The Composition and Functioning of Retail Areas

Karlo Lugomer

Thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Geography
University College London

August 5, 2019
I, Karlo Lugomer, 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 work.

Signed
Abstract

Urban retail areas are prone to constant structural and functional changes. These dynamics are at the same time recorded and reinforced by technological advances that have supplemented traditional data sources such as censuses and small area surveys with more spatially and temporally granular consumer data. At the same time, modelling methods based on aggregate spatial interactions are being displaced by micro-location methods. A direct consequence of such developments is the emergent capability of retail analysts to track how characteristics of retail areas change on a more frequent basis, in something approaching real time. This thesis makes use of the Local Data Companys expansive and annually updated inventory of retail unit occupancy and a major Wi-Fi sensor footfall database which continuously records the number of passers-by at over 600 locations throughout Great Britain at five-minute temporal resolution. The sensors have been placed using an agreed sample design, which guided the focus of the research towards urban high streets and shopping centres. The analytical part of this thesis begins by describing the structural characteristics of the British retail economy and its changes throughout the post-recession years. The signals received by the Wi-Fi sensors are assessed, validated using robust data cleaning and ground-truthing methods to create reliable footfall estimates. Next, the temporal variations of footfall at different microsite locations are analysed, and the classification of diurnal human activity patterns is developed for sites across Great Britain. Finally, the two main datasets are combined to discover how local retail composition and footfall are interrelated, which results in the development of a unified functional classification of microsite locations. This thesis contributes to our understanding of how retail areas work and change, with the goal of developing recommendations for improving both their management and the operational and strategic performance of the businesses located within them.
Impact Statement

One of the most important findings of this thesis is that the Wi-Fi sensor footfall dataset and retail unit dataset can be used for the improved early-stage assessment of different candidate locations for future stores. This is important, as different types of microsite locations attract consumers at different times of the day and week. Based on the local retail composition and other geographical factors, retailers can draw inferences about the most likely footfall patterns to be encountered in a given town centre or shopping centre. While the results demonstrated that modelling footfall patterns using ancillary variables is not as straightforward, our understanding of footfall at different microsite locations has been significantly improved. The knowledge about existing footfall profiles can also be used for better management of opening times of retail businesses, staffing and marketing strategies.

The Wi-Fi sensor data are useful for footfall analytics purposes and despite technological drawbacks and emerging obstacles, this source of data is likely to continue being insightful. However, in the era of automated data collection, there is a growing concern that researchers and practitioners are turning away from field validation of the acquired data. This research proves that the external validation of automated measurements and big commercial datasets, in general, remains instrumental in producing reliable and robust research results.

Finally, from the academic perspective, the thesis also emphasises the importance of function, which has long been a research object in geography but, until recent technological advances, its research was to a great extent limited to static snapshots of space. The activity patterns should be viewed as a crucial layer of any store site assessment and town centre development policy. In light of findings presented in the thesis, the concept of town centre vitality becomes even more complex because, all other things being equal, a town centre with negligible footfall at particular days of the week may not be seen as attractive for both consumers and retailers, leading to potential demise. Aside from the apparent relevance of the dimension of time in retail geography, the importance of micro scale was reinforced. The thesis, therefore, advises increasing research focus on the functioning of microsite locations, as opposed to solely focusing on the town centre structure.
First of all, I would like to thank my primary supervisor Professor Paul Longley for his pivotal role in managing my research project and providing continued guidance and support. If it had not been for his patience, knowledge and crucial role in securing the funding, this PhD thesis would not have happened.

Gratitude also goes to my secondary supervisor Dr James Cheshire who provided me with important advice in the early phase of my PhD that helped to steer my subsequent research in the right direction.

I would also like to thank the ESRC for providing the funding and the Local Data Company for granting me access to data. Having the opportunity to use such big, varied and commercially valuable and insightful datasets was definitely something that I would hardly be able to get elsewhere. Special thanks are due to Matthew Hopkinson, a former managing director with whom I collaborated since my MSc dissertation project, to Ivana Mogin for her initial explanations of the technology and data acquisition methodologies, to Barnaby Oswald for his support during the project, and everyone else who has been involved with the SmartStreetSensor project.

I would also like to thank my friends and colleagues in the Chorley Institute, in particular to Balamurugan Soundararaj and Dr Roberto Murcio with whom I collaborated on the project, but also, to everyone else in CDRC for the memorable three years. Gratitude also goes to my friends from the Department of Geography of the University of Zagreb, and all other friends who have been with me throughout my PhD journey. Amongst them, I would like to especially thank Anna Labetski for proofreading particular parts of the thesis.

Last, but not least, I would like to thank my family and my dear Valentina. It is with their support that I managed to go through the most difficult times during my PhD and finally get to the goal I have always dreamt of. I hope this work will make everyone proud.
Thesis Outputs

Peer-Reviewed Conference Papers


- Lugomer, K., Soundararaj, B., Murcio, R., Cheshire, J. and Longley, P., 2017. Understanding sources of measurement error in the Wi-Fi sensor data in the Smart City. GISRUK 2017, Manchester

Book Chapters


Conference Talks

- International Conference on Geographic Information Science, August 2018, Melbourne, Australia

- GIS Research UK (GISRUK), April 2017, Manchester, UK

- Oxford Retail Futures Conference, December 2016, Sad Business School, Oxford, UK
Contents

1 Introduction .......................................... 18
   1.1 The Role of Retailing in the UK Economy and Recent Trends .......... 19
       1.1.1 From Greatness to Disaster .................................. 19
       1.1.2 Some of the Key Issues of the UK Retail Industry ............. 20
       1.1.3 Vacancy Rates and Their Variation ............................. 24
   1.2 The Role of Footfall ...................................... 26
   1.3 Thesis Outline and Objectives .................................. 28

2 The UK’s Changing Retail Geography .................. 30
   2.1 Changing Spatial Processes and Methodologies in Retailing: an Early Literature Review ............................................. 30
       2.1.1 The Urban to Suburban Shift of Retail Location ................. 30
       2.1.2 New Methodologies in Retail Location Analysis: Intuition Versus Modelling ..................................................... 31
       2.1.3 The Rise of Micro-Scale ........................................ 32
   2.2 The Spatial Framework of UK’s Retail Geography Research .......... 33
   2.3 The Ongoing Evolution of the UK Retail Structure .................. 40
       2.3.1 Development of New Retail Formats ................................ 41
       2.3.2 Shopping as Leisure ............................................. 43
       2.3.3 Effects of the Global Financial Crisis on UK Retailing .......... 45
       2.3.4 The Impact of E-Commerce ..................................... 46
       2.3.5 The Rise of Convenience Culture ................................ 49
   2.4 Chapter Summary ............................................. 50

3 Established and New Forms of Consumer Data .......... 52
   3.1 The Residential Geodemographics Tradition ........................ 53
       3.1.1 What Are Geodemographics? ...................................... 53
       3.1.2 Open Geodemographic Classifications and UK Census Geographies ......................................................... 54
       3.1.3 The Impact of Catchment Demographics on the Economic Health of Retail Areas ................................................. 57
3.2 Activity-Based Representations ................................................. 58
  3.2.1 Time Geography ............................................................ 59
  3.2.2 Representing Daytime Demographics Using Census Data ............ 60
  3.2.3 Other Contemporary Sources of Consumer Data ....................... 63
3.3 Big Data and the Smart City .................................................. 63
  3.3.1 Big Data and the Smart City: Definition of Concepts .................. 64
  3.3.2 Form Versus Function: the Analysis of Configurations and Activity Patterns ......................................................... 65
  3.3.3 Measuring Footfall: an Overview of New Technologies ............... 66
3.4 Some Negative Side-effects of Big Data: Do We Know Everything Now? . 69
  3.4.1 Uncertainty and Consumer Data ........................................ 70
  3.4.2 Ethical Considerations .................................................... 71
3.5 Chapter Summary ............................................................... 73

4 Data Description and Quality Considerations 75
4.1 Retail Occupancy Database ................................................... 75
  4.1.1 Retail Attribute Data ..................................................... 75
  4.1.2 Spatial Hierarchy of Retail Areas in Great Britain .................... 76
  4.1.3 Structural Characteristics of Retail Centres in Great Britain ......... 80
4.2 SmartStreetSensor Footfall Database ....................................... 85
  4.2.1 The SmartStreetSensor Project Background ............................ 85
  4.2.2 The Sample Design and Spatial Distribution of Sensors ............... 86
  4.2.3 Probe Requests ............................................................ 95
4.3 Addressing Sensor Measurement Errors .................................... 97
  4.3.1 The Sources of Measurement Error in the Wi-Fi Sensor Data .......... 98
  4.3.2 Internal Validation ....................................................... 100
  4.3.3 External Validation ....................................................... 103
  4.3.4 The Final Data Preparation Steps ..................................... 114
4.4 Chapter Summary ............................................................... 117

5 Temporal Classification of Footfall Patterns Across British Retail Areas119
5.1 Time Series Classification: Methodological Overview .................. 121
  5.1.1 Representing Time Series and Data Pre-Processing .................... 122
  5.1.2 The Choice of the Distance (Dissimilarity) Measure ................. 123
  5.1.3 The Choice of Clustering Algorithm ................................... 126
  5.1.4 The Choice of Clustering Parameters and Cluster Validity ........... 128
5.2 Classifying Microsite Locations ............................................. 129
  5.2.1 Practical Implementation of the Distance Measures and Clustering Algorithms ......................................................... 129
# 5.2.2 Exploratory Cluster Analysis

5.2.3 Description of Clustering Results

5.3 Cluster Stability over Fridays and Weekends

5.4 The Special Case of London

5.5 Chapter Summary

6 Non-Retail Drivers of Footfall Variability

6.1 How Does Footfall Vary Within and Between Retail Areas?

6.1.1 Dealing with Proximal Sensors

6.1.2 Footfall Homogeneity in Retail Areas

6.2 Footfall Variability and Transport Networks

6.3 Workplace Demographics and Footfall Patterns

6.3.1 Associations Between COWZ-UK and Footfall Temporal Profiles

6.3.2 Associations Between LWZC and London Footfall Temporal Profiles

6.4 Chapter Summary

7 Local Retail Geography and Footfall

7.1 Footfall Variability Between Different Store Categories and Retail Area Formats

7.2 The Spatial Conceptualisations of Retail-Footfall Relationships

7.3 Linking the Existing Retail Occupancy and Footfall Temporal Classifications

7.4 Unified Retail-Footfall Classification

7.4.1 Footfall Variables

7.4.2 Retail Variables

7.4.3 Cluster Analysis

7.5 Chapter Summary

8 Conclusions

8.1 The Composition of Retail Areas

8.2 The Functioning of Retail Areas

8.2.1 Wi-Fi Technology and its Role in Footfall Measurements

8.2.2 Footfall Temporal Classifications

8.2.3 Footfall and its Non-Retail Drivers

8.2.4 Linking Retail Composition and Function

8.3 The Case For a New Activity-Based Retail Geography

8.4 Possible Future Extensions of This Work

8.5 Closing Statements

9 Appendices

Bibliography
List of Figures

1.1 Changes in the national living wage and national minimum wage in the UK (2010 – 2019) .................................................................................................................. 23
1.2 Changes in vacancy rates across regions of Great Britain between the first half of 2017 and the first half of 2018 ................................................................. 25
2.1 Three main formats of retail areas ................................................................................ 35
2.2 Example of retail area boundaries in Birmingham: contrast between older, DCLG-adopted boundaries (Thurstan-Goodwin and Unwin, 2000) and newer boundaries (Pavlis et al., 2018) ......................................................... 39
2.3 The forces of change in UK retailing .............................................................................. 41
3.1 The 2001 Output Area Classification – Greater London ............................................. 57
3.2 Spatial relations between Output Areas and Workplace Zones. Example of: (a) an area with a high density of workplaces (Central London) and thus many WZs in each OA; and (b) a primarily residential area (Sutton) where multiple OAs are contained within one WZ ................................................................. 61
4.1 Rank-size distribution of retail centres in Great Britain (2017) ................................. 79
4.2 Optimal number of retail centre size clusters (elbow method) ................................. 80
4.3 The structure of retail centres in Great Britain in 2017: top-tier level (classifications) ....................................................................................................................... 82
4.4 Ten most common retail categories across British retail centres (2014) ............... 83
4.5 Ten most common retail categories across British retail centres (2017) ............... 84
4.6 The structure of retail centres by size (A=largest retail centres) ............................... 85
4.7 LDC sensors used for the data acquisition .................................................................. 86
4.8 Spatial distribution of sensors in Great Britain and Greater London ...................... 89
4.9 Frequencies of retail categories in the sample of microsite locations in Great Britain ................................................................................................................................ 91
4.10 The number of installed sensors across CDRC Liverpool’s retail centre sub-groups ....................................................................................................................... 92
4.11 Temporal variation in the captured devices in Bayswater, London, 20 – 30 August 2016 (five-minute intervals) ................................................. 96
4.13 The sources of error in the Wi-Fi sensor data .................................................. 99
4.15 Frequency of measurement errors (raw sensor counts compared to the manual counts) ........................................................................ 104
4.16 Sheffield fieldwork location A ..................................................................... 105
4.17 Sheffield fieldwork location B ..................................................................... 106
4.18 London fieldwork location ............................................................................ 107
4.19 Temporal distribution of different components of the sensor counts at Sheffield location A (averages) .................................................. 108
4.20 Correlation between raw sensor counts and manual counts at Sheffield location B ................................................................................. 109
4.21 Correlation between processed sensor counts and manual counts at Sheffield location B ................................................................. 110
4.22 Temporal variation in the measurement error - Sheffield location B .......... 112
4.23 Temporal variation in the measurement error - London location ............... 112
4.24 Temporal variation in the error of estimate and manual count .................... 113
4.25 Distribution of z-scores for all location-weekday-hour triplets .................... 116

5.1 Line generalisation methods as a form of data dimensionality .................... 122
5.2 Differences between the Euclidean distance and Dynamic Time Warping .... 125
5.3 Example of two time series clusters produced by AR(2) and ARMA(2,2) models, respectively ................................................................. 129
5.4 The clustering methodology ........................................................................... 132
5.5 The two-cluster solution obtained through Ward’s method ......................... 136
5.6 The three-cluster solution obtained through Ward’s method ....................... 137
5.7 The four-cluster solution obtained through Ward’s method ......................... 138
5.8 The five-cluster solution obtained through Ward’s method ......................... 139
5.9 The six-cluster solution obtained through Ward’s method ......................... 140
5.10 The seven-cluster solution obtained through Ward’s method ..................... 141
5.11 The eight-cluster solution obtained through Ward’s method ...................... 142
5.12 The nine-cluster solution obtained through Ward’s method ...................... 143
5.13 Cluster dendrogram of the eight-cluster solution ........................................ 144
5.14 The National Footfall Temporal Classification: reference plots ................. 146
5.15 National Footfall Temporal Classification – Friday footfall profiles .... 148
5.16 National Footfall Temporal Classification – Saturday footfall profiles .... 150
5.17 National Footfall Temporal Classification – Sunday footfall profiles .... 151
5.18 London Footfall Temporal Classification - six-cluster solution (Monday – Thursday) ................................................................. 154
5.19 London Footfall Temporal Classification - Central London microsite locations 157

6.1 Average five-minute footfall in Central London during workdays (7 am - 7 pm) between 9 January and 19 March 2017. .......................... 161
6.2 Index of relative weekend activity in Central London between 9 January and 19 March 2017 ................................................................. 162
6.3 Relationship between the number of sensor microclusters and frequency of the dominant footfall profile in retail areas in England and Wales .... 166
6.4 Distribution in the frequency of the prevalent footfall profile across retail area classes ................................................................. 167
6.5 Relationship between the number of bus stops within 500 meters and footfall volume ................................................................. 173
6.6 Relationship between the distance to the nearest bus stop and footfall volume 174
6.7 Relationship between the number of car parks within 500 meters and footfall volume ................................................................. 175

7.1 Monday-Thursday average footfall pattern observed in front of two different retail units in the same high street (High Holborn, Central London) (2015 - 2017 overall average) ................................................................. 189
7.2 Temporal profiles of the most common retail categories across Great Britain 190
7.3 Temporal profiles by retail area formats ............................................. 191
7.4 Buffer-based models of sensors’ retail environments ............................ 193
7.5 Retail area-based models of sensors retail environments ...................... 195
7.6 Absolute frequency of retail categories with a particular number of retail units in the 500-meter retail environment sample of sensor microsite locations 201
7.7 Variation in the proportions of retail categories across retail environments (500 m network distance) of the sensor microsite locations .......... 206
7.8 Dendrogram of the chosen five-cluster solution obtained through Ward’s method ................................................................. 210
7.9 Pen portraits of the five-cluster solution obtained through Ward’s method 212
7.10 The spatial distribution of retail-footfall clusters across Great Britain (excluding London) ................................................................ 215
7.11 The spatial distribution of retail-footfall clusters in Greater London ........ 216
7.12 The spatial distribution of retail-footfall clusters across Central London .. 217
9.1 Sample design for sensor locations: England . . . . . . . . . . . . . . . . . 242
9.2 Sample design for sensor locations: London . . . . . . . . . . . . . . . . . 243
List of Tables

3.1 COWZ-UK supergroups ................................................. 62
4.1 Metadata for the Local Data Companys 2017 retail dataset .......... 77
4.2 The number and average size of the four hierarchical categories of retail centres in Great Britain .................................................. 80
4.3 The structure of retail centres in Great Britain (2014 – 2017) ........ 81
4.4 CDRC Liverpool’s retail area classification pen portraits of the selected groups (CDRC Liverpool, 2018) .................................................. 93
4.5 Average measurement errors for sensors in Sheffield (7 Sep – 9 Sep 2016) and London (12 Dec 2016 and 19 Jan 2017) ......................... 114
5.1 The amount of clustering structure found in different hierarchical clustering algorithms with: (a) Euclidean distance, (b) DTW (warping set to one hour) used as distance measures .............................................. 134
5.2 Number of clusters at which too similar pairs of profiles start to appear in the solution ......................................................... 134
5.3 The breakdown of cluster cases (eight-cluster solution) ............... 145
5.4 The change of cluster membership between Monday-Thursday and Friday ................................................................. 149
5.5 The breakdown of cluster cases - Saturday .................................. 150
5.6 The breakdown of cluster cases - Sunday ..................................... 151
5.7 The change in the comparable cluster membership from Saturday to Sunday ................................................................. 152
5.8 Comparison of temporal clusters occurrence on (a) London level and (b) national level .......................................................... 153
5.9 The breakdown of clusters in the London Footfall Temporal Classification ................................................................. 156
6.1 One-way ANOVA table: the relationship between street hierarchy and footfall volume ......................................................... 168
6.2 Mean hourly footfall by street hierarchy ..................................... 169
6.3 Test of the statistical significance of the association between the OSM road network hierarchy and footfall temporal profiles ........... 170
6.4 Location quotients displaying the relative distribution of temporal clusters on each type of the street .......................................................... 172
6.5 Distribution of COWZ-UK demographic supergroups around sensor microsite locations ................................................................. 177
6.6 Location quotients displaying the relative distribution of temporal clusters in each COWZ supergroup ........................................... 179
6.7 Location quotients displaying the relative distribution of temporal clusters in each COWZ group ....................................................... 182
6.8 Location quotients displaying the relative distribution of temporal clusters for London in each LWZC supergroup .......................... 185

7.1 Cluster groups and subgroups from CDRC Liverpool’s retail area classification that are covered by at least five sensors .................. 196
7.2 Location quotients demonstrating the relative likelihood of an occurrence of a particular Monday-Thursday temporal profile in a particular retail classification subgroup (England and Wales) .................. 197
7.3 Location quotients demonstrating the likelihood of occurrence of a particular Saturday temporal profile in a particular retail classification subgroup (England and Wales) .......................................................... 198
7.4 Highly correlated pairs of retail variables ........................................ 206
7.5 Results of the cluster validation: top three cluster solutions obtained by different clustering methods and validated by different validation methods ................................. 208
7.6 Absolute and relative frequencies of retail-footfall clusters .................. 211

9.1 The number of installed sensors in Local Authorities and Council Areas in Great Britain (23 October 2017) .............................................. 233
9.2 Fieldwork measurements and calibration for the sensor locations in London and Sheffield ................................................................. 236
9.3 Distribution of sensors across different types of retail centres (CDRC Liverpool, 2018) ................................................................. 237
9.4 Test of the statistical significance of associations between the COWZ supergroups and temporal profiles ........................................... 239
9.5 Test of the statistical significance of associations between the COWZ groups (lower-tier level of the classification) and temporal profiles ................................. 239
9.6 The test of the statistical significance of associations between the LWZC supergroups and temporal profiles ................................ 240
9.7 The test of the statistical significance of associations between the LWZC groups (lower-tier level of the classification) and temporal profiles ................................. 240
9.8 The breakdown of the footfall temporal clusters based on the retail area format 241
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>CDRC</td>
<td>Consumer Data Research Centre</td>
</tr>
<tr>
<td>COWZ</td>
<td>Classification of Workplace Zones</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>Density-Based Spatial Clustering of Applications with Noise</td>
</tr>
<tr>
<td>DTW</td>
<td>Dynamic Time Warping</td>
</tr>
<tr>
<td>DWT</td>
<td>Discrete Wavelet Transform</td>
</tr>
<tr>
<td>ESRC</td>
<td>Economic and Social Research Council</td>
</tr>
<tr>
<td>GB</td>
<td>Great Britain</td>
</tr>
<tr>
<td>GDPR</td>
<td>General Data Protection Regulation</td>
</tr>
<tr>
<td>iOS</td>
<td>iPhone Operating System</td>
</tr>
<tr>
<td>LAD</td>
<td>Local Authority District</td>
</tr>
<tr>
<td>LDC</td>
<td>Local Data Company</td>
</tr>
<tr>
<td>LFTC</td>
<td>London Footfall Temporal Classification</td>
</tr>
<tr>
<td>LOAC</td>
<td>London Output Area Classification</td>
</tr>
<tr>
<td>LWZC</td>
<td>London Workplace Zone Classification</td>
</tr>
<tr>
<td>MAC</td>
<td>Media Access Control</td>
</tr>
<tr>
<td>MAD</td>
<td>Median Absolute Deviation</td>
</tr>
<tr>
<td>NFTC</td>
<td>National Footfall Temporal Classification</td>
</tr>
<tr>
<td>OA</td>
<td>Output Area</td>
</tr>
<tr>
<td>OAC</td>
<td>Output Area Classification</td>
</tr>
<tr>
<td>OSM</td>
<td>OpenStreetMap</td>
</tr>
<tr>
<td>OUI</td>
<td>Organisationally Unique Identified</td>
</tr>
<tr>
<td>PAM</td>
<td>Partitioning Around Medoids</td>
</tr>
<tr>
<td>PII</td>
<td>Personally Identifiable Information</td>
</tr>
<tr>
<td>PIP</td>
<td>Perceptually Important Points</td>
</tr>
<tr>
<td>UCL</td>
<td>University College London</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>WZ</td>
<td>Workplace Zone</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Shopping and leisure are important day-to-day individual and household activities and have thus become deeply enrooted into the global and local economies and people’s lifestyles. Retail units can be found all across urban and rural areas, which is not surprising given the omnipresent consumer demand they have to meet. However, some parts of space are especially known for their economic attractiveness and clustering of stores and leisure units. These areas of higher concentration of retail activity, such as high streets and purposely built out-of-town and in-town shopping centres and retail parks are termed retail areas and constitute a spatial framework of this research. What makes retail areas an interesting and important academic and professional research topic, is the fact that they are not spatially uniform nor static entities. The composition of stores varies between different retail areas, as local consumer demand which is inherently tied to the local residential and workplace demographics and socioeconomic characteristics, also varies spatially. However, as this research aims to demonstrate, it does not suffice to focus the retail geography research solely on how different places are similar or different to one another at any given point in time, but there is also an emerging need to acknowledge the significance of investigating short-term and long-term changes that take place in retail areas.

What is the composition of British retail areas and how does it change over different time horizons? When do people visit retail areas and how are those activity patterns related to local retail geography, demographics and other characteristics of space? The answers to these questions are sought in all subsequent chapters. In doing so, the combination of traditional data acquisition methods and new technologies are used, and their role in the improvement of our understanding of retail areas is critically assessed.
1.1 The Role of Retailing in the UK Economy and Recent Trends

The importance of the retail sector for the UK economy is substantial. In 2017 it employed 2.8 million people or 9.5% of the UK total, and its economic output as measured by GVA\(^1\) amounted to £92.8 billion or 5% of the UK total (Rhodes, 2018). A year later, the number of retail businesses equalled 319,125. Retail sales in the UK in 2017 were worth £395 billion (Rhodes, 2018).

These figures highlight the fact that retailing plays an enormous role in the UK economy, and that any widespread problems that hit the retail sector, are likely to affect other sectors or the economy as a whole in a negative manner. Some common problems that have recently put pressure on the retailers are briefly visited in the next section.

1.1.1 From Greatness to Disaster

Prior to the 2008 global financial crisis, the retail economy saw an extended period of economic growth, spanning over several decades. It was a period in which even those retailers with theoretically questionable future found a way to survive in the expanding market. However, when the economic downturn came, retailers were slashed. Many of them suffered substantial losses, while some of them were completely erased from the market (Portas, 2011), an event that somewhat mirrored what happened with weak start-ups during the 2000 dot-com or the Internet bubble burst. The decline of the UK retail economy was widespread, but some retail categories, particular chains, retail area types and regions were hit more than others.

For instance, retailers that sell products that were itself being increasingly digitised and thus threatened by the online competition, such as computer games shops, music shops, booksellers and newsagents, recorded a 13% fall in high street unit numbers between 2011 and 2013 (Grimsey Review, 2013). With a combined net loss of over 500 retail units across the country (13% fall for multiples and 6% fall for the independents), fashion retailers also took a big hit (Grimsey Review, 2013).

According to Local Data Company (LDC)\(^2\) data reported in the Grimsey Review (2018), 2,214 (-0.89%) retail and leisure units have been lost from 1,048 towns with 50 or more units between 2013 and 2017 (Grimsey Review, 2018). Furthermore, 18% of towns lost 10% or more of their chain retailers, and the magnitude of chain retailers closures was at highest levels since 2010 (Grimsey Review, 2018).

\(^1\)Gross Value Added, which is a measure similar to GDP.
\(^2\)Headquartered in London, LDC is a company concerned with both automated and manual collection of data on the composition, performance, locational suitability and footfall of retail and leisure units throughout the UK (Local Data Company, n/d).
In the period from the 2008 global financial crisis until 2019, some famous retail chains were not only forced to shut down many of their units, but they went bust, a prime example being Woolworths Group which entered administration in the UK in January 2009. Known for its pioneering role in the pound shop category, Woolworths fell victim to its own inability to respond to the emerging value-conscious customer base that was, in turn, captured by other new pound shop chains (Portas, 2011).

While the crisis was an important catalyst of the fall of some of the prominent retailers, this was hardly the only relevant factor. Even in the post-crisis era, Debenhams department store chain has struggled under onerous leases and debts that have reached £1.1 billion (Retail Gazette, 2019). This culminated in Debenhams recently announcing the closure of 50 of their stores, with 22 of them having been named in the press releases in April 2019 (BBC, 2019). A part of the reason for this radical move which would close almost a third of Debenhams’ stores in the UK lies in the fact that Debenhams heavily expanded their physical presence on the high street during the times when many customers started switching to online sales (BBC, 2019). This was a costly investment blunder, and consequences were devastating.

Leisure sector has also seen notable closures in recent years. For instance, Jamie Oliver Restaurant Group has been taken into administration in 2019. The problems for the celebrity chef’s restaurant chain multiplied after the Brexit vote, which caused depreciation of the pound and the consequential rise in prices of the imported ingredients (Fortune, 2019). Moreover, too high prices of food served in the chain, the rise of rent prices and the oversupply of the restaurant market, in particular, Italian food and burger chains, made the survival of Jamie Oliver brand on the high street nearly impossible (Fortune, 2019; Szmigin, 2019). Within the Italian food restaurant chains, Strada had to shut a third of its units at the start of 2018, Prezzo and Carluccio’s underwent restructuring, and same was true for the burger chains Byron’s and Gourmet Burger Kitchen (Fortune, 2019; Butler, 2018).

1.1.2 Some of the Key Issues of the UK Retail Industry

The aforementioned examples of chains that experienced a period of hardship typically did have some specific reasons for their respective declines. However, one can still identify a general set of factors that adversely affected UK retail businesses throughout the past ten years.

Rising Business Rates

Business rates (also referred to as National Non-Domestic Rates) are a tax on non-domestic, i.e. business properties, which is set by the national government and collected by the local authorities. They are calculated based on their ratable value, which is basically the annual
rent that the property could have been let for on the open market (Dudley Council, 2019). Their value changes based on the measures of inflation (Accounts and Legal, 2018).

The growth of business rates has been a well-covered issue, especially in the press (Butler, 2018; Eley, 2019; Stevens, 2018). According to Stevens (2018), business rate bill for the 2019/2020 was expected to increase £728.20 million, with London taking the biggest hit. Small retailers are particularly struggling to pay their bill, as this type of tax hits them before they make any turnover, let alone profit (Accounts and Legal, 2018). This is especially troublesome for the entire retail industry and the national economy because small businesses have an instrumental role in keeping the employment levels at record highs (Accounts and Legal, 2018).

While online retailers were enjoying by far lower rates compared to the brick-and-mortar businesses, an increase in their rates has already been scheduled for 2021 due to the rising rent of the distribution centres upon which online retailers base their operations (Eley, 2019). This was fueled by the expansion of value retailers and the consequential rise of rents of the warehouses located near motorway junctions (Eley, 2019).

On the other hand, it is worth stressing that despite the adverse impact of soaring business rates, the UK’s top retail centres remained strong in 2019, and this is especially obvious when it comes to retail centres in the areas where rent and business rates rose dramatically, such as London’s Knightsbridge and Chelsea (Eley, 2019; Harper Dennis Hobbs, 2019). The attractiveness of a retail destination was therefore still primarily associated with the wealth of the local population (Harper Dennis Hobbs, 2019).

Nevertheless, business rates are seen as an archaic tax system that has long been out of sync with the value that the local government provides to retailers in return for the large tax bill (Grimsey Review, 2018). It insufficiently follows the changes in the overall economy (Grimsey Review, 2013), and there have been incentives from the retail community to make it a variable form of taxation, by conducting more frequent property valuations and refining the rates in line with the prevalent market conditions (Grimsey Review, 2013). In addition, a newer edition of the Grimsey Review (2018) recommends business rates to be completely replaced with a more meaningful form of tax, such as property value tax or sales tax.

Effects of Brexit

After British voters chose to leave the European Union in June 2016, this sparked a downturn in consumer confidence and caused more general economic problems that also affected retailing.

There is evidence to suggest that Brexit is driving inflation in the UK, import prices have risen given the fall in the pound with prices rising faster than wages, causing households to tighten their belts on spending, especially on big-ticket items (Accounts and Legal, 2018). Tariffs of five percent or more on imported goods, poor product availability, wastage and a
labour shortfall were all seen as major potential issues by Danny Sperrin, business manager at management consulting group Newton (Nazir, 2019). Consumers are also becoming more cautious about spending, and this was reflected in the Christmas 2018 sales, which happened to be the worst since the global financial crisis a decade ago (FTI Consulting, 2019). Spending may fall even further in post-Brexit era, as food prices are expected to see the highest rise, going up to between 45 and 50 percent, and this could be even exacerbated by the fact that, since half of the food UK’s food comes through Dover and Folkestone, delays of imports and disruption of transport services may take an even bigger toll on retail prices (Nazir, 2019). Problems are expected not only on the import side but also on the export side, as leaving the Union, especially without a deal would mean increased time and costs in administration and customs clearance, which would, in turn, decrease the demand of UK goods and services (FTI Consulting, 2019). The UK would also lose influence on EU policy, standards and regulation setting, while increased compliance costs and product lead times for UK retailers would make them less competitive (FTI Consulting, 2019).

In terms of labour access and costs, UK retail sector, which currently relies upon over 300,000 EU immigrants, would be shaken by the inability to employ more cheap workforce and it would be forced to hire more domestic workers in order to ensure continued operations, leading to the rise in the average labour costs (FTI Consulting, 2019).

It should also be noted that there are considerable differences in the consequences of Brexit under the two main scenarios: no-deal Brexit and a reached deal scenario. Under the no-deal Brexit, consequences may be more severe with little time for retailers to react and adapt to the increased supply costs, whereas if the deal is reached, a pound may rebound, thus reducing the adverse effects on prices (Nazir, 2019).

**Minimum Wage Increases**

The minimum wage hike is known to have adverse effects on retail employment, and also, on total state employment growth (Partridge and Partridge, 1999). It affects shift patterns, reductions in supplementary benefits, and encourages the introduction of new technologies to increase efficiencies (Addleshaw Goddard, n/d), for example, by introducing the self-checkout machines.

In the UK, there are two main types of minimal wages: national living wage (NLW) that applies to the workers aged 25 and above, and national minimum wage (NMW), which applies to the workers aged between 21 and 24. Both of them grew considerably over the past decade (Figure 1.1). National living wage grew from £5.93 in 2010 to £8.21 in 2019, making it 4.27% annual growth rate. National minimum wage, on the other hand, rose from £5.93 to £7.70 in the same period, resulting in 3.32% annual growth rate (Minimum Wage, n/d).
Minimum wage increases were forecast to lift the retail prices by about 1.1% per year between 2016 and 2020. Prices of services with high labour content, such as courier deliveries and even click and collect services were expected to increase particularly (Bamfield, n/d). Moreover, some retailers were expected to cut more jobs in light of the increasing minimum wages, and this, in turn, accelerates the store closures, a process that has long been in full swing on the UK high street (Bamfield, n/d). This forecast turned out to be valid, as the twelve-month period up to April 2018 saw over 3,700 job cuts in the retail sector (Addleshaw Goddard, n/d). It can be concluded that while not decisive, minimum wage increases may accelerate the store closures, which in turn encourages the further decline of high streets over time.

**Effects of Competitive Pressures**

British retail areas evolved substantially at the end of the 20th century, with out-of-town retail provisions growing and causing town centres to decline (Wrigley and Lambiri, 2015). This spatial process, which will be covered in more detail in Chapter 2, led to closures of some high street stores.

The process was supplemented by the rising competitive pressures from new channels of retailing, such as e-commerce. Internet sales, measured as a proportion of overall retail sales in the UK surged from about 5% in 2008 to around 18% in 2018 (Rhodes, 2018). This leap was powered by the rising smartphone ownership in the corresponding period, as well as changing consumer habits (Grimsey Review, 2018). Some of the bricks-and-mortar, i.e. physical stores could not cope with the fact that many products could now
be more conveniently purchased online from the competitors that invested more into their respective e-commerce departments, new distribution services and specialised novel forms of marketing.

Development of online retailing did initially provide a great opportunity to increase sales volumes of particular types of products, and they also encouraged the development of warehousing and road transport sectors in order to meet consumers’ rising demand for home deliveries (Grimsey Review, 2018). Unfortunately, problems have also struck a hitherto flourishing logistics sector, as retailers have reacted to the pressures they face by shifting at least some the burden down the supply chain (Grimsey Review, 2018).

1.1.3 Vacancy Rates and Their Variation

By now, it is clear that retail sector goes through periods of ups and downs. But how do we quantify how well the retail areas are doing? While there is no universal and unbiased indicator for this purpose, several commonly used measures emerged in the retail industry. These include simple counts and rates of openings and closures in the area of interest, composite indicators measuring the economic health or vitality of retailing using different variables, vacancy rates and others (Local Data Company, 2018).

Defined as a proportion of unoccupied units or floor space in the defined area (Lugomer, 2015), vacancy rate is a popular metric, widely reported by the general and retail press (Retail Week, 2013; BBC, 2014; Guardian, 2019; Property Week, 2019). At the same time, it is a simple measure, and as such, it has been criticised for being a poor indicator when used in isolation. As Harper Dennis Hobbs (2019) notes, new projects usually require existing tenants to move out before construction can take place, so the raw vacancy rate figure can penalise a centre in the process of regeneration. In addition, vacancy rates are normally not standardised based on floorspace due to the lack of corresponding data, which may result in misleading conclusions. However, they still offer some basic insight into where retail is going, which is definitely a step forward from indulging in pure guesswork.

In general, shop vacancy rates in Great Britain grew nearly fourfold between the beginning of 2008 and February 2012, when they peaked at 14.6%, and subsequently declined (Grimsey Review, 2013). This trend continued throughout 2015–2017 and 2017 saw the lowest vacancies since 2009, coming down to 10.9% in the first quarter of 2017 (Local Data Company, 2018).

However, the changes in the economic health of retail areas in any period have not been structurally and spatially uniform, and they depended heavily on the local economic conditions, formats of retail areas and the strategies of key players in the retail economy which sought to open or close their units in different regions of the UK. That being said, smaller retail areas were traditionally more badly hit by closures. Comparing the data on vacant units across retail areas acquired by the Local Data Company (LDC) for the first
half of 2017 and the first half of 2018, it is evident that the vacancy rates of small retail areas (< 200 retail units) increased from 8.7% to 9.1%, whereas vacancy in the medium-sized and large retail areas remained more stable, reflecting the preferences of retailers to locate in the areas with a larger local demand and a higher chance to attract their target geodemographic groups (Local Data Company, 2018). Also, out of the three main formats of retail centres (town centres/high streets, retail parks and shopping centres\(^3\)) investigated by LDC, retail parks have had the lowest vacancy rates (5.6%) and have seen the largest relative increase, whereas shopping centres had the largest vacancy rates (13.1%) (Local Data Company, 2018).

From a regional perspective, vacancies of almost all regions remained stable in the 2017-2018 period, or they even increased, except for Scotland (a decline from 11.9% to 11.3%) and West Midlands (a decline from 14.2% to 13.7%) (Figure 1.2).

![Figure 1.2: Changes in vacancy rates across regions of Great Britain between the first half of 2017 and the first half of 2018](image)

Source: Local Data Company (2018)

These examples reveal the existence of commercially interesting trends retail landscapes have experienced in recent years. Since vacancy rates are frequently used in retail analytics as a proxy of health of the retail centre (Genecon, 2011; Economy Committee of the Greater London Authority, 2013; Morgan Stanley Research, 2014), either alone or together with some other indicators, it may be concluded that health of retail areas exhibits spatially variable trends, the identification and understanding of which is instrumental in the better

\(^3\)A closer look at the differences among these formats will be given in the next chapter.
retail planning and town centre management. How well a certain retail establishment or agglomeration of stores is doing, depends on a number of locational, demographic and socioeconomic factors that will be discussed in the upcoming chapters.

1.2 The Role of Footfall

Functional characteristics of urban areas and local activity patterns are some of the main factors that dictate the suitability of a site for a store location on very granular scales. Footfall\(^4\) plays an important role here, serving as a proxy for the activity and vitality of retail areas.

A high street footfall can easily fall despite the population growth, and this was seen to be the case during the 2008 economic downturn, as Genecon (2011) reports a 10% footfall decline in the UK high streets, excluding Central London (2008–2011). As demonstrated earlier, vacancy rates were on the rise during the same period, and footfall happens to be tightly associated with them.

According to Helen Dickinson, chief executive of the British Retail Consortium (BRC), this is because the empty shop fronts may deter shoppers from an area, which, in turn, causes further losses for the local retailers, potentially even more closures and further rise of vacancy rates, creating a downward spiral (Fish, 2019). This only confirms that footfall, as such, can be used as a more accurate and up-to-date measure of the potential customer base that uses the services of retail areas on a daily or a weekly basis. In other words, footfall research can be seen as a prerequisite for any sort of serious town centre planning or the locational analysis conducted by individual businesses.

How does the volume of footfall differ from location to location (region to region, city to city, neighbourhood to neighbourhood, street to street, or even building block to building block)? How does it vary throughout the day? What about the weekly or seasonal variations? Would a store benefit from changing its opening times to make the best out of the daily peaks times? To what extent do less frequent, but locally popular events alter the overall activity of a certain neighbourhood? How can we measure footfall and use it to better understand the functioning of retail areas?

Unprecedented technological growth at the end of the 20\(^{th}\) century and the dawn of the New Millennium have influenced our ability to try and answer those questions through the use of many conventional and newer alternative data sources, some of which will be

\(^4\)While it is a term frequently used for the number of people that enter a retail unit, in this thesis, it is used as the number of people that can be counted at a given point in front of a store. Such a focus will become more clear in the later chapters. However, for now, it suffices to say that measuring footfall in front of a store enables comparison with the actual sales of the same store. This provides retailers with an opportunity to compute conversion rates and better understand what proportion of consumers are, in fact, attracted by their store and purchase their products.
addressed in Chapter 3.

One of those examples has become particularly interesting in recent years and is a cornerstone of this thesis - the Wi-Fi sensors. They are capable of capturing probe requests from smartphones and other devices and, therefore, record temporal variations of activity in their immediate vicinity. Their popularity in academic literature and business strategies has grown mainly due to increasing smartphone ownership, which, in turn, enabled the exploitation of many advantages such type of dataset brings. Data tools and applications installed on smartphones enable us to acquire data to investigate their owners’ location and make inferences about the spatial and social aspects of their everyday life (Chittaranjan et al., 2013).

According to the Spring 2015 Global Attitudes survey, smartphone ownership rates have skyrocketed in many countries since 2013, and while particular growth has been recorded in developing economies, most developed countries remain on the top. For example, the percentage of German adults who have, at the time of the survey, owned a smartphone amounted to 60%, the corresponding figure in the UK was 68%, in the US 72%, in Australia 77% and in South Korea 88% (Poushter, 2016). The newer generations comprise significantly more smartphone owners, with 91% of the UK population aged 18-34 owning one (Poushter, 2016). According to Office for National Statistics, 97% of the population in the age group 16-24 use mobile devices (phones, laptops, tablets) to connect to the Internet ‘on the go’ and 33% of those aged 65 years and over, which is a significant increase from the previous years (Office for National Statistics, 2016b). The numbers clearly indicated that this type of technology would be used even more in the future and that the age-gap is likely to diminish further.

The implications of such technological advances for retail geography research are substantial. Novel data acquisition technologies such as GPS, Bluetooth and Wi-Fi, enable an automated, less expensive and a continuous collection of the vast amount of commercially valuable data on where people go. These advances enable retailers to analyse what is happening at the micro-scale, and not only at the national, regional or urban scales. In addition, apps used by smartphone owners collect attribute data about them, such as their demographics, lifestyles, preferences of certain products, all of which can be utilised for improving our understanding of consumer behavior and the economic state of the retail areas from an academic perspective, as well as the corporate decision-making processes from industrial perspectives.

At the same time, utilisation of such novel technologies in retail analytics comes at the right time, as many ongoing problems in the retail sector require solutions. Without accurate, reliable and continuously collected and updated data, such potential solutions would remain speculative or even impossible.
1.3 Thesis Outline and Objectives

This research contributes to three main streams:

1. a better understanding of the data quality issues and uncertainties associated with Big Data, in particular, data acquired by the Wi-Fi sensors;

2. spatio-temporal analytics of human activity patterns in urban areas;

3. understanding the structural and dynamical characteristics of retail areas.

An important note to make at the beginning is that this thesis was written as part of the wider project called The SmartStreetSensor project, which was conducted in collaboration with the industrial partner and data provider Local Data Company. As part of this academic-industrial partnership, several members of academic staff at the University College London (UCL), as well as the employees of LDC contributed to the development of the overall project. Nevertheless, this thesis is an original piece of scientific work, and as such, it tackled specific problems that were not tackled by other contributors.

Where a partial overlap with the works of other PhD students, postdoctoral researchers, professors and data scientists existed, this was explicitly noted, both in the brief overview of the thesis structure that follows, and where appropriate later in the thesis. It should be noted, though, that this was constrained to the specific sections of Chapter 4.

The thesis begins with two literature review chapters on past and current research in retail geography, geodemographics, geographic information science and data science that is deemed of particular relevance to retail analysis.

In Chapter 2, a more detailed overview of the changing nature of UK retail geographies is given, as well as how the retail area definitions and delineations and retail data analysis have been improved throughout the last few decades.

The review then goes on to describe traditional and new forms of consumer data in Chapter 3. At the beginning of the chapter, the emphasis is placed on the role of the traditional residential census data, how they have been successfully utilised for geodemographic segmentations; and the how their shortcomings were addressed over time by introducing the more appropriate datasets and underlying geographies such as workplace zones. The second part of the chapter discusses what changes the Smart City paradigm coupled with the adoption of the aforementioned new technologies brought into the consumer research. Underlying technologies and their respective strengths and weaknesses are laid out, with particular emphasis on Wi-Fi.

Next, in Chapter 4, the two main datasets used in the analytical part of the work are introduced: the LDC’s national retail unit dataset and LDC’s SmartStreetSensor footfall dataset containing continuous measurements of footfall across hundreds of locations in Great
Britain. Developed by other collaborators before the start of this PhD, the sample design which governed the choice of the Wi-Fi sensor microsite locations to be used in the analysis of footfall is laid out.

Finally, data quality issues pertaining to the footfall dataset are thoroughly described and addressed to make sure that footfall estimates are reliable and as close to the ground-truth as possible. This was structured into the sections on internal and external validation. An internal validation methodology was not an output of this thesis, but it was mostly a product of LDC’s in-house team, refined through a series of consultations with UCL researchers, including myself. The agreed methodology was then applied in this research, with any analytical results arising from it being an original product of this thesis. Similarly, designing and conducting special field visits for the purpose of data accuracy improvement in the External Validation subsection 4.3.3 and applying final data cleaning steps in The Final Data Preparation Steps 4.3.4 subsection were the original contents of this PhD research.

After introducing the datasets and their strengths and weaknesses, the footfall estimates are used to devise a temporal classification of diurnal footfall patterns for the chosen set of microsite locations (Chapter 5). In other words, the objective is to find out whether microsite locations differ in how footfall is temporally distributed at them and if so, what types of footfall patterns can we discern at the national level.

In Chapter 6, some relevant drivers of footfall volume and patterns are examined, and a series of hypothesis tests are run to inspect whether there exist statistically significant relationships between them and footfall. The variation of footfall within retail areas is also examined to answer the recurring question of whether retail areas are homogeneous in terms of activity patterns generated by consumers.

In Chapter 7, the analysis goes on to bring the two main datasets together by inspecting the relationships between the local retail geography and characteristics of footfall. Since those two datasets were built at different times and for different purposes, it is not as straightforward to link them, so several conceptual models for their linkage are presented. The final product of this chapter is the unified retail-footfall classification which extracts the main groups of microsite locations that share common characteristics of retail composition and function.

Finally, conclusions are drawn in Chapter 8, where the research results are further highlighted, with particular emphasis on the value of the Wi-Fi analytics and retail unit data for retail planning and decision-making.
Chapter 2

The UK’s Changing Retail Geography

This chapter defines some of the key retail geography terms that will be used throughout the rest of the thesis and explains some relevant trends in the UK’s changing retail geography. The chapter begins by reviewing some of the most relevant changes in a retail location in the post-war era, and also stresses some of the methodological changes and the role micro-scale has played in the early retail research. Next, the organising concept of the retail area is explored in more detail, and terminological clashes found in the literature related to the types of retail areas are laid out. The literature review then goes on to outline the trends in the British retail sector, focusing on the period from the 1970s until the present day.

2.1 Changing Spatial Processes and Methodologies in Retailing: an Early Literature Review

Intuitive logic and formal methodologies used to select sites for new stores changed significantly in the post-war United Kingdom. The consumers became markedly more mobile, and longer trips necessary to indulge in shopping activities were no longer seen as an obstacle. During that shift, retail geography, analytical methods with a particular emphasis on the geocomputation, evolved dramatically and this enabled new research avenues, some of which are presented in this chapter, whereas others are introduced in Chapter 3.

2.1.1 The Urban to Suburban Shift of Retail Location

After the 1960s, locational theories such as central place theory that aimed to explain and suggest the settlement patterns or the location of the retail property were discounted as being largely unrealistic, as they assumed that forces shaping the locational decisions were permanent rather than dynamic (Schiller, 2001). During the 20th century, a rapid road
network developments and a surge in car ownership coupled with an increase in ownership of refrigerators and freezers fueled the changes in consumer behaviour and spending, as well as the spatial processes that govern the retail site suitability (Bromley and Thomas, 2002).

For instance, before the widespread adoption of cars, town centres were considered as the most accessible parts of urban areas, whereas increasing car ownership made town centres increasingly congested and sometimes even least accessible. (Schiller, 2001). As Schiller (2001) then emphasises, not only that traditionally optimal locations have become less favourable for retailers, but also, the attractiveness of environment has emerged as one of the essential factors in determining the location for a new store. Whilst not always crucial, the evidence shows that the lack of good accessibility to the town centre may lead consumers to shop outside of their local area (Findlay and Sparks, 2009), and this, in turn, makes out-of-town developments so attractive.

