibited the most distinct temporal patterns between groups (see Chapter 6, Section

6.3). This created the ‘Supergroup’ level of the customer segmentation. To create the ‘Group’ level, individuals were segmented by matching weekday-weekend profiles. Data at the store Group level were not utilised for customer segmentation due to the large number of possible interactions this would produce (i.e. across 13 different types). However, these data were utilised to understand more specific store interactions when interpreting the characteristics of each segment. The process of aggregating all possible weekday-weekend membership combinations created 25 groups. However, frequencies of customers were not evenly distributed amongst these profiles, with many explaining a very small percentage of overall behaviour. Thus, in order to refine outputs, a method of selecting the most highly interacting pairs was necessary.

Table 6.3 illustrates the cumulative percentage of active customers explained by each combination. This allowed interpretation of the most highly interacting profiles. An identified trend was that interactions with the equivalent weekend profile were prominent. For example, the most frequent pattern identified for Supergroups 1, 2 and 5 was patronage to the same store type during weekends. This is likely, in part, a reflection of customers who visit the same store location during both periods. Conversely, Supergroups 3 and 4 did not show the highest interaction with their equivalent weekend profiles. This is consistent with the fact that these store types exhibit much lower weekend activity than other Supergroups, therefore, interactions with alternative store types during weekends were more prominent.

The proportion of customers explained by each combination provided insight into the most interesting groups for further analysis. Based on these observations, Group interactions explaining the highest 99% of customer behaviour were selected for inclusion in the segmentation. As can be observed from Table 6.3, groups falling outside of this threshold described a very small proportion of the customer sample. This included the majority of weekend interactions with Supergroup 4 stores and also eliminated consideration of Supergroup 3. Overall, 44,676 customers (0.95% of the active sample) were removed from the analysis by this refinement process. The resulting classification structure therefore described customer behaviour across 4 Supergroups (describing primary weekday store type membership) and 14 Groups (describing interactions between weekday and weekend profiles). This structure is depicted in Figure 6.4. Each level was assigned a unique ID, which is used as a reference in the proceeding evaluations.

Table 6.3: Proportion of customers explained by each weekday-weekend profile combination.

Supergroup (Weekday profile)	Group (Weekday-weekend interaction)	% Active customers	Cumulative %
2	2+2	34.78	34.78
5	5+5	28.47	63.25
1	1+1	9.10	72.35
5	5+2	5.45	77.80
2	2+5	5.19	82.99
4	4+5	3.26	87.25
1	1+2	2.60	88.85
4	4+1	2.37	91.22
4	4+2	1.88	93.10
1	1+5	1.78	94.88
2	2+1	1.55	97.44
4	4+4	1.16	97.60
5	5+1	1.04	98.64
5	5+4	0.40	99.05
3	3+2	0.23	99.28
3	3+3	0.21	99.49
3	3+5	0.15	99.64
2	2+4	0.13	99.78
3	3+1	0.06	99.84
2	2+3	0.06	99.89
5	5+3	0.04	99.93
1	1+4	0.03	99.96
1	1+3	0.02	99.98
3	3+4	0.01	99.99
4	4+3	0.01	100.00

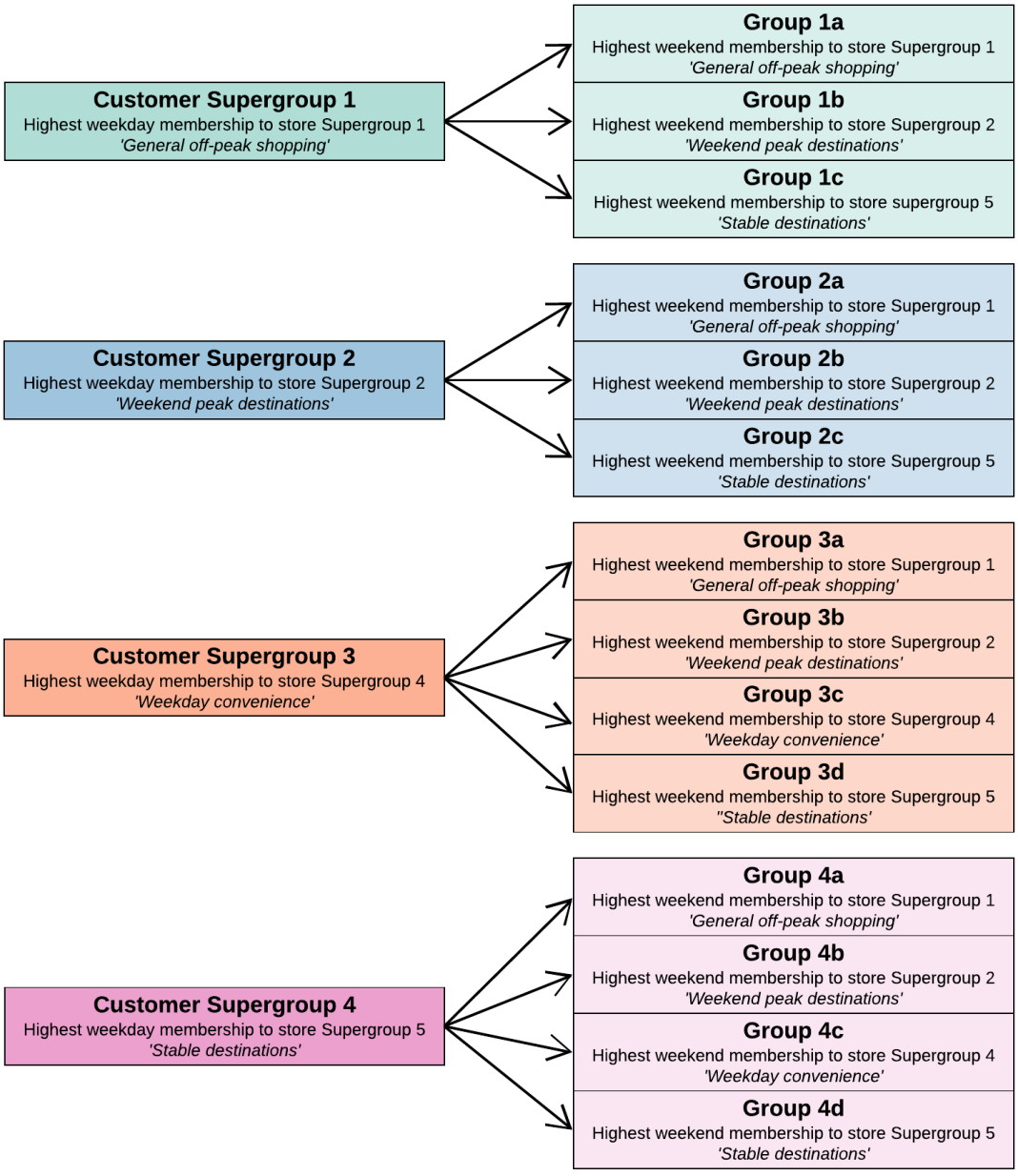


Figure 6.4: Overall structure of the customer classification.

6.2.2.4. Profile descriptions

The final consideration was the selection of appropriate variables to explore the characteristics of the resulting customer segments. Four criteria were used to create names and pen portraits for customer groups at each level:

1. *Temporal profiles* – the temporal profiles of each customer segment were analysed by obtaining transactional frequencies over time. This aimed to understand if individuals exhibited temporal profiles consistent with the stores that they patronised. Data were acquired over 10-minute intervals for each customer and aggregated by Group assignments. These data were treated with the same rate calculation process as outlined in Chapter 5 (Section 5.2).
2. *Demographic profiles* – descriptive statistics were obtained regarding age and gender characteristics provided in the customer metadata. As noted in Chapter 2, this sample was predominantly female. However, analysis of gender attributes aimed to provide insight into particular profiles that may exhibit higher proportions of male customers.
3. *Spatial profiles* - the geographic distribution of segments was investigated via linkage to customer postcodes and obtaining the percentage of customer Supergroup/Group per area (percentages derived from total customers per area). In the proceeding results, statistics are presented regarding the number of postcode units in which each Supergroup/Group dominated. However, for visualising results, the percentage of customers per Supergroup/Group per MSOA is used. This was in order to illustrate the geographical distribution of the resulting customer segments whilst adhering to disclosure control restrictions. The smallest number of active customers per MSOA was 24, therefore, any instances where data referred to less than 10 customers were excluded.
4. *Product consumption profiles* - product consumption trends were quantified for each Group by obtaining purchase frequencies per customer Supergroup/Group at the lowest level of the product hierarchy (see Chapter 3, Section 3.2.3.2, for an overview of the product hierarchy structure). For commercial disclosure purposes, outputs from the product category analysis were aggregated to Level 4 (describes 221 categories) and 20 categories selected that illustrated the most prominent distinctions in behaviour. These were also assigned custom category names based on their collective attributes. Spend characteristics were also obtained, such as the number of transactions, total spend and spend per transaction.

Product data required a process of rate calculation in order to account for underlying trends. For example, specific product categories exhibited higher demand overall, which inhibited the

identification of the unique consumption patterns of individuals. Similarly to previous data treatment methods, this trend component could be accounted for by weighting values by their relevant denominator (total number of products bought per category). In addition, data were weighted by total transactions per customer, to account for variations in individual transactional volumes. This avoided over-indexing of products as a result of frequently transacting customers. These data could be used to infer product consumption values that were significantly above or below the equilibrium of their relevant conditions.

6.3. Results

6.3.1. Customer Supergroups

The segmentation of customers based on their store visiting behaviours was able to identify socially distinct groups of individuals. Differences could be observed between the geographical distributions (see Figures 6.5 to 6.12), demographic attributes, temporal profiles (see Figure 6.13) and product consumption preferences (see Figures 6.14 to 6.18) of each segment. Product trends were derived from the rate-calculated data, and thus describe the categories for which each segment showed proportionally higher volumes of consumption. Table 6.4 provides overall profile descriptions, and Table 6.5 provides statistics of the age, gender and spend characteristics of each Supergroup. These illustrate the average age, average total spend per person, average total transactions per person, average spend per transaction and the percentage of females and males present. These profiles can be broadly summarised as follows:

- **Supergroup 1 – 'Rural Ageing Off-peak Shoppers' -** represented the oldest, lowest spending and most rural living of HSR customers, who primarily transacted during 'off-peak' periods (weekday mornings/afternoons and late mornings on weekends) and consumed higher proportions of 'Ageing Healthcare' products.
- **Supergroup 2 – 'Small Destination Shoppers' -** the largest Supergroup who primarily resided in small town areas, of a middle-aged to ageing demographic, transacted during late mornings/afternoons on weekdays and late mornings on weekends (also off-peak), showed a preference for general healthcare and cosmetics products and demonstrated fewer transactions, but larger basket sizes (i.e. higher spend/product volumes per transaction).
- **Supergroup 3 – 'Weekday Convenience Commuters' -** the youngest and smallest customer Supergroup (yet comprising the highest proportion of male customers), who primarily resided in urban areas and accessible commuter towns of varying size. This Supergroup demonstrated weekday convenience usage (i.e. between business hours) with a preference for food and drink/convenience products and evening peaks on weekends with a preference for 'Family Planning', 'General Healthcare' and

'Toiletries'. These customers were the most frequently transacting, yet exhibited the lowest spend per transaction.

- **Supergroup 4 – 'Large Destination Shoppers' -** exhibited the second lowest average age, predominantly resided in suburban and urban areas, primarily transacted during evenings on weekdays and afternoon to evenings on weekends. These customers showed a preference for cosmetics, general healthcare and beauty accessories and showed the highest spend over fewer transactions (i.e. larger basket sizes, less frequent spending).

Table 6.4: Summary of customer Supergroup attributes.

Supergroup	Supergroup summary
1 - Rural Ageing Off-peak Shoppers **(Highest weekday patronage to store Supergroup 1** 13.8% of active customers 15% of postcodes	Represented the oldest and most rural living customers and were the second smallest Supergroup (13.8% of the active customer sample). These individuals primarily transacted during 'off-peak' periods, such as weekday mornings/afternoons and late mornings on weekends. Product consumption was higher for 'Ageing Healthcare' type products. This was the overall lowest spending group, showing the lowest average total spend per person and spend per transaction of all Supergroups.
2 - Small Destination Shoppers **(Highest weekday patronage to store Supergroup)** 42.6% of active customers 47.9% of postcodes	This was the largest of customer Supergroups (42.6% of the active sample), which is reflective of the HSR market as there are a higher number of these store types. Represented customers living primarily in small town areas of a middle-aged to ageing demographic. These individuals primarily transacted late mornings/afternoons during weekdays and late mornings on weekends (also off-peak during weekdays). Product consumption was higher for general healthcare and general cosmetics. These customers exhibited fewer transactions, but larger basket sizes.
3 - Weekday Convenience Commuters **(Highest weekday patronage to store Supergroup 4)** 7.7% of active customers 2.9% of postcodes	These customers exhibited top weekday membership to store Supergroup 4 (Weekday Convenience) who primarily resided in urban areas and commuter towns. These customers represented the youngest and smallest customer Supergroup (7.7% of the active customer sample), and described the highest proportion of male customers. Weekday temporal consumption represented convenience usage between business hours and evening peaks on weekends. Product consumption was highest for food and drink, convenience accessories (such as umbrellas and hosiery), family planning, general healthcare and toiletries. These customers exhibited the highest transactions per person but the lowest spend per transaction, consistent with

	convenience shopping trends (i.e. frequent, small transactions of low basket size).
4 - Large Destination Shoppers **(Highest weekday patronage to store Supergroup 5)** 37.2% of active customers 34.2% of postcodes	These customers predominantly lived in suburban/urban areas and exhibited the second lowest average age. This was the second largest group (37.2% of the active customer sample), and contained the largest volume of male customers. Temporal consumption showed evening peaks on weekdays and afternoon to evening peaks on weekends. Product consumption was highest for cosmetics (general and premium), general healthcare and beauty accessories. These customers showed the highest spend over fewer transactions (i.e. higher basket sizes, less frequent spending). This is consistent with destination shopping patterns that are typical of these large urban store types.

Table 6.5: Summary of customer spend attributes (mean per person, per Supergroup).

Supergroup	Age	Total spend (£)	Total Transactions	Spend per transaction (£)	Gender (%)	
					F	M
1	52	990	65	15	94.5	5.5
2	47	1173	68	18	94.3	5.7
3	36	1213	81	15	84.9	15.1
4	40	1351	69	20	93.6	6.4

Supergroup 1 – 'Rural Ageing Off-peak Shoppers'

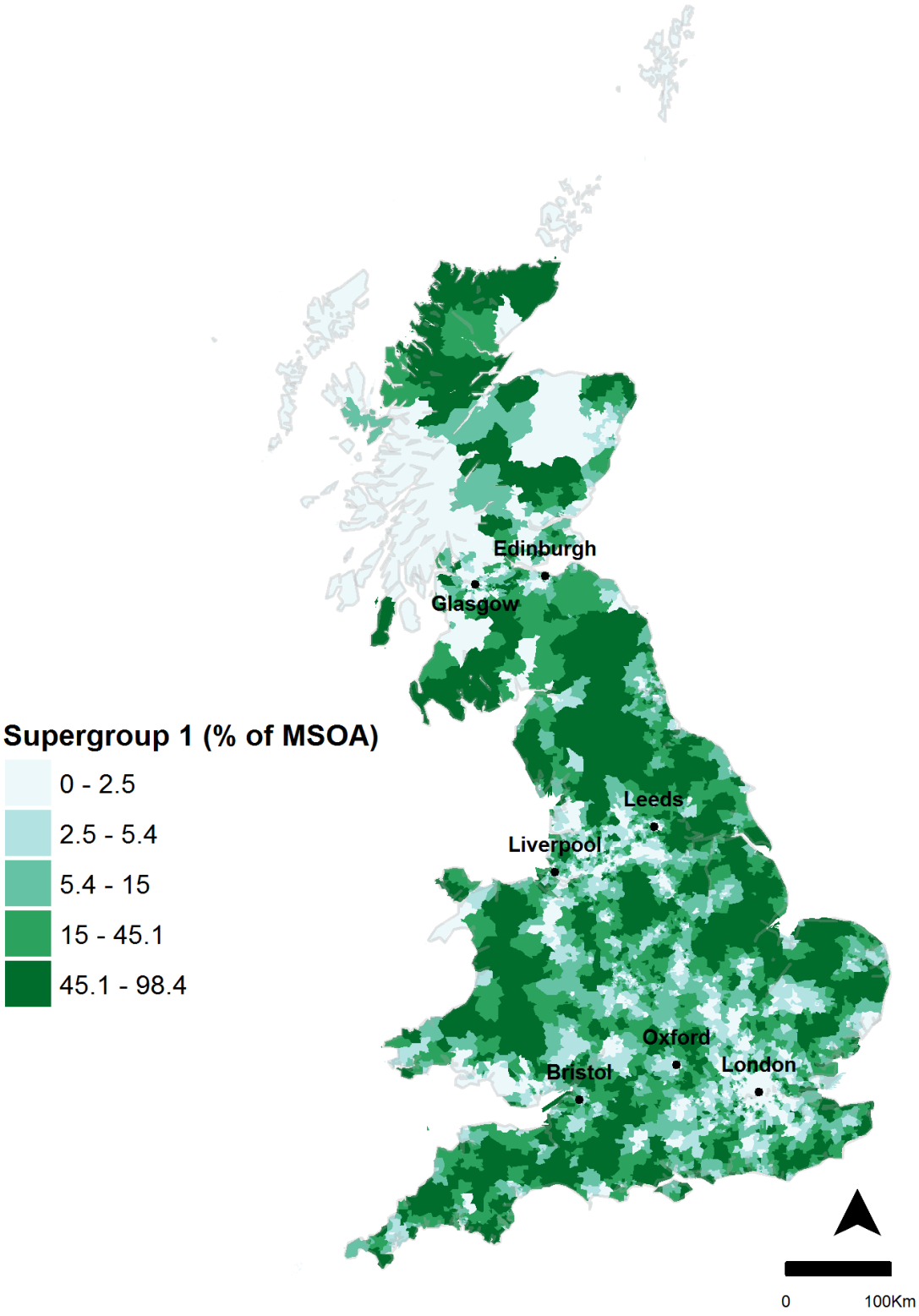


Figure 6.5: The percentage of customers in Supergroup 1 per MSOA, across Great Britain (quantile breaks).

Supergroup 1 – 'Rural Ageing Off-peak Shoppers'

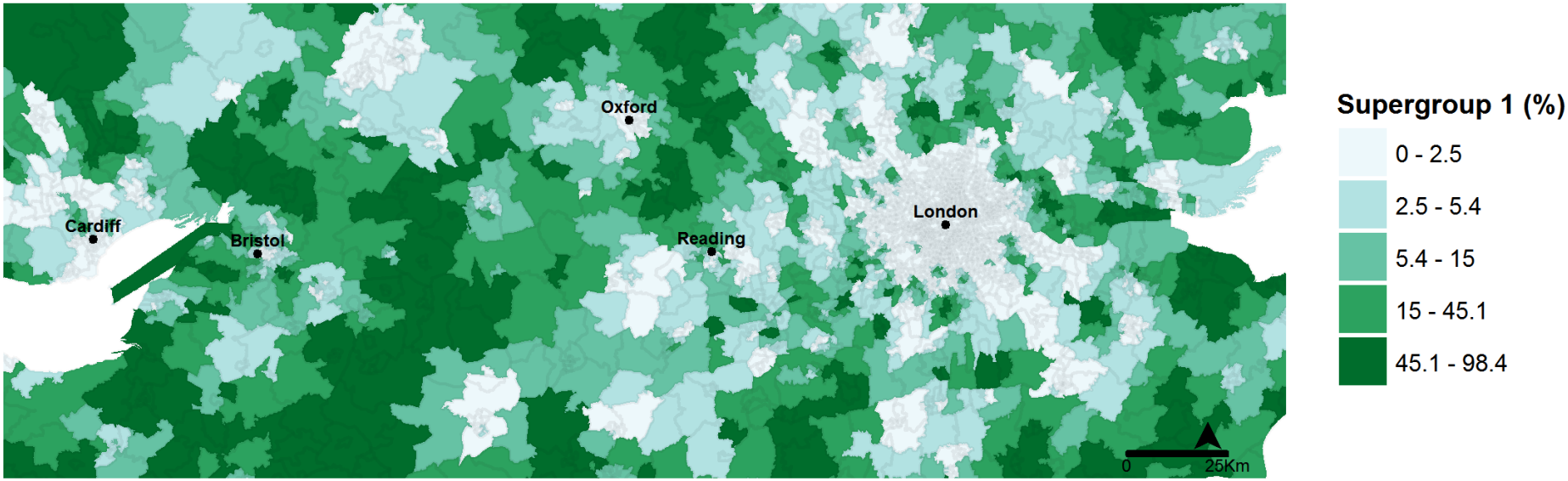


Figure 6.6: The percentage of customers in Supergroup 1 per MSOA across Southern England (quantile breaks).

Supergroup 2 – 'Small Destination Shoppers'

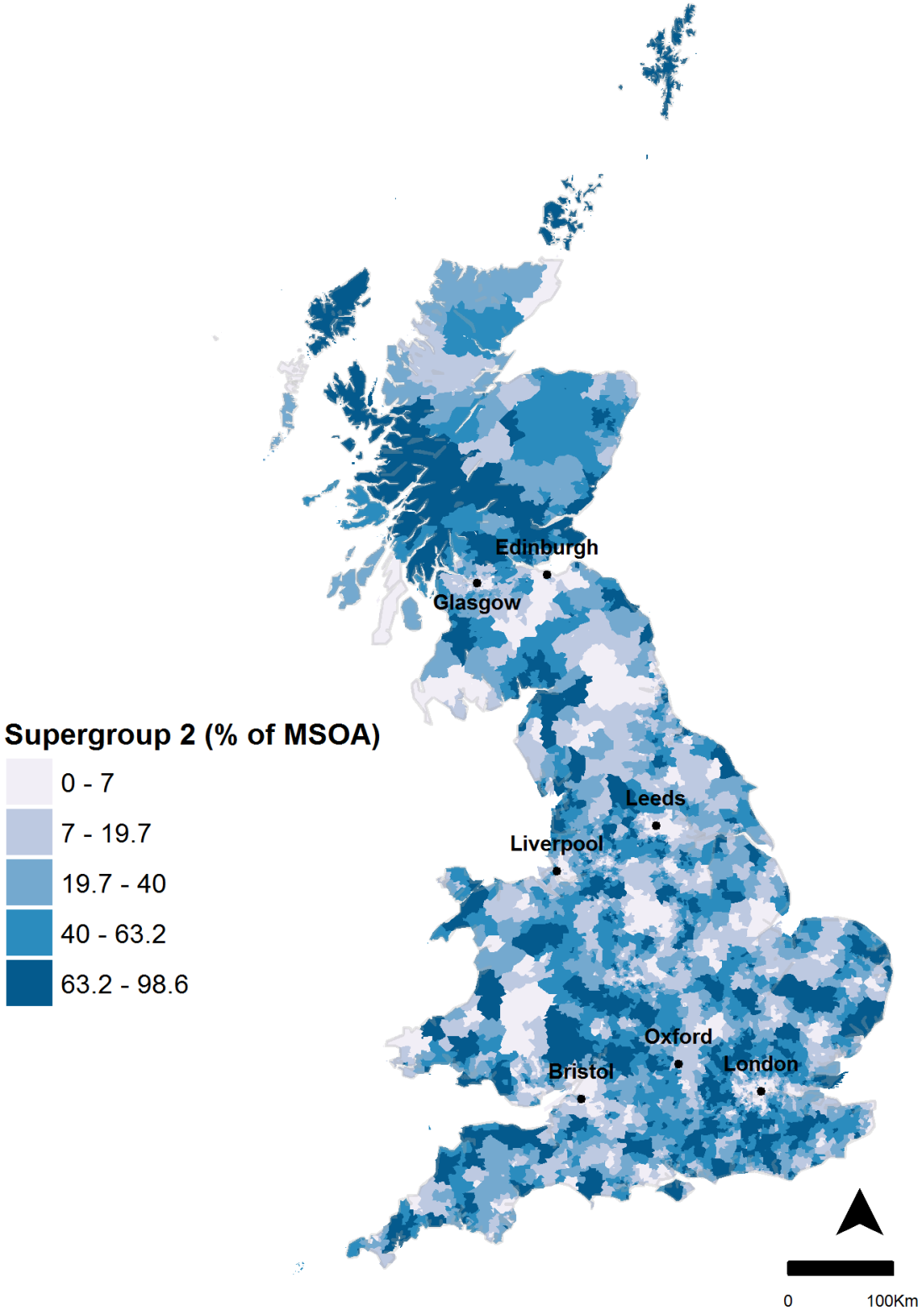


Figure 6.7: The percentage of customers in Supergroup 2 per MSOA across Great Britain (quantile breaks).

Supergroup 2 – ‘Small Destination Shoppers’

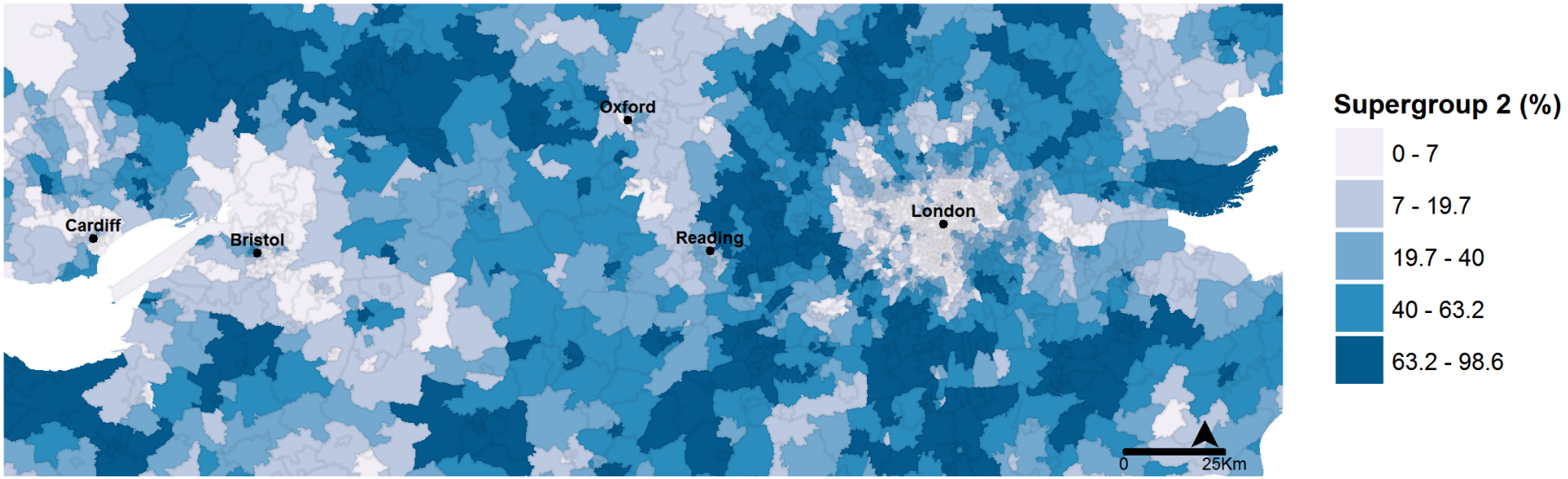


Figure 6.8: The percentage of customers in Supergroup 2 per MSOA across Southern England (quantile breaks).

Supergroup 3 – 'Weekday Convenience Commuters'

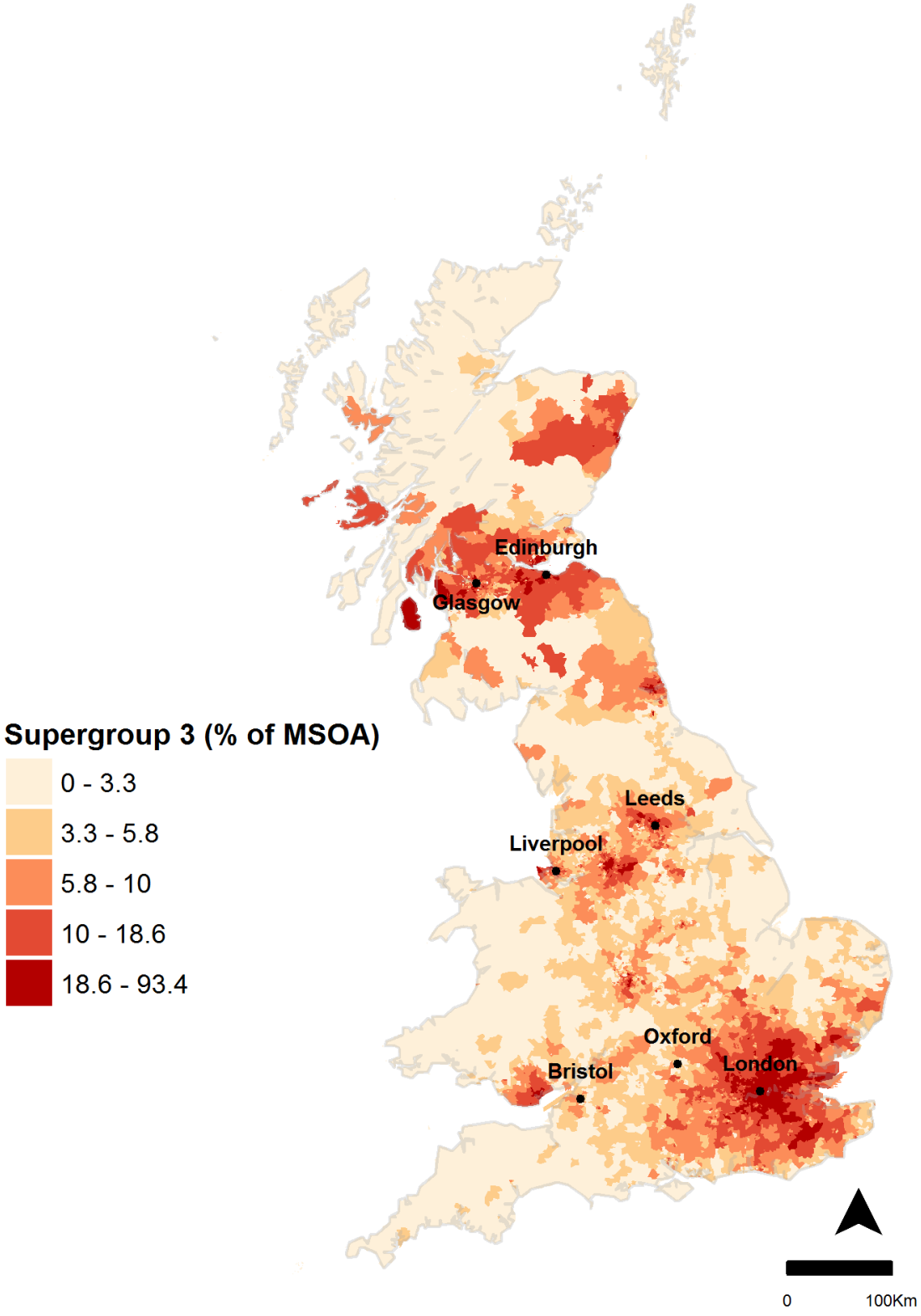


Figure 6.9: The percentage of customers in Supergroup 3 per MSOA across Great Britain (quantile breaks).

Supergroup 3 – 'Weekday Convenience Commuters'

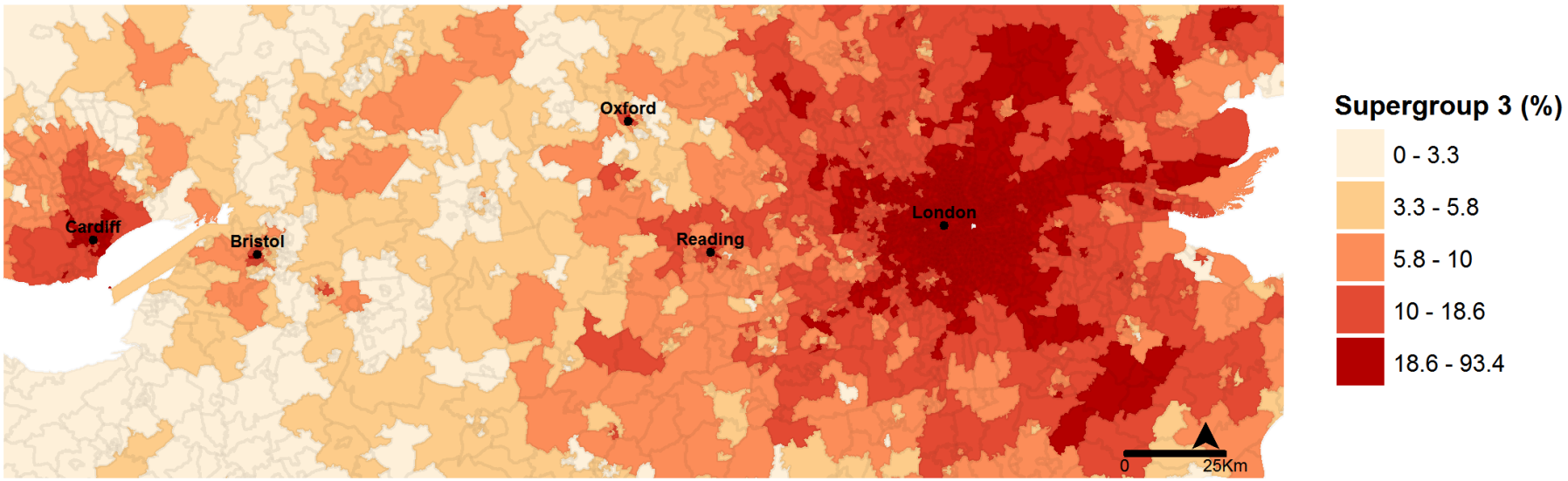


Figure 6.10: The percentage of customers in Supergroup 3 per MSOA, across Southern England (quantile breaks).

Supergroup 4 – ‘Large Destination Shoppers’

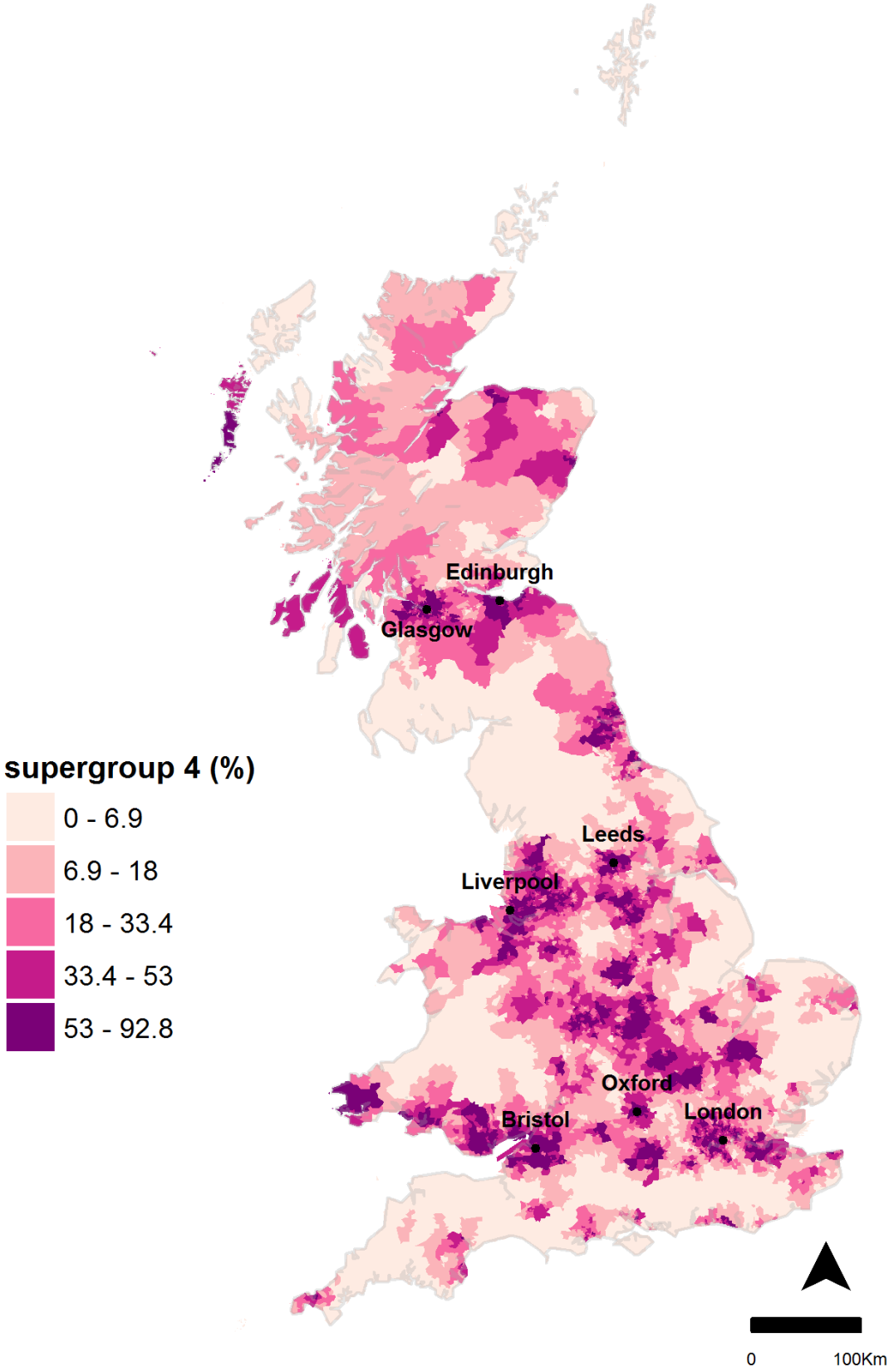


Figure 6.11: The percentage of customers in Supergroup 4 per MSOA across Great Britain (quantile breaks).

Supergroup 4 – ‘Large Destination Shoppers’

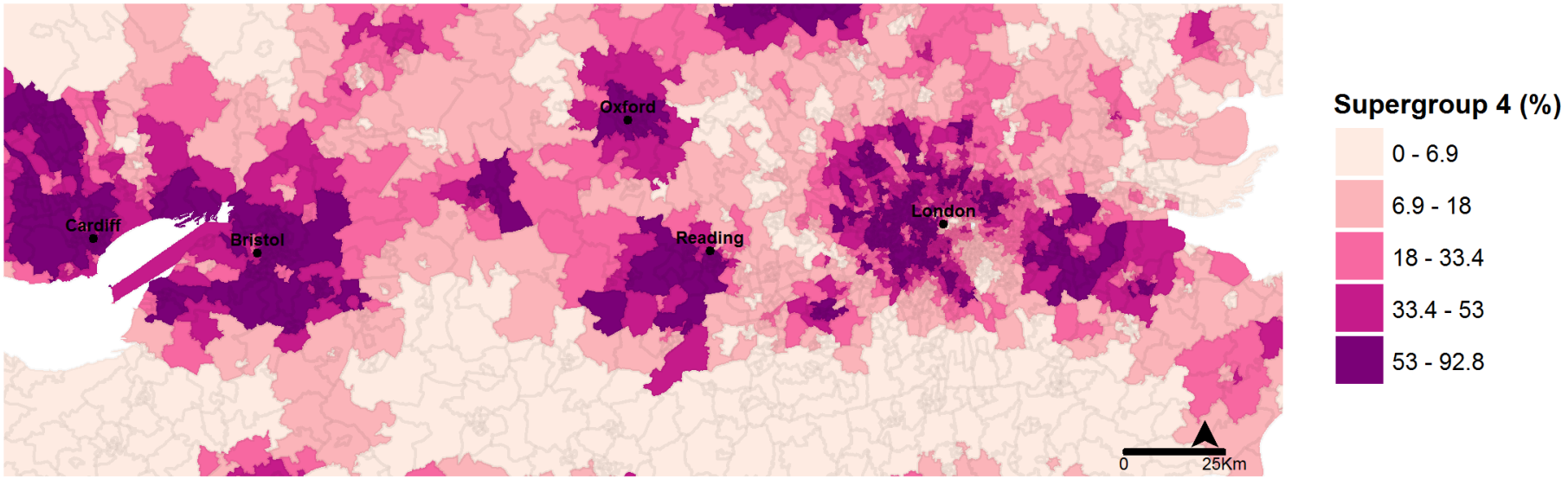
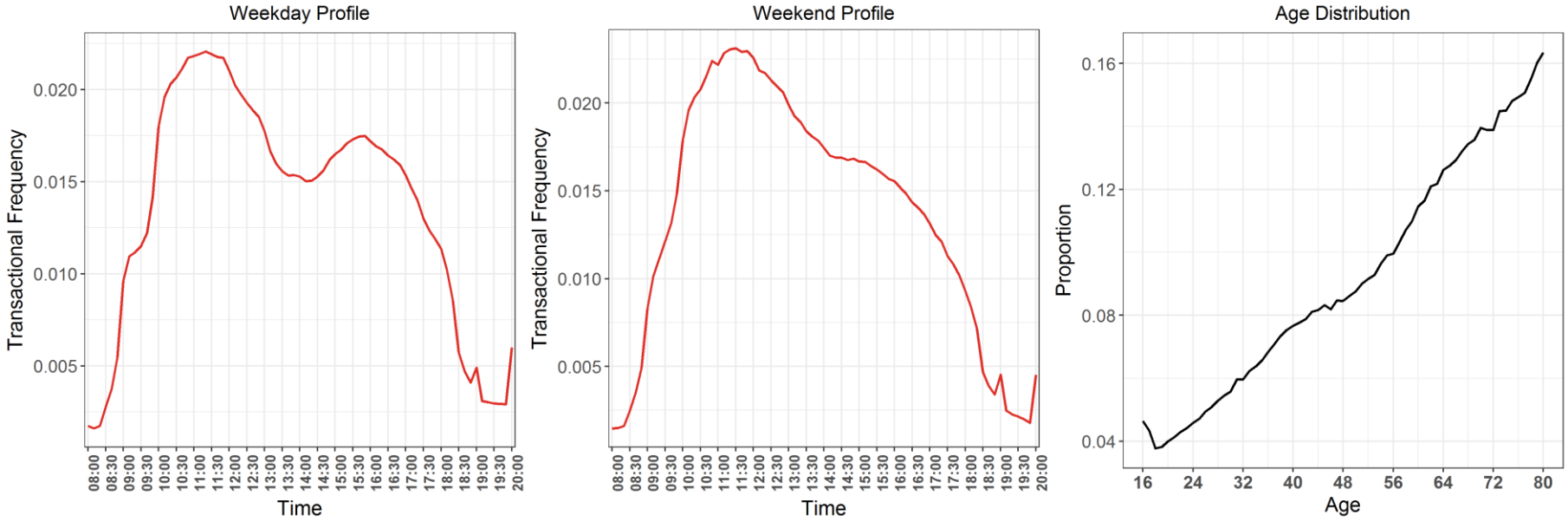


Figure 6.12: The percentage of customers in Supergroup 4 per MSOA, across Southern England (quantile breaks).

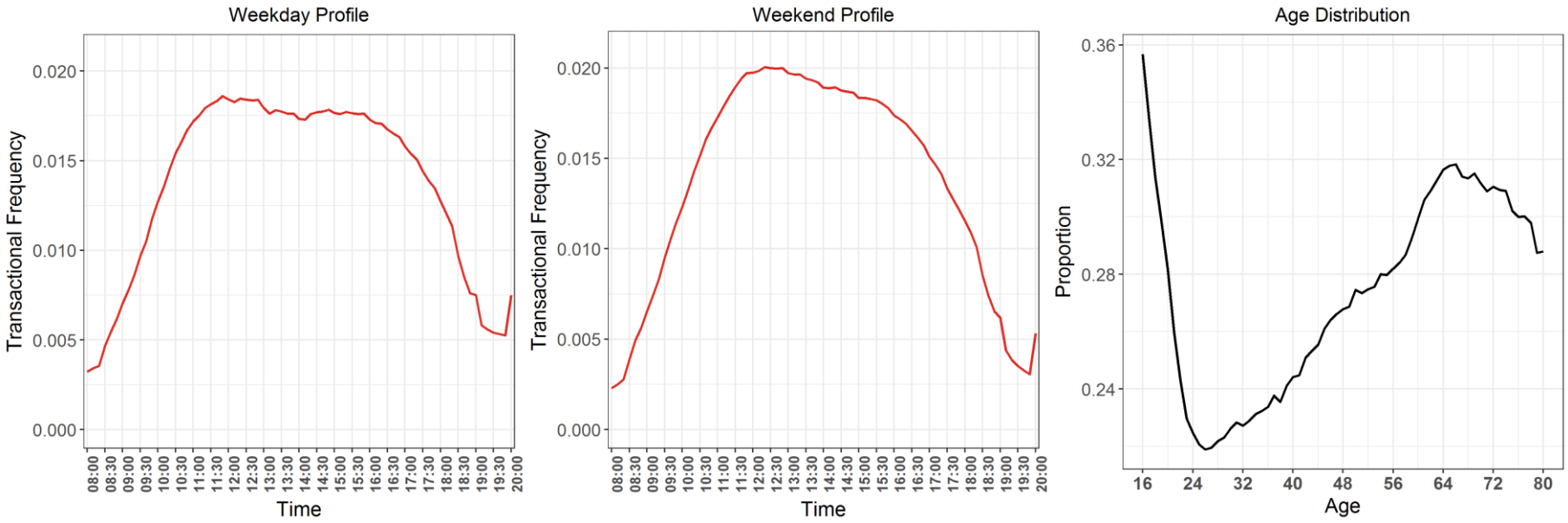
Supergroup 1 – 'Rural Ageing Off-peak Shoppers'.

a)



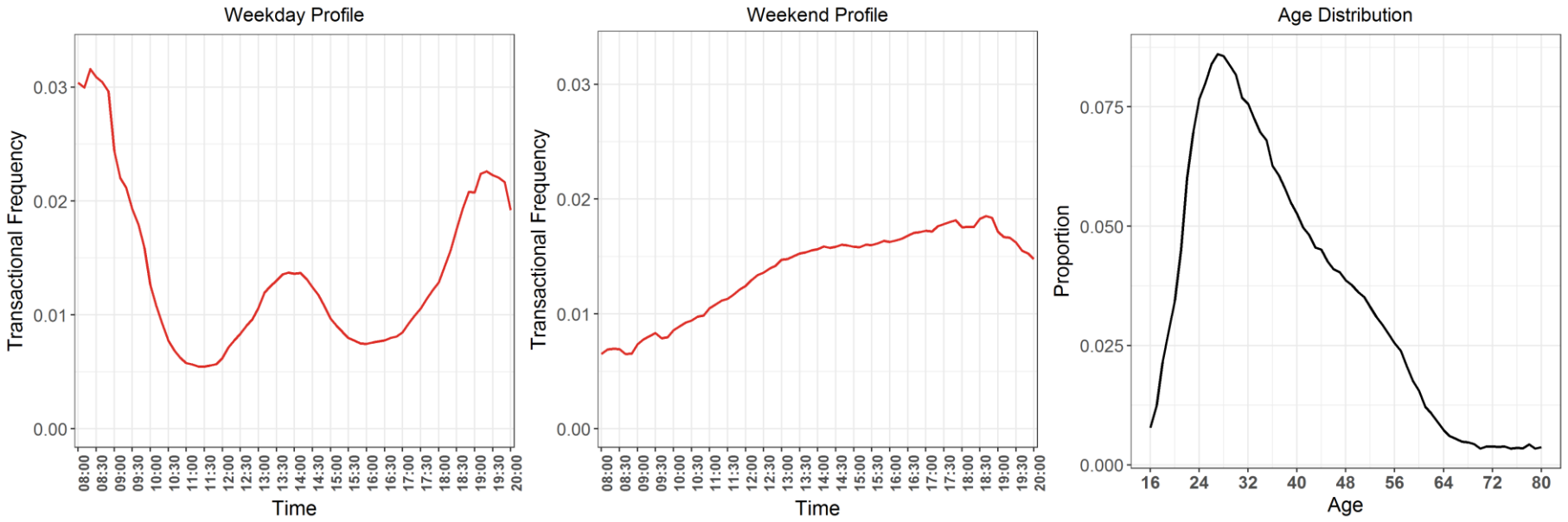
Supergroup 2 – ‘Small Destination Shoppers’.

b)



Supergroup 3 – ‘Weekday Convenience Commuters’.

c)



Supergroup 4 – 'Large Destination Shoppers'.

d)

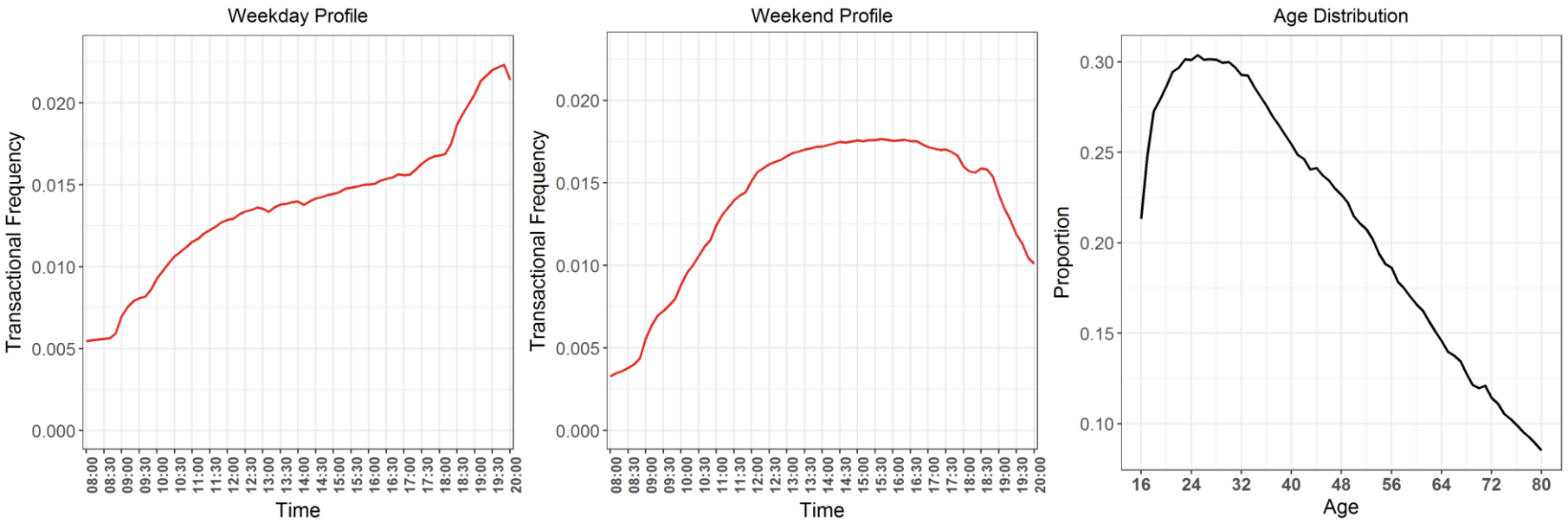


Figure 6.13: Time profiles (weekday, weekend, 10-minute intervals) and age distributions per customer Supergroup.

