UNIVERSITY OF LEEDS

Generating a Leeds specific open geodemographic classification

Amanda Leigh Otley

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Faculty of Earth and Environment

School of Geography

January 2021

Acknowledgements

This thesis has come to be with the kind support and assistance of many individuals. I would like to extend my sincere thanks to each and every one of them.

In particular, I would like to express my deep gratitude to my supervisors Michelle Morris, Andy Newing and Mark Birkin for their advice, support and enthusiasm throughout. Andy and Michelle's positivity, kind words and encouragement have kept my spirits high, and their knowledge, meticulous proof-reading, time and attention have been critical in helping me bring this thesis to fruition. Mark's expertise in geodemographics, insights, experience and practical advice has also been a great source of support. I would also like to acknowledge the advice and input of my Research Support Group members Lex Comber and Roger Beecham, both of whom have provided helpful and constructive feedback throughout this process.

This work has been funded through an Economic and Social Research Council (ESRC) White Rose DTP studentship and has been undertaken in partnership with Leeds City Council (LCC) and TransUnion. Huge thanks are due to Ruth Addison for her continued support, assistance and attention throughout. Many individuals at LCC have also worked hard to source and provide data and to share their knowledge with me, I would like to extend my gratitude to each, in particular, Qamran Hussain and Rachel Antcliffe. I would also like to acknowledge Malachi Rangecroft, Andy Peloe, and Jon Gill who were instrumental in helping establish the groundwork for the thesis. In addition, I must also acknowledge the numerous people who have been on the end of one (or several) of my many questions and who have been kind enough to offer insights and direction which have helped this work develop.

I would also like express thanks to all of the friends who have helped me both inside and outside of the University. Thanks in particular are due to Charlotte Sturley, Rachel Oldroyd, Fran Pontin and Vicki Jenneson for the invaluable tea breaks and the Shut Up and Write marathons, and to Rachel Hunter-Brookes and to Alex Bradley for keeping me distracted, and for keeping me going, when each were needed. Thank you is also due to Verity Tether, who I have been extremely lucky to have met and shared this experience with, and without whom the past four years would not have been the same. Her endless support and enthusiasm will always be remembered.

I am also indebted to my family, particularly my husband, Danny, my parents, my sister and Georgie-Leigh, for their limitless support and encouragement in everything that I do. I am deeply grateful for their belief in me which has encouraged me to believe in myself and the many ways in which they have provided practical and emotional support all along the way. This simply would not have been possible without them. I would finally like to acknowledge my grandma, who has always been there to celebrate our achievements and who I know will be with us when we can get together to celebrate this one.

Abstract

Stretched by increasing demand and decreasing budgets, like many local authorities, Leeds City Council have turned to geodemographics to support data-led decision making. As per the current trend for transparent research and policy development, the literature increasingly recommends open geodemographics for use in the public sector. However, the only open classification currently available, the 2011 OAC, which is derived at national level from decennial census data collected in 2011, has proven ineffective at identifying some of the unique multivariate local phenomena.

This thesis generates a new framework for a public sector focused place-specific geodemographic classification for Leeds. Primarily, the study introduces and explores the impact of making a methodological shift in geographic extent from national to local level. Secondarily, the research extends beyond traditional decennial census input data to include novel data from open and public sector sources. To support this extension, the work also investigates the potential of several Feature Extraction and Feature Selection methods to intelligently reduce the set of candidate input attributes by identifying those most capable of generating meaningful classification outputs to suit public sector requirements.

This thesis demonstrates that there is both scope for generating locally specific classifications with novel administrative data, and benefits to be gained, particularly in terms of identifying locally specific phenomena capable of enriching public policy and decision making processes. It also makes a strong argument for an increased emphasis on incorporating intelligent variable selection processes into geodemographic classification development.

The work has been completed during an ESRC collaborative PhD studentship in partnership with LCC and TransUnion. All developments made have primarily considered the needs of a public sector end-user, however, the outputs are transferrable and applicable beyond the public sector. Moreover, transparency and reproducibility has been prioritised to enable and support replications in other cities with similarly available data.

Contents

List of Figures

List of Tables

List of Abbreviations

ADR	Administrative Data Research UK
BCSS	Between Cluster Sum of Squares
BMRB	British Market Research Bureau
CDRC	Consumer Data Research Centre
COWZ	Classification of Workplace Zones
CRN	Classification of Residential Neighbourhoods
ED	Euclidean Distance
EF	Ecological Fallacy
EFA	Exploratory Factor Analysis
EPC	Energy Performance Certificates
ESRC	Economic and Social Research Council
FA	Factor Analysis
FE	Feature Extraction
FS	Feature Selection
GIS	Geographical Information Science
HMO	Houses in Multiple Occupation
IHS	Inverse Hyperbolic Sine
KMO	Kaiser-Meyer-Olkin
LA	Local Authority
LCC	Leeds City Council
LIDA	Leeds Institute for Data Analytics
LLTI	Limiting Long-Term Illness
LOAC	London Output Area Classification
LSOAC	Leeds Specific Output Area Classification
LVM	Latent Variable Models
MAUP	Modifiable Areal Unit Problem
ML	Machine Learning
NAP	National Action Plan 2019-2021
OA	Output Area
OAC	Output Area Classification
ONS	Office for National Statistics
PCA	Principal Component Analysis
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
SAA	Social Area Analysis
SDC	Statistical Disclosure Controls
SED	Squared Euclidean Distance
SIR	Standardised Illness Ratio
SMC	Squared Multiple Correlation
UPRN	Unique Property Reference Number
WCSS	Within Cluster Sum of Squares

Chapter 1 - Introduction

1.1 Research outline and context

1.1.1 A landscape for change

The primary aim of this thesis is to develop a process for building a locally specific geodemographic classification for the city of Leeds, in collaboration with Leeds City Council (LCC) and TransUnion.

Leeds is a diverse and multi-cultural city in the North of England with an increasing population of over 790,000 residents. LCC is the Local Authority (LA) for the city. In addition to being a forward-thinking and proactive LA (as demonstrated particularly in Chapter 5), LCC prioritises collaborative engagement with academia to develop responses to to social, environmental and economic challenges facing the city (Carroll and Crawford, 2020). TransUnion is a global information and insights company who develop and licence one of the most popular commercial geodemographic classification products, “CAMEO”, both in the UK and around the globe (see Section 3.2).

LCC desire a clearer, more holistic understanding of the resident population in the city, for a number of purposes, including informing decision making processes and policy development. Like many LAs, they have been seeking data-led strategies to achieve this objective and have turned to geodemographics to help provide some answers. Geodemographics offers a framework for assigning one of a set of classification labels to each pre-defined small-area geographies, often focused on residential neighbourhoods. The process offers an aggregate description of the individuals and households residing within each area, providing a useful metric for urban analysis which is applicable in a number of situations across the triad of academia and the public and private sectors (Harris et al., 2005).

Though the development of geodemographic classifications were once the exclusive realm of academics, following an adoption of methodologies in the commercial sector, focus and advancement in this field has almost completely shifted into the commercial Geodemographics industry, leading to a lengthy absence of academic interest and a stagnancy in progress within the academic environment (Singleton and Longley, 2009b). Nevertheless, scholarly research into geodemographic classification development has experienced somewhat of a comeback in recent decades, introducing a new, contemporary research agenda (Longley, 2005; Brunsdon and Singleton, 2015).

The development of the Output Area Classification (OAC) in 2001 (Vickers and Rees, 2007) (superseded by the 2011 OAC (Gale et al., 2016)), a freely available classification derived from data from the decennial censuses of England and Wales, Scotland and Northern Ireland, which was published alongside extensive supporting documentation detailing its development, marked a step-change in geodemographic classification openness (Singleton and Spielman, 2014) and offered an alternative to the commercial classifications which are licensable for a fee and which are built in a commercially sensitive black-boxed environment.

Despite broad academic literature citing the advantages of such openness, particularly signalling an opportunity for the public sector to benefit from the trust offered by such transparency (Brunsdon et al., 2018; Gale et al., 2016), the geodemographic classification market continues to be dominated by the proprietary commercial products which purport to deliver a superior outcome. This is true within the public sector, and LCC are no exception. Their adoption of geodemographics to achieve the objectives outlined at the outset of this section have been underpinned for many years by the use of a licensed commercial classification. However, there is a renewed impetus to re-consider the use of more open geodemographics within LCC, detailed below.

1.1.2 Public Sector requirements (specifically within LCC)

The emerging discussion regarding the continued use of the commercial classifications within LCC are threefold. Primarily, end-users within LCC who work closely with the classification in their decision making processes have raised concerns regarding discrepancies between the descriptive profiles assigned to many areas of the city and their own expert and local knowledge of these areas. There is a perception amongst these individuals that the available classifications are failing to capture nuances within the city and phenomena present in the population that is unique to Leeds. Thus, LCC are keen to identify a solution which resolves such a disparity.

Second, against a backdrop of public sector workers continuing to find innovative, regularly data-led, solutions to budget restraints on resources, LCC are considering the expense of licensing geodemographic classifications, and balancing this expense against the potential of the increasing volumes of rich population data routinely collected and stored internally. Restrictions in data accessibility, interest and investment, which has for the past five centuries seen geodemographic development fall almost exclusively in the realm of commercial scope (detailed in Chapter 2), are now being lifted by increasing public sector access to interesting and granular data. These considerations are leading LCC to wonder whether there might be scope for developing their own bespoke classifications, taking advantage of both the data and the expertise and knowledge held within the LCC itself, and in doing so, potentially also addressing the primary concern listed above.

Finally, there is a trend in the wider public sector in the UK, as data-led strategies have become more commonplace, to endeavour to adopt transparent processes to promote public trust (discussed in detail in Chapter 5). Though this notion has been particularly acute and has received more publicity throughout 2020 as the media and public alike have called for greater transparency of the data and thought processes which have underpinned many of the key decision processes throughout the COVID-19 pandemic, the promotion of transparency had been gaining in momentum for many years already. Consequently, criticisms have been raised regarding the use of metrics developed in black-boxed environments underpinning public sector decision making (see Section 2.4.3), presenting yet further encouragement for LCC to seek an alternative to the incumbent classifications used.

Until now, practical efforts to actually build an improved classification for use by the public sector, within LCC or in academia, has been limited. The motivations listed have yet to lead to practical solutions and there has been limited progression towards the generation of a viable alternative capable of challenging the persistent commercial dominance. Internally within LCC, at least, this might in part be attributed to a lack, or at least a perceived lack, of resources or expertise. Alternatively, it could in part be due to increased time pressures and decreased budgets, which have stretched the services of LAs over recent years. The limited practical progress in this direction which has been witnessed in academia will be considered in more detail in Chapter 2.

Nevertheless, the work contained in this thesis is seeking to address this. It will outline how the development of geodemographic classifications is prime for a new phase of development with these specific requirements of the public sector in mind.

1.1.3 Project origins

Concerns similar to those of LCC (outlined in Section 1.1.2) are echoed throughout the academic literature. These have been well documented since the turn of the century in several books and across a breadth of journal articles, each offering a critical commentary on the development of the current geodemographic landscape and recommending a range of potential avenues for improvement to the traditional practices. Notably, LCC's primary concern regarding a potential lack of local representation offered by the available classifications has been particularly highlighted as an issue. Specifically, several academics have suggested that the national extent, at which classifications have been traditionally derived, could contribute to the masking of local nuance, and as such, have recommended an exploration of the development of more local, place-specific classifications. In response, Singleton and Longley (2015) developed a London specific classification, the London Output Area Classification (LOAC), the success of which further propagated the discussion and introduced a potential solution to the concerns of LCC.

Moreover, public sector concerns, such as those of LCC, have not gone unnoticed within the commercial Geodemographics industry, who identify the public sector as an important market for their products. As an initial response, several commercial vendors have developed broad public sector, or public sector domain specific classifications targeted towards the specific needs of the domain, including CACI, Experian and TransUnion. However, these products continue to be developed within black-boxed environments and at a national extent. TransUnion recognise the additional work which is still required and are keen to seek further solutions to enable them to develop geodemographic classifications which meet the needs of these consumers.

To achieve these common goals, partnership between academia, industry and the public sector is encouraged (Longley, 2005). This thesis has been completed during an ESRC collaborative PhD studentship in partnership with LCC and TransUnion. Both LCC and TransUnion recognise the importance of collaboration and the potential of such an approach in this context. LCC increasingly prioritise collaborative partnership, particularly with local universities, evidenced in the 188 successful collaborative research projects which they have undertaken in partnership with the University of Leeds since January 2015, of which this work is one, valued at over £38m associated funding (Carroll and Crawford, 2020). Similarly, the commercial sector in the city is increasingly aware of the benefits available through partnerships with academia to achieve their own research objectives, demonstrated through the strong industry partnerships generated and maintained within the Leeds Institute for Data Analytics (LIDA) and the Consumer Data Research Centre (CDRC), where this work has been hosted. LIDA has offered a trusted middle ground for the work conducted within this thesis, facilitating a rare partnership across academia, the public sector (LCC) and industry (TransUnion), who have each identified the potential for working together to achieve common and individual goals benefitting each party and generating a wider public benefit.

This collaboration has introduced an opportunity to re-evaluate the perceived weaknesses and challenges of developing geodemographic classifications from a unique perspective underpinned by a rare combination of resources, both tangible, in terms of sharing previously siloed data, and intangible, in terms of a broad range of experience, expertise and knowledge. Moreover, this PhD format affords the much needed time to explore and seek outcomes which are independent of commercial expectations and are not constrained by the threat of reallocation of interest or funding, which is often present in the public sector.

1.2 Research aims and objectives

This research seeks to enrich the open geodemographic landscape in the UK by evolving beyond the existing practices adopted in the development of geodemographic classifications and improve the incumbent framework to increase their suitability and relevance, particularly with a focus on their use in the public sector. Though these considerations are made primarily with a focus on the city of Leeds, and generating improvements to benefit LCC, the work is to act as a case study which can be adapted to other geographical contexts.

The overall aims of this research are primarily:

- to present a review of the historic and contemporary practices involved in the development and public sector specific use of open and commercial geodemographic classifications in the UK.

- to develop and test new approaches for improving the incumbent standard framework, particularly focusing on a shift to developing place specific classifications, extending beyond the inclusion of solely census variables, particularly in the inclusion of other administrative data, and exploring more sophisticated approaches to input variable selection.

- to present LCC with an updated framework for generating more meaningful and relevant open geodemographics in the context of their use in developing public sector policy and service and resource allocation, offering an alternative to the currently favoured commercial classifications.
- to present TransUnion with a review and critique of novel methodologies for geodemographic classification development.

In order to meet these aims, this thesis addresses the following research objectives:

- To summarise the contemporary landscape of geodemographics in the context of their origins, precursors and historical development (Chapter 2).
- To present an overview of the standard framework traditionally adopted when developing a geodemographic classification, exemplified through a discussion of the 2011 OAC methodology (Chapter 3).
- To review the weaknesses of contemporary practices for developing geodemographic classifications as discussed in the literature, and the challenges which have underpinned recent stagnation in academic progress (Chapter 3).
- To review the literature documenting the weaknesses of using commercial classifications in public sector (Chapter 3).
- To further explore the benefits of a methodological shift in the geographic extent of geodemographic classifications from the traditional national level to a more local approach through a re-classification of the 2011 OAC exclusively for Leeds (executed first in Chapter 4 and maintained in further analysis executed in Chapter 5, Chapter 6, Chapter 7 and Chapter 8).
- To examine local and central government data infrastructure, specifically within LCC, alongside open data repositories to highlight and identify relevant data sources at the required geographical extent for inclusion in the development of future local geodemographic classifications (Chapter 5).
- To offer insight into the potential strengths and weaknesses of the data sources identified in the previous objective and present recommendations for implementing any

necessary improvements to better support such action in the future (Chapter 5).

- To investigate the practical scope for increased adoption of identified administrative datasets as input variables in the development of local geodemographic classifications, specifically exploring internal and open datasets available to LCC (Chapter 5).
- To practically evaluate the potential improvements to be gained from extending open geodemographics to include administrative and open data from sources beyond the traditional data source of the decennial census through a re-classification of the 2011 OAC (Chapter 6).
- To consider the necessity for improved variable selection techniques in the geodemographic classification development process, as identified in the literature and through an evaluation of the traditional approaches employed (Chapter 7).
- To develop a framework for using unsupervised machine learning techniques, namely Factor Analysis (FA), as a method of variable selection in the context of local geodemographic development (Chapter 7).
- To test the practical impact of employing the Factor Analysis (FA) framework (developed in the previous objective) as a variable selection technique in the development of a local geodemographic classification (Chapter 7).
- To consider the scope for introducing supervised Machine Learning (ML) techniques, namely Feature Selection (FS) techniques, as a method of variable selection in the context of local geodemographic development (Chapter 8).
- To test the practical impact of employing the Feature Selection (FS) techniques explored in the previous objective as a variable selection technique in the development of a local geodemographic classification (Chapter 8).
- To comment on the applicability of the approaches explored and tested throughout this thesis within LCC and the potential impact on local level decision making and policy development (Chapter 9).

These objectives are addressed systematically throughout this thesis, the structure of which is outlined in Section 1.3.