This had further consequences on the supply side, as there has been a shift from a larger number of smaller operations to the market dominance of a small number of major UK corporations (Schiller, 2001). This meant that retailers started investing larger amounts of money into individual locations, and if the corresponding locations later turned out to be failures, there was no simple way for that retailer to close the units without a considerable financial loss (Simmons, 1988).

However, benefits of opening up in the suburban locations seem to have outweighed risks. The out-of-town locations offered more space, both for new facilities and car parking (Bromley and Thomas, 2002), and these locations have, at the same time, become more accessible. Since one of the most important locational decision factors for retailers comprised accessibility and size of the shopping centres (Timmermans, 1986), the surge of out-of-town retail areas seems unsurprising.

However, this spatial process did not occur overnight, and it can be conceptualised by the so-called three waves of retail decentralisation as described by Schiller (1988). In the first wave, which commenced in the 1970s, food stores started purchasing vacant land by using the funds acquired by selling their town centre properties. Five to ten years later, the second wave followed, and it involved large retail warehouses selling bulky goods popping up at the suburban locations. Finally, the third wave encompassed clothing and quality comparison goods retailers, who also dispersed from the town centres (Schiller, 1988).

2.1.2 New Methodologies in Retail Location Analysis: Intuition Versus Modelling

The processes described in the previous subsection were accompanied by the changes in the methodologies of locational analysis, with GIS methods becoming increasingly available and important, making the decision-making process increasingly scientific, as opposed to it being art beforehand (Hernandez and Bennison, 2000). Store locations and locations of
competitors could now be mapped; road networks, land use, local demographics and other relevant layers of data could be added, and a more sophisticated spatial analysis could be conducted. In the 1990s retail sector was characterised by the increasing competition and falling margins, which meant that any competitive edge retailers could obtain in such an environment was deemed precious (Clarke, 1998).

Ironically, it took some time for such models to become widely applied in retail research and, especially, industry, so most of the literature making use of the more complex quantitative methods in retail location analysis comes from the newer dates. Examples of such methods applied in retail include spatial autoregression and geographically weighted regression (Ozuduru and Varol, 2011), multinomial logit models (Zhu et al., 2006), combining microsimulation with traditional spatial interaction models (Nakaya et al., 2007), machine learning approaches, such as neural networks (Byrom et al., 2001b), and many others.

According to surveys analysed by Hernandez and Bennison (2000), the majority of retailers at the turn of the century still resorted to subjective and intuitive approaches to solving locational problems. To a certain extent, this should not be seen as negative, as Davies (1976) pointed out earlier that newer technologies should not completely replace the retailer’s intuition and subjective judgement, but they should be used as the complement to either accepting or refuting the intuitive hypotheses. Wood and Browne (2007) emphasised the fact that the combination of the fieldwork, i.e. site visits and quantitative techniques is most likely to provide the most effective solutions.

2.1.3 The Rise of Micro-Scale

The growing wealth of data, analytical methods and computing power meant that we could now start conducting locational analyses not only on the coarser scales, such as national, regional or metropolitan scales, but it has become possible to come down to the neighbourhood, street or even a property level. The era of micro-scale research in retail geography has come. In his comprehensive literature review on retail micro-scale research, Brown (1994) recognised three different approaches to study of retail micro-scale location: theoretical, demand-side and supply-side perspectives. The theoretical approach comprises two long-established conceptualisations: bid rent theory and the principle of minimum differentiation (Brown, 1994).

Bid rent theory predicts that urban land uses change from the town centre towards the edge, reflecting the land prices. Since land in the centre is deemed most attractive due to its higher accessibility, it is often purchased by the highest bidders, i.e. the retailers that can derive the greatest utility from the location, such as department stores and speciality retailers (Brown, 1993). This results in a concentric pattern of urban land use, which was, to a certain extent confirmed in practice, but the theory also attracted heavy criticism due to its simplicity (Brown, 1993).
On the other hand, Hotelling’s principle of minimum differentiation posits that competing retailers selling the same type of product tend to cluster together at the centre of the market, and this is due to their continuous efforts to steal as many customers from their competitor as possible (Brown, 1993).

Demand-side studies range from the analysis of how consumers move and interact with space to micro-scale adaptations of the gravity model (Brown, 1994). For example, Foxall and Hackett (1992) used a wayfinding walk, a wayfinding commentary and a cognitive mapping exercise to assess the consumers’ perception of retail microsite location. They discovered that consumers tend to be more aware of high street stores located at the nodal and other prominent positions regardless of their function (Foxall and Hackett, 1992).

Finally, supply-side perspectives include the tenant placement policies practised by owners and developers of planned shopping centres, the micro-scale locational strategies of conglomerate retailing organisations and the influence of government legislation (Brown, 1994).

The early literature on micro-scale of retail location acknowledges that this level of detail in retail decision-making has been unfairly neglected, but it is interesting to note that retail microsite location analysis has not seen a lot of progress after the publication of Brown’s cited papers. In fact, theoretical approaches to which Brown referred to, such as the Hotelling’s theory or bid rent theory belong to those types of theories that Schiller (2001) viewed as simplistic and obsolete. Thankfully, further technological innovations at the beginning of the new century inspired retailers and academics to return to the microscale. This time, the vast quantities, high accuracy, speed and continuity of data acquisition, transfer and analysis served as the prerequisites for a more permanent interest and involvement in the micro-scale research. The primary datasets used in the analytical chapters of this thesis – an LDC’s retail unit database and the footfall database – both come down to micro-scale and will be introduced in Chapter 4.

The literature review now turns to some aspects of micro-scale, in particular, how individual retail units cluster together and how we can delineate those areas of higher concentrations of retail units in order to create a systematic spatial framework for the retail data collection and analysis.

### 2.2 The Spatial Framework of UK’s Retail Geography Research

Retailing is observed to concentrate more in certain areas than others. There has been a substantial interest in geography to try to delineate areas of dominant retail activity. This would enable academics and retailers to collect and aggregate data on retail land use at the level of purpose-built spatial units and enable comparisons of changes of retailing over time.
and between areas\(^1\).

As briefly mentioned in the introductory chapter, those areas will be referred to as *retail areas*, the term also used in the paper by Singleton et al. (2016b). The literature frequently mentions the term *retail centre* which is a wider term usually used to encapsulate all three retail formats commonly encountered in the UK (town centres, shopping centres and retail parks) (Dolega et al., 2016; Lugomer and Lansley, 2016); and it is thus similar in meaning to the term retail area. However, in this work, term *retail centre* is only used to denote relatively bigger retail areas, comprising at least 10 units (as suggested by Wrigley and Lambiri (2015)) and term *retail area* is a scale-free concept, which can contain anywhere from one stand-alone store to thousands of retail units. In other words, every retail centre is a retail area, but only those retail areas with higher importance in the British retail hierarchy should have the title of a centre. This is inspired by the Christaller’s central place theory (Getis and Getis, 1966) which uses the term central place for all the settlements that offer services to other surrounding settlements.

Retail areas can be divided into several spatial formats, based on their form and location. A categorisation employed by LDC, that has been followed in this thesis, classifies all the retail areas into: town centres (used interchangeably with the term *high street*), shopping centres\(^2\) and retail parks (Sparks, 2016). Their typical layout is shown in Figure 2.1, with additional remark that shopping centres often span across multiple floors.

\(^1\)These types of spatial units would be something comparable to the traditional census geographies, such as enumeration districts, census tracts, output areas, etc.

\(^2\)The term ‘shopping centre’ here implies the self-contained managed shopping centres.
Figure 2.1: Three main formats of retail areas

However, different terms can be found in the academic literature and professional reports, essentially referring to the same retail formats and these terminological clashes are now briefly visited. According to Reimers and Clulow (2004), terms town centre, city centre, high street, downtown and shopping strip may be treated as synonyms, even though further differences can be identified from other sources. High street is defined as a retail centre that serves the need of the local community. It is close to offices and other non-retail businesses and is accessible by various transport options (Deloitte, 2013). It normally consists of either one major high street or a collection of adjacent smaller streets, situated in a town centre. One should note that term town centre does not imply that major cities have only one such centre, but may have many of them. On the other hand, smaller or medium-sized towns may only have one.

The terms shopping centre and shopping mall are often close synonyms. They refer to a planned retail development comprising at least three shops, under one freehold, managed and marketed as a unit (Guy, 1994). They are typically covered and can be situated both within town centres or in suburban or out-of-town locations.

Retail warehouse is a term also used in the literature and it is mainly related to the concept of a retail park (Knight Frank, 2013), with retail park being a loose grouping of retail
warehouses and units at highly accessible suburban and out-of-town locations (Bromley and Thomas, 1988). They were originally developed at the converted industrial estates which ceased being in demand from their original industrial users (Schiller, 1988). Retail parks and shopping centres situated outside the city are sometimes collectively referred to as out-of-town centres (Knight Frank, 2013).

High streets play a major role in the development of sustainable urban communities by serving as places where residents meet, shop, simply travel through or pursue other activities (Jones et al., 2007). This said, retail centres of all types, including town centres, shopping centres and retail parks have evolved from being places where one goes shopping out of necessity to places with a pronounced leisure function (Deloitte, 2013).

Now that we have established what retail areas are, the next step is to explain how they can be delineated. If we imagine a set of 500 retail units scattered across the central neighbourhoods of a certain town, the reasonable question that comes up is – do all 500 retail units constitute a single retail area or should we break it down into multiple smaller retail areas? It is fair to say that delineating shopping centres and retail parks from the surrounding non-retail oriented areas is not a difficult task due to the fact that those types of retail areas have clearer physical boundaries. However, when it comes to town centres, the issue becomes more convoluted.

As Thurstain-Goodwin and Unwin (2000) note, everyone would agree that a town centre exists and it is also possible to tell when we are standing in one. However, it is very difficult to agree on its more accurate delineation and definition, as it can focus on different primary criteria. For example, a town centre may be defined as a retail core or an area with the highest density of buildings or an area with high density of offices. This all means that boundaries of town centres are fuzzy more than they are deterministic or in other words, they are geographic objects with indeterminate boundaries (Thurstain-Goodwin and Unwin, 2000; Burrough and Frank, 1996). In addition, the problem with the boundaries of town centres is that they have been static in a very dynamic period of change for urban area activities.

Nevertheless, despite all the obstacles, attempting to delineate town centres is a problem worth pursuing because such output is essential for statistics reporting and comparison (Thurstain-Goodwin and Unwin, 2000). Also, in more specific sense, availability of any resulting dataset would inherently help gain a better understanding of the relationship between use of retail space and changing consumer behavior (Pavlis et al., 2018).

Possibly the most widely used spatial delineation of town centres in the UK was developed by Thurstain-Goodwin and Unwin (2000) by using kernel density measures to generate continuous surfaces. These have been in use by the Department of Communities and Local Government (DCLG) (Dolega et al., 2016). Thurstain-Goodwin and Unwin (2000) emphasised that using the data gathered on the level of a typical census tract for the
purposes of town centre delineation is an approach that is doomed to fail because census tracts may easily be bigger than town centres as employment activities and land use vary at scales finer than that of a census tract.

That being said, the cited authors used unit postcodes (UPCs)\(^3\) and developed spatial density surface representations which were thought to have two major advantages compared to discrete boundaries, despite uncertainties they introduce through choice of arbitrary parameters. First, they enable utilisation of high-resolution data such as UPCs or satellite imagery-based land use data; and second, they enable the development and calibration of models of urban dynamics in which space is represented continuously and hence in a potentially richer way than the discrete zone-based methods (Thurstain-Goodwin and Unwin, 2000).

However, the boundaries drawn by Thurstain-Goodwin and Unwin (2000) were limited for the analysis of structural change of retail areas, as their detection was based on the combination of morphological and functional variables such as building density, socio-economic variables, diversity of building use, and tourist attractions (Mackaness and Chaudhry, 2011). Moreover, town centres in that case are defined more generally as areas of employment, thus including not only areas of higher concentration of retail activity, but also areas of prevailing office land use (Pavlis et al., 2018).

On the other hand, there has been a need for update of the boundaries, and retail unit datasets have become increasingly available and more expansive over the last decade (Pavlis et al., 2018). The same authors devised a new spatial clustering-based approach which tends to identify areas of larger concentration of retail units in British retail areas, thus focusing on retailing and services, rather than areas of employment in more general sense, as was the case with Thurstain-Goodwin and Unwin (2000) work. In doing so, they utilised an expansive national retail unit database\(^4\) created by the Local Data Company after nationwide field surveys conducted in 2015. They tested five different clustering algorithms (k-means, DBSCAN, random walk, quality threshold and KDE), each on the selected sample of eight retail areas of Great Britain that were deemed representative in terms of retail location density and size (Pavlis et al., 2018). The initial retail areas were selected based on the availability of qualitative data published by the local authorities that enabled the calibration and validation. Out of the five clustering algorithms, DBSCAN was chosen as a preferred method because it was found to be computationally most efficient and it dealt with outliers better than other tested algorithms (Pavlis et al., 2018). The authors then developed a modified version of DBSCAN to accommodate the fact that the original

---

\(^3\)According to Thurstain-Goodwin and Unwin (2000: 307): “UPCs are a generalisation of settlement geography designed for the speedy delivery of the Royal Mail. They represent the spatial average of around 14–17 properties on a postman’s delivery and are georeferenced by a single point reference.”

\(^4\)This database will be one of the two main proprietary datasets used for the analytical part of the research and will be introduced in more detail in Chapter 4.
DBSCAN method (as developed by Ester et al. (1996)) had issues identifying spatial clusters in areas of varying densities (Wang and Hamilton, 2003).

Examples of boundaries of retail areas situated in Birmingham derived both using the Thurstain-Goodwin and Unwin (2000) method and newer method developed by Pavlis et al. (2018) are given in Figure 2.2. As one can notice, Birmingham town centre, as defined by Thurstain-Goodwin and Unwin (2000), covers a substantially large area which comprises not only the most important high street, but also many different surrounding streets of different hierarchical level. Using the older method, only one retail area emerged in Birmingham. However, when newer, modified DBSCAN method was used, a single retail area ended up being broken down into seven constituent retail areas, some of which are neighbours and some of which are completely detached from one another. This is to a large extent a manifestation of the high granularity of LDC’s dataset, which enables detecting spatial continuities of retail activity on much finer spatial resolutions.
Given the fact that retail area boundaries derived by Pavlis et al. (2018) are more up-to-date and focus upon the areas of high presence of retail units, they were used as a spatial framework under which data on retail units were aggregated and further analysed in this thesis. After introducing the key terminology and defining retail areas, the next step is to describe what was happening within those areas throughout the past several decades, in
order to gain a better understanding of the most significant forces that have been shaping the retail economy.

### 2.3 The Ongoing Evolution of the UK Retail Structure

The second half of the 20th century saw a consistently increasing demand for retail prop-
erty (Hughes and Jackson, 2015) and consistent changes in UK retailing. According to
Wrigley and Lambiri (2014), the forces driving the retail economy can be divided into three
groups based on the time frame on which they operate: short-term factors such as effects
of economic shocks; medium-term factors such as the impact of out-of-town and off-centre
developments; and long-term factors such as demographic change, convenience culture and
the rise of e-commerce (Figure 2.3). Those groups of forces are interrelated in a complex
way and thus, they have been the topic of many research papers, especially in recent years.

In this section, the most important and the most widely covered changes that have taken
place are highlighted and described in more detail. The focus is on the more recent phe-
nomenons, whereas the introduction to broader geographical changes in retail geographies
that followed after World War II has already been given in Section 2.1. In some cases, ref-
erences were made to some other countries and global issues, as they provide an important
addition to a more general understanding of how main economic, social and political forces
shaped retail environments. The topics that were thereby explored were as follows:

- the establishment of new out-of-town retail formats and their more recent effect on
  the economic health of the traditional town centres;
- the emergence and development of the shopping as leisure concept;
- the impact of the global financial crisis on the UK retail economy;
- the rise of e-commerce and its implications for traditional bricks-and-mortar busi-
  nesses;
- the rise of the convenience culture.
2.3.1 Development of New Retail Formats

Development of out-of-town retail provision has already been introduced in Section 2.1, and this subsection expands on the topic by inspecting some of the negative consequences of the rise of those new retail formats, as well as the policies envisioned to mitigate them.

The emergence of competition from the out-of-town retail developments led to the demise of some of the town centres and this is particularly true for middle-order city centres in the national settlement hierarchy (Thomas and Bromley, 2003). During this period, there have been various attempts by successive governments to preserve the commercial and social identities of British cities by stopping the economic decline of town centres (Hughes and Jackson, 2015). Since town centres uses go beyond purely retail, as a consequence,
the need for management across a variety of stakeholders and initiatives arose in the past few decades (Wrigley and Lambiri, 2014) in the form of so-called town centre management (TCM). Town centre management can be defined as: “a comprehensive response to competitive pressures which involves development, management and promotion of both public and private areas within town centres, for the benefit of all concerned” (Wells, 1991: 24). Its growing importance at the turn of the century is reflected in the fact that the number of town managers in the UK grew from none in 1986 to 182 in 1996 (Pritchard (1996) cited by Warnaby et al. (1998)).

Town centre management schemes can be classified based on their position on the structural perspective continuum, in which either public (for example, forums) or the private sector (for example, Business Improvement Zones or BIZs) has potential to dominate decision-making (Warnaby et al., 1998). The majority of TCM schemes belong to the centre of the spectrum with the balance of the public and private sector in the organisational structure and example of such organisation are private limited companies (Warnaby et al., 1998).

Furthermore, TCM schemes can also be divided based on the way they are funded. According to Warnaby et al. (1998), funding can be either: (a) obligatory, in form of set rates which businesses pay to the local authority and which then distributes funds for the purposes of maintenance and planning; or (b) discretionary, in which case businesses give donations to the local authority on a one-off or recurring basis.

Apart from town management schemes, central government action also played significant role in managing the changes that hit town centres and arguably the most well-known one in the UK was the 1996 Planning Policy Guidance Note 6 (Department of the Environment, 1996), which aimed to restrict out-of-centre development by implementing the so-called sequential approach (Hughes and Jackson, 2015; Findlay and Sparks, 2007). The sequential approach implies giving town centres the first priority when selecting sites for opening new retail units. Only in the case that no locations are available, should a retailer choose to open elsewhere: with edge-of-centre locations having a secondary priority and out-of-town locations having a tertiary priority (Findlay and Sparks, 2007). In the latter two cases, a preference would be given to those sites that are either well connected with the centre via existing transport routes or that have good potential for forming such links.

Despite being partially successful for the capability of limiting the scale of retail decentralisation, the sequential approach had some flaws (Thomas and Bromley, 2003). Even though it did to a certain extent slow down the decline of town centres, restrictions imposed on out-of-town development were insufficient to reverse the process (Colliers (2011) cited by Hughes and Jackson (2015)). This is because there are other issues impacting the demise of town centres, such as the lack of regular-shaped sites, changing market conditions during drawn-out development periods and availability of car parks in town centres (Colliers (2011) and Guy (1994) cited by Hughes and Jackson (2015)). In addition, sequential
approach contained highly ambiguous definitions that enabled different local governments and town centre managers to address the issue in a non-uniform way. For instance, edge-of-centre locations were defined as places that are normally no more than 200 – 300 meters away from the town centre, with car parking facilities and good transport links with the town centre. Such definition is obviously open for different interpretations, as maximum distance from the town centre is ambiguously defined and also, it is not clear from where that distance should be measured (Thomas and Bromley, 2003). It is also worth stressing that town centres normally have complex land ownership, meaning that even the retail pitch most favoured by the government policies at the same time happened to be most constrained by the multiple land ownership (Adams et al., 2002). To illustrate how time-consuming and expensive the assembly of the land ownerships can be, the same authors provided an example of a major local development company which spent two and a half years assembling 37 different ownerships in the early 1990s to form an eight-hectare site in Stoke-on-Trent for commercial and retail development. The entire process cost a staggering £4.5 million (Adams et al., 2002).

After a period of relative decline in importance of town centres compared to out-of-town centres, the 2007 – 2013 period saw a revival of in-town shopping centres and growth of town centre sales in both convenience and comparison retailing (Wrigley and Lambiri, 2014). Moreover, food shoppers who previously used out-of-town centres and then switched to the town centres were found to exhibit linked-trip behaviour, resulting in more frequent visits to the town centre than before (Wrigley et al., 2010). It is also important to note that some specific retail categories helped high streets to re-establish themselves due to their traditional location and relevance for consumers. For instance, pharmacies have traditionally occupied important places on the high street and consumers prefer visiting high streets to out-of-town centres when shopping for pharmaceuticals and beauty products. (Wrigley and Lambiri, 2014).

### 2.3.2 Shopping as Leisure

Leisure floorspace and spend quadrupled in the 2002 – 2012 period (Hughes and Jackson, 2015). It is apparent that visiting a shopping destination, regardless of the retail area format, has evolved into something more than a shopping experience, with added elements of leisure (Bloch et al., 1991). This is not an entirely new phenomenon, as relative growth in consumer spendings on leisure goods and services and the simultaneous slide of spendings on convenience products, clothes and footwear was at full swing at the turn of the century (Office for National Statistics, 2006) and had commenced even earlier.

According to Jansen-Verbeke (1991), by the 1980s, leisure facilities had started to become ‘a must’ in the new retail environments. Not only did leisure become an important component of a local consumers’ experience, for both residents and workers, but it also
became an important attractive factor for tourists, especially domestic and international shopping tourists (Jansen-Verbeke, 1991; Law and Au, 2000; Turner and Reisinger, 2001). It was also noted that it became irrelevant whether the place of shopping was a simulation of, or a genuine cultural setting; whether it was newly built or not. Instead, the emphasis was upon diversified experience which, in effect, increased chances of visiting multiple different retail units (Jansen-Verbeke, 1991). Gratton and Taylor (1987) concluded that whilst major leisure attraction, such as exhibitions, can affect shopping patterns, the effects are smaller than anticipated, with retail revenues being unlikely to record spikes since most of the visitors’ expenditure would be on accommodation, food, drinks and events. Visitors’ shopping behaviour and expectations which are particularly relevant to those retailers operating in tourist destinations are also impacted by their cultural and ethnic background (Wong and Law, 2003). For instance, the cited study revealed that Western tourists visiting Hong Kong were more satisfied by the service quality, quality of goods, variety of goods and price of goods during their visit, as compared to the Asian travellers.

The impact of leisure on the consumer behaviour also heavily depends on other demographic characteristics, such as age. According to Anthony (1985), the longer time spent in the shopping centre, owing to a significant proportion of bars, cafes, restaurants and cinemas, does not necessarily have to increase spending and this was found to be particularly true for adolescents, who tend to visit shopping centres on a regular basis, i.e. at least once or twice a week, spend one to five hours there, but less than half of them would see shopping as a primary motivation for the visit. In other words, it seems that leisure units may serve to attract consumers who would not otherwise appear there at all, however, convenience and comparison stores would not benefit as much from such consumers as the leisure units themselves.

In some markets, leisure units may have negligible impact on motivating consumers to indulge in leisurely pursuits. Shopping for pleasure did not develop as much as a lifestyle concept in the less developed markets as it did in the developed ones (Babin et al., 1994) and an additional empirical proof for that claim was later given by Millan and Howard (2007) who investigated motives and behavioral traits of Hungarian shoppers (sample of 355 shoppers). Hungary has recently undergone a process of transition towards the free market economy and therefore started introducing the concepts of shopping as a hedonic experience. However, even while those social and economic processes were in full swing, Hungarians still saw shopping primarily as a functional experience, rather than hedonic. Some of the suggested reasons were lower purchasing power, smaller amount of available time in the struggling economy and the case that their leisurely expectations are not matched by the market supply. This all has important decision-making and shopping centre planning implications, as marketing strategies that aim to attract consumers within the shopping centre may not be effective due to the fact that such shoppers more often than not decide...
upon which products to buy before visiting a shopping centre (Millan and Howard, 2007).

Nevertheless, in the grand scheme of things, the emergence of the allied concept of shopping and leisure acted as a positive feedback loop by propagating the number of units offering food, drink or different forms of entertainment, which in turn generated more footfall. According to Howard and Gratton (2001), 56% of the money that people spend on leisure and away from home is on food and drinks and customers who stop for a meal or drink are more likely to spend more time and money in the shopping centre (Howard, 1993). Therefore, leisure units located inside shopping centres may benefit from taking advantage of lunchtime and evening footfall peaks even more than they would if they were dislocated from convenience and comparison stores (Vines, 2001). The key thing to keep in mind is that the impact of leisure on shopping behavioural traits is far from simple and it is modified by the vast array of factors pertaining to consumer demographics and broader economic context of the setting in which shopping takes place.

With this in mind and given the overall importance of retail areas as the loci where substantial proportion of a daily activity of local population takes place, change in their composition and economic health may greatly affect those activities. Therefore, it does not come as a surprise that much of the recent research in British retail geography has been concerned with the ways in which different retailers have responded to economic hardship, with special attention devoted to the greatest economic blow to the global economy in the 21st century, potentially even greater than the Great Depression (Chossudovsky, 2008) - the 2008 global financial crisis.

### 2.3.3 Effects of the Global Financial Crisis on UK Retailing

The 2008 global financial crisis struck the UK retail economy harshly. However, its effects were not spatially uniform. On a coarser scale, this can be attributed to the differential pre-crisis economic health of retail areas and, from a microeconomic perspective, to the fact that some retailers were better prepared and had devised robust strategies for dealing with the onset of economic hardship (Dolega, 2012).

During the crisis period, closures of retail units and the consequential rise of vacancy rates were found to be some of the typical processes at work in British retail areas (Wrigley and Dolega, 2011). The recession had more pronounced influence on those centres that experienced lower retail vacancy rates in the pre-recession period and, in effect, some retail centres were found to be much more resilient to the crisis (Tselios et al., 2018). Even though diversification, in terms of presence of both independent retailers and major chains, is often thought to be beneficial for the vitality of town centres (Wrigley, 2010), there are also studies that refute such claims by stating that underperforming town centres were no different in terms of retail diversity compared to the overperforming centres (BIS, 2014). Some measures that were found to be effective, on the other hand are: “investment in
temporary ‘crowd-pulling’ events to enhance the attractiveness of a centre; funding for town centre management; supportive car parking policies; schemes to strengthen the nighttime economy of town centres, etc.” (Tselios et al., 2018: 351). Also, whilst food superstores are capable of destroying small independent retailers in case their market dominance in the retail area becomes decisive, their presence is still viewed as beneficial because it can generate high levels of linked trips, which are deemed essential for town centre vitality (Wrigley et al., 2010).

Because of the combined effect of differential susceptibility of different retail categories to prevailing economic circumstances, changes in retail area structure occurred with particular types of shops filling in the vacated units\(^5\). As said before, not all retailers experienced the same degree of success in coping with macroeconomic hardship and this is also true for different categories of retailers based on the type of goods they sell. Convenience retail categories have shown much greater resilience\(^6\) to the macroeconomic shock, when compared to comparison retailers. More resistant to recession have also been smaller town centres, centres situated in southern parts of England, as well as centres with higher pre-crisis proportions of service and leisure units\(^7\) (Wrigley and Dolega, 2011).

Following the diminution of the crisis, retail areas have mostly returned to their pre-crisis trends (Wrigley and Dolega, 2011).

\(2.3.4\) The Impact of E-Commerce

Another important factor that influenced the direction in which UK retailing is heading is online retailing, the Internet retailing, e-tailing or e-commerce. Ever since the turn of the century, it has become impossible to read business newspaper without stumbling upon articles on major online retailers (Burt and Sparks, 2003). Since then, one can identify two distinctive models of retailing: brick-and-mortar and online retailing (Enders and Jelassi, 2000).

The more traditional, brick-and-mortar model of retailing implies that the setting where purchase of goods takes place is a physical store. The first notable advantage of this model is the fact that, compared to newly emerged online retailers, bricks-and-mortar stores have had presence on the high street for so long and thus have established brands that are well-known

---

\(^{5}\)For example, discount and charity shops came in place of comparison retailers which suffered comparatively greater loss (Wrigley and Dolega, 2011).

\(^{6}\)According to Reggiani et al. (2002), resilience is potentially a key topic on the study of how spatial economic systems respond to shocks and disturbances. Even though seemingly intuitive, concept of resilience is rather ambiguous and may be considered as both a positive and negative occurrence, with different meaning in different contexts (Martin, 2012). However, discussion of the conceptual details is out of the scope of this thesis.

\(^{7}\)Service and leisure units entail significant footfall generators, such as restaurants, cafés, cinemas, bars, etc.
to customers. Second, bricks-and-mortar businesses enjoy the benefit of having an existing infrastructure and distribution network and tend to attract customers not only directly, but also via the concept of linked shopping trips (Enders and Jelassi, 2000). On the other hand, online retailing offers wide reach, exhaustive product selection, smaller investments into physical infrastructure, unlimited opening hours, high degree of scalability in the likely event of demand expansion (Enders and Jelassi, 2000). At the same time, informational leverage that the Internet accessibility provides for increasingly large proportions of the population in most developed countries serves to offer a different kind of challenge: a more informed consumer, able and willing to challenge the status quo of retail brands (Reynolds, 2002).

In their review of the beginnings of online retailing, Doherty and Ellis-Chadwick (2010) said that first attempts of tech-companies to sell their products online date from the mid-1990s. However, back then, many problems were envisioned by the retailers that were thought to act as significant barriers, preventing the growth of e-commerce. According to Cockburn and Wilson (1996), some of the most burning issues at the time were found to be the Internet security issues and absence of online payment options. Nath et al. (1998) conducted a survey on the sample of executives of ten organisations in different sectors (transportation, food processing, hotel chain, finance, etc.) and found that whilst most of the surveyed managers were interested in utilising the opportunities that the Internet offered, they at the same time lacked the necessary knowledge and understanding of how exactly the Internet can be used to the advantage of their businesses. Besides the lack of key information, related lack of skilled personnel and the lack of legal framework that regulates the Internet space, the executives identified costs of using the Internet services as another crucial impediment, as setup, hardware and software, connection and maintenance all require dedicated funds (Nath et al., 1998). Even though the study also identified several positives (ease of access, low costs of advertising campaigns, improvement of the company’s image by establishing a presence online) (Nath et al., 1998), the prospects of online retailing were definitely uncertain at the time. In addition, it was unknown how consumers would react to such innovations and whether they would embrace the new mode of shopping (Burt and Sparks, 2003).

Uncertainties escalated during the dot-com stock market crash in April 2000 in which some of well-known businesses burned hundreds of million dollars in only a couple of years (Thornton and Marche, 2003). The cited authors studied whether online retail businesses that failed during the market bubble would otherwise have succeeded had the macroeconomic conditions been more favourable. They found that in most cases, famous pioneering e-tailers failed due to poor business models and in the event that the crash had not happened, they would have still struggled. Some of the notable mistakes that retailers made were (Thornton and Marche, 2003):
- developing a brand name and then offering different types of products that cannot be associated with the brand name;
- selling products that were simply more suitable for the bricks-and-mortar locations;
- initial availability of too large amounts of investments that were spent lavishly, leaving the new retailers incapable of strategic cuts during the subsequent market crash and the departure of investors;
- technical problems, security issues and similar factors pertaining to, at the time, yet uncharted territory of the newly emerged computer and the Internet technologies;
- general poor business planning, etc.

Concerns were also raised even about the scenario in which online retailing would eventually emerge as a successful long-term channel because this outcome would not be without potentially negative consequences. In particular, while Weltevreden (2007) did not find that online retailing significantly substituted visits to town centres, it was presumed that this could change in the long-run. In effect, some brick-and-mortar businesses could decline after being faced with increasing online competition, therefore also negatively affecting retail areas as a whole. There was a growing fear that physical stores may turn into some sort of ‘showrooms’ that will be used by customers only to browse the products and receive advice from the members of staff, but then they would return to their computers and finish the purchase online (Reynolds, 2002). Enders and Jelassi (2000) speculated that if physical stores remained important shopping destinations, those retailers who manage to combine them with e-commerce would flourish. Moreover, consumers are likely to start looking for new types of experience in physical stores which they would not be able to get online, especially in the event that a problem occurs and product cannot be bought online for whatever reason. Such experiences could be linked to the concept of shopping as leisure that was visited in one of the previous sections (Enders and Jelassi, 2000; Thornton and Marche, 2003).

An overview of presumed necessary future strategies that e-tailers need to adopt in order to succeed in the post-dot-com market conditions and in the long run (Thornton and Marche, 2003) were found to be largely correct. The cited authors established that businesses would have to develop plans for becoming profitable in the long run, otherwise, investors may not be interested. A slower growth and better planning were thereby expected and more personalised advertising campaigns, utilising newer technologies that the Internet had to offer were also thought to be essential. This all happened eventually and pessimistic predictions about the fall of e-commerce at the beginning of the 21st century did not come true. On the contrary, e-commerce has become a growing and reputable retailing model
and while at the same time it did pose a threat to the high street, a synergy between the two has become established. If one inspects what happened in the current century, the online population and the value of online transactions kept growing and albeit this happened at a slower rate than some have anticipated (Reynolds, 2002), this growth was still sufficient to eliminate a great deal of aforementioned uncertainties. Later, even throughout the period of the global financial crisis, it was found that the proportion of online sales in the total retail spend more than tripled between 2007 and 2014 (Wrigley and Lambiri, 2014). Moreover, competition amongst retailers targeting the same parts of the market became more fierce and strategies had to be developed that survey and exploit customer expectations (Wilson-Jeanselme and Reynolds, 2006), as this came to be particularly relevant for winning customers amidst strong competition.

While seeming as the potential death of the high street at first, it should be noted that town centres have not been decimated by online shopping services (Deloitte, 2013). According to Mesure (2011), Matthew Hopkinson, former managing director of LDC, explained this at the time as a result of differential presence of certain types of retailers online. For example, leisure retailers were not as struck by the online competition because their services were mostly provided on-site rather than online. Furthermore, ‘click and collect’ services show that online commerce may assist high street retailing, and not just oppose it, as it serves to attract people back to the physical stores (Wrigley and Lambiri, 2014). The recent literature refocused on the resilience of retail areas, so Singleton et al. (2016b), for example, constructed a composite measure of resilience of retail areas which measured how prone retail areas are to the impact of online retailing. This effect can vary from complementary to damaging and authors stressed that a number of assumptions were made about what types of retailers typically act as a positive or negative catalyst to change. The combined e-resilience measure revealed a geography where attractive and large retail centres such as the inner cores of large metropolitan areas and smaller more specialist centres serving convenience shoppers emerged as more resilient. On the other hand, many secondary and medium-sized centres were identified as more vulnerable (Singleton et al., 2016b).

Finally, according to the most recent Grimsey Review, the Internet sales, as a proportion of overall retail sales, increased from 10.4% in 2013 to 17.9% in 2017 and they are likely to continue growing. This, however, did not cause consumers to completely deter from high streets, as 85% of retail spend still touched physical store locations (Grimsey Review, 2018).

### 2.3.5 The Rise of Convenience Culture

Convenience foods can be defined as commercially prepared foods that enable the consumer to save time and effort in food activities, related to shopping, meal preparation and cooking, consumption and post-meal activities (Buckley et al., 2007). The prominence of convenience
foods arose as a consequence of a changing lifestyle, in which growing urban population ceased being interested in growing and preparing food but had the financial means necessary to buy more expensive processed foods (Tillotson, 2003).

It is therefore not surprising that convenience retail in town centres, both independently and corporately owned, has experienced significant growth over the past 15 years and this growth was sustained during the economic crisis and subsequent period of austerity (Wrigley and Lambiri, 2014). According to the survey conducted by Harris Interactive on behalf of The Grocer, consumers were found to be moving away from a regular big weekly shop and towards ‘as and when needed’ shopping trips (The Grocer, 2016). The main presumption for this process to take place is that local town centres either offer a range of goods and services simultaneously offered by the shopping centres and retail parks at the city edge; or that they offered goods and services that are unique or grounded in the culture of the local community (Wrigley and Lambiri, 2014).

Since convenience culture links convenience to the ‘local’ and ‘community’ (Wrigley and Lambiri, 2014), as a consequence, authenticity, traceability and ethical sourcing became important factors in consumers’ preferences and this in turn resulted in the rise of symbol group and independent retailers, who expanded even faster than the major corporate chains (Wrigley et al. (2009) cited by Dolega (2012)). A common opinion that small retailers stand no chance against the corporate food retailers has therefore been questioned or even rejected as not only recent academic literature, but also media reported fightback from the smaller entrepreneurs (Wrigley et al., 2010; Butler, 2015). This evidence further justifies the prediction that convenience stores could conquer almost a quarter of the grocery market share by 2019 (Gladding, 2016).

In effect, a significant increase in competition amongst convenience retailers was seen in recent years which is reflected in them stepping up their advertising campaigns and further optimising their locational strategies (The Grocer, 2016). The large grocery retailers were found to respond to the consumers’ changing preferences, however, this also varied from retail area to retail area, because different areas have different degrees of attractiveness for corporate retailers (Dolega, 2012).

2.4 Chapter Summary

This chapter defined some of the most frequently used retail geography terms that are used throughout the rest of the thesis. Retail areas were defined as the areas of dominant retail activity in both town centres and town edges (in which case a distinction is often made between shopping centres and retail parks). The importance of delineation of retail areas was stressed, with aim of having a consistent and robust spatial framework for data collection and spatial and temporal comparison of the composition and functioning of retail
areas. The examples of retail area delineation methodologies found in the literature were
described: a traditional one with broader use-case, and a newer one more appropriately
adapted to the research of retail activity.

The chapter also provided an overview of the most significant social, economic and
geographic forces that have shaped UK retail areas over the past couple of decades. It is
apparent that retailing is continuously going through a series of structural changes that
emerge as a result of changing local demographics, consumers’ lifestyles, behaviour and
preferences, as well as microeconomic and macroeconomic conditions. Understanding of
these factors is vital to the effective locational planning and implementation of marketing
strategies, which are, in turn, vital for the success of individual retail businesses as well as
the economic health of a retail area as a whole.

The most important factors that dictated where retailing in the UK was going over
the last few decades included the onsets of economic shocks (especially the dot-com bubble
burst and the 2008 global financial crisis). Second, development of new retail formats on the
city edges, which to a certain extent lead to a decline of town centres as primary shopping
destinations. Third, the e-commerce emerged and has grown substantially, which posed
threat to both town centres and out-of-town developments before eventually getting integ-
rated with bricks-and-mortar formats. Next, a leisure component as part of the shopping
experience became increasingly important. Finally, the rise of convenience culture has been
observed, which is at the same time related to the process of the return of consumers from
out-of-town centres to their local neighbourhood.

All of the described tendencies in development of the British retail landscape present a
complex system that needs to be thoroughly investigated and understood by analysing the
relevant underlying dynamic and structural factors. In the next chapter, the literature re-
view turns to some of those dynamic factors that are related to the consumers’ demographics
and activity patterns and ways these have been measured and conceptualised in traditional
and more recent literature for the purposes of target marketing, consumer behaviour and
mobility research.
Chapter 3

Established and New Forms of Consumer Data

Whilst the previous chapter introduced trends that govern the evolution of British retail areas, this chapter will be concerned with the underlying data sources that enable us understand that evolution. The spectrum of relevant factors influencing the state and trends of retailing is broad and many interesting datasets have been used in the past and developed recently to analyse them.

The chapter starts off by a brief overview of the traditional field of geodemographics. Knowing what type of people live in which areas has been one of the cornerstones of applied demographic research in the late 20th century and it remains a popular topic until the present day.

However, over time, newer forms of consumer data\(^1\) have emerged, that focus on activities, rather than on the static, infrequently updated snapshot of residential (nighttime) population in the form of the census. These are visited by going from the conception of Hagerstrand’s concept of time geography all the way to contemporary activity-based representations of consumers. These include demographic data gathered on the new census geographies such as Workplace Zones, as well as alternative datasets such as social media data, loyalty card data or other consumer data acquired by individual companies and local authorities.

After that, in the final section, the Big Data and Smart City concepts are introduced,

\(^1\)Throughout this thesis, consumer data are understood as all data sources that can be used to describe the characteristics of potential consumers in a given area. Even though the census of population may not always be labelled as ‘consumer data’ since it captures much wider population, here it is discussed as a traditional consumer dataset. This is because, by the end of the 20th century, most of the predictions of store patronage were modelled using the census data. Nevertheless, it is acknowledged that nowadays, when purpose-built consumer datasets are being created by the individual companies, the term is mostly used to refer to them and it would not normally be used to refer to the census.
as much of the methodological framework for subsequent work has been based on them. Increasing interest in the analysis of how urban areas function through analysis of human activity patterns is acknowledged and a special attention is given to the role that sensor systems play in the measurement of flows of people in the cities and thus, retail areas. Literature review finishes off by examining the negative consequences of the rise in volume of consumer data, primarily encapsulated by the problems of spatial data quality issues and emerging needs of preserving consumer privacy.

3.1 The Residential Geodemographics Tradition

In order to assess the suitability of a site for locating a particular type of store, demographic characteristics of consumers potentially visiting it need to be measured or estimated, along with other relevant socioeconomic and commercial factors. The most frequently used data source in that respect has traditionally been the census.

3.1.1 What Are Geodemographics?

Much of the applied analysis of census data originates from the field of geodemographics, which is defined as the classification of people based on where they live (Harris et al., 2005). The feasibility of such analysis is backed by Tobler’s First law of geography that states that “everything is related to everything else, but near things are more related than distant things” (Tobler, 1970). In the analysis of census demographics, this means that one can reasonably postulate that small administrative units such as unit postcodes or census output areas in the UK are homogeneous with respect to resident household structure (Longley, 2012).

The origins of modern geodemographic systems can be traced back to the early analysis of the patterns found in the census tract data in the USA from 1910 and early work of the urban ecologists of the Chicago school (Brown, 1991). However, in its newer form, geodemographics started its development in the era of rapid growth of popularity of quantitative geography and even more specifically, from 1966 when census data were made available in the format that could have been read by a machine (Brown, 1991). The first geodemographic classification in the UK was devised by Webber (1975). He later established the first commercially successful classification that became known under the proprietary name ACORN and marketed by the CACI. Later on, quite a few companies developed their own national geodemographic classifications and each was marketed as superior to others Brown (1991). However, as Brown (1991) noted in his review of the popular consumer segmentations (MOSAIC, PiN, Super Profiles and ACORN), it was very difficult to compare their

\(^2\)CACI is a company specialising in providing integrated marketing, location planning consultancy, network services and technology solutions (CACI, 2019).
quality and reliability as much of the methodology behind their creation remained unavailable. This is not surprising, as such segmentations were devised by private companies and have a tremendous commercial value.

Over time, a need for publicly available or ‘open-source’ national geodemographic classifications emerged (Longley, 2012). The idea is that they should make use of freely available data and share the complete methodology in a transparent way with all potential users that are then able to either use classifications as they are or create their own, using the underlying datasets. This presents a significant step in dissemination of geodemographic methods, which are no longer exclusive property of private corporations. In addition, and perhaps even more importantly, the existence of open-source geodemographics encourages users to be more critical to every classification out there, especially the proprietary ones whose provenance and methodologies often remain unknown or, at the very least, imperfectly understood.

3.1.2 Open Geodemographic Classifications and UK Census Geographies

Some of the most well-known and referenced national geodemographic classifications in the UK include the 2001 Output Area Classification (OAC) by Vickers and Rees (2007) and 2011 OAC by Gale et al. (2016). Output Areas (OAs) are the smallest census geography in the UK and were developed from postcode units specifically for statistical purposes. This is in contrast with the previously used Enumeration Districts (EDs) which were made for the conduct of the census in the field (Office for National Statistics, 2016d).

According to the Office for National Statistics (2016a), Output Areas were introduced in 1981 in Scotland and in 1991 in the rest of the countries of the UK. They were designed to meet the following criteria: (Office for National Statistics, 2016a)

- be as socially homogeneous as possible based on tenure of household and dwelling type (except in Scotland where this criterion has not been employed);
- urban/rural mixes were avoided whenever possible;
- it was aimed that they had approximately regular shapes and tended to be constrained by obvious boundaries such as major roads;
- and they had to have a minimum size to protect the data confidentiality. For example, in 2001, the minimum OA size was 40 resident households and 100 resident people, but the recommended size was rather larger at 125 households.

In 2011, some modifications took place on the boundaries of those OAs where significant population change had occurred since 2001 and in the case of OAs which were found to be socially heterogeneous. Since OAs tend to follow boundaries of the Local Authority Districts
(LADs), some OAs had to be modified in order to follow the changes in LAD boundaries. In effect, 2.6% of the 2001 OAs have been changed in 2011 (Office for National Statistics, 2016a) and the final figure comprised 171,372 OAs for England and 10,036 for Wales.

Various demographic and socioeconomic census variables aggregated at the level of the OAs can be used to create a nationwide geodemographic classification at fine spatial resolution. Vickers and Rees (2007) did that for the 2001 census data and documented the methodology in their paper. The methodology was complex and it involved many data wrangling steps before the actual cluster analysis took place. According to Vickers and Rees (2007), these steps include:

1. the initial choice of the underlying spatial units;
2. the initial choice of variables;
3. variable standardisation;
4. the choice of distance measure and clustering method;
5. the choice of the number of clusters;
6. interpretation, testing and replication.

It is important to note that there has been a lot of criticism of what has become known as data-driven social science that it does not always relate to the real world and, consequently, results of such approaches should be checked by comparing the results to what is observable in reality (Vickers and Rees, 2007; Pickles, 1995). Ground-truthing of the geodemographic classifications is usually conducted in a form of a consultation exercise whereby local residents, workers, or people otherwise familiar with neighbourhoods of interest are asked to confirm whether they agree with the group or subgroup to which their local area belongs. Outputs of such exercises, which were found to be very successful, include statistics about the success rate of matching of the classification results and respondents’ opinions; and also, a wealth of qualitative data in forms of comments and recommendations for future work (Vickers and Rees, 2011). Such consultations were found to be useful not only for ground-truthing purposes, but also for the preparation phase in making new classifications. For example, in their user engagement prior to the creation of new, 2011 OAC, Gale et al. (2016) found that users were looking for a general purpose classification conducted on the smallest areal level possible, without the need for being directly comparable to the previous classification and with preference of including the user-friendly pen portraits of clusters and interactive maps. They also found that cluster naming was very important to the respondents and that a measure of uncertainty of cluster assignment was desired. In the 2011 OAC conducted by Gale et al. (2016), 60 census variables entered the clustering and this resulted in 8 supergroups, 26 groups and 76 subgroups.
Although general-purpose geodemographic classifications have demonstrated utility, there are sound reasons for developing purpose-specific classifications (Singleton et al., 2016b). The variables that are selected for the specific-purpose classifications are easier to justify and their selection results in better clustering performance (Singleton and Longley, 2009). The purpose-specific classifications can be divided into those whose topic is specialised and those whose topic is generic, but are suited to a particular time period or area. For examples of the former, see the e-resilience classification created by Singleton et al. (2016b) and the classification of higher education opportunities by Singleton and Longley (2009).

The example of the latter is the Temporal Output Area Classification (TOAC), that focused on achieving stability over the intercensal period (Singleton et al., 2016a). Similar to this, some classifications were suited to particular area (see the London Output Area Classification (LOAC) (Longley and Singleton, 2016)). Petersen et al. (2011) emphasised that the special status of London in the national urban system comes from the fact that there is a high concentration of London neighbourhoods in only two of the seven nationally devised 2001 Output Area supergroups. According to Gale and Longley (2012), 56.1% of all the London output areas (2001) belonged to the Multicultural supergroup and 21.4% to the City Living group (Figure 3.1).

This shortcoming was addressed in the revised Output Area Classification for 2011 where population previously belonging to the single Multicultural supergroup was divided into Ethnicity Central and Multicultural Metropolitans (Gale et al., 2016). However, due to the need for even greater granularity, Longley and Singleton (2016) devised the London Output Area Classification (LOAC), which followed similar methodology and used the same groups of variables as the OAC. This resulted in eight supergroups with more even spatial distribution across London OAs.
3.1.3 The Impact of Catchment Demographics on the Economic Health of Retail Areas

How many people are likely to visit a store and who are they? The creation of catchment areas of individual stores has been a popular topic in quantitative geography ever since the 1960s and the so-called ‘Quantitative Revolution’. The idea of modelling catchment areas is to delineate the area from which consumers are likely to patronise a particular store, which in turn allows estimation of their number and demographic and socioeconomic composition. Huff (1964) has been commonly cited in the literature for his widely used probabilistic form of a spatial interaction model which computes probabilities that population of different output areas would visit a store, taking into account attractiveness of a store and distance from it as key factors. This can be done not only for the individual stores, but also for the retail agglomerations, such as retail areas (Dolega et al., 2016). The cited authors developed a modified Huff’s probabilistic method which calculates the attractiveness of the UK retail areas and estimates their catchments.