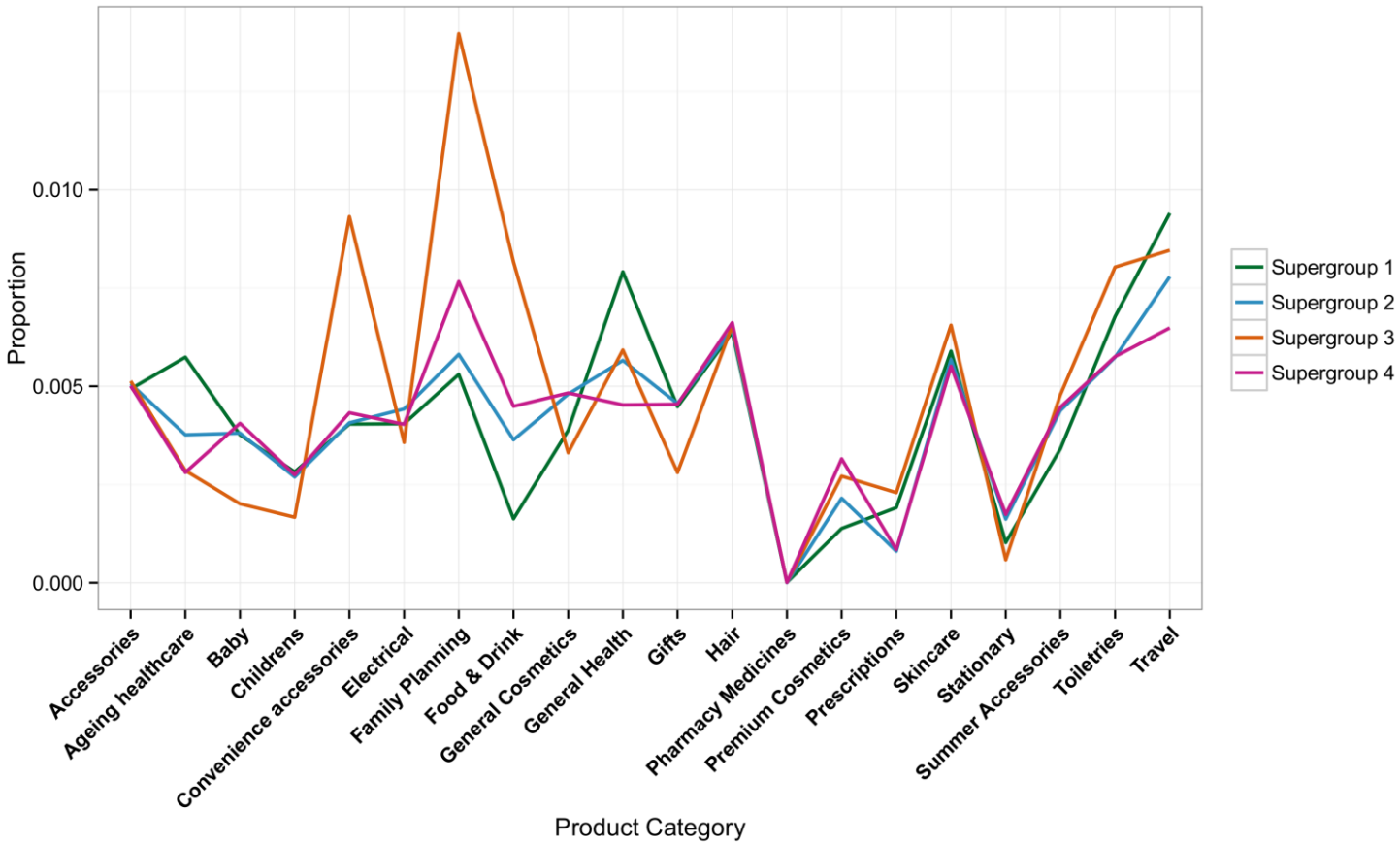


Figure 6.14: Product comsumption comparison across Supergroups.

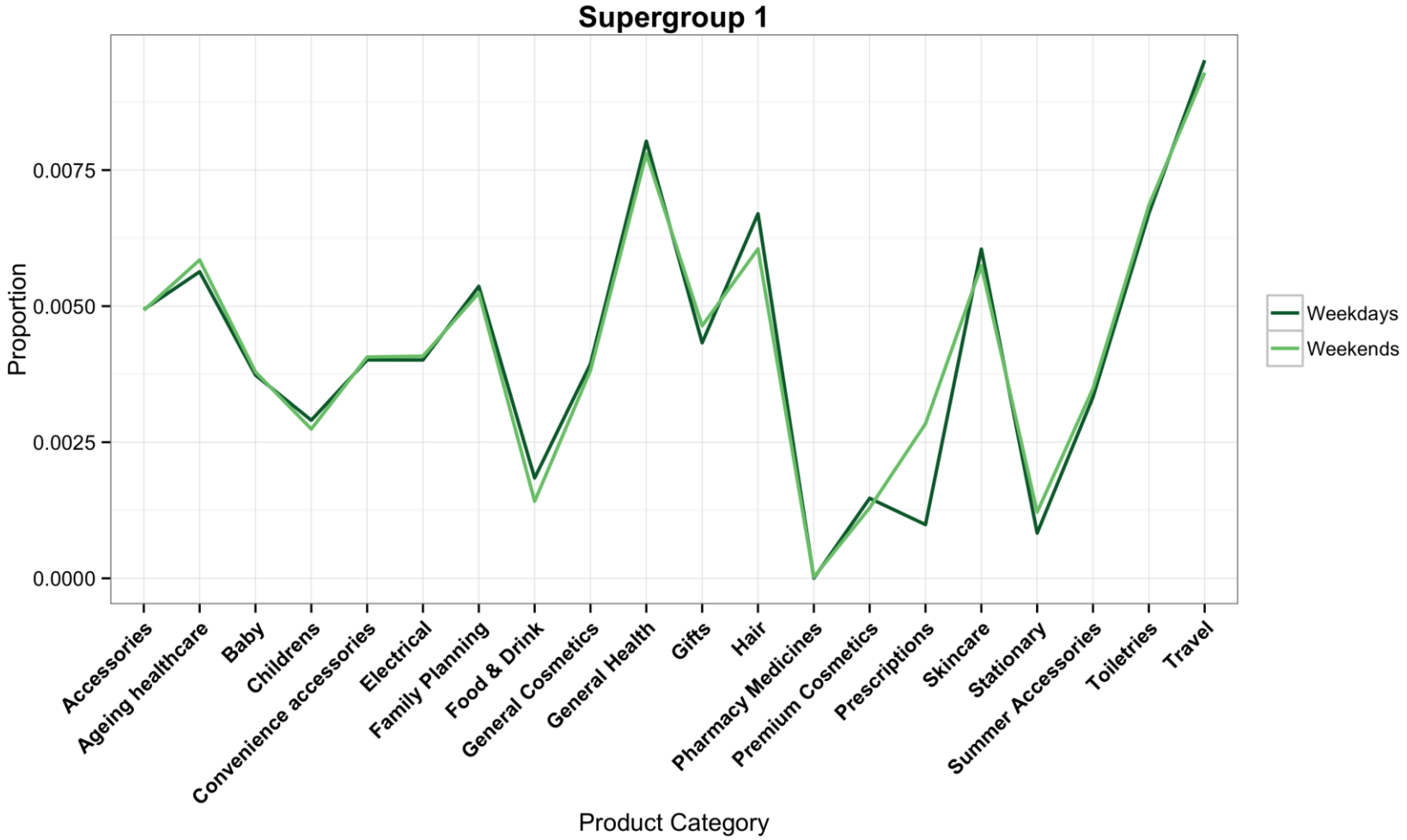


Figure 6.15: Weekday versus weekend product consumption comparison for Supergroup 1, (‘Rural Ageing Off-peak Shoppers’).

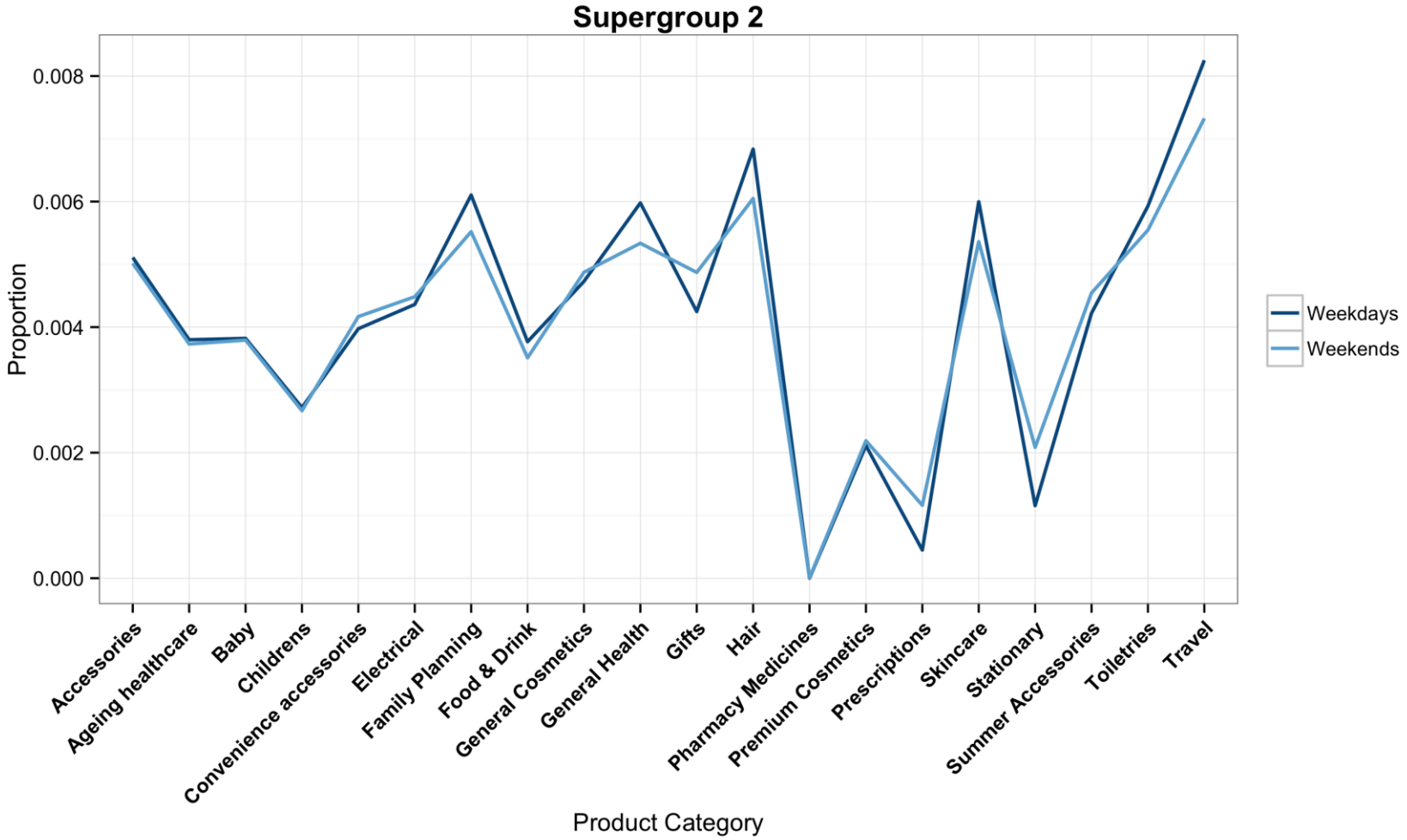


Figure 6.16: Weekday versus weekend product consumption comparison for Supergroup 2, ('Small Destination Shoppers').

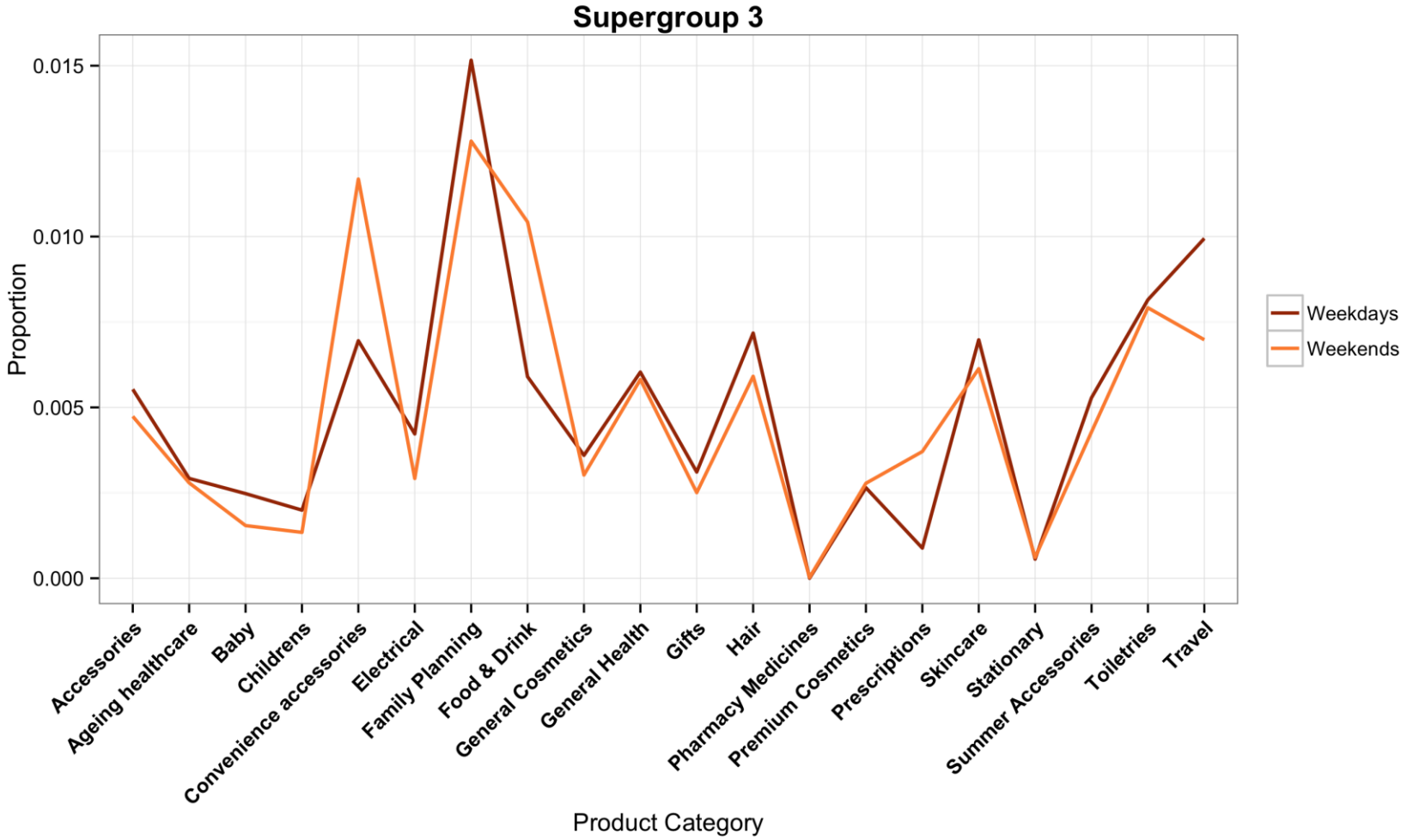


Figure 6.17: Weekday versus weekend product consumption comparison for Supergroup 3, ('Weekday Convenience Commuters').

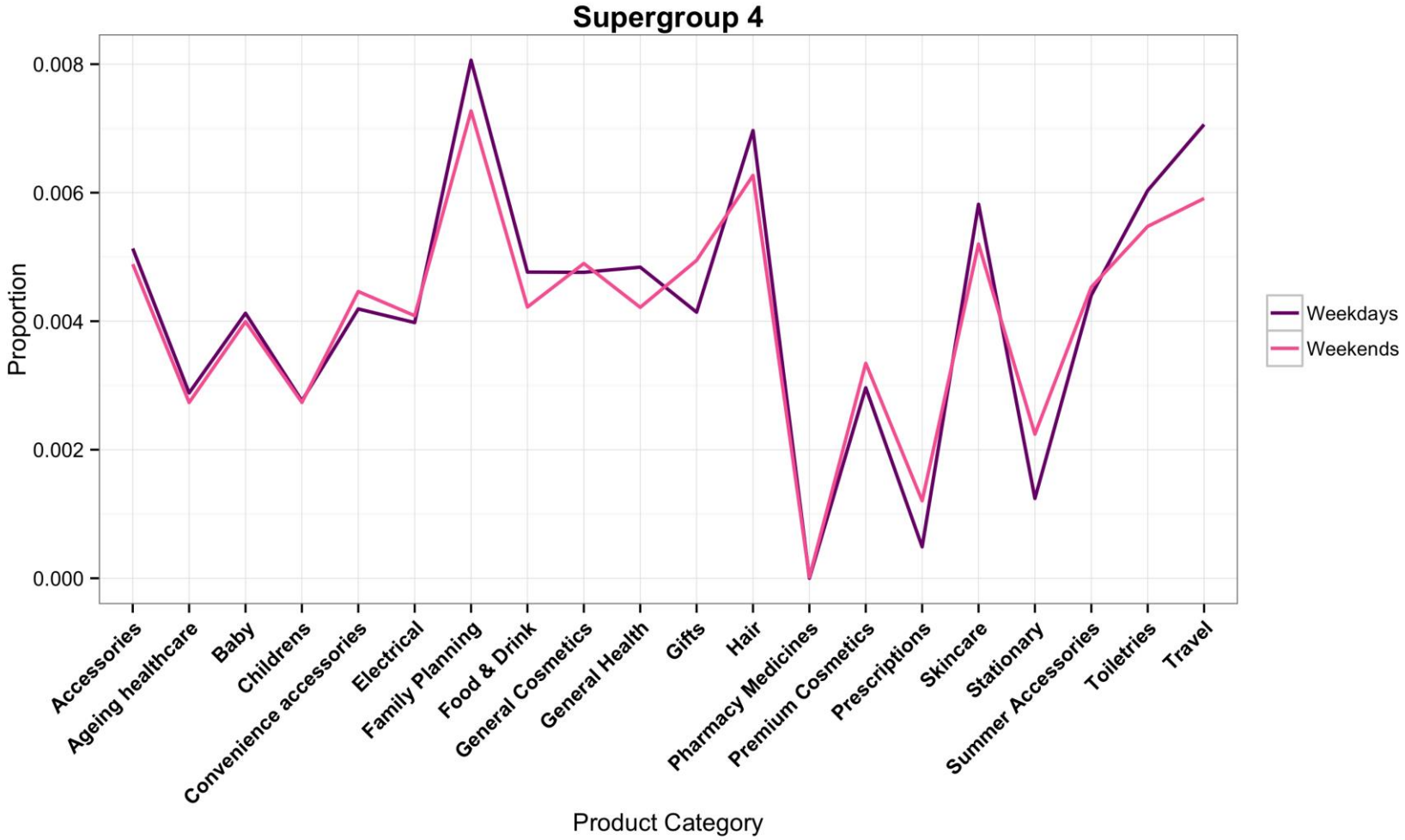


Figure 6.18: Weekday versus weekend product consumption comparison for Supergroup 4, ('Large Destination Shoppers').

Exploration of the differences between Supergroup characteristics suggested that the time and location that HSR customers transact may be predictive of their geodemographic attributes. For example, Supergroup 1 customers, grouped based on their shared primary usage of stores exhibiting off-peak shopping trends, described individuals of an ageing demographic who predominantly resided in the most rural areas. Supergroup 3 customers, grouped based on their shared primary usage of stores exhibiting weekday convenience trends, delineated younger customers living primarily in urban areas, who demonstrated significantly different product consumption and frequency of spend characteristics.

The product consumption analysis revealed a number of distinct patterns. Supergroup 1 demonstrated notably higher consumption of 'Ageing Healthcare' products. This included higher proportions of items such as anti-wrinkle skincare, heart health, hearing care, hair loss and incontinence products in comparison to other segments of the HSR population. Conversely, Supergroup 3 demonstrated higher consumption of food and drink, convenience accessories, which included items such as umbrellas and hosiery and 'Family Planning' related products. Supergroup 2's most prominent consumption was within 'General Healthcare' (including products such as painkillers, dental, first aid, coughs/colds and vitamins/supplements) and toiletries (i.e. deodorants, bath/shower, shampoo and conditioner). Supergroup 4 demonstrated the lowest consumption of healthcare oriented products and higher proportions of beauty and cosmetics, baby products and 'Family Planning'. These trends are investigated further in the proceeding Group level analysis.

Distribution measures provided insights regarding variation in behaviour within Supergroups, such as the most likely alternative store types to be patronised during weekdays and weekends. Figure 6.19 demonstrates the most prominent interactions, as measured by volumes of overall second-ranking store types within each segment. Activity was recorded within all types (which is expected when observing activities over a longitudinal period), however, varying preferences were evident. For instance, Supergroup 1 customers showed the highest likelihood to visit store Supergroup 2 ('Weekend Peak Destinations'), and Supergroup 2 customers showed the highest likelihood to patronise store Supergroup 5 ('Stable Destinations'). Observing these trends in combination with the geographical distribution of customer Supergroups suggested that propensity to visit alternative store types may be related to the accessibility of locations. For instance, members of Supergroup 1 (the most rural customers) were most likely to visit local small town destinations and Supergroup 2 customers, of whom were located in more accessible locations to urban areas showed a higher preference for larger, more urban destinations.

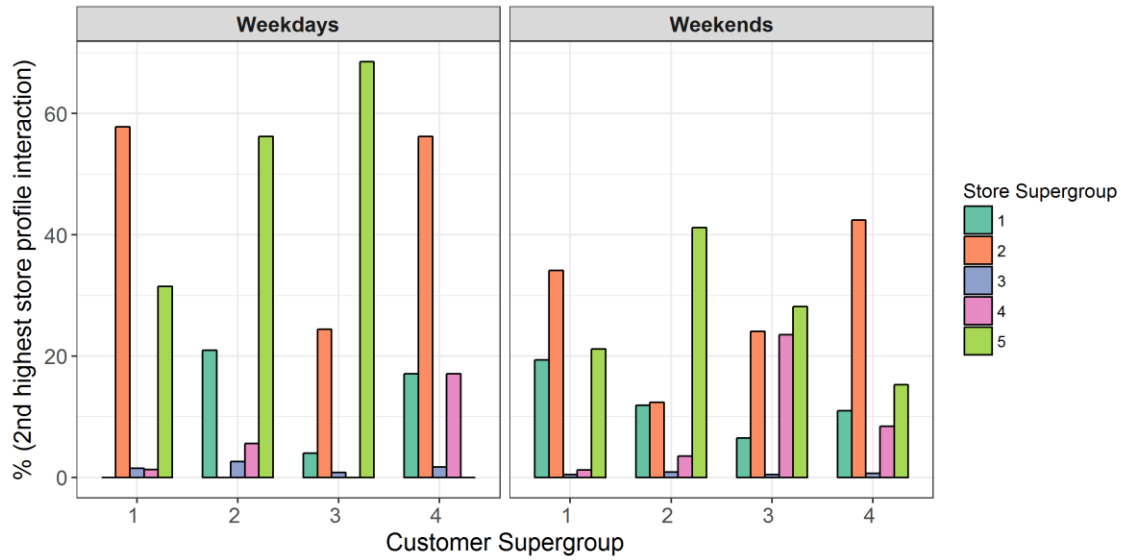


Figure 6.19: Percentage of interactions with second-ranking store profiles, per customer Supergroup.

Overall, Supergroup 2 customers showed the least distributed behaviour, suggesting that these customers were more likely to shop at the same store type (or location) during both weekday and weekend periods. It could be speculated that such dynamics are a reflection of differing retail compositions between the store types most prominent to each Supergroup. For example, customer Supergroup 2's primary stores were typically medium-large sized stores with larger retail offerings, of which may fulfill the majority of needs to this segment of customers. In contrast, Supergroup 1's primary stores were smaller, with more specific retail offerings (i.e. community stores or local chemists), which may encourage travel to alternative destinations in order to fulfil certain retail needs. These dynamics are explored further in Chapter 7.

6.3.2. Customer Groups

At the Group level, further differences could be identified between the geographical distributions, demographic attributes, temporal profiles and product consumption preferences of each Supergroup. Table 6.6 provides overall profile descriptions and Table 6.7 provides statistics of the age, gender and spend characteristics of each Group. For ease, profiles are summarised as follows:

- **Supergroup 1 ('Rural Ageing Off-peak Shoppers')**
    - **Group 1a – 'Stable Rural Ageing Health' –** exhibited the highest average age of Supergroup 1, the lowest spend and resided in the most remote areas. Patronage to alternative store types was minimal, activity was during off-peak periods and consumption was high for ageing healthcare products.

- **Group 1b – 'Rural, Weekend Small-town Shoppers' –** exhibited a high average age, lived in rural areas but demonstrated highest patronage to surrounding small town destinations on weekends. Activity was during off-peak periods and consumption suggested high usage for prescription collection during weekdays and healthcare essentials on weekends.
- **Group 1c – 'Rural Fringe, Urban Destination Shoppers' –** exhibited a high average age (but lowest of Supergroup 1)**,** primarily resided in 'accessible' rural locations (i.e. rural fringe) and demonstrated the highest patronage to urban destinations on weekends. Activity was during off-peak periods and consumption suggested high usage for prescription collection during weekdays and healthcare, cosmetics and children's products during weekends. This was the highest spending Group of Supergroup 1.

- **Supergroup 2 ('Small Destination Shoppers')**
  - **Group 2a – 'Rural, Weekday Small-town Shoppers' –** exhibited the highest average age of Supergroup 2, lowest spend and were geographically clustered on rural fringes surrounding small town areas. Activity was off-peak and primarily within small town stores during weekdays (rural stores during weekends). Product consumption indicated a mix of healthcare essentials and cosmetics during weekdays and primarily cosmetics on weekends.
  - **Group 2b – 'Stable Small-town Shoppers' –** the largest group that enclosed the majority of areas surrounding and within small towns. Activity was off-peak during weekdays (decrease at lunchtime) and late mornings on weekends. Patronage to alternative store types was minimal and product consumption was parallel with Group 2a (healthcare essentials and cosmetics based).
  - **Group 2c – 'Small-town, Weekend Urban Destination Shoppers' –** exhibited the lowest average age of Supergroup 2, the highest spend and were geographically clustered around small towns on urban fringes. Activity was during peak periods (weekday lunchtimes and weekend afternoons) and patronage was highest to large urban destination stores during weekends. Consumption was high for cosmetics on weekdays and a mix of cosmetics and general healthcare on weekends.
- **Supergroup 3 ('Weekday Convenience Commuters')**
  - **Group 3a – 'Rural Fringe Commuters' –** exhibited the highest average age of Supergroup 3, urban convenience shopping trends during weekdays and highest patronage to rural stores on weekends. Activity demonstrated the highest early morning peak of the commuter groups during weekdays, and mid-day peaks on weekends. Consumption indicated food and convenient essentials during the

week, and healthcare essentials, 'Travel' and 'Childrens' products on weekends.

- **Group 3b – 'Small-town Commuters' –** represented commuters residing in small town locations, who demonstrated convenience patterns during weekdays and afternoon peaks on weekends. Product consumption was highest for food and convenience essentials during the week, and 'Family Planning', healthcare essentials/toiletries on the weekends.
- **Group 3c – 'Stable Urban Workers' –** represented commuters residing in suburban/urban areas and had the second lowest average age of Supergroup 3, who demonstrated evening convenience patterns during weekdays and evening peaks on weekends. Patronage to alternative store types was minimal and product consumption was high for food and convenient essentials on both weekdays and weekends. This Group also exhibited the highest average transactions per person of all Groups but the lowest spend.
- **Group 3d – 'Urban-living, Weekend Destination Shoppers'** – exhibited the lowest average age of all Supergroups, primarily resided in urban areas, demonstrated convenience patterns during weekdays and patronised large urban destinations (i.e. flagships) during weekends. Product consumption was high for food and essentials during weekdays, and 'Family Planning', toiletries and convenient essentials during weekends. This Group exhibited the highest overall spend and spend per transaction of Supergroup 3.

- **Supergroup 4 ('Large Destination Shoppers')**
  - **Group 4a – 'Rural Fringe, Weekday Destination Shoppers'** – exhibited the highest average age of Supergroup 4, primarily patronised rural fringe retail park stores (store Group 5b) during weekdays and demonstrated the highest spend of those residing in rural areas. Activity was off-peak during weekdays (mid-morning to afternoon) and mid-morning to evening on weekends. Product consumption was high for premium cosmetics during weekdays and healthcare essentials during weekends.
  - **Group 4b – 'Urban Fringe, Weekday Destination Shoppers' –** exhibited the second highest average of Supergroup 4, who primarily patronised urban fringe retail park stores (store Group 5a) during weekdays. Activity was during peak periods on weekdays (lunchtimes and evenings) and mid-day to afternoon on weekends. Product consumption was high for cosmetics and gifting during weekdays and cosmetics and healthcare essentials during weekends.
  - **Group 4c – 'Urban Weekday Destination Shoppers' –** exhibited the lowest average age of Supergroup 4**,** primarily resided in urban areas and patronised

the ‘Urban Stable Destination’ store types (Group 5c) on weekdays. Activity demonstrated convenience patterns (parallel to Supergroup 3) on weekdays and evening peaks on weekends. Product consumption was high for food and drink and convenient essentials during weekdays and food and drink, ‘Family Planning’ and healthcare on weekends. This Group demonstrated a high number of transactions but low spend per person.

- **Group 4d – ‘Stable Urban Destination Shoppers’ –** exhibited a low average age, primarily resided in suburban areas and exhibited the highest spend per transaction of all Supergroups. Activity was primarily within the retail park oriented store types, and patronage to alternative store types was minimal. Weekdays demonstrated high afternoon/evening activity and weekend afternoon to evening peaks. Product consumption was highest for cosmetics, beauty accessories and ‘Family Planning’.

Table 6.6: Summary of Group level attributes.

Group	Group summary
1a **Stable Rural Ageing Health**	Demonstrated the highest average age of Supergroup 1, resided in the most remote areas and showed minimal patronage to alternative store types. Product consumption was highest for 'Ageing Healthcare' and 'General Healthcare' (on both weekdays and weekends) and customers exhibited off-peak transactional behaviour, such as during weekday and weekend mornings. This was the lowest spending Group.
1b **Rural, Weekend Small-town Shoppers**	Rural living customers that demonstrated highest patronage to surrounding small town destinations on weekends (store Supergroup 2). Weekday product consumption suggested high usage for prescription collection, general health and general cosmetics. On the weekend, 'General Healthcare' was prominent. Peak consumption times were weekdays late morning and weekends late morning to evening.
1c **Rural Fringe, Urban Destination Shoppers**	Demonstrated the lowest average age of Supergroup 1, primarily resided on rural fringes and demonstrated their highest patronage to accessible urban destinations on weekends. Weekday product consumption also demonstrated high usage for prescription collection but also, 'General Health', 'General Cosmetics' and 'Childrens'. On the weekend, consumption was high for higher value categories such as 'Premium Cosmetics' and 'Electrical'. Peak consumption times were late morning to afternoon, and weekend afternoons. This was the highest spending and youngest Group of Supergroup 1.
2a **Rural, Weekday Small-town Shoppers**	Rural living customers who demonstrated highest patronage to small town stores on weekdays. This Group had the highest average age of Supergroup 2, the lowest spend and were geographically clustered around small town areas. Peak consumption was late morning to mid-day during weekdays, and during early mornings on weekends. Product consumption indicated a mix of healthcare essentials and cosmetics during weekdays and primarily cosmetics on weekends.
2b **Stable Small-town Shoppers**	This was the largest Group that covered the majority of areas surrounding and within small towns. This Group showed minimal patronage to alternative store types and had a similar product consumption profile to Group 2a (healthcare essentials and cosmetics

	based). Peak consumption times were weekday mid-morning and afternoon (decrease at lunchtime) and late mornings on weekends.
2c **Small-town, Weekend Urban Destination Shoppers**	Exhibited the lowest average age and highest spend of Supergroup 2, and highest patronage to large urban destination stores during weekends (store Supergroup 5). This Group was predominantly clustered around the fringes of large urban areas. Weekday consumption was high for both general and premium cosmetics and on weekends a mix of cosmetics and general healthcare. Peak activity was weekday lunchtimes and weekend afternoons.
3a **Rural Fringe Commuters**	This Group exhibited the highest average age of Supergroup 3, most rural residential locations and demonstrated convenience trends during weekdays yet showed highest membership to rural Supergroup 1 stores on weekends. They demonstrated the highest early morning peak of the commuter groups during weekdays, and mid-day peaks on weekends. Product consumption indicated food and convenient essentials during the week, and healthcare essentials, 'Travel' and 'Childrens' products on weekends.
3b **Small-town Commuters**	This Group represented commuters residing in small town locations. These customers demonstrated convenience patterns during weekdays (morning, lunchtime, evenings - highest during early morning), and afternoon peaks on weekends. Product consumption was highest for food and convenience essentials during the week, and 'Family Planning', healthcare essentials/toiletries on the weekends.
3c **Stable Urban Workers**	These customers primarily resided in suburban/urban areas and had the second lowest average age of Supergroup 3. This Group showed minimal patronage to alternative store types. Weekday consumption demonstrated convenience patterns that peaked during evenings and also evening peaks on weekends. Product consumption was high for food and convenient essentials on both weekdays and weekends. This Group exhibited the highest average transactions per person of all groups but the lowest spend.
3d **Urban-living, Weekend Destination Shoppers**	These customers exhibited the lowest average age of all Supergroups and primarily resided in urban areas, showing convenience patterns during weekdays and patronising large urban destinations (i.e. flagships) on weekends. Weekday activity peaks were early morning, lunchtimes and evenings, and evenings on weekends. Product consumption was high for food and essentials during weekdays, and 'Family Planning',

	toiletries and convenient essentials during weekends. These customers exhibited the highest overall spend and spend per transaction of Supergroup 3.
4a **Rural Fringe, Weekday Destination Shoppers**	This Group exhibited the highest average age of Supergroup 4 and represented those patronising rural fringe retail park stores (store Group 5b) during weekdays. Consumption peaks were mid-morning to afternoon during weekdays and mid-morning to evening on weekends. They demonstrated the highest spend of those that live in rural areas. Product consumption was high for cosmetics during weekdays (premium and general) and healthcare essentials during weekends.
4b **Urban Fringe, Weekday Destination Shoppers**	This Group exhibited the second highest average age of Supergroup 4 and represented those patronising the urban fringe retail park stores (store Group 5a) during weekdays. Activity peaked during lunchtimes and evenings weekdays and mid-day to afternoon on weekends. Product consumption was high for cosmetics and gifts during weekdays and cosmetics and healthcare essentials during weekends.
4c **Urban Weekday Destination Shoppers**	This Group exhibited the lowest average age of Supergroup 4, primarily resided in urban areas and patronised the destination-convenience mix store types (Group 5c) during weekdays. Temporal consumption demonstrated convenience patterns on weekdays (morning, lunchtime and evening peaks – similarly to Supergroup 3 customers) and weekend evening peaks. Product consumption was high for food and drink and convenient essentials during weekdays (also indicating a convenience group) and food and drink, family planning and healthcare on weekends. These customers demonstrated a high number of transactions but low average spend per person.
4d **Stable Urban Destination Shoppers**	The largest of this Supergroup, these customers primarily patronised the same store type during weekdays and weekends. This included all 3 Group types in store Supergroup 5, but predominantly the retail park oriented Groups. These customers primarily resided in suburban areas and exhibited the highest spend per transaction of all Supergroups. Weekdays demonstrated high afternoon/evening activity (evening peak) and weekend afternoon to evening peaks. Product consumption was highest for cosmetics, beauty accessories and family during the week, and similar consumption on weekends.

Table 6.7: Summary of customer Group attributes (mean per variable).

Supergroup	Age	Total spend (£)	Total Transactions	Spend per transaction (£)
1a	54	940	64	14
1b	48	1057	63	16
1c	45	1152	64	18
2a	54	1027	62	17
2b	47	1173	68	18
2c	43	1220	66	19
3a	42	1093	76	15
3b	38	1231	81	15
3c	36	1136	83	14
3d	34	1238	80	16
4a	49	1196	63	20
4b	41	1277	65	20
4c	36	1043	70	16
4d	40	1376	70	21

These findings suggested that relationships exist between spatiotemporal consumption patterns and customer characteristics. The first key trend identifiable from these data was that the weekend profile of a customer appeared predictive of their likely area of residence, which in turn facilitated identification of demographic attributes. For example, in each Supergroup, the Group exhibiting the highest weekend interaction with rural stores (store Supergroup 1 – 'Weekday Off-peak Shopping') had the highest average age and the most rural postcodes. Conversely, those showing the highest weekend interactions with urban stores (store Supergroups 4 and 5) exhibited the lowest average age and the most suburban/urban postcodes.

The second key trend was that customers' weekday profile was indicative of where, and when, their weekday activities took place, providing further insight into their likely characteristics. For example, those primarily patronising 'Weekday Convenience' stores exhibited trends consistent with a working population (i.e. between business hours, in urban locations), whereas Supergroup 1 customers demonstrated opposing activity, during off-peak times (i.e. within business hours) in rural locations. Thus, the interaction of weekday and weekend profiles provided insights into both weekday activities and residential location types, which subsequently segmented customers into groups with distinct geodemographic attributes.

A prominent example is that of Supergroup 3 - 'Weekday Convenience Commuters' - where individuals who exhibited the same weekday activity patterns could be differentiated by the type of area they were likely commuting from through the addition of their weekend profiles. This was able to identify the older rural living versus younger urban living commuters. Furthermore, there were substantial differences between the product consumption patterns of these Groups, suggesting that demographic attributes may be correlated with increased levels of consumption within some categories. The extent to which this analysis was able to segment customers by socio-spatial characteristics can also be observed from the distinct geographical clusters of

individuals, presented in the proceeding sections. However, it should be noted that these distributions will be influenced by the differing sample sizes across customer Groups.

6.3.2.1. Supergroup 1 – 'Rural Ageing Off-peak Shoppers'.

Customers assigned to this Supergroup demonstrated highest patronage to store Supergroup 1 ('General Off-peak Shopping') during weekdays. Figure 6.20 demonstrates the age distribution and temporal profiles for Groups 1a, 1b and 1c. Figure 6.21 shows an overall comparison of product consumption between Groups and Figure 6.22 variations between weekdays and weekends. Product data were normalised by total purchase volumes per category, and therefore indicate products that were most prominent within this Supergroup in comparison to overall consumption. Figures 6.23 and 6.24 illustrate the volume of customers per Group across GB MSOA's.

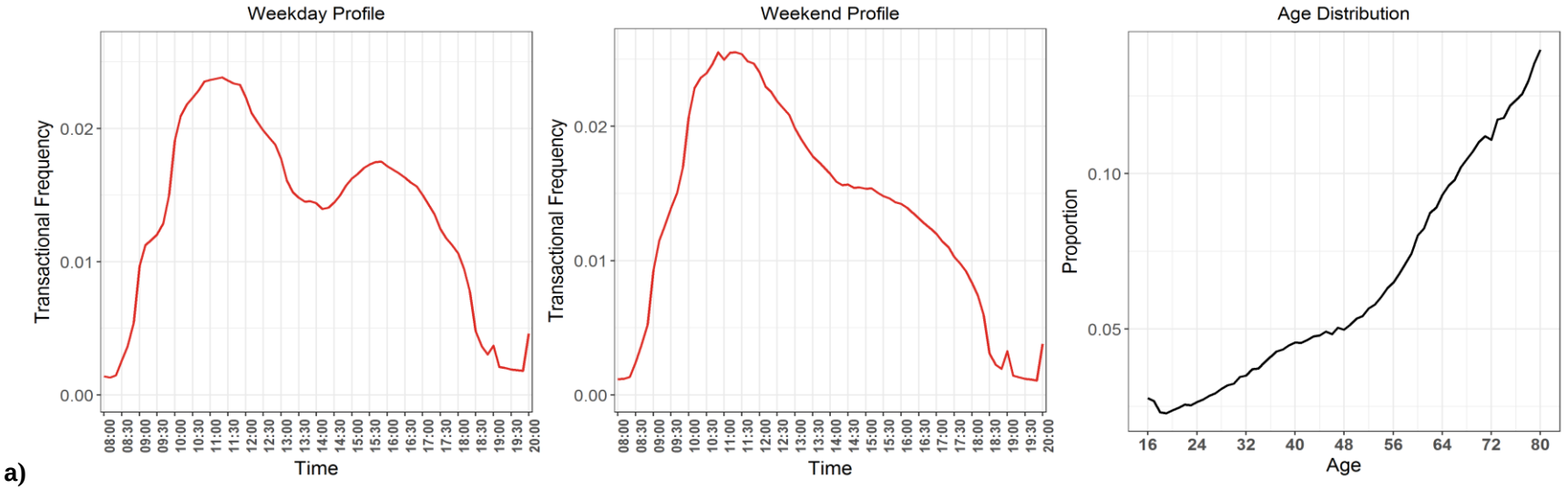
Distribution measures indicated that the majority of customers in this Supergroup interacted with 3 different store profiles overall on weekdays, and 2 on weekends, thus generally showing higher variation in behaviour during weekdays. Based on proportions of second-ranking profiles, customers were most likely to alternatively visit store Group 2a (predominantly medium-large town, health and beauty oriented stores) and to a lesser extent, Group 5b (predominantly rural fringe retail park stores) on both weekdays and weekends. However, these customers primarily patronised 'General Off-peak Shopping' stores, accounting for an average of 72.45% of their total transactions (only 10.5% of those assigned exhibited less than half).

This customer Supergroup had the highest average age of all Supergroups. As can be observed from Figure 6.20, all Groups exhibited a skew towards older cohorts, however those exhibiting highest patronage to urban destinations on weekends showed a slightly lower average age. The oldest demographic segment was Group 1a, who demonstrated minimal patronage to alternative store types. Product consumption for this Supergroup was higher for healthcare and pharmaceutical products, although this varied between Groups in line with what we might expect based on their demographic profiles. For example, Group 1a, who had the highest average age, showed the highest consumption within the 'Ageing Healthcare' category. Groups 1b and 1c both demonstrated high weekday consumption of 'NHS prescriptions', suggesting the primary usage for prescription collection. However, both segments demonstrated a mix of cosmetics and healthcare usage during weekend periods, with Group 1c showing higher consumption for more expensive, premium items (i.e. fragrance, premium cosmetics). Group 1c was also the only segment to show high consumption of 'Childrens' products.

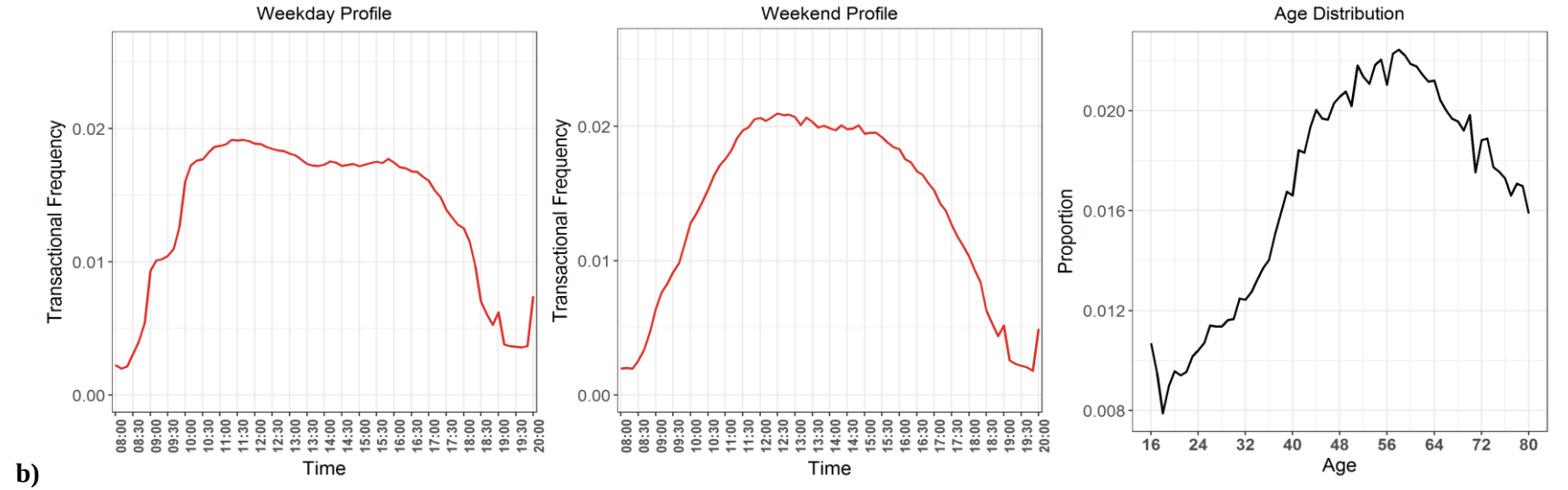
Distinctions were also evident between the temporal profiles of Groups. All customers in this Supergroup exhibited highest activity within off-peak periods, however, Group 1a showed the earliest peak during both weekdays and weekends, Group 1b demonstrated late morning to mid-

day peaks and Group 1c showed higher consumption mid-day to afternoon. This suggests that older HSR customers may have a preference for activity earlier in the day. Volumes of customers per Group across MSOAs suggested a relationship of proximity/accessibility to their preferred store types. For example, Group 1a were the most remote customers; Group 1b were rurally located yet within closer proximity to small towns and Group 1c closer to urban areas. This suggests strong relationships between spatiotemporal consumption patterns and geodemographic characteristics. This Supergroup dominated 15% of postcodes across GB. The majority of these were Group 1a customers (11.6% of postcodes) and the minority Group 1c (0.9%), however, this is largely reflective of differing sample sizes.

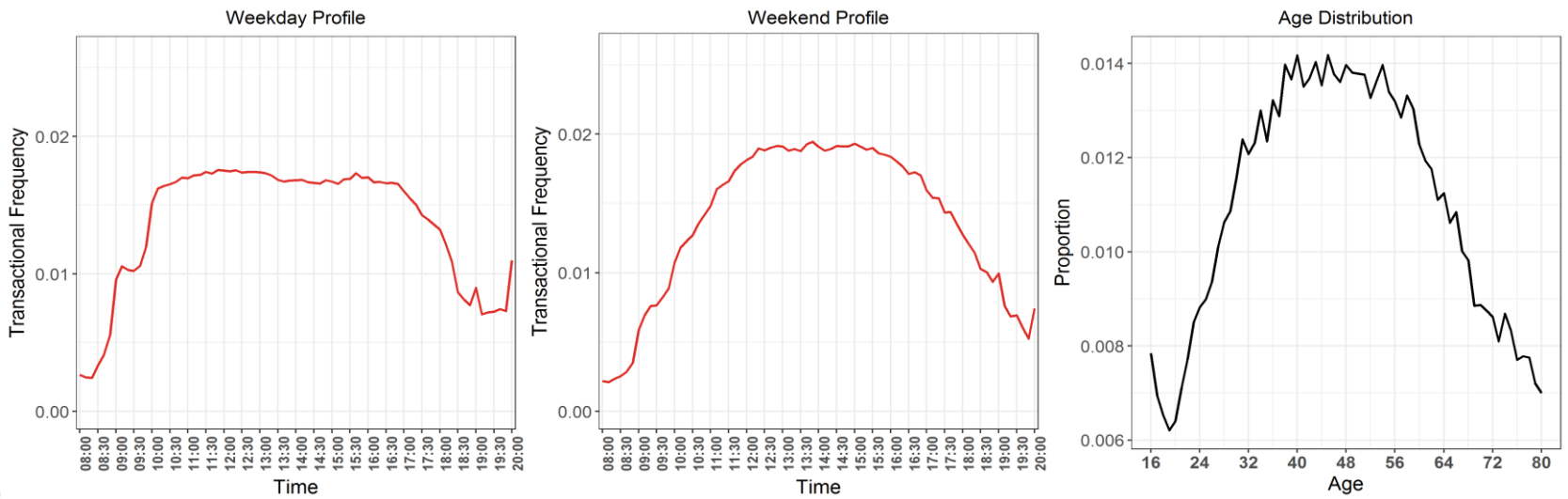
Overall, customer Supergroup 1 represented the HSR customers with the highest average ages and most rural residences. Group 1a represented the oldest cohort, which was also supported by their 'Ageing Healthcare' consumption habits. These were likely to be early riser off-peak shoppers and showed the least variation in store visiting behaviour. Group 1b represented a slightly younger demographic, who primarily resided on rural fringes surrounding small towns, and exhibited high prescription collection during weekdays and destination shopping in local small towns on weekends. These customers were likely to be most active within late morning to mid-day periods. Group 1c resided within closer proximity to urban centres and were the latest risers. These customers showed high prescription collection during weekdays and destination shopping in urban centres for cosmetics and healthcare essentials on weekends.


Weekday Profile
Weekend Profile
Age Distribution
Transactional Frequency
Proportion
Time
Age
0.00
0.01
0.02
0.05
0.10
08:00
08:30
09:00
09:30
10:00
10:30
11:00
11:30
12:00
12:30
13:00
13:30
14:00
14:30
15:00
15:30
16:00
16:30
17:00
17:30
18:00
18:30
19:00
19:30
20:00
16
24
32
40
48
56
64
72
80

a)


Weekday Profile
Weekend Profile
Age Distribution
Transactional Frequency
Proportion
Time
Age
0.00
0.01
0.02
0.008
0.012
0.016
0.020
08:00
08:30
09:00
09:30
10:00
10:30
11:00
11:30
12:00
12:30
13:00
13:30
14:00
14:30
15:00
15:30
16:00
16:30
17:00
17:30
18:00
18:30
19:00
19:30
20:00
16
24
32
40
48
56
64
72
80

b)



c)

Figure 6.20: Temporal profiles (weekday, weekend, 10-minute intervals) and age distributions for a) Group 1a, b) Group 1b and c) Group 1c.