1.3 Thesis structure and scope

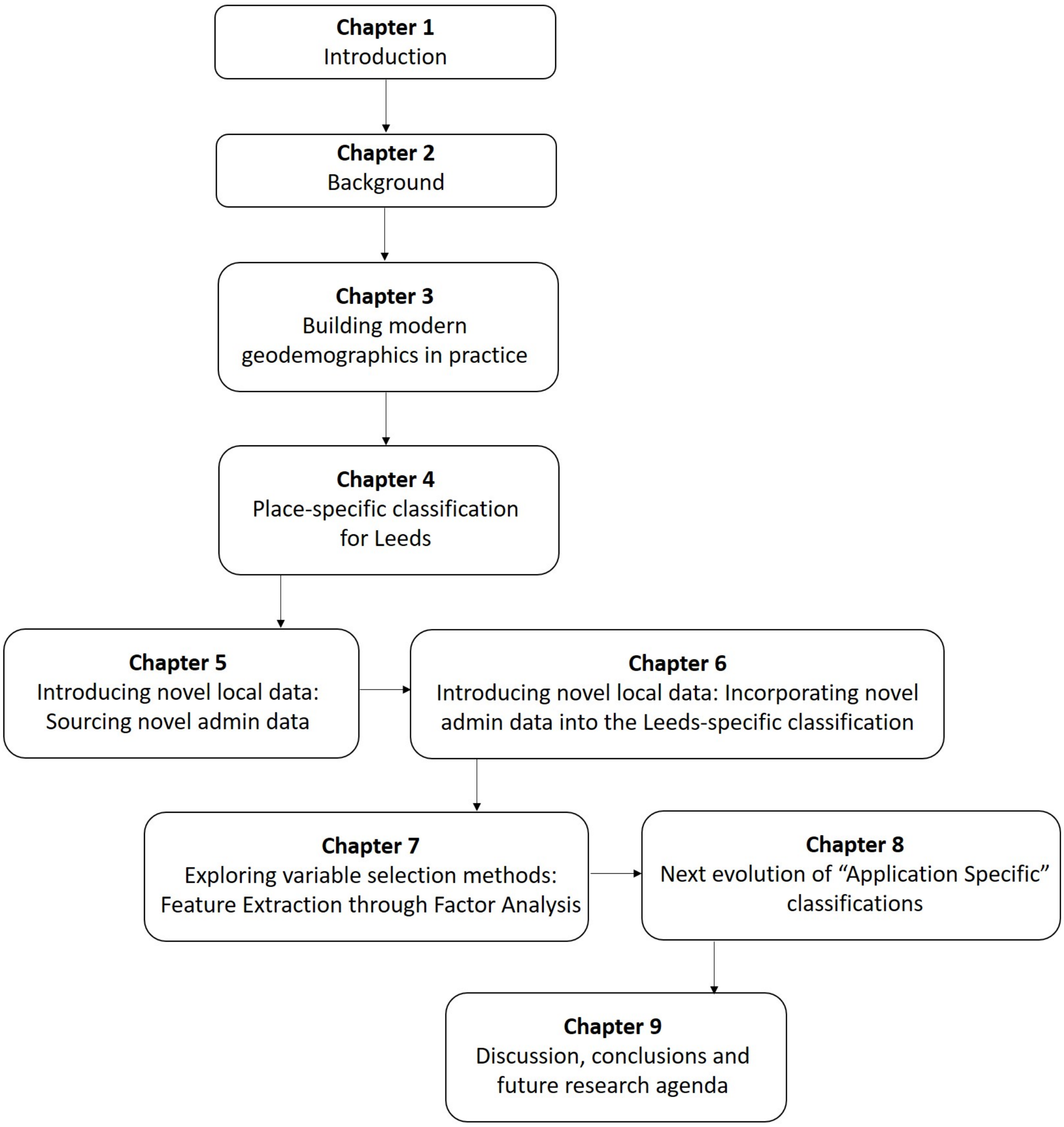

Chapter 1
Introduction
Chapter 2
Background
Chapter 3
Building modern geodemographics in practice
Chapter 4
Place-specific classification for Leeds
Chapter 5
Introducing novel local data: Sourcing novel admin data
Chapter 6
Introducing novel local data: Incorporating novel admin data into the Leeds-specific classification
Chapter 7
Exploring variable selection methods: Feature Extraction through Factor Analysis
Chapter 8
Next evolution of "Application Specific" classifications
Chapter 9
Discussion, conclusions and future research agenda

Figure 1.1: Thesis flow.

In meeting the objectives discussed in Section 1.2, the structure of this thesis is illustrated in Figure 1.1. First, Chapters 2 and 3 establish the study within the existing literature. Chapter 2 predominantly focuses on the origins and chronology of geodemographic classification development which have given rise to the contemporary landscape. The discussion particularly highlights the relevance of the historical progress and current status from a public sector perspective, documenting the full-circle taken from the roots of geodemographic precursors which focused largely on deriving an understanding of urban structure, through the decades of commercial dominance, and back to a recent resurgence of interest in local government. In doing so, this chapter situates the research in the context of developing modern geodemographic classifications for contemporary policy development and public sector decision making.

Chapter 3 situates the objectives of thesis more specifically within the established practical framework commonly adopted for developing contemporary classifications. This is exemplified through a documentation of the approach underpinning the development of the most widely used open geodemographic classification in the UK, the 2011 OAC. In this context, the strengths and weaknesses of the established practices are reviewed, as presented in the literature, alongside a discussion of the perceived challenges which have acted as barriers to further progress to date. From this discussion, several opportunities for improvement to the process of developing geodemographic classifications are identified as priorities, establishing the research agenda for the subsequent chapters in the thesis.

The primary development priority of this thesis, understanding the potential for achieving a superior classification output by shifting the scale of the classification from a national to a local extent to produce a classification for a single city, in this case Leeds, is investigated in Chapter 4. Extending the research of Singleton and Longley (2015) in their development of the LOAC, this chapter similarly generates a city-specific classification for Leeds (named LSOAC), deriving a new profile for each Output Area (OA) within the city and comparing the result to the established 2011 OAC profiles. This work presents a baseline for subsequent exploration of techniques to enhance elements of the framework adopted in the OAC, LOAC and within this chapter to seek to derive an improved Leeds specific classification development in the subsequent chapters.

The second priority identified involves the extension of the input variables underpinning the LSOAC generated in Chapter 4 with variables derived from sources beyond the decennial censuses. Chapter 5 presents a discussion of the literature supporting an adoption of administrative data in data analysis, and explores the scope for its use in this context, particularly considering the benefits of including relevant local attributes, based on a combination of LCC expert knowledge and guidance from the literature, which might lead to a more meaningful output. A case study which focuses on extending the variables included in the 2011 OAC to represent housing demographics practically demonstrates the work required to identify, gather and prepare relevant datasets from disparate sources for inclusion in a re-classification of the LSOAC, the conclusion of which is presented in Chapter 6 in comparison with the original LSOAC derived in Chapter 4. Novel datasets are sourced internally, from within LCC, and from open data repositories. Challenges encountered in the process of this research are documented across both chapters, and each include recommendations to facilitate improvements to data infrastructure and storage, collection, documentation and sharing procedures to support the implementation of similar processes as standard within LCC in the future.

As Chapter 5 and Chapter 6 introduce the potential for extending the input variable set in developing a new open geodemographic classification to any available relevant data at the appropriate geographic scale, a question arises around the definition and judgement of "relevant". Whilst the tradition has been to rely on the expert domain knowledge of the developer in the selection of input variables, Chapter 7 seeks to explore more statistical based techniques. This chapter looks back to unsupervised variable selection approaches widely considered, reviewed and tested in the research literature relating to the pre-cursors of geodemographics (identified in Chapter 2), namely the feature extraction method of Factor Analysis (FA), and addresses the weaknesses of the methodology which saw it subsequently fall out of favour in this context. In response, this chapter presents a clear framework with evidence based justifications to support the implementation of FA in the selection of variables for a Leeds specific re-classification which otherwise adopts the 2011 OAC methodology. As before, the classification is compared to the LSOAC classification derived in Chapter 4 to highlight the impact of such an updated approach.

Chapter 8 extends the investigations commenced in Chapter 7, in this case, seeking improvements to the variable selection procedure for deriving a geodemographic classification by considering the potential for introducing further meaning into the outcome by instead employing supervised machine learning techniques. Such an approach emerges from a recent trend in the academic literature for scepticism towards the utility of general purpose applications and to recommend, instead, a shift towards the development of domain specific alternatives, such as have become increasingly commonplace in the suite of commercial offerings. A review of the literature underpins a discussion of this trend and the changing practices with relation to developing bespoke, domain specific classifications in academic research to underpin ever more targeted public policy decision making. The research in

Chapter 8 extends this existing domain specific literature, which continues to rely on expert judgement within the variable selection process, by proposing a shift to developing application specific classifications capable of further increasing the discriminatory power of the resulting classification with respect to a given outcome of interest. Supervised machine learning methods are trained on this outcome and key variables driving the outcome are identified. These are subsequently used in another re-classification of the Leeds OAs, also based on the 2011 OAC methodology. The resulting classification is evaluated to present the potential of the method to offer a more discriminatory classification on the basis of the outcome, with which to support increasingly targeted decision making. The framework presented demonstrates a practical example supporting further development of application specific geodemographic classifications.

Chapter 9 presents an argument for local authorities such as LCC to consider substituting their dependence on proprietary commercial geodemographic classifications with the practice of developing bespoke city and application specific classifications based on the findings of this thesis. Chapter 9 also notes the contribution that this thesis has made to the academic literature in practically addressing some of the challenges which have, until now, largely been addressed theoretically.

In meeting the aims and objectives outlined in Section 1.2, this thesis approaches the research from the perspective of offering improvements to the standard framework for deriving geodemographic classifications to generate more meaningful and relevant outputs, particularly in the context of public sector application. Such an approach represents the strengths, interests and experiences of the author, academic supervisors and the research cluster within which this work was hosted. Whilst the thesis draws from the literature discussing the subsequent application of geodemographic classifications in both a public sector context and otherwise, the primary focus remains on the consideration and generation of improvements to the development and not the use of classifications, particularly in generating a local-specific classification for the city of Leeds for primary use by LCC (although some practical considerations in terms of their use or application are briefly considered in the later chapters). In doing so, this thesis contributes to a clear gap in the academic literature and seeks to offer insight relevant more broadly in the development of classifications for use in the public sector and otherwise, in Leeds and beyond.

1.4 Thesis contribution and potential impact

The outcomes of this thesis highlight some of the benefits that can be achieved through interdisciplinary research and close collaboration between academic institutions and the public and private sector. The results add to the limited practical exploration of the development of place-specific geodemographic classifications which incorporate novel locally specific data, particularly for use by the public sector. The findings have the potential to support local governments in developing their own bespoke, tailored city-specific geodemographic classifications. This has the potential to offer both a more meaningful and relevant output and a level of transparency and reproducibility in the development process which is necessary in the public sector, welcoming scrutiny and providing a greater level of trust in future results derived from their subsequent application. To date, there is no research in the UK jointly investigating these issues, supporting the need for the research presented in this thesis.

This project is thus well positioned to generate impact at a variety of levels. The research will be used to support specific case studies which themselves will generate specific impact within their application areas and with reference to Leeds which will have potential future wider application. The research outcomes present a considerable societal benefit in providing updates to the existing framework which are capable of deriving more meaningful classifications, with which to inform public sector decision making, providing an opportunity to change lives through improved targeting and evaluation of services, policy and interventions. Moreover, this work will assist LCC in uncovering the wider value of their data which could have broader societal benefits, including in guiding or supporting future census taking.

Furthermore, academic benefits are also available. The methodological considerations could both re-ignite and extend the debates regarding appropriate variable selection techniques, first with regards to Feature Extraction techniques such as Factor Analysis, which has appeared in several guises throughout the history of geodemographic classification development, but further, with regards to more novel Feature Selection techniques. This could herald a new era of academic innovation around locally-specific classifications, opening the door for future academic work, particularly considering the development of local classifications purposefully designed around specific applications, or the inclusion of ever more novel data.

Additionally, the outcomes offer commercially valuable insights regarding the potential benefits of place-specific geodemographic classifications and the inclusion of a wider source of administrative data, some of which is presently only available within local government. In turn, this could pave the way for a new generation of commercially exploitable city specific geodemographic classifications. Finally, this work will strengthen the links between the University of Leeds, LCC and Transunion, contribute to Leeds' role as a hub for data analytics, and act as a model for similar work elsewhere.

Chapter 2 - Background: There and back again

2.1 Introduction

The enduring desire to measure, model and describe populations is as prominent today than at any time before. The ability to do so is particularly key to public sector success. Geodemographic classifications offer a popular multivariate approach for understanding urban population structures and are widely adopted in public sector applications, and far beyond. However, the suitability of contemporary geodemographics in this context is being increasingly brought into question. The aim of this thesis is to therefore investigate the potential for a next phase of geodemographics which explicitly supports the unique requirements of the public sector, specifically LCC, and to suggest the direction such an evolution might take.

In achieving this target, it is necessary to first look back at the journey which geodemographic development has navigated thus far, and to take stock of the situation as it stands today. In so doing, it would be impossible, and also ill-advised, to consider the creation and specific applications of public sector developed geodemographics in a silo, separate from the shared past, present, and potential future alignments with advances in academia, and in the commercial Geodemographics Industry. The importance of this is evidenced in the cross-sector partnerships which support this research (described in Chapter 1). As such, this chapter presents an overview of the history, origins and contemporary landscape of geodemographics, which will necessarily span the triad of academia and the public and private sectors.

All of the literature here focuses primarily on the discussion of the *development* and *generation* of classifications, both theoretically and practically. This is not presented as a thorough review of the *applications* of geodemographic classifications, although there will be references to the applications in order to support the development of the most suitable classification, explicitly considering the practical needs and requirements of the public sector, and specifically LCC. This chapter is also not intended to act as a comprehensive history of geodemographic development, instead, its aim is to situate this thesis in the context of the established practices and the necessary future improvements identified.

Moreover, the literature considered here, and throughout this thesis, will focus on methodological developments made in the UK and the US, where the majority of the English language publications in the field also focus their attention. Geodemographic classification development in the two nations has been historically intertwined, supporting consideration of both, although there are many distinct features associated with each (Singleton and Spielman, 2014). As such, activity in the UK will be the primary focus, and any considerations of activity in the US will be largely limited to supporting this discussion, particularly in a review of historical developments.

Section 2.2 introduces key definitions and the theoretical concepts which underpin geodemographics. Section 2.3 summarises its historical evolution to the present day. Section 2.4 presents the contemporary geodemographic landscape in the context of the broader political and technological environment and concludes by reviewing the weaknesses of the contemporary practices and introducing alternative approaches which will be explored in more detail later in Chapters 4-8.

2.2 The theory of Geodemographics

The composition of urban societies can have a direct effect on the growth and evolution of cities. It therefore follows that an ability to identify and describe societal compositions within urban environments is crucial in effectively modelling, predicting and managing such an evolution. Naturally, this potential is of significant interest to the public sector who could benefit greatly from such an ability, but the interest is also shared by both the private sector and academia alike, each of whom have their own desires to better understand populations. These desires have stimulated a broad and enduring body of multidisciplinary research seeking to understand and describe urban community structures dating back over a century (Park et al., 1925; Longley, 2012).

The resulting research has advanced the understanding of residential population structures and sparked new fields of academic study (discussed in Section 2.3). The legacy of many of these contributions is embodied in the field of geodemographic analysis, the eponymous study of people by where they live, which has been established as the current standard for modelling resident populations in a social context based on demographic indicators, and the related commercial industry of Geodemographics which has since emerged, and which drives much of the requirement for continued research today. The theoretical constructs and the practical methods of geodemographics have been developed concurrently, each supporting the evolution and development of the other (Batey and Brown, 1995).

In terms of the theoretical foundations, human identity is defined by the characteristics which distinguish an individual or group (Longley, 2012). It is the desire to define and measure these distinguishing features and to unlock the group identities present within a population which have driven the research which both pre-dates and still continues in Geodemographic development today. Specifically, much of this research has been, and remains, grounded in the fundamental assumption that individuals sharing traits and characteristics tend to reside in close geographical proximity to one another, and vice versa. This theory extends *Tobler's First Law of Geography* (Harris et al., 2005, p. 16) which states:

> *"Everything is related to everything else, but near things are more related than distant things."* (Tobler, 1970)

Such a notion supports the idea that character attributes associated with individuals can be transferred to others residing within the same vicinity (Vickers and Rees, 2007). However, it is necessary to consider more than a single attribute to capture the complex social dimensions which define the holistic character of an area, as such, a multivariate approach is required (Beaumont and Inglis, 1989). Geodemographics offers a practical embodiment of this notion, seeking to detect such geographically defined social identity by identifying the distinguishing characteristics of the resident populations within small-area geographies to conveniently summarise the complexity of human populations.

In practice, this takes the form of a well-established and largely standardised framework, which receives population attributes and characteristics as inputs, and as an output, assigns summary labels to the geographical areas of study based on a grouping of areas which share prominent characteristics (see Chapter 3 for a detailed discussion of the framework). This process presents a tangible method for modelling complex urban systems whilst offering a workable foundation for deriving additional insights, with the resulting groupings facilitating a broader understanding of underlying socio-spatial patterns and offering a lens through which to identify structures within the population (Parker et al., 2007; Vickers and Rees, 2007).

The output labels generated as a result are commonly referred to as *geodemographic classifications*, with the methodological framework for deriving the classifications referred to as *geodemographic systems*, and their application, *geodemographic analysis*. Often the same terminology is used interchangeably to synonymously refer to each element, this is typically simply the umbrella term of just *geodemographics* (Blakemore and Masser, 1991).

Traditional, incumbent methodologies derive standard geodemographic classifications at a national extent. In the UK, this means assigning all small-area geographies within, however these are defined, to one of a finite set of nationally derived labels. As such, the small-area geographies are ordered into distinct and exhaustive category groups, typically through scientific clustering methods evaluating prominent shared area characteristics and attributes (Alexiou, 2017; Singleton and Longley, 2015). The resulting classifications offer a single, surrogate social measure for each small-area geography modelled in the form of a *classifier*. This classification grouping can subsequently be used to facilitate a broader understanding of underlying socio-spatial patterns and providing an increased awareness of the population, enabling a more efficient identification of actionable insights (Parker et al., 2007; Vickers and Rees, 2007; Savage and Burrows, 2007).

This potential for using geodemographic classifications to derive insight, and the versatility of the approach, have promoted its employment in a broad range of applications across academia and the public and private sectors, alike. Despite the application of geodemographics not being the focus of this thesis, some example applications and uses of geodemographics to derive insights do receive attention in Chapter 8. The evolution of this activity, the speed and trajectory of which will be told throughout this chapter, has consequently encouraged the elevation of geodemographics to a preferred geocomputational tool amongst sociologists, geographers, urban planners, policy makers, marketers and beyond (Berry and Smith, 1971; Vickers and Rees, 2007).

Yet the desire to better understand human systems predates this present proposition, which is relatively recent in its establishment (Harris et al., 2005). In the lifespan of research into population structures, geodemographics represents just the most recent phase within a history of related concepts, each of which have shaped both its theory and practical implementations and resulted in the practice of today. These historical developments are chronicled in the next section.