Regardless of whether the individual stores or retail agglomerations are subjects of the
analysis, the examination of demographic factors that influence their economic health has been an important topic (Lugomer, 2015). The majority of past works focused on determining the extent to which demographic and other variables influenced store performance.

Birkin et al. (2004), for instance, investigated a set of factors which determine convenience store performance of petrol forecourts and they consider demographic components of essential importance, as 80% weight has been assigned to them in the model. Among them, the local demand, including both local residents and workforce, has been considered as most important contributor to retail performance, followed by the variables such as car ownership, proportion of students and certain geodemographic groups coming from the ACORN classification.

Roig-Tierno et al. (2013), on the other hand, gave greater importance to location and competition. Genecon (2011) listed some economic factors, but also stressed the importance of demographics, acknowledging the fact that they have been frequently used in studies of that sort.

In a more focused study, Lugomer and Lansley (2016) researched associations between the demographic characteristics of both residential and workplace population in retail centre catchment areas, and their economic health. Demographic variables were found to account for up to about 24% of the variation of the economic health of retail centres. This means that, while local demographics cannot be used alone to explain the spatial variation of the health of retail centres, findings suggest that even a single demographic variable may explain around 20% of this variance. The characteristics of the nighttime (residential) and daytime (workplace) population were found to equally contribute to overall retail centre health and the most important variables were socioeconomic, having to do with the employment sector and level of education (Lugomer and Lansley, 2016).

It is indisputable that residential demographics play a very important role in understanding how well retail businesses perform and how they may be performing should they decide to locate a new store in a particular area. This notion sparked a wide commercial use of catchment modelling and assessment of catchment geodemographics in locational analysis (CACI, 2014; Geolytix, 2017).

### 3.2 Activity-Based Representations

Despite wide academic and commercial applications of geodemographic classifications and catchment modelling, the analysis can be only as good as the underlying data and census data have a number of limitations.

One of the most acknowledged drawbacks of census data is relative infrequency of update, with UK collection taking place every ten years and requiring additional two years for processing the data (Harris and Longley, 2002). In the meantime, particularly big and
dynamic urban areas are more likely than not to experience dramatic changes in population size and structure, even though literature suggests that some attributes, such as spatial patterns of poverty in inner London may change very little, even over the course of a century (Orford et al., 2002).

Next, census data fail to capture many daily patterns of human activity, as they are usually focused upon the characteristics of the residential or nighttime population. According to Longley et al. (2015b: 482): “Life chances and identity are shaped by household and residential structure, but also by interactions with employment locations and social and cultural facilities that are not local and are not constrained to nighttime residence.” There are two main reasons why focusing the acquisition and subsequent analysis of the census data only on the nighttime population is no longer good enough.

First, the majority of the applications of demographic data require knowledge of the local demographics during the daytime, rather than during the night. Retailing is probably one of the most illustrative examples, as working hours of most high street retail and leisure units are dictated by demand, and thus daytime hours.

Second, local population during the daytime may be very much different than the population that resides there. Most of the local residents work elsewhere and are therefore absent throughout the day, making limited or no purchases in the local shops. Any conclusions that are being made by analysing residential geodemographics could therefore be misleading and cause substantial revenue miss for a retailer who fails to acknowledge this discrepancy.

These drawbacks of census data based on the nighttime population motivated the invention of different alternative and enhanced types of demographic datasets that are either based on census yet again but are daytime-oriented, or they are based on proprietary commercial data sources that utilise relatively new technologies and acquisition methods. Both groups of data sources that aim to solve the problems with conventional census data are oriented towards activity-based representations because they honour the fact that areal characteristics change over time owing to their varying functions.

3.2.1 Time Geography

The temporal dimension has been incorporated in geographical research of human activities ever since the dawn of the concept of time geography conceived by T. Hägerstrand (1970), which recognises that individuals operate within temporal as well as spatial constraints on their behaviour (Miller, 1991). This pioneering concept concentrated on studying the movement of individuals throughout space and time and visualising them in the form of the space-time prism. Its most valuable aspect is that it allows the direct incorporation of accessibility into locational analysis (Miller, 1991).

The fact that people move throughout space as well as time, was utilised even in the more traditional forms of consumer data, such as previously introduced census data and
geodemographic systems. The need for going beyond the collection of data that describe residential or nighttime population was acknowledged and this resulted in the creation of new UK census geography that aggregates data on the daytime population (see Subsection 3.2.2).

Apart from enhancing the already existing, traditional sources of consumer data, some completely new opportunities emerged in the research of activity-based concepts of human behaviour. Miller (1991) noted the growing importance of geographic information systems (GIS) and presumed that various users would become increasingly able to solve problems related to the movement of people through space and time, as there would be more computing power available, as well as data on very fine spatial and temporal scales. At the turn of the century, this happened, fueled by the advent of new technologies and emerging Big Data and Smart City discourses. Analysing and understanding movements of people has therefore become possible on very granular levels, as H. Miller envisioned, and it was anticipated that these data would attain great commercial value.

Human mobility tends to be spatially and temporally highly regular, as people tend to have a set of frequently visited locations (Gonzalez et al., 2008a). That being said, people do not choose places of shopping and eating randomly, but their choice is based on the popularity of different places; or in other words, if a place is already popular, it will attract even more people (Hasan et al., 2013). The existence of patterns enables marketeers and researchers investigate the spatial processes that govern the functioning of urban areas.

The rest of this section and chapter is organised as follows. First, representations of daytime population are introduced (Subsection 3.2.2. Second, the new forms of consumer data are listed in the Subsection 3.2.3, while details of new technologies and data acquisition methods that enabled them to be collected are subject of the Section 3.3.

### 3.2.2 Representing Daytime Demographics Using Census Data

If a locational analyst observes a flow of people walking past a store in the town centre, some of whom eventually enter and make a purchase, the most striking questions that occurs to them are probably: “who are those consumers? What is their demographic composition?” It might seem sensible to try and estimate that information using the residential census data and apply spatial interaction modelling approaches as previously described. However, as said before, traditional census data gathered on the OA level only display who lives there. Given the fact that people commute to work, taking trips of various distances, the composition of people who work in the local area of interest and are, therefore, present there during the daytime, may be completely different.

These downsides of residential census data motivated the construction of the so-called Workplace Zones (WZs), which are geographical units similar in size to Output Areas, but which record the size and structure of the local workforce, rather than residents. In the
context of the census, a ‘workplace’ can be defined as a place of work recorded by a worker on their census form Cockings et al. (2015).

Workplace Zones (WZs) were designed from the subdivisons of the Output Areas called postcode-level building blocks as a new UK census geography and created in such way to reflect the density of workplaces (Mitchell, 2014). That being said, areas with high density of workplaces, such as Central London, may have many workplace zones comprising a single output area. On the other hand, predominantly residential areas, such as South London, may have a single workplace zone divided into many underlying output areas (Singleton et al., 2017a) (Figure 3.2).

![Figure 3.2: Spatial relations between Output Areas and Workplace Zones. Example of: (a) an area with a high density of workplaces (Central London) and thus many WZs in each OA; and (b) a primarily residential area (Sutton) where multiple OAs are contained within one WZ.](image)

*boundaries source: Office for National Statistics (2016c) and Office for National Statistics (2017)*

In the first instance, the Classification of Workplace Zones (COWZ) was created for England and Wales only, using the similar methodology to the one used for the Output Area Classification (OAC) (Gale et al., 2016). The reason for that was to enable the consistency between the classifications and to use previously obtained findings from the creation and applications of the OAC (Cockings et al., 2015).
The COWZ methodology was largely similar to the previously described OAC methodology developed by Vickers and Rees (2007). According to Cockings et al. (2015), it included choosing the set of meaningful variables by applying the domain knowledge and inspecting the variables’ spatial and statistical distributions. Variable transformations were applied and the trial set of variables was reduced after checking for multicollinearity. This was followed by the standardisation of the variables and adoption of k-means algorithms to generate top-level clusters. The algorithm was then repeated on the top-level clusters to achieve the lower-level clusters until newly generated clusters made little sense or made little contribution to the clusters interpretability (Cockings et al., 2015). After that, the identical methodology was iterated and applied to the census data from the entire UK, with some variables defined differently for Scotland and Northern Ireland (Office for National Statistics, 2018). This resulted in the COWZ-UK, which distinguishes between seven different daytime geodemographic supergroups (top level of the classification hierarchy, see Table 3.1) and 29 groups.

While results obtained this way are very similar when COWZ-EW (England and Wales) and COWZ-UK are compared, some differences in cluster membership and naming conventions do exist on a group (lower-tier) level, reflecting the differences in demographic characteristics of the entire UK on one side and England and Wales on the other side (Office for National Statistics, 2018).

In a similar fashion to LOAC and in line with the acknowledgements of importance of having purpose-built classifications (Singleton et al., 2016b), a special classification was also derived for those WZs situated in London, resulting in the so-called London Workplace Zone Classification (LWZC) (Singleton et al., 2017a).

However, even though workplace zones happen to be much more suitable census geography for the task of retail planning, some typical drawbacks of census data were not

<table>
<thead>
<tr>
<th>Supergroup label</th>
<th>Supergroup name</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Retail</td>
</tr>
<tr>
<td>B</td>
<td>City and business parks</td>
</tr>
<tr>
<td>C</td>
<td>Metro suburbs</td>
</tr>
<tr>
<td>D</td>
<td>Suburban services</td>
</tr>
<tr>
<td>E</td>
<td>Manufacturing and distribution</td>
</tr>
<tr>
<td>F</td>
<td>Rural</td>
</tr>
<tr>
<td>G</td>
<td>Servants of society</td>
</tr>
</tbody>
</table>

source: Office for National Statistics (2018)
eliminated by their introduction. Workplace zone data would still be updated every ten years and for most contemporary commercial applications, that is far from optimal.

### 3.2.3 Other Contemporary Sources of Consumer Data

In view of the limitations of the census as a data source for geodemographic analysis, attempts have been made to estimate population size through manually collected data, such as individual surveys (Zandvliet and Dijst, 2006; Jiang et al., 2012). However, while they enable the collection of data in more frequent time intervals compared to those based upon censuses, their conduction requires substantial financial and time resources, especially if they are undertaken on a large scale. In addition, even if a survey under consideration does capture large number of respondents, that number would still be just a small sample of what is collected during census data collection.

Recent technological advances have introduced many other new forms of consumer data, that go beyond both census and manual surveys and address their drawbacks. Such modern consumer datasets are useful because they are generated passively, for instance, simply by posting on social media or through store transactions; and this enables reliable stream of data on a range of attributes that are otherwise difficult and costly to obtain (Lansley and Cheshire, 2018). Examples of sources for such data are: social media (Adnan et al., 2014; Longley et al., 2015b; Lansley and Longley, 2016; Lloyd and Cheshire, 2017), GPS, Wi-Fi and Bluetooth tracking (Zheng et al., 2008; Vazquez-Prokopec et al., 2013; Chon et al., 2012; Malinovskiy et al., 2012), loyalty card transactions (Byrom et al., 2001a; Lloyd and Cheshire, 2017), travelcard usage (Ma et al., 2013; Lathia and Capra, 2011; Foell et al., 2014), smart meter data (Ushakova and Mikhaylov, 2016), online reviews (Ye et al., 2009) and many others.

All of these sources of consumer data can be used for representing population and their activities, reduce our dependence on theories, small samples, and official surveys and reveal otherwise unobservable trends about the population (Lansley and Cheshire, 2018). The amount and variety of newly available data opened up many interesting research avenues which commonly fall under the remit of the so-called Big Data and Smart City discourses.

### 3.3 Big Data and the Smart City

“We don’t yet know whether, and how, the adoption of ‘on the go’

3 technologies can boost town centre vitality.” (Wrigley and Lambiri, 2014)

This statement perfectly exemplifies that there is a great demand in applications of new technologies to finding answers to the burning questions in retail planning and management.

---

3 ‘On the go’ technologies imply everyday devices such as smartphones and laptops that are easily transferrable, connected to the Internet and can be used away from the desktop.
However, given the rapid pace at which technology develops nowadays, this statement has already been outdated at the time of the writing of this thesis. The toolsets of Big Data and Smart City discourses have made the research of dynamic processes shaping the rapidly changing spatial structure and function of urban areas not only possible, but also necessary to tackle a number of practical issues contemporary cities are facing. To mention a few: study of flows of people throughout urban areas, understanding how humans use space for commercial purposes, traffic congestion, air pollution, and many more (Becker et al., 2011).

3.3.1 Big Data and the Smart City: Definition of Concepts

There is no clear definition of Big Data, but initially, the concept was motivated by the fact that the amount of newly generated data has become so large that software engineers had to restructure the existing computational tools (Mayer-Schönberger and Cukier, 2013). That being said, there are three distinctive characteristics that make Big Data different from traditional data, often referred to as ‘3V’: volume, velocity and variety (McAfee et al., 2012; Zaslavsky et al., 2013).

Figures illustrating how voluminous Big Data are demonstrate are striking. Examples from 2012 state that 2.5 exabytes of data were being generated every day, with the number doubling approximately every 40 months; and that Walmart collects more than 2.5 petabytes of data from their customer transactions every hour (McAfee et al., 2012). There has been a marked transition from collecting data only about the small sample of individuals to data about almost everyone of interest, for example, all tap-ins and tap-outs at train station gates, all vehicles on a certain road, all pedestrians at a given microsite location, etc. (Kitchin, 2016). This definitely highlights the extent to which data acquisition has gone with the intensive worldwide use of the newest technologies.

When it comes to velocity, it is speed and efficiency of managing such quantities of data, that enabled us perform real-time or near real-time analyses (McAfee et al., 2012). This transformation from slow and sampled data to fast and exhaustive data has been enabled by the roll-out of a raft of new networked, digital technologies embedded into the fabric of urban environments that underpin the drive to create smart cities (Kitchin, 2016).

Finally, a third ‘V’, variety, has to do with the appearance of new types of datasets, many of which were mentioned in Subsection 3.2.3 on other modern forms of consumer data.

The technological advances resulting in the ability to collect such data coupled with the heavy institutional and organisational development spawned a new concept, termed ‘Smart City’, but also occurring in the literature under the names intelligent, digital, wired, creative, innovative and cultural city, tending to link the technological information revolution and economic, social and cultural change in the cities (Hollands, 2008). With the abundance of terms for the intrinsically same concept, there exist many different, but still commonly vague and problematic definitions of the smart cities (Hollands, 2008). Those theoretical debates
are out of the scope of this thesis and, for now, it suffices to say that the concept generally encompasses cities in which ICT is merged with traditional infrastructures, coordinated and integrated using new digital technologies (Batty et al., 2012).

### 3.3.2 Form Versus Function: the Analysis of Configurations and Activity Patterns

Leading role in the smart cities belongs to advanced sensing systems (Hancke and Hancke Jr, 2012). Since they are capable of collecting the data continuously and in large quantities, this gives rise to a new set of opportunities of researching urban processes, rather than focusing simply on forms. Measurement of urban form can yield generalised insights about the form, and, therefore, the functioning, of urban areas (Longley and Mesev, 2000: 487), however, sensors enable us to measure the processes directly, rather than having to rely on the inferences drawn from the structural characteristics of the urban landscape (cf. Hillier et al., 1976; Hillier and Hanson, 1989).

The latter idea has been enforced in the so-called space syntax concept, which is to a much lesser extent a geographical approach, and to the largest extent architectural concept. The main object of analysis within space syntax research is the configured space, or in other words, continuous space that has been turned into a set of topologically connected discrete units (Bafna, 2003). One of the key premises of spatial syntax is that sociologically relevant aspects of space can be identified by analysing the configurations of the built environment, i.e. morphology of the buildings, building blocks, streets and other morphological elements of the urban areas. This stems from the fact that advocates of the space syntax find topological relationships of component units more essential as compared to other relevant spatial attributes (Bafna, 2003).

Some works, particularly that of Ratti (Ratti, 2004b; Ratti, 2004a) have been quite critical of the tools used in the space syntax approach, which reportedly discards the precious metric information and it is also sensitive to boundary conditions, leading to difficulties in applications such as pedestrian modelling. Publication of his view was followed by the rejoinder of the conceptual pioneers of the space syntax (Hillier and Penn, 2004) and then Ratti’s further response (Ratti, 2004a) in which he continued highlighting some of the issues of the space syntax approach when analysing urban form, such as drawing conclusions that expected pedestrian flows may differ only due to minuscule geometrical differences in the configuration of the street.

The possible advantage of the sensor-based datasets, such as Wi-Fi sensors and smartphone-based data lies in the fact that they are capable of capturing both temporal variation of processes taking place around and within the otherwise static urban built environment. In other words, configuration of a certain high street goes through very little geometrical change in years, but at the same time, changes in the function, such as diurnal,
seasonal and long-term changes in footfall (Mumford et al., 2017; D’Silva et al., 2018), driving the retail sector and overall local economic vitality may have been at work and space syntax may not detect them. This further motivates human activity patterns research in a more direct sense.

Human activity patterns in urban areas have long been interesting to public institutions and private corporations, as their characteristics serve as an important factor of public policy and business strategy implementation, as well as the architectural design of the public spaces and buildings. As Utsch and Liebig (2012: 173) note:

“Everything provided to the guests depends on pedestrian movement. locations of information desks, shops or toilets depend on the reachability and quantity of persons, path-widths of the corridors in a stadium depend on people’s quantity as well, synchronization of digital signage and audio-guides depends on the average pedestrian speed, mobile phone networks are planned according to the expected movements and even locations of advertisement billboards are placed such that they achieve highest visit potential.”

The literature review now goes on to describe newer technologies that are commonly adopted for the measurement of footfall and, thus, description and explanation of the functioning of retail areas.

### 3.3.3 Measuring Footfall: an Overview of New Technologies

Even before the rise of the popularity of the Wi-Fi sensors and similar systems, cellular network data (calls and SMS traffic) had already been used to unveil city dynamics (Becker et al., 2011). ‘Call Detail Records’ (CDR), i.e. information on the location and time of calls can be used either alone, or combined with demographic data of the telecommunication companies’ customers to better understand their characteristics and spatio-temporal mobility (Reades et al., 2007; Becker et al., 2011; Ćamilović et al., 2009). However, this type of approach has coarser spatial and temporal granularity of identification of phone users. This is attributable to the fact that CDR is generated only when a call or SMS transaction has been made, and only with the spatial resolution of a cell tower sector, or with an order of magnitude of one square mile (Becker et al., 2011). Nevertheless, these shortcomings have not prevented a rather effective use of the cellular data transmission distributions in time and space (Pulselli et al., 2008; Isaacman et al., 2011; Pulselli et al., 2012).

More recently, research on how calls, SMS traffic and GPS signals can be used to estimate spatio-temporal mobility has been increasingly supplemented with the usage of Bluetooth and Wi-Fi technologies (Gonzalez et al., 2008b; Wu et al., 2013; Utsch and Liebig, 2012; Afanasyev et al., 2010). Wi-Fi network usage can be classified into three categories based completely on client device type: laptops, fixed-location devices and smartphones. Each
device type shows distinctive mobility patterns and geographic locality that follows distribution of functions of different areas of the city (Afanasyev et al., 2010). Wi-Fi sensor data may also be used to find out the total time smartphone owners spend in stores equipped with Wi-Fi sensors (Manweiler et al., 2013) or to analyse pedestrian flows more broadly (Fukuzaki et al., 2014; Schauer et al., 2014).

This subsection provides a brief overview of the three technologies commonly considered as options for measurement of footfall: Bluetooth, Wi-Fi and GPS.

**Bluetooth** is a wireless communication system designed for short range communication, which is normally up to roughly 10 meters, although it may be bigger. A device which wants to connect with another device via Bluetooth sends out an inquiry packet in the immediate environment. However, in order to respond to those packets, most devices need to be manually made visible by the user (Schauer et al., 2014).

**Wireless sensor networks** consist of wireless sensor nodes, which are devices equipped with a processor, a radio interface, an analog-to-digital converter, multiple sensors, memory and a power supply (Hancke and Hancke Jr, 2012). Devices such as smartphones attempt to connect to the Wi-Fi networks by consistently scanning the environment by sending out the probe requests to the wireless access points, such as sensor nodes (in the context of this report, simply referred to as sensors), at regular intervals varying between 30 and 120 seconds and depending on the device (Fukuzaki et al., 2014). Scanning for the surrounding wireless access points takes place even when a device is already connected to a certain Wi-Fi network, therefore searching for other available networks with possibly better signal (Bonn et al., 2013). At the same time, sensors send out beacon frames approximately every 100 milliseconds, advertising their presence (Schauer et al., 2014). Along with a couple of other technical pieces of information, probe requests contain data which uniquely identify the manufacturer of the device and the actual individual device, all contained within the 12-character address known as *Media Access Control* (MAC) address.

**Global positioning systems** (GPS) are navigation systems that makes use of satellites orbiting the Earth to determine the locations of objects and people carrying a GPS receiver, either as a stand-alone device or as a device embedded into other devices, such as smartphones.

Evidence in the literature suggest that all three technologies may be effectively used to estimate pedestrian flows. However, each exhibits somewhat different behaviour in terms of wave propagation, even at the same times (Dimitrova et al., 2012).

A study conducted by Schauer et al. (2014) showed that, when sensor readings and resulting estimate of flow of people is compared to the ground-truth, Bluetooth is significantly less accurate than Wi-Fi, to such an extent that it is unsuitable for the practical purposes of pedestrian flow estimation.

A study conducted at a major German airport resulted in a Pearson correlation coef-
icient between the Bluetooth estimate and the ground-truth being equal to 0.53, whereas correlation between Wi-Fi-based estimates and the ground-truth amounted to 0.61. Correlation was further improved up to 0.75 by incorporating a so-called hybrid approach of the improvement of flow estimation. This approach combines the time at which MAC addresses were recorded by the sensor A and sensor B, enabling direction determination and this is further supplemented by making use of the received signal strength indication (RSSI) threshold (Schauer et al., 2014).

“The pedestrian flow from node A to node B is then expressed as the number of unique MAC addresses in a specified time interval containing a time delay between the nodes and at least one capture with an RSSI value over a certain threshold for both nodes.” (Schauer et al., 2014: 173)

The same article also reports difficulties in choosing an appropriate threshold, however, this does not prevent this type of data from making it usable and useful.

The possibility of achieving a relatively strong correlation between the flows observed by the sensor and the ground-truth, however, without a perfect accuracy further motivates investigation of some of the factors that could be responsible for this deviation. Some important factors that are physical properties of radio wave propagation have been highlighted in research by Dimitrova et al. (2012):

“Radio signals coming from the Bluetooth and Wi-Fi devices propagate differently, as there are factors on the transmitter and receiver side, and during the propagation itself, affecting the transmission. Signal strength varies less between sensors of the same type than between mobile devices from different manufacturers, multipath propagation seems to have strong effect on signal strength and radio signals experience distinct propagation conditions in different directions.” (Dimitrova et al., 2012: 136)

The radio waves are affected by the presence of obstacles altering the radio wave trajectories and therefore modifying the signal strength. This is particularly relevant when the store or the shopping mall closes and shutters come down. However, adverse effects on the accuracy of measurements are apparent even during the store opening times, as certain obstacles are always present between a Wi-Fi sensor installed in a store and a smartphone owner whose movement is being recorded.

“The most common source affecting the wave trajectory distortion is metal equipment that induces huge signal reflections, preventing it from reaching areas theoretically within range. Devices functioning at frequencies close to Wi-Fi frequencies also distort the signal by covering it with great noise.” (Lassabe et al., 2005: 384)
As regards the propagation medium, humidity, potentially present in the air to a greater or lesser extent, impacts Wi-Fi signal propagation, as it absorbs part of the signal. However, this is unlikely to significantly degrade the Wi-Fi signal on the short ranges, on which Wi-Fi signals are typically propagated (Ireland, n/d).

Even though Wi-Fi has certain technological drawbacks, one of its key advantages, as compared to also rather popular GPS technology is that, for an effective tracking, GPS requires users to install specific applications, making it impractical for gathering larger quantities of trajectory data (Fukuzaki et al., 2014). At the same time, Bluetooth and Wi-Fi may be used without the awareness of the pedestrians if their devices have not been disabled\footnotemark{} (Schauer et al., 2014).

To a certain extent, large corporations such as Google may get around this limitation of a GPS technology due to the fact that their applications or even smartphones are used by a vast number of people. This enabled Google to launch the so-called Popular times and visit duration option, in turn allowing anyone to check the times when certain venues get exceptionally crowded, thus suggesting time for more enjoyable visit (Google, n/d).

The exact number of people, though, cannot be discerned from the histogram and graph only displays the information on the general temporal distribution for each day of the week, and, moreover, not for all the businesses of anyone’s interest, but only for those for which sufficient amount of data are available (Google, n/d).

Foursquare is another example of the Internet service which gathers a lot of geotagged data and makes it possible to analyse how popularity of different types of venues across cities change throughout time, with each location having a different temporal signature (D’Silva et al., 2018).

One of the drawbacks of using GPS in the research of activity patterns is that it can be seriously inaccurate when positioning is undertaken in the indoor environments (US Army Corps of Engineers, 2003). Wi-Fi has similar issue, which stems from the fact that sensors may capture probe requests from all the devices listening for the access points and they may be situated in different properties adjacent to the sensor location. This may lead to overenumeration of the pedestrians, or in other words, a variable number of false-positives.

3.4 Some Negative Side-effects of Big Data: Do We Know Everything Now?

It is apparent that the advent of new technologies and associated data acquisition methods and frameworks enabled far better understanding of local population from many different

\footnotetext{\(\text{This, on the other hand, spawns certain privacy concerns whenever there is a possibility of identifying an individual through the use of sensor data alone or by using an ancillary dataset from another source. See Subsection 3.4.2 on ethics considerations for a brief overview.} \)
perspectives. It has also enabled a range of different stakeholders, from retailing to transport and planning to make better informed decisions. However, with great power comes great responsibility and this is especially true with the processes of collecting, analysing and interpreting data, as well as presenting the results to the wider professional or academic audience or members of the general public. Being unaware, deliberately or accidentally, of misrepresentations that can be made by ill-informed approaches to dealing with newly emerged big geographic datasets, can result in costly erroneous business decisions or spread of misinformation. Some of the main issues that are explored here can be broadly fitted into two main categories: issues related to digital representations and uncertainty of spatial data and ethics violation issues.

3.4.1 Uncertainty and Consumer Data

The success of analysis conducted on any of the previously introduced traditional and new forms of consumer datasets will be largely dependent on their quality. Longley et al. (2015a) emphasise that spatial data, vector or raster, are never full representations of reality, i.e. they are always abstractions of a complex world and no matter what degree of detail is achieved when collecting them, the level of detail can almost always be even greater. A prime example used by Longley et al. (2015a) to demonstrate this is measurement of the length of Maine coastline. Drawing upon the fractal theory, the length of a given section can be anywhere between infinitesimal and infinite, depending on the resolution at which measurement takes place. Gathering, manipulating and analysing spatial data can be done in infinitely many ways and this has always presented a challenge for the researchers (Longley, 2012).

Since consumer data that are of interest to retail geographers are, as well, georeferenced, they also suffer from the same issues as the broader group of all spatial or all geographic data. According to Lansley and Cheshire (2018: 4) there are two main issues with consumer data representations: “Firstly, they are limited by demographic biases since no samples are objectively random. Secondly, as the data are often by-products of transactions of some kind, there are frequently data quality issues when they are repurposed.” In addition, the results of analysis can be strongly affected by outliers (Lansley and Cheshire, 2018).

Ironically, even though there has been a tremendous increase in the computational power and the amount of available data, as acknowledged in the previous sections of this chapter, we seem to be more uncertain about quality of their digital representations (Longley et al., 2015a). Analysis without accommodating data uncertainty can severely limit their usefulness, while on the other hand, proper understanding and conceptualisation of uncertainty enables more confident decision-making (Fisher, 1999). This is even more crucial given the fact that small errors may propagate through a model eventually resulting in unexpectedly large deviations from the true values in the model outputs (Evans, 2012).
Uncertainties occur in geographic data in three layers, cumulatively distorting the reality and therefore we can talk about uncertainty in conception, representation and the analysis of geographic phenomena. Conceptual uncertainty can be related to unclear definition of the spatial units of analysis or vagueness and ambiguity of the attributes (Longley et al., 2015a).

Different digital representations of geographic data in the conceptual forms of discrete objects and fields also result in different approaches to managing uncertainty, as does the belonging of the data set to nominal or continuous measurement scales (Longley et al., 2015a).

Finally, uncertainty also occurs in the analytical step of research and, therefore, results. Results may be validated internally or externally.

**Internal validation** can be done by cleaning the data so that, following the end of the procedure, results display the most realistic outcome. In contrast to internal validation, **external validation** is concerned with comparing the results to other data sources or the ground-truth (Longley et al., 2015a) and makes sure that internal validation has been done as thoroughly as possible.

It is important to stress in the end that error, often understood as a difference between value measured by the observers and instruments (Longley et al., 2015a) is often sought to be eliminated and does not have the same meaning as uncertainty. Apart from being a broader term, meaning of which has been discussed in this section from a geographic perspective, uncertainty in physical sciences is commonly also understood as a range of values in which the true value is asserted to lie with some level of confidence (NDT Resource Center, n/d).

In this thesis, one of the main tasks and focuses is on seeking a better understanding and attempts to eliminate the error of Wi-Fi sensor footfall measurements. This, in turn, means that while the process of making the measurements as close as possible to the ground-truth may be quite complex, we are still mainly aware of the underlying drivers of inaccuracy and are able to quantify the discrepancy between observed and true values.

### 3.4.2 Ethical Considerations

One can argue that even though we still do not know everything due to the imperfections of big data, this does not mean that, on the other hand, we did not start knowing too much.

The growing popularity and development of data mining technologies brought serious threat to the security of individuals’ sensitive information (Xu et al., 2014). Privacy is considered a basic human right in democratic countries, it is enshrined in national and supra-national laws (Kitchin, 2016) and therefore it needs to be particularly taken care of when creating, disseminating and using sensitive data for different purposes.

The companies and agencies who exploit new technologies possess a vast quantity of highly detailed spatial behaviour data from which lots of other insights can be inferred.
This is evident in most automatically collected ‘big datasets’, but has probably received most public attention when it comes to data collected online, either through informed or assumed consent. Informed consent implies obtaining full permissions from the study participants to use their personal data (Crow et al., 2006) and, as such, it is rarely achieved online. Web services, such as social media, normally obtain the so-called assumed consent from their users by asking them to agree to the terms and conditions of the service (Leak, 2017). In experiments where there is a greater control over the sample design, transparency can be more easily implemented as part of an ongoing informed-consent process, involving a session providing participants with information about what kind of data is being collected and what the goals of the study are (Harari et al., 2016).

However, large-scale studies employing automated data collection, such as social media, Wi-Fi analytics and other types of sensing systems can be more difficult when it comes to managing the user privacy. The issue is that Wi-Fi sensors, in particular, acquire probe requests from smartphone owners at all times, without asking them for consent. A common argument in favour of utilising such data acquisition methods is that no Personally Identifiable Information (PII) are being collected this way, but only technical information such as smartphone’s unique MAC address, signal strength and similar (see Chapter 4 for more technical information using the specific, SmartStreetSensor footfall dataset as an example).

Nevertheless, unless addressed in a more systematic way, there are ways one can still identify an individual due to high spatial and temporal granularity of such data. There has been a growing concern not only of the collection and misuse of the PII, but also of the fact that non-PII can become PII if various pieces of information are put together (Farshidi, 2016; Crawford and Schultz, 2014). As Leak (2017) notes referring to geotagged Twitter data, even though PII such as age, gender and ethnicity may not be displayed to the users who analyse the data, the high precision of GPS devices that come down to ±10 meters can enable us to infer this information from the location from which a person regularly tweets. When it comes to Wi-Fi, unless MAC address is randomised and regularly changed, it can be linked to an individual using other available data. For example, if a Wi-Fi sensor is installed within the store by the store owner, timestamps with MAC addresses can be linked to the transaction data at the tills, which can, especially during off-peak times, presumably lead to accurate identification of an individual. In addition, any kind of data being collected can, in certain circumstances, be accessed by the police and security forces through warrants and used for security and governance purposes (Kitchin, 2016).

In this project, Wi-Fi sensors installed within stores were used for footfall estimation, thus initially raising similar privacy concerns. The methods used to collect and process the data and the way privacy was preserved are described in Chapter 4.
3.5 Chapter Summary

This chapter reviewed traditional forms of demographic data used to estimate store site suitability, as well as new forms of consumer data and technological advances that made their acquisition possible.

There has been substantial academic and commercial interest in the field of geodemosgraphics which analyses and classifies people based on where they live. Knowing about the local demographic composition of potential consumers enables retailers to better plan locations of their stores and a range of spatial analysis methods have been developed over the past decades to estimate individual store and retail area catchments and their characteristics. New census geographies have recently been developed to solve the long lasting problem with residential census data that they describe the nighttime residential, rather than daytime population relevant for most commercial applications. This was, in a sense, perpetuation of the activity-based representations of human behaviour that have been gaining popularity ever since the Hägerstrand’s concept of time geography, which honoured the fact that people’s activities can be recorded not only based on their spatial coordinates, but also points in time, in the form of a space-time prism.

Towards the end of the 20th century, the opportunities that census data have to offer have become progressively exhausted and, at the same time, a need for more accurate estimates and indicators of footfall and catchment characteristics emerged. This demand was met by the dramatic increase in computational power of modern computers, the rise of the Internet and its services, increase in smartphone ownership across all age groups, development of different sensing technologies and other factors. The vast and varied new forms of consumer data were obtained quickly, continuously, at extremely granular spatial and temporal scales and at lower cost compared to the traditional surveys. The new datasets are varied and come from a range of different sources such as social media, smart meters, card transactions, GPS, Bluetooth and Wi-Fi tracking and many others. Rather than being constrained by a single snapshot of data as was the case with the census, it has become possible to measure how human activities fluctuate from day to day, from hour to hour and even from minute to minute, depending on the data source.

Nevertheless, these novelties are not without issues. Even though our knowledge about the world around us has been profoundly expanded in the era of Big Data and Smart City discourses, we are still unable to get all the answers as most data, no matter how big and varied, are collected for a predetermined purpose and even at that stage they are not free of sampling biases, measurement errors, outliers, missing data and similar problems. There are many uncertainty issues that need to be carefully taken into consideration and managed when analysing big consumer datasets and interpreting results.

On the other hand, even in the case when we know or attempt to know everything,
we may run into a brick wall of ethical questions, which sensibly prevent data owners from disseminating personally identifiable information to the third parties. The increased amounts of consumer data being collected, whether it is with informed or assumed user consent or without any consent at all, may leave consumers vulnerable to privacy violation and personal information leakages. Sufficient care must be taken to protect their data and achieve a balance between being too intrusive and failing to adopt innovations in data collection and analysis, to the advantage of both retailers and consumers.
Chapter 4

Data Description and Quality Considerations

The purpose of this chapter is to introduce the two main proprietary datasets that will serve as cornerstones of the analytical part of the thesis. Both datasets have been created and are maintained and updated by the Local Data Company (LDC). The chapter can be broadly divided into three sections. In the first part, LDC’s retail unit database in introduced, alongside with the corresponding metadata and the brief introduction into what that dataset tells us about the changing nature of the retail geography of Great Britain. The second section dives into the second big dataset that will be used in the thesis: the SmartStreetSensor footfall database. Finally, special attention is given to the footfall data quality issues, factors influencing them and proposals on how to rectify them.

4.1 Retail Occupancy Database

The retail unit data are collected by the LDCs team of field researchers every year and stored at the individual retail unit level (point data). In our case, data on the UK retailing sector for 2014 and 2017 will be used to analyse the structure and dynamics of British retail geographies\(^1\).

4.1.1 Retail Attribute Data

It is important to note that retail units contained in the dataset do not represent every single existing unit in the UK, but are the ones that form the previously defined retail areas. Since those areas are the local cores of retail and leisure industry, it can be argued that such sample is representative of retailing on a national scale and can be used to describe

\(^1\)The 2009 data were also available, however, they contained a much smaller sample, as will be described later.
its structural characteristics and temporal dynamics. The 2017 dataset contains the data about 546,722 units in Great Britain of which 523,260 are retail units and the remaining records are labelled as non-retail units on either a classification, i.e. top-tier level (these include shopping arcades, markets and shopping centres, i.e. not individual retail units) or a category level (for example, advice centres, community centres, architects, property consultants, etc.). The metadata of the 2017 dataset are presented in Table 4.1.

The data available to us were collected in three instances: in 2009, 2014 and 2017, although, they are not completely comparable, as the spatial extent captured by the LDCs field research grew over time. Consequently, the 2017 dataset contains substantially more records, but this cannot be solely attributed to the growth of the number of retail units over time, but it is the result of a larger amount of acquired data in the field research process. The 2009 database contains 282,686 records, while the 2014 database displays data on 479,266 units.

The individual retail units from the database can be spatially grouped to produce retail areas, i.e. areas of increased concentration of retail activity, as has been described in the previous chapter and using the methodology devised by Pavlis et al. (2018). Once we do that, we can start describing the general structure of UK retail areas or in other words, how frequently different retail classifications and categories occur across the country.

Out of 523,260 units recorded in the 2017 version of the LDCs retail database, 365,950 of them (69.9%) were situated within the boundaries of a retail area, whereas the remaining units were either remote stand-alone stores or were simply situated in the areas of too small concentration of retail units that hence did not qualify as distinct retail areas. There are 349,915 units are situated in town centres (high streets), whereas the remaining 16,035 units located are located in other retail area formats (shopping centres and retail parks). In 2014, there were fewer units (479,266) and only 247,076 were situated within retail areas and had full locational data (latitude and longitude) attached. Due to the incompleteness of the 2014 dataset, two snapshots are not completely comparable. However, general structural characteristics will still hold and are worth investigating. All subsequent analysis in this section was therefore performed on a subset of retail units that can be georeferenced and that are situated in the retail areas.

4.1.2 Spatial Hierarchy of Retail Areas in Great Britain

The distribution of the size of retail centres\(^2\) can be investigated in a similar way as the distribution of sizes of any other economic or urban agglomeration. Traditionally, urban geography recognises two types of distribution of the city sizes: rank-size distribution in which the distribution of the cities ordered by their population size is truncated log-normal

\(^2\)In this section, the term ‘retail centre’ is used instead of the term ‘retail area’ because the emphasis is on the hierarchy and role of retail areas as central places.
Table 4.1: Metadata for the Local Data Companys 2017 retail dataset

<table>
<thead>
<tr>
<th>Field</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupier ID</td>
<td>Nominal</td>
<td>A unique string of numbers representing the occupier of the retail unit</td>
</tr>
<tr>
<td>Business Name</td>
<td>Nominal</td>
<td>Name of the business</td>
</tr>
<tr>
<td>Postcode</td>
<td>Nominal</td>
<td>Locational information</td>
</tr>
<tr>
<td>Town</td>
<td>Nominal</td>
<td>Locational information</td>
</tr>
<tr>
<td>Classification</td>
<td>Nominal</td>
<td>The top-tier level at which LDC categorises businesses. The six classifications include: Convenience, Comparison, Leisure, Services, Miscellaneous and Non-Retail.</td>
</tr>
<tr>
<td>Category</td>
<td>Nominal</td>
<td>The second-tier level at which LDC categorises businesses (44 groups). For example, one of the categories in the convenience classification is named Groceries, Supermarkets &amp; Food Shops.</td>
</tr>
<tr>
<td>Subcategory</td>
<td>Nominal</td>
<td>The third-tier level at which LDC categorises businesses (438 groups). For example, previously mentioned category Groceries, Supermarkets &amp; Food Shops is broken down on the subcategory level into Convenience Stores, Supermarkets, Ice Cream Parlours, Grocers, Greengrocers &amp; Fruitsellers, etc.</td>
</tr>
<tr>
<td>Latitude</td>
<td>Numeric/continuous</td>
<td>WGS84 latitude of the given unit. Five-decimal precision.</td>
</tr>
<tr>
<td>Longitude</td>
<td>Numeric/continuous</td>
<td>WGS84 longitude of the given unit. Five-decimal precision.</td>
</tr>
<tr>
<td>Floorspace ($m^2$)</td>
<td>Numeric/continuous</td>
<td>Size of the unit in square meters (two decimals). Available for 211,347 or 38.7% of all the units (including the non-retail ones).</td>
</tr>
</tbody>
</table>

_data source_: Local Data Company (2017)
According to Zipf (1949), after whom rank-size distribution is frequently called, the rank-size distribution is formed by two opposing forces which he calls forces of unification and forces of diversification. These forces represent two opposite ways in which economical location of producers-consumers can be found out (Das and Dutt, 1993). Forces of diversification are related to the proximity of the raw materials to the centres of secondary production. When the economy is dependent on the sources of raw materials, a large number of smaller settlements emerges close to them (Zipf, 1949). However, at some point along the urban economic development timeline, demand for a greater variety of raw materials grows, and the probability of finding such diversity of goods at one location diminishes. At the same time, there is a need for minimisation of the cost of transport of the goods from the centres of production to the consumers and that cost is minimised in the large urban concentrations of population. Zipf terms those forces that influence the formation of a smaller number of larger markets, forces of unification (Zipf, 1949). According to Berry (1961), rank-size distribution occurs when many economical and political forces influence the urban pattern, whereas primate city-size distribution occurs due to a few, but strong forces.

Rank-size distribution can be spelt out by the following equation (Brakman et al., 1999):

$$R^q_j M_j = Co$$

where:
- $R_j$ is the rank of the city $j$;
- $M_j$ is the size of the city $j$;
- $Co$ is the constant;
- $q$ is the parameter that can be described as a ratio between the Zipf’s Forces of Diversification and Forces of Unification$^3$.

If we order British retail centres by size (the number of retail units within them) and plot the national rank of each retail centre against its size on a double logarithmic plot, we observe a pattern displayed in Figure 4.1$^4$. It can be discerned from the plot that retail centres in Great Britain display a pattern resembling rank-size distribution, at least for the largest centres. The largest retail centre, Central London, is more than two times as large

---

$^3$Typically, if $q=1$, the literature tends to term such distribution Zipf’s distribution and if $q$ differs from unity, a broader term rank-size distribution is used (Brakman et al., 1999).

$^4$Note that only the largest retail areas, i.e. centres, containing at least 100 retail units were displayed on a plot. The reason for that is that distribution starts falling sharply after that due to a relatively large number of very small centres that have fewer than 100 units, but over 10 units - a threshold usually used to define a retail centre (Wrigley and Lambiri, 2015).
as the second largest centre. After the sharper drop with several largest centres, the size continues to decline with rank at a slower rate, until it reaches a point around 100th rank where sharper drop occurs.

![Graph showing rank-size distribution of retail centres in Great Britain (2017)](image)

**Figure 4.1:** Rank-size distribution of retail centres in Great Britain (2017)  
*data source:* LDC retail database (2017)

It is also interesting to try and divide the set of retail centres by size into several meaningful hierarchical categories. Central London was excluded, as it includes 7101 units, which is more than twice as much as in the City of London retail centre. The remaining retail centres were clustered using k-means clustering algorithm. The number of clusters was set to four, as suggested by the elbow method (Figure 4.2). It can be seen from Table 4.2 that the number of largest centres is the smallest, whereas the number of smaller centres is larger. This is in line with expectations from the central place theory devised by W. Christaller (Getis and Getis, 1966).
4.1.3 Structural Characteristics of Retail Centres in Great Britain

What is the composition of the typical retail centre in Great Britain and how did that change over the course of several post-recession years (2014-2017)? The purpose of this subsection is to continue the description of retail data by inspecting structural characteristics of retail centres and their underlying formats: town centres, shopping centres and retail parks.

The first thing to consider structure-wise in the 2014 and 2017 retail datasets is to

---

**Figure 4.2:** Optimal number of retail centre size clusters (elbow method)

**Table 4.2:** The number and average size of the four hierarchical categories of retail centres in Great Britain

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Number of centres</th>
<th>Average size</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>18</td>
<td>1359</td>
</tr>
<tr>
<td>B</td>
<td>192</td>
<td>483</td>
</tr>
<tr>
<td>C</td>
<td>714</td>
<td>195</td>
</tr>
<tr>
<td>D</td>
<td>2325</td>
<td>44</td>
</tr>
</tbody>
</table>
observe the general nationwide structure of the top-tier categories (classifications) of retail industry in Great Britain. According to the LDC data, comparison stores were the most frequent amongst the retail centres of Great Britain in 2017, making (28.7%) of all the retail units (Table 4.3; Figure 4.3). They are closely followed by services (25.4%) and leisure units (23.4%). Discounting miscellaneous classification, convenience appears to be significantly under-represented as compared to the other main groups of the retail industry. This probably is not so unexpected, again given the spatial coverage of the LDCs dataset, which is inherently targeted towards understanding the town centres along with the shopping centres and retail parks. Since convenience stores have lower gravitational power, there will be many of them throughout Great Britain, but at the same time, many of them will be situated in the residential areas outside of town centres. For this reason, they are represented by a small proportion in this sample. The remaining three classifications (comparison, services and leisure) are similarly frequent. A period between 2014 and 2017 saw a slight decline in the percentage of vacant units and comparison stores. On the other hand, services grew by 2.2 percentage points, and leisure units increased by 1.1 percentage points.

Table 4.3: The structure of retail centres in Great Britain (2014 – 2017)

<table>
<thead>
<tr>
<th>Classification</th>
<th>Percentage (%)</th>
<th>Change (percentage points)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2014</td>
<td>2017</td>
</tr>
<tr>
<td>Comparison</td>
<td>30.9</td>
<td>28.7</td>
</tr>
<tr>
<td>Service</td>
<td>23.2</td>
<td>25.4</td>
</tr>
<tr>
<td>Leisure</td>
<td>22.3</td>
<td>23.4</td>
</tr>
<tr>
<td>Vacant</td>
<td>12.9</td>
<td>10.4</td>
</tr>
<tr>
<td>Convenience</td>
<td>8.8</td>
<td>9.2</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>1.9</td>
<td>2.9</td>
</tr>
</tbody>
</table>

*data source: LDC retail database (2014, 2017)*
Taking a step further, if we consider the most frequent retail categories (lower hierarchical level than classification), their order was similar in 2014 and 2017 (Figures 4.4 and 4.5). The main difference was a markedly higher percentage of vacant units on the national level in 2014, which made the category ‘vacant units’ more frequent in the British retail space than any other retail category. Disregarding vacant units, the most common retail categories in both years were hairdressers, health and beauty, cafes and fast food, fashion and restaurants. The only other difference between the two years is the fact that estate agents jumped from the 10th to the 8th position in 2017.

Figure 4.3: The structure of retail centres in Great Britain in 2017: top-tier level (classifications)

*data source:* LDC retail database (2017)
Figure 4.4: Ten most common retail categories across British retail centres (2014)

data source: LDC retail database (2014)
Finally, we can relate the structural characteristics of retail centres with their size. In the previous subsection, centres were classified into four hierarchical groups based on their size. If we compute the average composition of every hierarchical level (Central London excluded from the sample), we end up with the structure shown in Figure 4.6. The largest retail centres in the country (labelled as level A centres) have a decisively larger proportion of leisure units than other hierarchical levels and a smaller proportion of services. The smallest, local retail centres (C and D) are characterised by having the highest percentage of convenience stores and smaller percentage of comparison and leisure. This is in accordance with the expectations that emerge from the traditional spatial interaction models, which state that larger (and more attractive retail centres) have larger catchments and attract consumers from further areas (Dolega et al., 2016).
4.2 SmartStreetSensor Footfall Database

The second main proprietary dataset collected and maintained by the LDC is the SmartStreetSensor footfall database. It contains detailed measurements of the number of pedestrians walking past hundreds of different urban locations across England, Scotland and Wales that were acquired by the Wi-Fi sensors. In this section, a detailed overview of the project and the dataset is given.

4.2.1 The SmartStreetSensor Project Background

The project that this research is part of has been named ”The SmartStreetSensor Project” and results from a collaboration between University College London Consumer Data Research Centre (UCL CDRC) and industrial partner Local Data Company. The aim of the project is to measure footfall, i.e. the number of pedestrians walking in front of the stores across Great Britain using the Wi-Fi sensors and the fact that many pedestrians own a smartphone. The phone of each pedestrian periodically scans for Wi-Fi networks, and it does so by sending out probe requests in the surrounding space, some of which are then cap-
tured by the Wi-Fi sensor (Figure 4.7). The probe requests can then be summed and used to estimate the actual footfall at the given locations. However, the initial counts of probe requests have to go through a rigorous set of internal and external validation algorithms that will be introduced later in the chapter.