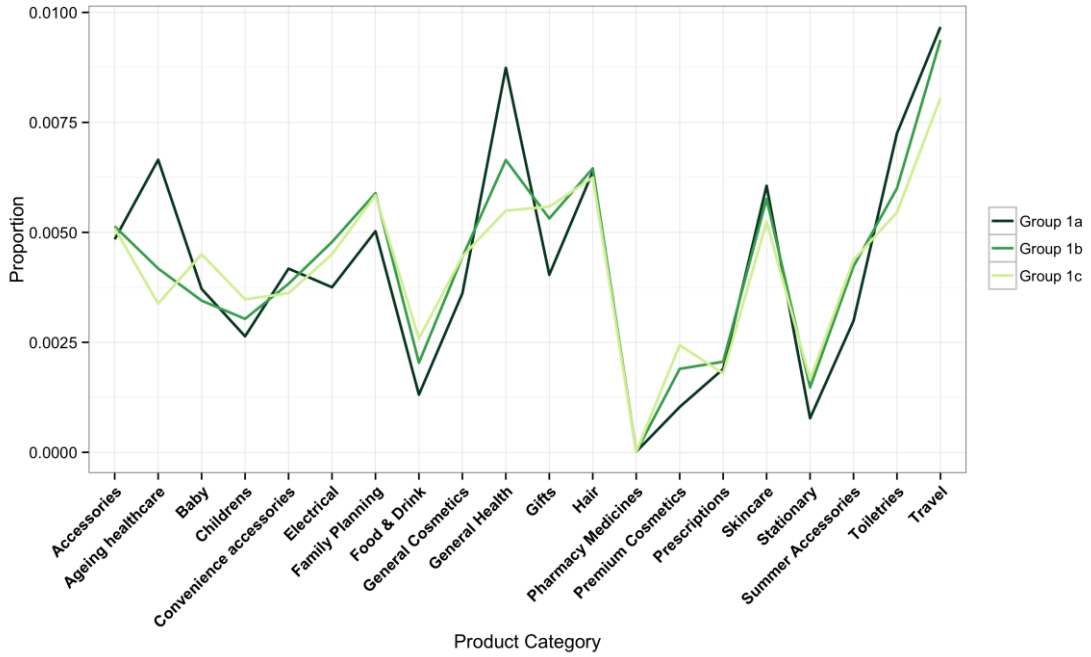
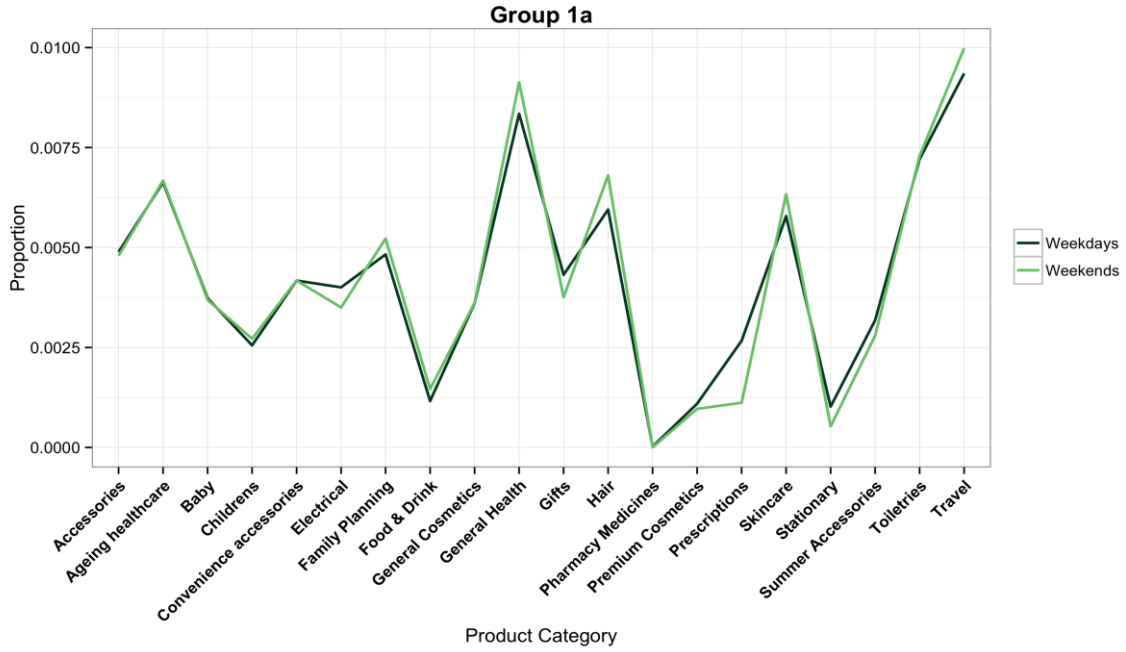
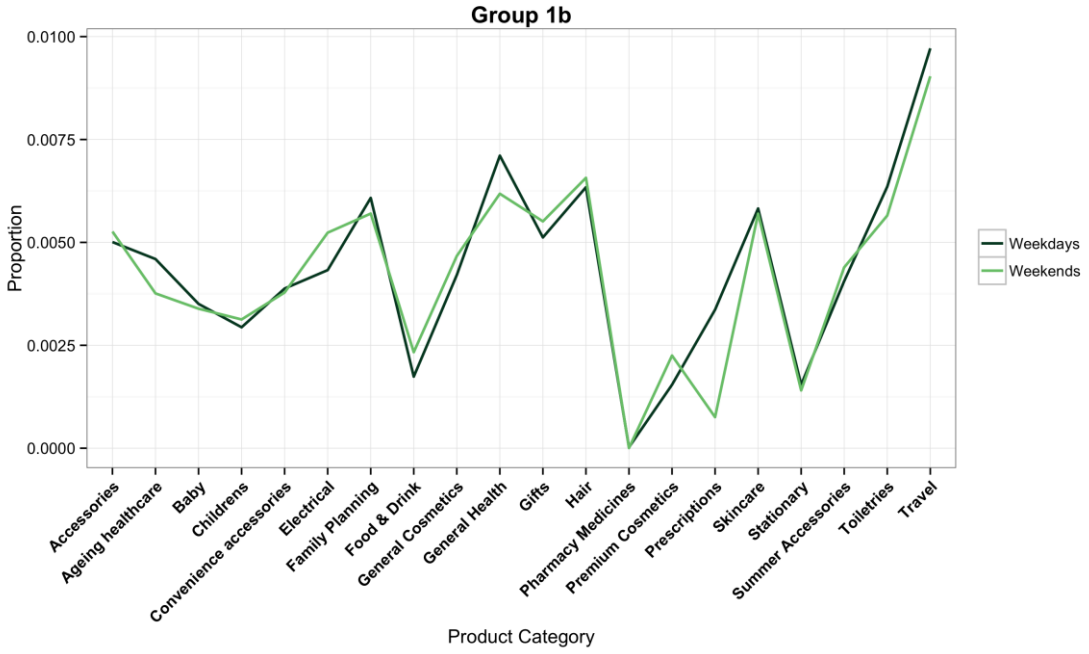


Figure 6.21: Comparison of product consumption (proportions) across Groups in Supergroup 1.

a)



b)



c)

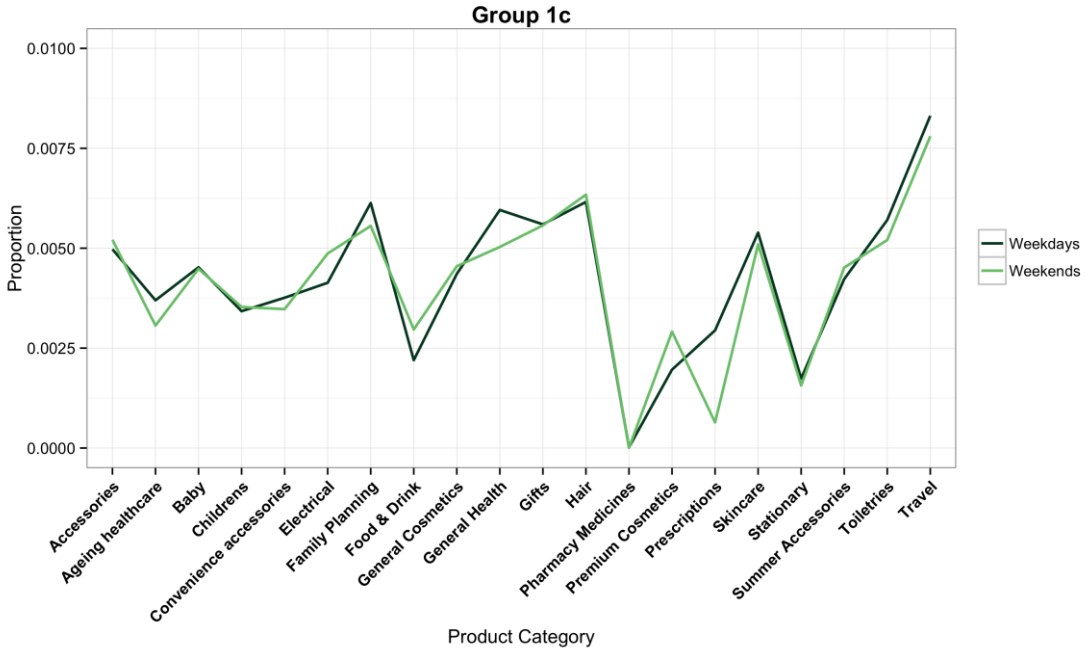


Figure 6.22: Comparison of product consumption (proportions) during weekdays and weekends for a) Group 1a, b) Group 1b and c) Group 1c.

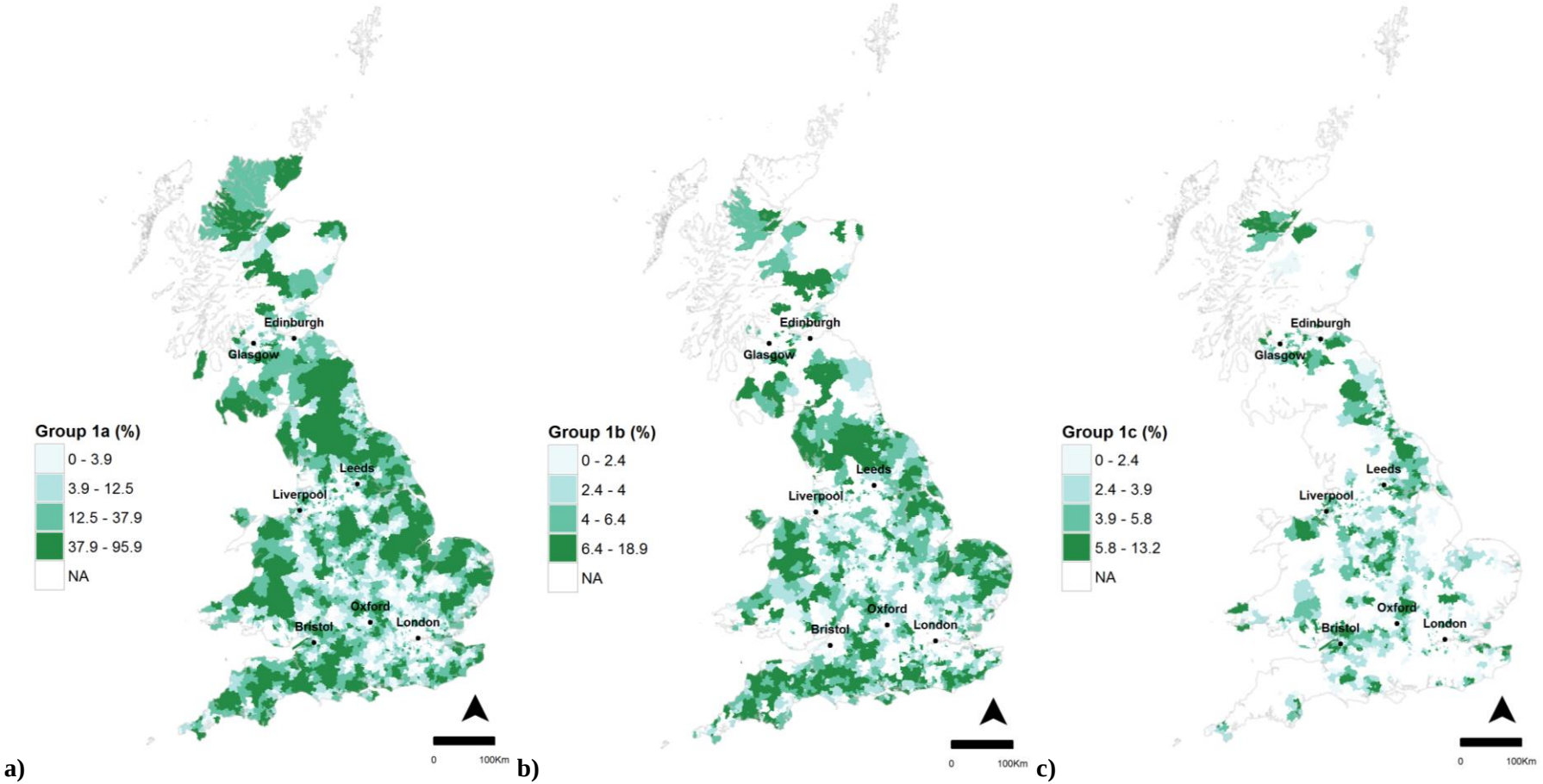
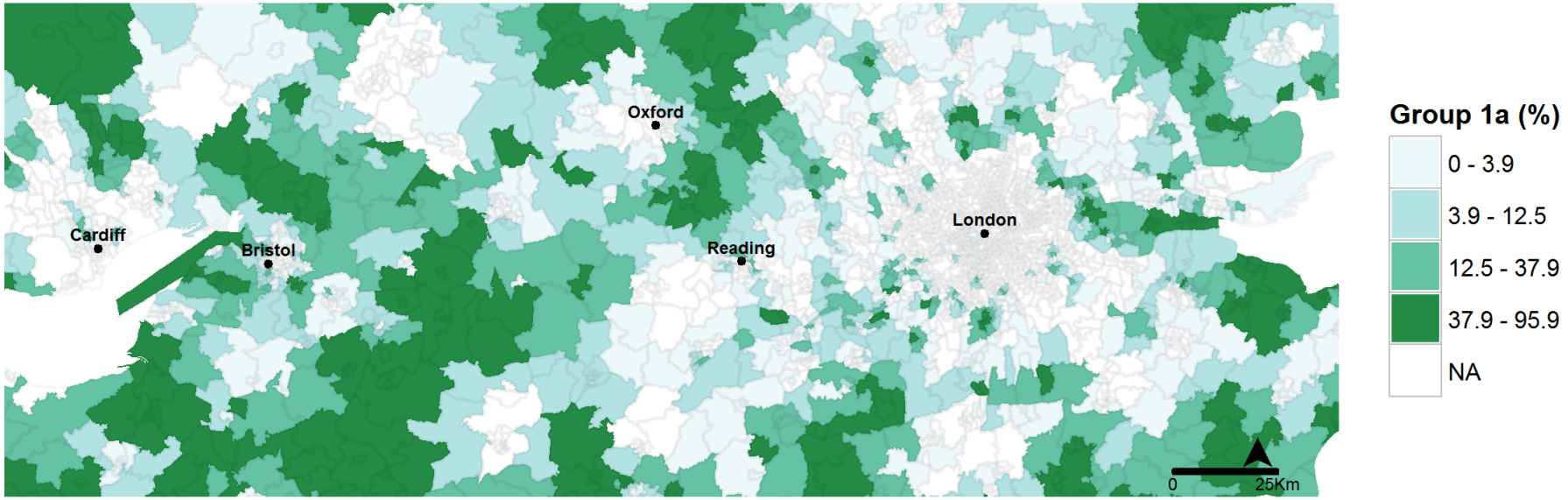


Figure 6.23: The percentage of customers per MSOA in a) Group 1a - 'Stable Rural Ageing Health', b) Group 1b - 'Rural, Weekend Small-town Shoppers' and c) Group 1c - 'Rural Fringe, Urban Destination Shoppers', across Great Britain (quantile breaks). ***'NA' = no customers in group present.***

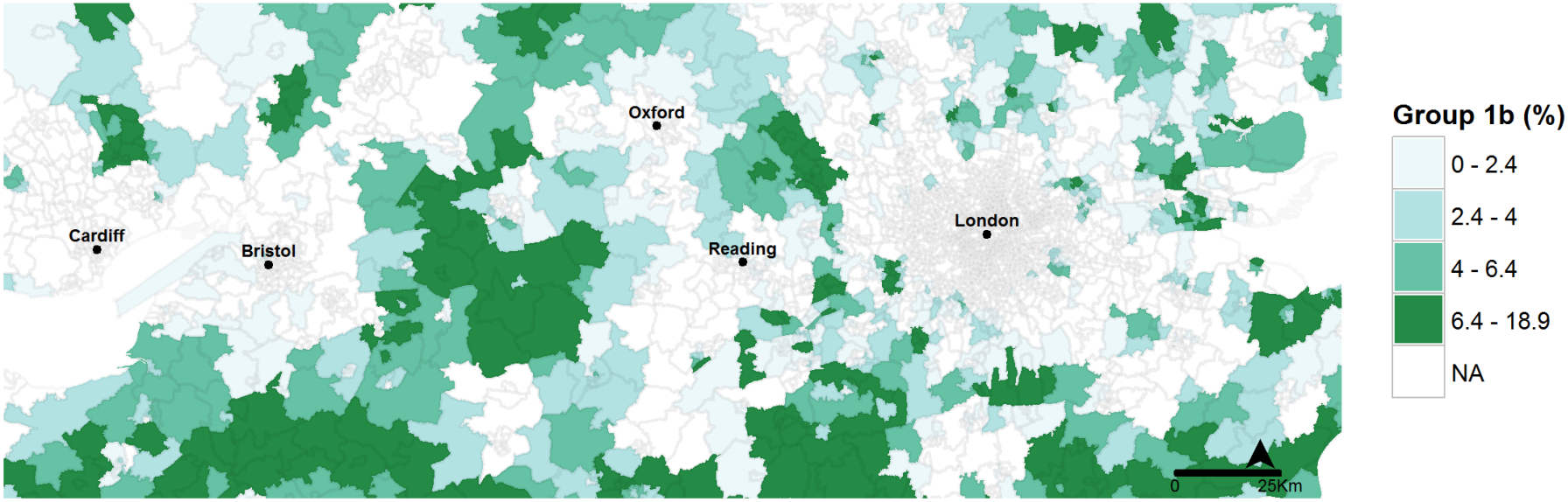
Group 1a – 'Stable Rural Ageing Health'



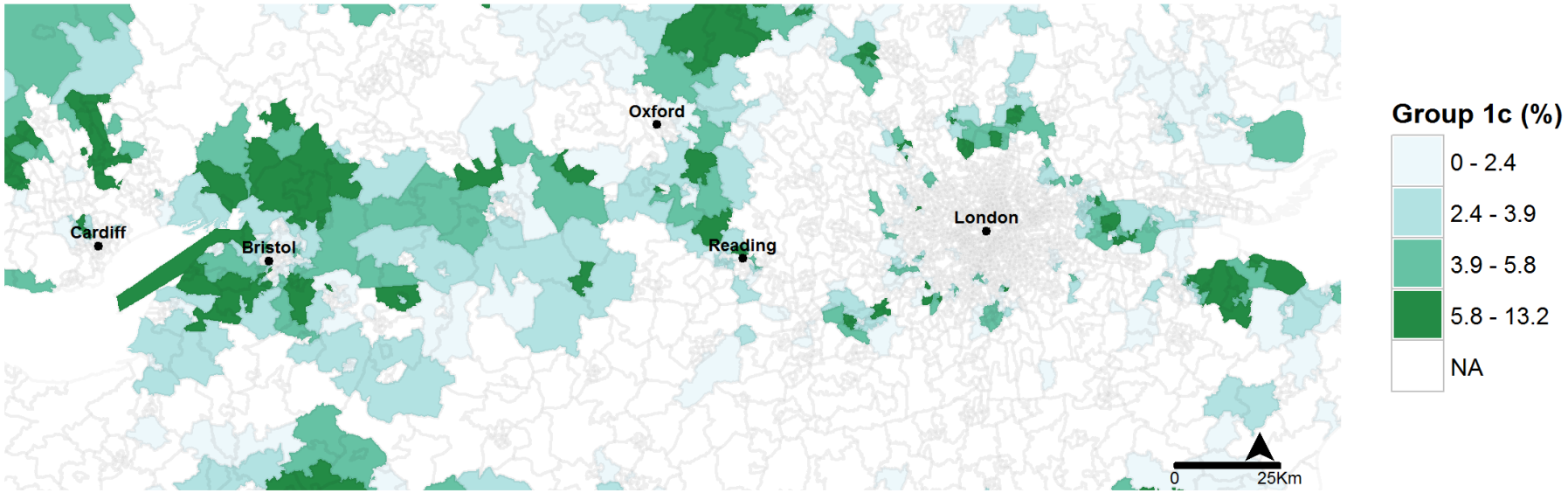
a)

Group 1b – 'Rural, Weekend Small-town Shoppers'

b)



Group 1c – 'Rural Fringe, Urban Destination Shoppers'



c)

Figure 6.24: The percentage of customers per MSOA in a) Group 1a, b) Group 1b and c) Group 1c, across Southern England (quantile breaks). ***'NA' = no customers in group present.***

6.3.2.2. Supergroup 2 – 'Small Destination Shoppers'.

Customers assigned to Supergroup 2 demonstrated highest patronage to store Supergroup 2 ('Weekend Peak Destinations') during weekdays. Figure 6.25 shows the age distributions and temporal profiles for Groups 2a, 2b and 2c, Figure 6.26 a comparison of product consumption between groups, Figure 6.27 product consumption during weekdays and weekends, and Figures 6.28 to 6.29 illustrate the volume of customers per Group across GB MSOA's.

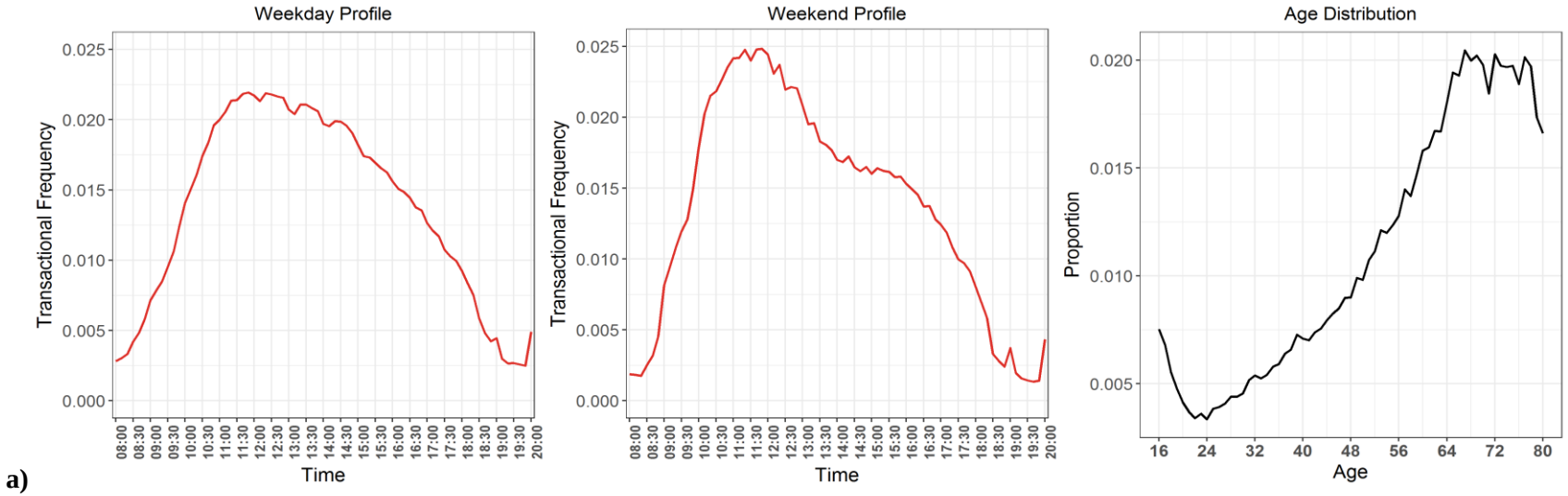
Distribution measures indicated that the majority of customers in this Supergroup interacted with 2 different store profiles overall on weekdays, and 2 on weekends, thus, generally showing lower mobility during weekdays than all other Supergroups. Based on proportions of second ranking profiles, customers were most likely to alternatively visit store Groups 5a and 5b (rural or urban fringe retail park stores) or to a lesser extent, Group 1a (predominantly small high street stores) during both weekday and weekend periods. These trends may be a reflection of easy accessibility to both urban and rural locations. However, these customers predominantly patronised 'Weekend Peak Destinations', which accounted for an average of 78.25% of their total transactions (only 10.7% of those assigned exhibited less than half).

Customers in this Supergroup showed the second oldest average age overall. Demographic attributes exhibited the same correlation with weekend profiles as were evident in Supergroup 1. For example, Group 2a, who primarily patronised rural (store Supergroup 1) stores on weekends, demonstrated the highest average age, and those visiting urban destinations (store Supergroup 5) the lowest. Product consumption for this Supergroup was highest for cosmetics and healthcare essentials (i.e. pain relief and hay fever), yet demonstrated higher beauty/cosmetic usage than the healthcare/pharmaceuticals oriented customers of Supergroup 1. Product consumption varied between Groups, with Group 2a demonstrating high healthcare product consumption, Group 2b a mix of cosmetics, beauty and healthcare and Group 2c predominantly cosmetic and beauty consumption. This indicated a relationship with customer demographics and product consumption, such as ageing populations being primarily healthcare focused and younger segments cosmetics and beauty focused.

Supergroup 2 dominated the largest number of postcodes (47.9%). The majority of these customers exhibited the behaviours of Group 2b (45.1%) and the minority 2a (0.9%). These dynamics were largely a result of differing sample sizes. The distribution of these areas at the Group level could be differentiated in a similar fashion to that of Supergroup 1, of which observations suggested was a reflection of accessibility to their preferred store types. For example, Group 2a resided in the most rural locations, consistent with their primary patronage of rural stores during weekends. Group 2b were clustered around small-town areas, and Group 2c around urban fringes. Observation of temporal profiles at the Group level also indicated

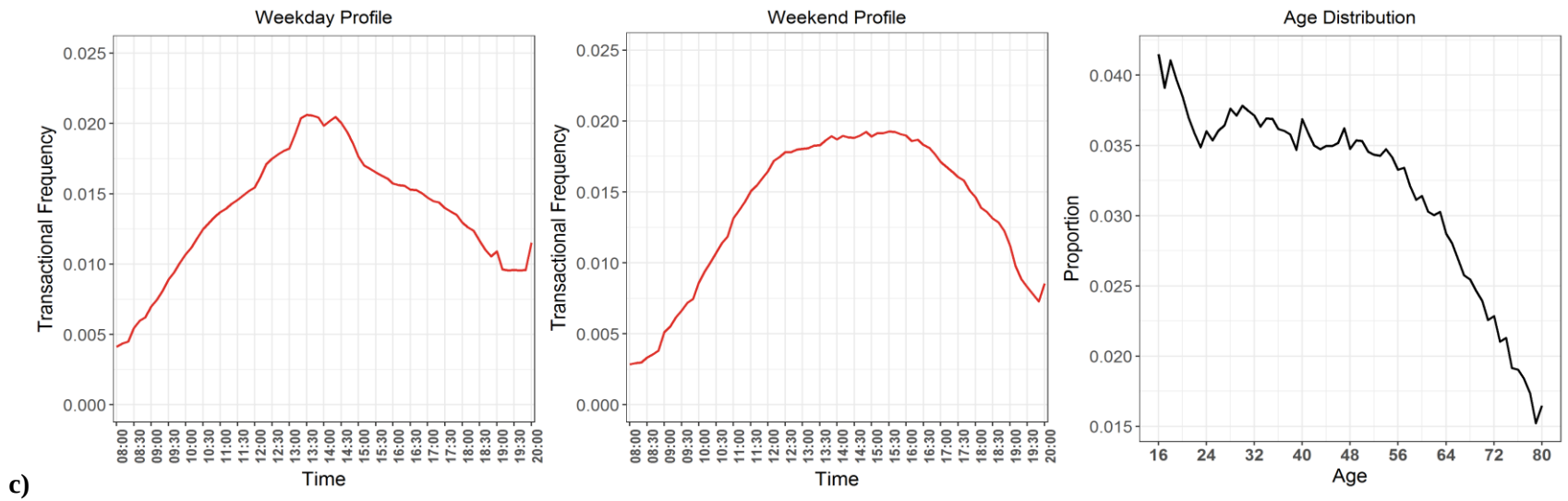
parallel trends to that of Supergroup 1 in terms of relationships with age and peak consumption times. Group 2a exhibited the earliest peaks of this Supergroup, with weekday peaks late morning to mid-day and early risers on the weekend. Group 2b also exhibited early peaks but increased activity during afternoons. Group 2c showed the latest peak consumption of mid-day to afternoon on both weekdays and weekends.

Overall, these customers represented rural fringe, middle-aged to ageing HSR customers who were geographically clustered around small town areas. They demonstrated off-peak consumption trends during weekdays (similarly to Supergroup 1), yet a higher consumption of cosmetics and beauty products. However, distinctions could be made between the consumption patterns of Groups, such as the oldest Group (2a) demonstrating higher consumption of 'Ageing Healthcare' products and the younger customers (2c) of cosmetics, reaffirming the relationship of demographic and product consumption observed thus far.


Weekday Profile
Weekend Profile
Age Distribution
Transactional Frequency
Proportion
Time
Age
0.025
0.020
0.015
0.010
0.005
0.000
0.28
0.24
0.20
08:00
08:30
09:00
09:30
10:00
10:30
11:00
11:30
12:00
12:30
13:00
13:30
14:00
14:30
15:00
15:30
16:00
16:30
17:00
17:30
18:00
18:30
19:00
19:30
20:00
16
24
32
40
48
56
64
72
80

a)

b)



c)

Figure 6.25: Temporal profiles (weekday, weekend, 10-minute intervals) and age distributions for a) Group 2a, b) Group 2b and c) Group 2c.

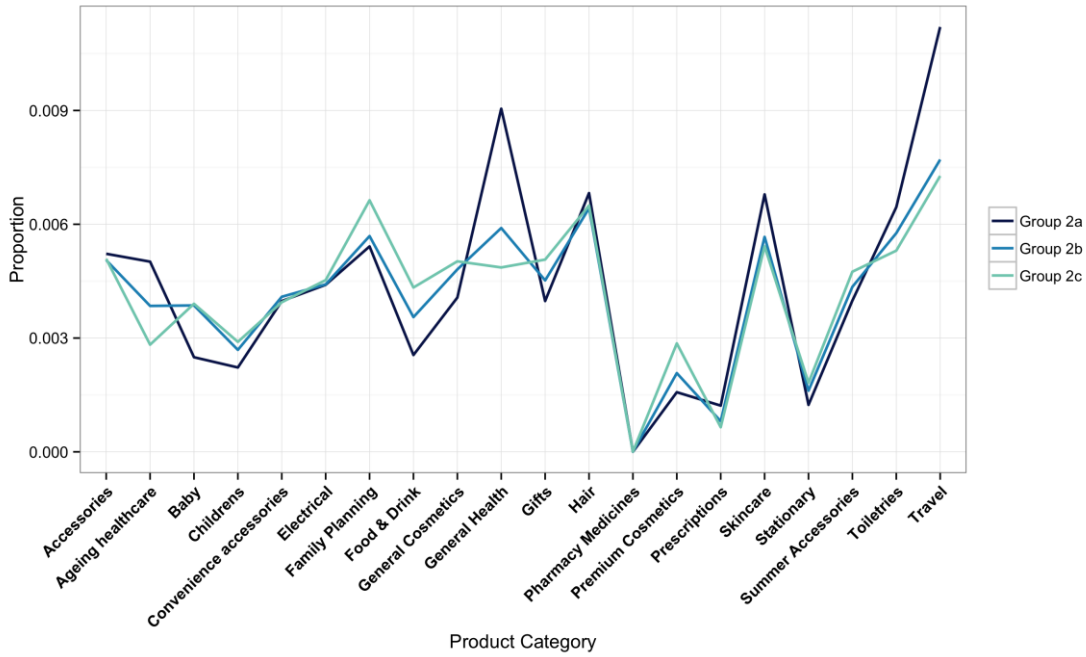
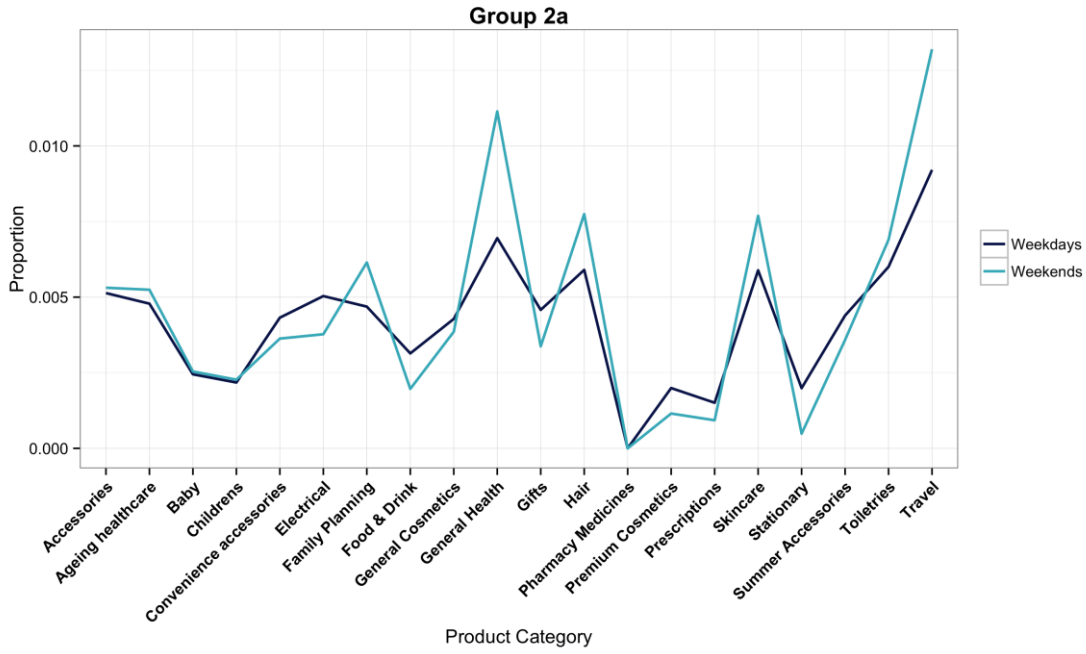
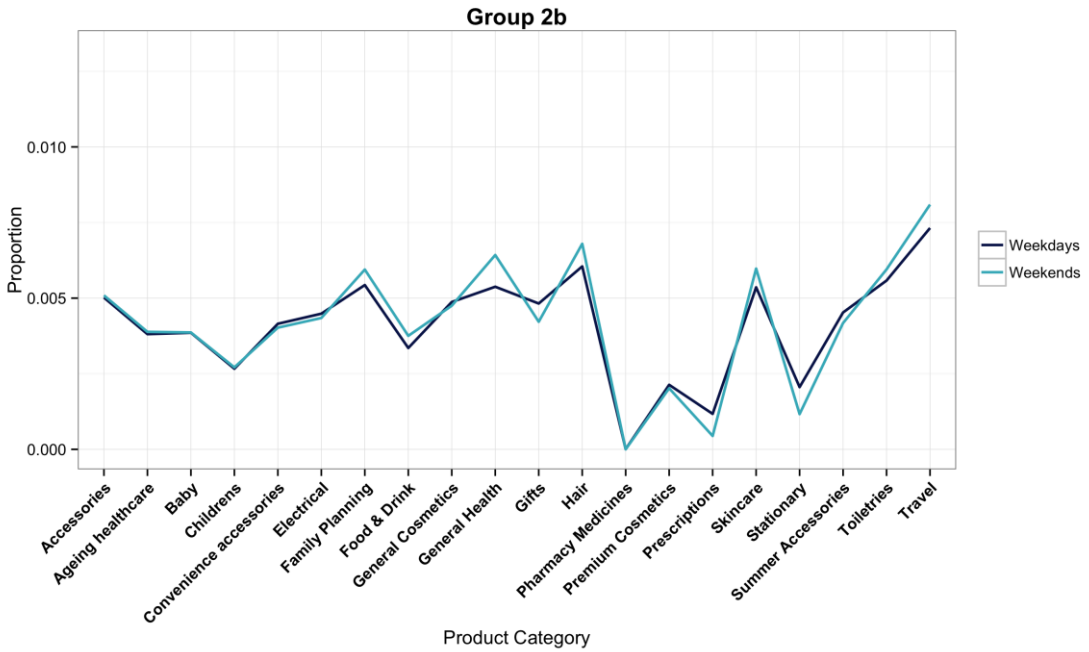


Figure 6.26: Comparison of product consumption (proportions) across Groups in Supergroup 2.

a)



b)



c)

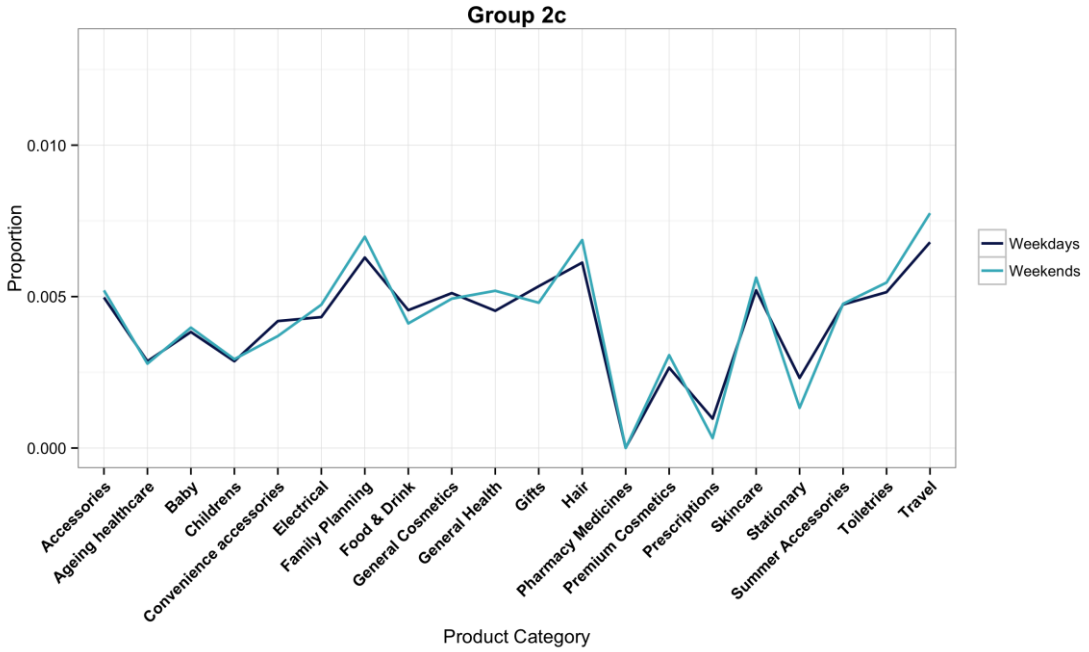


Figure 6.27: Comparison of product consumption (proportions) during weekdays and weekends for a) Group 2a, b) Group 2b and c) Group 2c.

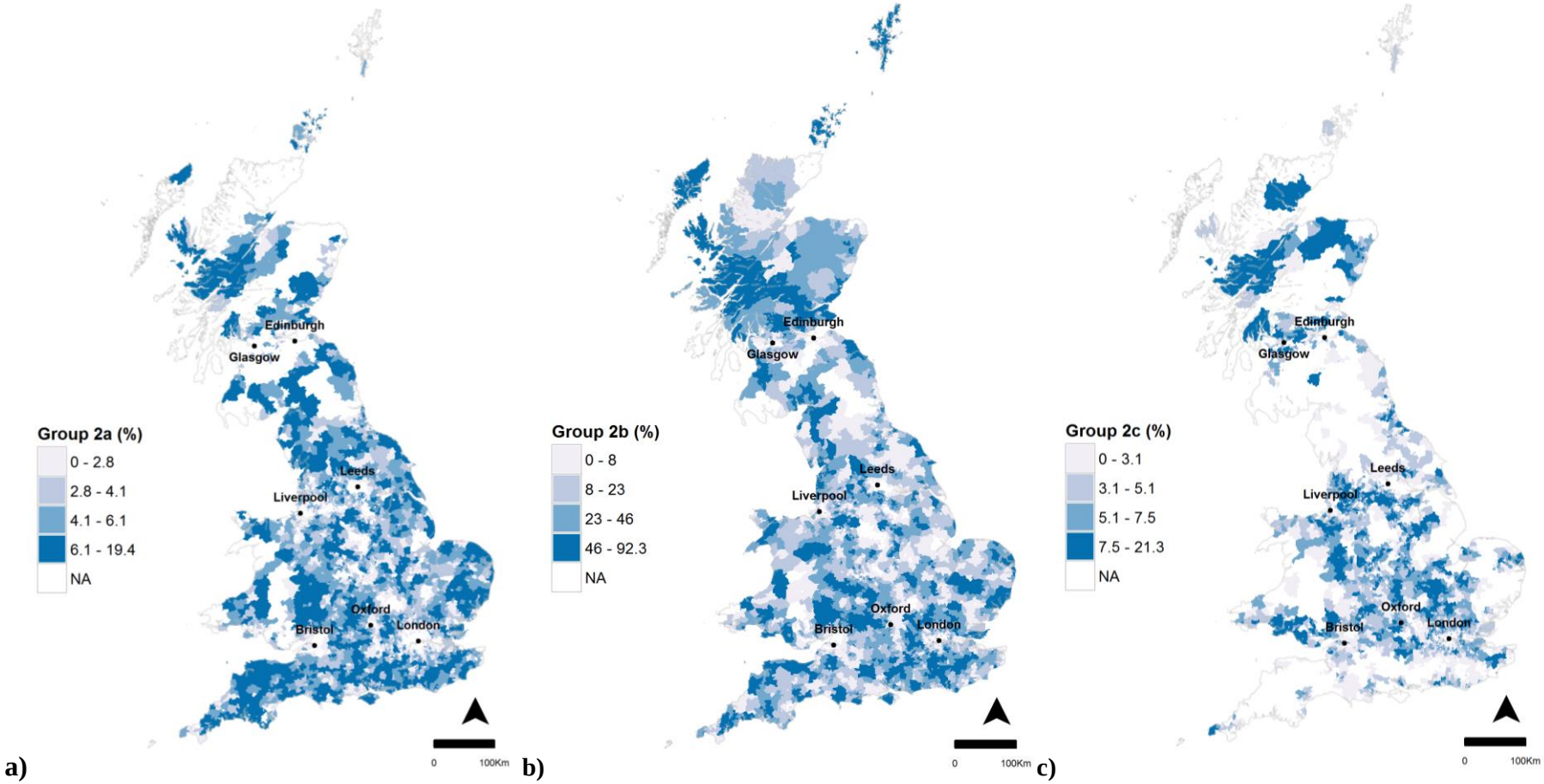
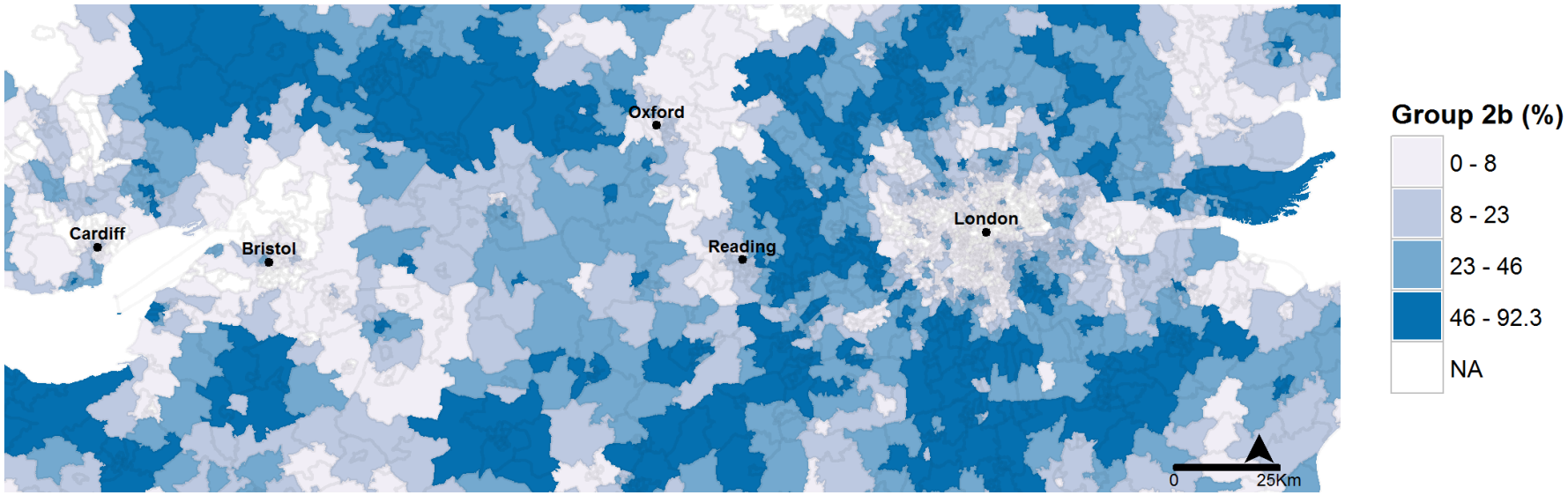


Figure 6.28: The percentage of customers per MSOA in a) Group 2a - 'Rural, Weekday Small-town Shoppers', b) Group 2b - 'Stable Small-town Shoppers' and c) Group 2c - 'Small-town, Weekend Urban Destination Shoppers', across Great Britain (quantile breaks). ***'NA' = no customers in group present.***

Group 2a – 'Rural, Weekday Small-town Shoppers'

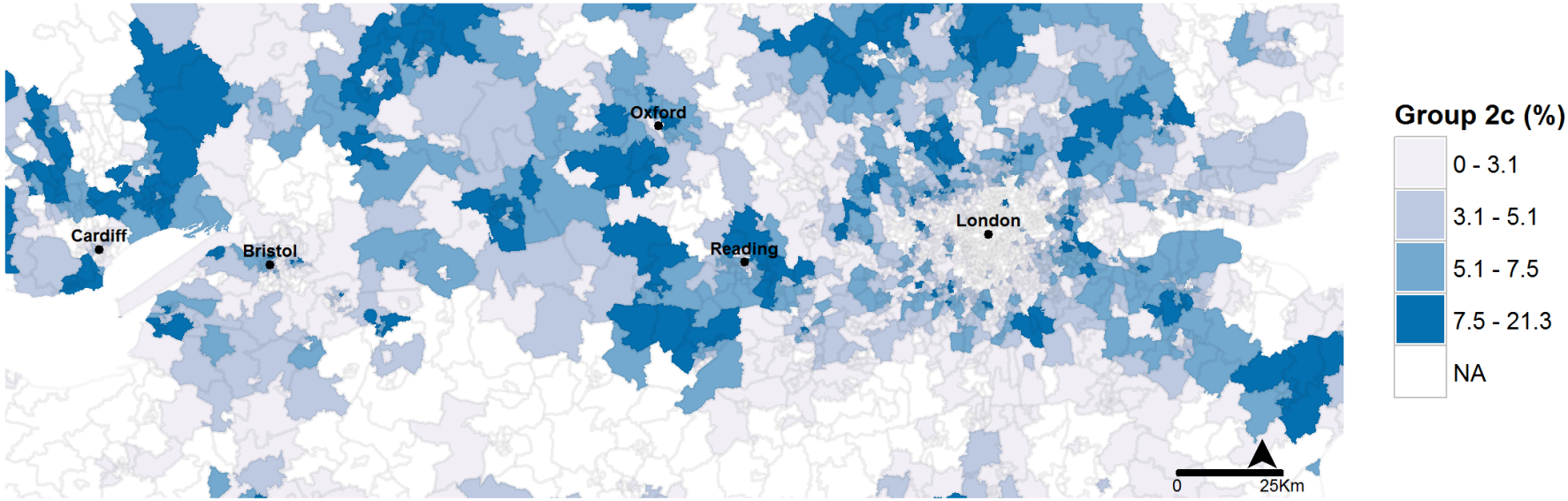
a)

Group 2b – ‘Stable Small-town Shoppers’



b)

Group 2c – 'Small-town, Weekend Urban Destination Shoppers'



c)

Figure 6.29: The percentage of customers per MSOA in a) Group 2a, b) Group 2b and c) Group 2c, across Southern England (quantile breaks). ***'NA' = no customers in group present***

6.3.2.3. Supergroup 3 – 'Weekday Convenience Commuters'.

Customers assigned to Supergroup 3 demonstrated highest patronage to store Supergroup 4 ('Weekday Convenience') during weekdays. Figure 6.30 shows the age distributions and temporal profiles for Groups 3a, 3b, 3c and 3d, Figure 6.31 a comparison of product consumption between Groups, Figure 6.32 product consumption during weekdays and weekends, and Figures 6.33 and 6.34 illustrate the volume of customers per Group across GB MSOA's.

Distribution measures indicated that the majority of customers in this Supergroup interacted with 3 different store profiles on weekdays, and 2 on weekends, thus showing higher behavioural variation during weekday periods. These customers were most likely to alternatively visit stores in Supergroup 5 ('Stable Destinations') and to a lesser extent Supergroup 2 ('Weekend Peak Destinations'), during both weekday and weekend periods. However, these customers predominantly patronised 'Weekday Convenience' stores, which accounted for an average of 70.54% of their total transactions. This was the lowest percentage observed across all Supergroups (20.2% of those assigned exhibited less than half), suggesting these customers show higher overall variation in behaviour than other segments.