2.3 Urban Analysis to Modern Geodemographics: A short history

Academic consensus ages the field of geodemographics in its current form at approximately five decades, at which time the term *geodemographics* itself was coined. However, this is just the accepted inception date of *modern* geodemographics. The roots of this modern interpretation antedate this period of development by many decades. Moreover, the practice has continued to thrive and has further evolved in the time which has since passed.

The active desire to identify and differentiate patterns within populations has encouraged iterative study since the early twentieth century under a series of different guises, with each phase of research influenced by and building upon the ideas and contributions of the last. Though not entirely analogous with any single one, the geodemographics of today is widely acknowledged as having its origins in each. The chronology of this evolution has been repeatedly well-documented throughout the decades in the supporting academic literature. On each occasion, the progression from early geodemographic precursors to the field as it exists today have been detailed and critiqued in the research, clearly outlining the contribution of each phase of research and its influence on the next. Though each review of the chronology is necessarily selective and incomplete in its account, a largely consistent timeline emerges, punctuated with the same seminal research studies and succinctly describing a sequential progression of the theoretical ideas and development of the practical processes leading to those employed today (Batey and Brown, 1995; Harris et al., 2005; Singleton and Spielman, 2014; Webber and Burrows, 2018).

A comprehensive review of this literature reveals several key themes, from which a complementary meta-narrative emerges. This meta-narrative presents the progress experienced by geodemographics as a consequence of the circumstances which enabled and encouraged each significant development from one evolution to the next. Particularly, it highlights that advancements occurred intermittently, as interest in the understanding of urban populations fell in and out of sync with changing social and academic trends. Additionally, advancements are also revealed to have occurred in parallel to increased capabilities in data availability, improvements in statistical methodologies and advancements in emerging technology. This is documented in phases throughout the literature. Intentions have likewise oscillated across this period, between satisfying purely theoretical curiosity and informing practical applications. This interest has influenced a breadth of diverse research throughout this time (Singleton and Spielman, 2014; Harris et al., 2005).

2.3.1 Origins of urban theories and early typologies

The Geodemographics of today is widely considered to have had its origins in the academic study of Social and Urban Research, which was similarly focused on deriving an awareness and understanding of social patterns and their implications (Burrows and Gane, 2006). Progress up to the birth of modern geodemographics came in waves, emerging from and occurring largely within academia, was focused on contemporary local, often city-level, problems, and was motivated by aspirations to deliver public benefit. As cities expanded and developed, new methods of understanding societal structures were required and enabled, incrementally, by the evolution in statistical abilities and computational capacity, and early data-driven methods for generating general principles about residential populations within urban societies, simplifying complex community structures into typologies, began to develop (Singleton and Spielman, 2014).

Influence of Charles Booth's descriptive map of London poverty

Most scholars today attribute the earliest precursor of modern geodemographics to a practical public health study by social reformer Charles Booth, aimed at mapping poverty in London in the late nineteenth century (Harris et al., 2005; Webber and Burrows, 2018). In the first mapping exercise of its kind (Davies, 1978b), Booth set about developing techniques to identify and represent indicators of deprivation, defining and mapping patterns in the ethnographic, social and religious makeup of households in London (Vaughan, 2018) based on data from school board home visit records (Harris et al., 2005).

This work was crucial in demonstrating a pragmatic application of urban research upon which future research could build (Davies, 1978b), and set a new foundation for understanding population structures (Alexiou, 2017; Batey and Brown, 1995). Whilst seeking to represent spatial variation in a combination of population characteristics, Booth developed and presented a new methodology for quantitatively measuring multiple social features.

Though pioneering, this study was far from the mainstream at this time. The work was privately funded (Harris et al., 2005) and relied heavily on Booth's ability to collect and handle the data required. In the context of the period in which the work was carried out, it was cutting edge. In some ways, Booth's ideas came almost *too* early. His efforts pre-dated the development of many fundamental statistical methods, which he could have made use of in his endeavour (Davies, 1978b). As a result, the methods were somewhat crude in comparison to modern demographics, however, the resulting outputs offered insights upon which spatially aware public policy could be developed, as has become the best practice of today.

Human Ecology

Booth's initial priority of simply measuring and observing, whilst important, did not seek to *understand* the basis of the structures which emerged. He did not hypothesise theoretically as to the causes of the patterns which his studies identified or link his findings to social theories (Davies, 1978b). In contrast, the next wave of progress led by sociologists Ernest Burgess and Robert Park throughout the 1920s and 1930s, was almost entirely rooted in developing an understanding of the social theory which underpinned the structure of cities, in particular, in the American city of Chicago which was undergoing major societal reconstruction at the time as a result of increasingly mobile populations (Harris et al., 2005).

In an era epitomised by mass migration and increasing segregation in the city, Burgess and Park, inspired by Booth's descriptive mapping methods (Davies, 1978b), and stimulated by the release of aggregate census statistics in the US (Batey and Brown, 1995), developed new theories and empirical methods for conceptualising and modelling the emerging and distinctive residential patterns. Lending and re-purposing established ideas from other disciplines, notably the study of Ecology, and re-applying the theoretical concepts in the context of population analysis, the two presented extensive, ground-breaking ideas of social dynamics and urban structures to explain the socio-spatial organisation of the city (Harris et al., 2005). In doing to, they developed an entirely new research domain, later to be known as *Human Ecology.*

Adopting established concepts from biological ecology, Park (1925) re-considered the urban ecosystems. He proposed that competition for land and resources led to the organic emergence of cultural subdivisions in the city, which he labelled "Natural Urban Areas". These were mappable units of communal life, in which residents shared social characteristics dictated by common social pressures (Brown, 2011). Accordingly, he generated simplified typologies from complex populations, which are recognisable in the concept of today's geodemographics (Longley, 2005). Burgess. (1925) further extended the ecological analogy, depicting this spatial differentiation of residential land use in the form of five concentric rings (Figure 2.1) graduating out from the more socially and physically deteriorated city centre, to areas of greater prosperity situated at the edge of the city, based, primarily, on attributes of employment and wealth (Brown, 2011).

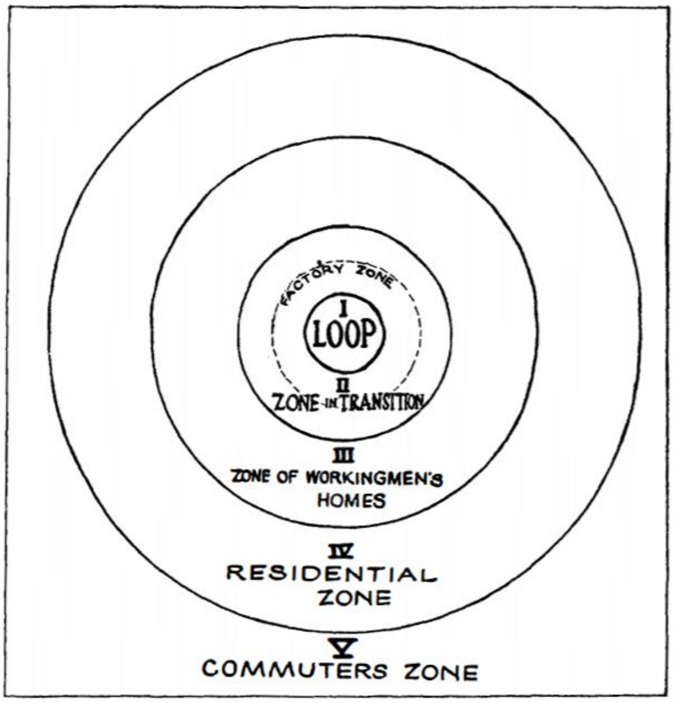


Figure 2.1: Diagram of Burgess' Concentric Zone Model (Burgess., 1925).

This seminal work not only offered the next iteration of Social and Urban Analysis but was to become the foundation of all future studies concerning city structures (Alexiou, 2017; Batey and Brown, 1995). The perspectives introduced are widely acknowledged as having inspired and paved the way for all subsequent research which has led to the field of Geodemographics today (Webber and Burrows, 2018).

Social Area Analysis to Factorial Ecology

Though the work of Park and Burgess undoubtedly progressed urban research, their efforts received as much criticism as praise. The theories supporting Burgess.'s (1925) concentric ring theory, in particular, were instantly accepted by some, and flatly dismissed by others. More nuanced evaluations largely picked holes not in the novel notion of the concentric patterns identified, but in more specific elements of the hypothesis. For instance, some challenged the circular shape proposed as a model of contemporary cities, and others, the physical legacy infrastructure which could affect such an outcome city-to-city (Quinn, 1940). One might summarise, therefore, that where balanced criticism was drawn, it largely concluded in favour of the theoretical direction of travel of Human Ecology, in considering the patterns of urban analysis, but invited improvements in developing more broadly appropriate methodologies.

Correspondingly, the next significant generation of socio-spatial research, *Social Area Analysis* (SAA) offered a new wave of methodological approaches (Singleton, 2014). Developed and evolved throughout the late 1940s and 1950s against a backdrop of increasing data availability and improving multivariate statistical methods, SAA brought an alternative approach for describing social divisions based upon a broader range of characteristics. In their seminal study, Shevky and Williams (1949) presented an initial framework for differentiating urban typologies within the city of Los Angeles based on a combination of three social constructs "Social Rank", "Urbanisation" and "Segregation". These constructs were elected as the key factors driving contemporary social patterns, and were composed of a set of related variables representing economic, family and ethnic indicators, respectively.

The resulting social area classifications were developed with less ambiguity than Park's Natural Urban Areas, affording them more applicability to other urban locations. However, criticisms were again raised, primarily regarding the empirical nature of the input variable selection criteria (Rees, 1972). Though these concerns were accepted and addressed by Shevky and Bell (1955), who retrospectively offered an alternative which involved a more statistically grounded variable selection methodology, their response was accused of simply seeking to act as a *post-facto* justification of the initial decisions made, rather than to select the appropriate variables anew (Robson and Robson, 1969).

Subsequent researchers sought themselves to develop more objective variable selection methods with which to evaluate the decisions of Shevky and associates and to underpin future social classifications. Notably the multivariate statistical methods of Factor Analysis (FA) (introduced in more detail in Chapter 7) became a popular methodology for identifying the underlying spatial structures and associated characteristics (Alexiou and Singleton, 2015). These explorations stimulated a further new field of research throughout the 1960s and 1970s, *Factorial Ecology.* Just like the development of Human Ecology and SAA before it, the emergence of Factorial Ecology was enabled by the increasing availability of census data, and relied essentially on further contemporary innovations in multivariate statistical methods which facilitated analysis of urban structure based on a greater number of variables than at any time before (Janson, 1980). Moreover, several studies were published presenting city-to-city comparisons of the approaches developed in an effort to identify a common, transferable city structure (Batey and Brown, 1995; Rees, 1972). However, many of the results instead indicated an absence of such a phenomena (Alexiou and Singleton, 2015). As such, the success of the outputs of the new field are still debated.

Nevertheless, experts agree that the conceptual innovations from which SAA and Factorial Ecology had emerged had introduced a paradigm shift from the descriptive nature of the methods of Booth, Parker and Burgess towards the adoption of complex quantitative, multivariate statistical processes (Reibel, 2011), which have remained the cornerstone of future population analysis.

2.3.2 The development of “Modern” Geodemographics

Early influences and applications of SAA in UK policy making

While generations of researchers in the US made contributions which had extended Urban Analysis far beyond Booth’s early work of mapping health inequality in London (Section 2.3.1), restrictions in data access had prohibited similar progress back in the UK, where the release of census data occurred much later (Batey and Brown, 1995). Nevertheless, when the first release of small-area aggregations of the 1951 census data eliminated these restrictions, activity and progress rapidly accelerated.

Motivated by objectives similar to Booth’s, to understand city structures to improve social outcomes, a series of research studies throughout the 1960s and 1970s saw academics in the UK draw on influences of SAA and Factorial Ecology to build upon, and push beyond, the platform that the US-based studies had constructed (Harris et al., 2005). Harris et al. (2005) highlight, several “pioneering” studies which are emblematic of the activity through this period, and each of which contributed to the wave of progress achieved. Many of the studies discussed in this review adopted the principles of Factor Analysis (FA) which had evolved through the Factorial Ecology research. However, in many cases these techniques were used simply to underpin emerging, more favourable clustering techniques. All of the studies highlighted by Harris et al. (2005) shared a common research agenda, seeking to address contemporary public sector concerns, and again, focused on deriving a taxonomy of a single city. However, each of the studies throughout this time demonstrated a contemporary shift in attention towards the analysis of increasingly smaller small-area geographies, in studies designed more purposefully around a specific application (Alexiou and Singleton, 2015).

One particular study highlighted by Harris et al. (2005) encompasses developments in the field which have been seminal in setting the background for the work contained in this thesis, and thus warrants particular attention here. This study occurred in Merseyside in the 1970s, beginning as the Liverpool Inner City Area Study, commissioned by the government as part of a wider project to identify the best way to revitalise inner city areas. Webber and Burrows (2018) describe in detail how the city's planning department contracted the Centre for Environment Studies, who took the unique approach of borrowing influences from SAA to develop taxonomies characterising the neighbourhoods in the city to be used to support policy decisions and allocation of government provisions. Though conceptually similar to its antecedents, the reviewers particularly highlight a move by the planning department to build a gazetteer of homes in the city to which the taxonomies of each home were assigned, enabling for the first time, a richer understanding of the use of services. Moreover, the methodology underpinning the generation of the taxonomies incorporated emerging techniques of cluster analysis, but foregoing FA, which was beginning to fall from favour.

In their discussion of the chosen methodologies of the Liverpool study, Webber and Burrows (2018) describe the renunciation of FA as "controversial" and contrary to previous studies at the time, however, a broader review of the Factorial Ecology literature indicates that this decision was reflective of a growing scepticism of FA in this context. Theoretical and methodological concerns were both noted (Berry and Kasarda, 1977; Alexiou and Singleton, 2015), which led to a subsequent demise in confidence in Factorial Ecology techniques (discussed more thoroughly in Chapter 7). Consequently cluster techniques quickly became the de-facto methodology. The employment of cluster analysis, which does not depend on pre-existing theories, but instead allows for the character of different areas to emerge based on a combination of the attributes, were found in the Liverpool taxonomy to generate more nuanced results than traditional approaches. These results afforded a more qualitative understanding of poverty than had previously been achieved, identifying different *types* of need, and not just different *levels*, and providing a metric which supported better, more targeted policy and decision-making (Webber and Burrows, 2018).

Based on the perceived success, the Centre for Environment Studies were subsequently commissioned to develop taxonomies in other areas, and later, to develop the first set of national taxonomies, labelled the Classification of Residential Neighbourhoods (CRN) (Webber, 1978). Where Factorial Ecology had failed to derive a taxonomy transferrable from city to city (Batey and Brown, 1995), the CRN seemingly offered just that. LAs were now presented with a choice, to use the national taxonomies or to commission their own local specific one. The CRN, which offered a framework enabling the national comparison of social need and thus better supported applications for government funding, proved the more popular choice, on balance, despite an acknowledged loss in local level detail (Webber and Burrows, 2018).

Evidently, the landscape of Urban Analysis was changed considerably by this study, and the cohort of others across this period (many more of which are considered by Harris et al. (2005)). It is also clear that the public sector was a significant beneficiary of this work. Yet immediately preceding the conclusion of the Liverpool study, a combination of simultaneously occurring events in the late 1970s was to effectively put an end the long history of enduring academic progress, and to call a halt on the focus on developing future taxonomies with the primary focus of generating social benefit.

The emergence of geodemographics and the Geodemographics Industry

In 1978, the national CRN came to the attention of the marketing industry (Webber and Burrows, 2018). Richard Webber, who had been working at the CES, presented the principles of the methodology to individuals at the British Market Research Bureau (BMRB), illustrating the transferability of the population typologies to support clear commercial benefits (Baker, 1991). A partnership between the two parties ensued, the outputs of which demonstrated the practical value of re-purposing the CRN, combined with data relating to consumer behaviour and media preferences, to generate commercially valuable insights not achievable through incumbent methodologies, and most importantly, demonstrating their potential profitability (Brown et al., 2000; Baker, 1991).

Taking this work a step further, Webber identified an opportunity to re-purpose the principles of the CRN within a commercial setting, specifically developing classifications of consumer types. Joining the London arm of the US based marketing, technology and data specialists CACI, he extended the company's existing practice of supplying retail clients with spatial census-statistics by developing and bringing to market "Acorn" (A Classification of Residential Neighbourhoods), the UK's first commercial *Geodemographic Classification*. The rich consumer profiles generated by Acorn have since instigated an innovative shift in marketing practices, informing increasingly targeted marketing activity (Harris et al., 2005; Webber and Burrows, 2018).

These developments catalysed a new era of commercial marketing strategies in the UK and breathed new life into the field of Urban Analysis, albeit, effectively ending the momentum of development in academia (Harris et al., 2005; Longley, 2012). In embracing these new practices, the commercial sector was rewarded with increased customer awareness and improved predictions of consumer behaviour, hitherto unachievable through the use of traditional univariate methodologies. Thus, this activity quickly attracted further interest and investment (Leventhal, 2016; Batey and Brown, 1995). This release was quickly succeeded by the rapid development of several competitor classifications throughout the 1980s, signalling the emergence of the lucrative *Geodemographics Industry* (Singleton and Spielman, 2014; Beaumont and Inglis, 1989), and introducing with it a set of terminology still adopted today. Similar developments were also taking place independently in the US with the development of the first US-based geodemographic classification, PRIZM (Webber and Burrows, 2018). Harris et al. (2005) present a comprehensive run through of the timeline surrounding the development and release of Acorn and competitors, including significant events and technological advancements propagating the commercial progress, and highlighting the commercial success achieved throughout these years.

In hindsight, this progression into the commercial sector was somewhat inevitable. The theoretical principles of Tobler's law (Tobler, 1970) which underpins the principles of typologies were already well established in consumer analysis, particularly relating to the influence of neighbourhoods and close associates on the consumption habits of an individual. Thus the commercial re-purposing of this analysis was a natural leap (Leventhal, 2016). Moreover, while the action taken at this time can be attributed to the foresight of a handful of individuals, including Webber, and their ability to see a commercial application for these established practices, in a practical sense, it was also seemingly prime time to facilitate this activity. Since CACI were already offering their clients access to census statistics to improve marketing strategies, their early generation of Acorn was a somewhat organic development.