![Image](image.png)

Figure 4.7: LDC sensors used for the data acquisition

*source:* Local Data Company (2015)

As the Figure 4.7 illustrates, sensors are most often installed as close as possible to the shop window inside the retail properties that have entered into either an individual or collective agreement (for example, the entire shopping centre) with LDC. The reason why the project focuses on the footfall estimation in front of the store rather than inside the store is that opportunities arise for comparing the footfall outside the store with the stores’ sales data. This potentially enables retailers to better understand conversion rates, i.e. what proportion of people that walk in front of the store enter the store and make a purchase.

The Wi-Fi sensors installed by the LDC’s field researchers were not randomly scattered across the country. The next subsection focuses on the spatial distribution of sensors and explains the sample design created by the CDRC which guided the sensor roll out.

### 4.2.2 The Sample Design and Spatial Distribution of Sensors

On 23 October 2017 there were 914 sensor locations recorded in the database. However, some of those sensors were not even active, some of them were active, but faulty, i.e. their measurements of footfall were later found to deviate too much from the ground-truth
regardless of the type of data cleaning or calibration employed and some sensors simply
had temporally unreliable measures, meaning that they go off at irregular intervals for
sometimes unpredictable amounts of time (e.g. when they are accidentally unplugged, or
shut down by an electric fault), but sometimes predictable (e.g. when the power is switched
off after the store closes). Therefore, the number of sensors which was eventually used for
all the subsequent analytical work of this thesis was comparatively smaller and amounted to
605. The number of sensors with valid measurements is actually higher than that, but the
period that was taken into consideration ran from 31 July 2015 till 31 August 2017 and some
sensors were installed very close to the latter date, meaning that there have not been enough
data for reliable conclusions about the footfall volume and patterns at those locations. The
reasoning for the choice of that particular time frame and detailed explanation of the data
processing steps will be provided later in this chapter.

For now, let’s start by outlining where those 605 sensors are located and what guided
the decision to install them there. It can be argued that choosing a viable set of locations
for the sensors is instrumental to the quality and reliability of results. For example, it would
make very little sense to locate all the sensors in the same borough of London or to locate
all the sensors in the shopping centres and arrange no installations in the town centres.
The objective should be to spread the sensor rollout across (a) the territory of the UK,
(b) different formats of retail areas, (c) different retail chains, (d) areas of different local
residential and workplace demographics and (e) areas of different local economic conditions.
That being said, before the sensor roll out started, the collaboration between CDRC and
LDC incorporated devising a locational sample design for the potential 1000 sensors. The
sample design for the locations outside London was created by L Dolega and A Singleton
of the University of Liverpool, whereas sample design for London was created by J Kandt
and P Longley of UCL before this PhD project commenced.

The criteria for the sample locations outside London (shown in Figure 9.1 in the Ap-
pendices), were dominant Output Area Classification (OAC) Supergroup\(^5\), town centre size
expressed by the number of businesses and town centre type (regional capital, market type
and other). The idea was to encapsulate town centres with a variety of size of the retail
supply and different geodemographic compositions.

The primary criterion for the locations in Greater London was, on the other hand,
population size of town centres’ respective catchment areas\(^6\) (Figure 9.2 in the Appendices).
At the same time, even coverage of different parts of London was also aimed to be achieved.

The total number of sample town centres was 63 out-of-London English locations (48
proposed locations and 15 ‘reserves’ in case the project expands, or if retailers become

---

\(^5\)As noted in the previous chapter, OAC is the geodemographic classification of residential population
based on the 2011 UK Census of Population.

\(^6\)The catchment areas for the national set of retail agglomerations were created by making use of spatial
interaction models, as presented by Dolega et al. (2016).
interested in installing additional devices in new areas), 21 town centres in London and additional couple of locations in Scotland (Glasgow, Edinburgh, Aberdeen and Stirling) and Wales (Cardiff, Merthyr Tydfil and Mold).

The next task is to further describe the combined consequences of the sample design and the actual sensor rollout. In other words, what types of locations were eventually covered and which types of locations are wholly unrepresented by the sample? This kind of description is important, as it highlights what kind of footfall and what portion of the British retail geographies this thesis investigates and conversely, what this thesis is not about, given the imperfect nature of any sample?

The sensor sample can be described in a multitude of ways:

1. by analysing the spatial distribution of sensors on a national scale;
2. by analysing how they are distributed across different categories of retail units;
3. by analysing how they are distributed across different spatial formats of retail areas (town centres, shopping centres and retail parks);
4. and by analysing how they are distributed across different structural types of retail areas.

**Spatial Distribution of Sensors**

First, the spatial distribution of installed sensors will be described. The first sensor, situated in Hammersmith, London, was installed in July 2015, with the total number of sensors gradually increasing. The majority of devices has been situated in Greater London (226, or 37.4%), with London Borough of Westminster having 72 devices installed, followed by Camden with 397. The spatial distribution of 605 sensors which are relevant for all the later analysis is displayed in Figure 4.8.

---

7See Figure 9.1 in the Appendices for the number of installed sensors in each Local Authority and Council Area in Great Britain.
Figure 4.8: Spatial distribution of sensors in Great Britain and Greater London
In terms of regional differences, 86.0% of the sensors were placed in shops in England, and it is important to note that 80 different local authorities have been covered. Scotland was covered by 74 sensors (12.2%) and Wales was only represented by 11 sensors. While the number of council areas/local authorities covered in Scotland and Wales was comparatively small (Edinburgh, Glasgow, Aberdeen, Aberdeenshire and East Renfrewshire in Scotland; and only Cardiff in Wales) the number of sensors in the larger cities there fares well when compared to other major cities in England, excluding London.

It is important to note that while the design was made and followed as much as possible, the actual sensor roll out did, in practice, depend on the generated interest from different retail chains and their local unit managers in every retail area. That is the main reason why a substantial number of sensors (29) was installed in the Ridings Shopping Centre in Wakefield, even though that was not envisioned by the initial sample design. In addition, on a micro-scale, even if there is an interest from a store manager to install the sensor inside the store, there may be some other obstacles preventing it, such as no electrical plug near the storefront. However, despite such pragmatic problems, initial sample design did achieve a spatially varied sample of sensors.

**Distribution of Sensors Across Categories of Retail Units in Which They Were Installed**

Next, the breakdown of retail categories by the number of sensors was assessed, and results are given in Figure 4.9. Of 605 considered locations in 2017, 21 sensors are situated in the vacant properties and the remaining 584 are situated in the retail units of known retail category and subcategory.

The most common retail category are Charity shops, comprising just over 16% of the retail units in our sample of microsite locations with installed sensors. They are followed by Fashion and general clothing shops (12.4%), Restaurants (10.4%) and Sports & hobbies shops (9.9%).

**Distribution of Sensors Across Different Retail Area Formats**

The selected sample of installed sensors is rather unevenly spread across three main retail formats. There is a decisively large number of sensors installed in the town centres (high streets) - 556 or 91.9%. The remainder of the sensors were located in the shopping centres (46 sensors), and only three sensors were situated in the retail parks. This already tends to suggest that this research will mainly be targeting functional characteristics of retail areas and their associated footfall patterns within the central parts of urban areas, rather than out-of-town shopping centres or retail parks.
Distribution of Sensors Across Different Structural Types of Retail Areas

Finally, sensor locations can be characterised not only based on the type of the retail unit in which they are situated but also based on the retail geography of their neighbourhood. While some general characteristics of the British retail economy had already been introduced in the previous section, in this instance, we will use the retail area classification which was created by the Consumer Data Research Centre (CDRC) Liverpool research team (CDRC Liverpool, 2018). The classification takes into account the structure of the retail occupancy (presence of different subcategories of stores), vacancy rates and crime and, although unpublished at the time of the writing of this thesis, its results were made available to us. The purpose of introducing this classification is to provide an overview of what structural types of retail environments were found to exist in Great Britain, with emphasis on what types of retail environments are particularly covered by the sensor sample design. After introducing it, references will be made to it from time to time in the later chapters in which some other useful datasets (primarily LDCs SmartStreetSensor and 2011 Census of Population) will be linked with it with the goal of expanding our understanding about the functioning of retail areas and changing nature of British retailing.

CDRC Liverpool’s retail area classification yields five groups and fifteen subgroups. The five groups are (1) Local Retail and Service Centres, (2) Retail, Shopping and Leisure Parks, (3) Leading Comparison and Leisure Destinations, (4) Primary Food and Secondary Comparison Destinations and (5) Traditional and Diverse High Streets (CDRC Liverpool,
Figure 4.10: The number of installed sensors across CDRC Liverpool’s retail centre subgroups


A detailed breakdown of the sensor retail group and subgroup memberships is given in Table 9.3 in the Appendices and a summary is presented on Figure 4.10. It is immediately apparent that the sample of sensors covers some of the subgroups more than others and some groups and subgroups are not represented at all. Only groups 3 (Leading comparison and leisure destinations) and group 4 (Primary food and secondary comparison destinations) are represented by more than five sensors. On a subgroup level, six of the fifteen subgroups from the retail area classification had less than five retail areas.

Out of 605 sensors, 489 of them were installed in the retail areas which were covered by the CDRC Liverpool’s retail area classification. This is not exhaustive, but main conclusions about the sample characteristics are still not expected to differ much from the case in which all sensors were included.

It can be seen that Local retail and service centres and Traditional and diverse high streets are vastly under-represented by the sensor sample, especially taking into account that they are the most frequently encountered retail groups across the British retail space. The former group is characterised by a relatively high prevalence of independent and small multiples with the absence of larger corporations. The latter group, on the other hand,
constitutes retail areas in suburban and rural areas which specialise in traditional grocery shopping and household services such as butchers, bakers, etc. In contrast to those under-represented groups, summary descriptions of the six most represented subgroup of retail areas in which sensors are located are given in Table 4.4.

Table 4.4: CDRC Liverpool’s retail area classification pen portraits of the selected groups (CDRC Liverpool, 2018)

<table>
<thead>
<tr>
<th>Group/subgroup</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group 3</strong>&lt;br&gt;Leading Comparison and Leisure Destinations</td>
<td>These are the areas of premium retailing characterised mainly by the prevalence of comparison and leisure destinations. They typically serve large catchment areas.</td>
</tr>
<tr>
<td>Subgroup 3.1.&lt;br&gt;Premium Shopping &amp; Leisure Destinations</td>
<td>They have a particularly diverse retail occupancy structure comprising different comparison, convenience and leisure retailers, with low vacancy rates and a higher share of department stores, upmarket fashion retailers, jewellers, health and beauty salons, and also services such as banks, travel agents, real estate and letting agents.</td>
</tr>
<tr>
<td>Subgroup 3.2.&lt;br&gt;Mass and Value Destinations</td>
<td>The local retail geography of such retail centres incorporates discount stores, charity shops, bookmakers and fast food takeaways, with a lot fewer premium restaurants, popular coffee shop chains and anchor stores. Catchments of these retail areas are less affluent and characterised by higher unemployment rate.</td>
</tr>
<tr>
<td>Subgroup 3.3.&lt;br&gt;Premium Out of Town Destination Areas</td>
<td>They are more affluent than Mass and Value Destinations and are characterised by the higher share of department stores, upmarket fashion retailers and national coffee shop chains.</td>
</tr>
</tbody>
</table>
Group 4  
Primary Food and Secondary Comparison Destinations

<table>
<thead>
<tr>
<th>Subgroup 4.1. Vibrant Urban Destinations</th>
<th>Group Primary Food and Secondary Comparison Destinations contains areas with relatively less affluent catchments and higher vacancy rates. Their function is mainly convenience and leisure and to a lesser extent comparison, unlike group 3.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgroup 4.2. Diverse District Destinations</td>
<td>These have a lower vacancy rate but are also less affluent on average. Both low-end and high-end retailers are less common.</td>
</tr>
<tr>
<td>Subgroup 4.3. Urban Value Destinations</td>
<td>They have more affluent catchments, more retail diversity, but still relatively high vacancy rates.</td>
</tr>
<tr>
<td></td>
<td>Less diverse, have higher deprivation levels and higher vacancy rates compared to Diverse district destinations.</td>
</tr>
</tbody>
</table>

Overall, we can state that the sensor sample design that was created and then applied during the sensor roll-out did confine sensors only to particular types of places. This means that signals that sensors measure and that will be used for the analysis of the temporal distribution of footfall in later chapters are not representative of neither the whole urban areas of Great Britain nor all retail cores of urban areas in Great Britain nor all retail area formats. Sensors from the SmartStreetSensor project primarily measure footfall in town centres (high streets) and to a limited extent shopping centres, while retail parks are almost completely excluded. The sample does not cover rural areas and suburban areas are vastly under-represented. The focus of the project is on the high streets that are either leading comparison and leisure destinations or primary food and secondary comparison destinations. We can also say that there is an overrepresentation of some of the categories of stores in which sensors are installed (particularly, charity shops, fashion retailers, restaurants and cafes and fast food), however, that bias is not going to be relevant because we aim to measure footfall in front of the store, hence not taking into account the footfall within the interior of the retail units.

Now that we described where and why sensors were installed, the next step is to explain what exactly constitutes the sensor database and what kind of data are being collected.
4.2.3 Probe Requests

Wi-Fi probes are detected and recorded by the sensor and sent to the server run by LDC. Data comprise the following components:

- device ID
- location ID\(^8\),
- timestamp (five-minute bins),
- MAC address
- signal strength
- the number of packets received from the detected device.

MAC address consists of twelve digits (48 bits). The first six digits are referred to as Organisationally Unique Identifier (OUI), as they uniquely identify the vendor of the device. The remaining six characters uniquely identify the device together with the OUI (What Is My IP, n/a). For instance, if a network adapter bears the MAC address 00-14-22-01-23-45, then the first six digits (00-14-22) represent OUI, which is tied to Dell Inc. (What Is My IP, n/a). This is an important piece of information contained within the MAC addresses, which will be utilised later during the sensor data cleaning procedure.

To preserve users’ privacy and to comply with newly adopted privacy preservation standard in the European Union (General Data Protection Regulation or GDPR), no additional, Personally Identifiable Information (PII) are collected, and MAC addresses are hashed at the sensor level before being sent to LDC cloud storage. From there, through a secure channel, they are sent to the CDRC secure servers, and further analysis is conducted there, using hashed addresses (Murcio et al., 2018). The data that are disseminated to the interested third parties, such as retail companies or academics, do not include MAC addresses or other sensitive information, as the only information being disclosed are estimated footfall counts on the five-minute or hourly temporal resolution.

Once available at the CDRC secure servers, probe requests can be aggregated into raw counts and we can inspect the temporal variation of the devices that were captured by the sensor, i.e. raw counts. These do not constitute footfall because at this point we are unsure what proportion of the captured devices are smartphones carried by the pedestrians that walk in front of the store. An example is provided displaying the short-term variation of the total raw counts of all the MAC addresses captured in five-minute intervals from Saturday

\(^8\)Locations are defined by the latitude and longitude and are important field, because some devices were installed at one location at one time and then they were deinstalled and installed at some other location. In such cases, it is important to use the data stream connected to a particular location, as footfall counts and patterns for a particular device might differ after its relocation.
Figure 4.11: Temporal variation in the captured devices in Bayswater, London, 20 – 30 August 2016 (five-minute intervals)


20 August 2016 until (and including) Tuesday 30 August 2016 in Bayswater, London. Time and location were chosen to capture both diurnal and day-to-day oscillations, a bank holiday (last Monday in August), as well as a major event.

The most prominent feature of the temporal variation of the detected devices are marked peaks on Sunday and Monday 28 and 29 August, respectively, when Notting Hill Carnival took place in the neighbourhood (Figure 4.11). A week before, on Sunday, as well as any other day, detected footfall was not as high, proving that such a high footfall has been caused by a rare event.

By taking aggregated hourly raw sensor counts instead of five-minute aggregates, distributions become more clear with less noise, while at the same time, specific aspects have been preserved (Figure 4.12). It can now be more easily seen that maxima during the workdays are higher than those recorded during the usual weekends (first Saturday and Sunday on the diagram). Among the workdays, Tuesday and Wednesday record highest peaks. Both diagrams very clearly show unusually high activity during the night between Monday (bank holiday) and Tuesday, at the very right. This is most likely due to the afterparties organised after the carnival.

Even though displayed plots only present a basic insight into footfall and they contain
counts of all the MAC addresses, including those that are of no interest\(^9\), results seem to be nevertheless interesting and corresponding to the reality.

### 4.3 Addressing Sensor Measurement Errors

To what extent do Wi-Fi sensors provide accurate counts of the number of passers-by, as validated by field observation and how can we identify and hence account for systematic discrepancies? This section describes the methodological framework for data cleaning and calibration.

Specific technical terminology will be used in the remainder of this chapter, raising a need for a brief overview of the terms and their definitions. Some of the most important terms used later on are defined as follows:

**Raw sensor count** is the total number of all the MAC addresses recorded by the sensor in a given period.

**Processed sensor count** is the generic term here used to describe sensor counts after

\(^9\)Plots presented later in the chapter will also demonstrate that patterns remain similar even after the data cleaning procedure and the only component which changes to a larger degree is the volume of flow, i.e. counts of the passers-by.
additional data cleaning, i.e. taking out the devices that are not smartphones, removing those devices that dwell at the same location for a longer period of time and estimating the number of devices that employ MAC randomisation.

Manual count or calibration count is the number of passers-by counted manually on site by a field researcher.

Adjustment factor or calibration coefficient are both terms used to describe a ratio between the manual count of pedestrians in a given period and the corresponding processed sensor count. If the adjustment factor is higher than 1, that means that there are more passers-by than the sensor detects in the same period, or in other words, undercounting is at work. If the adjustment factor is lower than 1, this means that the sensor count detects more devices than there are passers-by of interest, or in other words, overcounting is taking place due to the high relative number of other devices in the within the sensor range.

Measurement error is, here, defined as a deviation of a sensor count from a manual count, expressed in the form of percentages. While squaring of those deviations may be done, this is mostly avoided here, as the preservation of sign in front of the number representing the error is important for finding out whether the certain sensor is overcounting or undercounting at a specified location.

Private MAC address is a MAC address whose manufacturer can be identified through the vendor part of the address (OUI).

Public MAC address is a MAC address not specifically tied to any manufacturer and can be used by any device. Since we are unable to identify them, the best initial step in managing them is to remove them completely. However, since smartphones that use MAC randomisation also appear in the pool of these addresses, their number may need to be estimated and added back to the sensor counts. Public MAC can be identified from a second character of the vendor part of the MAC address.

4.3.1 The Sources of Measurement Error in the Wi-Fi Sensor Data

While Wi-Fi sensors may be used to estimate pedestrian flows, resulting measurements are in practice rarely, if ever perfect. The first step in understanding and addressing the error inherently present in this type of data should be the identification of the overall set of factors contributing to the deviation of the sensor measurements from the actual footfall.

That being said, there are two main groups of sources of error governing this discrepancy. Here, they are termed overcounting and undercounting factors and their breakdown is shown in Figure 4.13:

10There have been some cases where manufacturer can be identified even though device holds a MAC address falling into the range of public addresses, however, these cases are very rare.
Figure 4.13: The sources of error in the Wi-Fi sensor data

Overcounting factors cause sensors to count more MAC addresses than there are pedestrians passing by the front of the store. They comprise devices inside the store, as well as devices in the residential properties and offices in the immediate surroundings of the sensor. Those devices are not limited to smartphones, but also include printers, scanners, computers, wireless access points, etc. The key advantage to managing these sources of error is the possibility to programmatically remove them from the total footfall estimate by detecting vendor parts of MAC addresses belonging to the manufacturers of devices other than smartphones and by detecting and removing MAC addresses that remain near the sensor throughout longer periods. However, problems persist because some vendors have, in recent years, introduced randomisation of MAC addresses of their customers’ devices. For example, Apple introduced randomisation in iOS\textsuperscript{11} release, however, since this was only effective while the device was in the sleep mode, a more rigorous randomisation procedure was introduced with the launch of the iOS 9 (Vanhoef et al., 2016). Android devices of version 6.0 and above also perform randomisation for background scans (Android Developers, 2015). For these reasons, pedestrians owning a phone which randomises MAC need to be

\textsuperscript{11}iPhone Operating System
modelled separately.

Turning back to Figure 4.13, while partial or almost complete solutions exist for tackling the sources causing the sensor overcounting, the undercounting factors are much more difficult to manage. They were explored more thoroughly in Chapter 3 and include the proportion of pedestrians that either do not carry a smartphone, or have the Wi-Fi turned off, as well as any pedestrian of interest that has not been detected due to the earlier reviewed decay of signal strength with distance, presence of the physical obstacles in the store interior, as well as the variation in wave propagation (Jiang et al., 2015; Dimitrova et al., 2012). One important thing to note is that persons owning an Android smartphone of version 4.3 and above may be detected even if their Wi-Fi has been switched off, as long as the option of ‘Wi-Fi scanning’, which serves to improve the locational accuracy has been turned on (Murphy, 2013). Since this is default behaviour, it can be presumed that turned off Wi-Fi may not be as significant an issue as initially thought, especially given the high presence of Android on the smartphone market\textsuperscript{12}. However, the problem with it is that its impact is extremely difficult to estimate.

### 4.3.2 Internal Validation

Before calculating the measurement error, a decision needs to be made on the internal validation process. In other words, as has been acknowledged in the preceding section where key terms were defined, relevant MAC addresses taken from the sensor can be processed in different ways. Ideally, we want to make sure that only smartphones of pedestrians are detected, however, this is never going to be perfectly possible. Instead, the internal validation process seeks to minimise the possible error caused by the overcounting sources of error.

This said, the raw sensor counts were processed through the established internal validation methodology, devised by the team of LDC data scientists, after a series of discussions with UCL researchers, including myself. However, as explained in Section 1.3, applications of this methodology laid out below (Bayswater sensor example and initial measurement error calculations), as well as any other data processing steps conducted using this agreed methodology constitute original work. The internal validation methodology was structured as follows:

- All devices whose vendor part of the MAC address does not correspond to the relevant smartphone manufacturer were taken out. This removes distracting devices such as printers and WAPs.

- Next, if the device occurs at a location more than once and period between the first

\textsuperscript{12}According to IDC (2016), Android dominated the smartphone market in the third quarter of the 2016, with 86.8% prevalence rate.
and next occurrence amounts to ten minutes or less, that device will be counted only once. For example, if a particular MAC address is detected at 10:00, 10:10 and 17:00, it will be counted only twice because the 10:10 occurrence gets grouped with the first occurrence and it is counted as one. The purpose of the step is to eliminate the smartphone devices that are present throughout longer periods of and therefore do not represent the passers-by.\textsuperscript{13}

- Public MAC addresses were also taken out. In order to estimate the proportion of Apple smartphones (iPhones) in the number of public MAC addresses, August 2015 readings from one of the first installed sensors (in Hammersmith, London) have been taken. Since more rigorous randomisation has been introduced by Apple on September 16, 2015, with the launch of the iOS 9, the proportion of public addresses in the total raw sensor counts saw a sharp increase on the subsequent dates (Figure 4.14). After that, the increase continued at a more steady rate, as more users installed the newest OS version. Presumably, increase of the proportion of the public MAC addresses is primarily attributed to the appearance of iOS 9 and therefore, comparison of the corresponding proportions of public MAC addresses for August 2015 and any dates following its introduction allows for the estimate of the devices that randomise MACs. The proportion of public addresses for the first three weeks of August 2015 for each day of the week separately was computed by using the data from the Hammersmith sensor mentioned above. For any other device at any point in the future, the proportion of public addresses in Hammersmith in 2015 was subtracted from the current proportion of public addresses. That difference constitutes the estimate of smartphones which randomise MAC addresses.

- Finally, some five-minute intervals had missing data, so they had to be imputed. In the first instance, if the gap constitutes only one five-minute interval of missing data, the missing data will be interpolated by using averages of the three five-minute intervals before and after the gap as reference values. If the gap is bigger than that, historical data is copied to fill in the gap, i.e. we take the corresponding count for the corresponding day of the week from the first available week in the past.

Following the imputation of missing values, we obtain the so-called processed sensor count - a result of internal validation. The procedure so far can be summarised by the following equation:

\textsuperscript{13}It has been recognised that some people may walk past the store twice, even within a single five-minute period. They will be counted twice by the field researcher during the external validation, as it is hard to keep a track of unique people passing by. At the same time, data cleaning procedure will remove devices occurring multiple times during the same period. However, this is not too common occurrence and it is not expected to significantly impact the quality of the calibration.
Figure 4.14: Temporal variation in the proportion of the public MAC addresses in the total raw sensor count (Hammersmith, London between 22 July 2015 and 23 January 2017)

\[
\psi = Pub_{y,t_2} \left( \frac{Pub_{y,t_2}}{Raw_{y,t_2}} - \frac{Pub_{h,t_1}}{Raw_{h,t_1}} \right) + \left( Raw_{y,t_2} - Q_{y,t_2} - Rec_{y,t_2} \right) + \epsilon
\]

where:
- \( \psi \) is the processed sensor count;
- \( Pub \) is the number of public addresses;
- \( Raw \) is the raw (overall) sensor count;
- \( Rec \) is the number of recurring smartphones;
- \( Q \) is the number of devices such as printers, tablets, WAPs - which need to be eliminated from the counts;
- \( y \) is the sensor location;
- \( h \) is the location of the sensor in Hammersmith used as a reference for iOS 9 modelling;
- \( t_1 \) is the time reference referring to the August 2015;
- \( t_2 \) is the time reference referring to any period after \( t_1 \);
- \( \epsilon \) is the remaining error that was not accounted for by the calibration procedure. Its sources include passing traffic and undercounting sources of error, as well as any error in preceding estimates.

Patterns displayed for the Bayswater sensor in the previous section seem intuitive and logical. What about the counts? We first compare the raw counts, without any further processing, to the manual counts representing the ground-truth, at all locations. This does not firmly show how accurate the data are, as neither internal nor external validation have been applied yet, but it is considered to be a sensible preliminary quantitative indicator of the degree of error present.
Manual counts used in calculations were taken by the LDC’s field research team between 13 August 2015 and 20 October 2016. The sensors containing no data and sensors for which calibrations were done in five-minute intervals only were initially removed. Filtered data contained 1304 calibrations for 477 sensors. For most sensors, the database contained manual counts for three ten-minute periods in order to avoid taking the noisy sensor counts at finer temporal resolutions\footnote{At five-minute intervals, for example, an exceptionally large or small number of pedestrians passing by may significantly skew the ratio between the actual footfall and the footfall estimated by the sensor, which leads us to misleading conclusions about the error at a given location.}.

Maximum error amongst the overcounting sensors was 9942.86\% (!), whereas maximum error amongst the undercounting sensors was −85.2\%. The median error was 68.70\%, meaning that globally, sensors record 68.70\% more pedestrians than one can detect on site before any processing has been implemented. This distribution skewed towards errors with a positive sign is clearly depicted below (Figure 4.15). In order to make the visualisation more meaningful, errors with the value that is higher than +300\% have been taken out and they make 15.09\% of 477 sensors considered.

If we perform an internal validation on this dataset, the median error becomes negative (−33.61\%), which can be explained by the fact that after the overcounting sources have been accounted for, what remains is a set of factors that cause sensor to miss the passers-by.

### 4.3.3 External Validation

The processed counts acquired through agreed methodology laid out in the previous section are certainly more accurate representations of footfall than raw counts, but in most cases, they are still markedly different from the actual footfall. The next important step includes LDC’s field researchers visiting locations where each sensor was installed and manually counting the passers-by in specified periods, usually three periods of ten minutes\footnote{Note that the manual counts across all sensor locations were already available, i.e. collected by the LDC staff, but additional fieldwork was then conducted as part of this research. A part of this subsection was also published as a conference paper (see Lugomer et al. (2017))}. This makes the labour-intensive part of the external validation, whose overall purpose is to adjust the final processed counts from the internal validation so that they correspond to the ground-truth\footnote{As laid out in Chapter 1, conducting special fieldwork and analysing the variation of measurement error after the internal validation, including any other analysis in the remainder of this chapter and the subsequent chapters are the original contributions of this thesis.}.

This is done by calculating the adjustment factor $\alpha$:

$$\alpha = \frac{M}{\psi}$$

where:

$\alpha$ is the adjustment factor;
Figure 4.15: Frequency of measurement errors (raw sensor counts compared to the manual counts)


$M$ is the number of the passers-by counted manually on the street;

$\psi$ represents the processed sensor counts.

Processed sensor count is then multiplied with the adjustment factor to account for any remaining error.

At this stage of research, there has been a need for learning more about how well processed counts correspond to ground-truth, whether error displays temporal patterns and whether factors such as age structure of the passers-by may be influencing the error. This motivated field visits to certain sensor locations inside and outside of London and taking a larger number of manual counts spread throughout the day or even multiple days.

After a couple of test locations were visited within London, the main part of fieldwork took place at two sensor locations in Sheffield (7 – 9 September 2016) and one other London location (12 December 2016 and 19 January 2017)\(^{17}\).

\(^{17}\) All the fieldwork was undertaken during the workdays.
Fieldwork Sites Description

Sheffield location A (Figure 4.16) is situated on a busy road intersection, just next to the traffic lights. The pavement is relatively wide, and the surrounding retail properties comprise the British Heart Foundation, a bakery and a hair salon. Immediately above lie refurbished office floors which were, at the time of the fieldwork visit, advertised for letting. Together with the actual sensor location (Eurochange), the immediate environment of the sensor does not normally have many people inside the properties that would lead to significant overcounting. This may change to a certain extent, after offices have been let, though, and this is one of the reasons why an effective data cleaning algorithm is necessary to account for any similar urban functional changes.

Sheffield location B (Figure 4.17) is also situated in the centre of Sheffield, however, there is very little passing traffic in front of the store and street is mostly used by the cabs parked on the opposite side of the road. Property in which the sensor is located is Patisserie Valerie, particularly busy during the middays and early afternoons, meaning that there are more people inside, staying for a longer period of time and this may affect the counts. Surrounding properties are used by Barclays and Nottingham banks. The large glass building rising above used to be a nightclub that was permanently closed at the time of the field visit. Potential reopening or land use change may cause changes in observed counts in the long run.

Central London fieldwork location (Figure 4.18) has been on the very busy Tottenham Court Road, near Warren Street underground station. The sensor is situated inside the property shared between the Carphone Warehouse, Curry’s PC World and Google, which
implies that there is a high potential for severe sensor overcounting due to the presence of many additional devices. The pavement is as busy as the road, due to the proximity of the aforementioned underground station and University College London. Next to the actual sensor location, Sainsbury’s supermarket is situated, a significant lunchtime footfall generator.

Measurement Results and Calibration

Manual counts were taken at twenty-minute periods throughout the day for each of the two Sheffield sensors over the course of two and a half days\textsuperscript{18}. Additionally, 14 measurements in fifteen-minute periods over the two days were undertaken at the London location. Unfortunately, the sensor at Sheffield location A did not record any data during that period so that only long-term average sensor counts could have been compared to the manual counts there\textsuperscript{19}.

\textsuperscript{18}The first day comprised four measurements, starting from midday. All together, there were 16 measurements on each of the Sheffield locations: 4 on Wednesday, 6 on Thursday and 6 on Friday.

\textsuperscript{19}This once again raises an issue regarding the consistency of the device data acquisition. It is interesting that the sensor at Sheffield location A actually had a very consistent data feed, with data collected during the 112 out of 119 days in the 4 July – 30 October period, which made anticipation of possible missing of the data quite impossible.
Some Notable Patterns in Sensor Counts

During the data cleaning process, several different categories of MAC addresses that make up the total sensor count, as described in the previous chapter, were counted before being removed. What can be seen by comparing the raw sensor counts, smartphone counts (excluding iOS 9 devices), public counts and distinct smartphone counts is the coincidence of patterns (Figure 4.19). The same pattern occurs at Sheffield sensor location B as well.

An important direct implication of this is that there exists a relatively strong correlation between the processed sensor counts and manual counts, and this is also true for the raw sensor counts and manual counts. For instance, the Pearson correlation coefficient between raw sensor counts and manual counts amounts to 0.85, whereas processed counts amount to 0.76 (Figures 4.20 and 4.21). Both correlations are significant to 0.001 level. Similarly, for the London location, correlations amounted to 0.89 for both relationships (raw counts vs manual counts and processed counts vs manual counts). Similar strength of the relationship was reported by Schauer et al. (2014) who estimated the correlation at 0.75, even though they used a different approach.

Correlation drops down significantly if measurements from different sensors are com-

20If we take the four-month weekday average raw sensor count and assess its relationship with the manual counts, Pearson correlation coefficient will reach impressive value of 0.93. This demonstrates that even though measurement errors exist, can be relatively high in magnitude and temporally variable, pattern captured by the sensor even before additional data cleaning steps have been applied, corresponds very well to the reality.
bined\textsuperscript{21}, meaning that the relationship between sensor counts and the ground-truth is not universal.

**Addressing the Undercounting Sources of Error**

Earlier acknowledged existence of significant linear correlation enables us to apply the adjustment factor to bring the sensor data as close as possible to the manual counts. However, what is the appropriate value of the adjustment factor? Is there a single recommended value or is there a set of values that are a function of a certain factor? The key to this question is understanding of the error that remains after the data cleaning/internal validation process.

In the majority of cases, if the data cleaning algorithm is successful, any error remaining will be negative or at the very least, positive and near zero\textsuperscript{22}. Therefore, what remains in the calibration procedure is estimating the number of pedestrians who were not captured by the sensor. It has been presumed that the proportion of pedestrians who possess a smartphone

\textsuperscript{21}The corresponding Pearson correlation coefficients become 0.55 and 0.67 when all 30 measurements from Sheffield location B and London location are taken into account.

\textsuperscript{22}As previous analysis showed, this has not been the case for all sensors, highlighting the need for a more consistent approach.
and have a Wi-Fi turned off is very difficult if at all possible to directly measure. Instead, two possible partial remedies have been tested. The first one was trying to see whether the age structure of pedestrian flows significantly impacts the measurement error, as it was presumed to be the main demographic driver of smartphone ownership, and the second was to try and find systematic and predictable variations in the proportion of unexplained variation attributable to the undercounting sources of error.

Demographic characteristics of the flow may be estimated through traditional catchment modelling methods, however, even if positive results arose, one would not be able to firmly defend the outcome due to many limitations of such modelling. For example, census data are out of date and validity of the modelling outcome is difficult to prove. Surveying people during the fieldwork would also made very little sense given the fact that only a sample could have been taken and the total number of people surveyed in 15–20 minutes would have been too small. Therefore, the age of passers-by was hypothesised by visual inspection. Errors in estimating someone’s age may exist in some cases, they are not expected to be of significant magnitude.

---

23 A possible indirect solution would be conducting a survey and taking the corresponding proportion of the people who regularly turn their Wi-Fi off as a rule of thumb that can be applied to all the sensor locations.

24 While errors in estimating someone’s age may exist in some cases, they are not expected to be of significant magnitude.
adults. The proportion of elderly in the flow varied throughout the day with peaks around the midday, whereas the proportion of youth did not show a specific temporal pattern. As regards the correlations between the age structure and the measurement error, the conclusions were limited by the fact that sensor A was malfunctioned or turned off during those couple of days, so manual counts could have only been compared to the four-month averages.

Results of the correlation analysis are inconsistent. Sensor A, for example, shows a statistically significant ($p < 0.001$) and very high negative correlation ($-0.97$) between the age structure and the measurement error, meaning that higher proportion of elderly pedestrians in the flow coincides with the more negative measurement error. This is consistent with the hypothesis that elderly people may be responsible for the higher undercounting of the sensors. Sensor B, on the other hand, showed an insignificant correlation ($0.26$, $p \approx 0.32$). If the age structure of the flow is correlated with the calibrated sensor counts for sensor B, rather than the four-month raw averages, the correlation now changes sign and amounts to $-0.26$, however even this value is statistically insignificant ($p \approx 0.31$). The proportion of youth did not correlate well with the measurement error, which could be related to the fact
that their number was rather small anyway\textsuperscript{25}.

Taken together, these findings regarding the age structure indicate that the proportion of older adults comprising the flow can, to a smaller or greater extent contribute to the measurement error. Nevertheless, this impact appears to be spatially inconsistent and practical importance of this finding may be limited due to the fact that some locations have negligible numbers of the elderly and still have a temporally variable error (The sensor location in London is a good example.). Also, the number of elderly pedestrians would need to be collected at all sites at the same time calibration has been done.

Following this finding, a second possible way of modelling the proportion of error caused by the undercounting factors was explored. It had to do with trying to find the patterns in temporal variation of the error. Some motivation behind this is the hypothesis that as smartphone battery becomes drained, probe requests are sent out in the surrounding space less frequently, causing higher undercounting towards the end of the day.

In order to check this, we turn back to Sheffield sensor location B and the London location. Since the removal of overcounting sources of error has taken place prior to the calibration, remaining error is negative and its variation\textsuperscript{26} is shown on the figures 4.22 and 4.23.

As can be seen from the above diagrams, the remainder of the error after the data cleaning is negative. There are a few exceptions that can be explained by either the error in estimates of the raw sensor counts for those five-minute intervals when sensor was powered off; or by the error in estimation of the MAC randomising devices; or the fact that traffic flows have not been accounted for\textsuperscript{27}. The magnitude of error is still relatively high and changes significantly throughout the day. Also, the variation differs between sensor locations. Certain patterns can be discerned, however, they are not completely consistent. At Sheffield location B, the highest sensor accuracy is achieved in the morning. Middays

\textsuperscript{25}In fact, the correlation between the proportion of youth and error in the case of Sheffield A location amounted to 0.47, and in case of Sheffield B location to 0.46. However, these values were not statistically significant to a 0.05 level (p \approx 0.07)

\textsuperscript{26}A problem that arose at the fieldwork location in London was that, unlike the Sheffield sensor B, this sensor was not turned on during the periods when the store was closed. The sensor is usually turned on at various times between 7 am and 8 am. On December 12, 2016, during the second round of measurements, manual counting during the first calibration period started at 7:45 am, however, sensor only started recording the devices from 7:50 am. On such occasions, we do not know when exactly between 7:50 and 7:55 sensor started recording, owing to the temporal resolution of five minutes. Similar problem occurred the same day at 13:00. It also occurred on January 19, 2017 when sensor was powered on after 7:15, however, we do not know when precisely between 7:15 and 7:20 that happened. Missing figures were estimated by assuming that footfall was equal throughout three five-minute intervals. While this assumption to a certain extent adds a new degree of error, it is, at this point, important to have a relatively well covered different periods of a day, so estimation was preferred to omission.

\textsuperscript{27}The experimental evidence shows that devices may be captured by the sensor even when a person is on a bus.
Figure 4.22: Temporal variation in the measurement error - Sheffield location B

Figure 4.23: Temporal variation in the measurement error - London location
and early afternoons, when footfall peaks, have moderate measurement error. The highest errors are recorded during the evening calibration period (19:30 – 19:50), with the exception of Thursday when this was not the case. On the other hand, midday measurements at the London sample location are relatively inaccurate. Correlation between the manual counts and error (Figure 4.24) was also computed in order to inspect whether low counts coincide with higher errors. In the case of London location, there is a statistically significant correlation between the manual counts and error (−0.69, p ≈ 0.006). However, in the case of the Sheffield B location, this relationship becomes statistically insignificant, with p-value amounting to 0.16. This confirms that we cannot establish a relationship between the number of pedestrians on the streets and accuracy of measurements.

Eventually, since no strong correlations between the measurement errors and demographic factors and time could have been established, we decided to minimise the error by taking into account the average adjustment factor for all available fieldwork measurements for each location. The results of employing the adjustment factor are displayed in Table 4.5, whereas all fieldwork measurements along with the associated errors calculated during the internal and external validation are given in Table 9.2 in the Appendices. It can be seen that this kind of validation reduced the average measurement error on both locations substantially, bringing the absolute value of the measurement error from 68.46% to 0.83% in the case of the studied sensor in Sheffield and from 25.03% to 4.59% in the case of the

---

28 This is based on the second measurement day. The first day (12th of December) showed slightly different pattern, but this may have been due to the missing data that had to be, possibly erroneously, estimated.
Table 4.5: Average measurement errors for sensors in Sheffield (7 Sep – 9 Sep 2016) and London (12 Dec 2016 and 19 Jan 2017)

<table>
<thead>
<tr>
<th>Sensor location</th>
<th>Day</th>
<th>Measurement error (%) Without adjustment</th>
<th>Adjustment factor taken into account:</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheffield</td>
<td>Wed</td>
<td>-71.35</td>
<td>-47.67</td>
<td>75.2</td>
<td>-8.41</td>
<td></td>
</tr>
<tr>
<td>Sheffield</td>
<td>Thu</td>
<td>-67.04</td>
<td>-39.79</td>
<td>101.58</td>
<td>5.38</td>
<td></td>
</tr>
<tr>
<td>Sheffield</td>
<td>Fri</td>
<td>-67.75</td>
<td>-41.1</td>
<td>97.19</td>
<td>3.09</td>
<td></td>
</tr>
<tr>
<td>Sheffield</td>
<td>Average</td>
<td><strong>-68.46</strong></td>
<td><strong>-42.38</strong></td>
<td><strong>92.89</strong></td>
<td><strong>-0.83</strong></td>
<td></td>
</tr>
<tr>
<td>London</td>
<td>Mon (12 Dec)</td>
<td>-22.37</td>
<td>-33.12</td>
<td>33.14</td>
<td>8.32</td>
<td></td>
</tr>
<tr>
<td>London</td>
<td>Thu (19 Jan)</td>
<td>-26.85</td>
<td>-36.99</td>
<td>25.45</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td>London</td>
<td>Average</td>
<td><strong>-25.03</strong></td>
<td><strong>-35.43</strong></td>
<td><strong>28.56</strong></td>
<td><strong>4.59</strong></td>
<td></td>
</tr>
</tbody>
</table>

sensor located in London.\(^{29}\)

During the validation process, researchers should, therefore, whenever possible, take multiple counts throughout a day during the peak and off-peak times and then calculate the average value to minimise the measurement error (Lugomer et al., 2017). In the event that this is not possible, as it is a costly operation, several different periods of calibration need to be taken into account and averaged. The latter was the case in the SmartStreetSensor project, and such an approach was applied to all existing sensor locations.

### 4.3.4 The Final Data Preparation Steps

The period chosen for further analysis was since the installation of the first sensor (31 July 2015) until the end of August 2017. The reason is that sensor measurements became increasingly unreliable following September 2017 due to a sharp increase in the number of devices with randomised MAC addresses. This means that methods used for data cleaning and dealing with MAC randomisation need to be revised for future research taking into account post-August 2017 period. Furthermore, this is one of the reasons why the sensor data will be primarily utilised for the analysis of intraday and intraweek footfall patterns, whereas seasonal and long-term trends will not be investigated, but will be highlighted as the potential future extension of this work once the internal validation algorithm gets an

\(^{29}\)Adjustment factor for the London location amounted to 1.334 and it was calculated on the basis of the measurements in both days. Adjustment factor for the Sheffield sensor location B was also taken as an average of two days of measurements (8–9 Sep) and it amounted to 3.068.
For the purposes of this project, which is essentially researching the functional characteristics of retail areas, a period of up to 25 months of footfall data will suffice. Those sensors that were installed towards the end of the chosen time window were not taken into consideration unless they had at least one full week of data. In other words, they had to have at least one valid measurement for each of the 168 hours in a week after interpolation, but not necessarily measured in the same week. It could be argued that one measurement for each hour is insufficient as some of those data may be outliers. However, it was believed that a small amount of data is still better than no data and that the relative merits of keeping those sensors beat the threats of having a couple of unreliable measurements.

For those sensors that were installed near the beginning of the overall project, the entire period from their installation up until August 2017 or their deinstallation was taken into account. It is acknowledged that the effects of seasonality may influence the data for those sensors with a longer overall time taken into account. However, this was not thought to significantly influence the diurnal patterns (discussed in the next chapter) which were presumed to remain unchanged or barely changed over the course of two years. In addition, a decision had to be taken on whether to use the data from the same time period for all the sensors or whether to fetch as much data as possible for every given sensor. The latter option was deemed more appropriate because the former option results in discarding a vast amount of data.

The next important step in preparing the footfall data was detecting anomalies in the dataset. Since outliers could have adversely impacted any interpretation of the later results, we sought to remove them. Outliers were removed using the modified z-scores method that is based on the median and the median absolute deviation or MAD (Iglewicz and Hoaglin, 1993). The more traditional criterion which encourages to label values as outliers if they are at least three standard deviations away from the mean is based on the fact that they constitute only 0.13% of the values based on the properties of the normal distribution (Howell et al., 1998). However, the pitfall of that method is that this indicator itself is influenced by the presence or absence of extreme values and assumes the normal distribution (Leys et al., 2013). While MAD tends to be less efficient than the standard deviation in cases when data are normally distributed, this changes in favour of MAD when distributions bear heavier tails (Howell, 2014). Package ‘stats’ from R programming language, which was used in our case, defines MAD as:

\[
MAD_n = bM_i|x_i - M_j(x_j)|
\]

where the \(M_j(x_j)\) is the median of the original series and \(M_i\) is the median of the deviates

---

30 As will be noted in the Conclusions, a significant progress in that respect has been made in the past years and footfall counts have become reasonably reliable even after August 2017 (Soundararaj et al., 2019). However, this happened too late for this project to benefit from it.
from the median (R Documentation, n/a). The parameter $b$ usually amounts to 1.4826, a constant linked to the assumption of normality of the data, disregarding the abnormality induced by outliers (Rousseeuw and Croux, 1993).

In a given sensor dataset, z-scores were calculated within each location-weekday-hour triplet separately. The distribution of z-scores across the entire dataset is displayed in Figure 4.25.

![Figure 4.25: Distribution of z-scores for all location-weekday-hour triplets](image)

It appears that data are not entirely normally distributed and are skewed to the right (skewness = 0.89) and thus, outliers were defined as values that satisfy the following criterion, as suggested by (Leys et al., 2013):

$$value > |\text{median}(x) \pm 2.5MAD|$$

Using this outlier definition, 7.94% of the records were found to be outliers and they were removed. Where possible, they were replaced by interpolating the values in the same way as described in the section on internal validation.

Following the elimination of outliers, the earlier described time windows for all sensors
were taken, and hourly\textsuperscript{31} median of adjusted footfall counts was computed for each location-weekday-hour triplet separately, meaning for example that all footfall counts at location X, weekday Y and hour Z were considered as one group and aggregated accordingly.

Finally, sensors situated inside nightclubs were removed. This is due to the fact that their measurements were heavily influenced by exceptionally large crowd inflow during the event nights. Since this thesis is focused upon researching what footfall data tell us about activity patterns on the high street, rather than at a single point in front of a nightclub, such locations were usefully discarded. This entire set of data cleaning methods brought the number of sensors to be used for analysis to 605, as laid out earlier in the chapter when sample design was discussed.

Now that we prepared both datasets, we can finally decide which part of the retail unit data we are going to use together with the footfall data. Since sensor data were collected from 2015 and 2017, it makes the most sense to use a retail dataset from 2017. The temporal mismatch between the collection dates of the footfall and retail data is assumed not to have a major impact on the reliability of the results. This is because general compositions of the entire retail areas are unlikely to change dramatically over the two years, and most of the sensors were introduced later, thus being closer to 2017 and covering a smaller period of time with little change. Finally, it is not possible to achieve a perfect temporal match between the two datasets, as sensor data are continuously collected, while the same cannot be said for the retail dataset which represents snapshots in two parts of the year.

4.4 Chapter Summary

This chapter introduced the two main proprietary datasets that will be used in the subsequent chapters: the national retail units database and the SmartStreetSensor footfall database, both provided by the Local Data Company. Both datasets are big and varied: the retail unit database is available for multiple years and contains all retail units in town centres, shopping centres and retail parks in Great Britain. Retail areas vary in size and form a hierarchy, as expected by the rank-size distribution and central place theory. The national structure of retailing appears to change throughout recent years and while the structure of most frequent retail categories remains broadly unchanged, drop in vacancy rates appears to be the biggest element of the changing nature of the British retail geographies. The categorisation of retail units (breakdown on classifications, categories and subcategories) was thought to be useful for accurately depicting the structure of the local retail environments, especially in linkage with footfall data.

Footfall dataset, on the other hand, is even vaster and continuous in time. It was found

\textsuperscript{31}For most of the analytical purposes of this thesis, within-hour oscillations of footfall will be irrelevant and can be considered as noise. Therefore, hourly aggregates were given preference to five-minute counts.
that Wi-Fi sensors can be used to estimate footfall in front of the stores with reasonable accuracy, however, in order to get to that point, a complex set of data cleaning algorithms and ground-truthing steps had to be employed to tackle the sources of measurement error which cause either undercounting or overcounting of pedestrians. Internal validation consisted of removing all devices that were recognised by the Wi-Fi sensors that were not of interest. External validation tackled the issue that not every pedestrian can be captured by the sensor and it involved taking manual counts of footfall at the pre-specified times, comparing it to the processed counts derived by the internal validation and calibrating the final count accordingly.