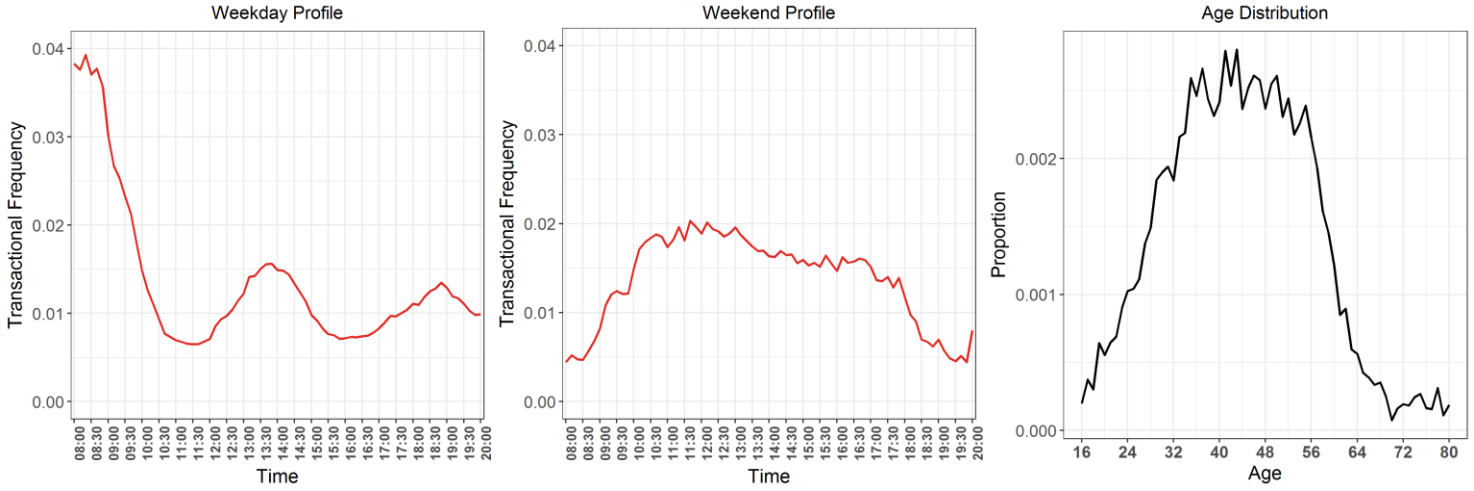
This Supergroup exhibited the lowest average age of all Supergroups, and temporal profiles delineated a working population who shop between business hours during weekdays. Product consumption during weekdays was relatively uniform across all Groups, showing high purchasing of food, drink and convenience items (such as umbrellas and hosiery). However, the inclusion of weekend profiles differentiated this Supergroup into Groups with unique activity and consumption patterns. Group 3a represented those with highest patronage of stores in Supergroup 1 on weekends and were the oldest and most rurally located of the commuter groups. Weekend consumption for this group was prominent for healthcare essentials and children's products. Group 3b represented those travelling from small commuter towns and were the second oldest commuter group. Weekend consumption was highest for 'Family Planning' and healthcare essentials.

Group 3c represented the smallest and second youngest group, living in suburban/urban areas and patronising the same store type on weekends. Product consumption showed similar usage to weekdays, such as food, drink and convenience essential items. This Group contained the highest proportion of male customers of all Groups. Finally, Group 3d represented the youngest Group, living predominantly in urban areas, and patronising large urban destinations on weekends. This was also the largest commuter Group. Analysis of behaviour of the store Group level indicated that the majority of these customers patronised the 'Urban stable destination'

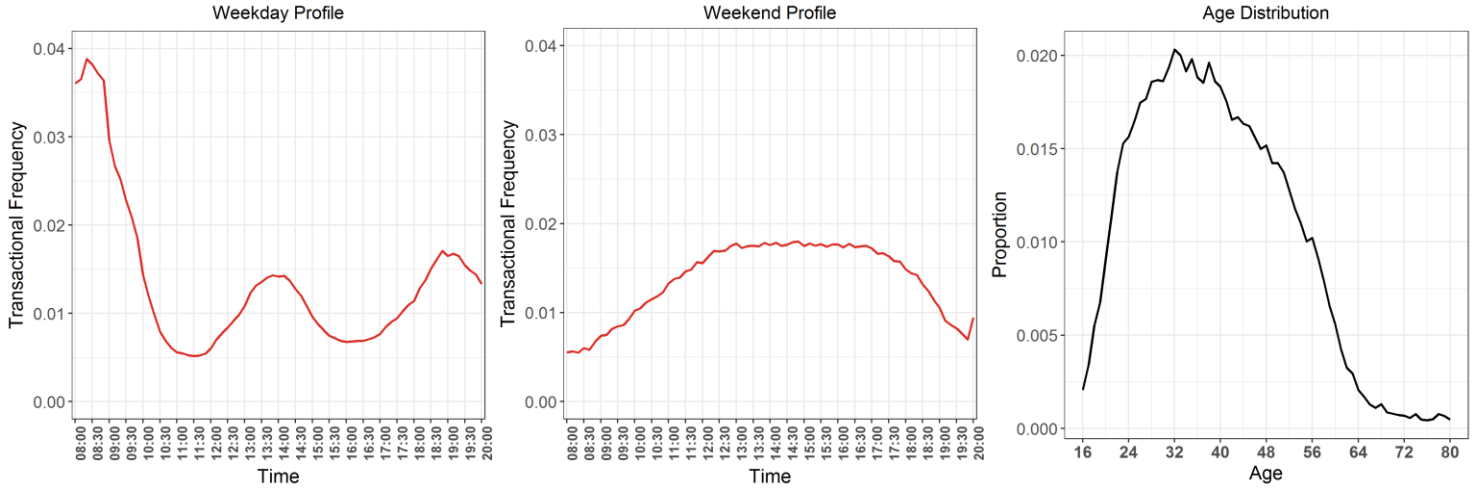
stores (i.e. city centre flagships) and the urban fringe retail parks. Product consumption was highest for healthcare essentials, beauty, food and 'Family Planning' on weekends.

This Supergroup dominated the smallest number of postcodes overall (2.9%), yet, they also exhibited the most distinct temporal patterns. All Groups demonstrated convenience patterns during weekdays (i.e. morning, lunchtime and evening peaks), however, Group 3a showed the highest and earliest morning peaks. This could be a result of longer commuting distances, and thus earlier activity periods. Activity in this Group also peaked during weekday lunchtimes and, to a lesser extent, evenings and during mid-day on weekends. Group 3b also demonstrated high early morning consumption (these customers may also be subject to longer commutes), but higher evening peaks than Group 3a. Weekend consumption was highest during afternoons. Conversely, Group 3c were most active during evenings on both weekdays and weekends. Group 3d demonstrated the most distributed behaviour throughout weekdays, with peaks during mornings, lunchtimes and evenings (the highest lunchtime peak of this Supergroup) and on evenings during weekends.

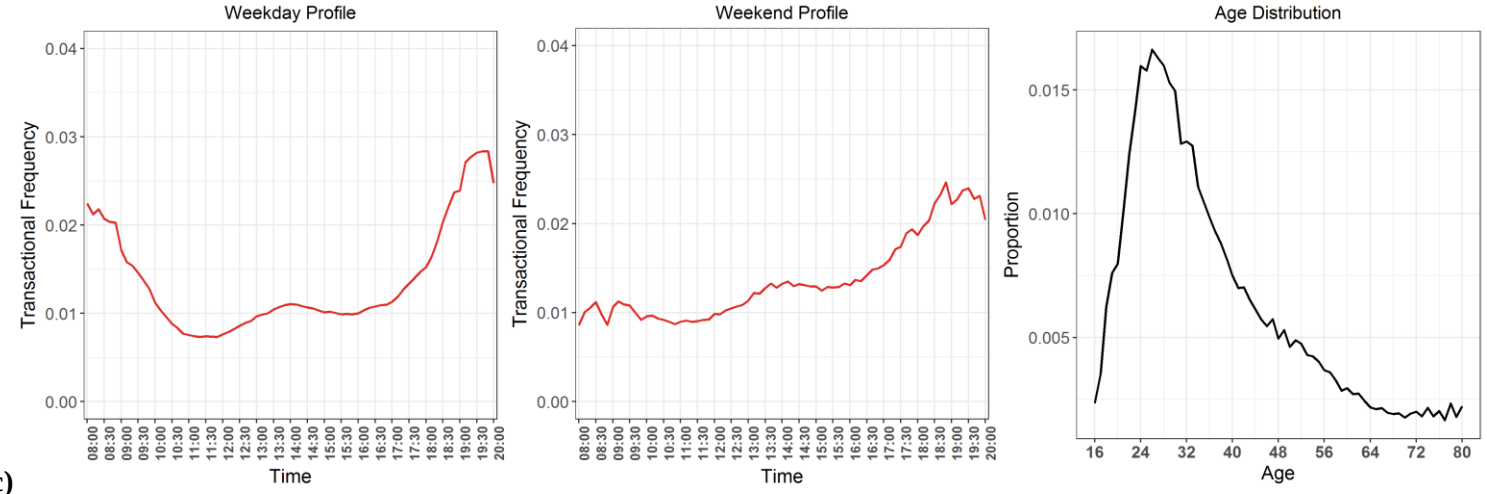
Overall, these customers represented a mix of rural fringe/urban customers who demonstrated the same distinct convenience shopping patterns during weekdays. Yet, analysis of their weekend consumption revealed differentiations in terms of residential location types, demographic attributes and product consumption patterns. During weekdays, all Groups demonstrated similar consumption patterns (food, drink and essentials). During weekends, older commuters demonstrated higher consumption for healthcare essentials and younger customers showed higher consumption for beauty and food.


Weekday Profile
Weekend Profile
Age Distribution
Transactional Frequency
Proportion
Time
Age

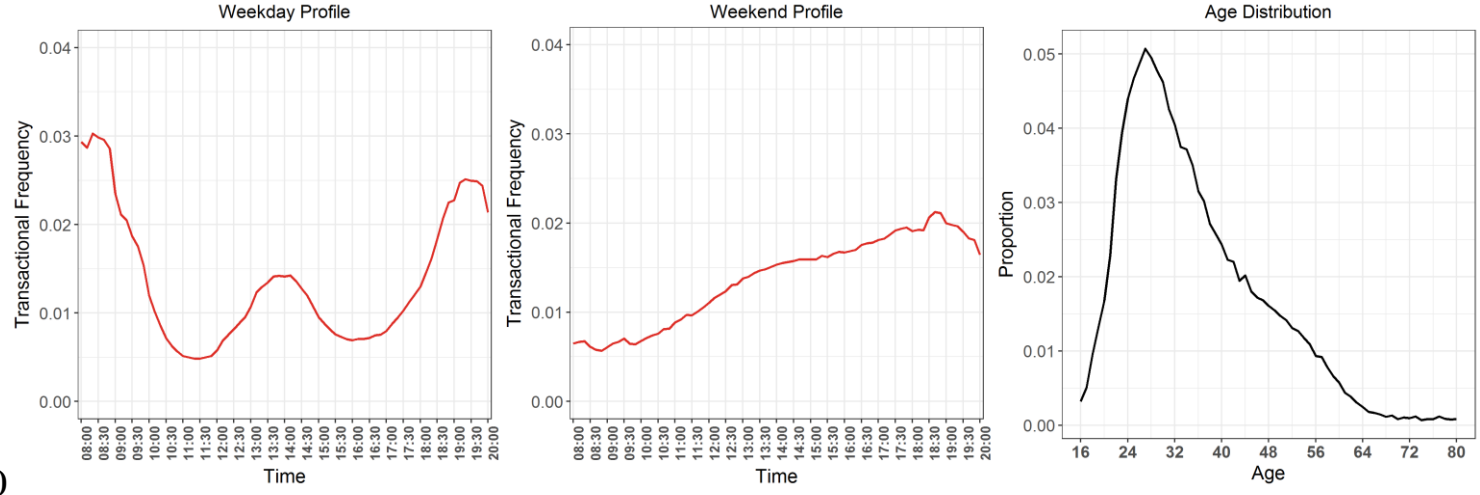
a)


Weekday Profile
Weekend Profile
Age Distribution
Transactional Frequency
Proportion
Time
Age

b)



c)



d)

Figure 6.30: Temporal profiles (weekday, weekend, 10-minute intervals) and age distributions for a) Group 3a, b) Group 3b, c) Group 3c and d) Group 3d.

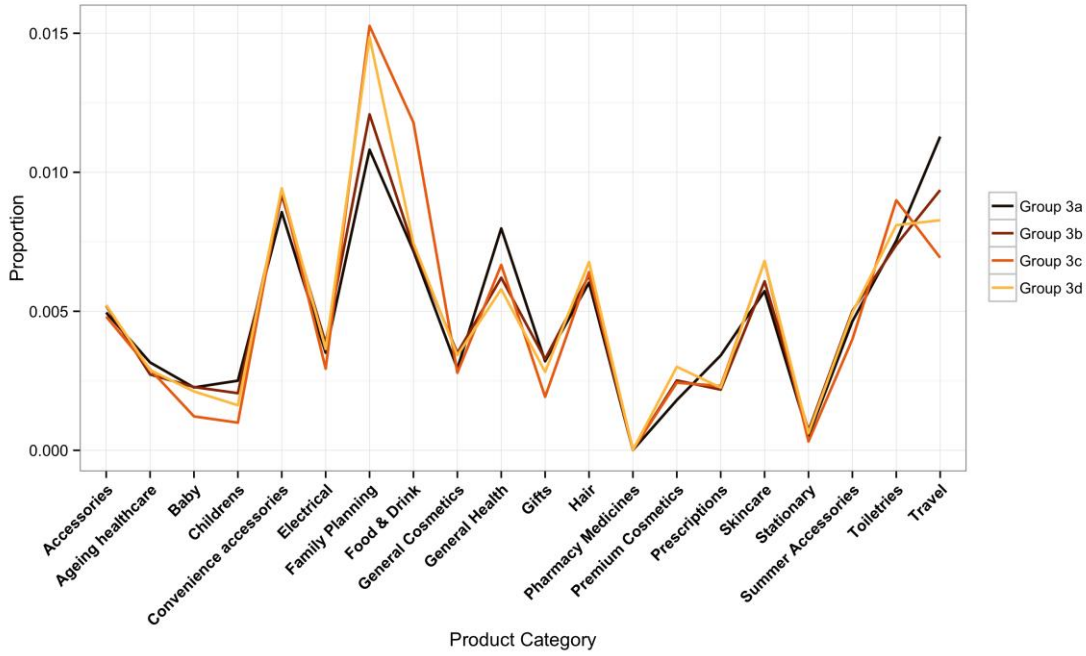
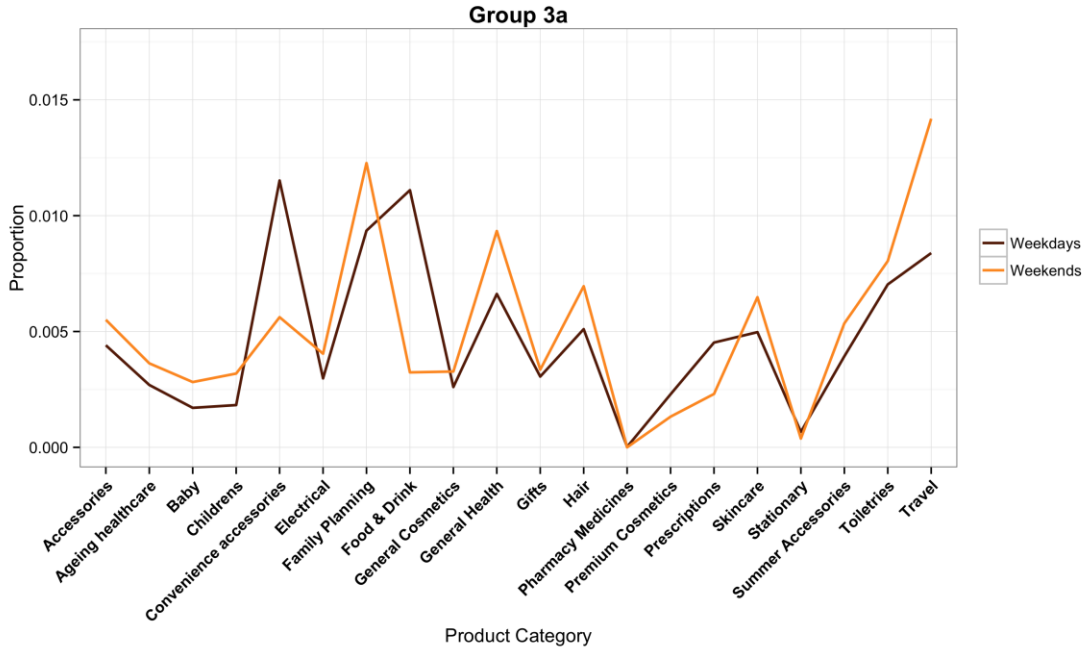
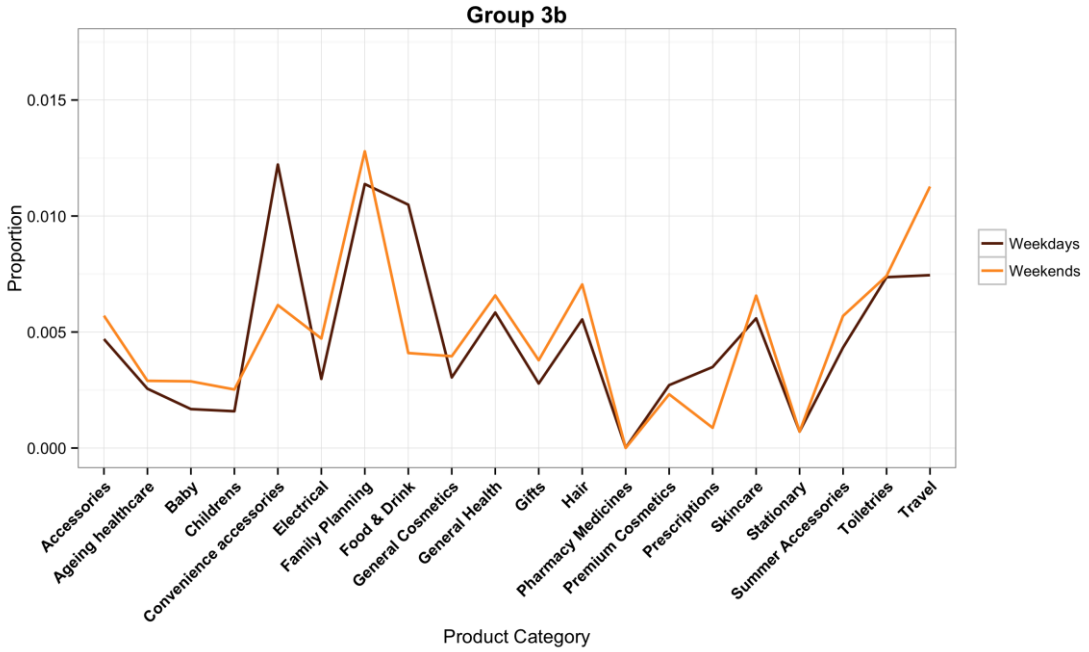


Figure 6.31: Comparison of product consumption (proportions) across Groups in Supergroup 3.

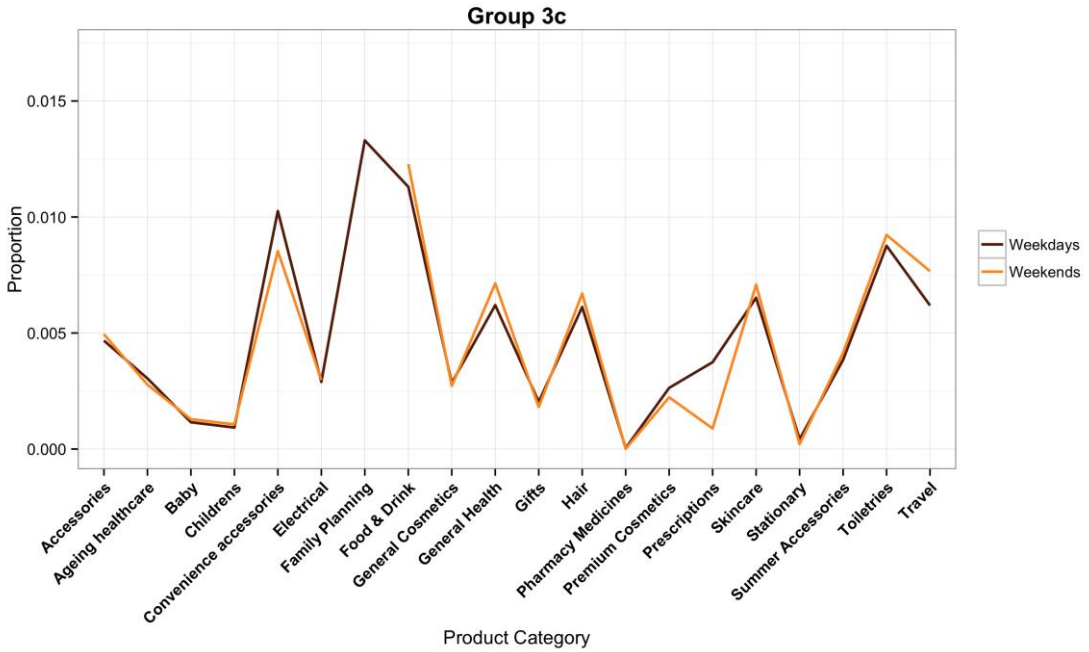
a)



b)


Group 3b
Proportion
0.015
0.010
0.005
0.000
Accessories
Ageing healthcare
Baby
Childrens
Convenience accessories
Electrical
Family Planning
Food & Drink
General Cosmetics
General Health
Gifts
Hair
Pharmacy Medicines
Premium Cosmetics
Prescriptions
Skincare
Stationary
Summer Accessories
Toiletries
Travel
Product Category
Weekdays
Weekends

c)


Group 3c
Proportion
0.015
0.010
0.005
0.000
Accessories
Ageing healthcare
Baby
Childrens
Convenience accessories
Electrical
Family Planning
Food & Drink
General Cosmetics
General Health
Gifts
Hair
Pharmacy Medicines
Premium Cosmetics
Prescriptions
Skincare
Stationary
Summer Accessories
Toiletries
Travel
Product Category
Weekdays
Weekends

d)

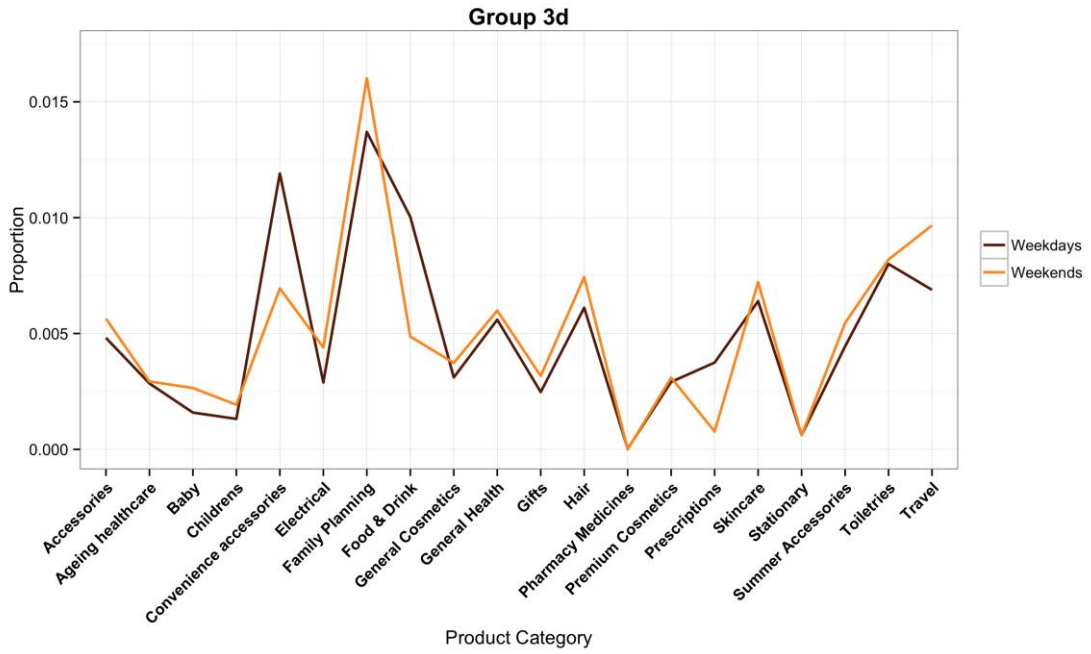
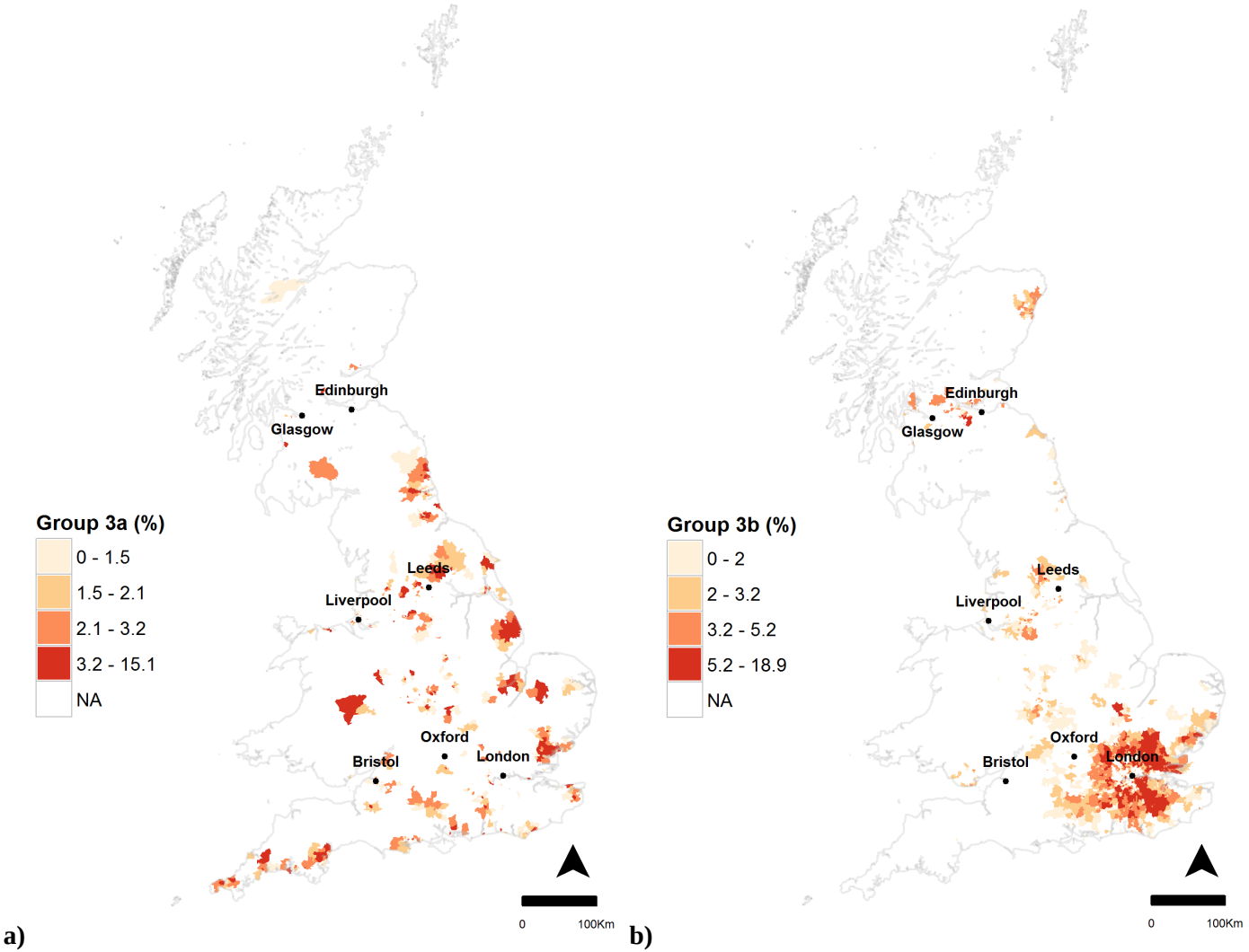


Figure 6.32: Comparison of product consumption (proportions) during weekdays and weekends for a) Group 3a, b) Group 3b, c) Group 3c and d) Group 3d.


Group 3a (%)
0 - 1.5
1.5 - 2.1
2.1 - 3.2
3.2 - 15.1
NA
Edinburgh
Glasgow
Leeds
Liverpool
Oxford
Bristol
London
0
100Km
a)
Group 3b (%)
0 - 2
2 - 3.2
3.2 - 5.2
5.2 - 18.9
NA
Edinburgh
Glasgow
Leeds
Liverpool
Oxford
Bristol
London
0
100Km
b)

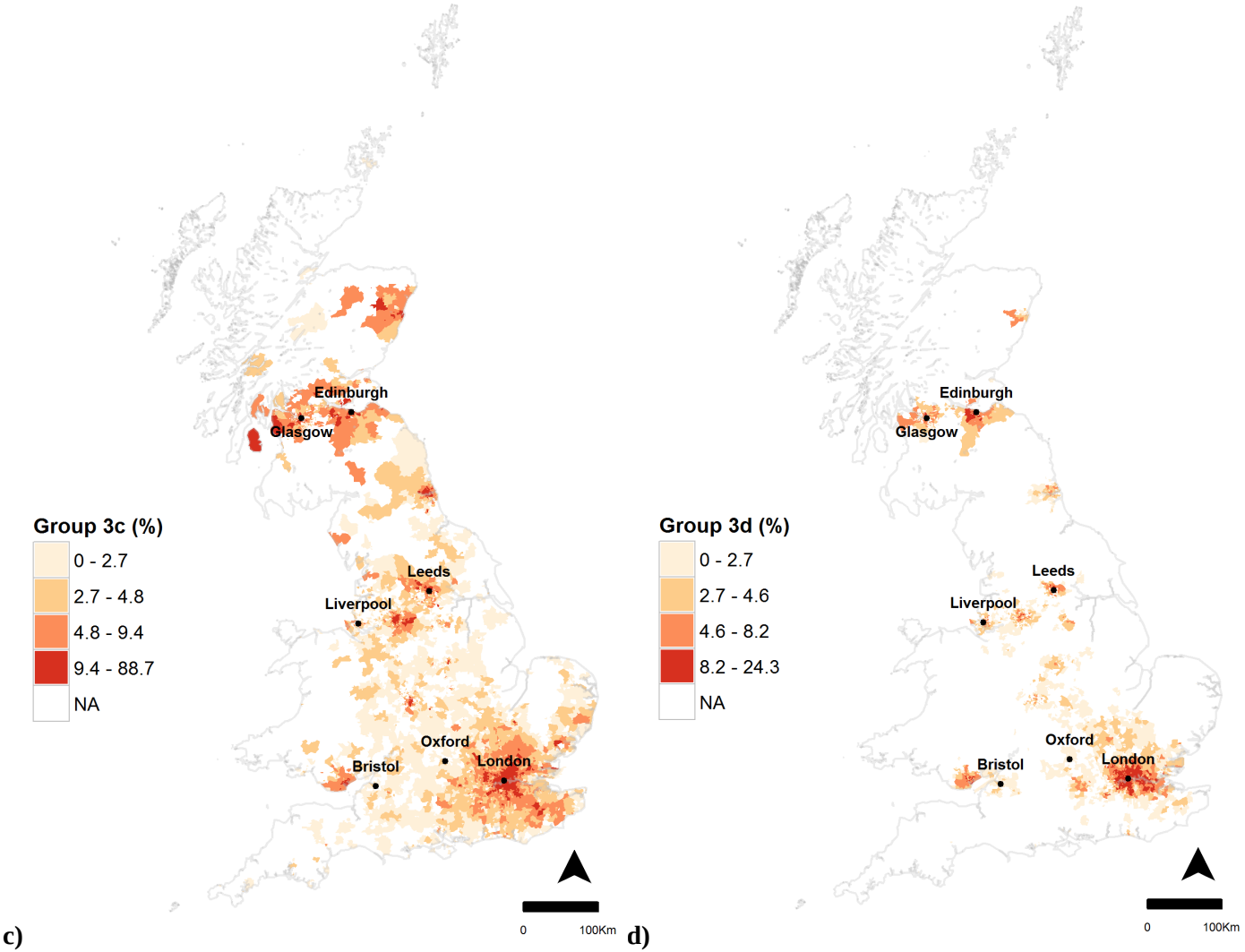
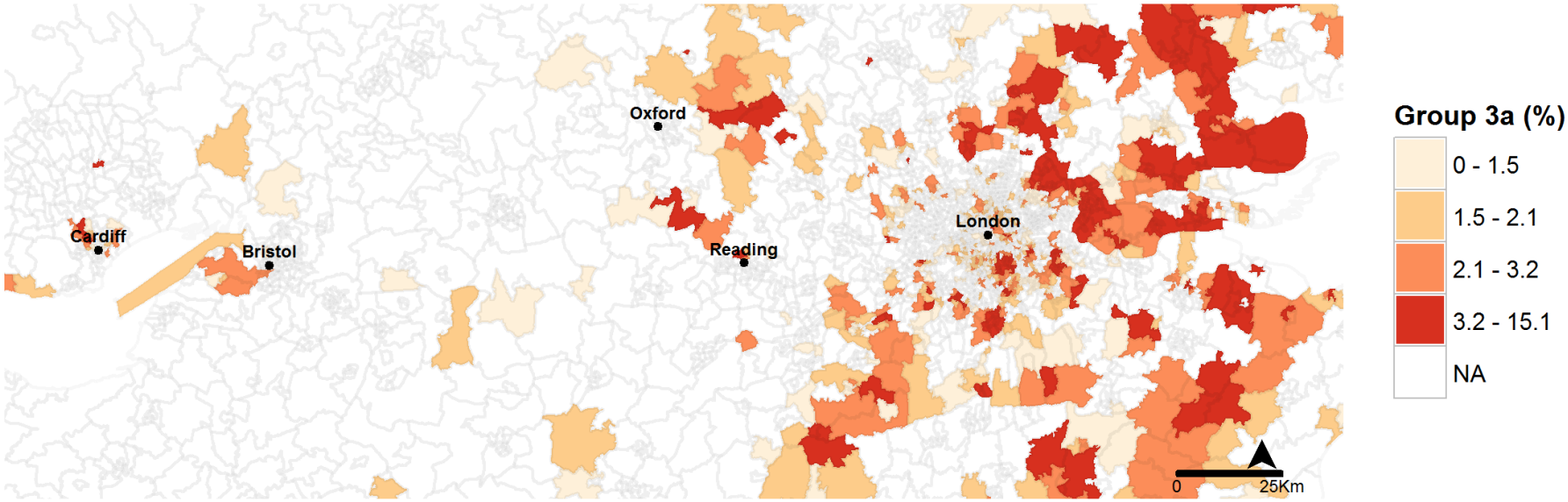


Figure 6.33: The percentage of customers per MSOA in a) Group 3a - 'Rural Fringe Commuters, b) Group 3b - 'Small-town Commuters, c) Group 3c - 'Stable Urban Workers', and d) Group 3d - 'Urban-living, Weekend Destination Shoppers' across Great Britain (quantile breaks). ***'NA' = no customers in group present.***

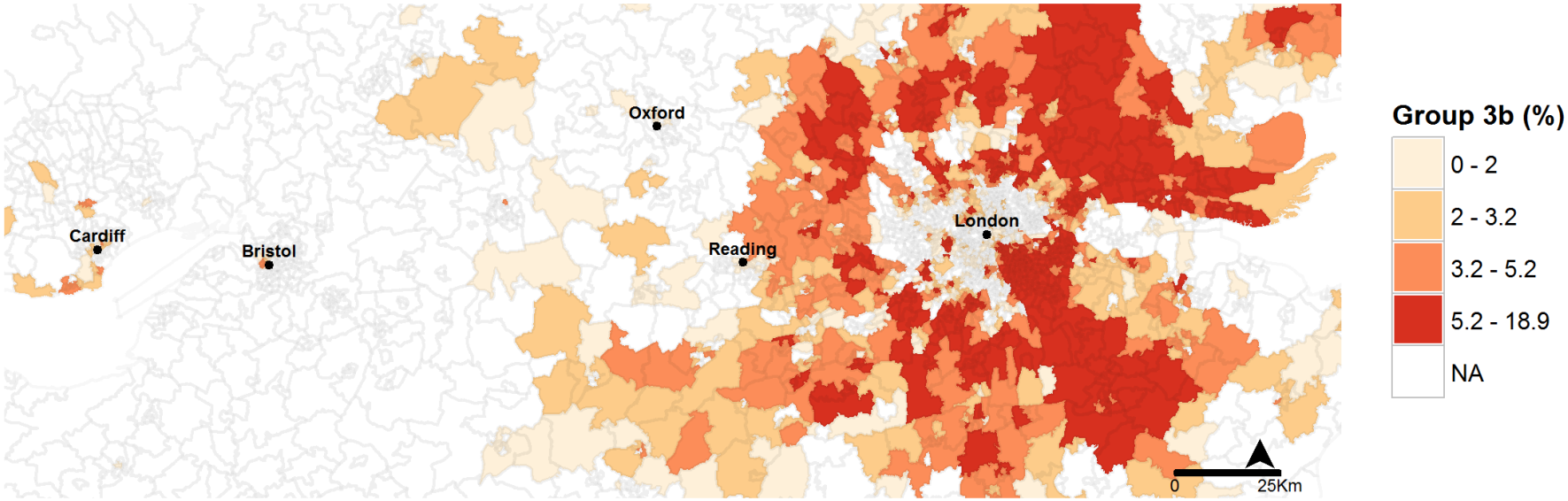
Group 3a – 'Rural Fringe Commuters'



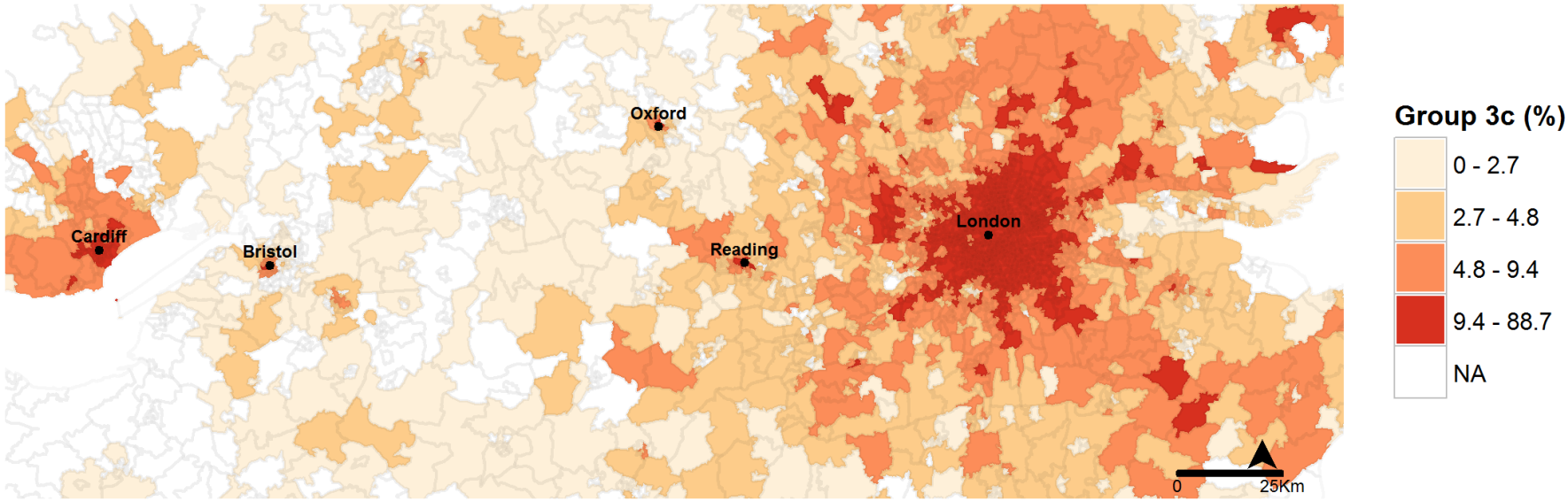
a)

Group 3b – ‘Small-town Commuters’

b)

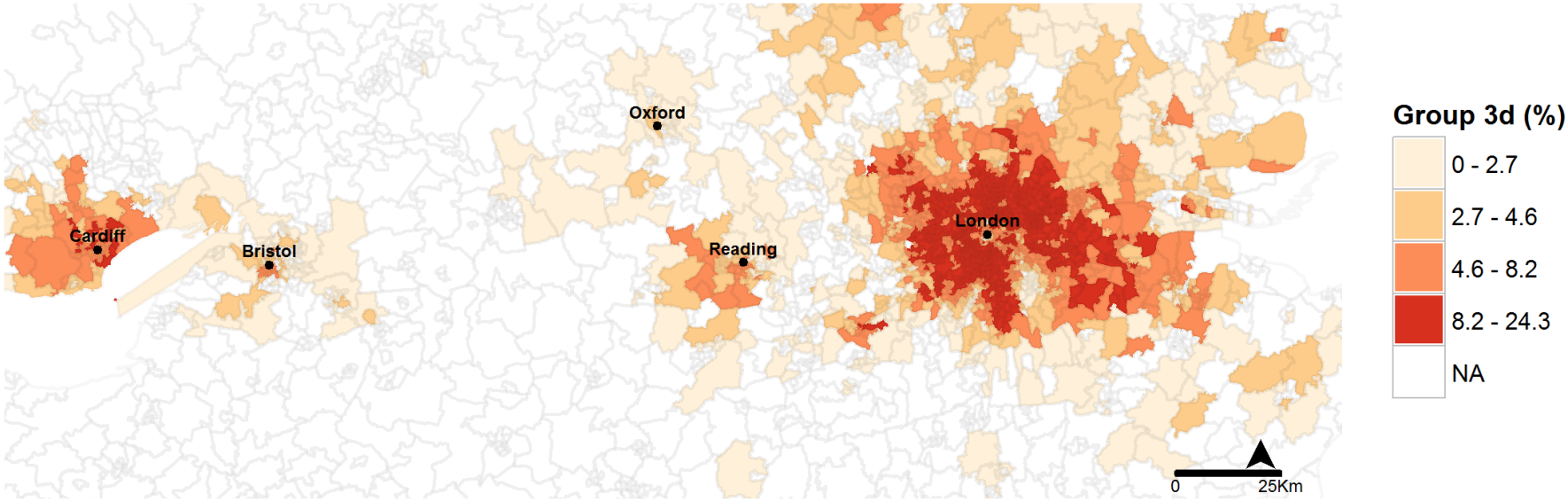


Group 3c – ‘Stable Urban Workers’



c)

Group 3d – 'Urban-living, Weekend Destination Shoppers'



d)

Figure 6.34: The percentage of customers per MSOA in a) Group 3a, b) Group 3b and c) Group 3c and, d) Group 3d across Southern England (quantile breaks). ***'NA' = no customers in group present.***

6.3.2.4. Supergroup 4 – 'Large Destination Shoppers'.

Customers assigned to Supergroup 4 demonstrated highest patronage to store Supergroup 5 ('Urban Stable Destinations') during weekdays. Figure 6.35 shows the age distributions and temporal profiles for Groups 4a, 4b, 4c and 4d, Figure 6.36 a comparison of product consumption between Groups, Figure 6.37 product consumption during weekdays and weekends, and Figures 6.38 and 6.39 illustrate the volume of customers per Group across GB MSOA's. This was the second largest Supergroup and represented 34.2% of postcode assignments.

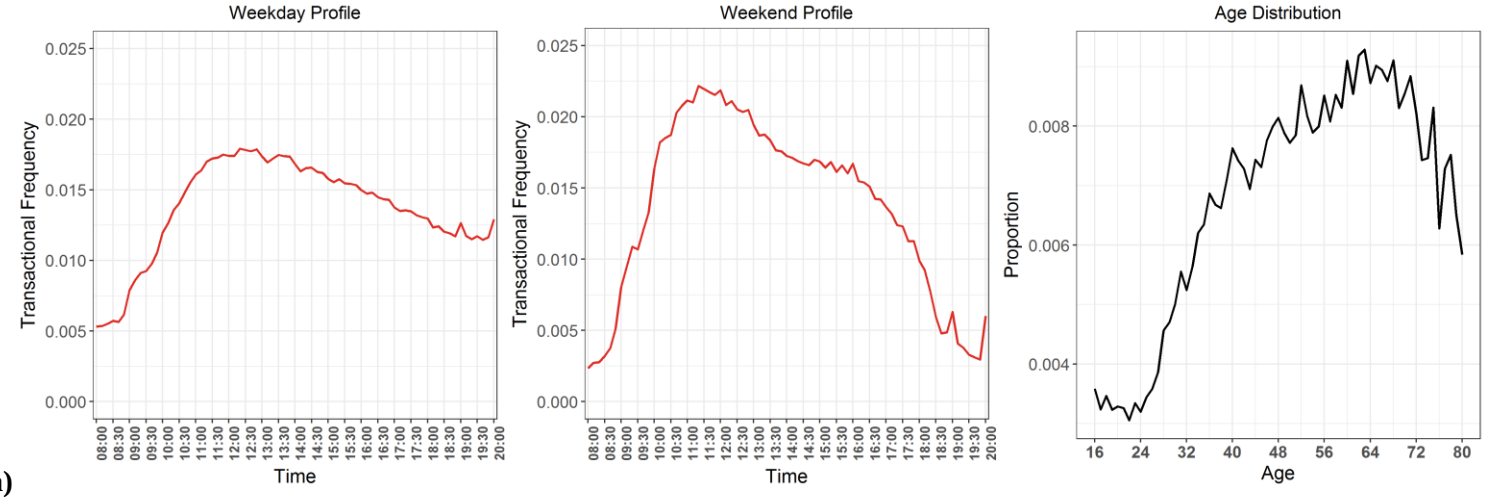
Distribution measures indicated that the majority of customers in this Supergroup interacted with 3 different store profiles during weekdays, and 2 on weekends, thus, similarly to the majority of Supergroups, showing higher variation during weekday periods. These customers were most likely to alternatively visit stores in Supergroup 2 ('Weekend Peak Destinations') during both weekdays and weekends. However, they predominantly patronised 'Stable Destinations', accounting for an average of 73.54% of their total transactions (10.6% of those assigned exhibited less than half). This Supergroup exhibited the second lowest average age overall, however, there was a higher range between Groups (the highest Group average was 49, lowest 36). In addition, due to the mix of store types apparent in store Supergroup 5 (i.e. retail parks versus urban flagships), there were greater distinctions in behaviour that required investigation of patterns at the store Group level.

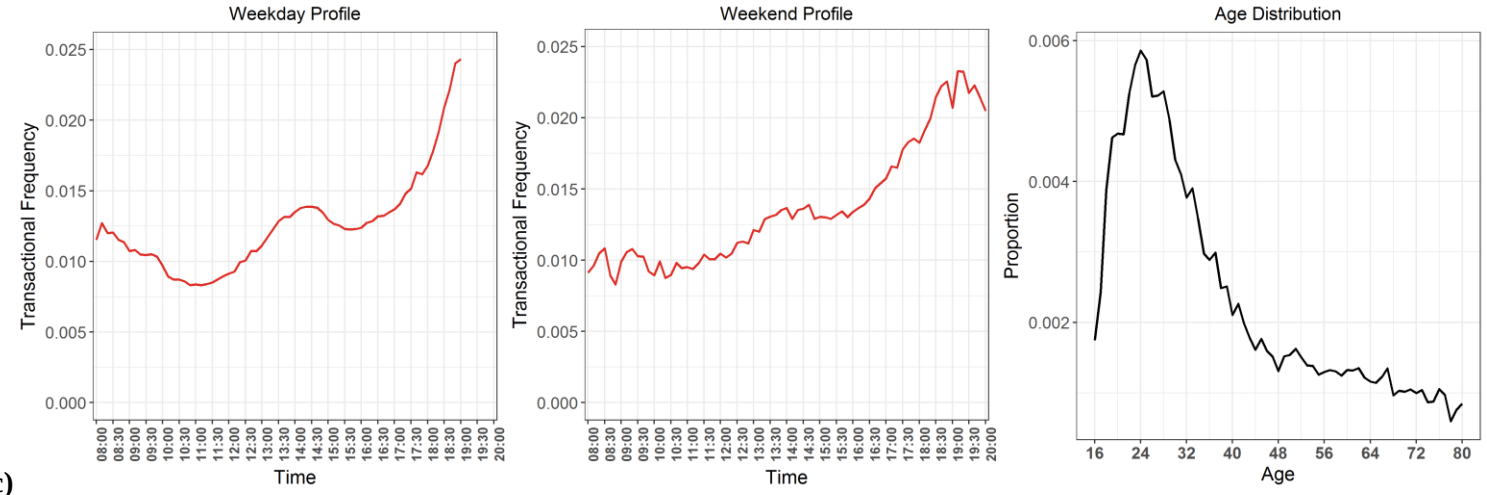
Consistent with previous observations, the oldest and most rurally located segment was Group 4a (of whom showed highest patronage to store Supergroup 1 during weekend periods). These customers demonstrated high cosmetics based consumption during weekdays, yet healthcare consumption during weekends and predominantly patronised store Group 5b (rural fringe retail parks). Group 4b exhibited the second highest average age, resided in urban fringe locations and predominantly patronised store Group 5a (urban fringe retail parks) during weekdays. Weekday consumption was high for cosmetics and gifting and weekend consumption a mix of cosmetics and healthcare essentials. Group 4c represented the youngest segment of this Supergroup, who primarily resided in urban areas and patronised Group 5c stores (destination-convenience mix urban flagships) during both weekdays and weekends. Product consumption was high for food, drink and essentials during both periods. Finally, Group 4d represented customers who patronised the same store type during weekdays and weekends. This predominantly described those patronising the same retail park stores during each period (i.e. either 5a or 5b). Consumption showed a mixture of cosmetics and beauty during both weekdays and weekends.