Mirroring previous phases of progress, new levels of computational power and statistical capability underpinned the proliferation of the next phase of advancement. The introduction and development of clustering algorithms for identifying consumer groups become commonplace, along with the inclusion of proprietary and ad-hoc population data to support and enhance the traditional census variables. This development enabled the classification of ever more granular geographies and adding increased flexibility into the models (Gale et al., 2016; Singleton and Spielman, 2014). Consequently, the largest producers and suppliers of geodemographic classifications in the UK became those with access to data, proprietary and otherwise, such as data warehousers and data vendors, including CACI, and later Experian and TransUnion[1]. Similar advancements were more difficult to achieve in academia, which faced limitations in computing power and more restricted access to data, even in access to the census data, which was still a paid-for service in the 1980s (Harris et al., 2005).

This 'consumer takeover' also completed the departure from the incumbent city-specific classifications to establish the practice of developing classifications for a national extent as the default practice. This development reflected a shift in focus from the locally specific issues which had dominated past research, and generating public sector benefits, to commercial applications, and unlocking the perceived commercial benefits to be gained in enabling inter-city comparisons (Alexiou and Singleton, 2015). Rather than developing an understanding of the population structures in a case study area, as per the approach of Booth, Park, Burgess and Shevky and associates (Section 2.3) single classifications were now routinely developed at the level of the UK as a whole. The national-extent introduced commercially favourable economies of scale, such as had been the experience of the LAs adopting the national CRN. For the end user, this presented a single, consistent product which could be easily understood and could be applied uniformly across the country, enabling much sought after national-level comparability. Though most commercial use-cases of geodemographic classifications remain undisclosed, Harris et al. (2005, p. 19) illustrate

[1] Both Experian and TransUnion are consumer credit reporting agencies with established data infrastructure and access to vast quantities of rich, transaction data. As noted in Chapter 1, TransUnion is also a partner in this research project.

these benefits in practice through a detailed commercial case study describing one end-user application of geodemographics by a UK restaurant chain who were able to geographically tailor their pricing in the restaurant market based on spatial consumer patterns.

The endurance of geodemographics and the Geodemographics Industry

The commercial geodemographic classifications continue to dominate the market today, with a sophistication yet unparalleled by free and open public sector or academically developed alternatives (Singleton and Longley, 2009b). The investment required to develop classifications to rival, or even match, the commercial outputs, in terms of purchasing data and the technology required for the development, presented an increasing barrier which led to a dark-age in academic study and a stagnation in non-commercial development and progress (Beaumont and Inglis, 1989; Singleton and Spielman, 2014). In the UK, this has created a monopoly of classification outputs from a handful of commercial vendors (Dalton and Thatcher, 2015).

Since the field was almost entirely displaced from its former position in academia, and its focus shifted from public sector problems and studies focused on deriving public sector benefit, most subsequent advancements have been led by the commercial sector in "black-boxed" environments. Having found the means and motivation to facilitate progress (Longley, 2005), geodemographic classification was able to thrive and remain at the cutting edge of innovation in the commercial sector throughout the 1980s, establishing a widely adopted framework for developing classifications (Harris et al., 2005). However, whilst this initially provided a new landscape in which to thrive, progress was relatively short-lived and limited in its nature. As suggested by the current use of the term "modern" (in the title of this section) to describe the practice of geodemographics as established in the 1970s and developed through the 1980s, the momentum of the innovation experienced during this transformative phase was not sustained, and limited advancements have been produced in the intervening period (Singleton and Longley, 2009b). Any advancements which have been made have been evolutionary rather than revolutionary (Gale et al., 2016), particularly with relation to the methodology underpinning their development (see Section 3.5 for a specific discussion of the advancements made). Relieved of the external scientific scrutiny which had befallen previous phases of development, but which also encouraged somewhat cyclical advancement, there has been limited impetus to drastically develop upon the long-established processes which continue to underpin a profitable industry. Nevertheless, the empirical success of the commercial industry endures, evidenced by the still thriving Geodemographics industry.

However, it must be noted that there are barriers to genuinely evaluating the commercial developments, for instance, the black-boxed development environments and commercial sensitivities might prevent the publication of any significant advancements. Such an evaluation is therefore limited to the learnings which can be derived from carefully curated marketing publications (which brings with it a fresh set of concerns, see Section 2.4.3). Nevertheless, a thorough review of the websites of the major providers and academic literature suggest that the standard framework adopted to develop contemporary classifications bares a close resemblance to the national-level cluster-based framework which was established in the early period of commercial geodemographic classification development (see Chapter 3).

Whilst public sector and academic interest and investment in the development of geodemographic classifications has remained vastly reduced since the commercial takeover in the 1980s, activity did not entirely cease. Most notably, in the period since aggregate statistics from the decennial censuses in the UK became free to access, a series of entirely census-based classifications have subsequently been generated and published for open use (Charlton et al., 1985; Blake and Openshaw, 2005; Vickers and Rees, 2007; Gale et al., 2016). The most recent iteration, the 2011 Output Area Classification (OAC) was developed through academic partnership with the Office for National Statistics (ONS), and consequently, has been published as an official ONS national metric for use in public sector research and planning (Samarasundera et al., 2010). This is the most popular and widely used open and free alternative available in the UK (Singleton and Longley, 2015). Details of the development of this classification have also been published with complete transparency. This documentation reveals that, similar to its predecessor the 2001 OAC, the 2011 OAC was also developed on a national-level cluster-based framework. Chapter 3 contains an in-depth review of this classification, and more broadly, the emergent standard classification at the root of both these open classifications and the commercial classifications on the market today.

However, though advancements in the practice of developing geodemographic classification have been limited, academic contributions have continued. Through this, a rich, critical, though hopeful, commentary on current condition of geodemographics has emerged. This body of work contains extensive documentation of the perceived limitations of contemporary practices and the stagnancy in progress, surmising that the building and application of geodemographics, for the first time in its long history, is no longer cutting edge, and that, particularly in the current era which is characterised by technological advancement, geodemographics seem to have fallen behind in comparison with the recent leaps forward in other data-led processes in other domains (Singleton and Longley, 2009b). However, far from being entirely negative, these discussions are also widely balanced with recommendations for potential future advancements, albeit rationalised by the challenges which have thus far prevented tangible, practical action in this direction (Longley, 2005, 2007; Singleton and Spielman, 2014).

2.4 The contemporary landscape and future directions

2.4.1 The “open data revolution”: A new data landscape

Recent decades have seen a dramatic increase in the availability, consumption and application of data for deriving crucial insights in a range of contexts. Substantial development in computing power and infrastructure alongside advances in software have made it easier than ever before to collect, store and manipulate data at unprecedented speeds and volumes (Dalton and Thatcher, 2015), both traditionally ‘standard’ forms of data and more complex forms, such as image and spatial data. The culmination of these factors has been an information, knowledge and data revolution permeating industries, governments and individuals alike (Mayer-Schönberger and Cukier, 2013).

Initially the private sector were best placed to take advantage of this emerging potential, both in the collection and storage of data and in its use, thanks to the interest and investment available in both time and resources. As such, the impacts of these advances have been most acute across a range of commercial industries. Many organisations were also quick to identify the value of their own data as an increasingly valuable asset. Consequently, an information economy soon developed in which organisations with access to data thrived. With early investment in the technology and data infrastructure required, the private sector were able to maximise on this commodity either through its direct sale, or in its adoption and analysis, generating wealth-creating solutions to wide-ranging real-world problems (Van Zyl, 2014).

Geodemographics was one such example of an industry which benefited greatly from this environment (Harris et al., 2005). Identifying and maximising on the potential of a combination of multivariate datasets almost five decades ago (see Section 2.3.2), commercial geodemographics was an early adopter and significant beneficiary of a data revolution which has since transformed 21st century life (Dalton and Thatcher, 2015). Since, it has continued to benefit, making continual improvements to the commercial classifications based on increased data access (Tate, 2018; CACI, 2020). Echoing the approaches demonstrated throughout its Urban Analytics ancestry, the industry advanced at the cutting edge of contemporary innovation, employing increasingly timely and novel data at greater volumes than had been possible before. In doing so, the commercial geodemographics outputs advanced beyond the achievable scope of the academic and public sector (Dalton and Thatcher, 2015).

However, progress is once again possible. Though slower to respond, the original "Data Revolution" has been followed by an "Open Data Revolution", which is bringing fresh possibilities (Benneworth et al., 2018). Where high levels of investment were once required to access the potential of such data, an increase in computing power at reduced cost has lowered the bar for accessibility, opening this potential up to a wider audience. Additionally, generally improved data infrastructure has facilitated easier linking of data, enabling increasingly sophisticated multivariate analysis at speeds which can even support real-time processing, often requiring nothing more than a common laptop (Singleton and Longley, 2009b). Now, those with access to data and the necessary creativity, expertise and technology can engage in this information economy and similarly benefit from its potential. Consequently, this has introduced fresh opportunities for innovation and a re-exploration of data-led approaches in academia and the public sector, particularly where collaborations can be arranged and resources shared (Benneworth et al., 2018).

Arribas-Bel (2014) emphasises the importance of these emerging data sources in conjunction with existing data and their potential for adding new insights for solving old problems, especially in understanding urban phenomena. As such, this new landscape has the potential to encourage the long-awaited next wave of development in geodemographics (Savage and Burrows, 2007; Singleton and Spielman, 2014). Though the commercial sector has once again responded much quicker, having already adapted their sophisticated data infrastructures to include the free and open data alongside the, now traditional, proprietary or paid-for transaction and consumer data (see Section 3.5.1), there is nevertheless an opportunity for the public sector and academia to respond with the development of more advanced free and open alternatives. Moreover, it is becoming increasingly important that they do, as will be discussed in Section 2.4.3.

The public sector in the "data revolution"

The public sector has been widely receptive to the new opportunities, understanding the potential of its own data and implementing new strategies to capitalise on this potential. Responding to criticisms of being "information rich but intelligence poor" (Local Government Association, 2013, p. 15), data focused philosophies have been adopted within central government departments and at a local level, alike. LAs across the UK have shifted towards the adoption of increasingly adaptive approaches for generating insights and informing decision-making processes, encouraging a culture of respect for data-led applications. The re-purposing of routinely collected government and LA data in deriving insights with which to inform public sector activity and policy development has become increasingly mainstream, reflecting the trends demonstrated in the private sector.

Intelligent, data-led techniques are being implemented within some local government departments as a matter of course, facilitating increasingly bespoke responses to public sector challenges, and in many instances, offering an overdue shift from traditional "one-size-fits-all" approaches (Department of Health, 2012). The aim of this shift is to use the newly available data to generate value for the public, wherever possible, either socially or economically, or in the improvement of public services (Arribas-Bel, 2014).

Specifically, utilising available data to cultivate efficiencies in the delivery of such services and the allocation of valuable resource has been high on the political agenda in the contemporary climate. Increasingly devolved LAs seeking efficient strategies for balancing growing demands and reducing budgets as a consequence of the recent recession have created fertile ground for such data-led innovation (Champion, 2014; Longley, 2005). This is likely to continue based on the bleak economic outlook generated as a result of the current COVID-19 crisis and its enduring consequence.

Additionally, the COVID-19 crisis itself has brought to the fore the importance of data in underpinning decision making processes, from the perspective of the government, LAs and society at large. The pandemic has amplified existing demands for complete transparency in the government's decision making processes and the open sharing of data and information

(Arribas-Bel, 2014). Requests have come from academia, opposition ministers, and the public more widely, and have been commonplace in the mainstream media and across social media platforms (Freeguard et al., 2020). In response, the government has sought to reassure with rhetoric containing claims of "following" and being "guided by the science", albeit not without criticism (Mathers, 2020), and with, on occasion, daily press briefings and the sharing of various statistics associated with the outbreak via *coronavirus.data.gov.uk*. There have also been publicly acknowledged recommendations for the UK government to recognise the value of internal data sharing practices and prioritise their improvement (Caldicott, 2020). Data-led public sector practices are becoming not only increasingly accepted, but expected, in today's culture, a progression which looks only set to continue.

Government data initiatives

To support the endeavours outlined, two decades of successive governments have developed and promoted a series of strategic data initiatives. Whilst the specifics have evolved, two core objectives have remained consistent, the first, to encourage improved decision making processes through the promotion and routine adoption of internal data-led approaches, and the second, to stimulate greater accountability and transparency, including in the sharing of public sector data, wherever possible and appropriate (The National Archives, 2013; Conservatives, 2015).

A 2017 strategic report commissioned and released by the current government (Cabinet Office, 2017) identifies data as a "critical resource" for achieving efficient, effective services tailored to meet the needs of the public. It also outlines a list of priority actions. These include, but are not limited to, the boosting of capacity for internal data use and the investment of data science and analytical capability across government to encourage mainstream data-led policy development, the provision of guidance on standards and best practice for data use for public sector bodies, the strengthening of open data assets, the creation of a Data Advisory Board and a Data Steering Group to govern the national data strategy and offer bleeding-edge thinking on data innovation, respectively, and the appointment of a Chief Data Officer to lead on the use of government data. Many of these intentions are later re-emphasised in the National Action Plan (NAP) 2019-21 (Department for Digital, Culture, Media Sport, 2019).

The main outcomes of these initiatives include the launch of *data.gov* in 2010 (which was supported with a re-launch in 2018), the UK government's open data platform for sharing data from within public sector bodies including central and local government, and the launch of *OS OpenData* in the same year, a similar platform for the release of free digital maps of Great Britain. Additionally, a further £8million of investment was announced in 2012 to enable public bodies to release open data to support commercial enterprise (Cabinet Office, 2012). In the same year, a further £10 million funding was also invested by the non-departmental public body the Technology Strategy Board (re-branded InnovateUK), part of UK Research and Innovation (UKRI), to support the launch of the Open Data Institutes (ODI) initiative (Guardian Government Computing, 2012), an independent conduit to the government and commercial organisations to facilitate the development of a reliable, open data ecosystem (Open Data Institute, 2020). Since, Administrative Data Research UK (ADR) has also established (in 2018) based on investment from the Economic and Social Research Council (ESRC). Its core function is to assist in linking together data held by different government departments and support the implementation of administrative data in research, though this initiative has arguably generated less success thus far, facing limitations in gaining access to timely and granular data, and in linking the data.

Nevertheless, these priorities of central government have also infiltrated the culture of data collection and the open sharing of data at local government level. For example, as part of their 'Smart Leeds' initiative (Smart Leeds, 2020), LCC regularly share routinely collected data on their own open data platform, Leeds Observatory (2020b) or in partnership with other local initiatives, for example Data Mill North (2020) (see Section 5.2.5).

All of this activity acknowledges the important role that data can play in a modern public sector. Whilst there are criticisms that the tangible outputs have not been enough, which will be discussed in more detail in Chapter 5, the actions taken to foster a nurturing environment have resulted in a wealth of open and public sector data available, outstripping the data freely and openly available at any time before. This landscape could be game-changing for open geodemographics, which have repeatedly thrived each time the open data landscape has advanced in the past.

2.4.2 Revival of interest in urban analysis in academia and the public sector

Geocomputation has been a major beneficiary of the data revolution. Commercially, location-aware data-led applications have transformed a range of industries from logistics to insurance and beyond (Lohr, 2015). Moreover, with much of the increasingly available data being intrinsically location-specific (Torun, 2016; Karimi, 2014), and spatial data offering some of the richest potential (Mayer-Schönberger and Cukier, 2013; Malloy, 2016), specifically in the public sector (Cabinet Office, 2017), these data have also frequently been at the heart of the public sector data revolution.

In turn, geodemographics has been a particular beneficiary of these advancements. Those with access to relevant, current and appropriately granular spatial data, and the ability to efficiently handle its computationally expensive analysis, have been able to thrive in this space (Longley, 2012). Consequently, players emerging as successful in the UK geodemographics industry over this period were, as mentioned previously, initially commercial organisations, primarily private sector data vendors and credit information companies such as CACI, Experian and TransUnion, who were able to capitalise on their means, including access to data, infrastructure, investment and expertise to develop a suite of competing products. By including new data into the building of their classifications, they incorporated a broader range of key population features, leading to timelier and increasingly granular results. Moreover, whilst the recent deluge of data available to the commercial geodemographics industry has supported the development of increasingly sophisticated classifications, the trend for data-led decision making has similarly increased demand for their products (Harris et al., 2005).

Public service reform and new local agendas (See Section 2.4.1), combined with pressure on local government to deliver demonstrable returns on their investments (Alexiou and Singleton, 2015), have similarly boosted the use of geodemographics in the public sector, generating a "renaissance" in applied social research (Harris et al., 2005; Singleton, 2016a). Efforts to adopt best practice demonstrated in the commercial sector, and reap benefits from the intelligence offered (Williamson et al., 2006), have propagated a fresh growth in interest in academic study and the public sector (Longley, 2005; Singleton and Spielman, 2014). Consequently, geodemographics are now positioned as a key component in insight generation (Local Government Association, 2013).

Many LAs, specifically, have identified geodemographic classifications as a useful, holistic variable when used to support increasingly bespoke local analysis to understand the spatial distribution of phenomena of interest to public sector agents (Batey and Brown, 2007), particularly with regards to health, crime and education (Batey and Brown, 2007) (see Chapter 8 for a discussion of examples). Treating the public as *consumers* of public services (Longley, 2005), local government analysts have adopted the methodology used to predict consumer behaviour to instead highlight the composition of demand for public sector services and resources, and derive insights with which to inform local government policy development (Harris et al., 2005; Burrows and Gane, 2006). Such "social marketing initiatives" (Brunsdon et al., 2018) have helped local policy-makers gauge social attitudes, and more intelligently develop strategies for service delivery and target the allocation of public resources (Longley, 2012).

This kind of activity has become an essential part of local government, particularly as they become increasingly devolved, raising expectations for more autonomous decision making. Employing what Longley (2012) describes as a "localism agenda" is difficult without a solid understanding of the population, including the social structures. This is only likely to become ever more important throughout the twenty-first century, for example, as we move to increasingly smart cities. For instance, Robinson and Franklin (2020) highlight the importance of positioning sensors carefully to measure specific populations, this requires a good basic knowledge of who the population are and how they are spatially distributed. This is just one example, but it highlights the continued and future importance of understanding social structures in the public sector.