The bottom line is that, even though the footfall data are very interesting and much can be done with them, a lackadaisical approach devoid of careful and rigorous data validation prior to the analysis would spark many issues with data interpretation and erroneous conclusions later on. To conclude, with Big Data come big opportunities, but also big responsibilities.
Classifications of areas and, less typically, microsite locations, have been a subject of a wide range of research papers in geography. The popularity of clustering can be justified by the fact that humans generally find it useful to simplify their representations of the world using classes. This need for interesting characteristics and patterns extraction has been made possible by the vast amount of readily available spatial data. (Halkidi et al., 2001).

That being said, the aim of geographical classifications is describing the extent to which place A is similar to place B. There is a vast number of possible variables that can be fed into such clustering algorithms in order to characterise space from morphological to functional, and from physical to social and economic. Regardless of the types of variables used, spatial classifications have been extensively covered in the past literature, comprising successful attempts to characterise both rural and urban areas and to target different topics of interest. For example, Lukić (2012) classified rural areas based on a range of demographic and socioeconomic attributes of residential population. Vickers and Rees (2007) and Gale et al. (2016) created geodemographic classifications of both rural and urban areas also based on the characteristics of residents. Alexiou et al. (2016) supplemented such classifications with variables describing built environment.

However, in the majority of geographical classifications, whether they are predominantly morphological or functional, much less attention was given to how attributes vary across time. Even where this has been the case, the typologies were mostly limited only to broad changes, such as dynamics recorded between several points in time (usually two or more years - for example, the change of the population structure between censuses conducted in 2001 and 2011 or land use and land cover changes between 1990, 2000 and 2010, etc.). It
is reasonable to assume, and it has been acknowledged that primary reasons for this were technological limitations, as available technologies were not capable of capturing the data at regular and fine intervals, while keeping the incurred expenses at reasonable levels.

Such limitations, regarded as crucial barriers in the past, have now been widely eliminated, and factors behind this progress have already been covered in the previous chapters. However, some of the contributors particularly relevant for this section include the rapid development and wide-scale adoption of smartphones and Wi-Fi, GPS and Bluetooth technologies. Together, they enabled collection of a high volume of data at fine intervals, while in a spatial sense coming even to the granularity of an individual.

The aim of this chapter is to use the previously cleaned, calibrated and further processed Wi-Fi sensor data to characterise urban microsite locations at which the sensors were installed based on the features of the recorded temporal signatures of footfall. In other words, the idea is to classify microsite locations based on how footfall varies throughout the day and week.

Knowing whether a microsite location attracts most passers-by around midday or during the weekend or weekday evenings or only during the peak commute hours is vital to understanding whether that particular location is suitable for a specific category of retail business. For example, pubs and bar operators will be more interested in the locations where footfall is significant in the evenings. This is contrary to the coffee shop operators such as Starbucks or Costa, which will seek to exploit the large flow of morning commuters and midday lunch and coffee consumers. Another reason why distinguishing between locations of different temporal functional characteristics is important is the opportunity to combine the derived location with the underlying social and economic characteristics of space. For example, individual retailers, retail estate landlords and town planners may be interested in finding out what characterises areas with relatively high footfall around midday, i.e. can we draw relationships between how and when people use space and characteristics of the local population and retail occupancy structure?

This chapter is structured as follows. First, a detailed overview of the time series cluster analysis methodological framework is covered, with special attention given to the myriad of available distance measures and clustering algorithms. This part is concluded by the discussion about which of these are thought to be most suitable for the defined objectives.

The chapter then goes on to describe the cluster centres of the chosen cluster analysis path – how they differ when the number of clusters parameter is changed and how much addition or removal of clusters contributes to the improvement of the interpretation of the underlying functional characteristics that those temporal profiles seek to describe. While a certain solution needs to be chosen in the end, some other solutions stemming from slightly different parametric setups are also briefly described.

The final part of the chapter contains a descriptive statistical analysis of the clustering
results, along with a brief insight into how temporal profiles in a big city such as London differ from the results of the classification conducted on a national scale.

5.1 Time Series Classification: Methodological Overview

Clustering can be defined as: “an unsupervised technique used to group together objects which are close to one another in a multidimensional feature space, usually to uncover some inherent structure which the data possess” (Brock et al., 2011: 2). Clustering of the time series has become a popular research topic in the 1990s (Keogh and Kasetty, 2003) and has gained even more traction in recent years. It has been applied in many fields such as economics, business, demography, climatology, geology, medicine, genetics and others (Bar-Joseph et al., 2002; Maharaj, 2000). Time series classification is distinct from other classification problems because the attributes are ordered (Bagnall et al., 2017). Because of this, a myriad of different distance measures has been proposed in the literature, both traditional ones and more specific ones tailored to the classification needs of the time series of different characteristics. A direct consequence of the existence of so many different approaches is the introduction of a certain degree of subjectivity in clustering, which persists regardless of the chosen methodology. This is a common problem, and it has been pointed out by many authors in both time series clustering domain (Maharaj, 2000) and clustering of other types of datasets (Cape et al., 2000; Lo Siou et al., 2011). While there have been attempts to eliminate some of the subjectivity concerns, the subjective decisions would re-emerge in some form even in the newly proposed solutions. For example, Maharaj (2000) presented a method used to programmatically suggest the number of clusters in the hierarchical clustering; however, one would still need to subjectively decide upon the significance levels. In the case of the majority of clustering approaches, subjectivity arises on multiple levels, and this is also the case in our research problem. The steps where the need for a subjective decision is encountered are the following:

- the choice of representation and data pre-processing approach,
- the choice of a (dis)similarity measure,
- the choice of a clustering algorithm,
- tweaking the clustering input parameters or making the decisions about the number of clusters before and after the initial clustering has been conducted.

Since each of the listed steps may take many different methods or parameters, the result is a very high number of possible combinations of parameters that could, theoretically, be taken. That being said, filtering out the methods that are unsuitable or less suitable to our given problem will be a high priority task which must precede any actual cluster analysis.
We will now give a thorough overview of each of those four sources of subjectivity and present some of the most popular methods used.

5.1.1 Representing Time Series and Data Pre-Processing

The first decision that needs to be made before the cluster analysis is the choice of the representation method. In other words, due to the high dimensionality of the time series data, there is often need to discard the irrelevant parts of the series or otherwise aggregating them to produce smaller datasets. Based on the nature of the data pre-processing, Liao (2005) recognises three main approaches to clustering of the time series data: feature-based, model-based and raw-data-based approaches. The first two groups of approaches aim to reduce the dimensionality of the data by either extracting the relevant features or otherwise transforming the data or by fitting a model and then using model coefficients as parametric inputs to clustering. Traditional methods which can be used in feature extraction are, for example, principal component analysis, sampling (Åström, 1969) and approximation of time series using a series of straight lines. The latter may be achieved by a number of methods, the most notable being identification of the perceptually important points (PIPs) (Chung et al., 2004) which is essentially similar to much older Douglas-Peucker line simplification algorithm (Douglas and Peucker, 1973) widely used in cartography (Longley et al., 2015a). The logic behind PIP algorithm is based on reordering the points that make up the time series by importance. The first two PIPs are the first and last point in the time series, while others are detected by first determining the point with a maximum distance to the first two PIPs and then by determining the point with a maximal vertical distance to the line joining its two adjacent PIPs. The process then continues until all points are reordered according to their importance or the time series has been reduced to the targeted number of dimensions (Figure 5.1) (Fu, 2011). While dimensionality reduction may be beneficial, especially when representing particularly noisy time series, the obvious drawback of feature-based and model-based approaches data transformations is a loss of some potentially valuable degree of information prior to clustering and problems selecting required parameters (Iglesias and Kastner, 2013).

Figure 5.1: Line generalisation methods as a form of data dimensionality

source: Fu (2011)
The raw-data-based approach, on the other hand, takes the raw data directly and feeds it directly into clustering algorithms. Going down this route, however, requires using similarity measures that are specifically adapted to work with time series data, rather than static data (Liao, 2005). Unlike feature-based approaches, raw-data approaches are more susceptible to noise and suffer from high dimensionality (Iglesias and Kastner, 2013).

5.1.2 The Choice of the Distance (Dissimilarity) Measure

The distance measure is the main input parameter taken by the clustering algorithms. It quantifies how dissimilar pairs of observations are - in this case, pairs of time series. Bagnall et al. (2006) posit that choice of distance measure is problem-dependent and they can be classified as (dis)similarities in either: (a) time, (b) shape or (c) change. Similarity in time can be regarded as a special case of similarity in shape so the two go under the common term shape-based methods (Bagnall et al., 2006; Montero and Vilar, 2014). The similarity in change is also called structural similarity, as its goal is to measure similarity in autocorrelation structure (Bagnall et al., 2006). The choice of similarity measure can also be guided by whether the data are discrete or continuous (Liao, 2005). Fu (2011) further divides similarity measures on those that are computed using the whole sequence and subsequence. However, the latter was found to be meaningless, i.e. the output of subsequence time series clustering was independent of the input and in many cases no different than random clusters (Keogh and Lin, 2005). Liao (2005), Montero and Vilar (2014) and Bagnall et al. (2017) provide a comprehensive and detailed treatment of many different similarity measures. Some of the commonly used measures were chosen and further described in the next section. Those measures are Euclidean distance, Dynamic Time Warping (DTW) distances, correlation and autocorrelation based distances, periodogram-based distances and similarity measures based on the discrete wavelet transform (DWT).

Euclidean Distance

A particular case of Minkowski metric, the Euclidean distance is computed as the square root of the sum of squared differences (or in other words, proximities) between two series (Jain et al., 1999). It is one of the most widely adopted distance measures in cluster analysis in general, and at the same time, it is conceptually simple and easy to implement (Fu, 2011), it works quite well when the dataset has compact or isolated clusters (Jain et al., 1999), and in some cases, it can be appropriately applied to time series classification. The latter was acknowledged by (Montero and Vilar, 2014: 2), as they denoted: “Similarities usually considered in conventional clustering could not work adequately with time-dependent data because they ignore the interdependence relationship between values. If, however, the objective is to compare profiles of series, then conventional distances between raw data (Euclidean or Manhattan, among others) evaluating a one-to-one mapping of each pair of
sequences can produce satisfactory results.” Despite its simplicity, caution must be taken when using Euclidean distance as to ensure that there are not any outliers significantly impacting the measures, a problem which is usually resolved by standardising the variables to a common range or variance (Jain et al., 1999). While the Euclidean distance is suitable for clustering time series based on time, it is also known to be sensitive to distortions on time axis, and therefore, alternatives are often sought when shapes of the time series which contain shift in phase are of greater importance (Ratanamahatana and Keogh, 2004).

**DTW Distance**

Non-Euclidean distance methods were developed to mitigate the shortcomings of the Euclidean distance. A very popular one is the Dynamic Time Warping (DTW). It is a shape-based method of time series classification which: “allows two time series that are similar but locally out of phase to align in a nonlinear manner” (Ratanamahatana and Keogh, 2005: 506). “The alignment works out to minimise the difference between the two time series and to this end, a matrix where the element of the matrix contains the distance (commonly Euclidean) between two points of each time series.” (Liao, 2005: 1862). The similarity in time is a special case of similarity in shape. Research has revealed that similarity in shape is superior to metrics based on similarity in time (Aghabozorgi et al. (2015) citing Ratanamahatana and Keogh (2005)).

In most cases, some invariance to mapping will be required by specifying the relatively narrow and different than zero warping window (Ratanamahatana and Keogh, 2005). A good illustration of differences between the Euclidean distance and the DTW is given in Figure 5.2 sourced from Ratanamahatana and Keogh (2004). With the Euclidean distance, change of the magnitude in time series A should happen at the same time interval for it to be recognised as the member of the same cluster as the time series B to which it is compared. DTW, on the other hand, allows specification of the width of the warping window, which, in turn, enables for the mapping the points between two time series that may not be aligned. In other words, if warping window is relatively wide, the two time series will be classified into the same cluster, provided that their shapes match, irrespective of time.

**Correlation and Autocorrelation-based Distances**

Correlation-based distances are based on the computation of the Pearson’s correlation coefficient measuring the correlation between two time series (Montero and Vilar, 2014). Equivalently, the autocorrelation-based distances are based on the comparison of autocorrelation functions of two time series. A weight matrix is commonly incorporated into the calculation, so that either geometric weights decay with autocorrelation lag, or it simply represents identity matrix which results in a distance corresponding to Euclidean distance between autocorrelation functions (Montero and Vilar, 2014).
Periodogram-based Distances

Unlike the previously discussed measures which are derived in the time domain, the foundation of periodogram-based distances is time series analysis in the frequency domain, commonly referred to as Fourier analysis of time series (Bloomfield, 2004) or spectral analysis (Jenkins and Watts, 1968; Priestley, 1981).

The purpose of spectral analysis is to describe the time series as a linear combination of multiple sine functions with different underlying frequency, amplitude and phase parameters. In more technical terms, the goal is to decompose the power of time series into its underlying harmonics, i.e. estimate the power spectrum for the given data (Koopmans, 1995). The power of time series is essentially a sum of powers of various frequency components of the time series, where each frequency component is independent of other frequency components (Koopmans, 1995).

This, in turn, enables identification of frequency parameters of the sine waves that together make up the considered time series. The graphical representation of the relationship between frequency and spectra is called periodogram. Unlike correlation based distances, the purpose of the distance measures from the spectral analysis is to identify the seasonal variations of different lengths, while in the autocorrelation and moving averages-based models such as ARIMA, the seasonal length is already known a priori (Hill et al., 2006). Caiado et al. (2006) derived a measure based on the spectra which can be used to distinguish between the stationary and non-stationary series.
Discrete Wavelet Transform (DWT) Measures

Discrete wavelet transform (DWT) technique works by replacing the original series by their wavelet approximation coefficients in an appropriate scale, and then by measuring the dissimilarity between the wavelet approximations (Montero and Vilar, 2014: 10). The most common case when DWT is used in time series classification is when we wish to choose an appropriate scale, and algorithm by Zhang et al. (2006) does this by making a compromise between two conflicting requirements: an efficient reduction of the dimensionality and preserving as much information from the original data as possible (Montero and Vilar, 2014).

5.1.3 The Choice of Clustering Algorithm

After choosing an appropriate distance measure, a clustering algorithm which will take the distance matrix as an input parameter must be chosen next. With many clustering algorithms to choose from, this is quite a difficult step as well. Parsimonious solutions, i.e. solutions with an optimal trade-off between simplicity and accuracy are preferred and therefore in the experiments with different methods later in this chapter, more common, simpler traditional clustering algorithms were tested with higher priority. According to Jain et al. (1999), clustering algorithms can be divided into four broad categories, based on the method used to define clusters: partitional, hierarchical, density-based and grid-based clustering.

The latter two groups of methods are commonly employed for spatial cluster detection. A popular example of density-based methods is DBSCAN (Ester et al., 1996), while quadrat method falls into the domain of grid-based clustering methods, whose idea is to divide the space into a set number of grid squares. Since the goal of temporal classification conducted in this chapter is to classify footfall distributions at discrete locations, and not to detect spatial clusters of certain phenomena, density-based and grid-based methods will not be investigated in further detail. Furthermore, clustering algorithms can be subdivided into crisp and fuzzy algorithms, where crisp ones are more widely adopted and assign each case to one derived cluster, whereas fuzzy algorithms tend to associated each case to multiple clusters, assigning the probability (Jain et al., 1999) of a case belonging to a particular cluster. A commonly adopted fuzzy clustering algorithm is fuzzy C-means (FCM) (Bezdek et al., 1984). FCM aims to find the most characteristic point in each cluster, label it as the cluster centre and it then derives the grade of membership for each cluster member (Halkidi et al., 2001). Next two subsections provide a brief overview of the nature of the partitional and hierarchical clustering algorithms.
Partitional Clustering Algorithms

Some of the most popular partitional clustering algorithms are k-means and Partitioning Around Medoids (PAM). The k-means is the most widely used clustering algorithm in the literature (Keogh and Lin, 2005). Its aim is “to divide M points in N dimensions into K clusters so that the within-group sum of square is minimised” (Hartigan and Wong, 1979: 100). Its disadvantages, however, are that the number of clusters needs to be prespecified (Bradley and Fayyad, 1998) and that clustering algorithm is guaranteed to converge at local, but not necessarily a global optimum (Keogh and Lin, 2005).

An algorithm closely related to k-means is Partitioning Around Medoids (PAM). It works by selecting the initial most representative objects, termed medoids and labelling all other points as non-selected. Each of the non-selected points is then allocated to the medoid to which it is the closest. A swap between the selected and non-selected object then occurs if such a swap would improve the quality of the clustering process (Halkidi et al., 2001; Mondal and Choudhury, 2013). As is the case with k-means, PAM is an iterative process and works until the optimal solution has been achieved and there is no change of medoid (Mondal and Choudhury, 2013).

A comparative study conducted by Mondal and Choudhury (2013) showed that the performance of these two partitional clustering algorithms heavily depends on the type of distance measure employed. For example, they show that k-means gives the most accurate results with Manhattan distance, while PAM gives the most accurate results with correlation distance used as a similarity measure. Depending on the distance measure, k-means can perform better or worse than PAM and therefore, as authors further emphasise, and what is particularly relevant for our clustering problem, is that searching for ‘the best’ clustering algorithm is fruitless and it beats the purpose of the exploratory nature of clustering. The clustering should thereby revolve around the incorporation of domain knowledge, finding the appropriate measure of similarity and validating clusters (Mondal and Choudhury, 2013).

Finally, partitional algorithms appear to perform better for isotropic clusters, while dense, clearly separated clusters of whatever shape are better detected by clustering algorithms which seek to maximise the minimum distance between the members of two distinct clusters, and thus are better captured by the hierarchical algorithms (Nagy, 1968; Jain et al., 1999).

Hierarchical Clustering Algorithms

The purpose of hierarchical clustering algorithms is the creation of nested grouping of patterns and similarity levels at which groupings change (Jain et al., 1999). Visual representation of such nested groupings is termed dendrogram and it can visually guide the researcher in the interpretation of the similarity relationships between different clusters.
yielded at different hierarchical levels.

The literature usually distinguishes between agglomerative hierarchical algorithms, which produce a sequence of clustering schemes of decreasing number of clusters at each step; and divisive algorithms, where, conversely, the number of clusters increases at each step, as a cluster from the previous step is split into two new clusters. The most popular hierarchical clustering algorithms are complete-link and single-link algorithms, with minimum-variance algorithms such as Ward’s method, also being commonly adopted (Jain et al., 1999).

With the complete-link method, the distance between pairs of clusters is characterised by the maximum of distances between all pairs of patterns drawn from the two clusters, while with the single-link method, a minimum of distances between pairs of patterns is sought (Jain et al., 1999). For most applications in which hierarchical algorithms are employed, the complete-link method gives better results than the single-link method (Jain et al. (1999), citing Jain and Dubes (1988)). It also tends to produce more compact clusters than the single-link clustering. On the other hand, Ward’s method (Ward and Joe, 1963) minimizes the total within-cluster variance so that at each step, the pair of clusters with minimum cluster distance is merged. This pair of clusters leads to a minimum increase in total within-cluster variance after merging (Charrad et al., 2014).

5.1.4 The Choice of Clustering Parameters and Cluster Validity

Even when the same clustering algorithm is adopted for the same data, a different set of clustering solutions can emerge. Hence, special attention needs to be given to the effective evaluation standards and criteria for deciding upon the clustering input parameters (Charrad et al., 2014). That being said, even the simplest clustering algorithms require pre-defining at least one or two parameters. For example, in the case of k-means, one would have to decide upon the number of clusters, as well as the initial seeds. While hierarchical clustering algorithms do provide dendrograms which can be used to identify a number of clusters, even there the final decision on which height to cut the tree at is due to a researcher.

Even though visualisation of the dataset usually aids in the validation of the clustering results, problems occur when attempting to visualise large multidimensional datasets (with more than three dimensions), as humans have trouble visually inspecting clusters defined by many variables (Halkidi et al., 2001). In order to get around those problems, recent literature came up with the quantitative measures which can be used to find the optimal number of clusters. Charrad et al. (2014) listed thirty clustering validity indices which can be used to assess the ideal clustering solution for any of the popular clustering algorithms and distance matrices, and made them available for practical experimentation in NbClust R package. The idea is to run the chosen clustering algorithm of interest and inspect how many of the clustering validity indices suggest a particular number of cluster solutions.
According to the majority rule, the more validity indices suggesting the particular number of clusters, the better the solution (Charrad et al., 2014). However, in some cases, the majority rule might not be as convincing, so in that case, five top performers in the study conducted by Milligan and Cooper (1985) (CH index, Duda index, Cindex, Gamma and Beale) are given stronger weight in decision-making.

### 5.2 Classifying Microsite Locations

In this section, we discuss the previously introduced time series classification methods and apply them to derive the temporal footfall classification of the sample of microsite locations in the British cities.

#### 5.2.1 Practical Implementation of the Distance Measures and Clustering Algorithms

With many methods of time series classification available and employed in different fields, Montero and Vilar (2014) compiled an R package TSclust comprising a series of methods used to compute a pairwise dissimilarity matrix between different time series. After that, they suggest using the computed dissimilarities as an input for the traditional clustering methods such as k-means, partitioning around medoids (PAM) or Ward hierarchical clustering. The aim is to use as simplest methods as possible while deriving a plausible classification. Out of all the previously introduced distance measures, not all are suitable for our dataset. The correlation-based similarity measures, for example, fail to capture relevant features of the Wi-Fi sensor data because they are concerned with the existence and strength of autocorrelation. Two time series can both have similar temporal model specifications, but be temporally misaligned, as Zhang et al. (2011) illustrated (Figure 5.3). This is not satisfactory in our case, where similarities in time are crucial. Moreover, the model-based methods were found to be more suitable for the long datasets of unequal lengths (Bagnall and Janacek, 2014; Bagnall et al., 2017), whereas in our case time series were pre-processed and converted into the series of equal length prior to clustering.

![Figure 5.3: Example of two time series clusters produced by AR(2) and ARMA(2,2) models, respectively](source: Zhang et al. (2011))
The literature suggested that shape-based methods such as Dynamic Time Warping distance (DTW) are potentially suitable distance measure candidates. The DTW, in particular, has become one of the most popular shape-based techniques and also one of the most accurate ones out there. Despite the recent advances in time series classification methods, DTW algorithm is still hard to beat, and while algorithms such as COTE (Collection of Transformation Ensembles) achieve higher accuracy, they are also computationally more intensive (Bagnall et al., 2017). In our case, clustering is based not only on shape but also time, as we are interested at what time a certain pattern is recorded (such as peak or trough during the day). This can be addressed by adjusting the value of warping window, a step which proved to be important in achieving the higher accuracy of clustering (Ratanamahatana and Keogh, 2005). The cited paper, therefore, confirmed the accuracy of clustering peaks when small size of a warping window (up to 10%, and 4% on average) is chosen. Contrary to that, setting no warping window or setting too wide warping window results in distortions. If warping constraint is set to 0, the case becomes identical to using Euclidean distance metric.

Shieh and Keogh (2008) further demonstrated that when dataset becomes even moderately large, there is very little difference in classification error rate between incorporation of the DTW and simple Euclidean distance. Therefore, if the dataset is sufficiently large (measuring thousands of cases or, sometimes, even just hundreds), suitable cluster cases can be found even without warping, which becomes necessary when datasets are very small (Shieh and Keogh, 2008). Our dataset (just over 600 time series cases after all the filtering steps) is somewhere on the borderline between the case in which Euclidean distance may be used and the case in which it may not produce satisfactory results. Experimentation will be necessary, and both methods will be tested to see which one yields better results.

As noted earlier, time series objects could be represented as feature vectors by extracting features from the time series, for example, by using principal component analysis. However, the experiments proved that using DTW on raw data presents a superior approach (Mueen and Keogh, 2016). It is because of this that dimensionality reduction was not conducted on our data. A further reason that supports this decision is the fact that the number of data points in each time series and each day of the week is not exactly that large, measuring 168 periods per week and only 24 per day.

Prior to computation of distance measures, data were range-normalised. Normalisation was conducted for each subsequence of the time series separately as suggested by Mueen and Keogh (2016), or in other words, for each location separately so that the maximal value of footfall for each day at each location amounted to 1. The literature demonstrated that normalisation presents almost a necessary step which significantly influences the accuracy of the subsequent clustering results and it needs to be done on each subsequence rather than the entire dataset as it is commonly misunderstood (Rakthanmanon et al., 2012). The
range standardisation (see equation below) works to convert the values of a variable to the values which fall within the [0, 1] interval (Vickers and Rees, 2007).

\[ R_i = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]

where:

- \( R_i \) is the range-standardised value;
- \( x_{\text{max}} \) is the maximum value taken by \( x \) in the data and
- \( x_{\text{min}} \) is the minimum value.

The methodology of microsite location temporal profiling (Figure ??) can, therefore, be summarised by the following steps, all of which were applied to the data cleaned through the procedures described in the previous chapter:

1. Range-normalise the data for each time series subsequence;
2. Compute DTW distances by varying the size of the warping window. Values of 0 (corresponding to the Euclidean distance) and 1 (hour) are to be tested. In a practical sense, dtw.basic function from the dtwclust R package will be used (Sarda-Espinosa, 2018);
3. Inspect which type of hierarchical clustering yields best solutions. A couple of common types are to be tested: average, single, complete and Ward’s. When a partitional algorithm such as k-means or PAM is tested, this step is skipped;
4. Input the computed DTW and/or Euclidean distances into the chosen clustering algorithm and decide upon the initial options for the number of clusters by running a set of thirty validation algorithms (Charrad et al., 2014);
5. Inspect the solutions that emerge from the different number of clusters and conclude at what number of clusters the clustering solution ceases being meaningful and useful;
6. Choose the optimal solution and proceed with the description of clusters.
During the initial experimentation and visual inspection of different outcomes of clustering, it was found that, regardless of the method used, weekend profiles tend to be different from the workday profiles. Weekend profiles normally tend to be one-peaked regardless of the profile of their workday counterparts. Also, Fridays are sometimes different than Mondays through Thursdays, and in some cases, minor differences may exist even among other workdays. This leads us to say that one cannot postulate that functional characteristics of a microsite location are static, as they appear to change depending on the day of the week. The goal of this classification should not be to characterise each location by a single representative profile, which would be an average of the entire week. While this is possible, a better approach would be to acknowledge the most relevant differences that exist throughout the week and create classifications for the characteristic groups of weekdays. For example, questions that are interesting in describing local functional characteristics, therefore, are:

1. What does the diurnal shape of the typical workday for a particular location look like?

2. What does the diurnal shape of the typical weekend for a particular location look
A further useful distinction could be made between Mondays through Thursdays on one side and Fridays on the other side, as Fridays may have pronounced afternoon and evening activity whereas their other workday counterparts may not have them. Saturdays and Sundays may also differ because nightlife activity (if it exists at a given location) becomes captured both in the early mornings and late evenings of Saturdays, but only in the early mornings on Sundays.

Having said that, the route that was taken included producing the average profiles from Monday to Thursday and then performing the cluster analysis. Separate clustering would be done for Fridays and separate for Saturdays and Sundays after which further conclusions could be drawn on whether Fridays are significantly different from the rest of the workdays at the majority of locations.

In the next section, we conduct the cluster analysis thorough experimentation with selected distance measures and clustering algorithms. In the first instance, clustering was conducted for the Monday-Thursday profiles\(^1\), while Fridays and weekend clusters will be derived separately afterwards.

### 5.2.2 Exploratory Cluster Analysis

Four agglomerative hierarchical algorithms were tested with the Euclidean distance and DTW taken as dissimilarity measures: average, single, complete and Ward’s method. The motivation behind this step was to reduce the number of clustering algorithms for the experimentation phase by finding the most appropriate algorithm from the family of the hierarchical algorithms. Ward’s method yielded the best results, as agglomerative coefficient\(^2\) amounted to 0.98. When DTW with the width of the warping window set to one hour to allow for a bit of invariance was tested, the agglomerative coefficients increased in each case. Ward’s method still came out as the most suitable candidate with 99% of the clustering structure detected (Table 5.1).

The clustering was initially conducted by using both Euclidean distance and DTW interchangeably and by using Ward’s algorithm, k-means and PAM interchangeably. In each case, all possible cluster solutions, taking the number of clusters between 2 and 12 were computed, and cluster centres were plotted.

A common denominator for all cases is that when clustering is conducted using two clusters selected as a parameter, one-peaked and three-peaked profiles emerge in all cases, without exception. When the number of clusters increases, cluster solutions start to differ. At some point, increasing the number of clusters results in at least one pair of profiles that

---

\(^1\)The summary of the methodology and results for the Monday-Thursday profiles were also published in the conference paper (see Lugomer and Longley (2018))

\(^2\)Values closer to 1 represent a larger amount of clustering structure found.
Table 5.1: The amount of clustering structure found in different hierarchical clustering algorithms with: (a) Euclidean distance, (b) DTW (warping set to one hour) used as distance measures

<table>
<thead>
<tr>
<th>Clustering algorithm</th>
<th>Agglomerative coefficient</th>
<th>Euclidean distance</th>
<th>DTW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.87</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>0.81</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>Complete</td>
<td>0.90</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Ward</td>
<td>0.98</td>
<td>0.99</td>
<td></td>
</tr>
</tbody>
</table>

visually look too similar. Table 5.2 shows at which point this happens in each clustering case. This is thought to serve as a useful guide in assessing the pragmatic importance of the cluster analysis results.

Table 5.2: Number of clusters at which too similar pairs of profiles start to appear in the solution

<table>
<thead>
<tr>
<th>Clustering algorithm</th>
<th>Euclidean distance</th>
<th>DTW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ward</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>K-means</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>PAM</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

With DTW, both partitional algorithms spawn two very similar profiles (three-peaked profile with increasing peaks) even when the number of clusters is low. When Ward’s algorithm is employed with DTW taken as a distance measure and dendrogram is cut yielding five clusters, two of those are single-peaked and too similar. It is important to note that increasing the number of clusters from there does not improve the solution, i.e. previously emerged pairs of too similar profiles remain similar. Euclidean distance gave generally more satisfying results, with Ward’s method and PAM giving most distinctive profiles. The ability of the Euclidean distance measure to pick up dissimilarity in time is thought to be important in discerning the functional differences between locations, as the warping conducted by the otherwise popular DTW method would ignore this time shift of the peak.

It is also important to note that even though clustering with Euclidean distance and
Ward’s method yields a pair of similar clusters when the number of clusters is set to five, increasing the number of clusters a bit further results in the creation of some new relevant profiles. This essentially means that information presented in Table 5.2 only serves as a suggestion and not as definite cut-off values for the maximum number of clusters, as further increase of the number of clusters may in some cases provide better results.

Every solution has certain advantages and downsides, so a compromise had to be made. Ward’s method was thought to be more consistent because, with each increase of the number of clusters, the newly produced cluster comes from one cluster which is its superordinate in the hierarchical tree. On the other hand, with partitional algorithms, newly created clusters result from taking some degree of information from every other cluster, rather than only one as is the case in hierarchical clustering. This is not thought to be ideal, as there will be many cases when profiles in the same cluster look too dissimilar. Having said that, since hierarchical algorithms provide a better control of the hierarchical relationships between clusters, clustering with Euclidean distance and Ward’s algorithm was chosen and the corresponding results are presented. The obtained cluster centres appear to be satisfactory as they are sufficiently distinctive, indicating different functional characteristics of different locations.

The next important step was to decide upon the optimal number of clusters, for which purpose, the previously reviewed validity indices were computed. While they did provide the initial indications of the quality of clustering solutions with each potential number of clusters, as will be demonstrated, such automated recommendations were not found to be satisfactory, and final decision was made by accommodating the potential of the clustering solution to effectively describe the functional characteristics of a given location. Among all tested indices from the NbClust R statistical package (Charrad et al., 2014), 10 (of 30) proposed 2 as the best number of cluster, five proposed 4 as the best number of cluster, while three of them proposed either 3 or 14. Unfortunately, statistically proposed solutions appear to be either overly generalised (2, 3 or 4 clusters), or too detailed (14 clusters). Therefore, the centres of each cluster solution were visualised to observe which set of clusters yields most interpretable results. This implies that we expect cluster solutions to be easily associated with some common aspects of the functioning of urban areas, i.e. they are expected to revolve around journey to and from work, lunchtime, start and end of leisure time, and perhaps aspects of the nighttime economy.

When the hierarchical tree is cut at the height which creates only two clusters, the corresponding clusters which emerge are one-peaked (although, with a slight lump which is close to being a local extremum) and three-peak profiles (Figure 5.5). However, even though cluster validity algorithms suggested two clusters as an optimal solution, interpretation-wise, this is not ideal and many distinctive functional characteristics become lost this way.

The three-cluster solution, for example, already subdivides the cluster 2 from the previ-
ous solution into two distinct cases: the ordinary single-peaked profile and the single-peaked profile with a secondary late afternoon minor peak (Figure 5.6).

Figure 5.5: The two-cluster solution obtained through Ward’s method
In the four-cluster solution (Figure 5.7), the two-peak cluster with much more pronounced afternoon peak appears. The difference between the clusters 2 and 4 in the four-cluster solution is also in the time the maximum is reached: in cluster 2 the maximum is reached at 5 pm, while in cluster 4 the same happens an hour later.

In the five-cluster solution (Figure 5.8), the existing differences between the three-peaked profiles are revealed: the cluster 2 displays gradual growth of footfall over the day, with each of the three peaks making a higher high (cluster 2). This is different from cluster 1 where mornings and middays are similar in importance, while afternoon peak slightly prevails.
Figure 5.7: The four-cluster solution obtained through Ward’s method
The six-cluster solution reveals functional differences between the locations with simple one-peaked profile (cluster no. 4 in Figure 5.9 with very short duration of the maximal footfall, and the locations which also peak around midday, but have an afternoon inflow of passers-by that produces something resembling secondary peak. However, in mathematical terms, this afternoon activity cannot qualify as a peak, but it is just a deceleration in the decline of afternoon footfall.

The seven-cluster solution (Figure 5.10) adds in the three-peaked cluster where midday peak is dominant and the other two peaks are far less significant.
Figure 5.9: The six-cluster solution obtained through Ward’s method
The eight-cluster solution (Figure 5.11) further establishes a cluster similar to the cluster 2 shape-wise, however, the difference in relative magnitude between the peaks is much larger, with morning peak constituting a smaller proportion of the daily footfall. It is also the case that the daily maximum is reached later (at 7 pm, as opposed to 5 pm – 6 pm in cluster 2).

And finally, for the sake of comparison, the nine-cluster solution is also displayed here (Figure 5.12). While clusters 1 and 2 share general shape characteristics, some minor differences can be identified. The difference between the peaks in cluster 1 is generally smaller, even negligible, while the afternoon peak in cluster 2 prevails. Moreover, the afternoon peak in cluster 1 is reached at 5 pm, while the same happens in cluster 2 locations at 6 pm. It is thought that the added value of the nine-cluster solution is negligible, as all the relevant and distinctive shapes have already been picked up. Therefore, the eight-cluster solution was eventually chosen.
Figure 5.11: The eight-cluster solution obtained through Ward’s method
5.2.3 Description of Clustering Results

If the eight-cluster solution is accepted for the National Footfall Temporal Classification, the results that thereby emerge are illustrated by the cluster dendrogram (Figure 5.13) and Table 5.3 which displays the exact number and percentage of cases in each cluster. Figure 5.14 is the copy of Figure 5.11 for an easier reference to the table.

The number of cases is unevenly distributed among eight clusters with clusters 5, 7 and 8 each capturing less than 5% of the total cases, whereas cluster 3 captures 27.9% of the cases. This is not necessarily a negative thing, as cluster 7, for example, presents a very specific profile which is markedly different than the rest of the cluster centres both in terms of shape and the time the daily maximum is reached (at 7 pm - later than the usual 5pm or less typical 6 pm). This said, in order to get a more even distribution of cases, the number of clusters would need to be reduced to six. This happens at the loss of some distinctive profiles which could be interesting in explaining the functional characteristics of certain places.

It is worth pointing out that one of the aims of this cluster analysis was validation of patterns that are known to exist because of the ways in which days are structured into
human activities. However, the ways in which these activities map onto retail areas had to be further examined in order to confirm those anticipated patterns, potentially discover new ones, find out how frequent each of them is and how they relate to other spatial attributes. It is already very interesting and to a certain extent unexpected that there are as many as eight possible ways of thinking about the footfall patterns in retail areas. And this is not even a complete picture, given the fact that Fridays are slightly different and weekends are substantially different, as will be demonstrated later on.

According to Table 5.3, the most common Monday-Thursday temporal profile in the retail areas of Great Britain (27.93% of the sampled microsite locations) is a two-peaked profile with a maximum around midday and late afternoon - labelled as Consistent Afternoons (profile 3 in Figure 5.14). However, unlike similar profiles, such as Lunchtime and Afternoon Rush Hour Spike, the drop of footfall during the early afternoon, i.e. between 2 pm and 5 pm is less significant, which means that such locations benefit from consistently high footfall throughout most of the afternoons. The mornings at such locations are not normally busy.

The second most common temporal profile (cluster no. 4, Midday Top, comprising 19.67% locations) is a simple one-peaked profile with maximum activity recorded around midday. Such locations likely attract lunch goers.
Table 5.3: The breakdown of cluster cases (eight-cluster solution)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Proposed name</th>
<th>Cases</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Commute and Lunch</td>
<td>84</td>
<td>13.88</td>
</tr>
<tr>
<td>2</td>
<td>Gradual Rise</td>
<td>80</td>
<td>13.22</td>
</tr>
<tr>
<td>3</td>
<td>Consistent Afternoons</td>
<td>169</td>
<td>27.93</td>
</tr>
<tr>
<td>4</td>
<td>Midday Top</td>
<td>119</td>
<td>19.67</td>
</tr>
<tr>
<td>5</td>
<td>Lunchtime and Afternoon Rush Hour Spike</td>
<td>29</td>
<td>4.79</td>
</tr>
<tr>
<td>6</td>
<td>Lunchtime with Secondary Afternoon Commute</td>
<td>90</td>
<td>14.88</td>
</tr>
<tr>
<td>7</td>
<td>Quiet Mornings, Busy Evenings</td>
<td>19</td>
<td>3.14</td>
</tr>
<tr>
<td>8</td>
<td>Busy Lunchtime with Secondary Commuting Peaks</td>
<td>15</td>
<td>2.48</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>605</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

The next cluster is *Lunchtime with Secondary Afternoon Commute* (profile no. 6) comprising 14.88% of the locations. It is a one-peaked profile with an increased activity in the late afternoon, during the afternoon rush hour. In these locations, however, commuters are not as numerous as in the case in some other locations, so pronounced secondary peak is not formed.

Similarly numerous, clusters *Commute and Lunch* (profile 1) and *Gradual Rise* (profile 2) account for 13.88% and 13.22% of the locations, respectively. Both are three-peaked profiles and are characterised by busier customer traffic at all three characteristic periods during the day - morning rush hour, lunchtime and afternoon rush hour. The difference is that Gradual Rise locations expect more customers towards the end of the day and intraday differences of footfall volume are not as pronounced. Commute and Lunch locations, on the other hand, have more pronounced peaks, and corresponding locations may expect similar footfall volume during all three periods. The peak in the late afternoon, though, seems to be slightly higher than other two peaks.

The profiles captured by the remaining three minor clusters are not as commonly encountered across the British retail space, however, since they are functionally specific, it is worth further investigating where and why they occur. As was already mentioned, cluster
Lunchtime and Afternoon Rush Hour Spike (profile 5) contains two-peaked profile microsite locations (4.79%) with a more significant drop in customer traffic after the lunchtime, as compared to the similar cluster 3 (Consistent Afternoons). Interestingly, these locations either do not record any peak during the morning rush hour or record a minor morning peak. At the same time, they do record a significant afternoon peak. At the same time, they do record a significant afternoon peak.

Cluster no. 7, Quiet Mornings, Busy Evenings (3.14%), is to a certain extent similar to the cluster Gradual Rise, but morning footfall is much smaller, and differences between the peaks are much more pronounced. Moreover, the maximum footfall is, on average, reached between 7 pm and 8 pm, which seemingly makes these locations more attractive for the dinner goers and pub goers.

Finally, occurring at only 15 locations (2.48%), cluster no. 8, Busy Lunchtime with Secondary Commuting Peaks, is characterised by its distinctive dominant lunchtime peak and two smaller peaks during the rush hours.
5.3 Cluster Stability over Fridays and Weekends

Even the initial visual inspection of the cluster profiles across the days of the week tends to suggest that significant differences exist between weekends and workdays. The differences are also expected between Fridays and the rest of the workdays. However, since they are not as obvious visually, a separate cluster analysis was conducted for each characteristic part of the week. Clustering results for Mondays through Thursdays have already been presented in the previous section, and we now proceed with describing how other three days of the week differ with respect to the diurnal footfall patterns. Cluster analysis was conducted using the same methodology as before, i.e. Euclidean distances were computed between each time series and subsequently fed into the Ward’s clustering algorithm.

The results suggest that Fridays are generally very similar to other days of the week (Figure 5.15). Similarly shaped profiles emerged when identical (n=8) or a similar number of clusters were chosen as in the case of Monday to Thursday profiles. A key difference is that in two clusters evenings appear to be on average significantly busier than in the rest of the clusters. Apart from that, cluster previously labelled as Busy lunch with both commuting peaks now emerges without the afternoon peak. Also, Monday to Thursday clusters Midday Top and Lunchtime with secondary commute now appear to be much more similar. However, the motivation behind keeping the two clusters as distinct is the fact that Friday cluster 7(6) shows higher footfall during the evening and drops slower in the afternoon when compared to cluster 4. Since most Mon-Thu and Fri clusters are comparable, the names assigned to the Friday profiles were generally very similar and, in some cases, identical.
Cluster membership for 408 of 605 (67.4%) locations remains unchanged between Monday-Thursday and Friday, i.e. when comparable clusters from both Monday-Thursday period and Friday are compared. It should be noted that this is an approximation because in both cases clusters were defined slightly differently due to the fact that footfall profiles for Mondays-Thursdays and Fridays differ at certain locations. However, besides the highlighted differences that emerge during the Fridays, the distinctive profile characteristics of each of the eight clusters have been preserved very well, so clusters can be deemed comparable for the most pragmatic purposes.

Table 5.4 displays a more detailed breakdown of how cluster membership changes between Monday to Thursday on one side and Friday on the other side. The matrix diagonal shows the number of locations which preserved their cluster membership and it is expectedly, the most dominant category in each cluster. While a change in cluster membership is uncommon for the locations that belong to clusters 5 to 8 on Mondays-Thursdays, clusters 3 and 4 do display more change and they are especially prone to swap with the cluster 6. This is somewhat expected to the imperfect nature of clustering, as some cases are inherently situated on the borderline between clusters and these clusters are to a large extent similar. They all have a prominent midday peak, with the main difference being in the duration of the peak or the existence of the secondary afternoon peak. The locations
Table 5.4: The change of cluster membership between Monday-Thursday and Friday

<table>
<thead>
<tr>
<th>Mon-Thu cluster</th>
<th>Friday cluster</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>52</td>
<td>3</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>23</td>
<td>87</td>
<td>2</td>
<td>16</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>169</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>89</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>6</td>
<td>119</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>21</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>9</td>
<td>1</td>
<td>74</td>
<td>1</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>59</strong></td>
<td><strong>85</strong></td>
<td><strong>99</strong></td>
<td><strong>100</strong></td>
<td><strong>69</strong></td>
<td><strong>140</strong></td>
<td><strong>36</strong></td>
<td><strong>17</strong></td>
<td><strong>605</strong></td>
</tr>
</tbody>
</table>

belonging to clusters 1 and 2 on Mondays-Thursdays in some cases tend to enter the cluster 5 membership because they all have high afternoon footfall and both clusters 1 and 5 have a marked decline between the peaks. Overall, the above table tends to demonstrate that the majority of cluster preserve their cluster membership between Fridays and the remaining workdays, but in cases where the exchange of cluster membership happens, it is mostly between the functionally similar clusters.

Unlike workdays, profiles for weekends are much more straightforward, owing to the fact that the majority of workday commuters either stay at home or travel to the town centres and shopping centres at different times of the day. For this reason, profiles on Saturdays and Sundays are mostly one-peaked with maximal activity around the midday, the three-peaked profiles generally do not exist and two-peaked profiles are different in that primary peak usually occurs in the evenings, while night also records a relatively large volume of footfall when compared to workday clusters. However, a distinction still needs to be made between Saturdays and Sundays. As has already been said, the Saturday evenings will be busy at some locations, while the same may not be true during Sundays.

Saturdays can be described by only three distinct profiles (Figure 5.16 and Table 5.5). Clusters *Consistent Saturday Afternoons* (cluster 1) and *Saturday Midday Top* (cluster 2) are mutually similar, as both are one-peaked profiles. However, some important differences can be discerned from the two plots. First, peak footfall at locations belonging to Consistent Saturday Afternoons lasts during the afternoon, while cluster Saturday Midday Top records a shorter peak restricted to hours around the noon, after which sharper drop occurs. Second, night/early morning hours in Saturday Consistent Afternoons have slightly higher footfall than Saturday Midday Top, signifying some impact of the Friday night out-goers. Finally, evenings are much busier at Consistent Saturday Afternoons locations, whereas Saturday Midday Top locations tend to be very quiet and see negligible volumes of footfall. Overall,
Table 5.5: The breakdown of cluster cases - Saturday

<table>
<thead>
<tr>
<th>Cluster no.</th>
<th>Cluster name</th>
<th>Number of locations</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Consistent Saturday Afternoons</td>
<td>213</td>
<td>35.2</td>
</tr>
<tr>
<td>2</td>
<td>Saturday Midday Top</td>
<td>356</td>
<td>58.8</td>
</tr>
<tr>
<td>3</td>
<td>Saturday Nightlife</td>
<td>36</td>
<td>6.0</td>
</tr>
</tbody>
</table>

Consistent Saturday Afternoons tend to be associated with more vibrant locations, both in terms of daytime visitors and nightlife. *Saturday Nightlife locations* (cluster 3), on the other hand, peak in the evening and display quite a large share of footfall around midnight. They also have a secondary peak around midday, meaning that they are relatively busy throughout the whole day and are presumed to be multifunctional.

Sundays can also be described by only three profiles, two of which are one-peaked and the third one exhibits high footfall during the early hours in the night, thus resembling the footfall distributions observed on Saturdays (Figure 5.17 and Table 5.6).

![Figure 5.16: National Footfall Temporal Classification – Saturday footfall profiles](image-url)
Table 5.6: The breakdown of cluster cases - Sunday

<table>
<thead>
<tr>
<th>Cluster no.</th>
<th>Cluster name</th>
<th>Number of locations</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Consistent Sunday Afternoons</td>
<td>249</td>
<td>41.1</td>
</tr>
<tr>
<td>2</td>
<td>Sunday Midday Top</td>
<td>315</td>
<td>52.1</td>
</tr>
<tr>
<td>3</td>
<td>Early Sunday Nightlife</td>
<td>41</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Cluster membership for 461 of 605 (76.2%) locations remains unchanged throughout the weekend, i.e. when comparable clusters from both Saturday and Sunday cluster-solution are examined. The most common change (about 56% of the weekend cluster changes both ways) was a swap between two one-peaked profiles. This difference is rather subtle, and it cannot be deemed significant functional change, because in both cases the footfall peaks around midday with the only difference being the more gradual fall of afternoon footfall in one case, as compared to the other. More interesting is the fact that some locations appear to belong to the cluster 3 on Saturdays, but then to one of the other two clusters with quieter nights on Sunday, and vice-versa. This is interesting because it tends to suggest
Table 5.7: The change in the comparable cluster membership from Saturday to Sunday

<table>
<thead>
<tr>
<th>Cluster membership change (Sat → Sun)</th>
<th>Number of cases</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 → 2</td>
<td>23</td>
<td>15.97</td>
</tr>
<tr>
<td>1 → 3</td>
<td>23</td>
<td>15.97</td>
</tr>
<tr>
<td>2 → 1</td>
<td>58</td>
<td>40.28</td>
</tr>
<tr>
<td>2 → 3</td>
<td>11</td>
<td>7.64</td>
</tr>
<tr>
<td>3 → 1</td>
<td>24</td>
<td>16.67</td>
</tr>
<tr>
<td>3 → 2</td>
<td>5</td>
<td>3.47</td>
</tr>
<tr>
<td>All changes</td>
<td>144</td>
<td>100</td>
</tr>
</tbody>
</table>

that if a particular microsite location appears to be busy during the early hours of Saturday, there is no guarantee that it will be as busy during the early hours of Sunday as well. It is, however, understood, that this is not a common occurrence, as it was detected at only 10.4% of the locations.