Temporal consumption varied substantially between these Groups. In line with previous observations of older cohorts, Group 4a demonstrated earlier activity, with peaks mid-morning

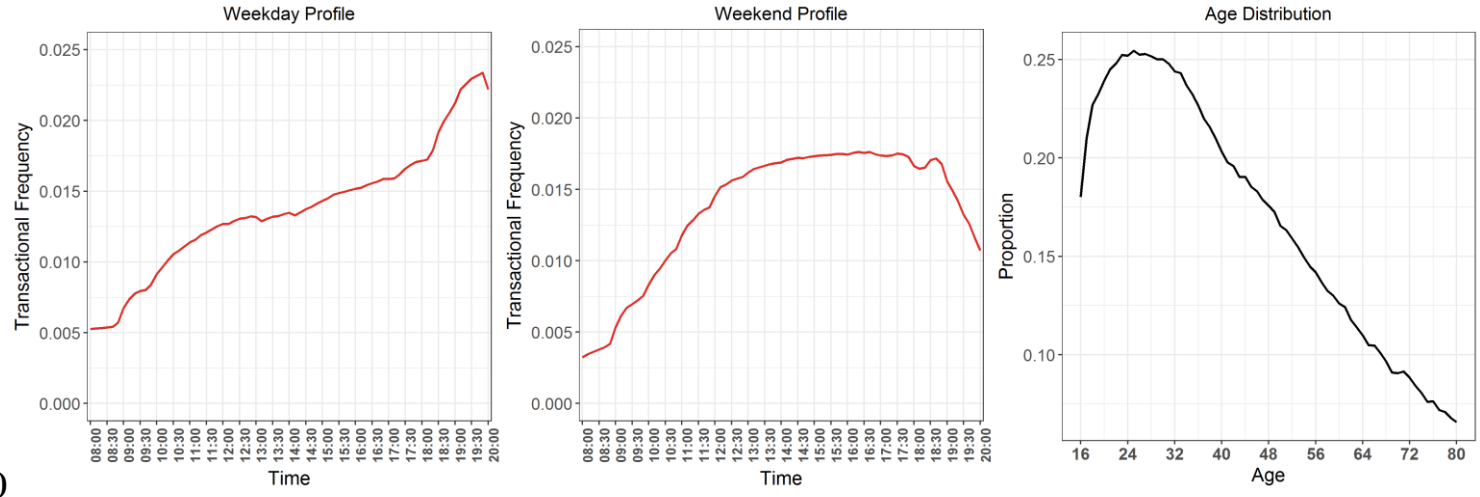
to afternoon during weekdays and midmornings on weekends. Group 4b demonstrated lunchtime and evening peaks during weekdays and mid-day to afternoon peaks during weekend periods. Group 4c demonstrated convenience usage (weekday early mornings, lunchtimes and evenings). These customers showed evening peaks during weekends, consistent with previous observations of younger age cohorts. Group 4d demonstrated afternoon to evening peaks (highest evening activity) and similar patterns on weekends.

Overall, these customers could be grouped based on their high patronage of large destination stores, however, the varying attributes of store Supergroup 5 (i.e. rural or urban fringe retail parks, or urban flagships) meant larger variation in characteristics were evident within this customer Supergroup. Consumption was highest overall for cosmetics and beauty items across all Groups.


Weekday Profile
Weekend Profile
Age Distribution
Transactional Frequency
Time
Proportion
Age
a)
b)



c)



d)

Figure 6.35: Temporal profiles (weekday, weekend, 10-minute intervals) and age distributions for a) Group 4a, b) Group 4b, c) Group 4c and d) Group 4d.

.

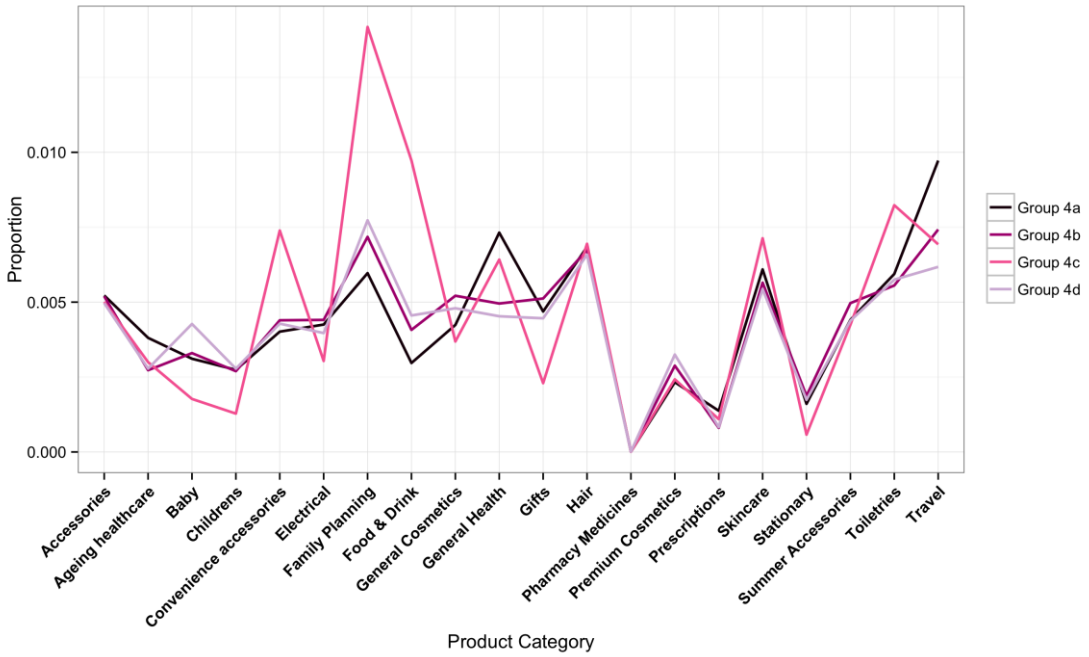
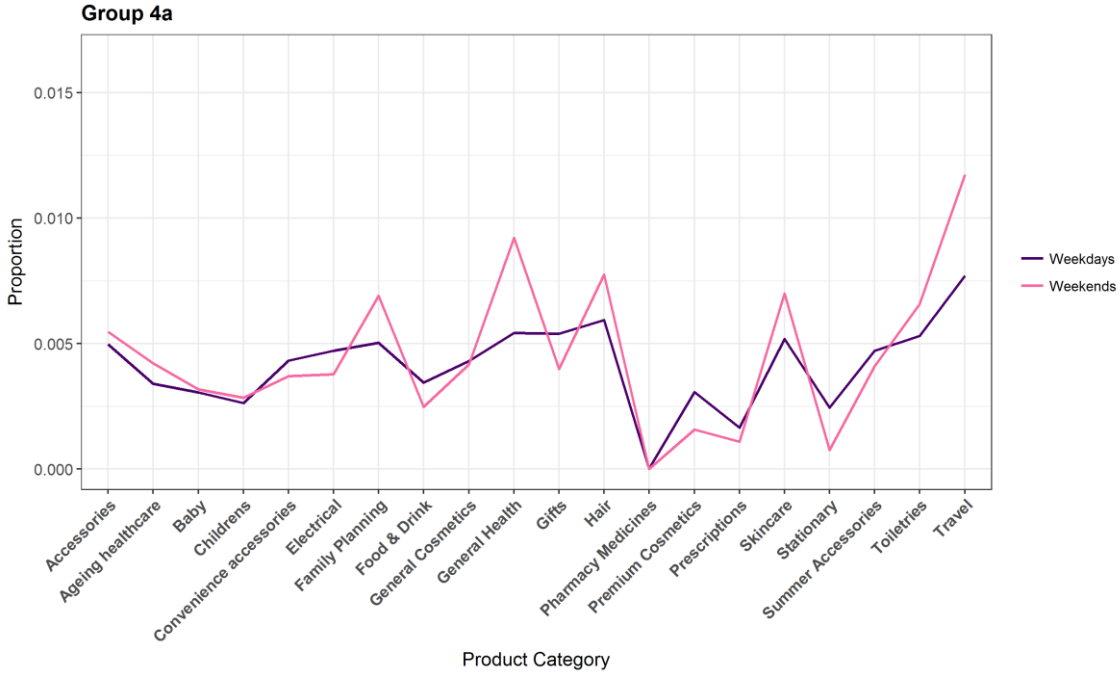
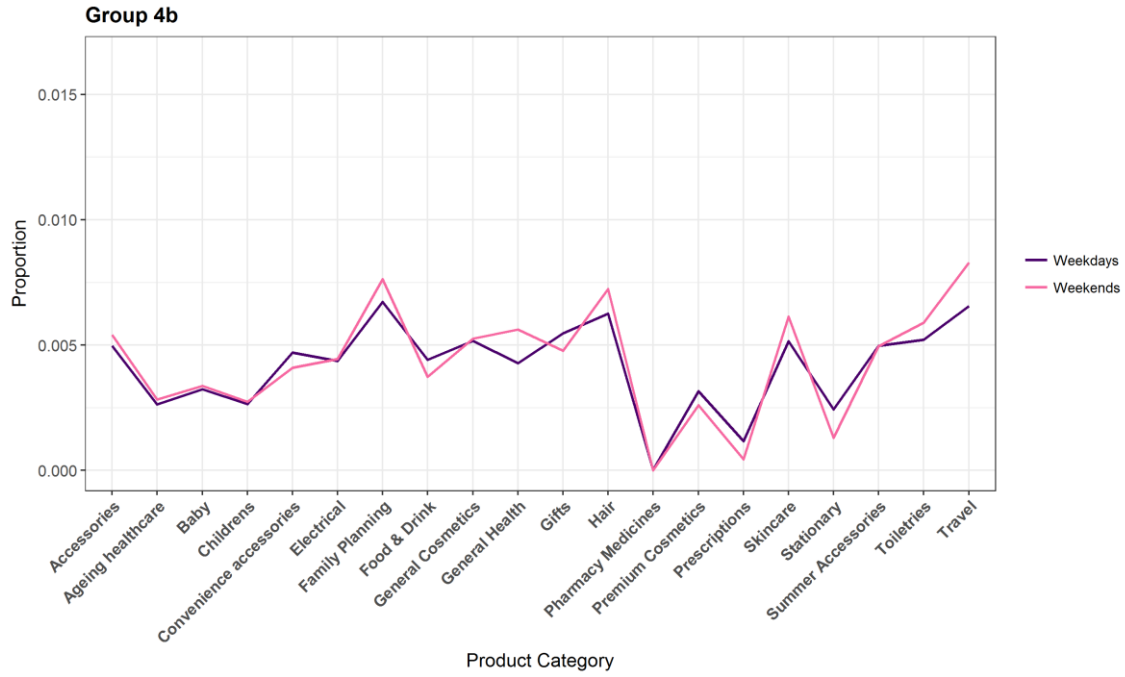


Figure 6.36: Comparison of product consumption (proportions) across Groups in Supergroup 4.

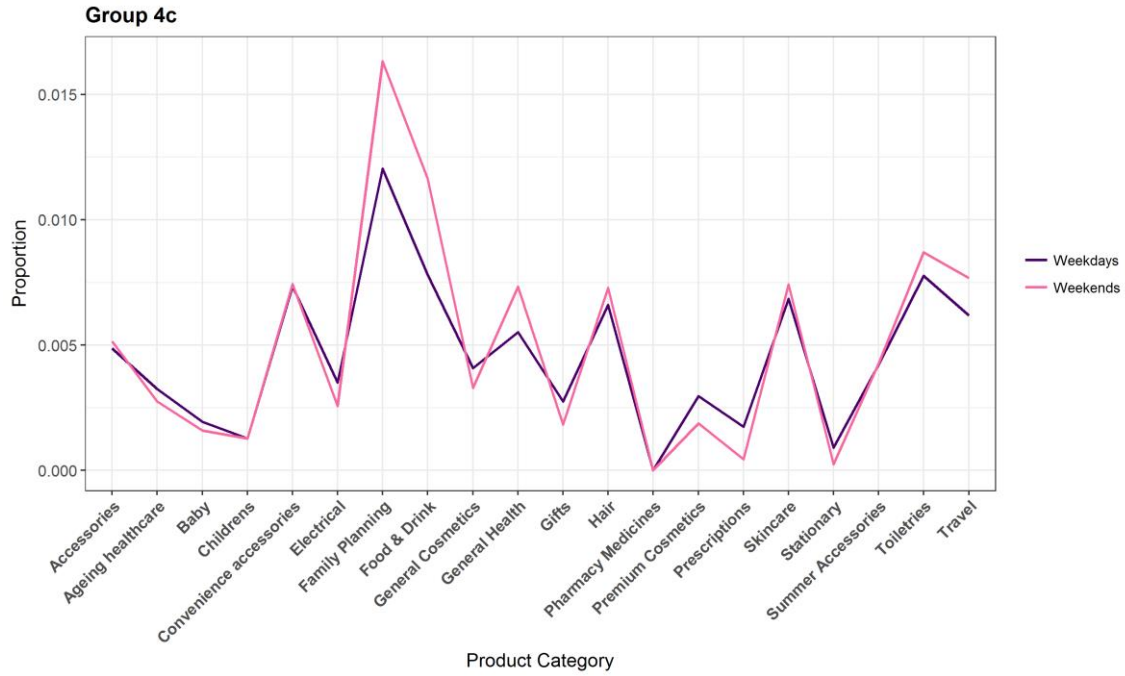
a)



b)


Group 4b
Proportion
0.015
0.010
0.005
0.000
Accessories
Ageing healthcare
Baby
Childrens
Convenience accessories
Electrical
Family Planning
Food & Drink
General Cosmetics
General Health
Gifts
Hair
Pharmacy Medicines
Premium Cosmetics
Prescriptions
Skincare
Stationary
Summer Accessories
Toiletries
Travel
Product Category
Weekdays
Weekends

c)


Group 4c
Proportion
0.015
0.010
0.005
0.000
Accessories
Ageing healthcare
Baby
Childrens
Convenience accessories
Electrical
Family Planning
Food & Drink
General Cosmetics
General Health
Gifts
Hair
Pharmacy Medicines
Premium Cosmetics
Prescriptions
Skincare
Stationary
Summer Accessories
Toiletries
Travel
Product Category
Weekdays
Weekends

d)

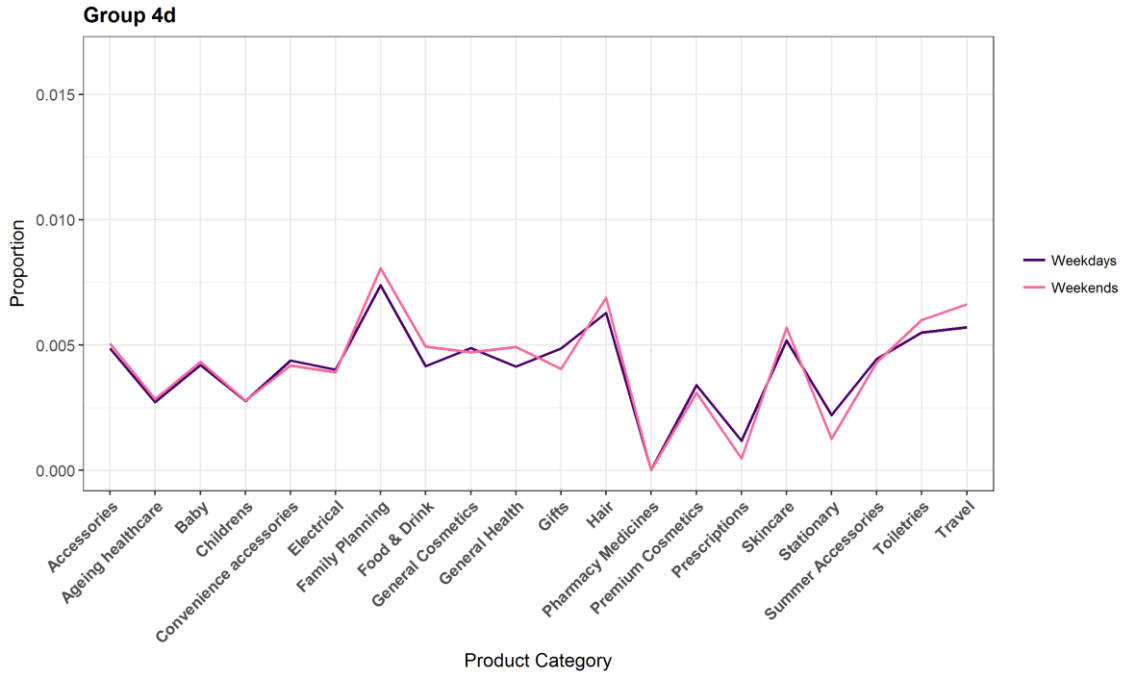
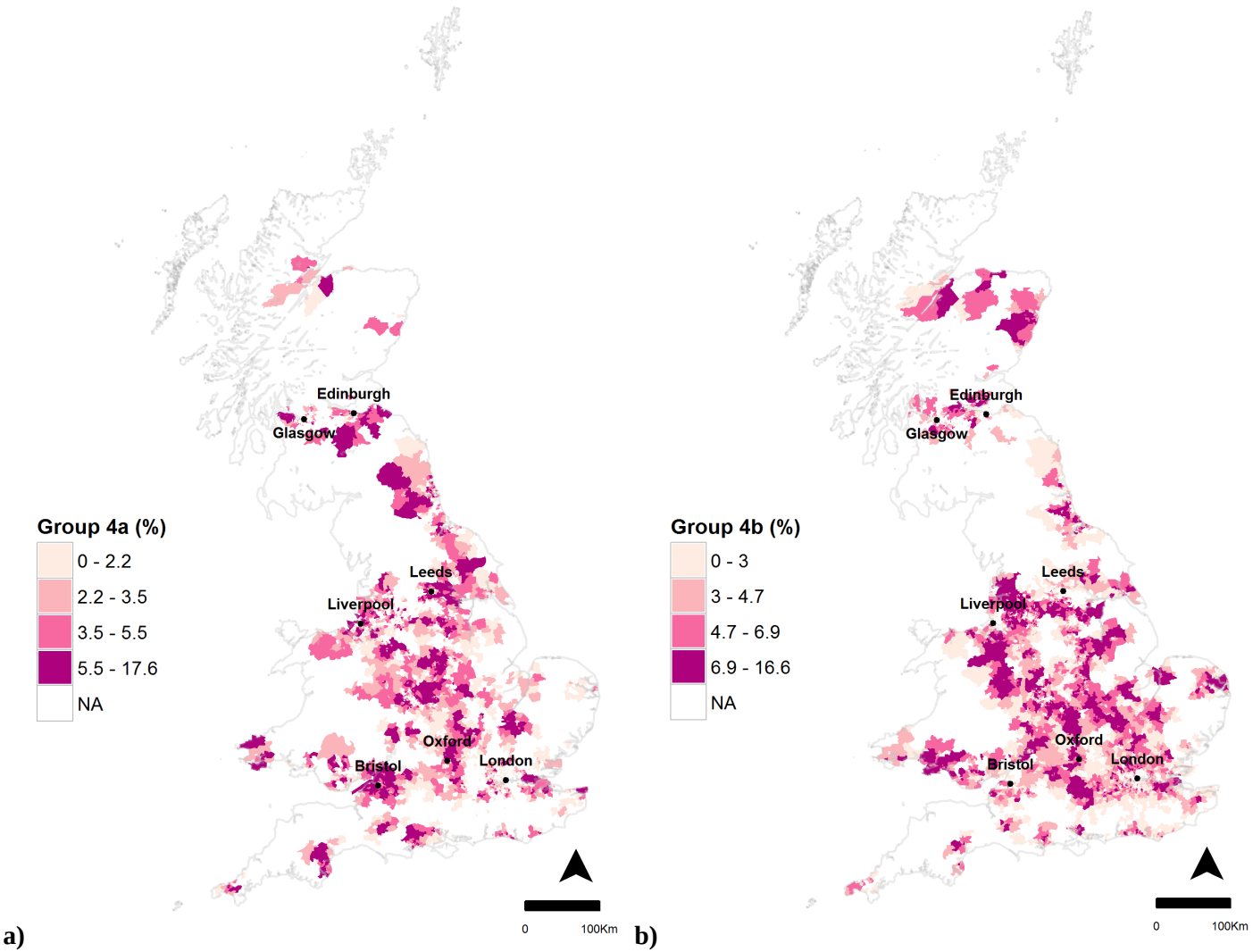


Figure 6.37: Comparison of product consumption (proportions) during weekdays and weekends for a) Group 4a, b) Group 4b, c) Group 4c and d) Group 4d.


Group 4a (%)
0 - 2.2
2.2 - 3.5
3.5 - 5.5
5.5 - 17.6
NA
Edinburgh
Glasgow
Leeds
Liverpool
Oxford
Bristol
London
0
100Km
a)
Group 4b (%)
0 - 3
3 - 4.7
4.7 - 6.9
6.9 - 16.6
NA
Edinburgh
Glasgow
Leeds
Liverpool
Oxford
Bristol
London
0
100Km
b)

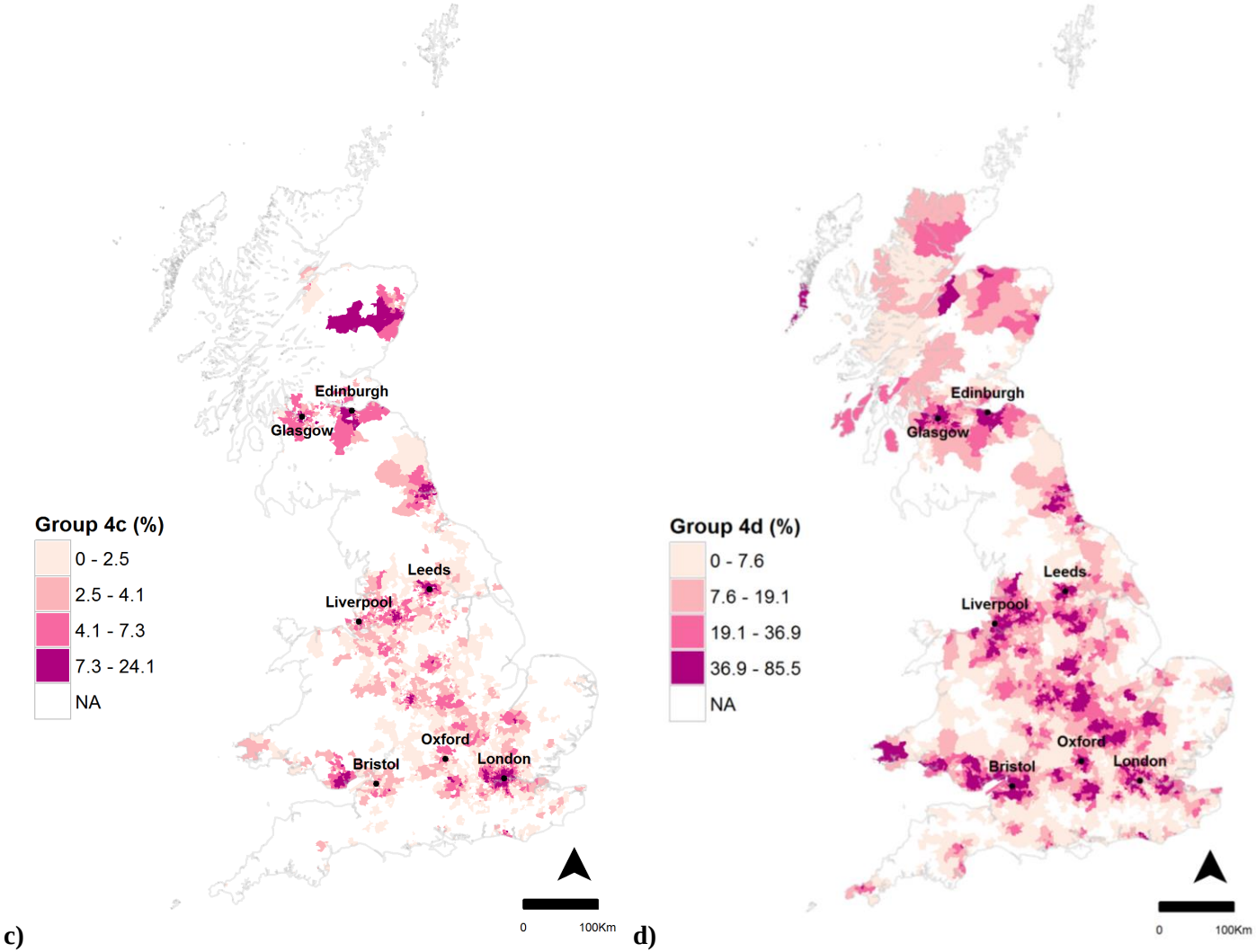
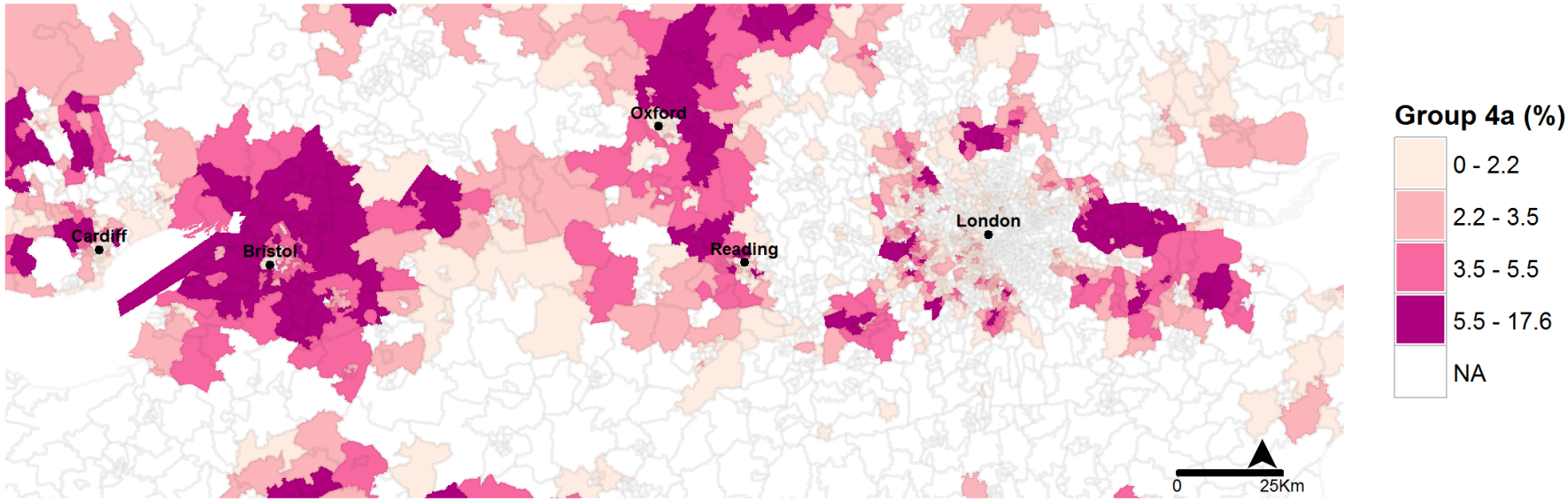


Figure 6.38: The percentage of customers per MSOA in a) Group 4a - 'Rural Fringe, Weekday Destination Shoppers', b) Group 4b - 'Urban Fringe, Weekday Destination Shoppers', c) Group 4c - 'Urban Weekday Destination Shoppers', and d) Group 4d - 'Stable Urban Destination Shoppers' across Great Britain (quantile breaks). ***'NA' = no customers in group present.***

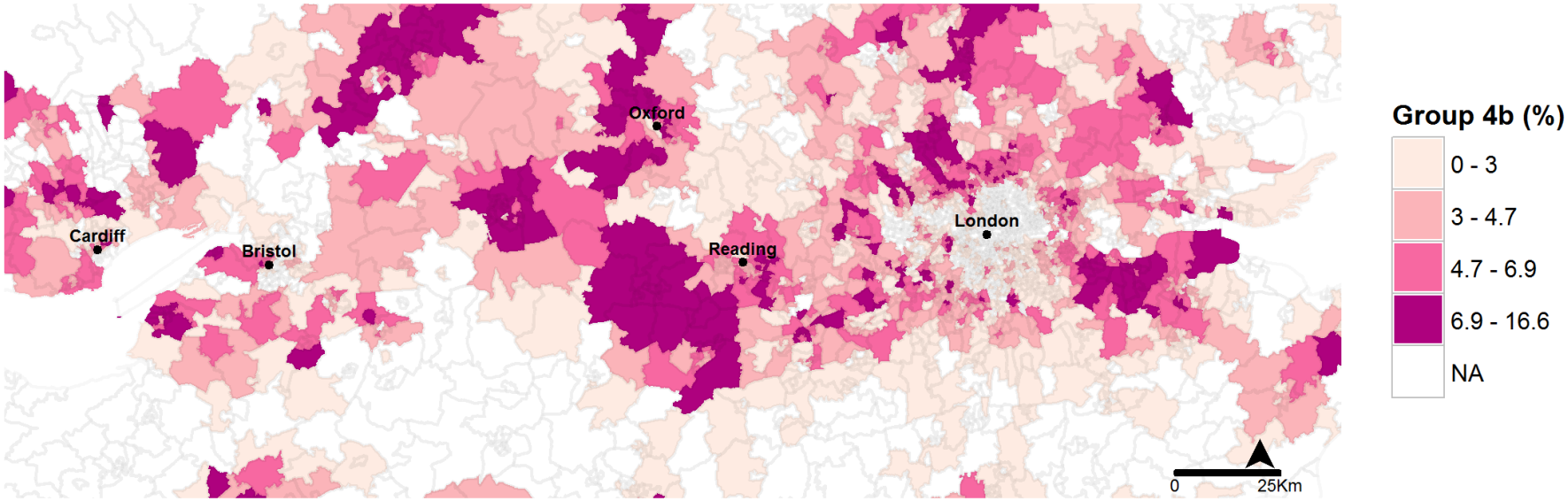
Group 4a – 'Rural Fringe, Weekday Destination Shoppers



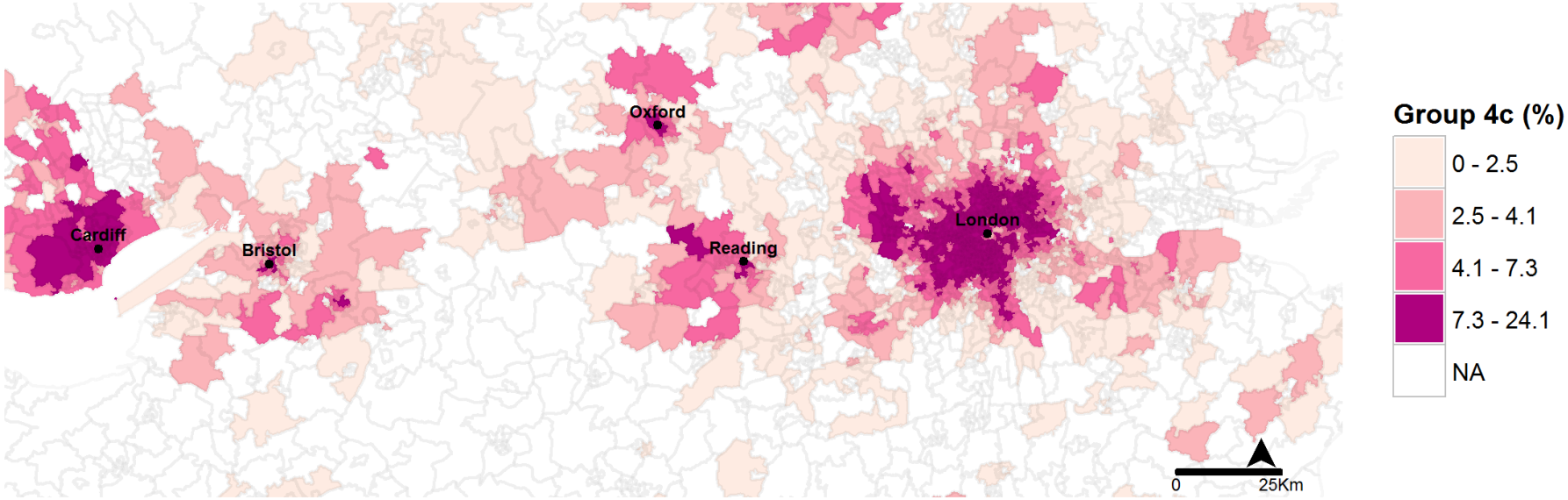
a)

Group 4b – 'Urban Fringe, Weekday Destination Shoppers'

b)

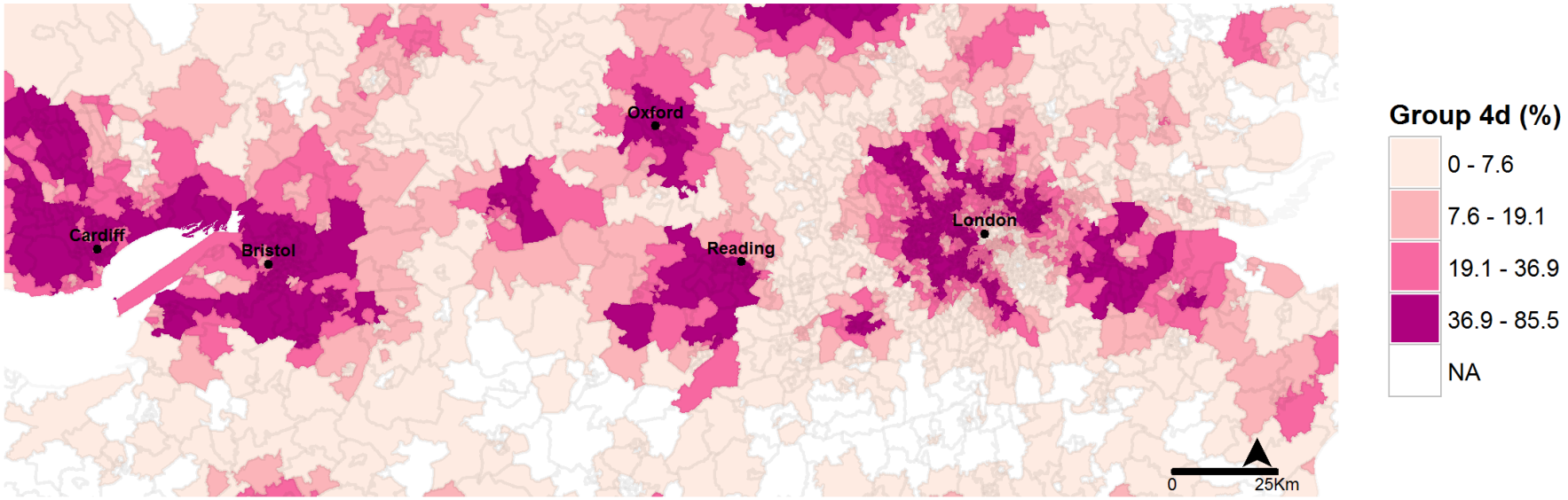


Group 4c – 'Urban Weekday Destination Shoppers'



c)

Group 4d – 'Stable Urban Destination Shoppers'



d)

Figure 6.39: The percentage of customers per MSOA in a) Group 4a, b) Group 4b and c) Group 4c and, d) Group 4d across Southern England (quantile breaks). ***'NA' = no customers in group present.***

6.4. Discussion and Conclusions

This analysis aimed to understand relationships between the temporal rhythms of HSR store locations the distinctive characteristics of the consumers that patronised them. This was achieved through the segmentation of HSR customers based on their patronage to stores that exhibited distinct temporal patterns. The results from these outputs revealed that a customer's weekend profile may be highly indicative of their residential location type and their weekday profile of their activity patterns. The combination of these attributes provided insight into the geodemographic characteristics of individuals and resulted in 4 distinct customer Supergroups and 14 Groups who exhibited unique temporal, demographic, geographic and consumption profiles. These patterns were able to be derived exclusively from the input of transactional frequencies exhibited by HSR stores over daily and weekly intervals.

These findings support the notions of rhythmanalysis and time geography, that focusing on everyday activities can reveal how the rhythms of people and places are ordered, and how these orderings may vary by social group. For example, whilst Chapter 5 indicated how temporal rhythms may indicate the characteristics of retail centres, this analysis demonstrated how the temporal profiles exhibited by those centres typically corresponded to that of the individuals who patronise them, and their distinct geodemographic and consumption characteristics could be inferred as a result. Furthermore, this information was distinguishable from a novel consumer dataset, which facilitated the analysis of spatiotemporal consumption patterns at a finer granularity than has previously been incorporated in population studies. From the perspective of geodemographics, whilst these insights may only be representative of HSR store locations and the HSR population, they support the general view that incorporating both spatiotemporal population activities, and alternative big data sources, may enrich our understanding of people and places. For instance, this work suggests that there may be alternative dimensions to identity that are not captured by static, residential based measures, and thus support the incorporation of alternative big data sources as indicators of social structure.

For retailers, who similarly continue to utilise static population measures, these insights suggest that exploiting the velocity of these data could aid identification of likely consumption patterns (including when, where and what customers are likely to consume), which could subsequently provide insights for optimisation of targeted marketing at specific times/store types. In addition, whilst not all customers could be utilised in the creation of these profiles (i.e. due to low transactional data), the assignment of profiles to postcodes provides a means of classifying customers based on their residential information and thus provides a means of inferring consumer characteristics in instances where there may be a lack of data pertaining to an individual.

However, from a wider perspective, these findings also build on the high street resilience challenges previously discussed in Chapter 5 (Section 5.4). For example, in addition to quantifying how people interact with various retail centres, there is a prominent gap in knowledge surrounding the demographic compositions of these locations, how this may evolve and change over time, and the impact of this on their resilience. This study shows an example of how elements of demographic composition may be inferred through the spatiotemporal analysis of loyalty card transactions. This could aid in understanding dynamics such as volumes of interaction generated by different demographic groups, how compositions change over time, impacts of accessibility in the context of local demographics, or how convenience trends may be defined for different types of customer. This may similarly aid in the modelling of locally relevant solutions to meet the needs of evolving consumer trends.

Limitations to acknowledge here are that firstly, these outputs present a bespoke representation of activities relevant to the activities of one retailer's loyalty card population. As outlined in previous analyses, this highlights a need for future endeavours of this kind, using datasets of differing representations, to further understand how the trends observed may be extrapolated to that of the general GB population. In addition, inherent bias in these data (see Chapter 2) may influence trends over a number of dimensions. For instance, behaviours may be over or under-represented in certain areas (which may become more apparent through active customer selection) and certain products may be over-indexed (i.e. higher value categories such as 'premium cosmetics' and 'Electrical'). Yet, unique patterns were still identifiable between customer segments that strongly indicated relationships between demographic profiles and consumption characteristics (such as 'Ageing Healthcare' products for older cohorts or 'Family Planning' for younger cohorts). Overall, these findings consolidate the potential benefits of incorporating temporal dynamics, as facilitated by novel consumer datasets, to understanding the organisation of societal flows, indicating that there may be alternative dimensions over which individuals can be represented in the study of people and places.

7. HSR Areas and Activities

7.1. Introduction

Chapter 6 demonstrated how distinct temporal rhythms may be related to the geodemographic characteristics of individuals. However, research has also indicated that the types of places people visit may be an equally important element in the study of people's spatiotemporal experiences (i.e. Kwan, 2012a), much of which will be influenced by the differing daily habitual obligations of socially distinct groups. For instance, people with different personal and household attributes may have different temporal routines and thus varying accessibility and environmental exposure influences (Neutens et al., 2010; Delafontaine et al., 2011). This would suggest that geodemographic indicators might be quantifiable not only by where people live and their temporal movements, but also by the places that they visit and when they visit them.

To date, enquiries regarding these types of daytime population activities have been limited by data availability and dynamics have been inferred from cross-sectional studies, where data are actively solicited via qualitative methods. As a result, insights have been derived from small sample sizes over limited time periods (Chen et al., 2016). However, the spatiotemporal attributes inherent in novel consumer datasets offer a framework for exploring the concepts of time geography and population activity patterns from a data-driven perspective, providing a much larger population sample than has previously been obtainable, over much more granular temporal intervals and longitudinal periods.

This chapter presents a broad exploration of the ability to extract daytime location-visiting activity patterns from loyalty card data, which has not previously been implemented using historically recorded consumption behaviours on a large scale. Based on previous gaps in population research, the focus here was on demonstrating how loyalty card data may be utilised to quantify differences in the types of places that distinct social groups interact with. In addition to this, research has suggested (i.e. Singleton and Longley, 2009; Webber and Longley, 2003) that locational context may influence the extent to which we can generalize the behavioural characteristics assigned to individuals within classifications. Therefore, a secondary aim of this analysis was to explore how the location-visiting characteristics of customer Supergroups and Groups varied over different geographical regions.

This was achieved by augmenting customer activity patterns with the Census based COWZ classification, which describes the characteristics of areas based on workday population

characteristics. Similarly to the work presented in previous chapters, this analysis was inherently limited to the study of 'places' in the specific context of retail centres in which HSR stores reside. However, these outputs served to consolidate the work of previous chapters and enforce the types of insights that can be extracted from loyalty card data. The implications of exploiting locally available time series data from alternative commercial sources are discussed both in terms of enriching geodemographic representations and understanding the resilience of high street economies.

7.2. Method

As outlined in Chapter 2, there are currently few widely recognised methods for analysing complex relationships among human space–time trajectories in big data, particularly in terms of reliable linkage to other relevant attributes, such as socioeconomic context (Kwan, 2013). Longley (2017) suggests how this can be achieved through triangulating big data sources with traditional administrative datasets in an attempt to quantify their socioeconomic value, their uses and to make sense of trends. Whilst there are many existing Census based geodemographic classifications, the majority of these have been derived from residential population databases. However, a viable alternative for contextualising daytime activities captured by loyalty card transactions is the COWZ Classification (Cockings et al., 2015). This was produced to provide users with unique insights into the characteristics of workers and workplaces, at a small area level, and thus describes socioeconomic phenomena during daytime periods.

As outlined in Chapter 3, the COWZ classification describes WZ's in England and Wales over 7 Supergroups and 29 Groups. WZ's are a unique geography, based on Census workplace data (information collected about workers and workplaces) created by splitting and merging OAs, which were designed to represent the geographical distributions of residents and residences. COWZ classifies WZs according to their similarity in terms of Census variables across 4 domains: composition of the workplace population, composition of the built environment, socioeconomic characteristics of the workplace population and employment characteristics of the workplace population (Cockings et al., 2011). This provided a socioeconomic framework for contextualising the characteristics of various HSR store locations. COWZ is based on data from the 2011 Census of England and Wales, therefore, Scotland were was excluded from these analyses. The method adopted to quantify these dynamics was a trip distribution analysis of interactions between the customer types derived in Chapter 6, and location types as defined by the COWZ classification. As illustrated in Chapter 4 (see Section 4.3.1.2), creating trip distribution matrices from these data provides a useful means of identifying volumes of interaction between origins and destinations in large datasets. In this context, the Supergroup or Group assigned to a customer represented origins, and COWZ area types in which HSR stores were located represented destinations.

The analysis presented over the subsequent sections describes differences in customer location visiting dynamics both overall and during different temporal intervals. The time periods of interest for this investigation were weekly (weekday versus weekend) and monthly, which aimed to identify broad differences in the location visiting behaviours of customer types during different times of the week (based on the previously observed importance of weekday versus weekend variations), and how these may also vary between months (or seasons). These relatively aggregate intervals were selected based on a compromise between gaining the lowest temporal granularity possible, whilst also ensuring disclosure control. For instance, data pertaining to hours per day, or days per week, resulted in low frequencies within some Groups. Therefore, temporal aggregations were necessary in order to present data at the lower Group level of the customer classification. Based on the aims, it was considered of higher importance to obtain non-disclosive data at this level, rather than utilising only the customer Supergroup level with more granular temporal intervals.

For the regional analysis, the temporally aggregated trip distribution data were obtained for each of the 10 regions across England and Wales (9 English regions, and Wales). Due to this extra level of segmentation, these data could only be presented at the customer Supergroup level to maintain disclosure control.

It should be noted that there would evidently be an uneven distribution of HSR stores across COWZ classes. For example, the classification contains specific 'Retail' oriented area types, which will inherently contain higher proportions of a high street retailer's store locations. However, due to the extensive network of this HSR, there were store locations present in every COWZ category (see Appendix 3), which allowed suitable comparison between location visiting characteristics. COWZ Supergroup and Group pen portraits were used to interpret results, which are freely available to download through the ONS (ONS, 2018a).

7.2.1. Data Preparation

For the weekly analysis, data were aggregated between the hours of 7am - 7pm during weekdays, and 8am – 6pm during weekends (weekend hours were reduced due to fewer stores being open during these early/late periods, resulting in disclosive frequencies within some Groups). For the monthly analysis, data were aggregated by month (Jan-Dec). Both datasets were derived from the full 2.5 years of transactional data. Trip distributions describing customer activities were generated between each customer Supergroup/Group and COWZ area type during the time periods of interest. Store locations were appended to COWZ area types by obtaining the WZ in which they were located and the class associated with it. Trip frequencies between customers within each Supergroup and Group could then be obtained via linkage to unique store IDs. This analysis aimed to quantify volumes of interaction between people and

place types, therefore, all trips generated by customers across the 2.5 year period were included in these distributions. The resulting frequencies thus comprised one count for each time an individual transacted in a given area.

Analysis of these data required weighting in order to extract distinct trends independent of underlying volumes of activity. For example, as illustrated in Chapter 3 (see Figures 3.13 and 3.14), HSR activity is higher during weekdays (mid-day to afternoon) and similarly on weekends, with lower overall magnitude. Monthly activity delineated peaks between April and September, and also during the Christmas period (see Figure 3.15). Regional trends demonstrated patterns consistent with differing population densities per area, such as increased volumes in the South, in particular, London (see Appendix 3). In addition, when comparing trends between Supergroups and Groups, weighting was also required to account for differences in sample sizes. Varying methods were applied to extract the desired trends relevant to each analysis, of which are described throughout the proceeding sections.

7.3. Customer Location Interactions

Figure 7.1 shows the percentage of total activity per COWZ group, by customer Supergroup and Figure 7.2 shows this at the customer Group level (weighted by total activity per COWZ Group to extract comparable trends). The key observation was that prominent differences existed between the characteristics of retail centress that different types of customer interact with.

Findings were consistent, in each case, with previously observed characteristics of each segment. For example, ‘Rural Ageing Off-peak Shoppers’ (Supergroup 1) accounted for the highest proportion of activity in the COWZ ‘Rural’ Supergroup. This included 79.8% of all activity that occurred within the ‘Rural with non-local workers’ group and over half of other rural areas including ‘Rural with Core Services’ (describing rural service centres, rather than the most remote areas) and the most remote area types - ‘Rural with Mining or Quarrying’ and ‘Traditional Countryside’. Other prominent activity was recorded within ‘Market Squares’ (primarily describing small town locations) and non-metropolitan suburban areas. The COWZ ‘Rural’ Supergroup is characterised by; affluent, older populations, high levels of working from home, locations in rural and suburban areas and, occupations dominated by agriculture, manufacturing/industry and education.

Conversely, ‘Weekday Convenience Commuters’ (Supergroup 3) demonstrated extremely minimal activity in ‘Rural’ areas. Yet, they accounted for 78.5% of all activity that occurred within the ‘Global Business’ group. These areas describe workplaces with the highest status occupations, exhibit very high percentages of commuters, and are located in centres of metropolitan cities (primarily London). These trends demonstrate how comparable relationships

can be drawn between demographic characteristics inherent in loyalty card data, with that of Census based classifications.

Supergroup 2 ('Small Destination Shoppers') demonstrated a preference for rural and suburban areas (i.e. their highest activity was within 'Metro Suburban Distribution', of which can be found scattered across the outer suburban areas of major metropolitan centres) and 'Large Destination Shoppers' (Supergroup 4) demonstrated higher interactions with both metropolitan centres and COWZ classes that represented HSR retail parks, such as 'Industrial Units'. Further distinctions were evident at the customer Group level, largely showing how area visiting dynamics are likely dependent on the accessibility of area types to the local population, in line with their geographical distributions, as illustrated in Chapter 6.

Whilst these trends may be anticipated to some extent, these analyses provide data-driven evidence for relationships between individuals that exhibit certain demographic characteristics and the area types in which they are most active. They further suggest that quantifiable differences exist between the types of retail centres that socially distinct groups visit, interact with and are exposed too.

7.3.1. Temporal Location Interactions

Examining activities over various temporal intervals identified further distinctions in behaviour between customer types. Figure 7.3 illustrates a comparison of weekday and weekend activity volumes across COWZ Groups, per customer Supergroup. Figure 7.4 demonstrates an example of the further distinctions that could be made at the customer Group level (other Group Figures are supplied in Appendix 4). Data were weighted by activity volumes per area, per time period, and subsequently transformed into percentages per customer Supergroup/Group, to extract unique trends. Variations were prominent across weekday and weekend periods both between and within customer Supergroups, highlighting how area visiting dynamics are likely interrelated with the accessibility of area types in terms of proximity, but also potentially in terms of the constraints imposed by the differing daily obligations of distinct groups.