In response, CACI, Experian and TransUnion have each developed a "Public Sector specific" classification. Marketed as suitable for the requirements of LAs, these seek to more appropriately cater for public sector demand. These products, and the displacement of academic attention following the commercial takeover which resulted in an almost complete vacuum of freely available alternatives, has seen many local governments, including LCC, frequently opt to license commercial outputs. However, LCC are growing increasingly discontented with this option, citing discrepancies and inconsistencies between the classification outputs and their informed expectations, based on local and expert knowledge (discussed further in Chapter 4). Moreover, the academic literature is growing increasingly sceptical of the employment of commercial geodemographic classifications in public sector applications, as detailed in the next section.

2.4.3 Critical evaluation of the use of commercial classifications in public sector applications

Though the theoretical methodology adopted in the development of the commercial classifications is seemingly transferable to the public sector, since the ideas found their origins in the public sector (Section 2.3.2), there is debate as to whether it is an appropriate practice to use commercial classifications to solve public sector problems, or whether there is a requirement to develop public sector specific alternatives. Whilst many employ these "off-the-shelf" classifications with limited scrutiny (Slingsby et al., 2011), others are beginning to increasingly question their general reliability, applicability and trustworthiness, and particularly their transferability for use in public sector contexts. This section presents four of the most commonly recurring arguments against the uncritical use of commercially developed classifications in public sector applications, found in the related academic literature.

1. Differing priorities and motivations

A major concern in the adoption of commercial classifications in the public sector is whether their use is theoretically appropriate, and whether commercial and academic practices are underpinned by different motivations. Fundamentally, despite the "general-purpose" label assigned to most commercial classifications, they are still designed with the primary aim of identifying distinct consumer groups, based on their consumer behaviours. Consider the subtle shift in intent from Booth's work in identifying different levels of poverty, to Acorn's identification of different types of affluence in its consumer profiles. As an illustration, Singleton and Longley (2009b) question whether parallels exist between holiday preferences and attitudes towards public health, security and education, and ask whether it is therefore appropriate to employ the same classifications to consider both.

Further, a report by the Local Government Association (2013) proposes a fundamental and important difference in the end-user intentions in commercial versus public sector applications which could undermine their suitability further. It suggests that the tendency for commercial classifications to help identify target consumers based on their likelihood to either be a current customer, or to be easily persuaded to convert into a customer, is based on a *presence* of something, for instance, their likelihood to buy a particular consumer good. However, this is claimed to be at odds with the needs of the public sector end-user, who is often seeking to identify those with an *absence* of something, or a particular need. This is not always the case, but in applications seeking to identify households in need of a service or resource, for example, the suitability of geodemographic classification outputs as provided by the commercial classifications is called into question.

Though the commercial sector has sought to quell these concerns with the development of Public Sector specific classifications, this debate has continued to become increasingly nuanced, with some extending the question to ask whether it is enough to offer any single classification with the expectation that it will be relevant and suitable to the entire, broad range of public sector applications. Returning to the above example, one might ask whether parallels actually exist in attitudes towards public health, security and education. If the answer is no, doubts might again be raised regarding the appropriateness of single Public Sector classifications produced by the commercial classification vendors in different public sector applications, and moreover, if the commercial sector is best placed to understand and account for the nuances of the public sector to develop classifications appropriate for the range of different circumstances (this discussion is extended in Chapter 8).

2. Inability to critically evaluate the development process

Critically, commercial classifications are developed in an opaque, "black-boxed" environment. In order to maintain the commercial sensitivity of the products (Singleton and Longley, 2009b), and due to the value of the data and processes, the specific details of their development are protected, kept from public knowledge or discussed using vague and ambiguous terminology (Burrows and Gane, 2006; Savage and Burrows, 2007). In this situation, the specifics of the methods are unknown, and the decisions made are not shared (Dalton and Thatcher, 2015).

The process of building a geodemographic classification can be extremely subjective, and each decision can dictate the final result. Often these decisions are explicitly led by the developers' experience and preferences, combined with an element of pragmatism (Singleton and Longley, 2009b). The choices made in selecting and executing the methods used can directly impact the outcome of the classification (Singleton, 2016a). Visibility of the justifications or explanations for each potentially crucial decision could therefore be extremely useful in understanding the method and final result, and ensuring that the justifications made are appropriate in the context of the application.

In the use of the commercial products, it is not possible to know which cluster methods, weighting techniques, variable selection methods or any other subjective decisions have been made throughout the process (Longley, 2012). The black-boxed development environment simultaneously erodes the ability to scrutinise the results and makes them impossible to reproduce (Longley and Singleton, 2009a). However, currently, verification of the commercial classifications occurs in-house, instead of in a formal scientific or peer-reviewed setting (Dalton and Thatcher, 2015). Whilst one might suggest that the end-user may be able to evaluate the result based on successes achieved in application, in many cases in the public sector, it is not appropriate to just attempt an application of the classification and evaluate afterwards whether or not it was successful. The stakes are often much higher in this setting, where applications have the potential to tangibly affect life chances (Longley, 2012; Singleton, 2016a), thus confidence in the metrics employed in public sector research is desirable at the outset.

Moreover, scrutiny and reproducibility of the results are core components in producing research with scientific integrity, and in being able to validate and trust the final results of the research to which the classifications are applied (Twa, 2019; Longley and Singleton, 2009a). Not only is it necessary in all science to have a healthy scepticism about results and to be able to explore them (Brunsdon et al., 2018), Longley (2012) specifically questions the ethics of basing decisions of public service allocation and public spending on a measure which cannot be interrogated. Finally, the fit of the result is also not shared as part of the licensing of commercial classifications, and any uncertainty in the model is withheld. Slingsby et al. (2011) highlight the potential for deriving more meaningful and well-informed public sector decision-making where uncertainty in a classification is transparent and can be evaluated. This is echoed by Singleton (2016a), who calls for all aspects of the build of a classification which is used in the public sector to be shared in the public domain, to increase social responsibility.

These are particularly timely concerns amidst the current trend for increasingly open public sector analysis and accountable decision making (see Section 2.4). Genuinely open decision making is hamstrung when it is based on metrics which are developed in a closed, opaque process. Moreover, Allen (as cited in Ghosh, 2019) warns of a “crisis” in scientific research where reproducibility is not prioritised, and implores analysts to check the reliability of patterns found in the use of big data processes to ensure that the outputs are genuine and not a consequence of the development process. Clearly the potential for this kind of validation is not possible in the use of black-boxed commercial geodemographics, and it is therefore difficult to assess how meaningfully the classifications represent real societal structures, rather than simply forming randomly as a consequence of the data (Longley, 2007).

3. Inability to assess the quality, provenance and contextual suitability of data used

Similar concerns are raised regarding the concealment of the data used in the development of the commercial classifications, the provenance and quality of which it is also not possible to assess (Singleton and Longley, 2009b). Since the quality of the analysis can be a function of the quality of the data (Harris, 1998; Parker et al., 2007), it is essential to understand any potential deficiencies in the data, such as in its completeness, or specific decisions which have been made in its preparation which could have introduced bias, to be able to have confidence in the outputs later derived. The end-user also needs to know the data used has been interpreted, processed and handled prior to input. If, for example, errors or bias has been introduced in the data cleaning or pre-processing phase, it is impossible to measure or adjust for the error to avoid further propagation if there is no visibility in these initial stages (Harris, 1998).

Moreover, the nature of geodemographic applications is such that end-users often append ancillary data to the results to infer further understanding in broader contexts. There is thus a risk that the same data could be used to both build and validate the classification, or to make predictions based on circular logic if the input data is unknown (Brunsdon et al., 2018). These concerns are further increased by the nature of many of the novel elements which underpin the sophistication of these classification methods. For example, there is inherent risk in the mixing of aggregate and individual data, and in the use of alternative data sources for which, unlike the traditional census variables, the stringency of the data collection procedures are not always guaranteed (Longley, 2007).

In addition, Harris (1998) warned against the replacing census based geodemographics with population insights generated from lifestyle data, on the basis that lifestyle data is likely to focus more on the affluent population, rather than those in need. Many commercial geodemographics are today increasingly built on lifestyle data (Harris et al., 2005). Since it is those who are most in need who are often of most interest to the public sector, one might want to be more considered in their use of lifestyle based geodemographics in this context, given Harris's (1998) caution, or, alternatively, one might hope that the necessary considerations have been made in the development of the classification, particularly in the development of the "Public Sector specific" options. It is, however, difficult to assess whether the necessary mitigation has been made, since the development of commercial geodemographics is black-boxed in nature.

4. Oversimplified outputs

There is an argument to suggest that the licensed products which release just the outputs is a welcome simplicity for some practitioners who might be overwhelmed with the detail of the development process, and instead, are simply looking for a functional, off-the-shelf metric to implement. However, the ability to review the process should be available for those who desire or require it (Singleton and Longley, 2009b; Singleton, 2016a). Moreover, it could be not only offensive, but misguided, to suggest that practitioners need to be kept apart from the detail. A study by Slingsby et al. (2011) revealed that end-users in the public sector felt that a broader, more contextual understanding of the final classification, achieved in this case by considering the fit of the 2001 OAC alongside the final result, aided them in a more accurate and considered application. Combined, these factors present a set of circumstances which could lead to a potentially dangerous misinterpretation of the outputs. Once again, the black-boxed development process of commercial classifications poses a risk to the accuracy of any insights subsequently drawn.

Contained herein is another demonstration of the advantage of open classifications in this context. The open nature of the 2011 OAC supports the exploration and encourages critical review of not only the result, but the procedure, and affords a level of scrutiny necessary to support their confident adoption in public sector use which is more in line with the history of free and open geodemographics developed for public benefit prior to the emergence of the commercial alternatives (Singleton and Spielman, 2014).

2.4.4 Weaknesses of national-level classifications and place-specific alternatives

One, more methodological, criticism of the commercial classifications which poses an additional potential weakness and which might further undermine their use in the public sector, relates to the national extent at which they are now commonly built. Concerns have particularly been raised regarding their ability to successfully discern local population structures (Singleton, 2016a; Singleton and Longley, 2015; Local Government Association, 2013). This is also relevant to the open 2011 OAC, which is likewise generated at a national level. Whilst much theoretical discussion exists in the published literature, LCC have also raised empirical concerns which have in many ways initiated the motivations of this thesis (see Section 1.1.3). The discussion and subsequent activity relating to this concern, in particular, therefore warrants the more in-depth review, presented below.

A summary of the core concerns

Criticisms specifically highlight the potential *masking* of unique local features by the national-level classifications, and thus the loss of critical, locally relevant information, particularly in small-area geographies which diverge from the national picture. This is noted as being a potentially more acute problem in large cities with distinct make-up (Singleton and Longley, 2015; Brunsdon et al., 2018), such as Leeds. Additionally, classifications derived at the national extent cannot incorporate local intelligence and nuance. For example, national-level classifications cannot account for the impact of public sector (or otherwise) initiatives deployed in a local setting, which might affect attitudes and behaviours in specific populations and encourage a response to council activity which might be counter to the national response, or counter to expectations based on the area's national classification profile. Moreover, geodemographic classifications already face problems of inclusion and exclusion, i.e. issues of misclassification, which can be further compounded by inaccurately defined groups (Petersen et al., 2010). Finally, from a broader perspective, the national-extent adopted in geodemographics is counter to the trend exhibited in contemporary theoretical developments in other fields of Geographical Information Science (GIS), such as Geographically Weighted Regression, which are beginning to consider a much more granular focus of geography (Singleton and Longley, 2009b).

It is therefore essential to ensure that the underlying methodology employed in geodemographic classification development is constructed to achieve results which are as accurate as possible. The national-level concerns have therefore led some to call for a shift towards developing classifications at a more granular, local level, generating place-specific classifications which are based on a single city and are designed to more appropriately reflect locally specific phenomena (Singleton, 2016a; Singleton and Longley, 2015; Local Government Association, 2013).

The emerging environment supporting a shift to local, place-specific geodemographics

Although Atlas (1981) and Openshaw et al. (1980) identified, three decades ago, that a shift to a local extent could naturally derive an entirely different cluster composition in the locality than is produced at a national extent, widespread arguments that the altered result could offer an *improvement* came much more recently, and practical exploration of this theory more recently still (Gale and Longley, 2012; Singleton and Longley, 2015).

As discussed in detail in Section 2.3.1, early precursors to modern geodemographics placed a much greater emphasis on developing a real awareness of the societies being modelled, and the underlying urban structures. Many of the seminal works in formerly vibrant areas of academic interest, such as Urban Ecology and SAA, from which modern geodemographics are widely acknowledged to have evolved, primarily took a local focus, seeking to understand the socio-spatial patterns underpinning specific cities. Much of this work was concentrated in the US in the mid-twentieth century, where unprecedented data facilitated a flurry of analysis in this domain (Batey and Brown, 1995; Harris et al., 2005). When the tide shifted towards the development of national-level classifications in the 1970s and 1980s, there was an acknowledgement that local-level insights could be sacrificed to enable national comparisons, but the benefits of the latter seemingly outweighed the losses endured (Webber and Burrows, 2018). These attitudes, which favoured national-level comparisons over local detail, prevail in the contemporary commercial sector and underpin decisions made in the development of the 2011 OAC (Singleton and Longley, 2015).

Yet, the situation today is very different. Swinney and Carter (2018) proffer that the composition of individual cities are more distinct than ever before, with the distinct evolution in different cities resulting in more complex societal structures. As Reibel (2011) notes, in the 1920s, it was possible to differentiate the population in fewer variables. In the 1950s, new suburbs emerged with new types of cultural subdivisions. Now, there is even more variation generated from the legacies of old suburbs and new developments, more identities inhibit different cities, based on more combinations of attributes all living alongside one another (Webber and Burrows, 2018). Consequently, there is a need to make sure that the classifications developed are equipped to handle the emergence of unique population structures, particularly in city environments.

Nevertheless, the suggestion to shift back to a city-level extent does not come without its own concerns. Chiefly, the shift to national-level classifications was made to support comparability at a national level, an ability which would be sacrificed with a shift back to a place-specific methodology (Singleton and Longley, 2015). But there is also an emerging sense that the necessity to choose either one extent or the other (national or local) no longer remains. Instead, it is both possible, and likely beneficial, to consider both approaches (Gale et al., 2012). As Singleton (2014) reflects, there is no single, observational reality, and that it may therefore be naive to expect a single classification system to satisfy all requirements. As such, it has become increasingly commonplace for commercial geodemographic systems to include a general purpose classification alongside a suite of bespoke classifications, modelling populations from the perspective of a range of specific domains (alternative, "domain specific" bespoke classifications will receive more attention in Chapter 3). It seems therefore appropriate to suggest that a place-specific classifier could, if necessary, take a place alongside national level classifiers in the arsenal of available tools for modelling populations both intra- and inter-regionally.

From a practical perspective, the development of place-specific open geodemographic classifications, particularly within or in partnership with the public sector, could be supported by the rapid generation of local level data increasingly available to LAs (Section 2.4.1). Since much of this data is only available to LAs at a local level, its use in the development of national-level classifications has been hamstrung. Additionally, alongside the proprietary commercial data, it is broadly understood that commercial classifications include openly available public sector data, such as information from the electoral roll and the Land Registry (Harris et al., 2005). It is also likely, although not possible to know for sure, that the Public Sector specific commercial classifications rely on this genre of data even more. Thus, the natural advantage of the commercial vendors, in terms of their access to data, is reduced, or even removed, in the context of building a classification on local, social data, given the access to the wealth of public sector data that the LAs themselves have. As such, open geodemographics is now in a new position, able to challenge the dominance of the commercial vendors who have benefited thus far from their superior access to national-level data, to begin to develop their own transparent alternatives to the commercial classifications.

Similarly, the technological advantages which the private sector has previously enjoyed have also been reduced, as alluded to earlier. The advent of free software which is capable of handling vast quantities of complex data, including statistical and GIS software, and which continue to advance, enable the easy generation of geodemographic classifications based on a large volume of data (Longley, 2012). Again, the reduced computational costs associated with developing local-level as opposed to national-level classifications will also level the playing field somewhat, giving rise to the potential for the academic and public sectors to begin to develop viable non-commercial alternatives.

Moreover, although ultimately developed at a national level, the latest iteration of Acorn does offer the input of slightly different variables region-to-region to account for cases where a particular dataset might not be available in all areas, for instance, between England and Scotland (CACI, 2020). The inclusion of some place-specific data in the development of the national-level Acorn challenges the entrenched methods which have rigidly prioritised maintaining a national consistency to enable cross-region comparisons, and which may thus far have limited local-level exploration. As such, it could signal a relaxing of the incumbent methodologies which have dominated since the late 20th century, as local level and local specific data becomes more readily available. However, the Acorn methodology is still focused on deriving the final classification at a national extent. This remains true for all of the main products of the commercial UK Geodemographics Industry.

A summary review of existing place-specific geodemographics

The situation described above has encouraged several studies developing local classifications to explore the potential benefits. One of the most commonly cited is Singleton and Longley's (2015) development of a London specific re-classification of the 2011 OAC, labelled the London Output Area Classification (LOAC) (see Section 4.3.1 for a more in-depth discussion). This study is complemented by a similar investigation which generated a similar re-classification of the 2011 OAC in the Liverpool city region (Singleton, 2016a). Both of these studies recorded improvements in the identification of unique, local population structures, and act as evidence to support a re-consideration of place-specific classifications elsewhere, particularly for deriving insights which could be crucial for the development of locally specific public sector strategies and policies.

These findings are echoed in a similar study undertaken within the public sector itself by Hull City Council, in partnership with Rushmoor Borough Council and Leicester City Council. Citing concerns with the local accuracy and reliability of both the open and commercial level classifications, and driven by empirical benefits identified in an initial development of a Leicester specific local classification, the researchers authored a local-classification development manual (Local Government Association, 2013), outlining a step-by-step guide to support other local governments to develop their own local classification based on the methodology adopted by the 2011 OAC. However, once again, the methodology was focused simply on input data sourced solely from the census.

Gale et al. (2012) extended this work one step further, prototyping a tool which supported end-users with limited technical expertise, specifically in the public sector, to more easily and flexibly derive bespoke versions of the 2011 OAC, namely local-specific classifications to co-exist alongside the national-level options. This tool, named “GeodemCreator”, also supported the inclusion of non-census data, expanding the expectations of open geodemographics in light of the new data landscape (Section 2.4.1). Although no further practical applications of either GeodemCreator or the documentation compiled by Hull City Council have been found in the published literature, these developments signal a change in thinking in geodemographic classification development, specifically for developing alternative options for applications in the public sector.