5.4 The Special Case of London

Do major, vibrant and cosmopolitan cities display different temporal patterns of urban activity on high streets and shopping centres? London is a dynamic cosmopolitan and business centre with much more complex demographics and population mobility. Therefore, there is a conjecture that not all of the eight average footfall profiles that emerged on a national level are relevant for Greater London and this assumption was inspired by the traditional special treatment of London in geodemographic classifications presented in the literature review. The assumption can easily be confirmed by inspecting the distribution of 226 sensor microsite locations within London (Table 5.8). The location quotients were computed by dividing the percentage of a particular cluster in London with the percentage of the corresponding cluster on the national level:

\[ Q_c = \frac{L_c}{N_c} \]

where:

- \( L_c \) is the percentage of sensor microsite locations belonging to the cluster \( c \) on the local level (Greater London);
- \( N_c \) is the percentage of sensor microsite locations belonging to the cluster \( c \) on the national level;
- \( Q_c \) is the location quotient. Values larger than 1 indicate that footfall pattern \( c \) is more...
Table 5.8: Comparison of temporal clusters occurrence on (a) London level and (b) national level

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Count</th>
<th>%</th>
<th>Count</th>
<th>%</th>
<th>Location quotient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commute and Lunch</td>
<td>68</td>
<td>30.09</td>
<td>84</td>
<td>13.88</td>
<td>2.17</td>
</tr>
<tr>
<td>Gradual Rise</td>
<td>55</td>
<td>24.34</td>
<td>80</td>
<td>13.22</td>
<td>1.84</td>
</tr>
<tr>
<td>Consistent Afternoons</td>
<td>45</td>
<td>19.91</td>
<td>169</td>
<td>27.93</td>
<td>0.71</td>
</tr>
<tr>
<td>Midday Top</td>
<td>4</td>
<td>1.77</td>
<td>119</td>
<td>19.67</td>
<td>0.09</td>
</tr>
<tr>
<td>Lunchtime and Afternoon Rush Hour Spike</td>
<td>15</td>
<td>6.64</td>
<td>29</td>
<td>4.79</td>
<td>1.39</td>
</tr>
<tr>
<td>Lunchtime with Secondary Afternoon Commute</td>
<td>14</td>
<td>6.19</td>
<td>90</td>
<td>14.88</td>
<td>0.42</td>
</tr>
<tr>
<td>Quiet Mornings, Busy Evenings</td>
<td>14</td>
<td>6.19</td>
<td>19</td>
<td>3.14</td>
<td>1.97</td>
</tr>
<tr>
<td>Busy Lunchtimes with Secondary Commuting Peaks</td>
<td>11</td>
<td>4.87</td>
<td>15</td>
<td>2.48</td>
<td>1.96</td>
</tr>
</tbody>
</table>

Taking a look at the table, it can be concluded that more complex three-peaked temporal profiles are much more common in London than throughout the entire Great Britain. This is especially the case with profile Commute and Lunch which is 2.17 times more common in London compared to the national average. On the other hand, cluster Midday Top is almost non-existent in Greater London, occurring at only 4 of 226 microsite locations, suggesting a dominance of multifunctional locations. Moreover, clusters Consistent Afternoons and Lunchtime with Secondary Afternoon Commute, which are also characterised by prominent midday footfall, along with the extended period of high footfall or secondary peak in the afternoon, are also less typical for central parts of the UK capital.

The next step is to create the London Footfall Temporal Classification (LFTC). An almost identical methodology was followed as in the GB-wide one. Cleaned and calibrated median hourly footfall measurements for each hour from Monday to Thursday were range-standardised on a scale of 0 through 1. A range of agglomerative hierarchical clustering
algorithms was tested and Ward’s clustering algorithm was chosen as the optimal\(^3\). The only difference to the original clustering undertaken for Great Britain was geographical extent of data, which captured only those sensor microsite locations located within the Greater London area. Sensitivity analysis was performed, and interpretability of different clustering solutions was tested. Cluster analysis resulted in six distinct groups of diurnal activity patterns (Figure 5.18).

Figure 5.18: London Footfall Temporal Classification - six-cluster solution (Monday – Thursday)

Already by a rough visual inspection, one can realise that a single-peaked cluster Midday

\(^3\)Agglomeration coefficient for the Ward’s method was the highest (0.96). It was followed by the complete method (0.88) and average method (0.79). Single method fared poorly (0.66)
Top from the GB-wide footfall temporal classification, which was found to be very rare in London, does not emerge as a distinct cluster.

Two other rare clusters (Consistent Afternoons and Lunchtime with Secondary Afternoon Commute) from the GB-wide classification are here merged into a single cluster (cluster 3, appropriately called High Afternoons, because this is the common feature of both equivalent clusters from the GB-wide classification).

Other clusters are relatively similar to the previously presented clustering solution, and little change can be seen.

Cluster 4 corresponds to cluster Lunchtime and Afternoon Rush Hour Spike from the nationwide classification, with the small difference that cluster average, in this case, comes with somewhat higher activity in the mornings. This distribution can, however, be found in some individual cases of Lunchtime and Afternoon Rush Hour Spike cluster of the nationwide classification, so it is not something completely new. The point is that London locations seem to impacted by the morning commuters more than the out-of-London locations.

Cluster 2 generally corresponds to Commute and Lunch from the nationwide classification, however, this time, the lunchtime peak is generally smaller than the morning and afternoon peak.

Table 5.9 shows the distribution of clusters in London. Proposed names for clusters were mostly inspired by the GB-wide classification with slight adaptations to reflect the specific characteristics of the footfall patterns in London. The distribution may be somewhat unexpected, as it is different from the original GB-wide classification. For example, there are more cases belonging to the Gradual Rise in London than in the nationwide classification. This is because there are fewer clusters in London case and, moreover, cluster centres are not identical due to the different (and smaller) data sample.

When it comes to spatial distribution of the clusters from the LFTC (Figure 5.19, one can notice that Quiet Mornings, Busy Evenings cluster appears almost exclusively in the heart of Central London (Soho area and some other nearby locations), which is expected due to the entertainment function those areas have. Gradual Rise locations are the most common, and they are scattered all across Central London, but they do appear to be less frequent in the touristic Central London areas such as Soho and Oxford Street. Both clusters have similarly shaped profiles, but Quiet Mornings, Busy Evenings have a much smaller significance of the morning footfall, and maximal footfall is shifted from the late afternoons (5 pm – 6 pm) towards evenings (7 pm). Clusters seem to group spatially and this is especially the case with the Quiet Mornings, Busy Evenings cluster (situated in

---

4If a higher number of clusters, such as eight is chosen, then separate cluster with three equally pronounced peaks emerges, however, in that solution there are two very similar three-peaked clusters with rising footfall towards the end of the day, so we decided to choose the simpler solution with more variance between clusters.
Table 5.9: The breakdown of clusters in the London Footfall Temporal Classification

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Proposed name</th>
<th>Frequency</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gradual Rise</td>
<td>96</td>
<td>42.48</td>
</tr>
<tr>
<td>2</td>
<td>Commute and Lunch</td>
<td>22</td>
<td>9.73</td>
</tr>
<tr>
<td>3</td>
<td>High Afternoons</td>
<td>55</td>
<td>24.34</td>
</tr>
<tr>
<td>4</td>
<td>Lunchtime and Afternoon Rush Hour Spike</td>
<td>30</td>
<td>13.27</td>
</tr>
<tr>
<td>5</td>
<td>Quiet Mornings, Busy Evenings</td>
<td>15</td>
<td>6.64</td>
</tr>
<tr>
<td>6</td>
<td>Busy Lunchtimes with Secondary Morning Commute</td>
<td>8</td>
<td>3.54</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>226</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

the central entertainment district) and Commute and Lunch cluster (situated along the Edgware Road in the west Central London). The fact that microsite locations that are closer to one another belong to the same temporal cluster is not surprising. However, what is perhaps more interesting is the fact that even at small distances, cluster membership may change. This is indicative of the impact of some specific characteristics on a micro scale, a significantly important factor retailers need to consider when making important locational decisions.
5.5 Chapter Summary

This chapter reviewed the traditional role of cluster analysis in geographical research and identified the potential which inclusion of time component may bring into our understanding of how places work. A range of commonly used time series classification methods was described, and several most suitable ones for the task were used to classify temporal dimension of the human activity patterns recorded at a set of retail microsite locations across British urban areas. The initial aim was to check whether different microsite locations in urban areas display different diurnal footfall patterns and if that is the case, the task was to check to what extent the readings from the Wi-Fi sensors can serve to derive such classification.

The cluster analysis proved that there exist significant differences among urban microsite locations in terms of footfall patterns. Eight clusters of distinctive functional characteristics for Mondays to Thursdays were identified. Fridays were found to display almost identical footfall patterns, with the only marked difference being higher evening activity in two of the clusters. The footfall patterns observed during the weekends are much less complex than workday footfall, and they are characterised by simple midday peak activity which is, at some places, accompanied by a nightlife peak.
Finally, this chapter also proved that London has different Monday to Thursday footfall patterns than the rest of the country. Having much more complex mobility patterns and distribution of workplaces, London is characterised by relatively higher importance of three-peaked footfall profiles and absence of the simple one-peaked profiles.

The next task is to combine the footfall classification with data such as road network hierarchy, proximity to bus stops, characteristics of local workplace population and the composition of local retail area. These associations are all investigated in the subsequent two chapters.
Chapter 6

Non-Retail Drivers of Footfall Variability

The footfall measurements obtained by the Wi-Fi sensors of the SmartStreetSensor project have a number of relevant applications. They are useful for measuring the site suitability, retail performance and conversion. In addition, they can be used to enable us to gain a better understanding of footfall at micro-scale and its interdependence with other relevant factors.

Amongst those factors, a particular emphasis will be on studying how the local transport geography, demographics and local retail geography affect footfall. The former two groups of variables are subject of this chapter, whereas local retail composition is explored later in the thesis.

The chapter is, thereby, organised around the following questions:

1. Do urban retail areas have homogeneous footfall, i.e. how does the footfall vary within and between the retail areas?
2. Does footfall vary by the levels of street network hierarchy and proximity to the public transport stations?
3. Does footfall vary by the type of local workplace demographics?

6.1 How Does Footfall Vary Within and Between Retail Areas?

One of the most relevant questions to be answered when exploring the spatial and temporal variations of footfall is how footfall varies and what causes its variation. As the previous chapter suggested, there exist eight different diurnal temporal profiles for Mondays through Thursdays, a similar set of eight profiles for Fridays and three distinct profiles for each...
Saturday and Sunday. A set of detected characteristic profiles appears to change when only London is taken into account, as Figure 5.19 in the previous chapter demonstrated.

Apart from the variations of diurnal footfall patterns, there exist spatial differences in overall volume of footfall observed across retail areas, as well as intraweek differences of footfall volume. Both were investigated as part of this project in a small case study conducted in the area of Central London and published in Murcio et al. (2018).

Figure 6.1 from Murcio et al. (2018) shows how average five-minute footfall varies throughout Central London. It can be seen that areas well known for their business are Soho (Central London) and Camden Town, as well as locations around some of the Tube and rail stations, with some notable examples labelled on the map (Victoria, Waterloo and Angel stations) (Murcio et al., 2018). The influence of station proximity is also seen on Edgware Road. Footfall around Edgware Road and Marble Arch Tube stations appears to be higher, while at the same time sensors between them record lower and relatively consistent and spatially comparable footfall. On the other hand, stores situated in quieter side streets or less attractive areas show lower footfall, including areas that may be near main attractions but outside main corridors – Tooley Street being a good example, situated behind the far more crowded Thames path near Tower Bridge (Murcio et al., 2018).

In addition to exploring the spatial variation of average volume, variation of volume throughout the week is also of particular interest for retailers. In this case, weekend footfall was compared to the workday footfall by dividing the average five-minute footfall on Saturdays and Sundays between 7 am and 7 pm by the average five-minute footfall on workdays between 7 am and 7 pm as follows:

\[
I_w = \frac{F_1(Sat-Sun,7-19)}{F_2(Mon-Fri,7-19)} * 100
\]

where:

- $I_w$ is the index of relative weekend daytime activity;
- $F_1$ is the average five-minute weekend footfall between 7 am and 7 pm;
- and $F_2$ is the average five-minute workday footfall between 7 am and 7 pm.
This kind of index does not necessarily tell us which areas get busiest during the weekend daytimes, but rather which areas have more pronounced daytime weekend activity relative to their daytime weekday activity. As Figure 6.2 shows, Soho, Camden Town and a location south of Hyde Park record higher footfall during the weekends than during the workdays, which can be attributed to their reputation as highly attractive tourist and/or recreational areas (Murcio et al., 2018). The results for Camden Town indirectly suggest where exactly the main attractions of the area (Inverness Street Market, Buck Street Market and Camden Lock Market further north) are located. Relative weekend activity is higher north of the Camden Town underground station and it diminishes in a southerly direction, i.e. away from the markets. On the other hand, many areas appear to be much busier during the workdays (Victoria, Waterloo and Tooley Street), where there are concentrations of working places, including the universities in Bloomsbury (Murcio et al., 2018).
Both figures, along with Figure 5.19 from Chapter 5 prove that both footfall volume and profiles vary geographically and in some cases, differences may be substantial even on the smaller distances. This is indicative of the pronounced importance of the micro-scale in retail locational planning. The rest of the section aims to explore further how footfall varies within the small areas with predefined boundaries - in this case, retail areas. Retailers that use the retail datasets aggregated on the retail area level, such as the one gathered by the Local Data Company, may wonder whether every point within a retail area records the same footfall profile. While it would be reasonable to expect that the entire neighbourhoods do have identical or at least very similar temporal distributions of footfall, this is purely an assumption that requires data-driven verification.

6.1.1 Dealing with Proximal Sensors

Before calculating the proportion of each temporal profile in the areas, proximal sensors need to be accounted for. If two sensors are located next to one another, they will measure almost
identical volumes and patterns, and if a retail area has many groups of proximal sensors, we will end up having a lot of redundancy in aggregated data. The step of aggregating either the footfall volume (numeric values) or temporal clusters to which locations belong (categorical values) of the adjacent sensors is essential before trying to draw any meaningful conclusions about how footfall varies within the retail areas. In this case, we are primarily concerned with temporal profiles.

A convenient approach that can be taken to address the data redundancy is to spatially cluster the sensors that are too close to one another. This way, if there are three sensors that are next to one another and therefore belong to the same spatial cluster, we will not record three separate observations that belong to temporal profile X, but only one as the other two measure the same pattern. This will be relevant when summarising the overall relationships on a national scale. The clusters of sensors that are very close to one another were termed spatial microclusters.

In order to spatially cluster proximal sensors, the DBSCAN method (Density- Based Spatial Clustering of Applications with Noise) (Ester et al., 1996) was employed. DBSCAN takes two parameters, namely Eps and a minimal number of points (MinPts). The minimal number of points is, as the name suggests, the smallest number of points that can qualify as a distinct cluster, while Eps parameter is the fixed threshold distance required to regard two sets of points as separate clusters. The algorithm works by detecting the ‘thinnest’ set of points, i.e. the least compact set of points from the dataset that is still compact enough to be considered a cluster, rather than noise. It then uses the Eps and MinPts parameters from that cluster as global parameters and applies them to all other sets of points (Ester et al., 1996). The sets of points may be merged into a single cluster if the minimal distance between one of their points is smaller than Eps parameter. On the other hand, “two sets of points having at least the density of the thinnest cluster will be separated from each other only if the distance between the two sets is larger than Eps” (Ester et al., 1996: 229).

The key disadvantage of this method, therefore, is the difficulty to choose the right parameters of Eps and minimum points neighbourhood. However, since these are dependent on the particular data, minimum points were easily chosen to be 1, as there are many cases of isolated sensors which should not be seen as belonging to any cluster with other sensors. Eps was set to 100 meters. The average length of road segments on which sensors are located was computed, and it amounted to 98 meters. It can be argued that characteristics of footfall are unlikely to change along the same road segment in most cases, however, when a junction is crossed, the pedestrian flow under consideration will be influenced by the inflow of pedestrians from the adjacent roads and so, volume and diurnal pattern may be modified. Further lowering of the Eps value results in some of the sensors that are situated along the same road segment and between the same two road junctions to fall into different microclusters and this is contrary to what we were trying to achieve here.
Conversely, raising the Eps value by a too high margin, results in fewer microclusters which may contain diverse footfall.

The DBSCAN method was applied to the set of 605 sensors, and this resulted in 378 distinct spatial microclusters. A relatively large number of sensor microclusters stems from the fact that there are 266 isolated locations, with no other locations within the viable distance. This is not to say that sensor microsite locations are so widely dispersed – it merely says that we aimed only to eliminate those locations that were very close to each other, in most cases adjacent. Of those 378 spatial microclusters, 86 of them are situated outside of any retail area. The majority of these locations are located in Scotland, and some of them are within England and Wales but outside of town centres. Next, in each spatial microcluster, the most common footfall profile was identified (the mode). If a microcluster had two or more equally common footfall profiles, then both, or all three were taken into account, as it may be possible that the microcluster has diverse temporal activity patterns. In most cases, however, it is understood that this is due to the small number of sensors. It happens, for example, that there are only two sensors within a microcluster of proximal sensors, one of which belongs to the profile A and another one to the profile B. Unfortunately, there is very little that can be done in this instance, so both profiles were regarded as typical, or most frequent, in that microcluster. Now that we had a set of measurements in each retail area that was coming from sensors that were thought to be sufficiently far apart, we could then proceed with inspecting how homogeneous retail areas are footfall-wise.

6.1.2 Footfall Homogeneity in Retail Areas

There are two ways we can assess whether footfall varies within the retail areas – with respect to the volume of footfall and concerning the footfall patterns. The former case was not explored here because it is obvious that numbers of passers-by will vary on different microsite locations even within the same town centre and to a large extent, this can be seen from Figure 6.1. Therefore, for now, the focus will be on whether each retail area has consistent footfall patterns. Of 106 retail areas, 59 have more than one microcluster of sensors (and 67 of them have more than one sensor). In 30 of those 59 retail areas, the majority of sensors (over 50%) belong to one of the eight footfall profiles, and in the remaining 29, the cluster memberships are more evenly distributed. This suggests that there is roughly an equal probability that a retail area will be homogeneous in terms of temporal distribution of footfall and that footfall will be heterogeneous. This finding might, however, be linked to the variable sample size of the sensors across all retail areas, so the next important step was to check whether footfall tends to be more or less homogeneous in a given retail area due to the smaller or higher number of sensors that were installed there. For this purpose, two terms are introduced and used to describe the degree of prevalence of any of the footfall profiles in a retail area:
A footfall profile is said to be prevalent (or dominant) in a retail area when it occurs more than any other footfall profile in that retail area, regardless of the percentage value. For example, if footfall profile A occurs at 30% of the sensor microclusters within the retail area and all other footfall profiles occur less frequently, then footfall profile A is said to be prevalent.

A footfall profile is said to be decisively prevalent (or decisively dominant) in a retail area if it is recorded at more than 50% of the sensor microclusters in that retail area.

The median number of sensors in retail areas where any of the footfall profiles is decisively prevalent, amounts to 3, while the median number of sensors in retail areas where there is a more even number of sensor microclusters falling into different footfall clusters (maximal percentage of any profile is 50% or under) amounts to 4. This is a very small difference and suggests that the sample size of sensors does not contribute to the homogeneity in footfall within a retail area. This finding is reinforced by the scatterplot showing the relationship between the maximal proportion of any temporal profile in a given retail area (if equal to 100, it means that all sensor microclusters in that particular retail area belong to the same profile cluster) and the number of sensor microclusters in the same retail area (Figure 6.3). Central London was excluded, as it was an outlier. Knowing that the number of sensor microclusters does not impact the recorded heterogeneity of retail areas importantly suggests that the pattern observed at the beginning of this section is not biased due to the sensor sample sizes. It is merely the fact that some retail areas are more homogeneous than the others, and retail areas are not, by definition, areas with uniform footfall.

There are examples of retail areas whose size is too big and inevitably results in various footfall profiles, the most notable being Central London. Central London is an example of a relatively large retail area containing 47 microclusters of sensors (14.9% of all the sensor microclusters that are situated inside any retail area). The most common profile in Central London is Commute and Lunch\(^2\), occurring at 34.0% of the locations. While Commute and Lunch is, in relative terms, the most common profile encountered in this retail area, none of the eight profile clusters has an absolute majority.

The final thing to check when it comes to homogeneity in footfall patterns was whether particular functional types of retail areas exhibited more or less homogeneous footfall. The percentage of occurrence of the prevalent footfall profile was plotted across the retail area subcategories, as derived by the CDRC Liverpool team (introduced in Chapter 4) (Figure 1).

\(^1\)Expectedly, there is a degree of subjectivity associated with these definitions. The main idea was to find a way of defining the most frequent footfall profiles in retail areas. However, there are different degrees of prevalence, such as an absolute or relative majority, and this guided the decision to define two similar concepts and use them in appropriate contexts.

\(^2\)The names of temporal profiles are as per National Footfall Temporal Classification in the previous chapter.
6.4). The resulting plot reveals that the frequency of the prevalent footfall profile does not depend on the retail (sub)class to which a retail area belongs. Premium Shopping and Leisure Destinations subcategory was vastly over-represented in our sample, so there is definitely room for re-assessment of this conclusion, but based on the data that were available to us, we cannot discern any particular evidence that retail occupancy characteristics are associated with particular degrees of local homogeneity in footfall.

A practical finding of this section is that retailers need to be aware of the fact that opening their store at different locations even within the same retail area might result in markedly different footfall and, consequently, conversion rates. Footfall-wise, some retail areas are more homogeneous than others, and this was not found to be a result of either variable sample size of the sensors (or their microclusters), nor it was found to be a direct implication of the local retail geography.

6.2 Footfall Variability and Transport Networks

Apart from analysing the spatial variation of footfall per se, one should consider some of the potentially most important factors driving its variation across the entire sample. This is the retail function (which is the subject of the next chapter) and variables related to the transport network and access points. The latter include the type of street on which store is located, the number of proximal public transport stations and distance to the nearest station.

In order to assess the impact of road network hierarchy on the footfall, OpenStreetMap (OSM) road classification was used (Open Street Map, 2018). The assumption is that road types differ in their width and, thus, pedestrian carrying capacity. Moreover, their locations
Figure 6.4: Distribution in the frequency of the prevalent footfall profile across retail area classes
*classification source: CDRC Liverpool (2018)*

with respect to the town centres differ. For instance, pedestrianised zones are likely to be encountered within the town centres, while the same cannot be said for the residential and other types of minor roads.

The types of roads on which the majority of sensors are located are primary, secondary, tertiary, unclassified, service, pedestrianised, residential, living street roads and footways. The former four categories are listed in order of their hierarchical importance in the national road network system. For instance, primary roads connect major cities and secondary roads link towns. Contrary to expectations, unclassified roads are not the ones of an unknown category, but their name is, instead, an artefact of traditional road network naming conventions. Based on the hierarchical importance, they are one level below the tertiary roads. Pedestrian streets mostly do not have passing vehicular traffic apart from limited times of the day and mainly include parts of retail areas. Footways are similar, but they are mostly narrower. Residential streets and living streets are similar categories that denote streets mainly situated in residential areas. Living streets are usually those streets where
Table 6.1: One-way ANOVA table: the relationship between street hierarchy and footfall volume

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>1.16E+07</td>
<td>5</td>
<td>2.31E+06</td>
<td>9.321</td>
<td>.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td>1.25E+08</td>
<td>505</td>
<td>2.48E+05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1.37E+08</td>
<td>510</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pedestrians have legal priority and car speeds are kept to a minimum (Open Street Map, 2018).

Prior to the inferential statistical analysis, sensors in shopping centres and retail parks had to be discarded due to their specific micro-location which cannot be conveniently associated with a particular street. Moreover, some sensors were located on the so-called service roads, which are defined as access roads to car parks, retail parks and similar types of estates (Open Street Map, 2018), meaning that they will not be as relevant for the pedestrian movement either.

After discarding those cases, we were left with 511 microsite locations. We then tested how: (a) volume of footfall and (b) Monday – Thursday diurnal footfall patterns behave with respect to the hierarchical level to which the local street belongs.

In order to inspect whether there is a statistically significant difference between the means of footfall volume for each of the road types, one-way ANOVA was conducted. The results are shown in Table 6.1. It can be seen that there is a statistically significant relationship ($p < 0.05$) between the street hierarchy and the average volume of footfall observed on them.

The observed ANOVA result is expected, and it can be further confirmed by looking at the mean hourly footfall by the levels of street hierarchy (Table 6.2). Primary streets are the busiest type of street in the UK national road network (679 passers-by per hour), and they are closely followed by the pedestrian streets. The finding that pedestrian streets are typically busier than the secondary streets can be attributed to the fact that they are located in the heart of town centres with pronounced retail and workplace function. Interestingly, they are followed by residential streets. However, one needs to bear in mind that OpenStreetMap defines these types of streets as the ones that do not connect different settlements, but are rather lined with housing and serve as access roads to residential houses (Open Street Map, 2018). This means that they can be situated not only in suburban areas but also in town centres just next to the primary or secondary roads. Finally, as expected, secondary, unclassified and tertiary streets record the lowest volumes of footfall, on average. The key takeaway from the two tables presented hitherto is that street hierarchy happens to coincide with the volume of footfall to a large extent. Roads that have a bigger significance
in the national road network system (excluding those that are exclusively used by vehicular traffic, such as motorways) and that are situated in the town centres, at the same time tend to have dense flows of pedestrians.

Next, we consider whether the street hierarchy influences the footfall patterns. When investigating whether there is a statistically significant relationship between the two categorical variables, the chi-square test is typically used but it may sometimes be replaced by the so-called Fisher’s exact test (Fisher, 1970). It was initially designed for statistical tests on the 2 x 2 contingency tables. Nowadays, the test is not limited only to 2 x 2 tables because Freeman and Halton (1951) developed a method which enabled its implementation on bigger tables (West and Hankin, 2008). Compared to chi-square test, it is recommended that Fisher’s exact test is given preference when the sample size is too small, however, there are various opinions in the statistics literature about what constitutes a ”small enough sample”. For example, McDonald (2015) recommends using Fisher’s test as long as the sample size is smaller than 1000. Wong (2011) suggests using it if over 20% of the expected values are lower than 5 and according to Kim (2017), when no frequencies are equal to zero. Moreover, if in doubt when it comes to the sample size, Fisher’s test is to be preferred to the chi-square test (Wong, 2011).

Unfortunately, while Fisher’s exact method enables dealing with tiny sample sizes, it is computationally very intensive and requires a lot of computing power. In our case, tables are much larger than tables that are usually discussed when talking about these methods in the literature. For example, street hierarchy levels paired with temporal profiles result in a 6 x 8 table. According to Mehta and Patel (2013), when data are too unbalanced or sparse for asymptotic methods to give reliable results and, at the same time, when data are too large for exact methods to work, Monte Carlo p-values can be computed. "The Monte Carlo option can generate an extremely accurate estimate of the exact p value by sampling tables from the reference set of all tables with the observed margins a large number of times. Provided each table is sampled in proportion to its hypergeometric probability,
Table 6.3: Test of the statistical significance of the association between the OSM road network hierarchy and footfall temporal profiles

<table>
<thead>
<tr>
<th>Value</th>
<th>df</th>
<th>Asymptotic Significance (2-sided)</th>
<th>Monte Carlo Significance (2-sided)</th>
<th>Monte Carlo Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>191.220</td>
<td>35</td>
<td>0.000</td>
<td>.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>195.278</td>
<td>35</td>
<td>0.000</td>
<td>.000</td>
</tr>
<tr>
<td>Fisher’s Exact Test</td>
<td>186.714</td>
<td>.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>2.088</td>
<td>1</td>
<td>0.148</td>
<td>.153</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>511</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

the fraction of sampled tables that are at least as extreme as the observed table gives an unbiased estimate of the exact p value.” (Mehta and Patel, 2013: 27).

To sum up, Fisher’s exact test with p-values estimated through Monte Carlo method was found to be the most appropriate method for assessing the statistical significance of the relationship between the street network hierarchy and the footfall temporal clusters.

The null and alternative hypotheses were formulated as follows:

- \( H_0 \): There is no statistically significant relationship between the hierarchical level of the street and Monday to Thursday diurnal footfall pattern recorded on the location on that street.

- \( H_1 \): There is a statistically significant relationship between the hierarchical level of the street and Monday to Thursday diurnal footfall pattern recorded on the location on that street.

It appears that there is a statistically significant relationship between the street network hierarchy and the observed temporal patterns of footfall in retail areas 6.3. Therefore, we reject the null hypothesis and conclude that different levels of street hierarchy do correspond to different diurnal patterns of footfall.
To inspect further in what way street hierarchy and footfall patterns are interrelated, the location quotients comparing the frequencies of footfall profiles observed at individual types of streets and on the national level were computed and can be defined as per the equation below:

\[
LQ = \frac{P_{str}(Y,S)}{P_{national}(Y)}
\]

where:
- \(LQ\) is the location quotient;
- \(P_{str}\) is the proportion of sensors that have footfall profile \(Y\) and are situated at the street hierarchical level \(S\);
- \(P_{national}\) is the proportion of sensors that have a footfall profile \(Y\) on a national level.

The most interesting findings arising from the computed location quotients from Table 6.4 can be summarised as follows:

- The footfall on the streets of the highest level of the hierarchy (primary and secondary roads) is more likely to be three-peaked (clusters 1, 2, 8 - refer to the GB-wide footfall pattern classification in Chapter 5 for the names and more specific characteristics of each cluster), reflecting the importance of such roads for the passing footfall of commuters.

- The footfall on the streets of the lower level of the hierarchy (tertiary and unclassified) is less complex, and it is often single-peaked with maximum footfall around midday or during the prolonged period of an afternoon.

- Pedestrianised streets seldom have three peaks and always record either single midday-peaked profiles or midday-peaked profiles with secondary late afternoon peak. This most likely suggests that they are generally not thoroughfares for commuters (at least not morning ones. They are either lunchtime destinations, midday shopping destinations or, marked by the high occurrence of cluster 6, after-work shopping destinations.

Following the inspection of how footfall varies by street hierarchy, we now inspect whether the density of public transport stations influences the footfall volume. Bus stops were taken from the National Public Transport Access Nodes (NaPTAN) dataset (Department for Transport, 2014), while train stations (inclusive of rail, London underground and overground stations) were taken from the LDCs dataset (Transport category). The use of the LDC’s dataset is justified by the fact that it is more comprehensive and already tailored for this task. Since some areas may have higher relative importance of private transportation to work, car parks were also included from the LDCs database. The kind of

---

3In LDC dataset, for example, underground stations are represented as a single location, whereas NaPTAN dataset contains points for every entrance, which was considered to be redundant.
Table 6.4: Location quotients displaying the relative distribution of temporal clusters on each type of the street

<table>
<thead>
<tr>
<th>Street hierarchy</th>
<th>Footfall temporal cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Primary</td>
<td>2.30</td>
</tr>
<tr>
<td>Secondary</td>
<td>1.03</td>
</tr>
<tr>
<td>Tertiary</td>
<td>0.80</td>
</tr>
<tr>
<td>Unclassified</td>
<td>0.55</td>
</tr>
<tr>
<td>Pedestrian/Footway</td>
<td>0.39</td>
</tr>
<tr>
<td>Residential/Living street</td>
<td>0.96</td>
</tr>
</tbody>
</table>

information we were interested in as regards the nodes of public and private transportation in the cities is their overall number around the sensors and the distance to the closest one. It is presumed that if there are many bus stops around the sensor, an entire area will be markedly influenced by the footfall generated by commuters and, therefore, footfall could be rather high. Same can be presumed for the sensors that are located just next to the bus stops, as opposed to being hundreds of meters away from them.

Bus stops, train stops and car parks whose network distance from a sensor was 500 meters or lower were selected for further analysis. It happens that out of 605 sensors, 604 of them had at least one bus stop in their vicinity, 591 had at least one car park and only 265 had a train station of any type. Overall variation of the number of train stations was very limited, with the cases mostly being either zero, one or two; with further fifteen locations having three train stations and five locations having four train stations in their vicinity. For these reasons, the number of proximal train stations or distances from the train stations were not examined. As before, to avoid data redundancy coming from the adjacent sensors, individual sensor microsite locations were merged into spatial microclusters and averages were computed for each of them.

As scatterplots on Figures 6.5 to 6.7 show, there is no clear correlation between either the number of bus stops nearby and the average volume of footfall. Same can be said for the relationship between the distance to the closest bus stop and footfall volume and the relationship between the number of car parks and footfall volume. It is understood that different car parks have different sizes and could, thus, impact the local footfall differently. However, the lack of strong correlation for the bus stops is surprising at first. It can probably be explained by the fact that at most of the points throughout town centres, especially in major cities which comprise a significant proportion of the sensor sample, there will always be quite a few bus stops in the immediate vicinity. Whether there is only one bus stop within 500 meters from the sensor or five of them - will not make a substantial difference, since even one bus stop can generate significant footfall, especially if it is of the higher order.
of importance in the urban bus network.

Figure 6.5: Relationship between the number of bus stops within 500 meters and footfall volume
Figure 6.6: Relationship between the distance to the nearest bus stop and footfall volume
6.3 Workplace Demographics and Footfall Patterns

The question of how local demographics impacts the various features of retail performance is one of the key questions retailers seek to answer and apply to the areas where they operate. That question was previously explored in past literature (Lugomer and Lansley, 2016). The newly available commercial data sources that were explored in Chapter 3 provide us with new opportunity to explore if local demographics is related to changes of footfall, which in turn presents one of the key indicators for retail planning.

Berry et al. (2016) demonstrated that workplace zone data and associated classification of workplace zones provide us with an opportunity to enhance our understanding of the drivers of store performance and observed store trading patterns. Their research was a case study of a set of chosen areas in Inner London and was undertaken in collaboration with The Co-Operative Group, which is a major UK grocery retailer (Berry et al., 2016). A step further from there could be to see how Classification of Workplace Zones (COWZ) is related to the footfall temporal patterns, and in doing so, to use a larger sample, scattered across the UK to reflect the variety of demographic characteristics of the entire country, and not
just (Inner) London.

For this purpose, the Classification of Workplace Zones for the UK (COWZ-UK) that was introduced in Chapter 3 was used to characterise the demographics of workplace population. As said, COWZ-UK outlines seven demographic supergroups and underlying 29 groups. The reason why COWZ-UK was relevant here and some other traditional census-based areal classification such as OAC were not, is the fact that OAC is concerned with the nighttime, i.e. residential population. In this case, we focus upon measuring human activity patterns that are relevant for the retail sector and in the majority of cases, population that is captured by the sensors does not incorporate a significant proportion of residents, but workers that are present in the town centre during the regular business hours and immediately before and after them. To a large extent, this will not be true for the weekend footfall, which presumably constitutes non-workers coming to the town centre for shopping and leisure purposes, so we focused on linking the COWZ-UK and Monday-Thursday footfall patterns.

6.3.1 Associations Between COWZ-UK and Footfall Temporal Profiles

The COWZ-UK was compared to the temporal classification of retail microsite locations presented in the previous chapter. Since there are no sensors located in Northern Ireland, only workplace zones in England, Wales and Scotland were taken into account. Point in polygon spatial operation was conducted, which resulted in every microsite location being associated with one footfall temporal cluster and one COWZ-UK supergroup and the underlying group. It has been understood that there are limitations to such an approach, as it oversimplifies the mobility of people. Pedestrians passing by the sensor microsite locations will not only present workforce working in the same workplace zone in which that microsite location is located. This is especially the case in big cities where workplace zones may be very granular, and thus, adjacent workplace zones’ centroids are relatively close to each other. One way to overcome this issue is to represent footfall as the weighted sum of the population structures of the neighbouring workplace zones, rather than the population structure of a single workplace zone. The zones could be weighted by their centroids’ distances from the corresponding microsite location where footfall is being measured. And while this would be conceptually valid and obey Tobler’s law, it was concluded that it would also introduce additional layer of uncertainty, due to the inevitable subjectivity associated with tweaking of the weights and distance decay parameters. Moreover, COWZ-UK was conducted on workplace zone level, so reclassifying the data after what was considered to be subjective weighting, was not expected to add more value to the analysis.

Eventually, one should also consider the fact that footfall comprises not only the local workforce but also local residents, especially unemployed and those who work very close to their home; and also, tourists, especially in the centres of major UK cities, which this
Table 6.5: Distribution of COWZ-UK demographic supergroups around sensor microsite locations

<table>
<thead>
<tr>
<th>COWZ supergroup</th>
<th>Sensor count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Retail</td>
<td>312</td>
<td>51.57</td>
</tr>
<tr>
<td>B City and Business Parks</td>
<td>216</td>
<td>35.7</td>
</tr>
<tr>
<td>C Metro Suburbs</td>
<td>37</td>
<td>6.12</td>
</tr>
<tr>
<td>D Suburban Services</td>
<td>8</td>
<td>1.32</td>
</tr>
<tr>
<td>E Manufacturing and Distribution</td>
<td>4</td>
<td>0.66</td>
</tr>
<tr>
<td>F Rural</td>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>G Servants of Society</td>
<td>23</td>
<td>3.8</td>
</tr>
</tbody>
</table>

research focuses on. However, as discussed before, workplace zones are currently the best census geography representing the population observable locally during the daytime and, thus, working hours of the majority of retail units in which sensors were installed.

All seven COWZ supergroups were represented in the sample of sensor microsite locations. However, some of those supergroups were massively under-represented because they were more closely linked to the rural areas (Table 6.5). Those supergroups are Rural with five recorded cases and Manufacturing and Distribution with only four cases (of 605). The most widely represented COWZ supergroups are Retail (51.6%) and City and Business Parks (35.7%), which is expected, given that the initial sample of sensor locations was designed to capture mainly urban cores with the high importance of retail industry.

In order to quantify the association between the two examined categorical variables: COWZ supergroups and footfall temporal clusters, a range of methods were considered. Typically used chi-square test could not be conducted because expected counts in 81.8% of the cells were lower than 5. In such cases, we cannot trust the results on the statistical significance of the chi-square statistic and alternative methods should be used and this was formerly identified to be Fisher’s exact test.

The null and alternative hypotheses for this statistical test were formulated as follows:

- $H_0$: There is no statistically significant relationship between the demographic supergroups of the workplace zones and footfall temporal profiles.
- $H_1$: There is a statistically significant relationship between demographic supergroups of the workplace zones and footfall temporal profiles.

The similar test was also conducted for the group level (lower-tier hierarchy of the workplace zone classification) and footfall temporal patterns. Both hypotheses tests’ results
are presented in the Appendices 9.4 and 9.5. When the unbiased estimate of the p-value for Fishers exact test is given through the Monte Carlo method, its value is smaller than 0.001, which means that we can reject the null hypothesis. It is highly unlikely that the observed associations were due to random chance. In other words, the demographic composition of local workforce and diurnal variation of footfall at the corresponding local microsite location do seem to be significantly associated, and this is the case at both COWZ supergroup and group level.

Having said that, the next step was to further assess the pairwise relationships between the individual demographic supergroups and footfall profiles. The idea is to provide a further description of which footfall profiles tend to be found in areas of particular workplace demographics. This was done by calculating the percentages of each COWZ supergroup in each temporal cluster and then by calculating the location quotients. The location quotients, in this case, represent how common it is to observe temporal cluster Y in demographic supergroup X compared to the overall distribution in Great Britain. The values higher than 1 mean that a particular temporal cluster is more likely to be encountered in a particular demographic supergroup than it would be expected from the national average. The resulting relationships are displayed in Table 6.6.

Being the most common, demographic supergroup Retail mainly constitutes temporal clusters with midday peak followed by either minor, pronounced or none secondary peak.

City and Business Parks supergroup is associated with three-peaked clusters: the ones in which morning, lunchtime and afternoon rush hour footfall are equally pronounced, as well as those where footfall rises towards the end of the day. Otherwise relatively rare temporal cluster Busy Lunchtimes with Secondary Commuting Peaks (cluster 8) characterised by lower morning and afternoon rush hour peaks and significant lunchtime peak, is found exclusively in this supergroup. This is obviously a generalisation because there are several different three-peaked profiles and several different geodemographic groups within City and Business Parks supergroup, but further distinctions will be drawn later.

In Metro Suburbs, temporal profiles Commute and Lunch (cluster 1) and Gradual Rise (cluster 2) are more likely to be found than expected based on the national average.

Finally, Servants of society supergroup is associated with profiles Consistent Afternoons (cluster 3) and Busy Lunchtimes with Secondary Commuting Peaks (cluster 8): both having a significant midday peak.

Other COWZ supergroups (Suburban services, Manufacturing and distribution and Rural) were covered by the too small sample of sensors, so no conclusions can be drawn about them.

The next step incorporates breaking down the supergroups into groups in order to obtain a more granular description of local daytime demographics and a more thorough understanding of the computed associations. That being said, the next step is to provide
Table 6.6: Location quotients displaying the relative distribution of temporal clusters in each COWZ supergroup

<table>
<thead>
<tr>
<th>COWZ Supergroup</th>
<th>Temporal cluster</th>
<th>Total cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A Retail</td>
<td>0.16</td>
<td>0.24</td>
</tr>
<tr>
<td>B City and business parks</td>
<td>1.97</td>
<td>1.79</td>
</tr>
<tr>
<td>C Metro suburbs</td>
<td>2.14</td>
<td>3.07</td>
</tr>
<tr>
<td>D Suburban services</td>
<td>4.5</td>
<td>0.95</td>
</tr>
<tr>
<td>E Manufacturing and dist.</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F Rural</td>
<td>1.44</td>
<td>0</td>
</tr>
<tr>
<td>G Servants of society</td>
<td>0.31</td>
<td>0.99</td>
</tr>
</tbody>
</table>

the description of how each COWZ group is related to the individual footfall profiles and highlight how associations on a supergroup level are different to the associations on a group level. The methodology was identical as in the case of supergroups and corresponding location quotients are shown in Table 6.7 below. Note that some groups were represented by less than ten sensor microsite locations so they were not explored in more detail, and hence, they were not part of the table.

Starting from the groups within the Retail supergroup, the Shop Until You Drop group is found in the large in-town retail developments and out-of-town shopping centres. The local workforce encompasses higher than average proportions of Black and Asian ethnicities and high female participation (ONS, 2018). Such areas are primarily characterised by the high footfall around midday, with secondary afternoon peak (mainly profile 6 - Lunchtime with Secondary Afternoon Commute). The resulting location quotients are very similar to the Retail supergroup as a whole due to a high share of Shop Until You Drop group in it.

Eat, Drink and Be Merry group is associated with areas where retail and wholesale are to a large extent accompanied by leisure units such as restaurants and bars. There is a higher than average proportion of European workers, and also a higher degree of part-time workers, especially students (ONS, 2018). A higher importance of leisure units in these areas is reflected in the footfall. The temporal distribution of footfall in this group is similar to the previous one, but the main difference is in the fact that temporal profile 7 (Quiet
Mornings, Busy Evenings) is most commonly encountered in these areas, which means that footfall grows throughout the day and reaches highs of the day in the evenings.

Traditional High Streets and Low-Density Wholesale and Retail are both associated with the simple midday-peaked profiles (Midday Top). Those two groups also share similar workplace demographics, in that they both mostly employ a part-time workforce, students and White ethnicities. However, the former of the two groups is mostly found in non-metropolitan areas and has people that are working in retail, finance, insurance, food and accommodation sectors, while the latter group is mostly found at urban fringes and comprises primarily workers in wholesale and retail and semi-routine activities (ONS, 2018).

The Workplace Zones classified as Market Squares are almost exclusively related to simple one-peaked profiles (Midday Top). Found in smaller towns, this group mostly comprises White British workforce, with a high importance of financial and insurance activities, followed by retail, food and accommodation services. There is a high level of self-employed, but at the same time, a low level of those working from home (ONS, 2018).

Multiethnic Urban High Streets are the only group in Retail supergroup which has a higher likelihood of containing three-peaked locations with similarly high peaks of morning and afternoon rush hour and lunchtime (Commute and Lunch). However, Midday Top and Consistent Afternoons clusters are above average.

City and Business Parks supergroup comprises four COWZ-UK groups. The first one, Big city life, entails most of the London entertainment district and other areas in the centres of major cities with higher than average proportion of employees in food and accommodation services, but also ICT, financial and insurance. The group is also characterised by a high percentage of non-British workers and those who commute more than 20 km by public transport (ONS, 2018). The fact that entertainment districts are frequently associated with Big city life group is also reflected in higher than average proportion of profile 7 (Quiet Mornings, Busy Evenings) and profile 2 (Gradual Rise) microsite locations, so these areas are generally characterised by three-peaked profiles with gradual increase towards the end of the day and relatively higher late afternoon and evening activity.

The Global Business Group is almost exclusively found in London, especially its financial districts, and rarely in other metropolitan areas. It is characterised by a high percentage of managerial, administrative and professional occupations, workers in ICT, finance, insurance, science and technology. People working in these workplace zones commonly travel more than 20 km to work and have high qualifications (ONS, 2018). Commute and Lunch is the most typical, reflecting the similar impact of commute and lunchtime. Unlike Big city life, evening activity is not relatively pronounced, which signifies the lack of entertainment function.

\footnote{The number of groups is actually five, but one of them, Science and Business Parks, has no installed sensors.}
Administrative Centres are more tightly related to footfall profile 1 (Commute and Lunch) than Big City Life areas, reflecting the importance of rush hour and eating out during the lunchtime in footfall generation. Cluster 8 (Busy Lunchtimes with Secondary Commuting Peaks), albeit relatively rare in general, is mostly found in these areas.

Compared to Big City Life, Administrative Centres have no microsite locations with higher evening activity, which leads to the conclusion that those areas are much less associated to leisure, save lunchtime venues which are very important, especially due to the high presence of profile 8 (Busy Lunchtimes with Secondary Commuting Peaks). Demographically, Administrative Centres are very similar to Global Business, but the industry of employment is mostly different and focuses on public administration.

Regional Business Centres are found in business districts of many regional cities and are not typical for London. With the dominance of high-status occupations such as finance and insurance (ONS, 2018), these workplace zones are similar to Global Business, but apart from their location, they also differ in having more than average Consistent Afternoons microsite locations, with a prolonged period of high activity around midday and early afternoon. This tends to suggest that business districts in regional centres of the UK do not have as pronounced rush hour peaks as equivalent locations in major business districts of London.

In Metro Suburbs supergroup, only two groups have at least ten or more sensor locations, hence only they will be described. Independent professional metro services are geographically restricted to the inner suburbs of London and only a few other major cities and are characterised by the cosmopolitan, higher status workforce, with a high occurrence of the self-employed (ONS, 2018). The most typical footfall profiles are three-peaked ones, either with a similar volume of footfall at all three peaks or gradually rising footfall with maximum activity during the afternoon rush hour (profiles 1 and 2).

Metro Suburban Distribution industries are also associated with Gradual Rise profile, and in addition, to profile Lunchtime and Afternoon Rush Hour Spike. These workplace zones are located in the outer suburban areas of major cities and predominantly employ workers in transport and storage industries, with wholesale and retail, food and accommodation services also being above average. As regards the origin of workers, there is also a higher than average percentage of Black and Asian minorities, as well as post-2001 EU accession countries minorities (ONS, 2018).

In Servants of Society supergroup, only one underlying group had over ten cases, and that one was Public Administration and Security. Such workplace zones are characterised by high proportion of the white workforce in intermediate occupations: mainly public administration, defence and social security (for example, city halls, council offices and prisons) (ONS, 2018). The footfall is concentrated around midday and early afternoons, reflected in higher than average presence of Consistent Afternoons temporal profile.
Table 6.7: Location quotients displaying the relative distribution of temporal clusters in each COWZ group

<table>
<thead>
<tr>
<th>WZ Supergroup</th>
<th>WZ Group</th>
<th>Temporal cluster</th>
<th>Total cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retail</td>
<td>Shop Until You Drop</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.23</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.83</td>
<td>2.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>125</td>
</tr>
<tr>
<td>Retail</td>
<td>Eat, Drink and Be Merry</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.31</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.33</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.19</td>
<td>1.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.27</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td>Retail</td>
<td>Traditional High Streets</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.76</td>
<td>3.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>61</td>
</tr>
<tr>
<td>Retail</td>
<td>Multiethnic Urban High Streets</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.03</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.41</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>Retail</td>
<td>Market Squares</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.99</td>
<td>3.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>Retail</td>
<td>Low Density Wholesale and Retail</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.43</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>City and Business Parks</td>
<td>Big City Life</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.43</td>
<td>2.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.57</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.38</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3</td>
<td>1.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>106</td>
</tr>
<tr>
<td>City and Business Parks</td>
<td>Global Business</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.8</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.52</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.28</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.16</td>
<td>3.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>55</td>
</tr>
<tr>
<td>City and Business Parks</td>
<td>Regional Business Centres</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.73</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.35</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>City and Business Parks</td>
<td>Administrative Centres</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.7</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.4</td>
<td>24</td>
</tr>
<tr>
<td>Metro Suburbs</td>
<td>Independent Professional Metro Services</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.8</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.72</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.04</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Metro Suburbs</td>
<td>Metro Suburban Distribution Industries</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.72</td>
<td>4.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.36</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.09</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Servants of Society</td>
<td>Public Administration and Security</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.99</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.16</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
</tr>
</tbody>
</table>
6.3.2 Associations Between LWZC and London Footfall Temporal Profiles

One question which immediately follows from the results presented in the previous subsection is whether we could obtain added insight into the associations between workplace demographics and footfall patterns if London was observed on its own. London is a dynamic cosmopolitan and business centre with much more complex workplace demographics and population mobility. In order to answer that question, London Workplace Zone Classification (LWZC) (Singleton et al., 2017b) and London Footfall Temporal Classification (LFTC) from Chapter 5 were compared in a similar fashion as COWZ-UK and the National Footfall Temporal Classification.