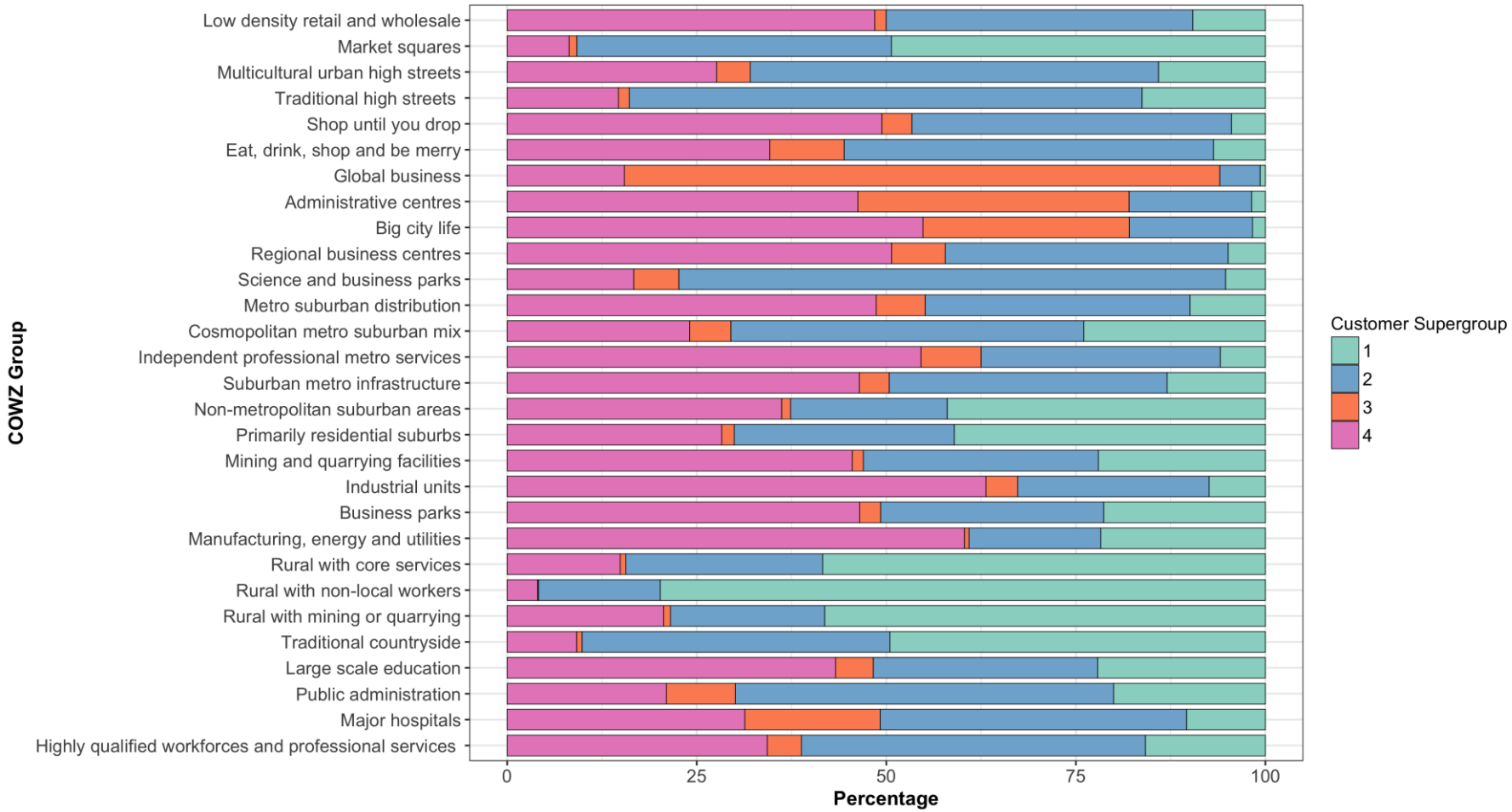


Figure 7.1: Percentage of total activity per COWZ Group, by customer Supergroup.

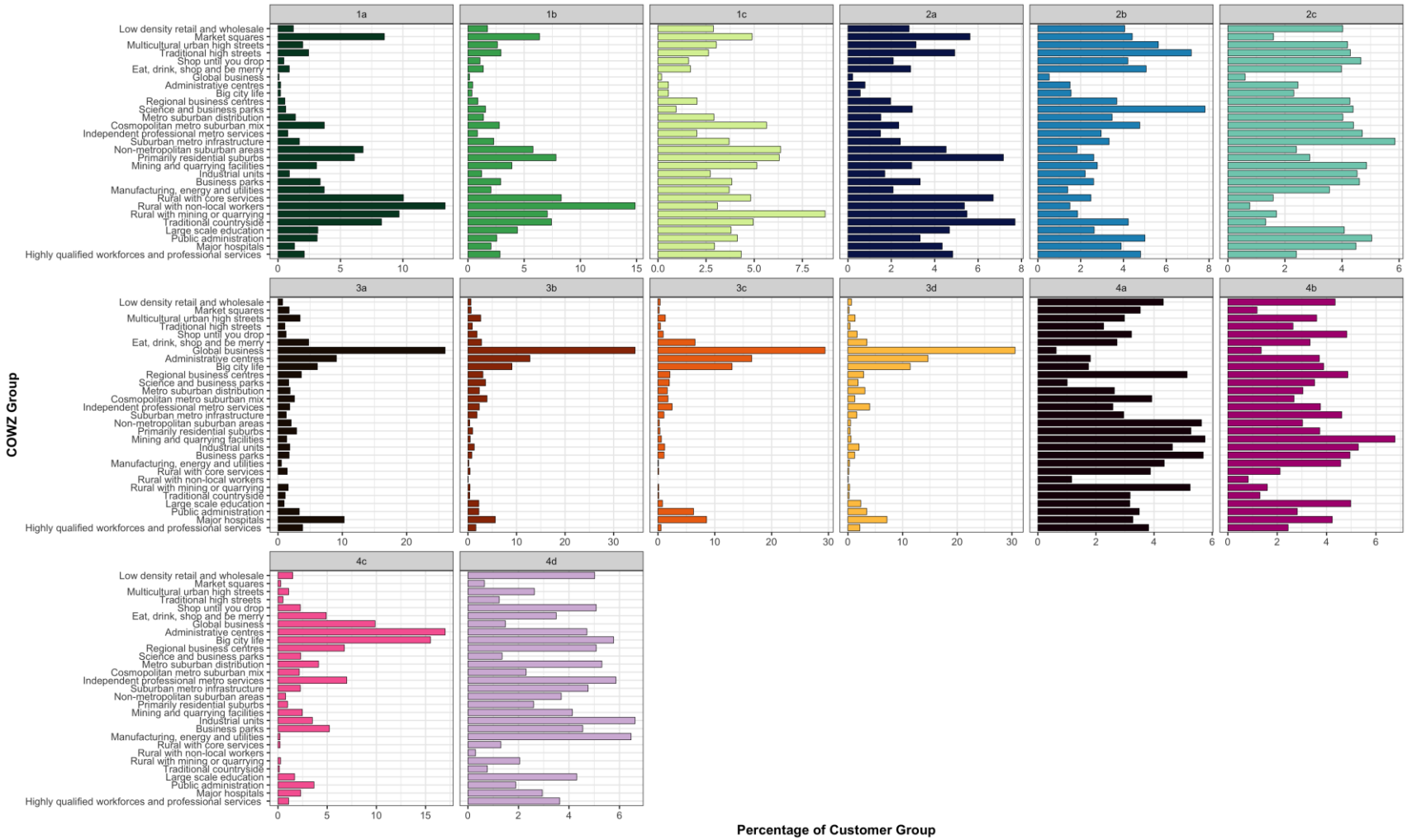


Figure 7.2: Percentage of customer Group activity, per COWZ Group (weighted by total activity per COWZ Group).

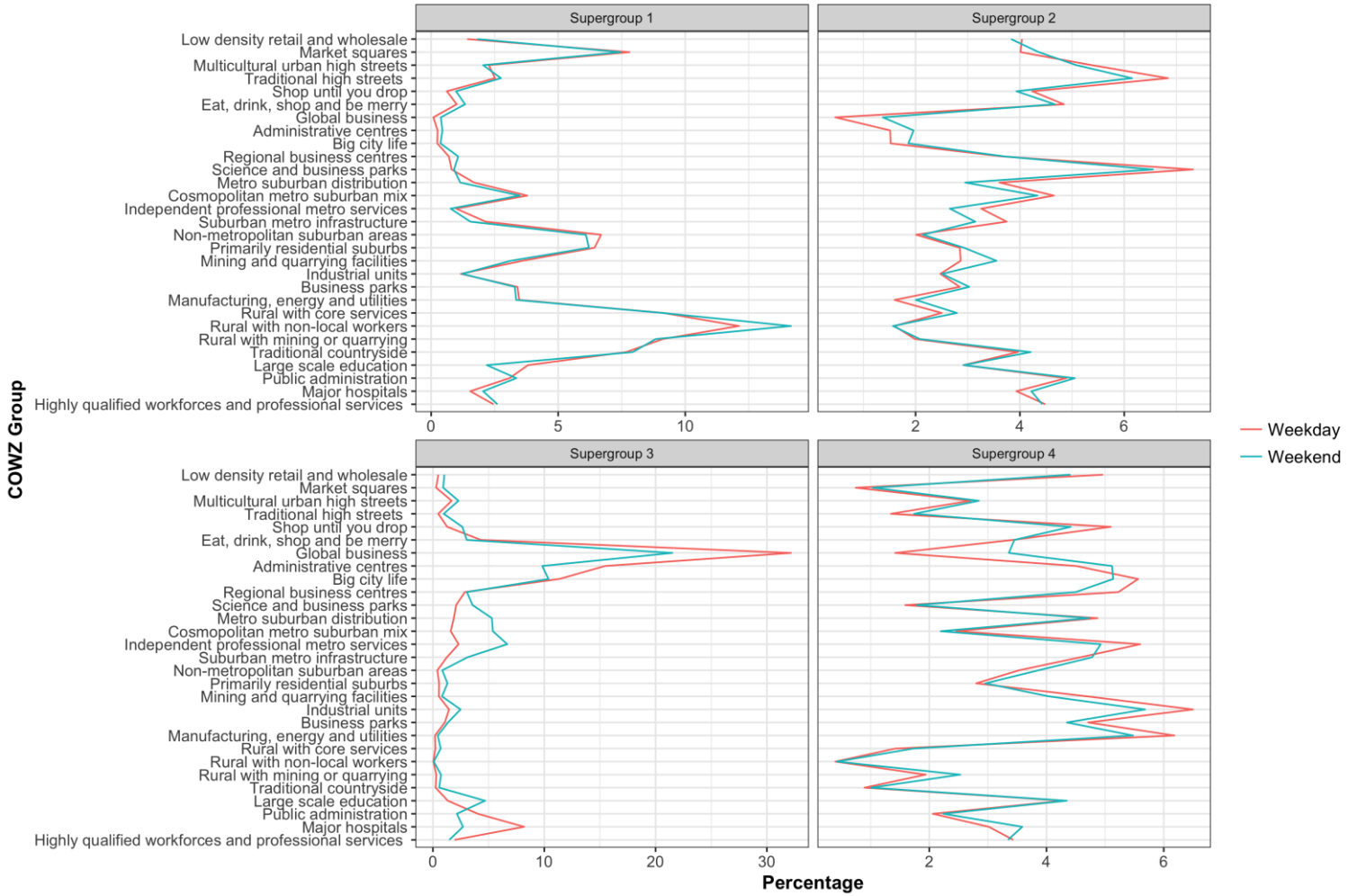


Figure 7.3: Supergroup area visiting characteristics during weekdays and weekends (Note: Scales are varying to illustrate fluctuations within each Supergroup).

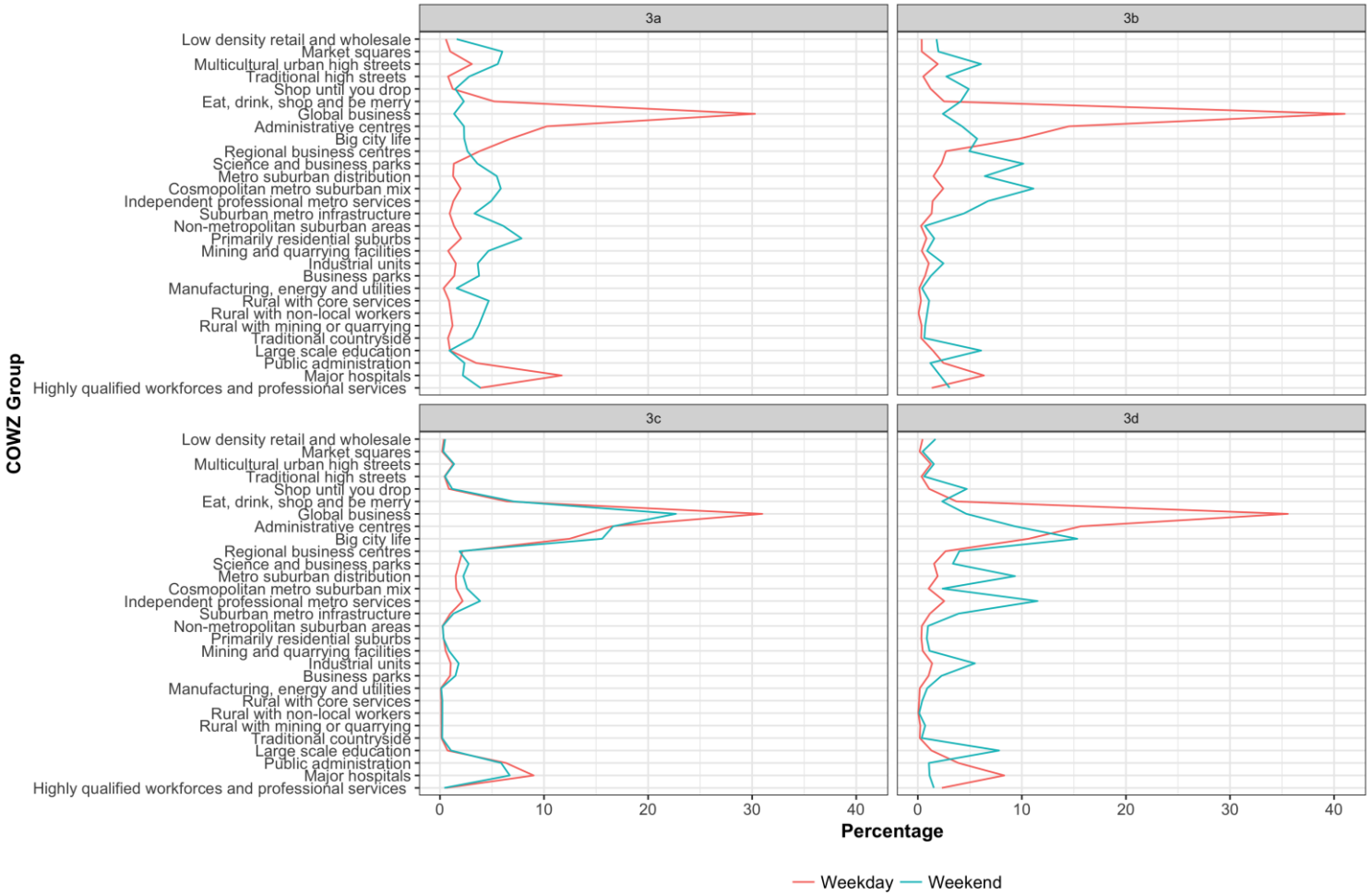


Figure 7.4: Supergroup 3 area visiting characteristics during weekdays and weekends (Note: Scales are varying to illustrate fluctuations within each Group).

Examples can be drawn from the obvious contrast in behaviour between Supergroups 1 and 3, where Supergroup 3 generated high volumes of interaction within central urban 'Top Job' areas during weekdays due to working hours, in contrast to Supergroup 1 customers who interacted prominently with local 'Rural' COWZ classes during the same time period. However, during weekends, when Supergroup 3 customers may not be constrained by workday activities, interactions were prominent within areas most proximal to their residential locations. In particular, the Rural Fringe Commuters (Group 3a), demonstrated activities within the same rural area types as Supergroup 1. Conversely, Group 3c ('Stable Urban Workers'), the youngest demographic group, showed increased activity within leisure/tourist oriented retail centres such as 'Eat, drink, shop and be merry', during weekend periods.

Other general observations from these data were that firstly, there was an overall higher propensity to patronise more prominent retail destinations during weekends. However, the characteristics of these areas varied between customer types (i.e. small towns, retail parks or city centres, depending on the customer Group). Secondly, there was an evident preference for customers of an older demographic to patronise more rural locations (i.e. market towns), and those of a younger demographic to patronise urban locations (i.e. city centre flagships). These trends demonstrate how the temporal analyses facilitated by loyalty card data may provide a useful source of recording variations in population mobility and how flows of socially distinct groups may be organised during various temporal intervals.

In addition to this, analysis of monthly fluctuations revealed that location visiting behaviours may also vary between seasonal periods. Figure 7.5 demonstrates monthly activity patterns by customer Supergroup (weighted by total volumes per month and per Supergroup). Overall variation in activity was minimal, however, there were evident differences between customer types. For example, 'Rural Ageing Off-peak Shoppers' demonstrated a decline in activity during the Winter-Spring months (yet a peak in November), 'Large Destination Shoppers' showed an increase during the Christmas period and 'Weekday Convenience Commuters' showed the most variable behaviour, demonstrating peaks in February and October, followed by a decline during the winter months/Christmas period.

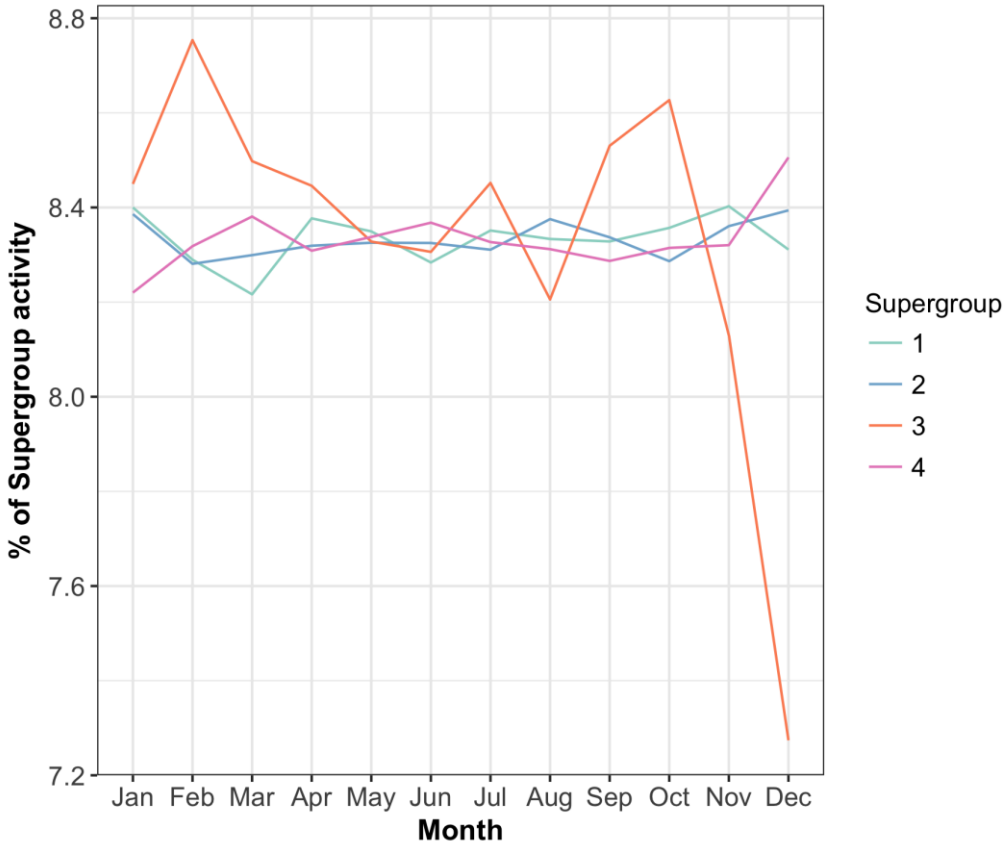


Figure 7.5: Monthly activity volumes by customer Supergroup.

Yet, examining area-visiting characteristics over monthly periods revealed distinctive variations in behaviour. Figure 7.6 shows monthly variations in activity volumes per COWZ area type, at the customer Supergroup level. Figure 7.7 shows an example of variations at the customer Group level, for customer Supergroup 4 (variations within other Supergroups are supplied in Appendix 5). Data were weighted by total activity per COWZ group, per month and transformed to percentages of totals per customer Supergroup, to extract unique fluctuations.

The most prominent trend across the majority of Supergroups/Groups was that variation in behaviour was identifiable within the Christmas period (October to December), spring (April) and summer (July to September). In retail terms, it is anticipated that the Christmas period would see increased levels of consumption. However, there were distinct patterns between the areas visited by different customer types during this period. For 'Rural Ageing Off-peak Shoppers', the most rurally located customers, increases in activity were visible within the most proximal major retail centres, for example, urban/suburban centres and retail parks. Activity declined in their usual, 'Rural' destination types during this time. 'Weekday Convenience Commuters' also demonstrated lower activity within their usual destination types during this period ('Top Job' areas), and an increase across a variety of alternative area types (primarily suburban). Both 'Small Destination Shoppers' and 'Large Destination Shoppers' demonstrated an increase in activity within their usual destinations during this time.

These trends indicated that, firstly, more rurally located customers may be motivated to travel further to larger retail areas during the Christmas period. Secondly, the activities of the urban commuter community became more prominent within non-urban destinations. Taking into account the previous weekday versus weekend area-visiting behaviours of this Supergroup, this is likely a reflection of increased destination shopping on weekends, which would likely take place within the area types accessible to their residential locations, rather than workplace locations. Thirdly, for those that already demonstrate their highest usual activity in 'destination' type areas (i.e. retail parks, or town centres) activity increases within these same locations. Thus, the uniform trend, across all Supergroups, is that an increase in activity occurs within centres with a larger retail offering during the Christmas period. For Supergroups 2 and 4, this is discernable as an increase in their usual area-visiting activity. For Supergroups 1 and 3, whose usual store locations likely offer a more limited range, both in terms of HSR store sizes and the local retail offering as a whole (i.e. Supergroup 1 – local, rural stores, Supergroup 3 – urban, convenience stores), a change in area-visiting behaviours can be identified.

A second identifiable trend from these data was fluctuations within summer months, primarily July to September. This was present to some extent across all Supergroups, however, most prominent for Supergroups 2 and 4. Patterns showed an increase in activity within their likely residential areas and a decrease in likely workplace areas during these months, similar to the changes in area visiting that are observed between weekdays and weekends. For example, Supergroup 2 showed an increase in activity within 'Rural' area types, and a decrease in more urban/suburban destinations. For Supergroup 4, the same patterns were evident relative to their more urban fringe oriented residential locations.

It is speculated that these behavioural fluctuations could be influenced by contextual changes that occur during these months, primarily educational term times in England and Wales. These typically occur between July and September, and may influence the mobility characteristics of certain segments of the population. For example, those employed in educational occupations would not be working during this period, parents may be more likely to take holiday leave from occupations, and urban living students often migrate back to their destinations of origin. These dynamics may also be correlated with fluctuations evident in April (which represents the UK Easter holiday period), during which a significant decline was evident within 'Large Scale Education' centres at the COWZ Group level. 'Rural' area types also saw a general increase during summer periods across many Supergroups/Groups, which could also be a result of UK tourism activity patterns. Further seasonal distinctions within each Supergroup were discernable at the Group level (see Figure 7.7). This demonstrated how the patterns observed thus far occur to differing extents between more distinct segments of customers.

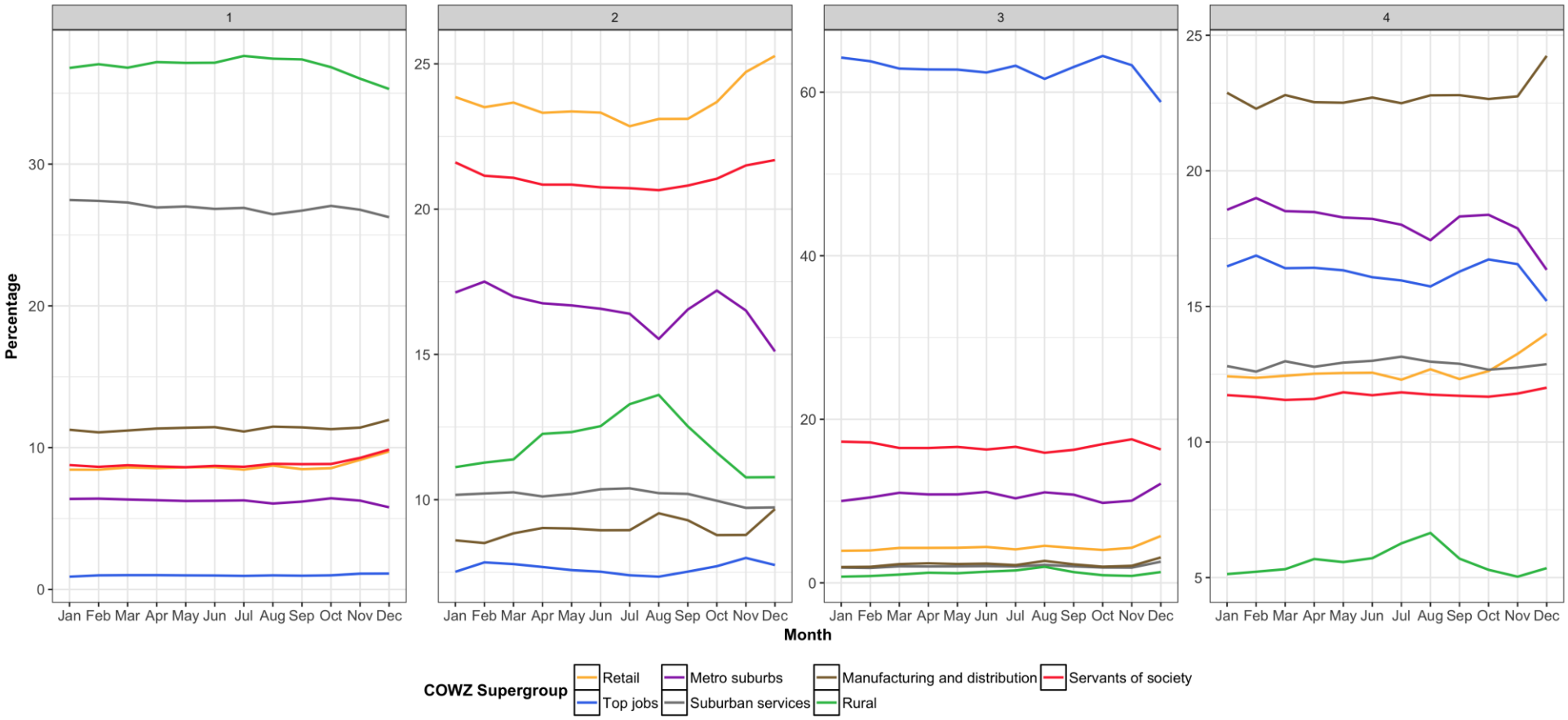


Figure 7.6: Monthly variation in area-visiting activity per COWZ Supergroup, by customer Supergroup. (Note: Scales are varying to illustrate fluctuations within each Supergroup).

Supergroup 4 – 'Large Destination Shoppers'

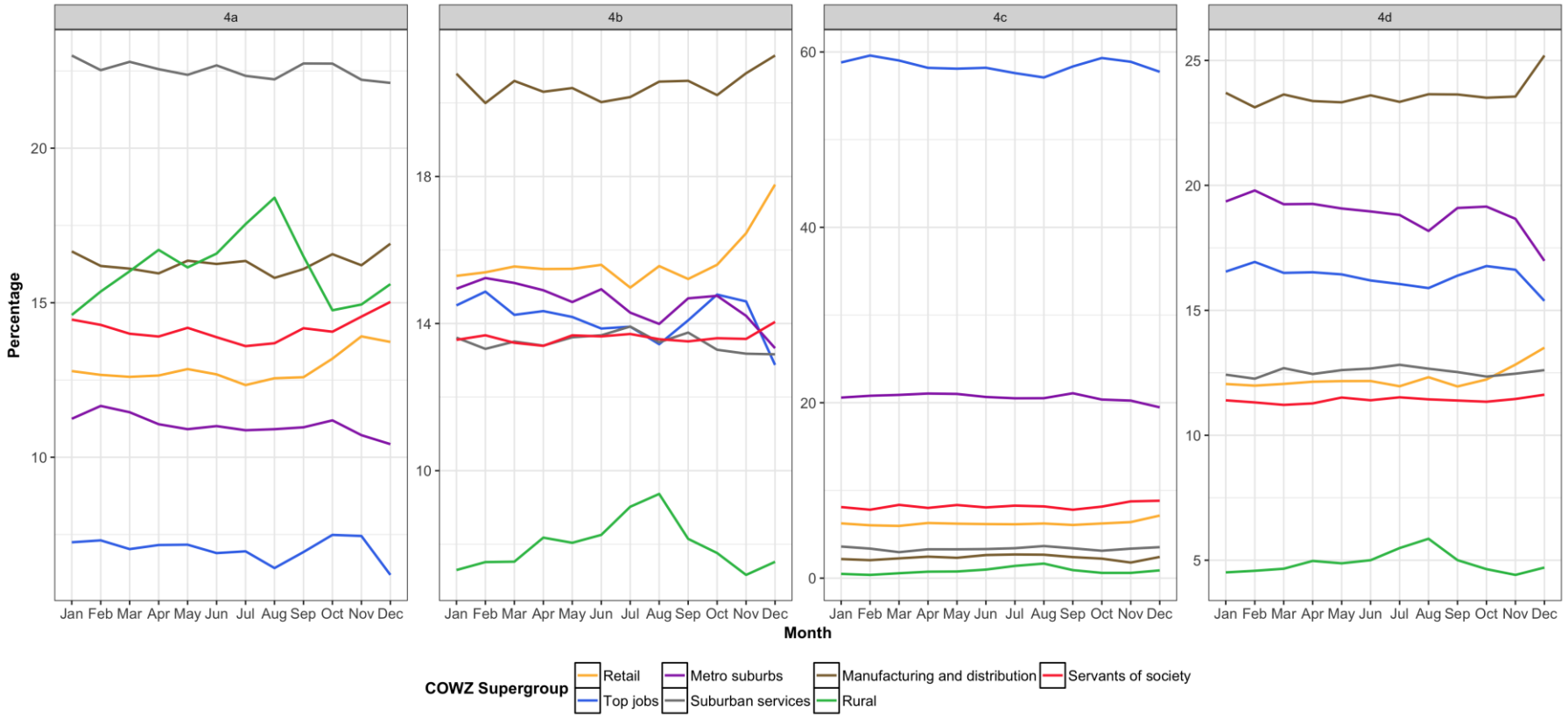


Figure 7.7: Monthly variation in area-visiting activity per COWZ Supergroup, for Groups 4a ('Rural Fringe, Weekday Destination Shoppers'), 4b ('Urban Fringe, Weekday Destination Shoppers'), 4c ('Urban Weekday Destination Shoppers') and 4d ('Stable Urban Destination Shoppers'). Note: Scales are varying to illustrate fluctuations within each Group.

These trends indicate that, whilst distinct behaviours can be identified between weekday and weekend behaviours throughout a yearly period, there may be seasonal periods in which these behaviours change due to contextual influences, such as educational term times. Across many Supergroups, usual area-visiting activity patterns indicated opposite trends to a usual working week. However, these findings also demonstrate that these fluctuations, in addition to the changes evident during Christmas periods, vary between customer Supergroups and Groups. The identification of these trends demonstrated further relationships between the distinct temporal rhythms of different customer types, which provided an enriched description of their daytime activity dynamics.

7.4. Regional Variation

The final analysis aimed to investigate the extent to which the trends observed thus far occurred consistently, within customer Supergroups and Groups, across different regions. Figure 7.8 presents firstly, overall levels of activity recorded within COWZ Supergroups (Figure 7.8a) and Groups (Figure 7.8b), within each region across England and Wales. Figure 7.9 then demonstrates differences in area visiting by COWZ Supergroups (Figure 7.9a) and Groups (Figure 7.9b). To extract unique trends, data were weighted by both overall activity volumes per region, and per COWZ destination.

As is evident from Figure 7.8, overall volumes of interaction with COWZ area types were not consistent across regions, with many regions exhibiting minimal activity within area types that were dominant elsewhere. For example, customers in the North East demonstrated higher activity within ‘Retail’ areas, most prominently within ‘Low density retail and wholesale’ and ‘Shop until you drop’. Higher interactions with ‘Rural’ types were evident within the North West, South West and Yorkshire and The Humber. Within the North West, this was predominantly within the ‘Rural with mining or quarrying’ Group. Interactions with ‘Manufacturing and distribution’ areas dominated the East Midlands, and also the West Midlands and Wales. This was primarily within ‘Business parks’ in the East Midlands, ‘Industrial units’ In the West Midlands and ‘Mining and quarrying facilities’ in Wales. Within London and the East, activity was most prominent within ‘Top Job’ areas, primarily ‘Science and business parks’ in the East, and an even distribution between ‘Global business’, ‘Administrative centres’ and ‘Big city life’ in London. Activity in ‘Metro suburbs’ was also most prominent in London (in addition to the East, and the South West). The largest proportion of activity recorded in Wales took place within ‘Servants of society’ areas, predominantly within the ‘Public administration’ and ‘Highly qualified workforces and professional services’ groups.

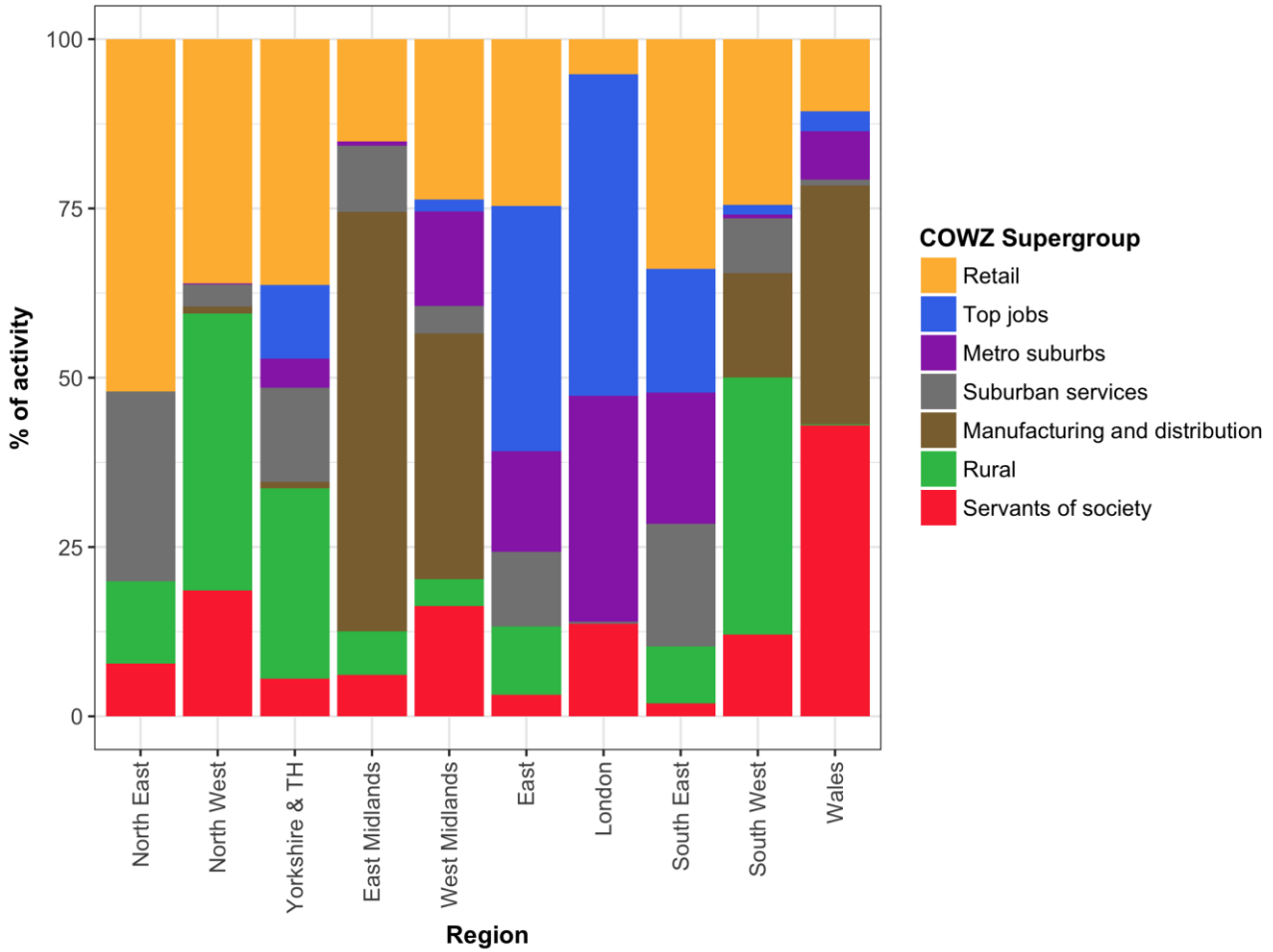


Figure 7.8a: The percentage of overall activity per COWZ Supergroup, by region.

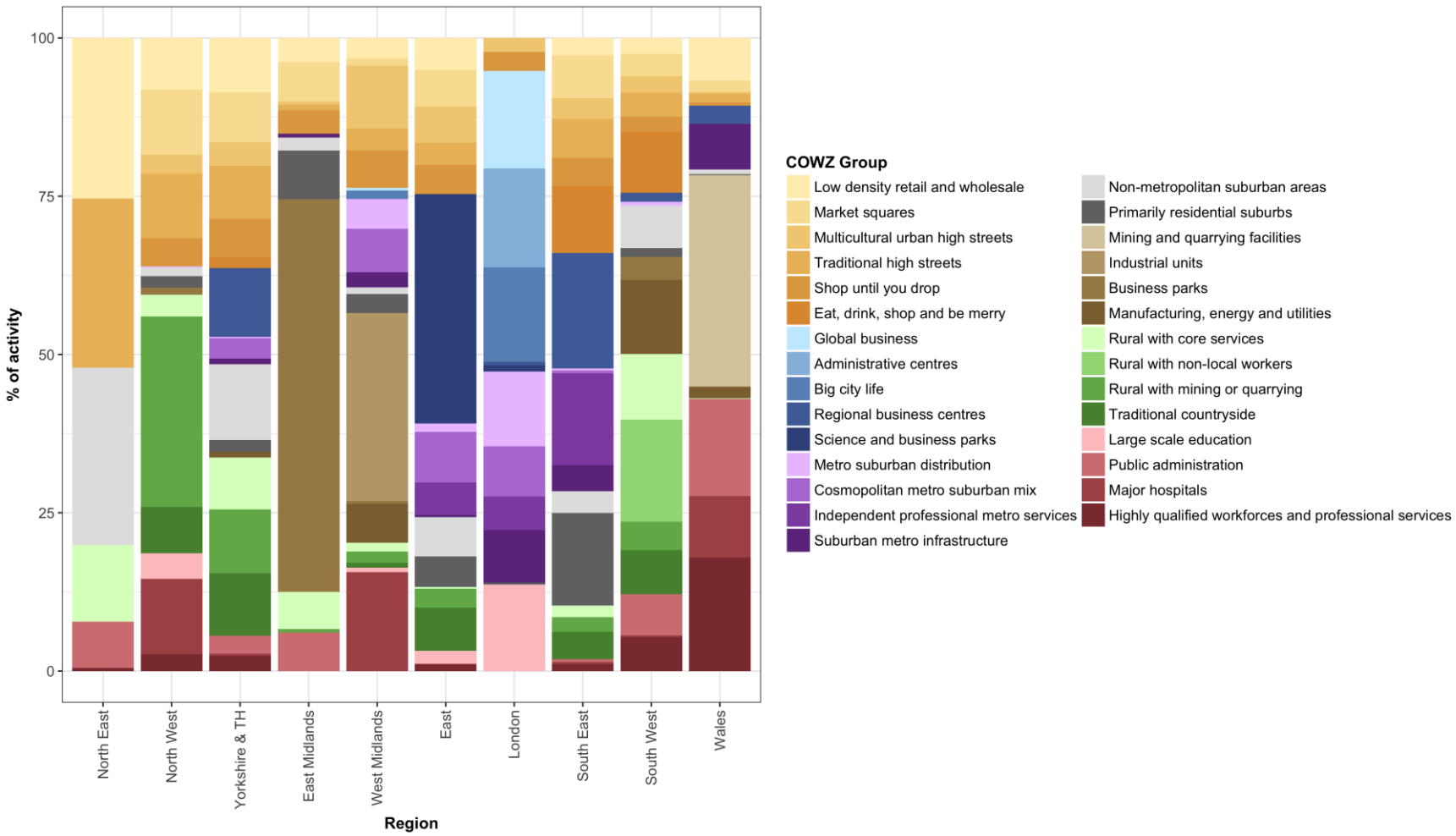


Figure 7.8b. The percentage of overall activity per COWZ Group, by region.

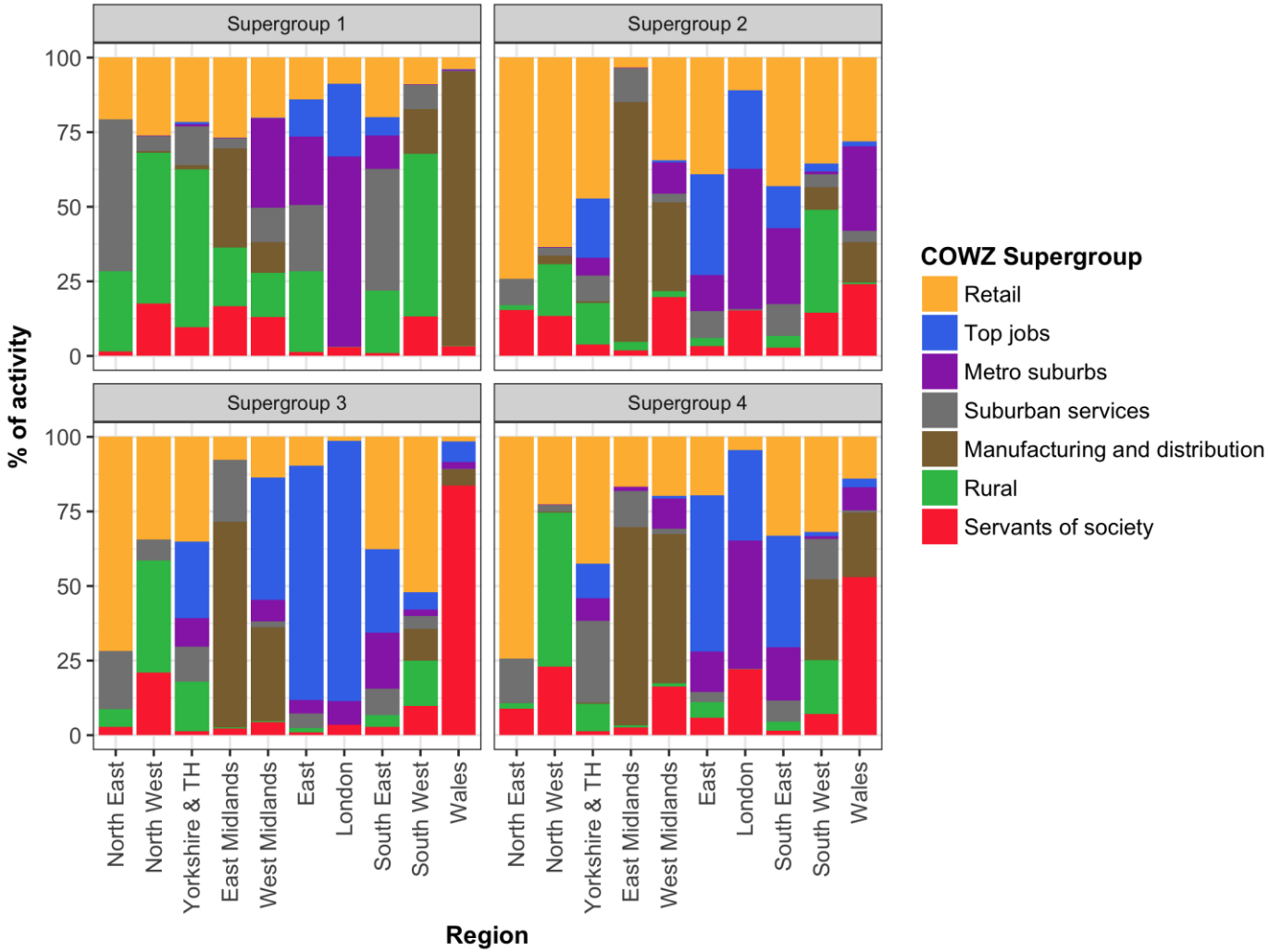


Figure 7.9a. Regional variation in location visiting behaviour – Supergroup level.

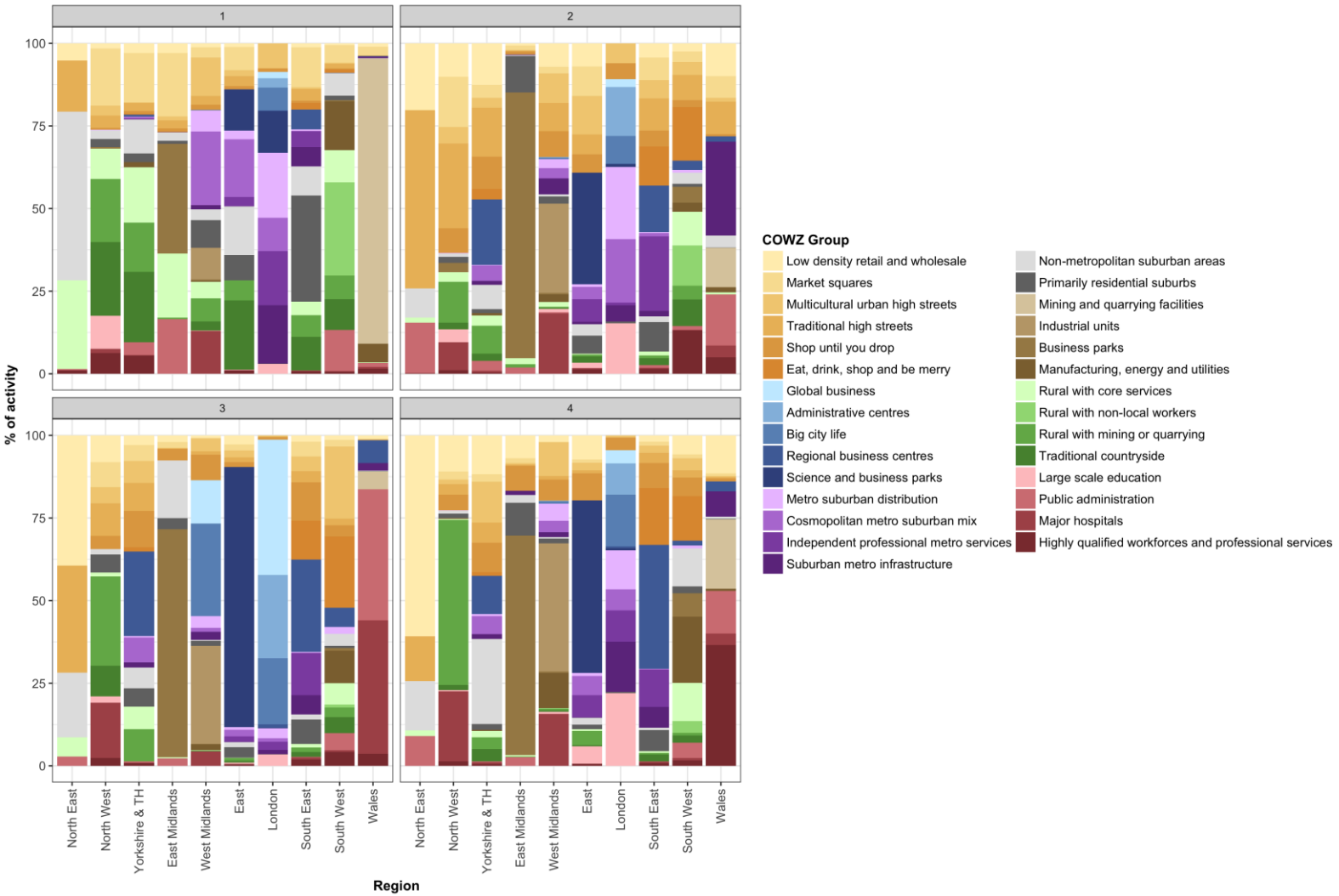


Figure 7.9b. Regional variation in location visiting behaviour – Group level.

These trends suggested that firstly, overall levels of activity within different retail centre types may be biased by the geographical and occupational characteristics of different regions. For example, the North East comprises of the most metropolitan areas in the North (i.e. in the counties of Tyne and Wear and Teeside) and the East Midlands and Wales both comprise of many industrial towns. London is dominated by a high density of urban business centres, and the East and South East both consist of outer suburbs of metropolitan London and also business centres (i.e. 'Science and business Parks' and 'Regional business centres') located in close proximity to the capital, for example, within Surrey and Essex. Business centres within Wales are also more oriented towards 'Servants of society' occupations, with the majority of high status occupations operating from London.

Analysis of Supergroup location visiting characteristics between regions demonstrated large variations in behaviour. These variations suggested strong relationships with the different types of areas that may be accessible to customers within each region, as delineated by Figures 7.9a and 7.9b. Despite this, Supergroups within the same region still demonstrated distinct patterns. The most prominent examples are differences in behaviour within the same region between Supergroup 1 and 3. For instance, Supergroup 1 customers in Wales demonstrated 91.2% of their activity within the 'Mining and quarrying facilities' group. In contrast, Supergroup 3 customers demonstrated 83.7% of their total activity within 'Servants of Society'– primarily the 'Major hospitals' and 'Public administration' groups. In London, the largest proportion of Supergroup 1's activity took place within metropolitan suburbs, whereas 87.3% within 'Top job' areas for Supergroup 3 customers.

The primary observation from these trends was that whilst customers within the same Supergroup show differences in behaviour by region, this is likely due to the types of location that dominate the surrounding areas. General behaviours within each Supergroup were still consistent with the characteristics observed thus far. For example, in each region, Supergroup 1 showed the least interaction with core urban areas and a preference for the more rural/ less metropolitan areas, relative to each region (e.g. 'Manufacturing and distribution' in Wales, 'Metropolitan suburbs' in London). Supergroup 3 customers showed a high preference for the urban centres, relative to each region (i.e. 'Top jobs' in London, 'Retail' in the North East and 'Servants of Society' in Wales). Across all regions, Supergroups 2 and 4 demonstrated a preference for area types that were consistent (i.e. within close proximity) with their small town (Supergroup 2) versus urban fringe (Supergroup 4) locations.

These analyses highlighted that whilst customer segments demonstrated similarities in temporal consumption patterns, demographic attributes, geographic distributions, product consumption habits and area-visiting preferences, there may be differences in the types of retail centre they

interact with across different geographical regions. This is likely an effect of the types of area that are accessible to them and may influence the extent to which customers in different regions are exposed to different location types. Thus, whilst preferences may be evident between distinct groups of customers, the area types that fulfill these preferences may vary between regions.