Debenham (2002) and Debenham et al. (2003) have also published work developing postal sector level classifications for just Yorkshire and the Humber. These studies linked census data with other administrative data, primarily focusing on work-place-characteristics, and the “demand” for and “supply” of services in an area. Although the census data was expanded with public sector administrative data, the studies leaned more towards a commercial end-user. While this work acts as a precursor to some of the work contained in this thesis (which will be highlighted throughout), there are several key differences. As administrative data was added, the performance was evaluated against a benchmark local classification, generated to replicate the national-level MOSAIC. However, the black-boxed development of MOSAIC inhibits an exact local-level replica, thus it is not possible to compare local vs. national level to explicitly attribute any change in performance to the change in extent. The subsequently developed 2011 OAC (see Chapter 4) has supported the development of an exact local-level replica of the 2011 OAC which can be comapred to national-level and used as a benchmark as new data is added. Moreover, Yorkshire and the Humber is a geographically large and diverse region, incorporating large urban conurbations, rural areas and seaside towns. The unique local phenomena which this thesis is trying to capture is likely to necessitate a much more granular focus, as per Singleton and Longley's (2015) recommendations to explore place-specific classifications at a city level.

Finally, Burns et al. (2018) offer a recent example of a classification, which focuses specifically on the city of Leeds, matching the study area of this thesis. In this example, however, the local extent is somewhat secondary. The focus of the work is on generating a novel classification for *individuals*, based on individual-level data taken from the 2001 census. As such, this is not a place-specific classification in the same sense as has been discussed in the previous examples, nor is it really presented as such. Although the classification was derived at a Leeds-specific extent, it is difficult to evaluate it as such and to identify the impact of the change in overall extent, since the shift to individual-level data likely had a more substantial impact than the place-specific element. Moreover, the data upon which this classification was derived is now two decades outdated. As such, whilst there are definite learnings to be taken from the example, the specific classification derived is also likely to offer a similarly outdated outlook of the residents of Leeds. Additionally, several weaknesses were identified within the development of the classification, and extensions to the methodology were recommended. Consequently, the output derived is not considered currently as a ready place-specific alternative to the 2011 OAC and the commercial classifications.

2.5 Conclusions and context

This chapter has summarised the chronology of geodemographic classification development from its early precursors to the present day, particularly through the lens of its development, or the relevance in its application, in the public sector. In doing so, a narrative has been presented documenting the advancements and subsequent evolutions which have been periodically stimulated by contemporary demand and innovations in technology, statistics and data availability. The present-day conclusion details a situation in which the commercial sector have dominance in the geodemographics market and continue to be widely adopted in public sector applications, despite reservations relating to their appropriateness, their cost, the black-boxed nature of their development, and the relevance of their nationally derived results at a local level, espoused by both academics and local government analysts.

Although positive, the developments outlined have seemingly not encouraged a mass departure from the use of national-level commercial classifications in the public sector. Seemingly, until now at least, some loss in transparency and local-level insights have been a price that many are still, figuratively and literally, willing to pay for the increased sophistication offered (Longley, 2012), suggesting that further work is required to generate a more attractive alternative. In particular, instead of opting to follow the guidance set out by Hull City Council (Local Government Association, 2013) to derive a Leeds-specific alternative of the 2011 OAC in-house, LCC continue to licence costly national-level commercial classifications developed in a black-boxed environment and which seemingly mask local insights. However, their partnership on this thesis indicate a belief that the available solutions, including the Hull City Council option, do not yet meet the requirements for developing a viable open, place-specific classification capable of challenging the dominance of the commercial incumbent, but that this could be a prime time for developments which might.

This summary has also highlighted the advancements in technology and data availability, set against the current political landscape, which present ideal conditions to support a new phase in development and the creation of not only a valid, but essential and more relevant, open and place-specific alternative to the established practices and classifications on the market. This thesis, therefore, will extend the options described in Section 2.4.4 by taking a more in-depth look at the framework employed and the methodologies used. It will test the capability and suitability of the data infrastructure within LCC to expand the input data beyond the reliance on census data, and address the challenges which arise in doing so, particularly in terms of the increased requirement for sophisticated variable selection methodologies which will be introduced (Longley, 2012). It will also consider other common criticisms of the established practices of geodemographics, in addition to those presented in this chapter, to understand whether further evolution is needed, beyond the change in geographic extent and the inclusion of novel public sector data. In order to offer an open alternative which is able to genuinely compete with the dominant commercial products, the issues with the 2011 OAC which have made it less favourable than the licensed options, and the practical challenges which underpin the enduring stagnancy in academic development need to be understood and addressed. This will begin in Chapter 3 with a detailed overview of the established practices, followed by a review of the criticisms and challenges which have repeatedly featured in the geodemographics literature over the past two decades, but which have not been addressed in the open geodemographic alternatives previously presented.

Summary

This chapter has presented a review of the relevant chronology of the field of geodemographics to-date, including a consideration of its origins and precursors. This work outlines the historic and present-day context in which this thesis is positioned, and provides an initial introduction to the challenges which have led to the commencement of this study. As such, this review has focused on the emergence and contemporary circumstances of geodemographic classification development with a particular interest on public sector involvement and the oscillation between the development of classifications at a local and national extent. The next chapter will extend this review, which has been largely theoretical, to similarly contextualise this thesis in terms of the current practical and methodological practices and challenges of the field.

Chapter 3 - Building modern geodemographics in practice

3.1 Introduction

Chapter 2 presented a summary of the evolution of geodemographics from its precursors in Urban Analysis to its position as an established and trusted framework for summarising complex population structures. However, details relating to the tangible practice of developing geodemographic classifications, which comprises a statistical methodology wrapped in a series of qualitative and subjective decisions, were not discussed. Claims that geodemographics are straightforward to develop or simple and easy to understand are commonplace, however, these claims undervalue both their appeal and the complexity of their development (Harris, 2001). Substantive work goes into deriving such seemingly simple, modestly sophisticated results, including careful considerations informing decisions built on expert knowledge, understanding, experience and skill. With each decision the resulting output becomes more or less relevant, more or less accurate, more or less objective, and thus more or less appropriate. It is therefore essential to have a full understanding of the traditional process and its potential weaknesses to be able to confidently generate a new classification.

Modern geodemographics are founded upon a standard framework (henceforth referred to as the "Standard Framework") which is customised in the creation of each classification by the many decisions made by the classification developers at each stage of the process. In recent years the proprietary commercial classifications claim to have made several extensions to this framework, improving their outputs to develop competition in the marketplace, whilst the practices underpinning open geodemographics have remained broadly consistent for a number of decades. This is often by design, to support longitudinal comparison (Gale et al., 2016). However, whilst tradition dictates much of the contemporary methodology, some question whether such a legacy-based process remains relevant in the 21st century (Openshaw, 2001; Singleton and Spielman, 2014), whether the next evolution of progress in the field should continue to adapt and extend the traditional framework, or whether it is time to seek a more revolutionary approach, and in each case, which are the most critical developments to be addressed.

This chapter begins in Section 3.2 by briefly introducing the most popular of the available geodemographic classifications in the UK. Section 3.3 presents an outline of the Standard Framework which underpins the building of these current classifications, detailing the distinct stages of the build and the decisions to be made at each stage. Section 3.4 demonstrates the Standard Framework in practice through a detailed run-through of the 2011 Output Area Classification (OAC) development process. Section 3.5 summarises some of the significant extensions that the proprietary geodemographic classifications in the commercial sector has made, beyond the Standard Framework. Section 3.6 reviews the criticisms raised against the traditional process as it stands, alongside the challenges which have been presented in the literature as barriers which have limited advancements up until now, (extending the discussion initiated in Chapter 2). Section 3.7 considers some potential forward-thinking extensions to the Standard Framework. Finally, Section 3.8 details the outline of the project documented in this thesis in the broader context of the recommended next steps for geodemographics as a whole.

3.2 Available classifications (UK)

There are several proprietary geodemographic classifications available on the commercial market in the UK. The following list comprises the most popular of the offerings, as identified by Webber and Burrows (2018).

Acorn from CACI: Acorn was the first consumer classification in the UK and markets itself as "the leading geodemographic segmentation of residential neighbourhoods in the UK" (CACI, 2019, p.3). It is developed at a national extent, is available at household and postcode levels, and is "no longer reliant on census data" (Tate, 2018). It is dynamically updated and uses proprietary commercial, public sector and open data. CACI also offer bespoke versions of Acorn to suit individual client needs and data, and domain specific variants for sectors including "health", "retail" and "leisure" (CACI, 2020).

Mosaic UK from Experian: "Mosaic" is derived at a national extent for the UK, and is one of several versions available in 29 countries worldwide. The product is continually updated, and the product is re-released with updates twice per year. Classifications are derived for individuals, households and postcodes based on a mix of proprietary commercial, open and census data (Experian, 2009). A "family" of Mosaic classifications are available alongside the original. These are optimised for specific sectors, including the public sector, the "digital consumer", and "shopper" segmentation (Experian, 2021).

CAMEO from TransUnion: Generated at national level, CAMEO has a global reach with a product in 36 different countries, including the UK, from where it originates (TransUnion, 2021). A universal International CAMEO is also available to support consumer comparisons across the 36 countries (TransUnion, 2020). The output is updated dynamically to support the movement of individuals between "segments" (classification groups) depending upon their changing circumstances (Sleight, 2014). It is available at individual, household and postcode levels.

Censation from AFD Software: Censation is marketed as a "simple but effective geodemographic classification system", which is generated at a national extent for all UK postcodes. It is not licensed as a stand-alone product, but is appended to ADF Software's proprietary "Address Management" products. The classification focuses on levels of wealth, life-stage, and any "distinctive characteristics" represented in a postcode. It uses data from the census, the Land Registry, and information gained from face-to-face interviews (ADF Software, 2021).

P^2 People and Places from Beacon Dodsworth: P^2 offers a classification for UK postcodes and administrative boundaries. It is generated at a national extent based on census and lifestyle data extracted from the Living Costs Food Survey and British Population Survey to categorise the behaviours, attitudes and lifestyles of UK consumers (Beacon Dodsworth, 2021a,b).

Sonar from TRAC: Sonar is a consumer classification of UK households generated at a national extent. It combines census data, Land Registry data, Council Tax Bands and Benefit claimant data with proprietary data to classify the households into categories relating to lifestage and wealth (Griffiths, 2020).

As discussed in Chapter 2, in the open geodemographics market, there is one main popular classification available, the 2011 OAC. This is discussed in detail in Section 3.4.

In spite of initiatives by some of these vendors, including CACI and Experian, to waive license fees for academic research purposes (Webber and Burrows, 2018), the main aim of these organisations is to generate revenue. Thus, whilst the outputs can be accessed, the available “literature” about these classifications is generally marketing material. As such, specific detail relating to the methodologies used to develop the classifications is scant. However, most of the above classifications seemingly adopt some customisation of the general, common framework outlined in the next section. However, CACI do claim to have developed beyond the “traditional approach” (CACI, 2020), although again, specific examples of how are not forthcoming. It is therefore not possible to know if their departure from the traditional approach is revolutionary, or represents a perceived substantial customisation of the Standard Framework discussed in Section 3.3.

3.3 The “Standard Framework”

3.3.1 Overview

There is an established framework which underpins the development of most geodemographic classifications, both commercial and otherwise. The framework scaffolds a multivariate clustering process by which classification groupings are generated. The process has remained largely the same since the 1970s (Gale et al., 2016), though each application will be necessarily unique to the developer (or set of developers), introducing an element of customisation supporting competition in the geodemographics market.

The specifics of each instance are mostly proprietary (Parker et al., 2007), as such, commercial sensitivities restrict their publication, though novel elements are frequently referred to in the marketing literature, promoted as unique selling points. Otherwise, Harris et al. (2005) is potentially the most comprehensive single source of this information, sharing snippets of information relating to the internal constructs underpinning the commercial products, as published across the various marketing literature. Though some of the specifics of this information is likely to be dated, given the considerable advancement in data availability, discussions with TransUnion suggest that the core of the framework discussed here has fundamentally maintained the same.

Though there is limited reference material to draw from (Murphy and Smith, 2014), there is a broad identification across the literature, academic and commercial marketing literature (Burns et al., 2018; Vickers and Rees, 2007; Openshaw, 2001), that the typical process comprises some version of the following core steps:

STEP 1: Define the geography and purpose

STEP 2: Identify data and refine variable choice

STEP 3: Data preparation

STEP 4: Select and run multivariate analysis

STEP 5: Repeat analysis (if hierarchical clustering desired)

STEP 6: Link outputs to geography

STEP 7: Produce descriptions of the classes (pen portraits)

STEP 8: Validate and enrich the results

The specific circumstances of each classification development will dictate a varying emphasis placed on each of the different steps.

The next sections provide detail of each step, first generally (in Section 3.3.2), and then in practice in the context of the development of the 2011 OAC (in Section 3.4). This detail not only summarises the discussions in the literature surrounding each step, it also acts as a comprehensive reference for the practical development of classifications and the decisions made in Chapters 4-8.

3.3.2 Summary of each step

STEP 1: Defining the geography and purpose

As outlined in Chapter 2, geodemographic classifications are typically derived at a national extent, generating a single classification for the entire UK, for instance. However, there remain decisions to be made with regards to the geographic boundaries of the geographic units (such as small-area geographies) to which the classification groups are to be assigned. Decisions are often founded on a combination of the intended purpose of the classification and a range of practical and theoretical considerations.

Arguments have been proposed for and against both coarser and finer geographic levels, even down to classifying individuals. For example, too coarse a geography could see the smoothing away of local heterogeneous patterns (Singleton and Spielman, 2014), and could introduce inefficiencies in the mis-classification and subsequent mis-targeting of individuals (Batey et al., 2008). Moreover, coarser geographies might increase the risk of introducing geostatistical biases (Harris, 1998), or falling foul of the common geographical scaling and aggregation problems, the Ecological Fallacy and the Modifiable Areal Unit Problem (MAUP). These two commonly cited pitfalls of geodemographics can be summarised as the potentially incorrect assumption that the characteristics associated with a given cluster represent every individual resident in the population assigned to that cluster (Harris et al., 2005), and as the problems which can arise in the generation of arbitrary small-area boundaries with which to distinguishing one area from the next (Dalton and Thatcher, 2015), respectively.

The consequences of the MAUP are difficult to understand and address (Martin and Bracken, 1991). It is widely understood, however, that every caution must be taken to ensure that variation identified between small-area geographies is genuine and is not artificially created or emphasised by the boundary divisions (Vickers and Rees, 2011). Nevertheless, boundaries are often imposed arbitrarily, concealing, or misrepresenting the population (Langford and Unwin, 1994). Similarly, though Williamson et al. (2006) suggest that the effects of EF might be less pronounced in geodemographics than in other spatial analysis, effects still pose a risk to the accurate development and use of geodemographics, and there is currently no solution to the problems that these effects cause (Dalton and Thatcher, 2015). Thus both the EF and the MAUP warrant careful consideration in the development of geodemographic classifications and in the decisions of the geographic scale used.

Burns et al. (2018) suggest that finer level classifications, particularly classifying at the level of the individual, will reduce the associated risks of EF. However, such decisions have their own inherent risks. Too fine a geography in a classification could introduce sampling errors resulting from small population numbers (Singleton and Spielman, 2014), and individual level classifications have also received criticism for their failure to account for the influence of an individual's environment and social interactions on their behaviour, attitudes and preferences (Harris, 1998). This might be particularly pertinent when considering characteristics such as employment indicators, which might have more contextual meaning at a coarser geographic level (Webber, 2004). Moreover, Burns et al. (2018) note the challenges of accessing data at an individual level scale.

Despite the concerns of the EF and the MAUP, it is still possible to derive a sense of the average characteristics found within an area by considering the individuals within. This is often desirable in practice, to generate order from the chaos of individual level data and remove some of the noise (Harris, 2001). Thus Burns et al. (2018) concede that the selection of the geographic scale of a classification thus necessitates a "trade-off", where some features must be prioritised over others. As such, many of these decisions are intertwined with the purpose of the classification.

In the commercial classifications, postcode areas are typically selected, although many also now offer household and individual level classification options (see Section 3.2). In practice, since public sector decisions and policy are often made in-line with administrative boundaries, high levels of granularity are often less important in the public sector (Webber, 2004). The mis-targeting of some individuals within these areas is an unfortunate but unavoidable consequence of local government activity, and thus cannot be necessarily addressed by individual level classifications. Thus, in open geodemographics, the selection often aligns with census boundaries, supporting the easy use of census data. This is often Output Areas (OAs), owing to these being the smallest geography at which the aggregate data from each census is released, offering the most granular detail and acting as building blocks for higher level census geographies (ONS, 2016a).

STEP 2: Identify data and refine variable choice

The next step is to identify relevant datasets and derive the input variables upon which to build the classification. This is arguably the most important step in terms of ensuring that the results are contextually meaningful (Murphy and Smith, 2014). The developer must decide the type of data that is to be included, and in which format, and identify relevant datasets. The potential data must then be assessed for inclusion based on its contextual and mathematical relevance.

The data identification and variable selection is a typically pragmatic process, dependent on a balance of convention, data availability and developer preference (Brunsdon and Singleton, 2015). Empirical selection methods informed by expert knowledge and context specific literature, particularly in the case of bespoke or domain specific classifications, are regularly combined with some scientific methodology, including data reduction practices, most popularly, correlation analysis. Murphy and Smith (2014) present a detailed discussion and exploration of the suitability of possible statistical methods.

It is essential to be selective in deciding which data to include and exclude, particularly in the age of “Big Data” and the increasingly infinite options potentially available (Longley, 2007). Desirable variables are those with the greatest potential for identifying differentiation between the geographic areas. The selection of variables is thus a complex and nuanced element of the process. The challenges and considerations which must be made are discussed in more detail in Section 3.6.

STEP 3: Data preparation

It is not always appropriate to use the data which has been selected for inclusion in its raw form. A detailed inspection of the data is required to understand which, if any, data preparation procedures are needed. This can include plotting the distribution, further correlation analysis or mapping the data.

A review of the academic literature and discussions with TransUnion indicate that a variety of approaches can, and often are, taken to prepare data for input into a geodemographic classification system. Requirements can be strongly dictated by the data and circumstances of the development process.