First, the goal was to test whether there existed a significant relationship between the LWZC supergroups and groups and the temporal cluster from LFTC to which the corresponding microsite locations belong. The null and alternative hypotheses had the same form as before:

- $H_0$: There is no statistically significant relationship between demographic supergroups (or groups) of the London workplace zones and London footfall temporal profiles.
- $H_1$: There is a statistically significant relationship between demographic supergroups (or groups) of the London workplace zones and London footfall temporal profiles.

Monte-Carlo simulation was conducted, and an unbiased estimate of the p-value for Fishers exact test was derived, with results laid out in the tables 9.6 and 9.7 in the Appendices. The p-value was smaller than 0.001 in both cases (LWZC supergroups and groups), which means that the null hypothesis can be rejected. There exists a significant relationship between the London workplace demographics and London diurnal footfall patterns. It is highly unlikely that this association exists due to random chance.

We now turn to describe what added value comparing LWZC to London Footfall Temporal Classification brings as opposed to simply relating the two classifications on the national level. The location quotients were computed to describe relative frequencies of different London footfall temporal profiles across different LWZC supergroups\(^5\) and results are displayed in Table 6.8. The associations between the LWZC groups (lower-tier level) and LFTC could also be considered, but there were found to be five (out of sampled nine) LWZC groups with five or fewer cluster members. This number was considered to be too small for those groups to be properly interpreted and only the four remaining groups could

---

\(^5\)LWZC uses somewhat different terminology for the hierarchies than COWZ. Top-level clusters are called groups in the LWZC (compared to ‘supergroups’ in the COWZ) and lower-level clusters are called subgroups (compared to ‘groups’ in the COWZ). However, in this work, terms supergroups and groups were adopted in both cases for consistency reasons.
be used for that purpose, which would make the interpretation pretty much identical to the interpretation on the supergroup level.

LWZC classifies workplace zones into five supergroups: Residential Services, City Focus, Infrastructure Support, Integrating and Independent Service Providers and Metropolitan Destinations. Only one of those supergroups (Residential Services) is not represented by the sample of London sensor locations.

City Focus areas are hosts of specialised professional activities, general support services and retail activities. They are important for nighttime economies and workers in all of these activities mostly fall within the 25-39 age group (Singleton et al., 2017b). These areas are the only ones with overrepresentation of the cluster 1 of the LFTC (Gradual Rise) and there is also above average representation of Lunchtime and Afternoon Rush Hour Spikes.

Infrastructure Support mostly encapsulates workers in the sectors such as transport, utilities and the retail, with high participation of Asian minorities (Singleton et al., 2017b). Infrastructure workers seem to generate Busy Lunchtimes with Secondary Morning Commute and High Afternoons footfall profiles.

Integrating and Independent Service providers are characterised by high levels of self-employment, and a significant number of part-time workers. They may be working from home, or travel to deliver services to local communities (Singleton et al., 2017b). Both Commute and Lunch and High Afternoons are typical for these areas, which reflects the nature of employment. High levels of the self-employed workers working from home would explain rather high representation of High Afternoons in which there is no pronounced rush hour footfall in the streets because the workforce is at home. On the other hand, typical three-peaked footfall which characterises Commute and Lunch cluster is probably typical for those areas with local workers who do not work from home, but commute at least on short distances and thus get captured by the sensors.

Finally, Metropolitan Destinations are heavily concentrated in the Inner London, with the prevalence of international workers who often reside in Central London and work in the retail sector and high-value services (Singleton et al., 2017b). Given that Gradual Rise and Quiet Mornings, Busy Evenings locations are often found in Central London; it comes as no surprise that those two clusters are over-represented in the Metropolitan Destinations.

It appears that analysing the relationship between LWZC and LFTC did not add much insight when compared to the analysis of the correlation between the COWZ-UK and national footfall patterns classification. Most relevant relationships have already been extracted in the previous section and they mostly have to do with the fact that areas with increased retail and leisure activity include a higher proportion of microsite locations with higher footfall towards the end of the day.

Furthermore, from the presented descriptions in the last two subsections, it seems that local retail geography helps much more in explaining footfall patterns than local workforce
Table 6.8: Location quotients displaying the relative distribution of temporal clusters for London in each LWZC supergroup

<table>
<thead>
<tr>
<th>LWZC Supergroup</th>
<th>London Temporal Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>City Focus</td>
<td>1.82</td>
</tr>
<tr>
<td>Infrastructure Support</td>
<td>0.37</td>
</tr>
<tr>
<td>Integrating and Independent Service Providers</td>
<td>0.3</td>
</tr>
<tr>
<td>Metropolitan Destinations</td>
<td>0.7</td>
</tr>
</tbody>
</table>

demographics. For example, even though it was found that Big city life COWZ group is mainly associated with those temporal profiles that are characterised by increasing footfall towards evenings, that could have probably been concluded by simply examining the relative frequencies of types of retail and leisure units in the immediate vicinity such as bars and pubs. This served as a motivation for the next chapter where local retail geography is paired with footfall to gain a better understanding of how retail areas function.

6.4 Chapter Summary

This chapter explored how footfall varies across retail areas. The emphasis was placed on the variation of footfall per se, as well as how some of the relevant microlocational factors influence footfall volume and patterns.

While locations that are close to one another are more likely to have the same footfall profile, on the scale of retail areas, one can still find a multitude of different observed profiles. Retail areas, as they are currently delineated and used for various academic and industrial purposes of data collection, storage and business analytics are not functionally entirely homogeneous areas, and it is not possible to generalise footfall patterns on such large areas.

Footfall varies based on the level of road network hierarchy of the streets on which sensors are situated. Streets of higher hierarchy record higher footfall, typically concentrated during the morning and afternoon rush hours, along with the lunchtime peak. Conversely, streets of lower hierarchy record lower footfall, which is, at the same time, less complex with peaks around lunchtime and extended period throughout the afternoon. The number and distance from the nearest public transport stations were not found to be correlated with footfall volume, which is presumed to be the consequence of their widespread are relatively uniform presence throughout town centres.

Temporal profiling of footfall combined with conventional demographic classifications provides us with an interesting insight into when areas with different demographic and socioeconomic characteristics of local workforce generate peak footfall. The relationships
between footfall and workplace demographics are not as straightforward, which reflects the complexity of urban workers behaviour which is not dependent solely on their demographic characteristics.

Nevertheless, a statistically significant relationship between the COWZ and the footfall classification was established, as some footfall profiles are more likely to occur in some areas than other. This finding is an interesting preliminary indicator of what type of footfall retailers may expect, or what type of footfall retailers are highly unlikely to expect in different workplace zones when planning to locate their businesses. Moreover, the benefit of using COWZ is the fact that it is an open source classification. It can therefore freely be used both as a primary dataset in researching what type of population is present in any workplace zone throughout the day and as an ancillary dataset paired with other data sources, such as the ones on human activity patterns used in this thesis.
Chapter 7

Local Retail Geography and Footfall

In Chapter 4 on data characteristics, two main datasets were introduced: the SmartStreet-Sensor footfall database and the national retail unit dataset, both acquired by LDC. Up to now, the retail unit database has not been used much in this thesis aside from the introductory description of the state and changes of UK retail geographies in the post-2008 recession period. In this chapter, the aim is to include retail data into explaining the variations of footfall in retail areas. As stressed in the previous chapter, there were some preliminary indications that local composition of retailing may have much more defined impact on when people visit those areas, as compared to some other contributing factors such as transport and local demographics.

The aims of this chapter were, therefore, as follows:

1. Investigate how footfall varies based on the types of stores and formats of the retail areas in which sensors are located.

2. Build and assess different approaches to collating retail composition data and footfall data. What is the best way of using footfall data to characterise a predetermined set of retail areas or newly defined retail environments around sensor microsite locations?

3. Determine what type of footfall occurs in what type of retail environments and how footfall data can enrich the retail functional classifications. Can retailers expect to find particular characteristics of footfall in the particular types of town centres and shopping centres? This was done in two different ways:

   (a) by using footfall data, i.e. footfall temporal classification to profile an existing classification of retail areas (CDRC Liverpool, 2018);
(b) by creating a completely new, unified retail-footfall classification, which will take
footfall data as part of the classification process, together with variables de-
scribing retail composition. The new classification will highlight the functional
differences that exist across the retail space in Great Britain and could be used
for inferring the likely footfall patterns for the retail areas in which no sensor
data are available.

7.1 Footfall Variability Between Different Store Categories
and Retail Area Formats

Can we expect identical footfall patterns in front of the identical store categories and general
retail area formats (high streets, shopping centres and retail parks)? This subsection further
investigates the dependence of footfall on those factors across all sample microsite locations.
It is useful for a retailer to know whether their retail units are located at the microsite
locations recording similar or markedly different footfall patterns; appropriate or unsuitable
footfall patterns.

Temporal patterns observed in front of a store are not necessarily identical to the pat-
terns observed inside the store. Chapter 4 explained in detail the process of sensor data
acquisition, cleaning and calibration, and one thing to remember is that measurements are
calibrated for measuring the number of pedestrians passing by the store, on the street pave-
ment or in front of the store inside the shopping centres. Customers who dwell longer and sit
inside the retail unit are not taken into account multiple times. Moreover, the adjustment
factor that was derived for dealing with undercounting sources of measurement error (see
Chapter 4) was computed by taking into account the number of passers-by in front of the
store, rather than inside. That being said, it becomes easier to comprehend why recorded
footfall patterns in front of the store may not be entirely relatable to the type of the retail
unit in which a sensor was installed.

If two sensors are situated in, for example, a takeaway food shop and a phone shop
directly next to it or very close to one another on the same high street, those two sensors
are expected to measure the similar diurnal footfall pattern if algorithmically cleaned and
calibrated correctly, even though the function of the underlying retail units are different.
The example of this is provided in Figure 7.1 which displays the range-normalised Monday to
Thursday average footfall for a phone shop and a fast food takeaway at High Holborn street
in Central London. One would probably not expect that high footfall at fast food takeaway
in the mornings, because the unit is not a coffee shop or a convenience store. However,
since sensors measure the flows in front of the store, they will be similar for both units,
with minor difference being that fast food shop still has a more significant lunchtime peak,
meaning that some proportion of footfall going past (and inside) the fast food takeaway
does not also go past the phone shop. It may be the case that the local workers working in the surrounding offices only walk to the fast food shop to have lunch and then go back, therefore avoiding passing in front of the phone shop.

![Footfall pattern graph]

Figure 7.1: Monday-Thursday average footfall pattern observed in front of two different retail units in the same high street (High Holborn, Central London) (2015 - 2017 overall average)

While this real-world example does demonstrate that store category may not be related to the footfall profile observed in front of the store, further validation of the claim is required on the bigger amount of data. We now present associations between ten retail categories with the highest sensor coverage and the corresponding footfall patterns. After assigning each retail unit its corresponding Monday to Thursday footfall profile, it appears that results are rather complex as there is a lack of a simple one-to-one relationship (Figure 7.2). Some notable findings include the fact that out of 15 Busy Lunchtimes with Secondary Commuting Peaks locations, seven are located in fast food takeaways and two more in different types of restaurants (pizzeria and subcategory named ‘other restaurants’). However, at the same time, we cannot conclude that this is a uniform profile associated with fast food shops, coffee shops and restaurants, as Figure 7.2 suggests that multiple different profiles are observed at those types of retail units.

Next, restaurants, in general, have a very low proportion of Midday Top footfall profiles, which is unexpected due to the fact that they should cater for consumers looking for lunch and dinner. On the other hand, this may be due to the fact that their function as a dinner place is much more pronounced, highlighted by the prevalence of Gradual rise and Quiet Mornings, Busy Evenings temporal profiles. This is interesting, as it suggests that restaurants do have some sort of lunchtime peak and late afternoon or evening peak, but simple one-peaked profiles are not typical for the footfall recorded in front of them. Both groceries and cafes fast food restaurants have a substantial proportion of Commute
and Lunch profiles. Charity shops, footwear and sport hobbies retail categories have a substantial proportion of Consistent Afternoons profiles, meaning that their peak footfall lasts a bit longer throughout midday and early afternoon. Both Commute and Lunch and Gradual rise clusters are typically encountered in convenience stores (10 of 11 convenience stores from the sample belong to one of those two clusters).

Figure 7.2: Temporal profiles of the most common retail categories across Great Britain

The next useful distinction is analysing the profiles based on the format of a retail area at which they were recorded. As said before, retail areas across Great Britain are usually classified as either town centres (high streets), shopping centres or retail parks. Unfortunately, only three retail parks were captured by the 605 locations under considerations so, due to a small sample, no conclusions can be drawn about the frequency of occurrence of different temporal profiles in retail parks. As regards shopping centres, it was found that the most common profile in the shopping centres, by far, is a simple single-peaked profile (Midday Top). While the most common profile in the town centres, i.e. high streets is the profile with a consistently high footfall throughout the whole afternoon (Consistent Afternoons) (Figure 7.3 and Table 9.8 in the Appendices). Apart from the single-peaked profile, a single-peaked profile with the afternoon inflow of commuters is also relatively common in the shopping centres.
This tends to demonstrate that shopping centres are less typically busy during the mornings and footfall tends to fall after the midday. At best, late afternoons are only busy enough to slow down the fall of the daily footfall, but insufficiently to produce the second prominent peak. Many of the shopping centres under consideration are in fact in-town shopping centres, which tends to suggest that customers prefer to shop there during their work break, whereas visiting them for after-work shopping is less typical. While some interesting findings can be obtained by observing the previous two charts, the results were expectedly complex. The complexity of results is inherently due to the fact that footfall patterns are the result of the overall features of the local neighbourhood and cannot be simply explained by the sole retail unit in which a sensor was installed. This especially becomes apparent in the sample of sensors installed in the shopping centres where Midday Top cluster is prevalent regardless of the type of the store at the microsite location. The function of the individual retail units is important to a certain extent, but in order to acquire a better insight into how characteristics of retail environments influence footfall recorded in front of the store, data should be aggregated in some way. This aggregation can be done either by taking into account more stores around the sensor to define its retail environment or by associating the footfall data with retail unit data on the retail area level. The ways footfall and local retail geography can be collated are laid out in the next section.

7.2 The Spatial Conceptualisations of Retail-Footfall Relationships

The principal reason why there are multiple, arguably, valid approaches to linking retail and footfall data lies in the fact that there are different ways to conceptualise the relationship between the footfall and the surrounding retail structure. The key word in the last statement is ‘surrounding’ because there are infinitely many ways to define what makes the
environment of a particular microsite location.

Should one relate measured footfall only with the type of store in which a sensor is located, or should we take into account all the retail units in the given neighbourhood? We saw in the previous section that the former variant is invalid and should not be applied due to the nature of SmartStreetSensor data acquisition process. If the latter approach is valid, there is a question how big the neighbourhood of each sensor should be and how we go about weighting different retail units based on their relative contribution to footfall at a distant point. That being said, the next step is to come up with a methodological framework which enables collating those inherently different datasets. Here we adopted the term retail environment to describe the immediate neighbourhood, regardless of the size, of a retail unit at which footfall is measured. We propose five ways in which we could conceptualise the relationship between retail environment E and footfall pattern at the microsite location Y. In first three conceptualisations (termed buffer-based models), the starting point from which retail environment is defined is sensor microsite location, whereas in the last two models (termed retail area-based models) retail environment is predefined area – a retail area. In all five models the focus is on the local retail geography. However, the variables that could in practice be taken into account involve not only the presence and absence of the particular types of retail units, but also factors such as vacancy rates, densities of points of interest, distances to the bus or rail stations, etc.\footnote{The latter was, however, not found to be relevant (see Chapter 6).}

**Model 1: A zero-dimensional retail environment.** A retail environment E of the sensor microsite location Y is simply a unit in which a sensor was installed. The retail environment, in this case, is zero-dimensional, i.e. a single point (Figure 7.4). Every retail unit is paired with the footfall profile of the sensor which is installed inside that retail unit. The assumption here is that footfall signature recorded at footfall measurement point is mainly or solely driven by the function of a retail unit in which sensor is placed. As the previous section demonstrated, this assumption cannot be validated by the data and, therefore, a zero-dimensional retail environment model approach should not be used when attempting to link retail and footfall data sets – at least not when footfall is calibrated the way it was calibrated in this project.

**Model 2: Equal-weight model.** A retail environment E of the sensor microsite location Y constitutes a number of individual retail units around the sensor, within a fixed Z-meter buffer. Equal weight is assigned to every retail unit, meaning that presence of every retail unit is expected to equally contribute to the observed footfall pattern, regardless of the distance, as long as the retail unit is within the predefined distance threshold. Here, distances are defined as the lengths of the shortest paths between the given sensor and a retail unit along the road network.

**Model 3: Decaying-weight model.** Retail environment E of the sensor micros-
ite location Y constitutes a number of individual retail units around the sensor, up to a predefined distance threshold, but greater weight is assigned to the closer units. The motivation behind this model is the fact that retail units directly adjacent to the unit in which sensor is installed clearly have a decisive role in defining the measured footfall pattern as compared to the retail units that are much further away. Decaying-weight model overcomes this issue by placing diminishing weight on the more distant units, but it contains a high added degree of subjectivity because we need to decide how to model the distance decay function. Literature provides some examples of building distance decay functions for modelling pedestrian movements in cities for various trip purposes (Iacono et al., 2008). Longley et al. (2015a) also provide an overview of typical distance decay functions: linear (e.g. in describing the attenuation effect of noise over distance), negative power (e.g. in describing how population density declines with distance from historic Central Business Districts) and negative exponential function (often used in retail geography to describe diminishing store patronage with distance from it) (Longley et al., 2015a).

Model 1: A zero-dimensional retail environment

Model 2: Equal-weight model

Model 3: Decaying-weight model

All weights are equal
Beyond retail environment $E$, $w = 0$

Figure 7.4: Buffer-based models of sensors’ retail environments

The following conceptualisations come with retail data collected on the retail area level. Retail areas should be classified based on their retail composition. After that, the retail area class can be paired with either each sensor microsite location separately on the one-to-many basis (model 4), or to the aggregate average footfall profile computed from the raw data on
the one-to-one basis (model 5).

**Model 4: Footfall cluster structure.** Retail environment E of the sensor microsite location Y is the entire local retail area. The data representing the retail characteristics of retail areas need to be used to classify retail areas first. Class from the resulting functional classification to which a retail area E belongs is paired with each sensor microsite location Y (Figure 7.5) on a one to many basis. After that, we sum the number of each type of footfall profiles that occur in each type of the retail area on the national level.

The model comes with issues mostly related to sampling. Some retail areas were found to have none or only one sensor, which brings up the question whether that single available microsite location with data is representative of that retail area. Moreover, a number of sensors in most of the retail areas is generally small and in some cases, there is a similar or even equal number of observed clusters in the same retail area.

**Model 5: Average retail area footfall.** Retail environment E of the sensor microsite location Y is the entire local retail area as in the previous case. An average profile across all the microsite locations that have footfall data is computed for each retail area. In a literal sense, this means that raw footfall counts of every hour of the day need to be averaged across all locations within a retail area, thereby creating a unique average diurnal profile. All average profiles on the national level are then reclassified if substantial difference compared to the existing national footfall profile classification emerges and then paired with the clusters to which their retail areas belong.

The problem with this approach is the possible lack of usefulness of averaging the footfall data within the entire retail area. One of the relative merits of using the footfall data is its granularity and the fact that it can be used to assess the suitability of locating the store on a microsite level. Averaging the raw sensor measurements between locations removes that benefit and, besides, averaging may cancel out peaks of troughs in footfall counts and lead to loss of commercially valuable information. This emphasises the point that the task of generating the single most representative footfall for a wider area such as retail area is pointless and footfall should always be analysed on a microsite level.
This concludes the discussion of features and relative merits and disadvantages of each of the five proposed models of conceptualising the retail - footfall relationships. We now proceed with using them to assess the relationships between the characteristics of local retail environments and footfall patterns. As said in the introductory paragraph of this chapter, two approaches were taken. Approach A included linking the existing retail area classification and the most frequent footfall profiles within them (linkage of the two categorical data sets). Approach B involved cluster analysis of the individual retail and footfall variables (linkage of the two sets of individual numerical variables) The Approach A was conducted first, as it was easier to implement, and made use of the readily available related research results conducted by the CDRC Liverpool.

7.3 Linking the Existing Retail Occupancy and Footfall Temporal Classifications

The linkage of the existing retail area classification and our footfall temporal classification was conducted so that all temporal profiles of the sensor microclusters were joined to the corresponding retail areas in which they were located. Referring to the previous section, the linkage Model 4 (Footfall cluster structure) was used. The reason for this choice is
Table 7.1: Cluster groups and subgroups from CDRC Liverpool’s retail area classification that are covered by at least five sensors

<table>
<thead>
<tr>
<th>Cluster group</th>
<th>Cluster subgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Leading comparison and leisure destinations</td>
<td>3.1 Premium shopping &amp; leisure destinations</td>
</tr>
<tr>
<td></td>
<td>3.2 Mass and value destinations</td>
</tr>
<tr>
<td></td>
<td>3.3 Premium out of town destination areas</td>
</tr>
<tr>
<td>4. Primary food and secondary comparison destinations</td>
<td>4.1 Vibrant urban destinations</td>
</tr>
<tr>
<td></td>
<td>4.2 Diverse district destinations</td>
</tr>
<tr>
<td></td>
<td>4.3 Urban value destinations</td>
</tr>
</tbody>
</table>

Based on CDRC Liverpool (2018)

that by using it, we do not need to compute an unreliable average profile or decide upon one single representative footfall profile for each retail area (as is the case with Model 5). As we saw earlier in Chapter 5, that would not be meaningful anyway as retail areas are very often heterogeneous footfall-wise. The first three, buffer-based models of sensor retail environments were not suitable in our Approach A simply because the existing retail classification was conducted on a retail area level and not for the sensor neighbourhoods defined by the arbitrary distance threshold. The CDRC Liverpool’s retail area classification to which footfall classification will be linked has already been introduced in Chapter 4. It is worth reminding the reader that not all the groups and subgroups of retail areas were covered by our sensor sample. Out of six groups and fifteen subgroups, only two groups and six subgroups of retail areas have at least five sensors installed in them, and they are listed in Table 7.1.

Frequencies of different temporal profiles\(^2\) occurring in each of the functional subgroups of retail areas were summed on a national level. We also computed the percentage of each of the eight footfall profiles on the national level regardless of the retail area subgroup, so that we can compare whether the occurrence of footfall profiles is under or above the national average across retail subgroups. The values comparing the frequencies of different footfall profiles in different types of retail areas and frequencies of different footfall profiles in the entire country are termed location quotients and can be defined similarly as in the section on footfall and street hierarchy:

\[
LQ = \frac{P_{\text{subgroup}}(Y, E)}{P_{\text{national}}(Y)}
\]

where:

\(^2\)To avoid data redundancy, sensor microsite locations were clustered into sensor microclusters, as described in Chapter 5.
A brief summary of the results obtained by analysing the location quotients across retail area subgroups is given. The full results for the associations between the local retail geography and Monday to Thursday footfall patterns can be found in Table 7.2. Furthermore, the results displaying the associations between the local retail geography and Saturday footfall patterns can be found in Table 7.3. Individual cases summed in the last column of both tables present sensor microclusters, instead of individual sensors.

Table 7.2: Location quotients demonstrating the relative likelihood of an occurrence of a particular Monday-Thursday temporal profile in a particular retail classification subgroup (England and Wales)

<table>
<thead>
<tr>
<th>Retail subgroup</th>
<th>Temporal profile cluster (Mon-Thu)</th>
<th>No. of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3.1</td>
<td>0.8</td>
<td>0.84</td>
</tr>
<tr>
<td>3.2</td>
<td>0.47</td>
<td>0.91</td>
</tr>
<tr>
<td>3.3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4.1</td>
<td>5.49</td>
<td>0</td>
</tr>
<tr>
<td>4.2</td>
<td>1.92</td>
<td>1.61</td>
</tr>
<tr>
<td>4.3</td>
<td>2.36</td>
<td>1.73</td>
</tr>
</tbody>
</table>

To sum up the findings from both tables, it appears that different types of retail environments do tend to generate footfall differently at different times of the day.

Areas oriented towards convenience and food generally contain more locations whose footfall follows the meal times during the day. This is particularly true for all sensors located in the group 4 (Primary food and secondary comparison destinations) where three-peaked profiles such as Commute and Lunch and Gradual rise occur more frequently than it would be expected from the national average.

Comparison-oriented retail areas generally have simpler footfall profiles, with very much defined midday peak. This is especially the case with those comparison-oriented retail areas that are associated with weaker economic retail health (for example subgroup 3.2 Mass and value destinations which is characterised by higher vacancy rates and has footfall concentrated in the very short period around midday during both workdays and weekends).

There is, however, a distinction between such declining comparison destinations and premium comparison destinations. Those comparison destinations with pronounced leisure
Table 7.3: Location quotients demonstrating the likelihood of occurrence of a particular Saturday temporal profile in a particular retail classification subgroup (England and Wales)

<table>
<thead>
<tr>
<th>Retail subgroup</th>
<th>Temporal profile cluster (Sat)</th>
<th>No. of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3.1</td>
<td>0.94</td>
<td>1.01</td>
</tr>
<tr>
<td>3.2</td>
<td>0.53</td>
<td>1.39</td>
</tr>
<tr>
<td>3.3</td>
<td>1.76</td>
<td>0.68</td>
</tr>
<tr>
<td>4.1</td>
<td>2.93</td>
<td>0</td>
</tr>
<tr>
<td>4.2</td>
<td>1.58</td>
<td>0.79</td>
</tr>
<tr>
<td>4.3</td>
<td>1.13</td>
<td>0.92</td>
</tr>
</tbody>
</table>

component tend to record increased footfall towards evenings both during workdays and weekends and have a higher significance of the nighttime economy. For example, Premium shopping leisure destinations (subgroup 3.1) are the only retail areas where Monday to Thursday temporal profile with marked evening activity (Quiet Mornings, Busy Evenings) appears more frequently. Moreover, when it comes to weekend footfall patterns, Saturday nightlife temporal profiles are more likely to occur there than in the other retail areas.

Linking the existing retail area classification and existing footfall profiles within the retail areas tends to demonstrate some generic patterns. This is useful in understanding how retail space functions, but it also casts light on the fact that footfall is far too complex to be accurately predicted by linking the two previously built classifications, one of which having been built without envisioning the potential future linkage with footfall. A simple one-to-one relationship between the retail area and footfall temporal classifications does not exist. Our view at this point is that involvement of footfall into the existing retail area classification would not work, and findings from the section about footfall homogeneity and this section confirm this.

Retail areas are predefined areas delineated based on spatial continuity of retail units, and our aim should be to create custom areas of interest for the research of footfall, rather than use the spatial units not suitable for the task. This is not to say that retail areas are not useful or outdated concepts: they are still a useful framework for data acquisition and tracking of the evolution of the structure of high streets, retail parks and shopping centres.

Therefore, the next logical step to overcome these issues is to try and devise a new classification involving both retail and footfall components from scratch.
7.4 Unified Retail-Footfall Classification

This section aims to derive the unified retail-footfall classification. Instead of linking the existing classifications and describing the likelihoods that different footfall profiles will occur in different retail area type, the task here is to create a new classification which will take individual retail and footfall variables. Since one of the conclusions of the previous sections was that there is no point in trying to come up with the representative average footfall profile of entire retail areas, this will not be a retail area classification, but rather a classification of sensor microsite locations. For that purpose, Equal-weight model from Section 7.2 was employed to define the retail environment of each microsite location. The idea is to then calculate the proportion of different types of retail units that make up the retail environment and use them as variables which will enter the classification. One should note that even with Equal-weight model there are, theoretically speaking, infinitely many ways of defining the sensor retail environment because there are infinitely many cut-off distances that can be chosen to define the boundary of the retail environment. In the end, the 500-meter network distance was chosen as a cut-off, as it was thought that it incorporated a sufficient number of retail units into every sensor retail environment and it was not too wide. An origin-destination network distance matrix was generated, and retail units that were further away from any sensor were filtered out. Now that we defined the retail environment of all 605 sensor microsite locations, the next step was to generate continuous numerical variables relevant to the unified classification. There are, therefore, three main groups of variables: footfall and retail. We will start by introducing the footfall variables and then move on to the retail ones.

7.4.1 Footfall Variables

It is important to emphasise that results of the Footfall Temporal Classification from Chapter 5 will not be used here. The goal is to make all variables continuous in order to enable the use of one of the commonly used clustering algorithms. In the case of footfall, that means that we need to approach differently in describing it. Rather than stating that microsite location Y has temporal profile X, we aim to derive a series of simple temporal indicators which represent the volume and temporal distribution of footfall at a microsite location.

Essentially, there are three groups of temporal variables that need to be created:

- the variables representing the intraday oscillations;
- the variables representing intraweek oscillations;

\[^3\text{In addition, some other, ancillary variables were considered, but were eventually not used. A note on them is given in the subsection on retail variables.}\]
• the average volume of footfall

The variables representing how footfall varies within a day were simply taken as proportions of footfall in particular parts of the day. For this purpose, eight Monday-Thursday temporal profiles were visually inspected to see when footfall usually records peaks or troughs during the typical workdays. It can be concluded that they appear at well-defined times and thus, footfall variables to enter the classification are, therefore:

• % of daily footfall between 8 am and 10 am (morning peak);
• % of daily footfall between 10 am and 12 pm (late morning trough);
• % of daily footfall between 12 pm and 2 pm (midday peak);
• % of daily footfall between 2 pm and 4 pm (afternoon trough);
• % of daily footfall between 5 pm and 7 pm (late afternoon peak);
• % of daily footfall between 7 pm and 9 pm (evening weekday activity)

Unlike Monday through Thursday profiles, weekend profiles are less complex and the most distinctive feature that can be taken from them is the presence or absence of marked nightlife activity. Daytime is almost always characterised by the midday peak and apart from the duration of maximal footfall, there is minimal variation of daytime footfall across locations. The variable representing nightlife activity that was taken from weekend footfall was the proportion of Saturday footfall between 8 pm and midnight in the overall Saturday footfall.

In terms of intraweek variations of footfall, only one variable was created and it was the index of relative weekend activity that was introduced by Murcio et al. (2018). Relative daytime weekend activity was expressed as a ratio between the average footfall on weekends between 7 am and 7 pm; and average footfall on workdays between 7 am and 7 pm. Values higher than 100 represent locations in which weekend daytimes are busier than workday daytimes, and vice-versa. Finally, while daily and weekly temporal profiles of footfall are essential, retailers are often even more interested in the overall volume, which is relatable to the potential sales and revenue. Median weekly footfall was calculated to characterise the volume at every given location.

7.4.2 Retail Variables

As regards retail variables, the first decision to make was whether to consider using retail categories of the units or subcategories. Retail subcategories were eventually not taken as variable candidates, because that level of the retail data set hierarchy contains too many
potential variables, and their count per microsite location would end up being too low, with too many zero-values. Therefore, the proportion of retail categories occurring within the retail environment of each sensor microsite location were taken into account as variable candidates, however, before their inclusion into the cluster analysis, we had to check if they satisfy some of the standard conditions. First, variables that do not occur frequently, i.e. their overall absolute count is low were also considered for removal. Some categories predominate in the sample of 500-meter retail environments where sensors are installed. These are cafes and fast food restaurants (6852 units), hairdressers, health and beauty parlours (6721), restaurants (5588) fashion and general clothing (5360), and bars, pubs and clubs (3947). The distribution of frequency of the retail categories is displayed in the histogram in Figure 7.6 to decide whether there is a particularly prominent cluster of categories that have a small number of units that could be easily removed. The binwidth was 100, and it can be seen from the left side of the histogram that there are five retail categories with 100 or fewer units. The 100-200 bin contained three categories, and the following two bins (200-300 and 300-400) contained four retail categories each. The removal of both latter two bins was deemed as too drastic data reduction, but 200-300 bin contained some retail categories that were not thought to be of great importance for our classification, so those four categories were removed.

Figure 7.6: Absolute frequency of retail categories with a particular number of retail units in the 500-meter retail environment sample of sensor microsite locations

The list of relatively infrequent categories (< 300 units in the sample) contained the following: Royal Mail delivery offices (14 units), petrol filling stations (33), auto and accessories (35), pet shops (77), car and motorbike showrooms (78), household and home shops (107), shopping centres and markets (144) and florists (199), butchers and fishmongers (228), off
licences (232), auto services (276), department stores and mail order (287). In addition, categories medical, miscellaneous and non-retail were removed because they are either not retail units or they are too diverse - for example, miscellaneous category contains units such as credit unions, libraries, storages, photographers, driving schools, etc. Second, the number of variables can be further brought down by taking a look at their overall variation. For this purpose, boxplots were created for each retail classification (top-tier hierarchical level of the retail data set) (Figure 7.7), and those variables with the smallest variation were considered for removal. The specific characteristic of value distributions present across all retail categories is a common occurrence of outliers. The span of proportions of every retail category is relatively compact in most cases, with some locations markedly standing out. The examples of the latter are fashion, hairdressers, health and beauty, bars, pubs and clubs, cafes and fast food and restaurants. Based on the comparatively low variability, discount stores may be removed. However, they were seen as a valuable indicator of the state of the high street. According to PwC (2016), discount stores, along with health care and fast food units tend to occupy the previously vacated units. Also, discount stores registered the greatest drop in 2017 in terms of their Compound Growth Annual Rate, to 2% from the 5.1% five-year average (Grimsey Review, 2018). Removing them from classification would neglect their importance as the bellwether of the local retail economy. Similar can be said for pawnbrokers, so they were also kept.

On the other hand, locksmiths and shoe repairs and travel agents were removed due to their low variability and smaller expected contribution to the interpretability of the final classification. Interestingly, the footwear category also exhibits small overall variation, unlike (general) fashion. Footwear retailers were not removed, due to the fact that they can be usefully merged with general fashion.

In addition to retail units, there are some other geographical factors potentially important for the explanation of local footfall patterns, for example, presence or absence of non-retail points of interest such as tourist attractions and proximity to the stations of public transport. The problem with the inclusion of tourist attractions into the analysis is their either sparse or widespread distribution. For example, the main tourist attractions are heavily concentrated in Central London and to a lesser extent in a few other cities. Public and private transport is, on the other hand, present everywhere throughout town centres. While variables associated with transport such as distances to the closest bus stops and a number of proximal bus stops are poorly correlated with footfall as was seen in Chapter 6, they would therefore not violate multicollinearity rules for cluster analysis. They were initially incorporated into the classification, but no interesting insight was thereby produced and some useful results that otherwise emerge when they are not included, become obscured.

---

4This fact might call for a special classification conducted only for London, as was done in the case of LOAC and LWZC.
After recursive testing of the different combinations of variables that could be fed into the classification along with the aforementioned retail and footfall variables, it was decided that public and private transport variables should be omitted from the final list of variables.

Next, all the candidate variables from both two categories of variables had to be checked for data redundancy issues. Accounting for multicollinearity presents an important step that precedes cluster analysis, and there are many ways how this can be done. However, the common denominator of all the popular approaches is that they entail a certain degree of subjectivity when deciding upon the cut-off values. This is the case because there is no uniform way of deciding what degree of correlation between variables can be classed as too high. Due to its simplicity, the Pearson correlation matrix was generated prior to cluster analysis in order to inspect pairwise correlations between proportions of each retail category in retail areas.

Several different cut-off values were tested on the Pearson correlation matrix. Tabachnick and Fidell (2013) suggest that correlation coefficients among independent variables should be smaller than 0.9. In our case, the main idea is that the cut-off does not have to be particularly high, as exceptionally high correlations (0.8, or 0.9 and above) are not to be expected in social sciences. The first cut-off value that was tested was 0.7 and by selecting it, only one pair of retail variables appear to be flagged as highly correlated: fashion-footwear (0.77). The high positive correlation between the proportion of general fashion and footwear retailers is unsurprising, given the fact that they are comparison shops offering similar types of products. Rather than removing one of them, it makes sense to sum their proportions and create one category out of them. Among footfall variables, many of them are mutually correlated. Volume indicator, relative weekend activity and proportion of morning footfall are not highly correlated with any other footfall variable.

If we lower the absolute value of the correlation coefficient limit to 0.6, we find no other highly correlated retail variables. There are a few additional pairs of footfall variables with a high correlation that are detected now in the [0.6, 0.7] interval: early afternoon vs. Saturday evening (-0.68), evening vs. midday (-0.68), Saturday evening vs. late morning (-0.67), Saturday evening vs. midday (-0.63) and early afternoon - late afternoon (-0.63). However, it would not make sense to discard all of those variables, because then we would be left without indicators of diurnal temporal distribution of footfall.

Furthermore, if 0.5 is set as a cut-off value for the Pearson correlation coefficient, then a few more highly correlated pairs of variables emerge. All pairs (disregarding pairs of footfall variables) which recorded correlation higher than 0.5 are listed in Table 7.4, and all of the listed correlations are significant to a 0.05 level.

Shoe shops and fashion shops were merged, whereas coffee shops and fast food restaurants were separated because they were expected to impact the footfall differently. DIY stores and sports hobbies stores were dropped. Restaurants were kept because they were seen as
Figure 7.7: Variation in the proportions of retail categories across retail environments (500 m network distance) of the sensor microsite locations

<table>
<thead>
<tr>
<th>Correlated pair</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fashion - Footwear</td>
<td>0.77</td>
</tr>
<tr>
<td>DIY - Launderettes</td>
<td>0.58</td>
</tr>
<tr>
<td>Sports &amp; Hobbies - Footwear</td>
<td>0.54</td>
</tr>
<tr>
<td>Restaurants - Late morning footfall</td>
<td>-0.54</td>
</tr>
<tr>
<td>Restaurants - Late afternoon footfall</td>
<td>0.51</td>
</tr>
</tbody>
</table>
an essential variable associated with upmarket areas and/or areas that are leisure-oriented. 
Aside from the reduction in the number of variables, 16 locations were also removed prior 
to the cluster analysis, because they were found to have too few (< 20) retail units in their 
immediate retail environment (500 m network distance). Keeping them would cause some of 
the clusters centroids to have outlying values. After the number of variables was reduced to 
ensure that only the most relevant ones remain, the values of all variables were standardised 
by their conversion into the z-scores:

\[ z_i = \frac{x_i - \bar{x}}{s} \]

where:
\( z_i \) is the computed z-score;
\( x_i \) is the value of an observation;
\( \bar{x} \) is the mean of a variable across all location;
\( s \) is the standard deviation.

7.4.3 Cluster Analysis

Footfall and retail variables all entered the cluster analysis. Typical clustering algorithms 
introduced earlier in Chapter 5 were tested, together with the most common distance meas-
ures. A specially developed R package ClValid developed by Brock et al. (2011) was used 
in this case along with NbClust (Charrad et al., 2014). NbClust does not contain function-
ality necessary for validating clusters that emerge from the Partitioning Around Medoids 
(PAM) method, while ClValid does have it, and in this cluster analysis, PAM could not be 
ruled out as a viable method at the very beginning, as was the case with Footfall Temporal 
Classification. Similarly, as in the case of the footfall temporal classification, the results of 
this assessment are simply algorithmic suggestions and may not always be optimal for all 
pragmatic purposes. That is why they were taken as initial guidance, and the final number 
of clusters was determined based upon the inspection of the characteristics of the derived 
clusters and their usefulness in the context of retail geography and retail planning.

ClValid clustering validation package contains validation methods that can be divided 
into three groups: internal (connectivity, silhouette width and Dunn index), stability (av-
erage proportion of non-overlap, average distance, average distance between means, figure 
of merit) and biological methods (BHI and BSI) (Brock et al., 2011). Internal validation 
methods reflect the compactness, connectedness, and separation of the cluster partitions; 
whereas stability measures compare the results from clustering based on the full data to 
clustering based on removing each column, one at a time (Brock et al., 2011). Biological 
validation evaluates the ability of a clustering algorithm to produce biologically meaningful 
clusters (Brock et al., 2011) and due to its specific applications that are not relevant to
Table 7.5: Results of the cluster validation: top three cluster solutions obtained by different clustering methods and validated by different validation methods

<table>
<thead>
<tr>
<th>Validation method</th>
<th>Rank/number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>APN</td>
<td>hierarchical-2</td>
</tr>
<tr>
<td>AD</td>
<td>pam-9</td>
</tr>
<tr>
<td>ADM</td>
<td>hierarchical-2</td>
</tr>
<tr>
<td>FOM</td>
<td>pam-9</td>
</tr>
<tr>
<td>Connectivity</td>
<td>hierarchical-2</td>
</tr>
<tr>
<td>Dunn</td>
<td>hierarchical-2</td>
</tr>
<tr>
<td>Silhouette</td>
<td>hierarchical-2</td>
</tr>
</tbody>
</table>

geography, it was not employed in this case. That being said, internal and stability validation was conducted by taking into account Euclidean distance as a dissimilarity measure and three clustering methods: PAM, k-means and Ward’s hierarchical clustering. A range of the proposed numbers of clusters from the [2,9] interval was fed into the three selected clustering algorithms, and the top three suggested combinations of clustering methods and the optimal number of clusters were given below for each of the aforementioned validation methods (Table 7.5).

It can be seen that some of the solutions prevail regardless of the validation method used. For example, hierarchical clustering with only a few clusters is the most common solution. K-means emerges as an optimal solution when the number of clusters is 2 or 9, whereas PAM is only an optimal solution when the number of clusters is 8 or 9. This set of solutions is seen as problematic because there is a question of pragmatic usefulness of having only two or three clusters, whereas eight or nine clusters may seem a lot - depending on the cluster profiles and their interpretation.

The lists of the best performing clustering solutions can be unified into one list that shows which clustering algorithm and number of clusters comes out as a clear winner. This can be done by finding a list of solutions that minimise the distance between itself and the individual lists for each validation measure (Brock et al., 2011). Here we used Spearman's footrule distance as a distance measure and cross-entropy Monte Carlo algorithm. The aggregated list of optimal methods thereby obtained is as follows:

---

5 Average, single and complete hierarchical clustering were also considered. Agglomeration coefficients were calculated for those three methods, and the Ward’s method and Ward’s ended up having the highest agglomeration power (0.96). Conversely, the single method yielded the value of the agglomeration coefficient of 0.80, average method 0.82 and complete method 0.86.
1. Hierarchical with 2 clusters
2. Hierarchical with 3 clusters
3. K-means with 2 clusters
4. Hierarchical with 4 clusters
5. Hierarchical with 5 clusters
6. Hierarchical with 6 clusters
7. Hierarchical with 7 clusters
8. Hierarchical with 8 clusters
9. Hierarchical with 9 clusters
10. PAM with 9 clusters

Ward’s clustering occupies the most of the list top. Two and three-cluster solutions are statistically by far optimal solutions and they are achieved by both k-means and Ward’s clustering. Overall, we can conclude that generally, for our data set, the optimal cluster solution is yielded by a hierarchical method and comprises only a few clusters. Partitioning clustering methods appear to prefer more clusters in the solutions, but they are statistically less optimal. Priority was eventually given to those cluster solutions with a moderate number of clusters. Hierarchical (Ward’s) five-cluster solution (Figure 7.8, Table 7.6) was accepted as the most appropriate one. It was ranked fourth in the unified list but it was still one of the top suggested solutions, and it provided solid findings. Other solutions were also tested, and it was found that clusters ceased being meaningful and adding to the overall interpretability of the classification once the number of clusters reached 6. On the other end of the spectrum, solutions with only a few clusters do not generate a lot of insight, as retail space is more diverse than it would be suggested by characterising it by only two or three clusters.
Figure 7.8: Dendrogram of the chosen five-cluster solution obtained through Ward’s method
Table 7.6: Absolute and relative frequencies of retail-footfall clusters

<table>
<thead>
<tr>
<th>Cluster label</th>
<th>Cluster name</th>
<th>Occurrences</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Morning Coffee and Evening Dining</td>
<td>135</td>
<td>22.9</td>
</tr>
<tr>
<td>B</td>
<td>Nightlife and Entertainment</td>
<td>129</td>
<td>21.9</td>
</tr>
<tr>
<td>C</td>
<td>Midday Retail Environments with Limited Leisure Facilities</td>
<td>112</td>
<td>19.0</td>
</tr>
<tr>
<td>D</td>
<td>High-Footfall Weekend Comparison Destinations</td>
<td>142</td>
<td>24.1</td>
</tr>
<tr>
<td>E</td>
<td>Workday Services and Convenience Destinations</td>
<td>71</td>
<td>12.1</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>589</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

According to Table 7.6, the clusters are relatively evenly distributed across the British retail space. The most frequently occurring cluster is High footfall weekend comparison destinations (24.1%). It is followed by Morning Coffee and Evening Dining (22.9%), Nightlife and Entertainment (21.9%) and Midday Retail Environments with Limited Leisure Facilities (19.0%). The least frequent cluster is Workday Services and Convenience Destinations (12.1%).

Next, profiles of all five clusters were plotted (Figure 7.9) and described. Retail variables occupy the upper part of the y-axis, whereas footfall variables occupy the lower part of the y-axis. One important point to make is that the focus of pen portraits below is on retail function. However, it is difficult to ignore other functions that are implicit in the temporal profiles.

**A - Morning Coffee and Evening Dining** cluster is characterised by a higher than average proportion of coffee shops, fast food restaurants, general restaurants and accommodation. Footfall has above average concentration in the rush hour times (mornings, late afternoons), and also in the evenings, and in relative terms, footfall is not as pronounced around midday. Footfall volume is average and relative weekend activity is significantly below average. Altogether, this means that cluster A locations are popular eating out destinations. During the daytime, they are busier over the workdays and less busy during the weekends. However, weekend evenings are important periods of higher footfall. Comparison

---

^Note that this distribution of frequencies is influenced by the sampling bias, i.e. the fact that many sensors are relatively close to each other and thus may have a very similar structure of retail units in their 500-meter retail environment, and thus belong to the same cluster.
Figure 7.9: Pen portraits of the five-cluster solution obtained through Ward’s method
shopping is below average and vacancy rates slightly above average.

**B - Nightlife and Entertainment** cluster is similar to cluster A in a sense that it is also leisure-oriented. However, the difference lies in the somewhat higher importance of some comparison categories and the nighttime economy and entertainment. There is a higher than average proportion of bars, pubs and clubs, entertainment category and restaurants. These high streets and shopping centres have good economic health, as vacancy rates are below average. The daily footfall concentration is shifted towards the end of the day (late afternoon and evening), and this is even more pronounced than around the cluster A locations. The high importance of Saturday night footfall reflects the specialisation of these areas for nightlife entertainment.

**C - Midday Retail Environments with Limited Leisure Facilities** have a higher percentage of hairdressers and beauty salons, pawnbrokers, discount stores and charity shops. There is also an above average presence of electrical goods shops, tobacconists and newsagents and finance. There is a substantial absence of restaurants, cafes and fast food restaurants, bars, pubs and hotels. Groceries and supermarkets are also below average. Vacancy rates are above average. Footfall is concentrated in the first half of the day, with late mornings and midday being the busiest. The volume of footfall is below average.

**D - High-Footfall Weekend Comparison Destinations** are predominantly comparison shopping destinations comprising a relatively high percentage of fashion and footwear, jewellers, clocks and watches shops, electrical, chemists, tobacconists and newsagents and gift shops. Similarly to Cluster C, typical daily footfall is concentrated around midday, but unlike cluster C, mornings record much less footfall. Footfall volume is the highest across five clusters and so is the relative weekend activity. Therefore, cluster E locations are typical weekend daytime comparison shopping destinations with relatively insignificant role of leisure units and evening footfall and below average vacancy rates. The fact that comparison-oriented centres have the highest footfall is consistent with the comparable studies of footfall in the UK retail areas Mumford et al. (2017).

**E - Workday Services and Convenience Destinations** specialise in different kinds of services (estate agents, launderettes, hairdressers and beauty salons) and also have a higher than average percentage of grocery stores and supermarkets. Footfall volume is generally lower and weekend daytimes are also less busy than the national average. Vacancy rates are slightly above average.