7.5. Discussion

These analyses provided data-driven evidence for the daytime activity patterns of socially distinct consumer groups, through integration with the Census based COWZ classification. Findings suggested that quantifiable differences exist between the types of retail centres that these groups visit, interact with and are exposed to. In addition, the proximity, and thus accessibility of area types may also be influenced by regional characteristics. For example, customers within different regions may be inherently subjected to area characteristics as a result of regional differences in geographic composition and the types of occupation that dominate the local areas. This appeared to play a key role in the extent to which customers assigned to the same Supergroup or Group interacted with location types. Despite this, distinct preferences within each segment were still evident (such as customers of an older demographic being active in the most rural locations, and vice versa). However, the area types that fulfil these preferences may vary between regions.

Whilst many observed trends might be anticipated, these insights serve to further enforce the potential benefits of using novel consumer datasets to extract trends that are not obtainable via traditional methods. For example, they demonstrate how expected relationships can be drawn between demographic characteristics evident in loyalty card data, with that of Census based classifications. Yet, loyalty card data may provide enriched descriptions of geodemographic phenomena not provided in traditional sources, such as daytime activity patterns. These findings therefore have promising implications for enriching current geodemographic representations through the creation of bespoke consumption indicators, which are not based solely on residential geographies. This would be of relevance to a wide range of users who continue to utilise classifications based on residential based activity, including academia, government and commercial organisations.

These outputs have a number of implications for the exploitation of locally available time series data from alternative commercial sources. For instance, in combination with outputs from previous chapters, detailed descriptions of likely activity patterns are provided for each distinct customer segment, regarding both the location types that customers may be most likely to visit and also the time periods in which they are most likely to be active. From a retail perspective, these insights could inform the most likely times and locations in which distinct customer types

may be most engaged. Analysis of how their product consumption varies during different time periods could inform both store logistics and personally tailored offers that are both spatially and temporally relevant. Group 3a ('Rural Fringe Commuters') provide a notable example – where customers may be engaged with convenient essential items, or food and drink, between business hours within urban centres during weekdays, yet, healthcare products within rural area types during weekends.

Furthermore, incorporating trends over seasonal periods offers further insights into customer types; for instance, offers aimed at students based on their overall activity patterns may be redundant outside of term times. To enrich understanding of the times in which different groups of customers may be most receptive to engagement, further analysis could aim to identify time periods in which the highest spend, or volume of transactions is generated for each individual. This would also offer insights into the area types in which customers may be most effectively incentivised to visit during their less active periods. Finally, evidence regarding the temporal rhythms of socially distinct groups may aid understanding of the evolving convenience market, such as how differing daily obligations may influence consumption habits and thus how 'convenience' behaviour may be defined differently for customers who are subject to varying temporal constraints. Temporal differences in activities between customer types also allowed for differentiation between those engaging with the same area types. For example, Group 1a and Group 3a customers may both be active in rural location types during weekends, however, Supergroup 1 customers demonstrated a preference for early mornings, whilst Group 3a customers demonstrated activity later in the day. These trends not only have implications for retailers, but also serve to enforce the previously outlined implications for addressing UK high street resilience.

It is finally important to acknowledge that whilst these insights can be applied to the HSR population, caution should be given when extrapolating insights to the GB population. For example, potential over-indexing of trends could have occurred in the seasonal analysis if this sample contained a higher proportion of individuals employed in educational occupations (i.e. teachers) or those with a particular family/household structure (i.e. parents). Thus, attempts should be made to integrate alternative big datasets, with differing population representations, to understand how consistent these patterns may be. This would also be an important development for gaining a wider understanding of high street resilience factors. Yet, access to these data facilitated preliminary investigation of their potential for informing dynamic population activities, demonstrating their prospective uses for addressing current gaps in knowledge as a result of the data-scarce era.

8. Discussion, Applications and Research Prospects

8.1. Introduction

Endeavours in population studies to date have been limited by a lack of data by which human behaviours can be studied, which has inevitably restricted our understanding of the complex dimensions that encompass human identity. The aim of this thesis has been to explore the potential of a novel consumer dataset to enrich our understanding of population processes, whilst also comprehending their inherent limitations. Preliminary appraisal of the dynamics of these data confirmed many anecdotal beliefs that data quality and issues of uncertainty are evident across a number of dimensions. This highlighted that whilst many of these are identifiable through standard data cleaning practices, many will not be, and temporal errors (where attributes may change between the time of collection and time of data usage) are substantially harder to isolate. This may be particularly troublesome in the context of customer address attributes, which are key to utilising such data in this context. However, it was also demonstrated how address errors identified here were not random, and we can therefore make attempts to control for them. In addition, a large proportion of address attributes in these data are likely to be correct.

Although the identification of uncertainty emphasises negative aspects of these data, the proceeding chapters then sought to remedy this through exploration of their use for population insight. The endeavours of Chapters 5 to 7 were justified by the various limitations of population studies outlined in Chapter 2. Most notably, how conceptualising people and places in static spatial terms may disregard the important role that temporal dynamics can play in the functioning and organisation of societies. The analyses here provide evidence that loyalty card data facilitate opportunities to enhance human geographical understanding. Outputs suggested that exploiting the inherent velocity of novel consumer datasets may enrich our understanding of exchanges between people and consumption spaces, in addition to how these dynamics vary among distinct social groups, and over various temporal intervals. Combined, these insights endorse that there may be alternative geodemographic dimensions over which individuals can be represented, and that these kinds of insights may now be obtainable through the emergence of novel big data sources.

The proceeding chapter seeks to consolidate the contributions of this thesis by firstly, reflecting

on the methods applied and acknowledging important limitations. Following this, the applications and implications of this work are discussed in the context of loyalty card data, but also more broadly in terms of integrating novel datasets in social science research. Applications are also discussed according to various agendas such as those of academia, governments and retailers. Discussions are concluded by identifying paths for future research.

8.2. Reflection on Methods

Due to the infancy of this area of research, methods applied throughout this thesis were largely exploratory in nature and in many instances, it was necessary to develop and employ novel heuristics. The most notable context in which novel heuristics were required was for the quantification of address errors presented in Chapter 4. This represented an arguably ad-hoc approach to resolving this issue, however, was deemed the most practical solution in the absence of viable alternatives. A number of pragmatic steps were administered to create this method, of which required a vast number of considerations. This included the formalisation of existing knowledge across the domains of travel behaviour, human mobility and retail catchment areas, considerations of spatial scale, definition of thresholds, and creation of an algorithm that could detect patterns in this dataset. Each stage required subjective decisions in regards to what might produce the best outputs and various options were tested through the process of iterative implementation and truth propagation.

As outlined in Chapter 4, it is inevitable that this method could be improved in a number of ways. On reflection, it is likely that incorporating a temporal dimension into these analyses may have improved its efficiency. For example, the subsequent chapters demonstrated that differentiations between weekday and weekend activities were important in delineating behavioural characteristics. Thus, integrating this concept would likely provide a more detailed understanding of what constitutes normative behaviour for different areas, during different time periods. Despite this, the construction of this method provides evidence of how we can apply Miller and Goodchild (2015)'s *knowledge solution* in order to attempt to clean spurious patterns in these data. It also supports the notion of reconfiguring traditional scientific method to provide valuable insights that 'knowledge-driven science' would fail to generate (see Chapter 2, Section 2.1.2). For example, this process was abductive in nature, yet utilised an inductive approach by integrating guidance from existing theory, which was then deductively validated. This produced insights that were not practically obtainable via traditional methods.

The second area of analysis that required methodological considerations was the classification of spatiotemporal trends. For the classification of stores, only temporal frequencies were used to segment data. It should be noted that classifying these data using a broader number of variables (such as integrated with built environment or demographic characteristics) would evidently

produce more descriptive outputs. However, the aim of this analysis was not to produce comprehensive classification outputs, but rather to investigate the utility of incorporating time into conventional, static representations of places, given its limited appraisal in both public and commercial settings. Future endeavours could aim to create more comprehensive classifications by integrating temporal elements with static spatial measures.

It should be further noted that elements of temporal aggregation were utilised, of which can be criticised for its potential to simplify key trends. However, this was a necessary measure in order to make sense of trends in these data, and the scale of aggregation utilised was considered sufficient enough to extract general trends in line with the broad aims of this work. That being said, future endeavours, with more specific and less global focus, may benefit from analysis over less aggregate temporal intervals. In relation to this are the methodological limitations imposed by assigning generalised profiles to large numbers of individuals, potentially giving rise to ecological fallacy. This represents a commonly recognised issue when utilising classifications, however, similarly, as the aims of this work were to identify the types of population insight that we may be able to extract from loyalty cards, attempts had to be made to summarise complex patterns in order to understand the general trends in these data.

Methods utilised in other areas of this analysis were largely descriptive in nature. For example, trip distribution matrices provided one of the most useful measures for extracting activity patterns in this context. This method was utilised to quantify interactions between MSOAs and stores locations in the identification of address uncertainty (Chapter 4), between customers and store types in the customer classification (Chapter 6) and between customer types and COWZ area types in the analysis of location-visiting dynamics (Chapter 7). This is a relatively basic measure, involving obtaining frequencies of interactions between origins and destinations. However, as noted in Chapter 2, there is a need for more research that can apply descriptive data-driven methods to provide robust measures of activities such as counts, distance and time from large samples. This work demonstrates an example of applying this and how descriptive measures can provide a valuable means of summarising complex interactions in large datasets.

8.3. Limitations

This thesis presents a broad range of insights, not only in terms of who loyalty card data best represent, but also with regards to the types of population insight we may extract. Despite this, as with any analysis, this work has limitations. In this case, limitations are primarily a result of the inherent, uncertain nature of these data, but also the availability of suitable reference data.

There are various areas in this thesis whereby uncertainty was a factor. This includes, firstly, issues of data quality, which as outlined in Chapter 3 are not always quantifiable. Examples of

this include entry errors, where values are admissible but incorrect, processing and assignment errors, of which will remain undetected due to not knowing the processing these data have undergone, and also temporal errors, such as are evident in loyalty card address attributes. Therefore, whilst in each case careful measures were applied in an attempt to account for error, uncertainty is inherent in these data and there are inevitably instances where identification is not possible.

Further uncertainty is apparent when utilising the metadata provided by the HSR to interpret analysis outcomes. For example, the product category classification represented aggregate groups of products, of which the original composition was unknown. Thus, there is uncertainty surrounding the precise products belonging to each category, and the potentially subjective nature in which they were assigned. Whilst insights derived from these data followed expected trends and allowed interpretation of generalised behaviours, this factor ultimately limits the extent to which we can make inferences about people's product consumption dynamics. In addition, as noted in Chapter 3, many of these definitions were unclear (i.e. 'Miscellaneous') and therefore were not utilised. Similarly, the provenance of the HSR store type classification was largely unknown. Whilst these were useful for interpreting the characteristics of store clusters in Chapter 5, the variables and processes involved in creating these were not provided.

Issues of representativeness are also prominent in loyalty card data. Whilst biases in data are both recognised and understood in traditional social science practices (i.e. with protocols in place to measure and address the issues), there are currently no formalised frameworks for their quantification in this context. Attempts were made here to understand the inherent bias of these data, largely through linkage to existing national statistics. Yet, caution should be given when extrapolating these outputs to the behaviours of the general GB population. A prominent example where this may influence outcomes was in the analysis of location-visiting (Chapter 7), where prominent educational term time fluctuations might be attributed to an over-representation of those with a particular household structure or occupation. This highlights the need for future research to replicate such analyses using novel datasets that offer differing representations, in order to understand if such trends are consistent. Furthermore, the need to reduce the customer sample to only the most 'active' customers for a number of analyses may intensify the effects of representation, particularly in terms of affluence (i.e. these are generally higher spending customers).

Further limitations of these data arise from the uneven distribution of data across individuals and stores as a result of differing motivations to participate. This ultimately means that not all customers and store locations are equally represented by these data in terms of the true consumption that occurs in each case (i.e. transactions without a card may still occur). Whilst attempts were made to eliminate customers with minimal activity through active customer

selection, this also increased the number of areas for which no data were available. Thus, it can be concluded that whilst the total number of individuals present in loyalty card databases may be voluminous, the volume of those exhibiting quality behavioural data may be much smaller. Furthermore, dynamics such as the inherent nature of the HSR store network created bias in many analyses. For example, when utilising these data for customer profiling (Chapter 6), there were many more of certain store types (i.e. small high street stores and chemists) and subsequently, higher volumes of transactional data for the segments of the population that patronise them. This affected the size of the resulting customer Groups. These issues represent inherent limitations that require consideration when utilising loyalty card data.

It is finally important to note that the data provided by the HSR only represented a portion of the total data they collect. Whilst this still provided a large, detailed and longitudinal source for analysis, there were a number of areas in which access to even more longitudinal records may have improved outputs. For example, as previously noted, the nature of these data meant that individual trajectories are sparse in some cases, and demonstrated relatively high intervals between transactional events. Access to more longitudinal data would increase the number of 'active' customers that may be extracted, which would inevitably enrich the analyses in this thesis. In addition, there were many instances where uncertainties in analysis were potentially a reflection of the sample rather than true customer behaviour. For example, it is highly likely that the trip distributions utilised to identify address uncertainty (Chapter 4) would be greatly enriched by the provision of more longitudinal data. In addition, this may have shed light on uncertain behavioural trends observed in the customer classification (see Chapter 6, Section 6.2.3). Despite this, access to this sample provided a means to test the implementation of novel heuristics and provided a sample size much larger than has previously been achievable.

From a wider perspective, it is important to remember the fundamental limitations of utilising only one retailer's loyalty card data, such as that they do not represent full individual consumption characteristics. Thus, the outputs presented in this thesis only represent one dimension of these individuals' consumer profiles. This further highlights the need to support the insights produced here with evidence from alternative data sources. In addition, there are a number of dimensions in which issues of uncertainty cannot be ascertained through data-driven measures. For example, as outlined in Chapter 2, whilst research indicates behavioural differences between loyalty card members and non-members, quantifying this using the data alone is fraught with difficulties, and may require incorporating qualitative studies if aiming to gain an exhaustive description of scheme member characteristics.

In summary, this thesis highlights important considerations for the adoption of such data as indicators of social and spatial phenomena, primarily that analyses will be dependent on the data available, which may limit the scope of insights than can be derived.

8.4. Applications and Implications

Despite the acknowledged limitations, this thesis provided many valuable insights for the integration of loyalty card data in social science research. The first notable applications relate to data quality and representation challenges. There has been a fundamental lack of understanding of these issues in regards to loyalty card research and also consumer data more widely. This work provides data-driven evidence for these properties, and provides a valuable point of reference for those wishing to employ such data for insights regarding the general population. It is hoped that this work can progress future research by outlining pragmatic steps and the necessary considerations in order to reliably integrate these data. However, these insights are not only relevant for this cause, but also for understanding the potential uses of different types of consumer data to social problems. For example, if aiming to utilise loyalty card data to inform particular phenomena, it is critical to understand who the data are representative of. If the individuals of interest are male or of low socioeconomic status for example, then the target population may be largely missed in these data. This extends the importance of ascertaining representativeness beyond performing robust analysis to their wider uses as a source of population data.

This prospect highlights an important need to compile formal frameworks that benchmark and outline the applications of various forms of consumer data. In this thesis, it is hoped that this has been established for the provenance of loyalty card data. The provision of such information in an accessible and standardised format may provide both academics and industry with a consistent point of reference whereby the aforementioned biases may be recognised and addressed. These endeavours would ultimately allow us to build a more wide-ranging representation of populations from novel data sources. The creation of such ‘meta-data’ would be an important step in the transition from traditional population data, as is one of the primary motivations of data collaboratives such as the CDRC.

Succeeding this is the prospect of achieving effective data linkage from these multiple sources. For instance, through an improved understanding of the merits of each data source, and through cooperation with data providers, there are substantial opportunities to incorporate new forms of data in support of existing population datasets. It should also be noted, that, these biases may not necessarily represent a limitation in the context of linking datasets. For example, whilst the focus of this thesis was ascertaining the characteristics of the loyalty population in comparison to the general population, the prospect of data linkage may enable an amalgamation of datasets that provide a more detailed representation of their target populations, that exceed the representations of data produced to capture entire populations using a few select variables (i.e. censuses).

In regards to the specific role in which loyalty card data will play in this, it is clear that considered independently, these data are insufficient to effectively describe the general population. However, the analyses conducted here outline ample ways in which these data provide rich insights regarding their target population. Given the pharmaceutical nature of this retailer, it may be that more intensive analysis of product consumption patterns could be particularly useful for insights regarding population health characteristics. Yet, they also show value over many dimensions regarding the extraction of population activities. For example, shedding light on population flows, spatiotemporal activity patterns and location type visiting characteristics. These insights may facilitate greatly enriched representations of people and places that move away from traditional, residential-based, static measures. This provides further evidence that the linkage of novel datasets may facilitate better representations of the population, over a wider number of dimensions, than conventional measures are currently able to.

Overall, whilst further developments are needed, this work provides positive evidence for the ability to extract population insight from novel consumer datasets. Collating such data offers prospects for filling the data void that occurs during inter-censual periods and ultimately moving away from the 10-yearly census approach, a prospect highlighted by the objectives for the Census 2021 (Stillwell, 2016) through the integration of address-level administrative, commercial and open-data sources. However, key to this integration will be the ability to link data efficiently and accurately, and it is further highlighted here how preliminary data treatment is necessary to ensure the veracity of the data being integrated.

From a more theoretical perspective, this thesis provides ample evidence for the potential benefits of incorporating a temporal dimension into population studies, as is facilitated by novel datasets. The quantification of customer activity patterns over various dimensions illustrates how our current understanding of transactions between human identity and places, as is a core feature of geodemographics, may be greatly enriched by adopting this perspective. For example, it was illustrated here how there are differences in the types of retail centre that socially distinct groups of consumers interact with and are exposed too. These dynamics may vary according to the area types that are accessible to individuals both in terms of proximity and the temporal constraints imposed on different social groups. Substantial differences were evident between different customer types, indicating that these might be important alternative dimensions of human identity, and thus an integral part of understanding geodemographic phenomena. Similarly to the aforementioned applications, linkage between datasets is necessary in order to build a more complete representation of how these dynamics vary between different segments of the population. However, this research provides data-driven evidence for the existence of dynamic relationships between time, consumers, retail centres and identity factors, of which

traditional geodemographic representations fail to capture. As noted throughout, these types of analyses may also have broad implications for understanding high street resilience, which would be of interest to both public bodies endeavouring to revitalise high street economies, and retailers endeavouring to increase engagement with physical stores.

Whilst the primary focus of this work was the generation of social science insights, there were also prominent applications for retailers. Firstly, the identification of uncertain postcodes in these data has implications for consumer data collectors and users who are operating reliant on consumers keeping up to date address records. Results provided insight into the extent of customers who no longer live at their stated address, and will therefore no longer be correctly identified for location-based targeting efforts. In addition, this work demonstrated how these errors are not random, and thus attempts can be made to identify the customers most at risk of these uncertainties in order to mitigate negative effects. Secondly, the classification of both HSR stores and customers by time suggested many benefits of integrating a temporal dimension into commercial analyses. For instance, it was demonstrated how this may greatly enrich descriptions of store locations in terms of how and why their consumers utilise those consumption spaces during different time periods.

In addition, insights from the customer classification had numerous implications for optimised targeted marketing that could be both spatially and temporally relevant to individuals. This demonstrated how such analyses can aid identification of the most likely times and locations in which distinct customer types may be most engaged, which could inform both store logistics and personally tailored offers. The evidence regarding the temporal rhythms of socially distinct groups may also aid understanding of the evolving convenience market, such as how differing daily obligations may influence consumption habits and thus how ‘convenience’ behaviour may be defined differently for customers who are subject to varying temporal constraints. The creation of these profiles at the postcode level also provides a means of classifying customers with minimal data (for example, new or unengaged individuals) based on their residential information in order to identify potential paths of engagement.

Broader implications of this work relate to current perspectives on access, privacy and ethics in this novel data era. For many of the applications outlined above to be enabled, there is a fundamental need for organisations to facilitate data access to researchers. It is hoped that this research provides an example of how consumer data can be as much of a public asset as they are commercial, in addition to how they can be utilised on behalf of public and social good in a context that protects both personal and commercial integrity. Benefits for organisations may well arise from such collaborations, for instance, many may not be fully aware of the precise sectors of society that they serve, and deeper insights may be generated through researchers with broader timescales in comparison to commercial analysts, who may be faced with time

and profit optimisation constraints. It is finally important to note that undoubtedly, the prospect of data linkage brings further ethical and disclosure challenges to the fore. Inevitably, the greater information we gather about individuals, the greater the risk of personal identification. Therefore, in considering future directions we must also remain aware of the risks that may occur inadvertently through the analysis and linkage of new data. However, under the condition of careful controls, it is clear that such endeavours could be hugely beneficial for both social science research and retailers alike.

8.5. Future Prospects and Closing Remarks

In many respects, the social sciences are only just beginning to understand the applications of consumer data and their potential for the study of human populations. In reference to the applications outlined above, future research should continue to develop our understanding of the dynamics of alternative datasets through the creation of 'meta-data' and progress towards data linkage in order to build a wide-ranging representation of populations from novel data sources. This would also facilitate progression of methods regarding how to handle and analyse data with various attributes. For example, the method applied here to detect uncertainty in addresses is limited in its applicability outside of this specific context. Yet, exploration in alternative settings could facilitate a greater understanding of how uncertainties can be managed and ultimately move towards the creation of more universally applicable methods. Obviously, such progressions are reliant on commercial bodies opening up their data to such causes.

In the context of loyalty card data, a notable area of future research relates to the largely exploratory nature of this thesis. For example, the aims here were focused on addressing preliminary issues of implementation, in addition to understanding what they may contribute to social science endeavours more broadly. This approach was necessary to some extent, given the infancy of this area of research and lack of prior knowledge in this particular context. However, using these foundations as a benchmark, future work should aim to explore the data-driven applications of loyalty card data in more specific contexts. For instance, it may be useful to develop case studies in which these data are applied to particular problems at a less global scale, such as with a specific social/public issue in mind or a specific segment of the loyalty population. This would inevitably highlight more concrete scenarios in which these data may provide social value. It is hoped that this work provides the basis for understanding what types of insight may be applicable.

In closing, the use of new forms of data for the generation of population insight are still very much in their infancy. Whilst the potential of these data is evident, further rigorous assessments are needed regarding the applications and limitations of the various forms of data being generated, and caution is necessary when aiming to make inferences about the population at

large. It is also critically important that analyses of this nature endeavour to achieve outputs that are both informative and safe, especially where data linkage is concerned. Clearly, loyalty card data are by no means a viable substitute for conventional population survey data in isolation, yet, such research agendas do present new ways in which we can investigate social and spatial processes and highlight promising applications for the use of large consumer datasets in social science research.

9. References

Adey, P. (2010). *Aerial life: Spaces, mobilities, affects*. John Wiley & Sons.

Adler, T., & Ben-Akiva, M. (1979). A theoretical and empirical model of trip chaining behavior. *Transportation Research Part B: Methodological*, *13*, 243-257.

Aitchison, J. & Greenacre, M. (2002). Biplots for compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 51*, 375-392.

Aitchison, J. (1986). *The statistical analysis of compositional data.* Chapman and Hall, London

Aitchison, J. (2008). The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. *CODAWORK'08, Girona, Spain: Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona*. Available at: https://bit.ly/2GEhbBm.

Alhadeff-Jones, M. (2016). *Time and the rhythms of emancipatory education: Rethinking the temporal complexity of self and society.* Taylor & Francis.

Alheit, P. (1994). Everyday time and life time: On the problems of healing contradictory experiences of time. *Time & Society*, 3, 305-319.

Allaway, A. W., Berkowitz, D. & D'Souza, G. (2003). Spatial diffusion of a new loyalty program through a retail market. *Journal of Retailing, 79,* 137-151.

Allaway, A. W., Gooner, R. M., Berkowitz, D. & Davis, L. (2006). Deriving and exploring behaviour segments within a retail loyalty card program. *European Journal of Marketing, 40,* 1317-1339.

Anderson, C. (2008) 'The end of theory, will the data deluge makes the scientific method obsolete?', *Edge*, [Online] Available at: https://bit.ly/2q9y3sy.

Anderson, J. (1971). Space-time budgets and activity studies in urban geography and planning. *Environment and Planning A, 3,* 353-368.

Anderson, K. & Kerr, C. (2001). *Customer Relationship Management*. London: McGraw-Hill Education.

Ashby, D. I. & Longley, P. A. (2005). Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS*, *9*, 53-72.

Atkins, K. G., Kumar, A. & Kim, Y. K. (2016). Smart grocery shopper segments. *Journal of International Consumer Marketing*, *28*, 42-53.

Bailey, T. C., & Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. Harlow, Essex, England: Addison Wesley Longman.

Batey, P., & Brown, P. (1995). From human ecology to customer targeting: the evolution of geodemographics. *GIS for business and service planning*, 77-103.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., ... & Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics*, *214*, 481-518.

Behr, M., & Gober, P. (1982). When a residence is not a house: Examining residence-based migration definitions. *The Professional Geographer, 34*, 178-184.

Berry, M. J. A. & Linoff, G. S. (1996) *Data Mining Techniques: For Marketing, Sales, and Customer Support*. Wiley.

Berry, T., Newing, A., Davies, D., & Branch, K. (2016). Using workplace population statistics to understand retail store performance. *The International Review of Retail, Distribution and Consumer Research*, *26*, 375-395.

Birkin, M., Clarke, G., Clarke, M., & Culf, R. (2004). Using spatial models to solve difficult retail location problems. *Applied GIS and spatial analysis*, 35-56.

Birkin, M., & Culf, R. (2001). Optimal distribution strategies. In *Regional Science in Business* (pp. 223-241). Springer, Berlin, Heidelberg.

Bollier, D., & Firestone, C. M. (2010). *The promise and peril of big data*. Washington, DC: Aspen Institute, Communications and Society Program.

Box-Steffensmeier, J. M., Freeman, J. R., Hitt, M. P., & Pevehouse, J. C. (2014). *Time series analysis for the social sciences*. Cambridge University Press.

Brace, C., & Geoghegan, H. (2011). Human geographies of climate change: Landscape, temporality, and lay knowledges. *Progress in Human Geography*, *35*, 284-302.

Brenner, D. J., Staley, J. T., & Krieg, N. R. (2005). Classification of procaryotic organisms and the concept of bacterial speciation. In *Bergey's Manual of Systematic Bacteriology* (pp. 27-32). Springer, Boston, MA.

Bridson, K., Evans, J. & Hickman, M. (2008). Assessing the relationship between loyalty program attributes, store satisfaction and store loyalty. *Journal of Retailing and consumer Services, 15,* 364-374.

Buliung, R. N., Roorda, M. J., & Remmel, T. K. (2008). Exploring spatial variety in patterns of activity-travel behaviour: initial results from the Toronto Travel-Activity Panel Survey (TTAPS). *Transportation*, *35,* 697.

Burns, L. P. (2014). Geodemographics: *Creating a classification at the level of the individual.* PhD thesis, University of Leeds. [Online]. Avaliable at: https://bit.ly/2NJGhFT

CACI, (2014). *ACORN User Guide*. [Online]. Available at: https://bit.ly/2En4QPU.

CACI, (2018). *Workforce Acorn*. [Online]. Available at: https://bit.ly/2Jj4DkC.

Candela, L., Castelli, D., Pagano, P., (2012). Managing big data through hybrid data infrastructures. *ERCIM News*, *89,* 37–38

Cardot, H., Cenac, P., & Monnez, J.-M. (2012). A fast and recursive algorithm ´ for clustering large datasets with K-medians. *Computational Statistics & Data Analysis*, *56,* 1434–1449.

Carmona, M. (2015). London's local high streets: The problems, potential and complexities of mixed street corridors. *Progress in Planning, 100,* 1-84.

CDRC, (2018). The General Data Protection Regulation & Social Science Research, Consumer Data Research Centre. [Online]. Available at: https://bit.ly/2Nm5UwU.

Chen, C., Gong, Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. *Transp. Res. Part A*, *44*, 830-840.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies*, *68*, 285-299.

Chen, X. & Clark, J. (2016). Measuring space–time access to food retailers: a case of temporal access disparity in Franklin County, Ohio. *The Professional Geographer*, *68*, 175-188.

Clarke, G. (1999). Geodemographics, marketing and retail location. In Pacione, M. (Ed), *Applied Geography*. (577-592). London: Routledge.

Clarke, G., & Clarke, M. (2001). Applied spatial interaction modelling. In *Regional science in business* (pp. 137-157). Springer, Berlin, Heidelberg.

Cockings, S., Harfoot, A., Martin, D., & Hornby, D. (2011). Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 Census output geographies for England and Wales. *Environment and Planning A, 43*, 2399-2418.

Cockings, S., Martin, D., & Harfoot, A. (2015). *A classification of workplace zones for England and Wales (COWZ-EW)*. University of Southampton, Southampton, UK.

Consumer Data Research Centre (2018). *Data Service User Guide, Version 6.0*. [Online]. Available at: https://bit.ly/2IrKt6A.

Conway, J., Eddelbuettel, D., Nishiyama, T., Prayaga, S. K., & Tiffin, N. (2012). *RPostgreSQL: R interface to the PostgreSQL database system*. R package version 0.3-2.

Cortiñas, M., Elorz, M., & Múgica, J. M. (2008). The use of loyalty-cards databases: Differences in regular price and discount sensitivity in the brand choice decision between card and non-card holders. *Journal of Retailing and Consumer Services*, 15, 52-62.

Crang, M. (2001). Temporalised space and motion. *Timespace: Geographies of Temporality Eds., J. May, N. Thrift, (Routledge, London)*, 187-207.

Crouter, A. C., & McHale, S. M. (1993). Temporal rhythms in family life: Seasonal variation in the relation between parental work and family processes. *Developmental Psychology, 29*, 198.

Dalton, C. M. & Thatcher, J. (2015). Inflated granularity: Spatial "Big Data" and geodemographics. *Big Data & Society*, 2.

Darden, W. R. & Ashton, D. (1974). Psychographic profiles of patronage preference groups. *Journal of Retailing, 50*, 99-112.

Data Protection Act (1998). *GOV.UK*, [Online]. Available at: https://bit.ly/2kV9S1T.

Debenham, J. (2001). Understanding geodemographic classification: Creating the building blocks for an extension. [Online] *The School of Geography, University of Leeds Working Paper*. Available at: http://www.geogleeds.ac.uk/wpapers/02-1.pdf.

Delafontaine, M., Neutens, T., Schwanen, T., & Van de Weghe, N. (2011). The impact of opening hours on the equity of individual space–time accessibility. *Computers, environment and urban systems, 35,* 276-288.

Demoulin, N. T., & Zidda, P. (2009). Drivers of customers' adoption and adoption timing of a new loyalty card in the grocery retail market. *Journal of Retailing, 85,* 391-405.

Dennett, A., & Stillwell, J. (2011). A new area classification for understanding internal migration in Britain. *Population Trends, 145,* 146-171.

Diez Roux, A. V., & Mair, C. (2010). Neighborhoods and health. *Annals of the New York Academy of Sciences, 1186,* 125-145.

Digital Economy Act, (2017). *Digital Economy Act 2017.* [Online]. Available at: https://bit.ly/2qjlURv.

Dolega, L., Pavlis, M., & Singleton, A. (2016). Estimating attractiveness, hierarchy and catchment area extents for a national set of retail centre agglomerations. *Journal of Retailing and Consumer Services, 28,* 78-90.

Dorotic, M., Bijmolt, T. H. & Verhoef, P. C. (2012). Loyalty programmes: current knowledge and research directions. *International Journal of Management Reviews, 14,* 217-237.

Dowling, G. R. & Uncles, M. (1997). Do customer loyalty programs really work?. *MIT Sloan Management Review, 38,* 71.

Dramowicz, E. (2005). Retail trade area analysis using the Huff model. *Directions Magazine*. [Online]. Available at: https://bit.ly/2IsKOGs.

Dugmore, K., Furness, P., Leventhal, B., & Moy, C. (2011). Beyond the 2011 census in the United Kingdom: with an international perspective. *International Journal of Market Research, 53,* 619-650.

Dutcher, J. (2014) What Is Big Data? *Data Science Berkeley Blog*. [Online]. Available at http://datascience.berkeley.edu/what-is-big-data/.

Edensor, T. (2012). The rhythms of tourism. In *Real Tourism* (pp. 68-85). Routledge.

Edensor, T. (2016). Introduction: Thinking about rhythm and space. In *Geographies of rhythm* (pp. 13-30). Routledge.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy, 14*, 1-24.

Ellegård, K., & Vilhelmson, B. (2004). Home as a pocket of local order: everyday activities and the friction of distance. *Geografiska Annaler: Series B, Human Geography*, *86*, 281-296.

Elwood, S. & Leszczynski, A. (2011). Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum, 42*, 6-15.

Esayas, S. Y. (2015). The role of anonymisation and pseudonymisation under the EU data privacy rules: beyond the 'all or nothing' approach. [Online]. *European Journal of Law and Technology*. Available at: https://bit.ly/2GUtZa6.

Experian, (2017). *Mosaic: The Consumer Classification Solution for Consistent Cross-channel Marketing*. [Online]. Available at: https://bit.ly/2GzUnmh.

Fagan, C. (2001). The temporal reorganization of employment and the household rhythm of work schedules: The implications for gender and class relations. *American Behavioral Scientist*, *44*, 1199-1212.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review, 1*, 293-314.

Farag, S., Schwanen, T., Dijst, M. & Faber, J. (2007). Shopping Online and/or in-store? A structural equation model of the relationships between e-shopping and in-store shopping. *Transportation Research Part A: Policy and Practice*, *41*, 125-141.

Farber, S., Neutens, T., Miller, H. J. & Li, X. (2013). The social interaction potential of metropolitan regions: A time-geographic measurement approach using joint accessibility. *Annals of the Association of American Geographers, 103*, 483-504.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*, 15.

Firmani, D., Mecella, M., Scannapieco, M., & Batini, C. (2016). On the meaningfulness of "big data quality". *Data Science and Engineering, 1*, 6-20.

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology*, *25*, 435-437.

Gale, C. G., Singleton, A., Bates, A. G., & Longley, P. A. (2016). Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science, 12,* 1-27.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35,* 137-144.

GDPR, (2018). *Guide to the General Data Protection Regulation (GDPR).* Information Commissioners Office. [Online]. Avaliable at: https://bit.ly/2slkeJJ

Godichon-Baggioni, A., Maugis-Rabusseau, C., & Rau, A. (2017). Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. *arXiv preprint arXiv:1704.06150.*

Gómez, B. G., Arranz, A. M. G. & Cillán, J. G. (2012). Drivers of customer likelihood to join grocery retail loyalty programs. An analysis of reward programs and loyalty cards. *Journal of Retailing and Consumer Services, 19,* 492-500.

Gong, H., Chen, C., Bialostozky, E., & Lawson, C. (2011). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems, 36,* 131-139.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature, 453,* 779.

Goodchild, M. F. & Longley, P. A. (1999). The future of GIS and spatial analysis. *Geographical information systems, 1,* 567-580.

Goodchild, M. F. (2013). Prospects for a space–time GIS: Space–time integration in geography and GIScience. *Annals of the Association of American Geographers, 103,* 1072-1077.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. Spatial statistics, 1, 110-120.

Government Statistical Service (2014). *GSS/GSR Disclosure Control Guidance for Microdata Produced from Social Surveys*. [Online]. Available at: https://bit.ly/2En070B.

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography, 3,* 255-261.

Granickas, K. (2013). *Understanding the impact of releasing and re-using open government data*. [Online]. Available at: https://bit.ly/2dGwo8J.

Gren, M. (2001). Time-geography matters. *Timespace: Geographies of temporality,* 208-225.

Guptill, S. C., & Morrison, J. L. (Eds.). (2013). *Elements of spatial data quality*. Amsterdam: Elsevier.

Guy, C. M. (1998). Classifications of retail stores and shopping centres: some methodological issues. *GeoJournal*, *45*, 255-264.

Guy, C. M. (1991). Spatial interaction modelling in retail planning practice: the need for robust statistical methods. *Environment and Planning B: Planning and Design*, *18*, 191-203.

Hägerstrand, T. (1970). What about people in regional science? *Papers in regional science*, *24*, 7-24.

Hanson, S. (2010). Gender and mobility: new approaches for informing sustainability. *Gender, Place & Culture, 17*, 5-23.

Harris, R. & Jarvis, C. (2014). *Statistics for geography and environmental science*. Abingdon: Routledge.

Harris, R., Sleight, P. & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting (Vol. 7)*. Chichester: John Wiley and Sons.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*, 100-108.

Heerde, H. J. V., & Bijmolt, T. H. (2005). Decomposing the promotional revenue bump for loyalty program members versus nonmembers. *Journal of Marketing Research*, *42*, 443-457.

Heuvelink, G. B. M. (1999). Propagation of error in spatial modelling with GIS. *Geographical information systems, 1*, 207-217.

Hey, T., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery (Vol. 1).* Redmond, WA: Microsoft research.

Hiziroglu, A. (2013) Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications 40*, 6491–6507.

Hornsey, R. (2010). *'He who Thinks, in Modern Traffic, is Lost': Automation and the Pedestrian Rhythms of Interwar London.* University of Minnesota Press.

Huff, D. L. (1963). A probabilistic analysis of shopping center trade areas. *Land economics, 39*, 81-90.

Humby, C., Hunt, T. & Phillips, T. (2004). *Scoring points: How Tesco is winning customer loyalty.* London: Kogan Page.

Isard, W. (1956). Regional science, the concept of region, and regional structure. *Papers in Regional Science, 2,* 13–26.

Kamenjuk, P., Aasa, A., & Sellin, J. (2017). Mapping changes of residence with passive mobile positioning data: the case of Estonia. *International Journal of Geographical Information Science, 31,* 1425-1447.

Kandt, J. (2015a). *Geodemographics and spatial microsimulation: using survey data to infer health milieu geographies.* [Online]. Available at: https://bit.ly/2HdYEg8.

Kandt, J. (2015b). *The social and spatial context of urban health inequalities: towards an interpretive geodemographic framework*. (Doctoral dissertation, UCL).

Kang, H., & Scott, D. M. (2010). Exploring day-to-day variability in time use for household members. *Transportation Research Part A: Policy and Practice, 44,* 609-619.

Karr, A. F., Reiter, J. P., & Lane, J. (2014). *Using statistics to protect privacy*. *Privacy, Big Data, and the Public Good Frameworks for Engagement.* Cambridge University Press, Cambridge.

Kawachi, I., & Berkman, L. F. (Eds.). (2003). *Neighborhoods and health.* Oxford University Press.

Kelling, S, Hochachka, W, Fink, D. (2009). Data-intensive Science: A new paradigm for biodiversity studies. *BioScience, 59,* 613–620.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography, 3,* 262-267.

Kitchin, R. (2014a). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1, 2053951714528481.*

Kitchin, R. (2014b). *The data revolution: Big data, open data, data infrastructures and their consequences.* London: Sage.

Kotler, P. (2002). *Marketing places*. New York: Simon and Schuster.

Kwan, M. P. (2009). From place-based to people-based exposure measures. *Social science & medicine, 69,* 1311-1313.

Kwan, M. P. (2012a). How GIS can help address the uncertain geographic context problem in social science research. *Annals of GIS, 18*, 245-255.

Kwan, M. P. (2012b). The uncertain geographic context problem. *Annals of the Association of American Geographers, 102*, 958-968.

Kwan, M. P. (2013). Beyond space (as we knew it): toward temporally integrated geographies of segregation, health, and accessibility: Space–time integration in geography and GIScience. *Annals of the Association of American Geographers, 103*, 1078-1086.

Kwan, M. P., & Weber, J. (2008). Scale and accessibility: Implications for the analysis of land use–travel interaction. *Applied Geography, 28*, 110-123.

LaBelle, B. (2008). Pump up the bass—Rhythm, cars, and auditory scaffolding. *The Senses and Society, 3*, 187-203.

Lager, D., Van Hoven, B., & Huigen, P. P. (2016). Rhythms, ageing and neighbourhoods. *Environment and Planning A, 48*, 1565-1580.

Lane, J., Stodden, V., Bender, S. & Nissenbaum, H. (Eds.), (2014). *Privacy, big data, and the public good: Frameworks for engagement.* Cambridge: Cambridge University Press.

Laney, D. (2001), 3-D Data Management: Controlling Data Volume, Velocity and Variety, *META Group Research Note, February 6*. [Online]. Available at https://bit.ly/2H2kMMS:

Lansley, G. (2014). Evaluating the Utility of Geo-referenced Twitter Data as a Source of Reliable Footfall Insight. [Online]. *In Proceedings of the Association of American Geographers AGM 2014, Tampa, USA.* Available at: https://bit.ly/2HcAkeH.

Lansley, G., Li, W., & Longley, P. (2017). Representing Population Dynamics from Administrative and Consumer Registers. *In Proceedings of Geographical Information Science Research UK (GISRUCK) 2017, Manchester, UK.* Available at: https://bit.ly/2GzXsa1.

Lansley, G., Wei, Y., & Rains, T. (2015). Creating an Output Area Classification of Cultural and Ethnic Heritage to Assist the Planning of Ethnic Origin Foods in Supermarkets in England and Wales. [Online]. *In Proceedings of Geographical Information Science Research UK (GISRUCK) 2015, Leeds, UK.* Available at: https://bit.ly/2IyfQwD.

Larsen, J., & Urry, J. (2016). *Mobilities, networks, geographies.* Routledge.

Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T. M. T., Dousse, O, Eberle. J., & Miettinen, M. (2012). The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing* (No. EPFL-CONF-192489).

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science, 343*, 1203–1205

Leak, A. B. (2017). *Applications of new forms of data to demographics* (Doctoral dissertation, UCL).

Leenheer, J., Van Heerde, H. J., Bijmolt, T. H., & Smidts, A. (2007). Do loyalty programs really enhance behavioral loyalty? An empirical analysis accounting for self-selecting members. *International Journal of Research in Marketing, 24*, 31-47.

Lefebvre, H. (2004). *Rhythmanalysis: Space, time and everyday life*. A&C Black.

Leventhal, B. (2016). *Geodemographics for marketers: Using location analysis for research and marketing*. Kogan Page Publishers.

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: methods and applications*. John Wiley & Sons.

Lloyd, A., & Cheshire, J. (2018). Detecting Address Uncertainty in Loyalty Card Data. *Applied Spatial Analysis and Policy*, 1-21.

Longley, P.A. (2005). *Geographic information systems and science.* John Wiley & Sons.

Longley, P. A. (2017). *Geodemographic Profiling*. The International Encyclopedia of Geography.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A*, 47, 465-484.

Longley, P., Cheshire, J., & Singleton, A. (Eds.). (2018). *Consumer Data Research*. UCL Press.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic information science and systems*. John Wiley & Sons.

Longley, P. A. & Harris, R. J. (1999). Towards a new digital data infrastructure for urban analysis and modelling. *Environment and Planning B: Planning and Design, 26*, 855-878.

Loomis, C. P. (1946). Political and occupational cleavages in a Hanoverian village. *Sociometry, 9*, 316-33

Loukides, M. (2010). What is data science? The future belongs to the companies and people that turn data into products. [Online]. *An OReilly Radar Report*. Available at: https://oreil.ly/1faH0Xu.

Loyalive, (2015). *Loyalive – An Introduction*. [Online]. Available at: https://blog.loyalive.com/intro/.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA*, 1, 281–297.

Maritz Research (2006), Understanding your customers – choice, experience, loyalty. *The Research Report, Vol. 19.*

Martin, D. (1998). Optimizing census geography: the separation of collection and output geographies, *International Journal of Geographical Information Science, 12*, 673–685.

Martin, D. (2000). Towards the geographies of the 2001 UK Census of Population. *Transactions of the Institute of British Geographers, 25*, 321-332.

Martin, D. (2002). Geography for the 2001 Census in England and Wales. *Population Trends, 108*, 15.

Martin, D. (2006). Last of the censuses? The future of small area population data. *Transactions of the Institute of British Geographers, 31*, 6-18.

Martin, D. (2010). Understanding the social geography of census undercount. *Environment and Planning A, 42*, 2753-2770.

Martin, D., Nolan, A., & Tranmer, M. (2001). The application of zone-design methodology in the 2001 UK Census. *Environment and Planning A, 33*, 1949-1962.

Martin-Fernandez, J., Barcelo-Vidal, C., & Pawlowsky-Glahn, V. (1998). A critical approach to non-parametric classification of compositional data. In *Advances in data science and classification*, 49–56. Springer.

Matthews, S. A. (2008). The salience of neighborhood: Some lessons from sociology. *American Journal of Preventive Medicine, 34*, 257–59.

Matthews, S. A. (2011). Spatial polygamy and the heterogeneity of place: Studying people and place via egocentric methods. In *Communities, neighborhoods, and health*, 35-55, Springer, New York, NY.

Matthews, S. A., Detwiler, J. and Burton, L. (2005). Geo-ethnography: Coupling geographic information analysis techniques and ethnographic methods in urban research. *Cartographica, 40,* 75–90.

Mauri, C. (2003). Card loyalty. A new emerging issue in grocery retailing. *Journal of Retailing and Consumer Services, 10,* 13-25.

Mayer-Schonberger, V. & Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work, and think (1st ed.).* Boston: Houghton Mifflin Harcourt.

McLachlan, G. & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology, 27*, 415-444.

Mennis, J., & Mason, M. J. (2011). People, places, and adolescent substance use: integrating activity space and social network data for analyzing health behavior. *Annals of the Association of American Geographers, 101*, 272-291.

Meyer-Waarden, L. (2007). The effects of loyalty programs on customer lifetime duration and share of wallet. *Journal of Retailing, 83,* 223-236.

Miller, H. J. & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal, 80,* 449-461.

Miller, H.J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science 50,* 181-201

Miller, M. M., Gibson, L. J., & Wright, N. G. (1991). Location quotient: A basic tool for economic development analysis. *Economic Development Review, 9,* 65.

Milne, G. R. & Culnan, M. J. (2004). Strategies for reducing Online privacy risks: Why consumers read (or don't read) Online privacy notices. *Journal of Interactive Marketing, 18,* 15-29.

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology, 9*, 15-24.

Neutens, T., Delafontaine, M., Scott, D. M. & De Maeyer, P. (2012). An analysis of day-to-day variations in individual space–time accessibility. *Journal of Transport Geography, 23,* 81-91.

Neutens, T., Schwanen, T. & Witlox, F. (2011). The prism of everyday life: towards a new research agenda for time geography. *Transport reviews, 31*, 25-47.

Neutens, T., Versichele, M., & Schwanen, T. (2010). Arranging place and time: A GIS toolkit to assess person-based accessibility of urban opportunities. *Applied Geography, 30*, 561-575.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life.* Palo Alto, USA: Stanford University Press.

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus, 140*, 32-48.

Nowok, B., Raab, G. M. & Dibben, C. (2015). Synthpop: Bespoke creation of synthetic data in R. [Online]. *Journal of Statistical Software*. Available at: https://bit.ly/2HaR0TW.

NRS (2018), *National Records of Scotland.* [Online]. Available at: https://www.nrscotland.gov.uk/

O'Brien, O., & Cheshire, J. (2014). Mapping Geodemographic classification uncertainty: an exploration of visual techniques using compositing operations. [Online]. In *Workshop "Visually-Supported Reasoning with Uncertainty", GIScience*. Available at: https://bit.ly/2GCzwil.

O'Brien. O., (2016). *Population Density and Urban/Rural Classification.* [Online]. Available at: https://bit.ly/2qdi3qD.

ONS (2014). 2011 *Census: Internal and international migration for the United Kingdom in the year prior to the 2011 Census.* [Online]. Available at: https://bit.ly/2q5U1xI.

ONS (2017*). Local Authority Districts: Methodology for the 2011 area classification for local authorities*. [Online]. Available at: https://bit.ly/2ElF66N.

ONS (2018a). *2011 Census Data.* [Online]. Available at: https://bit.ly/2fsjDkI.

ONS (2018b). *2011 Census Origin-Destination Data User Guide.* [Online]. Available at: https://bit.ly/2uLr12V.

Openshaw, S. (1977) A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling, *Transactions of the Institute of British Geographers, 2*, pp. 459–472.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and planning A, 16*, 17-31.

Openshaw, S. (1989). Learning to live with errors in spatial databases. In Goodchild, M. & Gopal, S. (Eds.), *Accuracy of spatial databases*, 263-276. London: Taylor & Francis.

Openshaw, S., Blake, M. & Wymer, C. (1995). Using neurocomputing methods to classify Britain's residential areas. *Innovations in GIS, 2*, 97-111.

Palarea-Albaladejo, J., Martín-Fernández, J. A., & Soto, J. A. (2012). Dealing with distances and transformations for fuzzy c-means clustering of compositional data. *Journal of classification, 29*, 144-169.

Pawlowsky-Glahn, V., & Buccianti, A. (Eds.). (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana Delgado, R. (2007). Lecture notes on compositional data analysis. [Online]. Available at: https://bit.ly/2EmVfIR.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data.* John Wiley & Sons.

Prensky, M. (2009). H. sapiens digital: From digital immigrants and digital natives to digital wisdom. *Innovate: journal of online education, 5*, 1.

Press, G. (2014). 12 Big data definitions: What's yours? [Online]. *Forbes.* Available at: https://bit.ly/2GW1zMR.

Quan-Haase, A., & Sloan, L. (2017). Introduction to the Handbook of Social Media Research Methods: Goals, Challenges and Innovations. *The SAGE Handbook of Social Media Research Methods*, 1.

R Core Team (2018). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing, Vienna, Austria.* Available at: https://www.R-project.org.

Rainham, D., McDowell, I., Krewski, D. & Sawada, M. (2010). Conceptualizing the healthscape: contributions of time geography, location technologies and spatial ecology to place and health research. *Social science & medicine, 70*, 668-676.

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public opinion quarterly, 76*, 163-181.

Richardson, H. M. (1940). Community of values as a factor in friendships of college and adult women. *The Journal of Social Psychology, 11*, 303-12

Riddlesden, D. (2014). *Internet User Classification User Guide.* [Online]. Available at: https://bit.ly/2IvL4EA.

Roorda M. J., & Ruiz, T. (2008). Long- and short-term dynamics in activity scheduling: a structural equations approach. *Transportation Research, Part A: Policy and Practice, 42*, 545-562

Sanford, M. (2008) *Consumption and the Urban Milieu: Using Consumption as a Measure of Similarity for Defining Urban Neighborhoods* (Doctoral Dissertation. The University of Chicago).

Savills, (2005). The Blackburn with Darwen Shopping Study 2005–2016. *Savills, London.* Available at: https://bit.ly/2q9mWA6.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in higher education, 44*, 409-432.