The following processes traditionally feature in this phase (where necessary):

- *Adjust the data type* – to generate data in a suitable format for the subsequent analysis.

  It might be necessary to change the data type, for instance, to calculate percentages or ratios, or maybe create a composite measure from several related variables.

- *Transform the data* – to normalise, reducing or removing any skew within the data.

  Skewed data can negatively affect the cluster assignments in the common cluster methods (Gale et al., 2016) (discussed further in STEP 4). For example, the potential occurrence of illness often increases with age. As such, a high prevalence of illness in a specific geography could simply be a reflection of an aged population, rather than an indication of poor health in the community. Standardising for age can assist in highlighting where there genuinely is higher than expected illness, irrespective of age. As such, input data is regularly normalised to account for potential skew.

- *Standardise the data* – to set all data to a single, common scale.

  It is difficult to mathematically measure *similarity* when comparing attributes measured on different scales. The importance of this in terms of the process of geodemographic classification development will be highlighted in STEP 4, where the classification procedure is introduced. This could be achieved with z-scores or range standardisation, if dealing with outliers.

Alternatively, some developers favour an application of Principal Component Analysis (PCA) to the raw data to generate composite "components" from the individual variables, which can be used as replacement inputs. This process was commonplace in Factorial Ecology, and is adopted and encouraged most recently in Brunsdon et al. (2018). In doing so, the PCA could render the above procedures unnecessary, since it can automatically account for any scaling issues and pull the data into a standard measure across the components produced. However, there is an argument that the practice could reduce any interesting dimensions which are present within the data (Harris et al., 2005). The detailed procedure for doing so, and the perceived benefits and limitations of PCA components vs. individual variables are still for debate, and will be considered further in Chapter 7.

The common processes outlined, and their impacts, are discussed in more detail in Brunsdon and Singleton (2015). Again, each of these decisions are nuanced and steeped in complexities. It is necessary to take the time to understand the data to be able to appropriately apply these steps. For example, there could be instances where normalising the data or removing outliers is not desirable, if, for instance, doing so could reduce or remove theoretically interesting real-world phenomena represented in the data (Reibel, 2011). There is thus nuance in the potential implications of the decisions made during this step. Some conflicting opinions between experts and alternative methods proposed for data preparation are discussed more thoroughly in Section 3.6.3.

Whilst the above discussion covers the most common data concerns, there are other potential issues highlighted within the literature which must also be considered at this stage. Often these limitations in the data are not so easily addressed. For example, census data is susceptible to under-counting, where measures are taken to maintain anonymity in areas with low counts for specific attributes. This is rarely, if ever, adjusted for in the adoption of census data (Voas and Williamson, 2001). Additionally, though it would be recommended, no classification currently accounts for error or uncertainty in the input data (Singleton and Spielman, 2014). It is evident, therefore, that the developer must have a broad set of general data literacy, an understanding of common pitfalls in data analysis (both within geodemographic classification development and beyond), and the knowledge and skills to reliably and responsibly work with the data and to prepare it appropriately. These recommendations will be starkly evidenced and further emphasised in practice in Chapter 5.

STEP 4: Select and run multivariate statistical analysis

The term *geodemographic classifications*, is somewhat of a misnomer. Though called a classification, in reference to the act of assigning each area to one of a set of groups, or classes, the technical process through which this is achieved is typically a *clustering* process. The subtle but important distinction lies in *classification* processes assigning objects, (small-area geographies in geodemographics), into a set of *pre-defined* classes, whereas, in a clustering process, the objects are grouped into a set of *clusters* defined *during* the process, not driven by no a-priori assumptions. Classification processes are therefore *supervised* Machine Learning (ML) methods. Conversely, clustering processes are more typically *unsupervised.* In line with convention, the terms *classification process* and *cluster method* will be used synonymously throughout this thesis to describe the process of deriving the classification output via a clustering technique, unless otherwise stated. Similarly, *class*, *cluster* and *groupings* will be used interchangeably.

Unsupervised learning techniques within geodemographics gained in popularity throughout the 19th century, initially by means of Factor Analysis (FA) and Principal Component Analysis (PCA) (Longley, 2012), giving rise to the era of Factorial Ecology, and later with the adoption of clustering methods, which have become the de-facto (Reibel, 2011) (see Section 2.3.1). Of the four different types of clustering methods, *partitioning, hierarchical, density based* and *grid-based*, which each differ in their mathematical detail (Jain et al., 1999), modern geodemographic classification systems most typically employ the partitioning method of *k-means* clustering, which seeks to split the data into k homogeneous groups with maximal heterogeneity between groups, in doing so, identifying internal structures. Several comprehensive explanations of the process are offered in the literature, including a detailed summary presented in Alexiou and Singleton (2015).

The method requires the number of expected clusters (k) into which to split the data as an input, to be decided by the developer. Whilst various techniques can be used to derive the optimal cluster number, a *scree plot* (also known as an elbow plot), is most commonly used. An example of this can be found in Section 4.4.1. However, whilst the plot provides a guide, there is some subjectivity employed in its interpretation. Whilst some have explored different methods, such as Reibel (2011), who discusses the development of a potentially more statistically robust method which tests and compares the impacts of different numbers of k, there currently remains no objective method of choosing the number of clusters. As such, these guides are often applied in context to generate a final decision. As an example, Batey et al. (2008) suggest that too many clusters are not conducive to mapping, and thus, one might choose to optimise the number based on a combination of the mathematical guidance and the experience of the end-user. However, since the intention is rarely to simply produce a mappable output, such a prioritisation seems poorly justified. Though it could be claimed that outputs that are clearly mappable might be easier for an end-user to interpret, one might be more justified in considering a cluster count which drives a better cluster performance, and would ultimately produce more appropriate decision-making, especially for use in the public sector.

An iterative process is next performed to split the data into the k clusters. In summary, a *seed* is randomly generated as a starting point for each cluster, the objects (small-area geographies) are then assigned to the closest seed by way of some geometric distance measure based on the multidimensional space between the object attributes. Typically, either Euclidean Distance (ED) or Squared Euclidean Distance (SED) are employed, with the latter used in the development of both the 2001 OAC and 2011 OAC (Gale et al., 2016; Gale, 2014). However, there could be a requirement to find a more appropriate measure as data becomes bigger and more complex (Everitt and Dunn, 1991). A review of some commonly used distance measures can be found in Gale (2014).

After the initial assignment to the seeds, a cluster *centroid* is calculated for each cluster (of objects assigned to each seed) from the average of the objects based on the initial cluster placement. Each object is then re-assigned to the closest cluster centroid. This process is repeated iteratively until convergence, at which point no objects are re-assigned. In doing so, *similar* objects are assigned to the same cluster. Often the full procedure will be repeated with x different initial random seed placements, at the developer's discretion. The best 'fit' of the x tests will then be taken as the final result, whereby the best fit is evaluated on the outcome with the greatest within-cluster homogeneity and between-cluster heterogeneity (Gale et al., 2016; Reibel, 2011). Mathematically, this is demonstrated through an evaluation of the *closeness* of the clusters using a Within Cluster Sum of Squares (WCSS) analysis, measuring the distances *within* each cluster, and a Between Cluster Sum of Squares (BCSS) analysis, measuring the distances *between* each cluster, respectively. The final clusters generated are mutually exclusive and exhaustive, with every object assigned to a single cluster.

Since the process is dependent on the geometric distance between variable attributes, the k-means methodology is potentially vulnerable to the impact of skewed data distributions, hence the strong emphasis in the data preparation phase (STEP 3).

STEP 5: Repeat analysis (if a hierarchical classification is desired)

In some classification systems, there is a desire to generate *hierarchical classifications* with several cluster levels with varying degrees of granularity, to present an increased flexibility for future applications (Gale et al., 2016). This is the case in the 2011 OAC, which assigns OAs to 8 Supergroups, 26 Groups and 76 Subgroups. This is achieved with a *hierarchical clustering*, whereby clusters are generated recursively by either merging or splitting existing clusters (Gale, 2014). Specific hierarchical clustering algorithms such as Ward's hierarchical clustering algorithm do exist, these have been used to construct classifications in the past. However, in the development of the 2011 OAC, for example, the cluster process described in STEP 4 is simply repeated for the OAs in each of the 8 Supergroups, re-clustering the OAs within the Supergroups into a total of 26 Groups, which are then split again into Subgroups (Gale et al., 2016). This is not an essential step in the Standard Framework if a hierarchical classification is not desired or required.

STEP 6: Link outputs to geography

Though there is a geography assigned to the input attributes, the classification process itself is a-spatial, up to this stage. The geography does not become important until the results are mapped (Harris, 2001). The assignment of the classification outputs back to the underlying geography is essentially the step which makes the *geo*demographic classification geographical, and underlies their core popularity. In a public sector sense, services are consumed by people with needs and requirements. The spatial element of geodemographics supports the identification of sub-populations with differing needs, which can be located when mapped (Longley, 2012).

Such a process can also help when seeking to develop an understanding of the resulting classification. Some attitudes and needs might be inherently dictated by the physical environment in which the individuals find themselves, for instance in broad urban and rural differentiation, or coastal and non-coastal environments, or at a more local level, based on the specific circumstances of the immediate built environment within which one resides (Batey et al., 2008). This kind of contextual understanding can only be developed when the results are mapped back to the geography.

STEP 7: Produce descriptions of the classes (pen portraits)

Often, rather than simply sharing the raw results, the developer will provide an empirical interpretation of the results as an output. As a minimum, the classification groups (or *classes*) are typically given a descriptive name and summary of the *average* resident. In the commercial setting, these outputs regularly also include photographs of individuals, properties and goods intended to visually represent the key attributes of the population assigned to the class, and additional information, primarily related to consumer behaviours, derived from the linking of ancillary data sources. These are often referred to as "pen portraits" and are considered useful in facilitating and supporting subsequent applications of the classification.

Usually, an index is derived based on the average of each of the variables within a cluster (the local mean) against the average across all clusters (the global mean), which is typically standardised to 100 (Harris et al., 2005). Variables in each class which are "over-indexing", i.e. generating an index score of above 100, are deemed to be the key attributes for the class, and underpin the descriptions provided in the pen portraits. The suite of outputs produced by the 2011 OAC, the primary free and open classification in the UK, are presented in Section 3.4 as a tangible example.

This process has some potential weaknesses to which careful attention should be paid. There are some innate risks in identifying clusters based on individual variables, or even groups of individual variables, in this way. These extend beyond the potential introduction of the Ecological Fallacy (see Section 3.3.2). Notably, the outputs could be potentially misleading. Simply because a class has a comparably higher than average prevalence of a particular attribute, does not necessarily equate to the class having an objectively high prevalence in and of itself, nor does it guarantee that it is the most noteworthy attribute in the class.

Moreover, naming the classes can be a difficult task and even controversial if the connotations are negative (Gale et al., 2016). Although the names are only intended to be indicative of the seemingly important attributes of a class, and are to be used in conjunction with the more detailed summaries which should also be provided (Harris et al., 2005), consultations with end-users of the 2001 OAC suggested that the names of the classes often guide the end-user's opinion of the class. Thus, it is imperative to get this step right, to ensure that the names are suitable and do not mislead, and to avoid stereotyping (Brunsdon et al., 2018). Poorly executed naming could have detrimental consequences on the effective use of the classification (Vickers and Rees, 2011). In response to the warnings outlined, the website on which the 2001 OAC was published (the pre-cursor to the 2011 OAC) opted to drop the names from the descriptions (Vickers and Rees, 2011), however, the names were once again adopted in the 2011 OAC.

In whichever way the outputs are derived, it is necessary that they be accessible and easily understood by a broad range of audiences (Batey et al., 2008). This is particularly important where the intended end-users are public sector practitioners who's primary role might not necessarily be as geodemographics experts. In such an instance, the outputs should be driven by the needs of the end-user, and not the developer. The specific needs will differ by user, but the broad needs should be met, and as a priority, the outputs should be clear and intelligible. This might include the recommendations made above, ensuring, as far as possible, that the results cannot be easily misinterpreted. This could be achieved with a clearer visibility of the process and the results.

STEP 8: Validate and enrich the results

It is critical to ensure that the results are sound if they are to be used to inform future applications. The nature of any clustering algorithm is that it will *always* generate a result, however, there is no guarantee that the result derived is meaningful or appropriate, either statistically or in terms of representing real-world structures. Any result is one of many which are mathematically possible, thus exploration and validation is essential. Validation should evaluate both the process and the results, assessing whether the clusters match real-world patterns, and checking that the results are not simply an artefact of the clustering algorithm. This is the foundation of ensuring good scientific analysis (Vickers and Rees, 2011).

However, validating the results of clustering algorithms is more complicated and nuanced than some statistical procedures (Harris et al., 2005). As with all other steps in this process, there is no one-size-fits all for validating the results. Often developers will begin with some rule-of-thumb checks, for instance, evaluating whether the resulting clusters are of generally even size. It is also possible to undertake external validation procedures, employing ancillary data which has not been used to develop the classification, to cross-validate the results and see if they represent meaningful divisions (Vickers and Rees, 2011). This is a popular technique in commercial classifications, with the results contributing towards the creation of the classification profiles, or pen portraits. However, such a technique cannot be used to explain causality, even where relationships are observed (Harris et al., 2005). Such analysis offers simply an identification of patterns among the classes and not an explanation of how or why the classes have formed as they have (Batey et al., 2008), though the ancillary data can add more context to support the development of theory in this vein.

Once the profiles have been generated, some recommend a more tangible, real-world approach, physically visiting an area, consulting with residents, or employing experts with local knowledge, to provide an informed opinion regarding the appropriateness of the result. Vickers and Rees (2011) promote this approach, which they call "ground-truthing", though suggest it is an under-utilised method. This idea, that expert opinions or those based on local knowledge should be validated gives credence to the opinions of LCC, who have highlighted potential concerns with the appropriateness of the available classifications for the city of Leeds, which have in part motivated this study. It is also likely that these concerns of experts are more objective than the concerns of residents. Parker et al. (2007) found in consultations with residents that they were able to remain objective about their neighbours, but less so about themselves, when evaluating classification results. Consequently, the classification development guidelines from Hull City Council (Local Government Association, 2013) (discussed in Section 2.4.4) recommend having classifications developed for public sector use "signed off" by expert "key stakeholders", which might include data teams and teams working in the community. However, Longley and Singleton (2009a) present a dispute to any distrust in public consultation, demonstrating a successful example of such a practice, and strengthening the case for consultation with a wide range of stakeholders *including* the public.

Several mathematical validation approaches are also presented in the literature. Whilst it is not possible to take advantage of some typical statistical validation techniques, such as an analysis of statistical significance or importance measures (Harris et al., 2005), there are statistically founded methodologies for internal validation available to validate cluster groups. The most common is the use of the SED (introduced in Section 3.3.2, STEP 4) as a proxy measure for 'uncertainty'. The SED can be used as a *dissimilarity* measure comparing the attributes found within each individual small-area geography with the average of all of the small-area geographies in the cluster group within which it has been assigned, also known as the *cluster centroid* (Gale et al., 2016). Clusters where the attributes of each small-area geography within are 'close' (with a lower SED) represent a good fit. However, in classifications where the clusters are not close, or small-area geographies are also close the centroids of clusters to which they have not been assigned, the classification is considered to have high 'uncertainty', and is thus a poorer result (Slingsby et al., 2011).

The SED is published as an output in the 2011 OAC to enable this form of validation. Again, the SED is just one available measure of geometric distance. There is a suggestion that other measures, such as Mahalanobis distances, might become a more appropriate measure, favoured for its ability to more capably handle data containing multiple correlated variables (Gale et al., 2016). Thus, there remains scope for exploration in this area, particularly as the data involved becomes more complex.

3.3.3 Inherent subjectivity in the Standard Framework

The ability to outline a standard framework, as per the summary presented in (Section 3.3.1), gives an impression of ease, but in practice, this is not so. The complexity is increasingly evident in the attempt to present a basic summary of each step in the previous section (Section 3.3.2). The framework outlined provides the groundwork for most available geodemographic classifications, as mentioned, but it is clear that there are *many* decisions to be made to implement these guidelines in practice, and the outcome will be dependent on the interpretation and customisation of these steps. This requirement introduces an inherent subjectivity into the process.

Multiple decisions are required at each stage. These can be context dependent and might be dictated by a need to balance pragmatic possibilities with the requirements of the end user or with maintaining consistency with existing practice or traditions. As indicated, these decisions also rely heavily on the experience and preferences of the developer (Brunsdon and Singleton, 2015). Whilst caution is taken in the decisions made, the ever present need for pragmatism characterises the difference, and explains some of the compromises accepted, between geodemographics in theory and geodemographics in practice. The practical challenges, constraints and necessary compromises will be presented further in Section 3.6.

The process as a whole is complex and must be well understood to produce a reliable and meaningful output. The resulting customisation accounts for the competition in the commercial market. Since all classifications are largely derivative of the same process, the differentiation is created in the decisions made throughout, namely, the choice of data, the preparation procedures, the cluster methods employed and the outputs developed (Gale et al., 2016). Though it is understood that the consequence of each decision could affect, or even dictate, the final result (Brown, 1991; Singleton, 2016a), there is limited discussion in the literature regarding the impact of *all* of the decisions that could be made in the design of a classification (Webber, 2004). Singleton (2016a) does offer a review of *some* of the typical choices which are made in the development process, and of their potential impacts, though concludes that it is not pragmatically possible to empirically test the impact of all decisions discussed in a single paper, or even a single doctoral thesis. Whilst this conclusion supports the decisions made later in this chapter to focus on limiting exploration and development to specific elements of the Standard Framework, it serves to emphasise the reliance placed upon decisions of the developer, and the subjectivity which is thus inherent in the development of geodemographic classifications.

3.4 Detailed run-through of 2011 OAC

3.4.1 Background

This process outlined by the Standard Framework above is best demonstrated through an example. As such, this section contains a review of the steps taken in the development of the 2011 OAC. This is made possible due to the freely and openly available assets and transparent, well-documented process. The supporting literature which is published alongside references the decisions made. The data and code, also published, outline the methodology and practical application. Every bit of the release was intended to encourage reproducibility, exploration and support open critique, in addition to enabling the future development of bespoke variants. The following review of these assets demonstrates each stage of the

Standard Framework (outlined in Section 3.3.1) in action, and highlights the frequency and potential impact of the developers' inputs in practice.