The spatial distribution of clusters across Great Britain is displayed in Figure 7.10. The reader should be reminded that only a sample of retail areas was taken nationally, although the sample was intended to be regionally representative. Since distances between the individual microsite locations are rather small given the scale of Great Britain, several town centres were selected as examples and shown at a larger scale. Throughout the Midlands and North England Midday Retail Environments with Limited Leisure Facilities
prevail, which are characterised by a higher proportion of retailers that usually focus on less affluent consumers (discount stores, charity shops, pawnbrokers). The exceptions in those regions are some of the major cities, such as Liverpool and Leeds, where comparison retailing function becomes more important and which have higher relative weekend activity compared to other locations. High-Footfall Weekend Comparison Destinations are also very frequent in the more affluent parts of the country in the south (Brighton, Reading, Oxford, Cambridge). Nightlife and Entertainment cluster is present to a certain extent in most of the cities, but in some of them, it is particularly marked (e.g. Edinburgh).

If we zoom in to Greater London area (Figure 7.11), we can see that all five clusters are still present, however, Midday Retail Environments with Limited Leisure Facilities become rare and only occur at the edge of Greater London (Bromley). Most microsite locations situated in districts outside of Central London belong to the Workday Services and Convenience Destinations, meaning that while there are some comparison retail units, there is higher than average incidence of services, groceries and supermarkets. High-Footfall Weekend Comparison Destinations are not typical for Greater London, and there are just several microsite locations that are an example of this functional cluster (parts of Kingston Upon Thames, Croydon, Bromley and Tottenham Hale).
Figure 7.10: The spatial distribution of retail-footfall clusters across Great Britain (excluding London)
Finally, if we move onto investigating Central London, two of the clusters that do not exist outside of Central London tend to occur there (Figure 7.12). Expectedly, one of them is Nightlife and Entertainment which is highly clustered in areas traditionally known for their vibrant nightlife (Soho and Camden Town). Morning Coffee and Evening Dining locations also appear only in Central London and not in Outer London and they can be found primarily around major transportation hubs (Paddington, Waterloo, Victoria, London Bridge) or simply locations in which commuting workers have decisive impact on the footfall patterns and retail supply (Tottenham Court Road, Holborn). This is in accordance with the characteristics of the diurnal footfall patterns of this cluster which include higher than average concentration of footfall during the rush hours and also markedly lower footfall volume during the weekends as compared to workdays. This is not to say that these areas are completely deserted during the weekends, especially the aforementioned train stations which are heavily used by the tourists and London residents looking for a weekend city break. Nevertheless, since a number of people going through those stations during the regular workdays is immense, it is relatively larger than footfall during the weekends. In addition, areas such as Holborn definitely get a lot quieter during the weekends simply because of their marked function of work, with leisure and shopping being negligible. Generally speaking, the area of Central London can be divided into three zones, each dominated
by one of the three retail-footfall clusters. Due to the high density of sensors and due to the fact that retail environment of every sensor was defined in a consistent way by taking into account 500-meter network distance buffer, we can clearly see where zone dominated by, for example, Nightlife and Entertainment ends and where another zone, for example, Morning Coffee and Evening Dining, starts. In this case, boundary can be established along Oxford Street.

Figure 7.12: The spatial distribution of retail-footfall clusters across Central London

One other interesting result of this cluster analysis is that microsite locations along Oxford Street or any other locations in Central London did not fall into the High-Footfall Weekend Comparison Destinations. Since Oxford Street is one of the prime examples of comparison shopping, this may seem peculiar, however, this cluster is characterised by significantly higher footfall during the weekends and Oxford Street is rather busy throughout the entire week (as demonstrated by the analysis of spatial distribution of the relative index of weekend activity by Murcio et al. (2018)). Moreover, Oxford Street is, as said, surrounded by zones with different functions (Morning Coffee and Evening Dining; and Nightlife and Entertainment). When the 500-meter retail environment is defined starting from the microsite location at Oxford Street, it will inevitably capture both neighbouring functional areas. We strongly believe that places of complex footfall and retail structure such as Oxford Street should be approached carefully when planning store locations, because
people walking through them may not have them in mind as a final destination, but they may just be on their way to one of the neighbouring areas – say, Soho – a popular nightlife destination.

7.5 Chapter Summary

This chapter explored the relations between the retail composition of the town centres and shopping centres in which sensors are located and footfall profiles recorded within them. Collating footfall and retail unit datasets per se presents a significant methodological challenge, and five different models of combining the data were suggested, described and shortcomings and advantages of each were laid out. Of the five suggested models, footfall cluster structure model was adopted for the initial linkage of the existing retail area classification and the existing footfall profiles classification. After that, equal-weight model was adopted for creating the unified retail-footfall classification.

There are several key takeaways of the analysis. Referring to the conclusions of Chapter 6, retail areas are not functionally entire homogeneous areas, which served as motivation to create a unified retail-footfall classification from a microsite location perspective. Rather than trying to relate all retail units confined within the boundaries of a retail area and some sort of average representative footfall of the entire retail area, it was found that it makes more sense to start from the microsite level by defining the area around each sensor location separately, although, with a common criterion.

Second, the composition of retail areas does seem to be related to the detected footfall patterns. This was confirmed by using both approaches. Approach A through the use of the separate existing retail area and footfall classifications found that retail areas specialising in convenience and food tend to be more closely associated with the three-peaked footfall profiles, while those specialising in premium comparison and leisure are the ones attracting after-work, evening and weekend nighttime footfall. Simple one-peaked profiles can be found in other types of comparison destinations characterised by less affluent catchments, higher vacancy rates, weaker economic conditions and lack of premium upmarket retailers.

According to approach B, characteristics of the retail environments and microlocal footfall can be explained by five main clusters in Great Britain, each having specific characteristics of diurnal and weekly footfall, average footfall volume and relative proportions of different types of retailers. This serves to prove that footfall is indeed related to the characteristics of microsite locations’ retail environments.

The methodological value of the findings of this chapter is proof that the infusion of footfall data can enrich what would otherwise be another conventional retail classification. The unified classification blends the short-term functional dynamics in form of diurnal and weekly footfall patterns with longer term economic dynamics proxied by vacancy and retail
occupancy composition. In broader sense, the classification improves our understanding of
the characteristics and value of different microsite locations for retail planning purposes.
That being said, one extension of this research could be to use the LDC retail unit database
to infer characteristics of areas in which there are no sensors. This would, however, be
restricted to the types of areas in which sensors are located, i.e. primarily high street and
to a certain extent stand-alone shopping centres, but not retail parks.
Chapter 8

Conclusions

Retail areas consistently go through series of structural changes. The new trends in retailing, which emerged over the past five decades, were all fueled by the changing consumer lifestyles and demand, technological advances, real estate and land prices, as well as local and national government planning policies. The out-of-town retail developments flourished, leaving traditional high streets vulnerable and causing them to decline. Furthermore, the rise of e-commerce posed a seemingly even bigger threat, given that the new retailing channel made possible to shop without even visiting a physical store.

Notwithstanding, the forecasts for the demise of town centres have proven to be unreasonably pessimistic. Online shopping proved to be a less sensible option for particular types of items, and moreover, the introduction of combined modes of shopping, such as online shopping with Click and Collect services, kept consumers in the high streets. Governmental planning policies ensured that town centres are given preference whenever possible when locating new retail developments. That being said, they still play a vital role as the places where people go not only for shopping but also, to indulge in leisurely activities. The emergence of the notion that people visit retail areas not only to go to the comparison and convenience stores, but also to dine out or visit a cinema, influenced the way retailers make locational decisions. It should, therefore, come as no surprise that there has been a growing demand, in both the academic and commercial sectors, for the better grasping and understanding of the complex spectrum of factors that drive the changing nature of retail areas.

In the age of modern technology and data-driven analytics, taking and studying the static snapshots of space, as was done in the traditional censuses and different sorts of purpose-built manual surveys, has become insufficient. Birthed in an era of technological advances, a range of new data acquisition methods emerged at the turn of the century. Most of them occurred under the common discourses of Big Data and Smart Cities, referring to the fact that data have become much more voluminous and have made the measurements of form and function of urban areas much more thorough, automated and cheaper.
This also meant that, with the dawn of Wi-Fi, GPS and Bluetooth technologies, as well as substantial rise in smartphone ownership, every person has become a carrier of a wealth of commercially valuable data. Data could now be passively acquired from consumers, as they moved throughout retail areas and as their smartphones sent out signals in the intermediate space. Furthermore, this incredible technological leap did not only enable us to gain more information about consumers but also about places where consumers go. The aim of this thesis was, therefore, to investigate the extent to which new forms of consumer data, in particular LDC’s expansive retail unit database and LDC’s SmartStreetSensor footfall dataset, can improve our understanding of how retail areas are structured and how they function.

This chapter concludes the thesis by summing up the most significant findings from earlier empirical chapters. The research presented in this thesis breaks new grounds in several different domains, as is described in Chapter 1. These domains are visited separately in the subsequent sections, followed by suggestions for future extensions to this research and concluded with some final remarks.

8.1 The Composition of Retail Areas

The previously listed technologies enabled retailers to gather vast quantities of data not only on consumers, but also on the built environment and individual objects, e.g. retail units. The creation of a national retail unit inventory, such as LDC’s retail unit database, may have been possible before those technologies developed but data acquisition and storage would have been rather slow, cumbersome and costly. Nowadays, a lot of data can be validated without ever leaving the office by consulting the equivalent existing databases and comparing them to what has been acquired. While this presents significant technological and methodological progress, the complete disconnection from the real world should not become a guiding principle in the creation of such large-scale databases.

Office-based data validation should, thereby, still be supplemented with some sort of ground-truthing exercises. For example, LDC have a team of dedicated field researchers who travel around the country and update the expansive retail unit database on a periodical basis. Small hand-held devices such as smartphones and tablets that are connected to the Internet on the go, along with the apps installed on them enabled the ground-truthing process to become smoother and quicker. This field survey-driven approach, to data collection and validation, makes for one of the greatest advantages of LDC’s dataset compared to any purely office-generated inventory of nationwide retail units.

It is important to restate, that even after the data are collected in an orderly fashion and validated using robust methodologies, they can hardly be compared across different parts of retail space if there is no established underlying spatial framework. This notion
motivated the introduction of retail areas – the areas of the highest density of retail land use.

The methodology for their delineation is not trivial and has also evolved over time through modification based on a different range of applications. When it comes to UK retail areas, their demarcation went from being based upon a greater variety of variables and thus being targeted towards the wider town planning applications (Thurstan-Goodwin and Unwin, 2000), to making use of the densities of stores as a primary input attribute (Pavlis et al., 2018). This shift in methodology made them more granular than before and more suitable for retail planning purposes. It also, presented the possibility of going beyond town centre envelopes, as was the case in earlier delineations, and moving towards the research of smaller retail areas, dislocated from town centres in the form of suburban centres or out-of-town planned retail developments.

In this thesis, LDC’s retail unit database was used to inspect the general composition of UK’s retail areas and how it changed over the recent years. It was discovered that retail areas vary in size and form a rank-size distribution. While the relative share of different categories of retailers remained unchanged throughout recent years, some minor changes in retail occupancy on a national scale are still evident and would otherwise remain unrecognised without such an expansive and varied dataset. The vacancy rates were found to be the most significant indicator of changing retail composition, as they fluctuated much more than the presence of different retail unit categories. While being relatively resistant to structural changes of occupancy, the composition of retail areas was found to exhibit substantial spatial variability. The evidence for this was found by analysing the composition of retail areas at a national scale. Comparison stores were found to be the most prevalent top-tier category of retailing in Great Britain, closely followed by services and leisure. The larger centres are characterised by a greater share of comparison and leisure units, whereas centres placed lower in the national spatial retail hierarchy display a relatively higher share of convenience retailing compared to larger centres.

Further evidence for the spatial variability of retail composition was found in the retail area classification conducted by CDRC Liverpool (2018) and this was also reiterated in the unified retail-footfall classification of microsite locations that was conducted in Chapter 7 of this thesis. This will be reflected upon, in more detail, later in this chapter.

8.2 The Functioning of Retail Areas

Apart from purely focusing on structure, this thesis made important progress in our understanding of the dynamics of retail areas. Several findings are discussed here: the advantages and disadvantages of deploying Wi-Fi sensors to measure passing footfall in front of retail units, the footfall temporal classification, the impact of different variables on footfall and
the unified retail-footfall classification.

8.2.1 Wi-Fi Technology and its Role in Footfall Measurements

The Wi-Fi sensor data presented a significant opportunity for the improvement of our understanding of how retail areas function. Footfall counts estimated using this technology display patterns generally corresponding to reality and enable the detection of spatio-temporal distributions of footfall, including the diurnal and weekly periodicities in function, as well as anomalies in the form of rare events.

However, when this relatively new type of Big Data is used to measure footfall, the measurements can significantly depart from the ground-truth, thus requiring a complex and robust validation methodology.

The factors contributing to the overall error of footfall estimation can be divided into two groups: those leading to overcounting (e.g. the presence of recurring smartphone and non-smartphone devices and smartphone devices in passing traffic) and those leading to undercounting (e.g. the presence of specific physical obstacles which distort the signal propagation or the proportion of people who turn off the Wi-Fi on their phone). Overcounting factors may be more easily tackled through the internal validation, i.e. data mining methods. This does remove a good proportion of error, but regardless of the data cleaning quality, further improvements of accuracy are necessary to account for any residual error which always exists. This is conducted by addressing the sources of error that lead to undercounting, and this is, in turn, achieved by visiting the individual sensor microsite locations and manually collecting the data used for validation purposes.

The magnitude of the residual error that remains after internal validation, may differ significantly from location to location. This is, in turn, caused by the fact that every microsite location has specific store layout characteristics, different positioning of the device inside the store, different thickness of the surrounding walls, as well as the varying degree of the presence of devices that may distract the signal propagation (Lassabe et al., 2005). At the time of the conducting of this research, issues described were minimised by taking into account multiple measurements, ideally spread throughout the day whenever possible, to capture the possible variations in measurement error.

Following that, an average adjustment factor may be calculated and applied to all processed counts to bring the counts as close as possible to the ground-truth (Lugomer et al., 2017). Note that this does not resolve the potentially high errors on fine temporal scales such as five or fifteen-minute intervals, however, it does significantly improve the accuracy of the data on coarser temporal scales and its validity for the overall assessment of footfall at different locations.

As technology undergoes rapid development, so does the methodology for processing the footfall counts. Even over the course of the three years during which this project took
place, some significant challenges and solutions have occurred. For example, MAC address
randomisation algorithms introduced by the leading smartphone producers made accurate
estimates of footfall much more difficult, especially from September 2015 and then even
more so from September 2017. In the meantime, Vanhoef et al. (2016) found that peri-
odic randomisation does not completely prevent us from identifying an individual, as probe
requests contain some other information that can be used for identifying devices, such as
Information Elements (IEs) fingerprints which can be combined with frame sequence num-
bers. My colleagues at UCL Soundararaj et al. (2019) have recently improved the internal
validation methodology originally developed by LDC, by clustering sequence numbers and
filtering out noise. Unfortunately, since this occurred in the late phase of the research, it
was not applied to the counts used in the empirical chapters presented earlier. However,
this is only a confirmation of how dynamic technological fields of science can be and how
there is further potential of utilising improved counts in future extensions of this research.

It can be stated that Wi-Fi technology has advantages and disadvantages when it comes
to urban and retail geography research. It is superior to Bluetooth technology when it comes
to passive data collection for footfall estimation purposes, as it appears to be more accurate
and more widely used. It also allows for much greater spatial and temporal resolution of
data acquisition compared to the employment of Call Detail Records. On the other hand, it
does require a set of fixed sensors that are prone to periodic and irregular failures and can
therefore only be effectively used around a set of pre-determined locations. This is not the
case with GPS, which is capable of capturing user location without any additional devices
on the ground apart from the phones themselves, and therefore represents a worthy rival
from a technological standpoint.

As a final note on the technological perspectives, even though the top smartphone
producers made Wi-Fi analytics substantially more difficult in order to preserve user pri-
vacy, workarounds were found that still enable reasonably accurate footfall estimates. It
is presumed that Wi-Fi probe requests will remain a useful and reasonably accurate proxy
of footfall and will see continued development in the academic literature and commercial
think tanks. After all, despite all the challenges and, at the time, seemingly insurmount-
able obstacles, our knowledge about these techniques and quality of applications have only
improved over time.

8.2.2 Footfall Temporal Classifications

How do retail areas function? This question was tackled by extracting the diurnal footfall
patterns that can be detected in the UK’s retail areas. The cluster analysis conducted on
605 microsite locations across GB retail areas demonstrated that footfall temporal profiles
are a function of location (Lugomer and Longley, 2018). In total, eight different temporal
signatures were identified when Mondays through Thursdays and Fridays were examined,
whereas three different temporal signatures could be discerned for both Saturdays and Sundays.

Inspired by Longley and Singleton (2016), London can be treated as a distinct entity due to its higher complexity of demographics, morphology and function. This treatment resulted in a special, London Footfall Temporal Classification, in which only six profiles emerged. Its most striking difference to the original nationwide classification is the absence of simple, one-peaked profiles. This suggests local prevalence of multifunctional locations, in which there is a mixture of commute and lunchtime activity; or commute, lunchtime activity and evening leisure, each with different relative heights of peaks, depending on the nature of individual microsite locations.

In a similar recent study, Mumford et al. (2017) used camera-based Springboard’s\(^1\) footfall data to extract different footfall profiles based on their monthly, weekly and diurnal distributions for 144 retail centres. When it comes to diurnal distributions, they did not consider intraweek differences which were found in this thesis to be substantial and their results incorporated only two one-peaked clusters with the only difference being volume of footfall during nighttime. In addition, any breakdown of the locations of counters by retail area type were missing in the cited report, so it is not completely clear whether the sample of counters represented only shopping centres or if high streets and/or retail parks were included as well.

That being said, the main advantages of this study, compared to similar studies of footfall temporal distributions in UK’s retail areas, includes much greater control over the sample design in the sensor rollout phase, wide geographical coverage, well-documented validation methodology and a focus on not only the top-tier levels of possible signatures such as identifications on one, two or three peaked-profiles, but also more subtle differences that emerge across space. Findings of this sort are relevant for retail planning, especially in staffing and supply planning before picking a new location, as well as after the opening of the new store (D’Silva et al., 2018).

8.2.3 Footfall and its Non-Retail Drivers

In order to gain a better understanding of what type of profile occurs at what types of locations, further linkages of function with composition and the assessment of spatial variability of footfall patterns was conducted.

It was found that retail areas are not functionally entirely homogeneous and it is not possible to generalise their footfall patterns. The footfall was found to vary based on the level of road network hierarchy, in that higher hierarchical level streets record higher footfall and multifunctional three-peaked temporal signatures. The number of bus stops around sensor

\(^1\)Springboard is a retail analytics company that provides location-based retail property and downtown performance insights (Springboard, n/d).
microsite locations, as well as the proximity to the nearest one were found to be independent of footfall volume, which is presumably due to their omnipresence in UK’s town centres. At the same time, when local workplace demographics is linked to the Monday – Thursday temporal profiles, a statistically significant association can be discerned.

However, it was thought at that point that local retail geography may be much more useful in explaining the local footfall patterns than workplace demographics classification such as COWZ. This is because the most useful bits that stem from COWZ, and that can be used to interpret footfall profiles are associated with local retail composition. For example, Big City Life COWZ group is characterised by a higher proportion of workers in accommodation and food industries (Cockings et al., 2015). However, there is a question whether we need COWZ, and its more detailed description of a local workforce, or if a simple note of categories of retail units from LDC’s retail unit database that exist around the corresponding microsite location, would be equal or even more helpful in explaining local footfall patterns. This is an important point to make, because even though COWZ is an open source classification and can be used by anyone, and even though it is a classification of rather small areas, there is a much greater degree of flexibility when using even more granular, microsite level data, such as LDC’s retail unit dataset.

The bottom line, of this subsection, is that footfall varies spatially, both in terms of volume and footfall patterns. It varies not only between retail areas, but also within retail areas at particularly fine scales. The findings on the drivers of footfall variability spark an interesting notion of differential suitability of retail areas for the measurement of composition on one hand and functioning on the other hand. The retail areas are suitable for the collection of data on retail structure. However, the functioning should be examined from the microsite perspective, as each retail area can be thought of as multifunctional or as a collection of functionally diverse microsite locations.

8.2.4 Linking Retail Composition and Function

One of the classic problems in geography is linking the spatial attributes at the micro and macro scales. Following the creation of footfall temporal classification, there is a question of how function, as represented by footfall at a microsite location, can be linked to composition, as represented by the proportions of different categories of retail and leisure units within a given retail area. Five different linkage models were suggested in Chapter 7, after which their relative advantages and disadvantages were laid out. The two main approaches were eventually adopted and each of them utilised a different model of linkage between composition and function.

The first one included making the footfall temporal classification exogenous to the existing retail area classification created by CDRC Liverpool (2018). This was done by a point-in-polygon operation in which each sensor microsite location, i.e. its temporal profile,
was assigned a class of its local retail area. It was found that three-peaked footfall profiles more commonly occur in retail areas whose composition is predominantly oriented towards convenience and food retailers. Amongst them, premium comparison and leisure-oriented areas are more tightly associated with those three-peaked profiles that have a higher concentration of evening activity. On the other hand, other comparison destinations mostly have a simple one-peaked profile, as well as less affluent residential population of their catchment areas and higher vacancy rates.

In the second approach, rather than using the existing classifications, a completely new, unified retail-footfall classification was created by using individual variables representing distribution of footfall over time and proportions of relevant categories of retail and leisure units. This was done using a different linkage model, where a network distance-based buffer was drawn around each sensor microsite location to define their respective retail environments. The classification was conducted for the whole of Great Britain and resulted in five distinct clusters. They range from multifunctional retail environments characterised by a prominence of leisure units and the nighttime economy to the low-footfall environments with single-peaked profiles, an absence of entertainment and leisure categories and a prominence of low-profile retailers.

One of the merits of creating a unified classification is the emerged capability of embedding the short-term functional dynamics into the composition, which results from medium and long-term forces shaping retail environments. The addition of footfall data to the retail composition dataset drew a more complete picture of the types of retail areas that exist in the UK’s retail landscapes and added a lot of insight into the assessment of town centre vitality.

### 8.3 The Case For a New Activity-Based Retail Geography

By now, it has become well established that different types of retail areas or retail environments of the microsite locations attract footfall at different times of the day and week.

The sum of the functions of distinct retail microsite locations in each retail area explains when a retail area is crowded with people, many of which are potential consumers. Even more importantly, it tells us when activity in the corresponding retail area dies out. This invites rethinking notions of the vitality of retail areas as not only defined by their composition, vacancies, openings, closures and local demographics, but also by their activity-based function. It can be argued that a retail area with higher and more consistent flow of consumers throughout the day, and especially, throughout the whole week is characterised by more favourable local economic conditions than a retail area which is deserted during the weekend (e.g. London’s financial district – the City of London).

Everything mentioned hitherto implies that there exist strong foundations for rethinking
traditional retail geography and reconceptualising it into what could be called the New Retail Geography. The New Retail Geography would not be oriented exclusively towards form, but also function, which inevitably honours the significance of micro scale and time dimension. While this could have only been a retail geographer’s dream even two decades ago, the advent of groundbreaking technological inventions made retail data analytics much more scalable. The finding that retail areas are functionally heterogeneous and that function might vary from street to street or even along the same street suggests that any locational decisions need to be guided by a sound understanding of the forces acting on the micro, as well as macro, scale. In the event that they have not adopted or purchased footfall data from third parties, retailers and town centre managers might want to resort to prediction based on the type of street or local workplace zone or local retail geography, by assembling the findings gathered in this thesis. Whilst creation of a statistical model that accurately predicts footfall patterns forms a basis for future extension of this work, some advances have already been made in the meantime (D’Silva et al., 2018), and their models may be strengthened by taking into account additional factors considered here.

Building upon the notion of the aggregate influence of micro and macro, in both spatial and temporal terms, retail areas should still be used for the collection of data on composition and research focus in retail geography should still be directed towards them. This is because the growing availability of data at a micro level does not refute the fact that those focal points of human activity are still concentrated in particular parts of urban areas, or even suburban areas when it comes to out-of-town developments. The most profound difference compared to the earlier research is the addition of the functional component, which is generally measured and verified on the micro scale. Therefore, future research in retail geography is expected to tackle scale dichotomy that has always been evident, but has now become manageable and, hence, researchable.

However, it is worth saying that applications of the microsite-driven analysis of retail area functioning go beyond pure academic and commercial applications and assist governmental decisions at various levels, from national to local, by informing them about where and what can be done to improve the vibrancy and vitality of different town centres. According to Portas (2011), high streets had much more trouble responding to economic hardship compared to out-of-town shopping centres, as they evolved without a single ownership, with managers having varying levels of responsibility and knowledge about consumers and the industry. Therefore, high streets traditionally lack vision of what they should look like and where they should be going in the future (Portas, 2011). This supports the efforts from the academic community and various stakeholders to regenerate the high streets and bring the consumers back to the high street. One of the prerequisites would be to accurately describe, quantify, analyse and monitor the state of the high street, and some of the findings of this thesis may be useful in that respect.
The popularity of different types of composite indicators in consumer research further motivates the potential creation of different indices of composition and function that can be used for the improvement of government statistics. Some of the simple examples of such indicators (refer to the index of relative weekend activity in Chapter 6) were introduced in this thesis and can be found in the literature (Murcio et al., 2018). Apart from this, such indicators and research findings can be infused into the well-established reviews of the state and trends in UK retailing (Grimsey Review, 2018) and projects that seek to identify the most crucial drivers of the performance of high streets (see, for example, the UK High Street 2020 project) (Parker et al., 2017). As stated at the beginning of the section, knowing when the area is vibrant and when the retailers cannot expect to benefit from the inflow of consumers is instrumental in determining the retail site suitability. Furthermore, the footfall classification can help the local government in identifying those retail areas which are deserted during the particular hours, for example, workday evenings. The unified retail-footfall classification and retail unit data alone can be used to determine the retail composition of an area and then inferences can be drawn about whether some types of units, such as premium restaurants or entertainment centres are significantly underrepresented and whether the encouragement of their openings may catalyse the town centre regeneration. The advantage of using such datasets and classifications lies in the fact that they can be periodically updated and this can help identify the changes, positive or negative, that occur as an effect of the adoption of new policies.

When it comes to ensuring the sustainable future for the high streets, there are some other more specific considerations that need to be given closer attention. The local government may be interested in using the findings of this research to build a revised methodology of determining the widely criticised and controversial business rates. Helen Dickinson, chief executive of the BRC said in 2019:

“For many retailers, business rates remain the single biggest tax imposed by the government. They are a levy on physical space that is paid in full regardless of whether a firm is in profit or in loss. Importantly, they are also borne disproportionately by retailers who represent 5% of the economy yet pay 2% of all business rates. If the government is serious about reversing the decline on our high streets, then reforming the broken business rates system would be an essential first step.” (Fish, 2019)

It is clear that the business tax system is unfair, and it is far from being market-friendly. While this is out of the scope of this thesis and it is not directly associated with footfall, the highest priority should be to change the definition of this tax, so that no retailers who registered a loss in a given year are liable to pay it. After the tax becomes restricted to profit makers, the footfall data could be used to classify retail areas and, ideally, micro-parts
of retail areas based on the volume of footfall. The higher business rates could be imposed
on the businesses that are situated on the main thoroughfares, whereas areas of substantial
decline and poor retail health may even be freed from this tax or assigned to the lower tax
band to increase their propensity to attract new investments. The unified retail-footfall
classification detected the retail microsite locations whose neighbourhood is characterised
by the above-average presence of low-quality retailers such as pawnbrokers, and the absence
of footfall magnets, such as pubs and restaurants (cluster Midday Retail Environments
with Limited Leisure Facilities). Knowing their whereabouts and characteristics aids tax
rate assignment.

Note that these are only initial suggestions, but separate and in-depth research should
be conducted to confirm their viability. For instance, a problem with this approach is that
stores may sit in the busy locations, but still attract too few customers and book too small
profits. It may seem fairer to apply a single tax rate to all retail businesses, exclusively based
on the profit they make, and completely irrespective of any other external factors, such as
the location, land prices, etc. A simpler tax system is always preferred by the businesses,
and whenever possible, its unnecessary components should be removed or simplified.

Next, when speaking about the future of high street, an important role is played by the
property landlords, who impact whether their leaseholders will succeed or struggle and fail.
The landlords may use either raw footfall data or the classifications of activity patterns and
the unified retail-footfall classification to determine or adjust their rental prices. This is true
for both the owners of residential and commercial property. Rental values are dynamic, as
they change along with the changes in the overall attractiveness, aesthetics and accessibility
of the neighbourhood and the microsite location. The frequently updated retail databases
and continuously updated footfall data such as those owned by LDC enable taking into
account frequent changes in local retail geographies and the activity patterns that can be
used to update the rental values. Without the knowledge of the temporal distribution of
footfall and the retail composition of the neighbourhood, the rent may end up being too
different from what the market expects. That being said, this research can be particularly
useful for determining the rental prices on the micro-scale. For instance, even though rental
levels are known to be notoriously high in Central London, there is an opportunity to
distinguish further between the microsite locations with higher and lower potential in the
generally upmarket neighbourhood. Similarly, declining locations are also not uniform, and
within them, one can detect the microsite locations offering better trading opportunities.
Landlords could benefit from this fact by negotiating higher prices for the better microsite
locations in otherwise poor retail environments. On the other hand, coffee shop retailers
may negotiate better rental prices if the corresponding unit does not belong to the Morning
Coffee and Evening dining retail-footfall cluster or any of the footfall profiles from the
footfall classification in which the activity is concentrated in the mornings.
Finally, it is worth stating that it is only with the combined effort from all of the involved stakeholders that town centres will become more vital and viable, thus attracting more people willing to spend more time and money there. This, in turn, initiates a positive feedback loop that motivates retailers to improve the quality of their products and services and which further increases the structural and functional strength of the UK’s high street.

8.4 Possible Future Extensions of This Work

While this thesis breaks ground in multiple different fields and contributes significantly to our understanding of the complex spectrum of forces that dictate the structural and functional vitality of retail areas, much can be done in the future to expand on the presented findings. Some of the suggestions were already touched upon directly or indirectly in this or previous chapters.

First, as stated above, at the time of writing the empirical chapters of this thesis, there was still room for improving the accuracy of footfall counts by amending the internal validation methodology. A lot of progress has been made in the meantime (Soundararaj et al., 2019) and this means that any research on long-term footfall patterns using Wi-Fi probe requests has become more reliable and feasible now. The improved comparability of footfall counts at markedly different points in time (such as different seasons or years), that are free from the adverse impact of MAC address randomisation issues and that do not violate user privacy, enables us to say a lot more about the changing nature of UK’s high streets.

Second, one of the primary commercial motivations behind the whole SmartStreetSensor project was enabling the calculation of conversion rates. There has been demand from various retailers to find out how many people walk past their stores and to compare those figures with the actual number of consumers walking in and making purchases. The LDC’s Wi-Fi sensors have already been utilised for that particular commercial purpose, and there is already ongoing academic work conducted at UCL which investigates patterns and drivers of sales and conversion.

Third, a sample design of sensors that was devised before this project started ensured wide geographical coverage and its findings are thought to be representative of the national high streets and, to a smaller extent, shopping centres. However, future research could capture a bigger sample of shopping centres and put a particularly greater focus on retail parks, as they were largely under-represented. The fact that sensor samples were biased towards town centres can be justified by the increased interest of the academic community, as well as local authorities, in attracting people back to the high street, following the growing competition that came from the out-of-town retail developments at the end of the century. That being said, while understanding how town centres work is thought to be of
key importance given the wider context of trends in the retail sector, there is an opportunity for investigating the extent to which retail parks are special in terms of function.

Last, but not least, another possible extension of this research could be using the LDC retail unit database, along with other ancillary publicly available datasets, to infer characteristics of areas in which there are no sensors. The SmartStreetSensor footfall dataset is proprietary, and deployment of new sensors requires not only the allocation of financial resources but also new deals with retailers who would be potentially interested in paying for the service and allowing sensors to be installed in their stores. However, the wealth of already available data and findings from this thesis are expected to significantly assist in the creation of footfall prediction models.

8.5 Closing Statements

Before this research started, very little was known about either Wi-Fi footfall analytics, the functional differences amongst the UK’s retail areas or the ways in which the wider area retail composition is linked to function at the micro scale.

While all of that was pretty much uncharted territory three years ago, we now have a more solid picture of the composition and functioning of retail areas. A lot of that advance is owed to the utilisation of the LDC’s retail unit database and the SmartStreetSensor footfall database. In the era of Open Data, the academic community tends to neglect the significance of proprietary data gathered by the individual corporations, when compared to the popular datasets available from local or national authorities such as the census. Proprietary consumer datasets may not be available to everyone, especially not without incurring a cost, but it has become more apparent over the years that they have shown immense utility and that they add value to our understanding of retail areas that would have otherwise remained under the radar.

Probably the most important output of this thesis, is the notion that function manifested in the temporal activity patterns should be a key component, next to the retail composition, when planning retail areas and individual store locations. Not only that, but the spatial micro scale should be studied in far more detail than has been the case so far.

It is acknowledged that the functioning of microsite locations has been traditionally overlooked, due to technological and data limitations, but this is a matter of the past. Nowadays, retailers and local authorities can, and should, carefully consider the role that the dimension of time plays in shaping retail environments on the whole spectrum of scales, from micro to macro. After all, time is, indeed, money.
## Appendixes

Table 9.1: The number of installed sensors in Local Authorities and Council Areas in Great Britain (23 October 2017)

<table>
<thead>
<tr>
<th>Local Authority / Council Area</th>
<th>Number of sensors</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Westminster</td>
<td>72</td>
<td>11.90</td>
</tr>
<tr>
<td>Camden</td>
<td>39</td>
<td>6.45</td>
</tr>
<tr>
<td>City of Edinburgh</td>
<td>33</td>
<td>5.45</td>
</tr>
<tr>
<td>Brighton and Hove</td>
<td>22</td>
<td>3.64</td>
</tr>
<tr>
<td>Nottingham</td>
<td>17</td>
<td>2.81</td>
</tr>
<tr>
<td>Glasgow City</td>
<td>17</td>
<td>2.81</td>
</tr>
<tr>
<td>Bromley</td>
<td>16</td>
<td>2.64</td>
</tr>
<tr>
<td>Manchester</td>
<td>15</td>
<td>2.48</td>
</tr>
<tr>
<td>Liverpool</td>
<td>15</td>
<td>2.48</td>
</tr>
<tr>
<td>Aberdeen City</td>
<td>15</td>
<td>2.48</td>
</tr>
<tr>
<td>Leeds</td>
<td>14</td>
<td>2.31</td>
</tr>
<tr>
<td>Southwark</td>
<td>14</td>
<td>2.31</td>
</tr>
<tr>
<td>Reading</td>
<td>13</td>
<td>2.15</td>
</tr>
<tr>
<td>Wakefield</td>
<td>13</td>
<td>2.15</td>
</tr>
<tr>
<td>Cheshire West and Chester</td>
<td>12</td>
<td>1.98</td>
</tr>
<tr>
<td>Hammersmith and Fulham</td>
<td>12</td>
<td>1.98</td>
</tr>
<tr>
<td>Kensington and Chelsea</td>
<td>12</td>
<td>1.98</td>
</tr>
<tr>
<td>Bristol, City of</td>
<td>11</td>
<td>1.82</td>
</tr>
<tr>
<td>Harborough</td>
<td>11</td>
<td>1.82</td>
</tr>
<tr>
<td>Norwich</td>
<td>11</td>
<td>1.82</td>
</tr>
<tr>
<td>Cardiff</td>
<td>11</td>
<td>1.82</td>
</tr>
<tr>
<td>Location</td>
<td>Value</td>
<td>Score</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Wiltshire</td>
<td>10</td>
<td>1.65</td>
</tr>
<tr>
<td>Kingston upon Thames</td>
<td>10</td>
<td>1.65</td>
</tr>
<tr>
<td>Southampton</td>
<td>9</td>
<td>1.49</td>
</tr>
<tr>
<td>Cambridge</td>
<td>8</td>
<td>1.32</td>
</tr>
<tr>
<td>Oxford</td>
<td>8</td>
<td>1.32</td>
</tr>
<tr>
<td>Sheffield</td>
<td>8</td>
<td>1.32</td>
</tr>
<tr>
<td>Birmingham</td>
<td>8</td>
<td>1.32</td>
</tr>
<tr>
<td>Lambeth</td>
<td>8</td>
<td>1.32</td>
</tr>
<tr>
<td>Islington</td>
<td>7</td>
<td>1.16</td>
</tr>
<tr>
<td>Newcastle upon Tyne</td>
<td>6</td>
<td>0.99</td>
</tr>
<tr>
<td>Haringey</td>
<td>6</td>
<td>0.99</td>
</tr>
<tr>
<td>Blackpool</td>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>Plymouth</td>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>County Durham</td>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>Gloucester</td>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>Taunton Deane</td>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>Croydon</td>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>Aberdeenshire</td>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>York</td>
<td>4</td>
<td>0.66</td>
</tr>
<tr>
<td>Leicester</td>
<td>4</td>
<td>0.66</td>
</tr>
<tr>
<td>New Forest</td>
<td>4</td>
<td>0.66</td>
</tr>
<tr>
<td>Warwick</td>
<td>4</td>
<td>0.66</td>
</tr>
<tr>
<td>City of London</td>
<td>4</td>
<td>0.66</td>
</tr>
<tr>
<td>Hounslow</td>
<td>4</td>
<td>0.66</td>
</tr>
<tr>
<td>Kingston upon Hull, City of</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>Windsor and Maidenhead</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>West Lindsey</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>Bradford</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>Ealing</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>Stirling</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>West Dorset</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>East Hertfordshire</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>Bolton</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>Trafford</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>Coventry</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>Solihull</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>Gateshead</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>Location</td>
<td>Count</td>
<td>Value</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Waltham Forest</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>Wandsworth</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>Warrington</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Southend-on-Sea</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Cheshire East</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Chelmsford</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Basingstoke and Deane</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Three Rivers</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Watford</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Dartford</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Dover</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Boston</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>East Staffordshire</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Spelthorne</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Bury</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Oldham</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Sandwell</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Walsall</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Barnet</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Bexley</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Brent</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Harrow</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Havering</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Lewisham</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Merton</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Newham</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Redbridge</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Richmond upon Thames</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>East Renfrewshire</td>
<td>1</td>
<td>0.17</td>
</tr>
</tbody>
</table>
Table 9.2: Fieldwork measurements and calibration for the sensor locations in London and Sheffield

<table>
<thead>
<tr>
<th>Sensor Location</th>
<th>Calibration date</th>
<th>Calibration start time</th>
<th>Calibration end time</th>
<th>Manual count</th>
<th>Processed count</th>
<th>Measurement error</th>
<th>Adjusted count</th>
<th>Final error</th>
<th>Absolute Improvement (in percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>12/12/2016</td>
<td>07:45</td>
<td>08:00</td>
<td>133</td>
<td>93</td>
<td>5.38</td>
<td>125</td>
<td>-6.51</td>
<td>-1.13</td>
</tr>
<tr>
<td></td>
<td>12/12/2016</td>
<td>09:15</td>
<td>09:30</td>
<td>421</td>
<td>394</td>
<td>-6.41</td>
<td>526</td>
<td>24.98</td>
<td>-18.57</td>
</tr>
<tr>
<td></td>
<td>12/12/2016</td>
<td>10:15</td>
<td>10:30</td>
<td>321</td>
<td>227</td>
<td>-29.28</td>
<td>303</td>
<td>-5.60</td>
<td>23.68</td>
</tr>
<tr>
<td></td>
<td>12/12/2016</td>
<td>13:00</td>
<td>13:15</td>
<td>506</td>
<td>301</td>
<td>-10.80</td>
<td>401</td>
<td>-20.67</td>
<td>-9.87</td>
</tr>
<tr>
<td></td>
<td>12/12/2016</td>
<td>14:45</td>
<td>15:00</td>
<td>381</td>
<td>286</td>
<td>-24.93</td>
<td>382</td>
<td>0.27</td>
<td>24.67</td>
</tr>
<tr>
<td></td>
<td>12/12/2016</td>
<td>17:45</td>
<td>18:00</td>
<td>486</td>
<td>443</td>
<td>-8.85</td>
<td>591</td>
<td>21.64</td>
<td>-12.79</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>07/09/2016</td>
<td>12:50</td>
<td>13:10</td>
<td>619</td>
<td>191</td>
<td>-69.14</td>
<td>586</td>
<td>-5.34</td>
<td>63.80</td>
</tr>
<tr>
<td></td>
<td>07/09/2016</td>
<td>14:30</td>
<td>14:50</td>
<td>441</td>
<td>103</td>
<td>-76.64</td>
<td>316</td>
<td>-28.35</td>
<td>48.29</td>
</tr>
<tr>
<td></td>
<td>07/09/2016</td>
<td>17:30</td>
<td>17:50</td>
<td>254</td>
<td>102</td>
<td>-59.84</td>
<td>313</td>
<td>23.19</td>
<td>36.65</td>
</tr>
<tr>
<td></td>
<td>07/09/2016</td>
<td>19:30</td>
<td>19:50</td>
<td>159</td>
<td>26</td>
<td>-83.65</td>
<td>80</td>
<td>-49.84</td>
<td>33.81</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>08/09/2016</td>
<td>08:40</td>
<td>09:00</td>
<td>137</td>
<td>75</td>
<td>-45.26</td>
<td>230</td>
<td>67.94</td>
<td>-22.68</td>
</tr>
<tr>
<td></td>
<td>08/09/2016</td>
<td>10:30</td>
<td>10:50</td>
<td>284</td>
<td>97</td>
<td>-65.85</td>
<td>298</td>
<td>4.78</td>
<td>61.07</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>08/09/2016</td>
<td>12:50</td>
<td>13:10</td>
<td>618</td>
<td>200</td>
<td>-67.64</td>
<td>614</td>
<td>-0.72</td>
<td>66.92</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>08/09/2016</td>
<td>14:30</td>
<td>14:50</td>
<td>462</td>
<td>121</td>
<td>-73.81</td>
<td>371</td>
<td>-19.66</td>
<td>54.15</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>08/09/2016</td>
<td>17:30</td>
<td>17:50</td>
<td>309</td>
<td>98</td>
<td>-68.28</td>
<td>301</td>
<td>-2.71</td>
<td>65.58</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>09/09/2016</td>
<td>08:40</td>
<td>09:00</td>
<td>155</td>
<td>72</td>
<td>-53.55</td>
<td>221</td>
<td>42.50</td>
<td>11.05</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>09/09/2016</td>
<td>10:30</td>
<td>10:50</td>
<td>324</td>
<td>88</td>
<td>-72.84</td>
<td>270</td>
<td>-16.68</td>
<td>56.16</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>09/09/2016</td>
<td>12:50</td>
<td>13:10</td>
<td>596</td>
<td>180</td>
<td>-69.80</td>
<td>552</td>
<td>-7.35</td>
<td>62.45</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>09/09/2016</td>
<td>14:30</td>
<td>14:50</td>
<td>438</td>
<td>108</td>
<td>-75.34</td>
<td>331</td>
<td>-24.36</td>
<td>50.98</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>09/09/2016</td>
<td>17:30</td>
<td>17:50</td>
<td>294</td>
<td>148</td>
<td>-49.66</td>
<td>454</td>
<td>54.43</td>
<td>-4.77</td>
</tr>
<tr>
<td>Sheffield B</td>
<td>09/09/2016</td>
<td>19:30</td>
<td>19:50</td>
<td>246</td>
<td>66</td>
<td>-73.17</td>
<td>202</td>
<td>17.70</td>
<td>55.47</td>
</tr>
</tbody>
</table>
Table 9.3: Distribution of sensors across different types of retail centres (CDRC Liverpool, 2018)

<table>
<thead>
<tr>
<th>Group</th>
<th>Subgroup</th>
<th>Total number of centres in GB</th>
<th>Number of sensors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Retail &amp; Service Centres</td>
<td>Local Upmarket</td>
<td>328</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Local Urban Convenience Centres</td>
<td>535</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Inner Urban Service Centres</td>
<td>329</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1192</td>
<td>2</td>
</tr>
<tr>
<td>Retail, Shopping &amp; Leisure Parks</td>
<td>Primary Parks</td>
<td>89</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Secondary Parks</td>
<td>172</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>261</td>
<td>1</td>
</tr>
<tr>
<td>Leading Comparison &amp; Leisure Destinations</td>
<td>Premium Shopping &amp; Leisure Destinations</td>
<td>217</td>
<td>372</td>
</tr>
<tr>
<td></td>
<td>Mass &amp; Value Destinations</td>
<td>193</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Premium Out of Town Destinations Areas</td>
<td>48</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>458</td>
<td>429</td>
</tr>
<tr>
<td>Primary Food and Secondary Comparison Destinations</td>
<td>Vibrant Urban Destinations</td>
<td>179</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Diverse District Destinations</td>
<td>243</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Urban Value Destinations</td>
<td>137</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>559</td>
<td>52</td>
</tr>
<tr>
<td>Traditional and Diverse High Streets</td>
<td>Traditional High Streets of Rural Britain</td>
<td>106</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Suburban &amp; Market Town High Streets</td>
<td>210</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Diverse &amp; Affluent Urban High Streets</td>
<td>165</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>640</td>
<td>5</td>
</tr>
</tbody>
</table>

237
| Indie & Value Oriented High Streets | 159 | 1 |
Table 9.4: Test of the statistical significance of associations between the COWZ supergroups and temporal profiles

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymptotic Sig. (2-sided)</th>
<th>Monte Carlo Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>308.198</td>
<td>42</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>334.545</td>
<td>42</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>Fisher’s Exact Test</td>
<td>315.043</td>
<td></td>
<td>.000</td>
<td>0</td>
</tr>
</tbody>
</table>

N of Valid Cases 605

35 cells (62.5%) have expected count less than 5.
The minimum expected count is .10.

Table 9.5: Test of the statistical significance of associations between the COWZ groups (lower-tier level of the classification) and temporal profiles

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymptotic Sig. (2-sided)</th>
<th>Monte Carlo Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>647.838</td>
<td>161</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>592.347</td>
<td>161</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>Fisher’s Exact Test</td>
<td>538.550</td>
<td></td>
<td>.000</td>
<td>0</td>
</tr>
</tbody>
</table>

N of Valid Cases 605

157 cells (81.8%) have expected count less than 5.
The minimum expected count is .02.
Table 9.6: The test of the statistical significance of associations between the LWZC supergroups and temporal profiles

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymptotic Sig. (2-sided)</th>
<th>Monte Carlo Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>95% Confidence Interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>116.800</td>
<td>15</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>102.957</td>
<td>15</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>Fisher’s Exact Test</td>
<td>94.003</td>
<td></td>
<td></td>
<td>.000</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>226</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11 cells (45.7%) have expected count less than 5.
The minimum expected count is .68.

Table 9.7: The test of the statistical significance of associations between the LWZC groups (lower-tier level of the classification) and temporal profiles

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymptotic Sig. (2-sided)</th>
<th>Monte Carlo Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>95% Confidence Interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>143.997</td>
<td>40</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>134.632</td>
<td>40</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>Fisher’s Exact Test</td>
<td>120.476</td>
<td></td>
<td></td>
<td>.000</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>226</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

40 cells (74.1%) have expected count less than 5.
The minimum expected count is .12.
Table 9.8: The breakdown of the footfall temporal clusters based on the retail area format

<table>
<thead>
<tr>
<th>Temporal cluster</th>
<th>Shopping centres</th>
<th>Town centres</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Percentage (%)</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>15.22</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>10.87</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>47.83</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>8.7</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>15.22</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2.17</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>100</td>
</tr>
</tbody>
</table>
Figure 9.1: Sample design for sensor locations: England

author: L. Dolega
Figure 9.2: Sample design for sensor locations: London

*authors:* J. Kandt, P. Longley
Bibliography


Findlay, A. and Sparks, L. (2009). Literature review: policies adopted to support a healthy retail sector and retail led regeneration and the impact of retail on the regeneration of town centres and local high streets.


253


Lugomer, K. et al. (2017). Understanding sources of measurement error in the Wi-Fi sensor data in the Smart City.


Ordnance Survey (2017). “OS VectorMap District”.

256


R Documentation (n/a). MAD function. URL: https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/mad (visited on 18th Feb. 2018).


— (2001). The Dynamics of Property Location: Value and the factors which drive the location of shops, offices and other land uses. Taylor & Francis.


Springboard (n/d). URL: https://www.spring-board.info/about-us (visited on 22nd May 2018).