Schönfelder, S., & Axhausen, K. W. (2016). *Urban rhythms and travel behaviour: spatial and temporal phenomena of daily travel*. Routledge.

Schwanen, T. (2007). Gender differences in chauffeuring children among dual-earner families. *The Professional Geographer, 59*, 447-462.

Schwanen, T., Banister, D., & Anable, J. (2012). Rethinking habits and their role in behaviour change: the case of low-carbon mobility. *Journal of Transport Geography, 24*, 522-532.

Schwanen, T., Kwan, M. P., & Ren, F. (2008). How fixed is fixed? Gendered rigidity of space–time constraints and geographies of everyday activities. *Geoforum, 39*, 2109-2121.

Shearer, C., Rainham, D., Blanchard, C., Dummer, T., Lyons, R. & Kirk, S. (2015). Measuring food availability and accessibility among adolescents: Moving beyond the neighbourhood boundary. *Social Science & Medicine, 133*, 322-330.

Sheller, M. (2011). Mobility. *Sociopedia. International Sociological Association.* [Online], 1–12. Available at: https://bit.ly/1JzNAWX.

Sheller, M., & Urry, J. (2006). The new mobilities paradigm. *Environment and planning A, 38*, 207-226.

Simma. A., & Axhausen, K., (2001) Successive days, related travel behaviour? *Arbeitsbericht Verkehrs-und Raumplanung*, 62.

Singleton, A. D., & Longley, P. A. (2009). Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Papers in Regional Science, 88,* 643-666.

Singleton, A. D. & Longley, P. A. (2015). The internal structure of Greater London: a comparison of national and regional geodemographic models. Geo: *Geography and Environment, 2,* 69-87.

Singleton, A. D. & Spielman, S. E. (2014). The past, present, and future of geodemographic research in the United States and United Kingdom. *The Professional Geographer, 66,* 558-567.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70,* 263-286.

Slingsby, A., Dykes, J. & Wood, J. (2011). Exploring uncertainty in geodemographics with interactive graphics. *IEEE Transactions on Visualization and Computer Graphics, 17,* 2545-2554.

Smith, A. & Sparks, L. (2009). Reward redemption behaviour in retail loyalty schemes. *British Journal of Management, 20,* 204-218.

Smith, A., Sparks, L., Hart, S., & Tzokas, N. (2003). Retail loyalty schemes: results from a consumer diary study. *Journal of Retailing and Consumer Services, 10,* 109-119.

Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania law review, 154,* 477-564.

Solove, D.J. 2013. Introduction: Privacy Self-Management and the Consent Dilemma, *Harvard Law Review, 126,* 1880-1903.

Song, T. Koren, P. Wang, A. L. Barabási (2010). Modelling the scaling properties of human mobility. *Nature Physics, 6,* 818-823.

Stillwell, J. (Ed.). (2016). *The Routledge handbook of census resources, methods and applications: Unlocking the UK 2011 census.* London: Routledge.

Strong, C. (2015). *Humanizing big data: Marketing at the meeting of data, social science and consumer insight*. Kogan Page Publishers.

Summerfield, M. A. (1983). Populations, samples and statistical inference in geography. *The Professional Geographer, 35,* 143-149.

Susilo, Y. O., & Axhausen, K. W. (2014). Repetitions in individual daily activity–travel–location patterns: a study using the Herfindahl–Hirschman Index. *Transportation, 41*, 995-1011.

Tauber, F. (1999). Spurious clusters in granulometric data caused by logratio transformation. *Mathematical Geology, 31*, 491–504.

Taylor, P. J. (1971). Distance transformation and distance decay functions. *Geographical Analysis, 3*, 221-238.

Tene, O., & Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology & Intellectual Property, 11*, xxvii.

TFL (2017). *TfL's free open data boosts London's economy*. [Online]. Available at: https://bit.ly/2JlJzdj.

Thrift, N. J. (1977). *An introduction to time geography*. Geo Abstracts, University of East Anglia.

Timmermans, H., Arentze, T. & Joh, C. H. (2002). Analysing space-time behaviour: new approaches to old problems. *Progress in human geography, 26*, 175-190.

Tinbergen, J. (1962). *Shaping the World Economy*. Twentieth Century Fund, New York, NY.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*, 234-240.

Tsagris, M. T., Preston, S., & Wood, A. T. (2011). A data-based power transformation for compositional data. *arXiv preprint arXiv:1106.1451*.

UKDS (2018a). *Key data*. [Online]. Available at: https://bit.ly/2uT6UQy.

UKDS (2018b). *UK Data Service.* [Online]. Available at: https://bit.ly/2EmmLqc.

Unwin, D. J. (1996). GIS, spatial analysis and spatial statistics. *Progress in Human Geography, 20*, 540-551.

Valentine, G. (2008). Living with difference: reflections on geographies of encounter. *Progress in human geography, 32*, 323-337.

Van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). *Analyzing compositional data with R (Vol. 122)*. Berlin: Springer.

Van den Boogaart, K. G., Tolosana, R., & Bren, M. (2015). *Compositional data analysis.* R-package software (Version 1.40-1).

Van der Klis, M., & Karsten, L. (2009). Commuting partners, dual residences and the meaning of home. *Journal of Environmental Psychology, 29*, 235–245.

Verhulst, S. G. (2015). *Data Collaboratives: Exchanging Data to Improve People's Lives*. [Online]. Available at: https://bit.ly/1DgtiKK.

Verhulst, S. G., Young. A., & Srinivasan. P. (2018). *An introduction to data collaboratives: Creating public value by exchanging data.* [Online]. Available at: https://bit.ly/2ElKDdq.

Vickers, D., Rees, P., & Birkin, M. (2005). *Creating the national classification of census output areas: data, methods and results*. Working Paper. School of Geography, University of Leeds.

Voas, D. & Williamson, P. (2000). The scale of dissimilarity: Concepts, measurement and an application to socio-economic variation across England and Wales. *Transactions of the Institute of British Geographers, 25*, 465-481.

Walker, S. J. (2014) Big Data: A Revolution That Will Transform How We Live, Work, and Think, *International Journal of Advertising, 33*, 181-183.

Wang, S., Shi, W., Yuan, H. & Chen, G. (2005). Attribute uncertainty in GIS data. *In International Conference on Fuzzy Systems and Knowledge Discovery, 614-623*. Berlin – Heidelberg: Springer.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*, 236–244.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*, 3-46.

Webber, R. & Longley, P. (2003). Analysis of similarity and proximity: their roles in the understanding of the geography of need. In Longley, P., & Batty, M. (Eds.) (2003). *Advanced spatial analysis: the CASA book of GIS*. Redmond, USA: ESRI, Inc.

Webber, R. J., Butler, T., & Phillips, T. (2015). Adoption of geodemographic and ethno-cultural taxonomies for analysing big data. *Big Data & Society, 2*, 2053951715583914.

Widener, M. J., Farber, S., Neutens, T., & Horner, M. W. (2013). Using urban commuting data to calculate a spatiotemporal accessibility measure for food environment studies. *Health & place, 21*, 1-9.

Wilson, A. G. (1971). A family of spatial interaction models, and associated developments. *Environment and Planning A, 3,* 1-32.

Wong, D. (2009). *The modifiable areal unit problem (MAUP).* The SAGE handbook of spatial analysis, 105-123.

Wright, C., & Sparks, L. (1999). Loyalty saturation in retailing: exploring the end of retail loyalty cards*?*. *International Journal of Retail & Distribution Management, 27*, 429-440.

Wrigley, N., & Brookes, E. (2014). *Evolving high streets: resilience and reinvention-perspectives from social science.* ESRC, University of Southampton.

Wrigley, N., & Lambiri, D. (2014). *High street performance and evolution: A brief guide to the evidence.* University of Southampton.

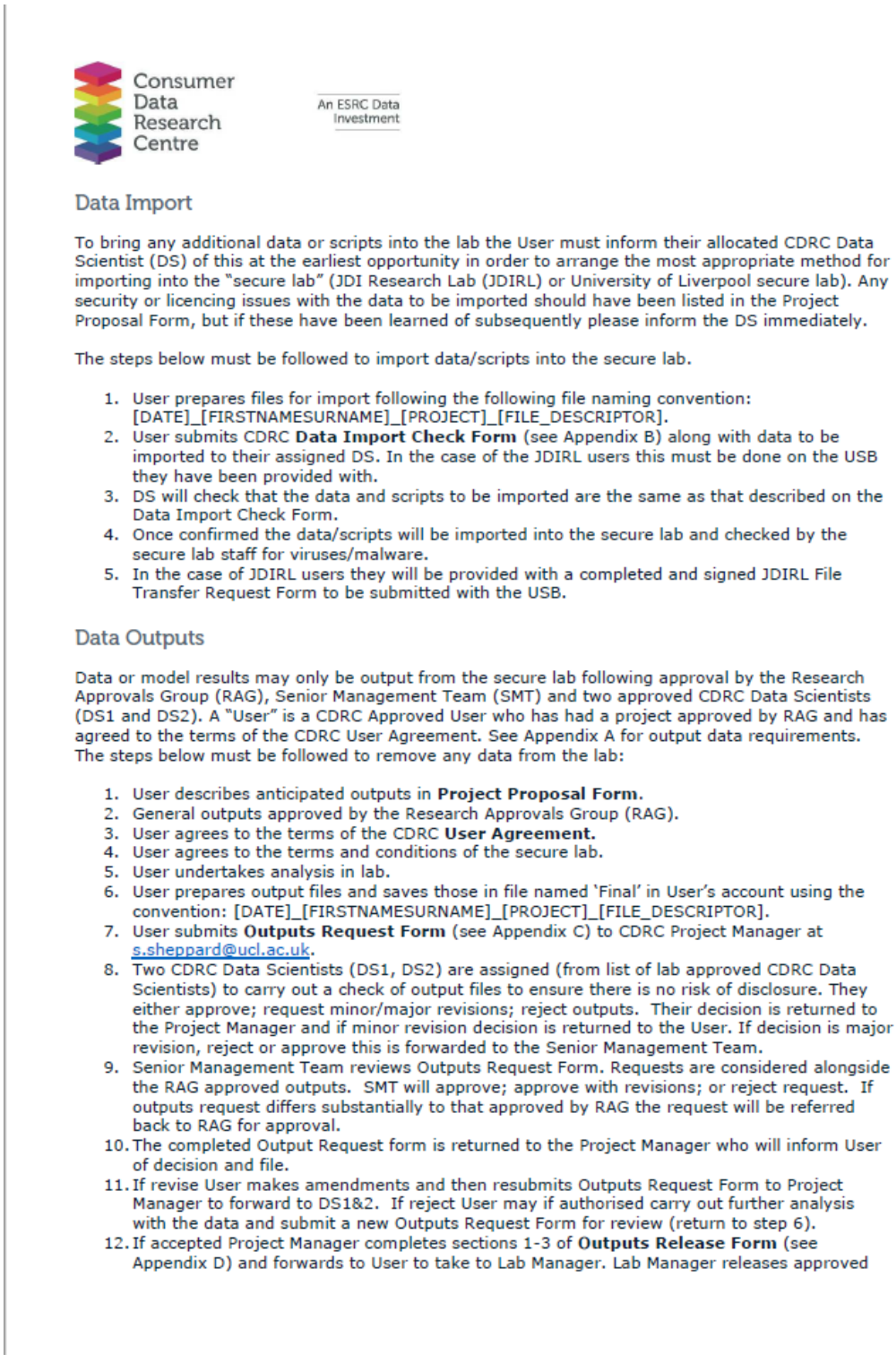
Wu, J. P., & Wei, S. (1989). *Time series analysis.* Human Science and Technology Press, ChangSha.

YouGov, (2013). *British shoppers in love with loyalty cards.* [Online]. Available at: https://bit.ly/2Hdsy4u.

Zhou, D., Chen, H., and Lou, Y. (1991). The logratio approach to the classification of modern sediments and sedimentary environments in northern South China Sea. *Mathematical Geology, 23*, 157–165.

10. Appendix

1. CDRC Statistical Disclosure control guidance material.



Consumer Data Research Centre

An ESRC Data Investment

Data Import

To bring any additional data or scripts into the lab the User must inform their allocated CDRC Data Scientist (DS) of this at the earliest opportunity in order to arrange the most appropriate method for importing into the "secure lab" (JDI Research Lab (JDIRL) or University of Liverpool secure lab). Any security or licencing issues with the data to be imported should have been listed in the Project Proposal Form, but if these have been learned of subsequently please inform the DS immediately.

The steps below must be followed to import data/scripts into the secure lab.

1. User prepares files for import following the following file naming convention: [DATE]_[FIRSTNAMESURNAME]_[PROJECT]_[FILE_DESCRIPTOR].
2. User submits CDRC **Data Import Check Form** (see Appendix B) along with data to be imported to their assigned DS. In the case of the JDIRL users this must be done on the USB they have been provided with.
3. DS will check that the data and scripts to be imported are the same as that described on the Data Import Check Form.
4. Once confirmed the data/scripts will be imported into the secure lab and checked by the secure lab staff for viruses/malware.
5. In the case of JDIRL users they will be provided with a completed and signed JDIRL File Transfer Request Form to be submitted with the USB.

Data Outputs

Data or model results may only be output from the secure lab following approval by the Research Approvals Group (RAG), Senior Management Team (SMT) and two approved CDRC Data Scientists (DS1 and DS2). A "User" is a CDRC Approved User who has had a project approved by RAG and has agreed to the terms of the CDRC User Agreement. See Appendix A for output data requirements. The steps below must be followed to remove any data from the lab:

1. User describes anticipated outputs in **Project Proposal Form**.
2. General outputs approved by the Research Approvals Group (RAG).
3. User agrees to the terms of the CDRC **User Agreement.**
4. User agrees to the terms and conditions of the secure lab.
5. User undertakes analysis in lab.
6. User prepares output files and saves those in file named 'Final' in User's account using the convention: [DATE]_[FIRSTNAMESURNAME]_[PROJECT]_[FILE_DESCRIPTOR].
7. User submits **Outputs Request Form** (see Appendix C) to CDRC Project Manager at s.sheppard@ucl.ac.uk.
8. Two CDRC Data Scientists (DS1, DS2) are assigned (from list of lab approved CDRC Data Scientists) to carry out a check of output files to ensure there is no risk of disclosure. They either approve; request minor/major revisions; reject outputs. Their decision is returned to the Project Manager and if minor revision decision is returned to the User. If decision is major revision, reject or approve this is forwarded to the Senior Management Team.
9. Senior Management Team reviews Outputs Request Form. Requests are considered alongside the RAG approved outputs. SMT will approve; approve with revisions; or reject request. If outputs request differs substantially to that approved by RAG the request will be referred back to RAG for approval.
10. The completed Output Request form is returned to the Project Manager who will inform User of decision and file.
11. If revise User makes amendments and then resubmits Outputs Request Form to Project Manager to forward to DS1&2. If reject User may if authorised carry out further analysis with the data and submit a new Outputs Request Form for review (return to step 6).
12. If accepted Project Manager completes sections 1-3 of **Outputs Release Form** (see Appendix D) and forwards to User to take to Lab Manager. Lab Manager releases approved

files on encrypted media either to User or to DS. If released to DS then they will arrange the delivery of the files to User by secure download.

13. User signs Output Release Form acknowledging receipt of data. A copy of the signed form to be returned to the Project Manager for filing.

Appendix A: Statistical Disclosure Control - Output Requirements

Outputs

Outputs requested should be 'finished outputs' i.e. the finished statistical analyses that you intend to present to the public. If requiring intermediate outputs for a particular reason e.g. to present initial findings then these may be considered if clearly presented and clearly explained. All outputs must be easy to read and interpret and how they are to be used explained.

Non-Disclosive Data

Taken from GSS/GSR Disclosure Control Guidance for Tables Produced from Surveys, October 2014

Social Surveys

- For the majority of surveys, outputs should be for large geographical areas, e.g. Country or Government Office Region, or in some cases Local Authority District (or equivalent). The level of geography should reflect survey design.
- Suppress or combine unsafe cells, i.e. where there are one or two units contributing to the cell.
- Where the sample size of a total or sub-total is one or two, suppress the whole row or column to which the total refers, including any zero cells (or combine neighbouring categories).
- In unweighted tables, cell suppression does not provide sufficient protection. Unsafe cells should only be combined with other cells.
- If unweighted sample base numbers are essential they should be conventionally rounded to base 10.
- Percentages may be released, provided it is not possible to deduce where only one or two units have contributed to the cell.
- Units may be individuals, families or households, communal establishments or any other unit whose confidentiality should be protected.

Subsamples

- For the majority of surveys, outputs should be for large geographical areas, e.g. regions, or in some cases Local Authority District (or equivalent). The level of geography should reflect survey design.
- Table design should be used to remove all unsafe cells, i.e. where there is one unit contributing to a cell. Variable categories should be combined or variables removed until only safe cells remain.
- Percentages may be released, provided it is not possible to deduce where only one unit has contributed to the cell.
- Units may be individuals, families or households or any other unit whose confidentiality should be protected.

Business Surveys: Magnitude tables

- A cell meeting both the following criteria is safe (otherwise the cell is unsafe):
  - there must be at least *n* enterprise groups in a cell (threshold rule)
  - the total of the cell minus the largest *m* reporting unit(s) must be greater than or equal to p% of the value of the largest reporting unit (p% rule)

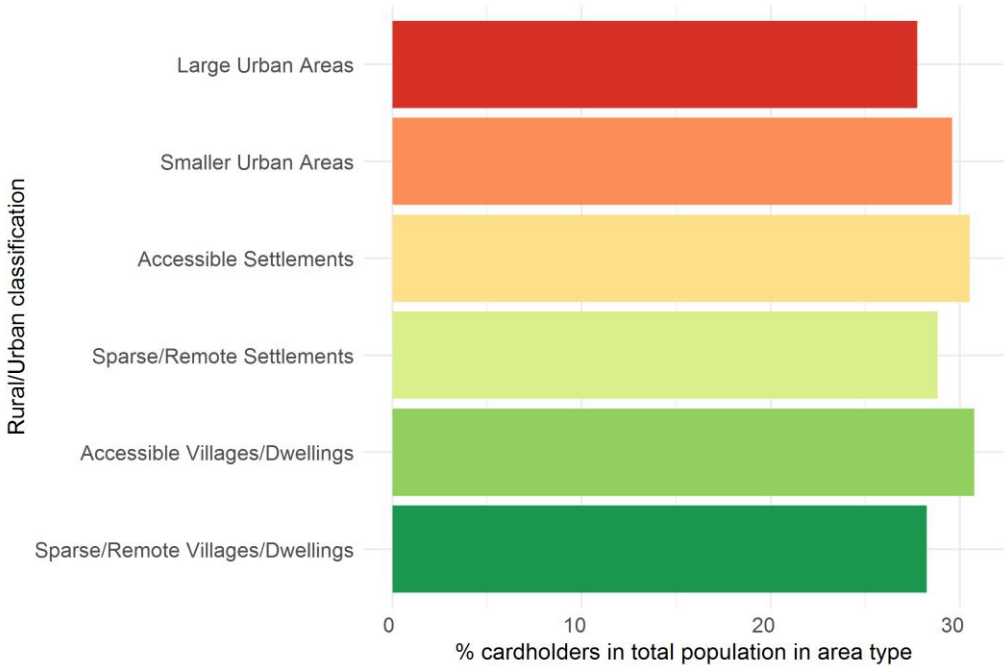
Note that the values of the p% and minimum threshold parameter *n* and m should remain confidential, since knowledge of these values reduces the protection. The choice of p, n and m would usually be decided by the Responsible Statistician. Typical examples would be 2,3,4,5 (for n), and 2,3 (for m) and 5% 10% ,15%, 20% (for p).

- Table design should be used first to reduce the number of unsafe cells in a table where this is consistent with the main uses of the data.
- Cell suppression is the standard method used to protect tables with unsafe cells. The unsafe cells are suppressed, known as *primary suppressions*. Other cells must be suppressed to prevent the values of the unsafe cells being calculated by subtraction from the marginal totals of the table. These are known as *secondary suppressions*.
- Cell suppression does not generally provide protection from disclosure by differencing. Tables should be published using fixed categories to avoid disclosure by differencing. For example the same geographies and SIC codes should always be used.

Business Surveys: Count data

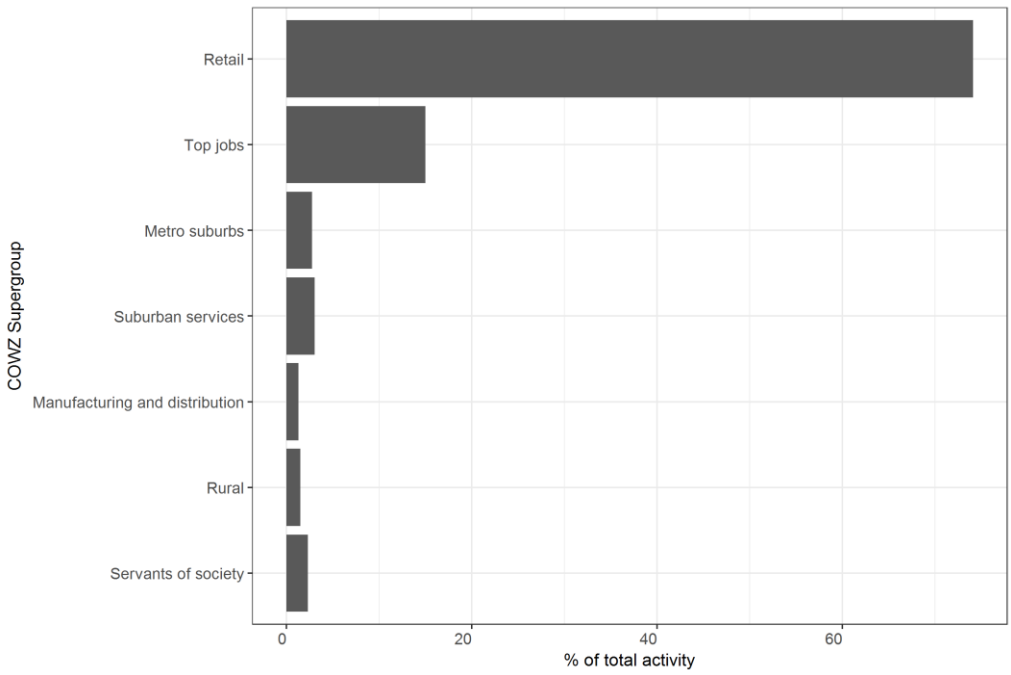
- Tables of count data are to be protected by redesign of the table to protect sensitive cells. If further protection is required other techniques such as controlled rounding to base 5 should be considered.
- Percentages or rates must be derived from rounded values.

2. **Comparison of the total Census population per area type with the total customers.**

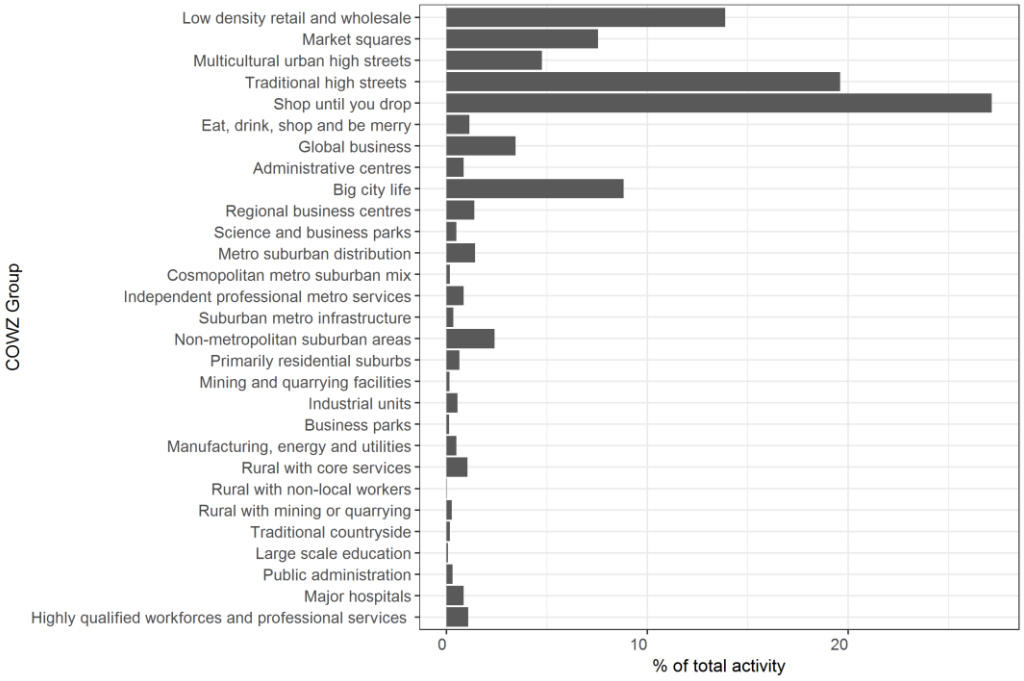


3.

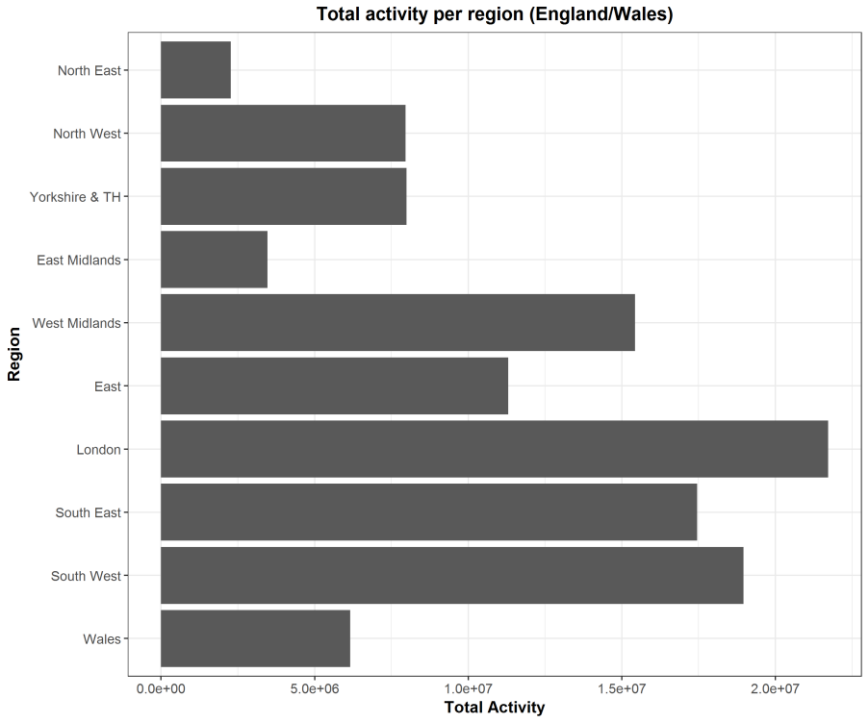
a. **The percentage of HSR stores per COWZ Supergroup.**



b. The percentage of HSR stores per COWZ Supergroup.

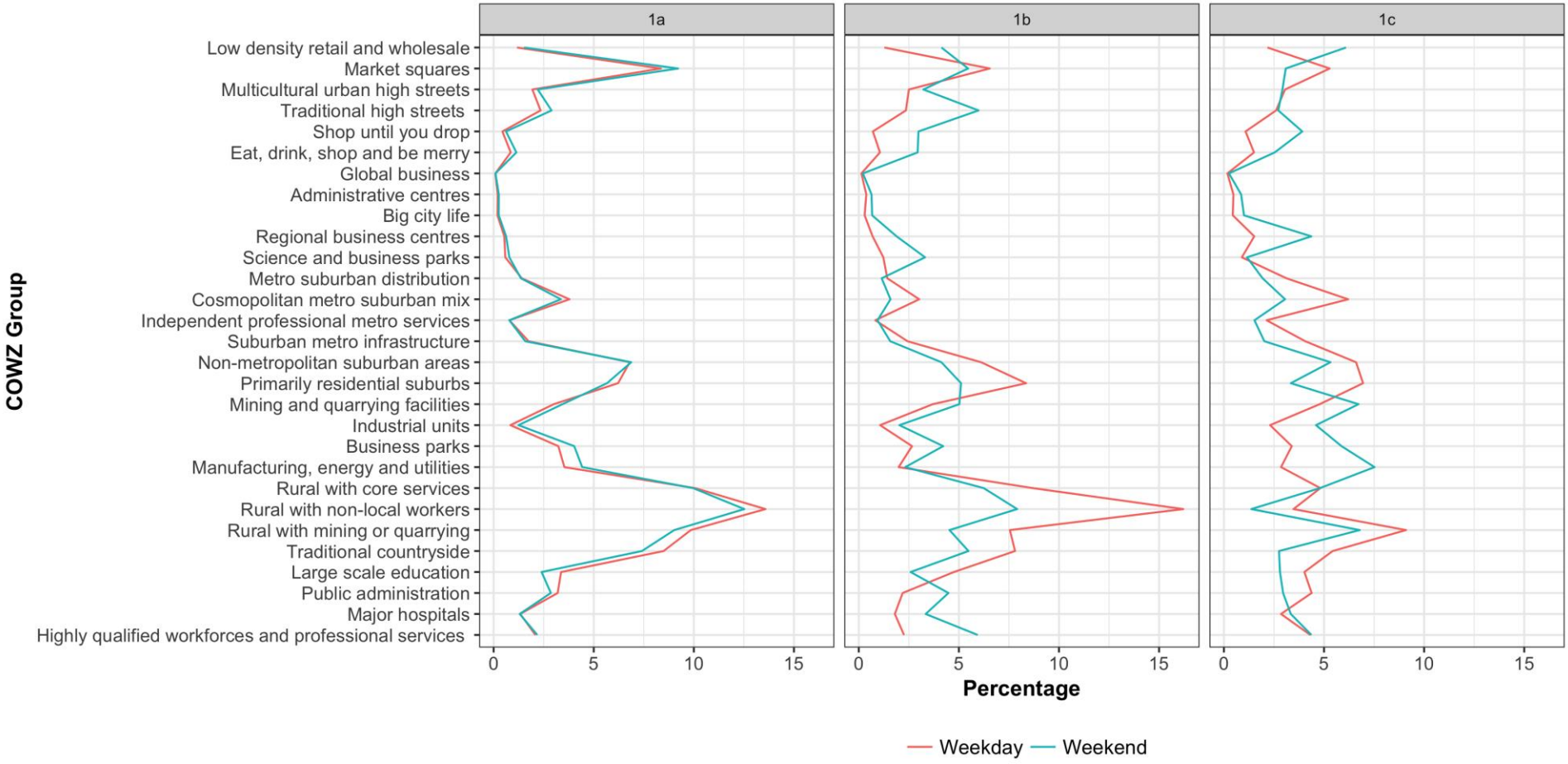


c. Total HSR activity, per region across England and Wales (count).

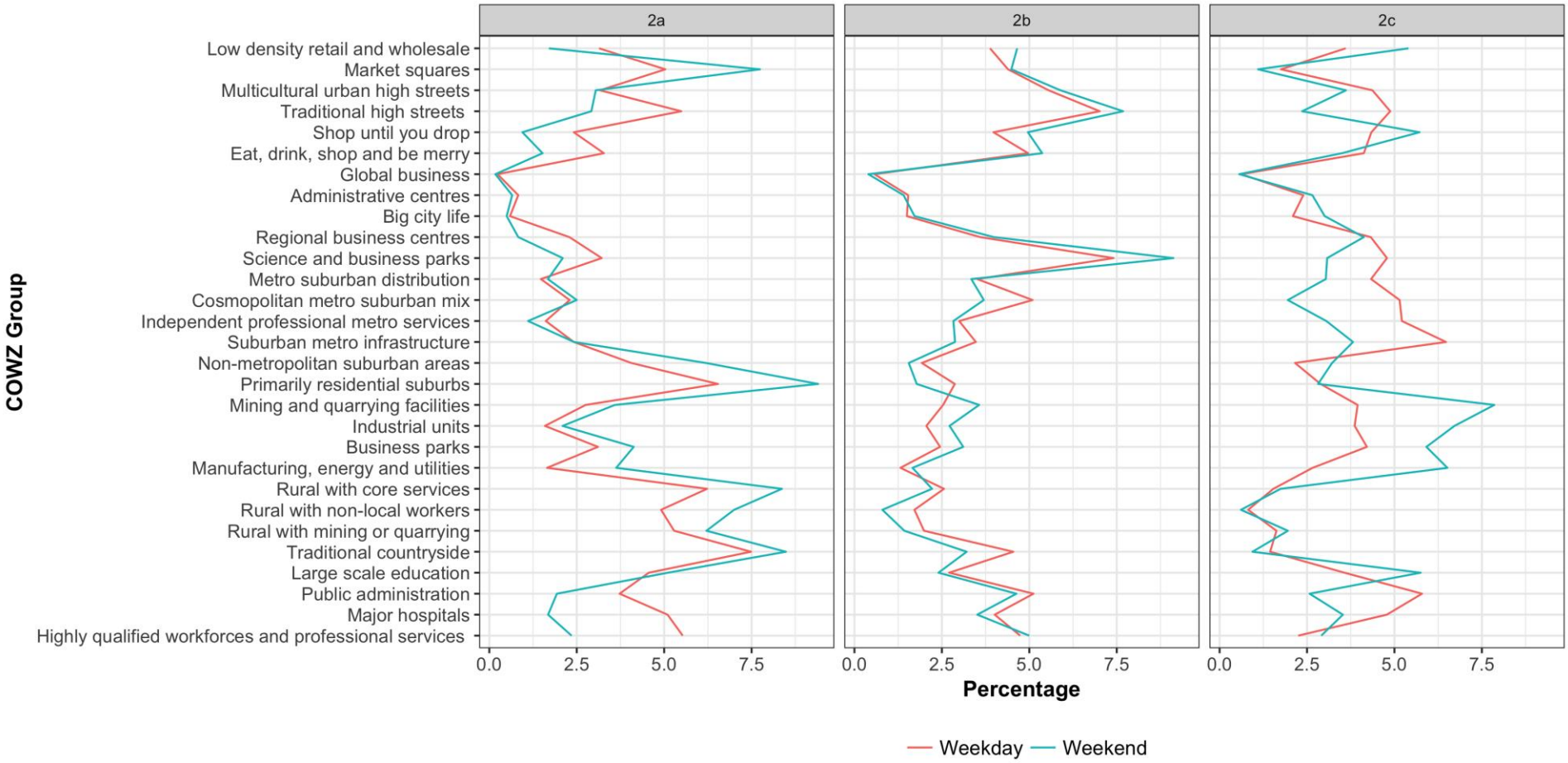


4.

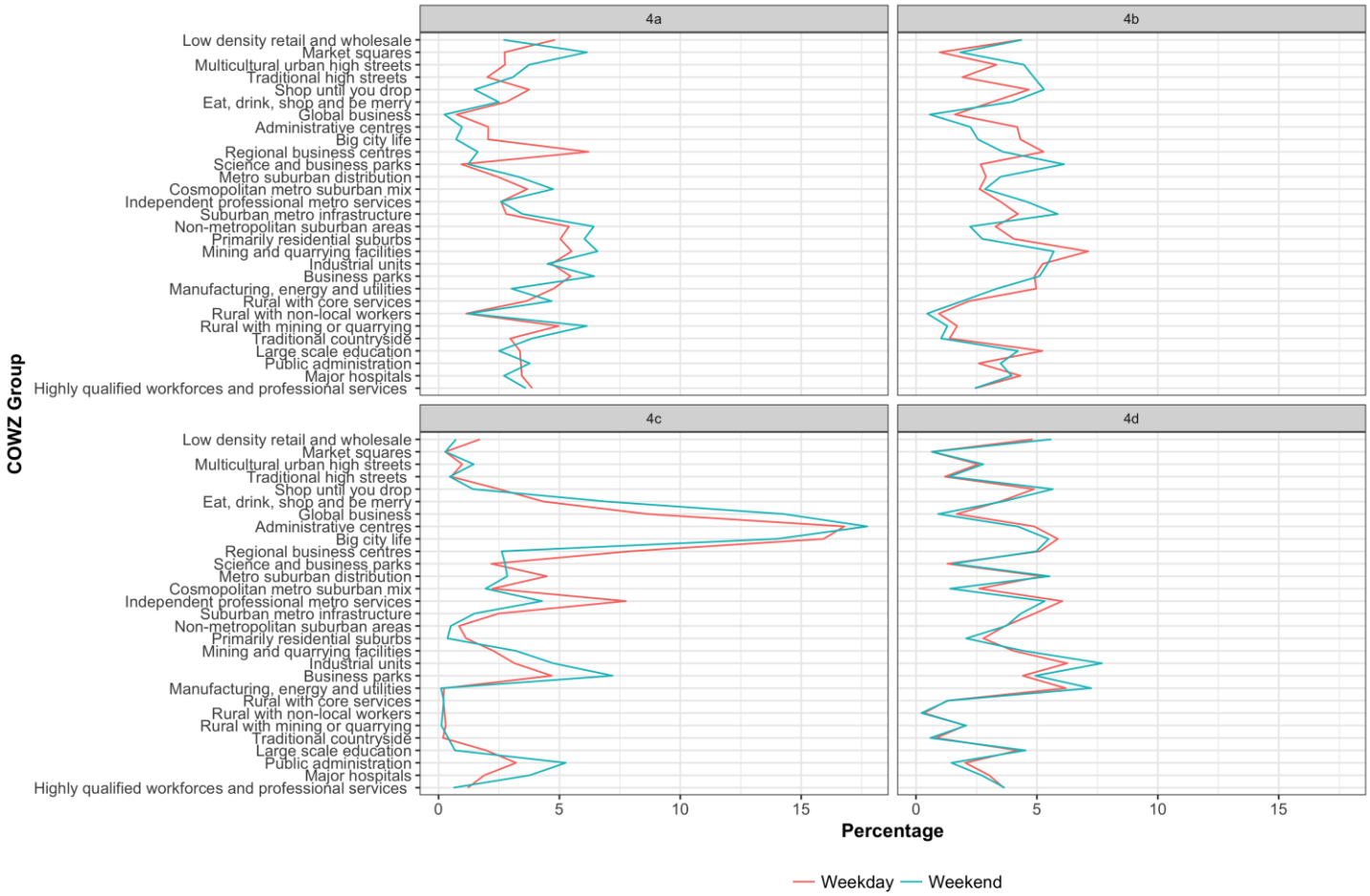
a. Supergroup 1 area visiting characteristics during weekdays and weekends (Note: Scales are varying to illustrate fluctuations within each Group).



b. Supergroup 2 area visiting characteristics during weekdays and weekends (Note: Scales are varying to illustrate fluctuations within each Group).

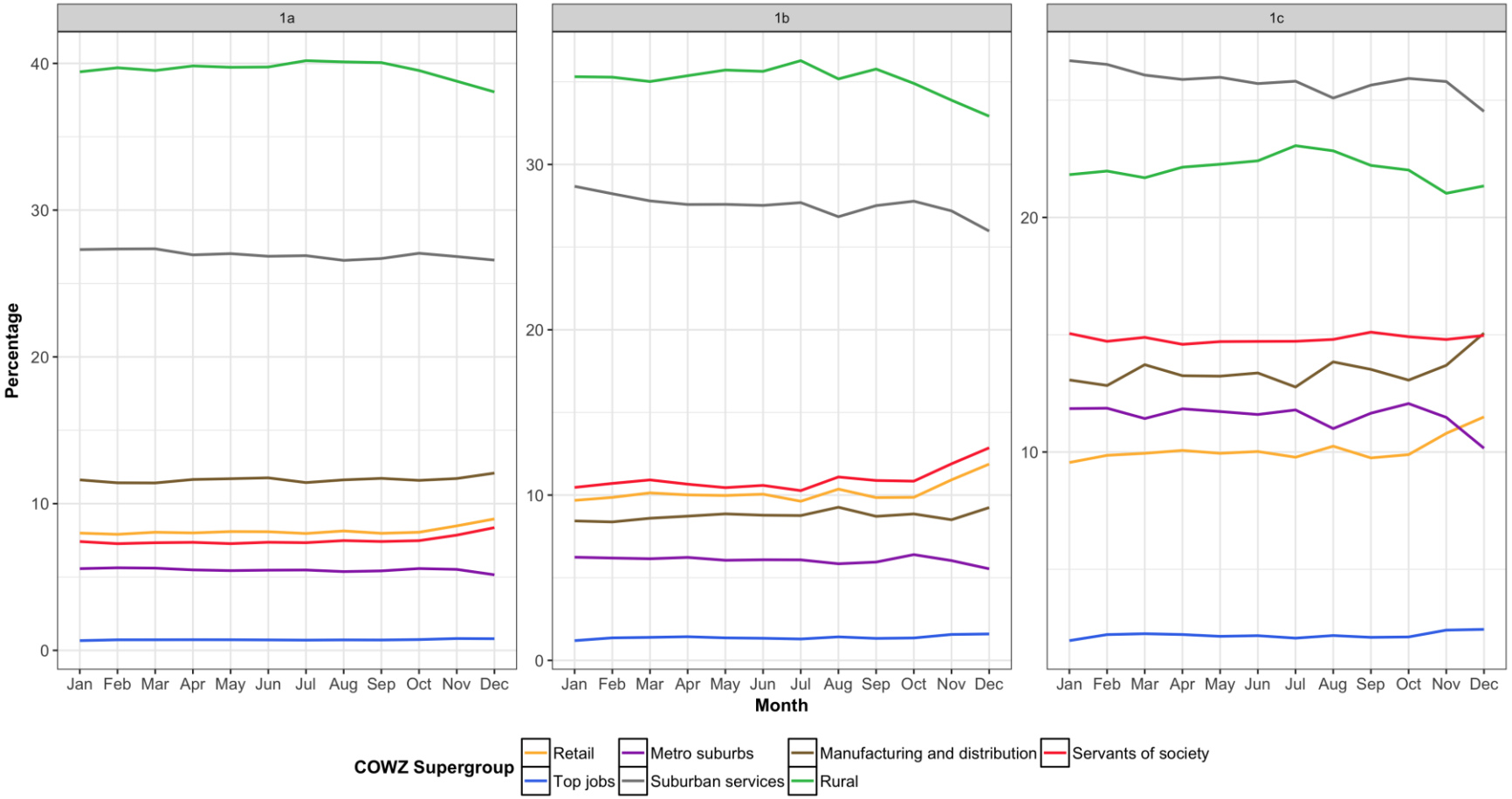


c. Supergroup 4 area visiting characteristics during weekdays and weekends (Note: Scales are varying to illustrate fluctuations within each Group).

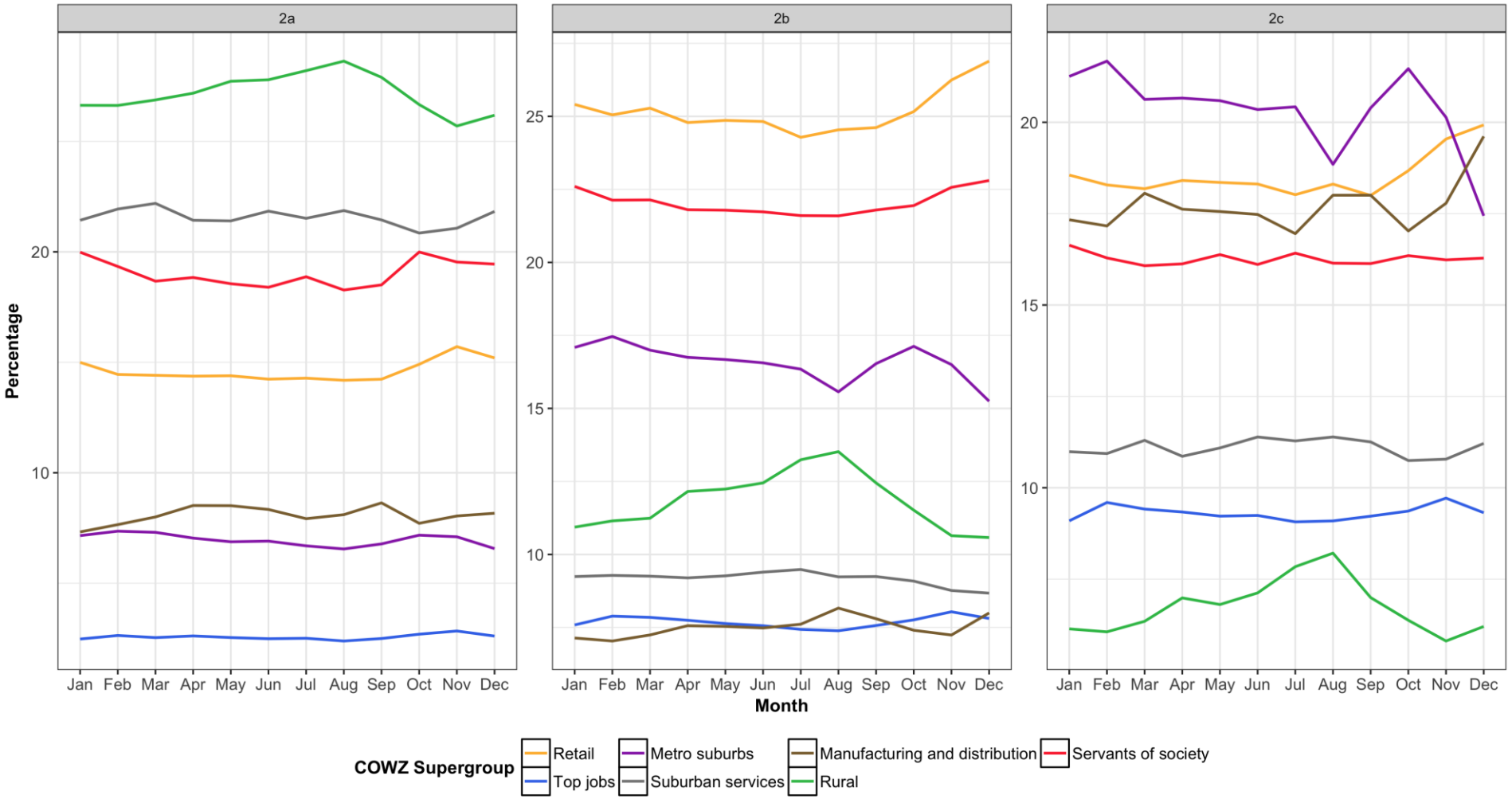


5.

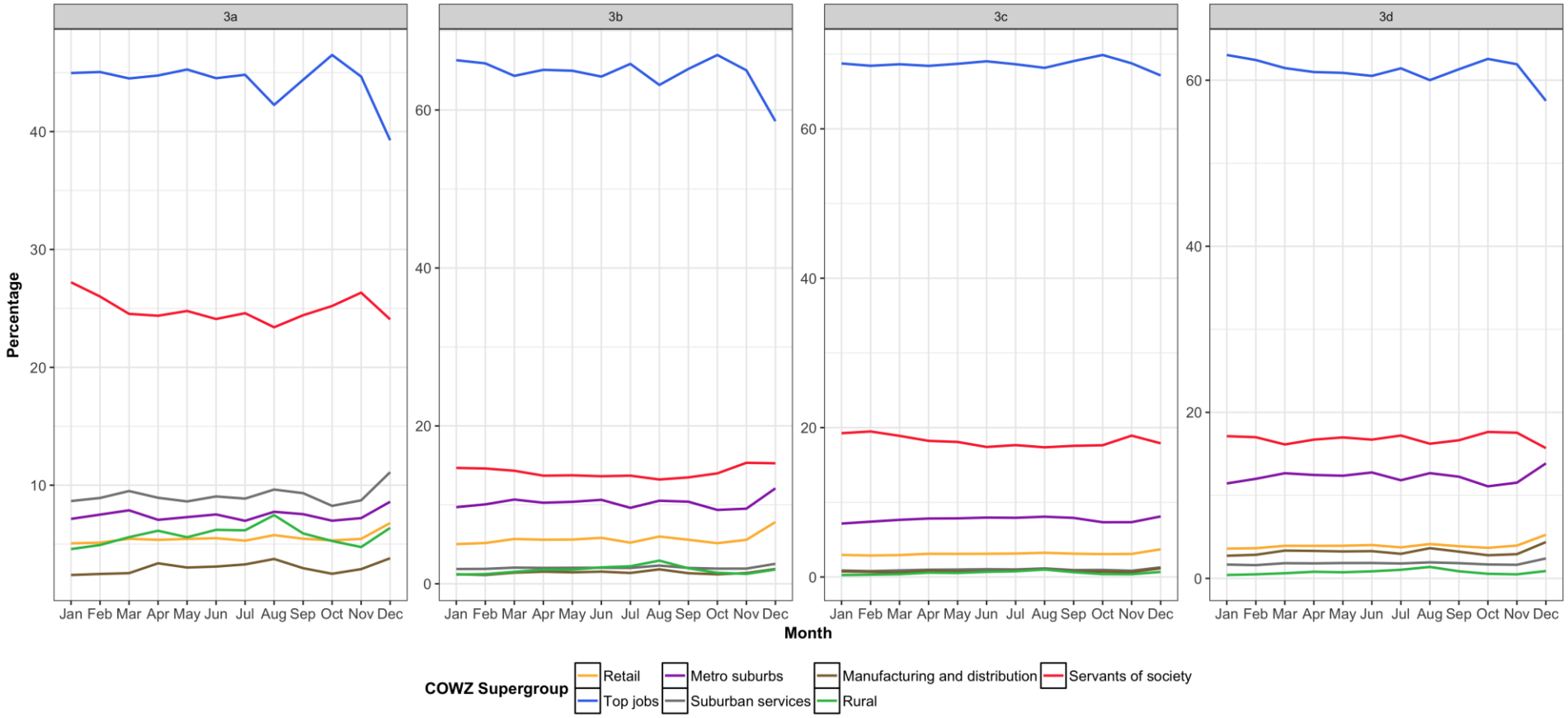
a. Monthly variation in area-visiting activity per COWZ Supergroup – Supergroup 1



b. Monthly variation in area-visiting activity per COWZ Supergroup – Supergroup 2



c. Monthly variation in area-visiting activity per COWZ Supergroup – Supergroup 3



6. Published Journal Papers

2018 Detecting Address Uncertainty in Loyalty Card Data. ***Applied Spatial Analysis and Policy***, (Lloyd, A., Cheshire, J.).