In addition to presenting a practical demonstration of the steps outlined, and highlighting the decisions made, this review will present a tangible reference for the discussion of the challenges and constraints which hamper innovation and progress in geodemographics (as per the literature), which is to follow in the final sections of this chapter. Moreover, the 2011 OAC, including the data, methodology and assumptions, will form the foundation of the preliminary substantive analysis carried out in Chapter 4 which will offer a first attempt at a place-specific geodemographic classification for Leeds. As such, a level of detail will be contained in this review of the 2011 OAC to explain the data, methods and assumptions adopted in this future work.

3.4.2 Detailed review

Summary of outputs:

The 2011 OAC, developed by Gale et al. (2016) in conjunction with the Office for National Statistics (ONS) is the most recent in the history of free and open national classifications for the UK. It is an evolution of the 2001 OAC, and was developed with maintaining consistency across the two classifications as a priority. The classification comprises 8 Supergroups, 26 Groups and 76 Subgroups, derived hierarchically from national-level comparison of carefully selected OA characteristics from the 2011 decennial census. It should be noted that all mentions of the census here, and throughout this thesis, collectively refer to the individual censuses of England and Wales, Scotland and Norther Ireland, from which combined census data is compiled, unless otherwise stated. Though the data for England and Wales is published in the same release, the data for Scotland and Northern Ireland are each released separately. However, the datasets are compatible and can be easily linked and used together, enabled by the detailed and clear supporting documentation published alongside each release.

The data included in the development of the 2011 OAC comprises 60 input variables (listed in full in Appendix B.1) from across the five domains inherited from the 2001 OAC: *Housing type*; *Housing composition*; *Employment/Education*; *Socio-economic indicators*; and *Demographic information.*

All of the decisions and justifications for the selection of the input variables and methodologies for the clustering and labelling of the output classification groupings are discussed in the supporting documentation, supporting reproducibility (Gale et al., 2016) and encouraging confidence in the results. The published assets and accompanying documentation for the 2011 OAC detail the entire process, explicitly outlining the methodology, alongside the thought processes and justifications of the decisions made by the classification developers (Singleton et al., 2016).

The 2011 OAC for Leeds at the Supergroup level is displayed in the Figure 3.1, alongside the names of the Supergroups and the percentage of Leeds OAs attributed to each. The classification was developed in R, an open source and free to use software, and all code, data and metadata were published online alongside the outputs.

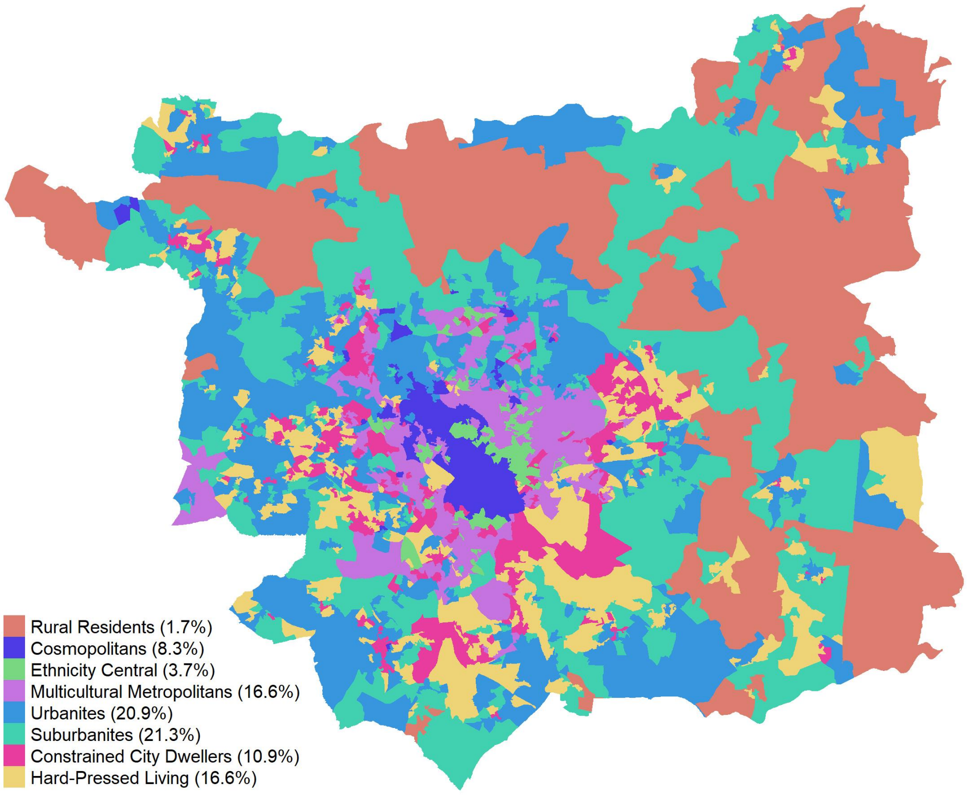


Figure 3.1: Distribution of the 2011 OAC Supergroups across Leeds OAs.

Development process:

The 2011 OAC is a general-purpose national-level classification of the UK, where classes are assigned at OA level. Since this is the lowest level at which the input data was available, there was no scope for generating the classification at postcode or even household level classification to match the granularity of the commercial classifications, or experimenting with input data including both very fine and coarser granularity (as discussed in the previous section).

Aggregate census statistics were selected as a convenient source of reliable openly available data with almost comprehensive coverage across the UK. It is well maintained and easily understood (see Section 5.2.4 for a more detailed discussion). The aggregate data contains population counts for each OA, representing the presence of each of the attributes measured in the census. To protect confidentiality, Statistical Disclosure Controls (SDC) (ONS, 2016b) are applied in instances of low counts, in which some records are swapped between areas. This is applied consistently across the four nations and will have some effect on the data, albeit limited.

The developers expressed an initial intention to include additional data from beyond the census (Gale, 2014), however, this was not pursued due to difficulties obtaining relevant open data with the same granularity and coverage for the whole of the UK, features which were prioritised and not compromised. Instead, recommendations were made in the literature for future adaptations of the 2011 OAC to seek to broaden the data in this way.

A set of 167 candidate variables were initially compiled from the census data, which were subsequently prepared and refined to a final set of 60 variables. Counter to the Standard Framework presented in Section 3.3, in this case, the data was prepared prior to variable selection, dictated by the requirements of the variable selection methods adopted. Both the data preparation and variable selection procedures comprised a series of statistical tests, aimed at introducing objectivity into the decision processes wherever possible. Incidentally, when completed, the results of the tests also revealed a consistency across the different tests, which is presented as evidence of genuine structures in the data transcending any one specific method, and acts as an early validation of the final classification.

The raw census counts were first converted to percentages, representing the presence of each attribute in each geography. A population density ratio was calculated, and a Standardised Illness Ratio (SIR) was derived, taking the census count representing Limiting Long-Term Illness (LLTI) in each OA and adjusting for age variation (as per the discussions in Section 3.3.2, STEP 3). The data was subsequently normalised to account for the non-normal distribution identified. Weighting the variables (as discussed as an alternative to normalisation in Section 3.3.2, STEP 3) was ruled out due to the unavoidable subjectivity of the method (Gale et al., 2016). Three transformation methods were tested, *log-10*, *Box-Cox* and *Inverse Hyperbolic Sine (IHS)*. Similarly, three standardisation methods were also explored for setting the data to a common scale, *z-scores*, *range standardisation* and *inter-decile range standardisation*. Every combination of the transformation and standardisation methods were tested and the results evaluated, both mathematically and contextually, before the IHS and range standardisation methods were selected for final use.

The variable selection process was led with two primary objectives. The first was to maintain some consistency with the 2001 OAC, wherever possible. The second was to achieve parsimony, selecting the minimum number of variables with the greatest potential for differentiating population across the 5 domains (Vickers and Rees, 2007), whilst ensuring a selection broad enough to constitute a fully descriptive general-purpose classification (Voas and Williamson, 2001). Pearson correlation analysis was employed to remove collinear variables, and thus reduce the potential for over-inflating the importance of such variables. This was followed by a cluster based sensitivity analysis, iteratively running k-means analysis, removing a single variable for each run through and observing the impact measured in the total WCSS (introduced in Section 3.3.2, STEP 4).

However, the results of these tests were merely used to guide the selection process. Highly correlated variables (with an absolute correlation greater than 0.6) were either removed or combined into a composite variable where variable shared same denominator, for example, producing age-bands from single age variables. However, some highly correlated variables were retained, at the developers discretion, where their removal was deemed to compromise the priority objectives, outlined above. Similarly, variables with limited or even negative impact in the sensitivity testing were retained if deemed contextually important to retain. This was the circumstance supporting the inclusion of many housing variables, in particular, many of which performed with statistically negative impact in the sensitivity analysis, yet were identified as important representations of the physical infrastructure, recorded as a key domain in the 2001 OAC, and thus important to retain (Vickers and Rees, 2007).

K-means clustering was further used to derive the final classification, employing an optimisation process selecting the best solution of 10,000 runs as the final result, based on the lowest total WCSS. Again, this was selected to offer continuity with the 2001 OAC, which was also derived from a k-means clustering. Further tests were considered to trial other clustering methods, but were ruled out due to restrictions in time and resources, and testing the data preparation and selection phase was given priority.

Using a bottom-up hierarchical clustering, such as Ward's algorithm, was rejected based on concerns that the approach which focuses on clustering the centroids as opposed to the OAs could result in clusters with reduced homogeneity, since the centroids are known to be poor representations of the entire cluster, thus suggesting an awareness that the resulting clusters could have a high degree of uncertainty which should not be propagated, if possible. Instead, a three tier top-down hierarchical clustering was carried out to derive the 8 Supergroups, 26 Groups and 76 Subgroups. Gale et al. (2016) note that this was a largely subjective decision rooted more in tradition, suitability for end-user, and on suitable cluster counts at each level, than on methodological justification.

The cluster results were then linked back to the OAs. The variables were indexed to identify the attributes driving each cluster (as outlined in Section 3.3.2, STEP 7) to support the development of class names and pen portraits, and maps and expert knowledge of local areas were drawn also upon to offer an empirical validation of the results. The literature does note that the national level of the classification was too broad to allow for a close observation of all areas, however, it hopefully suggests that the open and transparent nature of the outputs might allow for critique and open avenues of communication where suggestions for improvements or re-allocations could be made (Gale et al., 2016). At this stage, considerations were explicitly made in relation to this being a classification developed primarily for public sector use.

This was arguably the only stage in which the developers made decisions which were explicitly led by an intention for the classification to support a public-sector focused end-user. The exclusive use of openly available census data, and openly published code, naturally met the transparency requirements of such an end-user. Moreover, the intended purpose of the classification was to offer a general-purpose taxonomy which was likewise not explicitly developed for specific public sector application. However, the development of the classification names and descriptions necessitated such a specific focus. As a consequence, neutral terminology was cautiously selected to prevent the use of language which could be considered disparaging, and final selections were made under consultation.

This summary demonstrates the extent of the time and effort which was invested into the development of the 2011 OAC, and the non-trivial nature of the decisions which had to be made throughout the process. This is reflective of the non-trivial and complex nature of developing any geodemographic classification, even where carrying out the steps outlined in the Standard Framework, with some adjustments and context specific customisation. It also clearly indicates the influence of the developer throughout each stage of the process, and the reliance on subjective decision making, even where conscious efforts have been made to minimise such activity.

3.5 Commercial extensions

As outlined in Chapter 2, the balance of innovation in the development of geodemographic classifications in the UK has for some time been on the side of the commercial products, most notably produced by data-focused companies. The history of this progression is succinctly described by Singleton and Spielman (2014) in their comparative, chronological review of geodemographic activity in the UK and USA.

The official technical documentation provided by the developers, website copy, marketing brochures and supplementary user guides have been reviewed for the seven commercial classifications listed in Section 3.2 to gather as many learnings as possible from recent progress made. The significant developments and relevant features are summarised here. This summary explicitly considers the features that are innovative and unique to the commercial sector and are not present in the development of the 2011 OAC.

3.5.1 Input data

One of the earliest and most significant developments in the commercial classifications, as discussed in Chapter 2, was the inclusion of data which extended beyond the census. As software improved and computational power increased, success came to the commercial vendors who were able to adapt and extend their offerings to include timelier data from a broader range of sources. In contrast to census data, for which the accuracy and predictive powers were considered to be decreasing annually as it became increasingly outdated, and the context was more pertinent to the identification of deprivation, emerging consumer transaction data offered a new perspective with timelier, richer, and more relevant information regarding affluent activities, from which highly desirable insights into consumer behaviours were more readily extractable (Webber, 2004). Information such as house prices and buying habits were perceived as more able to help identify net worth and the spending power and preferences of prospective customers. As such, the geodemographic classification industry thrived through companies who owned and were able to take advantage of this kind of data to develop proprietary products with a commercial focus (Beaumont and Inglis, 1989).

Still today the legacy of this era is evident in the continued dominance of data vendors and warehousers (as outlined in Chapter 2). Though the census still runs through the core of the vast majority of these products, it is extended by a range of valuable ancillary data, both publicly and privately available, the volume and variety of which is as yet unmatched by solely open sourced alternatives, and at a velocity which supports "near real-time" updates (TransUnion, 2018). Moreover, the focus on including ever more insightful, novel and appropriate data is as strong today as it has been throughout the past four decades, as CACI claim to have progressed away from any reliance on the census, developing their Acorn classification from entirely non-census data (Tate, 2018; Sleight, 2014), a trend which seems only set to continue.

3.5.2 Methods

Though the proprietary commercial geodemographic classifications are seemingly largely developed using the same national level methodology as outlined in the Standard Framework, several extensions have been developed. For example, Experian now employ data at a mixture of geographic levels. Both postcode level Electoral Roll data and OA level Census data are included in its development, with different weights applied to each dataset (CACI, 2019). This flexibility enables the use of the more granular Electoral Roll data as a proxy by which to replace some of the traditional census variables, for instance, in checking addresses against who lived there and in surname analysis to act as a surrogate for information about marital status (Webber, 2004). This particular extension has enabled a widespread shift towards the development of classifications at postcode and household levels, common across all of the main classifications on the commercial market. This is a finer geographic level than the available open classifications.

As highlighted in Section 2.4.4, CACI can now develop Acorn with slightly different input variables region to region, to account for cases where a particular dataset might not be available for all areas, for instance, between England and Scotland (CACI, 2020). This has not been done for the purpose of developing place-specific classifications, but does introduce a relaxing of the traditional standards, moving away from the strict use of data available consistently across the entire extent, which could make the shift towards place-specific classification development a more natural progression.

TransUnion also claim to have introduced a flexible modelling system into CAMEO which enables regular updates and adaptations involving new input data (discussed in Section 3.2), which is not currently supported by open geodemographic alternatives, such as the 2011 OAC. This progression enables a widely recommended use of more timely data (discussed in Section 2.4.1). However, whilst the methodology which facilitates this is briefly summarised in the literature at a top level, it simply references the development of "building blocks" which facilitates the swapping of data in and out and from which the classification is subsequently derived. There is no distinct mention of any particular statistical or spatial method employed to achieve this. The documentation is also unclear regarding the selection, suitability, coverage, accuracy or quality of the input data used, or any transformation applied to particular datasets, for instance, methods to attribute sample surveys to the full population, if applicable. The extension offers a new way of thinking about flexibly generated geodemographics, but the proprietary nature is, at this time, inhibiting its direct replication in open geodemographics.

Additionally, as indicated in Section 3.2, several of the proprietary commercial classifications have also begun to develop geodemographics through different lenses to deliver bespoke, "domain specific" classifications. One such example are the Public Sector specific offerings, discussed in Section 2.4.2. The supporting documentation does not indicate whether new methodologies have been developed to support the creation of the domain specific classifications. However, since the development of new methodologies would likely be promoted (albeit guarding commercially sensitive information), it is likely that these have been developed with similar, or the same, methodology applied to context specific datasets, as per the methodology of developing domain specific classifications which have also emerged in academic research (discussed in Chapter 8).

3.5.3 Outputs

Licensed users of the proprietary classifications can benefit from static and interactive visualisations of the classification outputs, alongside detailed classification profiles, as outlined in Section 3.3.2 (STEP 7). These often include supplementary results and information derived by appending the classification to market research outcomes and ancillary datasets, presenting an inferred yet rich, holistic summary of the behaviours and preferences for each classification group.

Whilst these outputs are supported by user guides, attempting to aid in the accurate interpretation of the results, there is very little mention of the underlying methodologies, ancillary data or development process undertaken to deliver the resulting outputs. As discussed, the process of developing the classification is itself black-boxed. This is the same for the outputs produced. Though the websites offer some "technical" information, the details contained are somewhat vague, for instance, general sources of ancillary data included are listed alongside some domain headers, such as the census and *data.gov.uk* alongside "Shopping behaviour" data and "Income and savings" data. Details of specific variables are omitted. There are risks in developing the outputs in such a way. Obscuring their creation from view might lead to misunderstandings, and ultimately, their misuse. These risks will be discussed further in the next section.

3.6 Criticisms, challenges and constraints impeding practical developments

Though the framework is both well established and widely adopted, there is a great deal of discontentment with the process which have led to widespread calls for improvements to be made. Much of the disparagement dates back over many years, yet has led to little tangible action by way of practical advancement. However, academic interest in the development of geodemographic classifications was somewhat rejuvenated early in this century, as the appetite for their application in the public sector fortuitously coincided with the emerging trend for generating reproducible research. The growing culture for developing public policy in an open and transparent environment, and substantial advancements in the availability of data and low cost technologies with increasingly sophisticated GIS technologies (see Section 2.4.4). The resulting landscape led to several academics reclaiming responsibility for the development of open geodemographics capable of challenging the commercial dominance.

Nevertheless, the contribution of these academics has not yet been the generation of a new open geodemographic classification beyond the scope of the 2011 OAC. It has, however, led to the publication of a rich, largely theoretical, commentary on the contemporary status of geodemographics and anticipations of its future direction. Whilst these discussions have been essentially aspirational, including bold predictions for the future advancement of geodemographics (discussed in Section 2.3.2), and the tone is largely positive, the optimistic future-outlook is grounded in a counter-narrative, a pragmatic realism offering tangible explanations for the limited progress. This narrative reflects on the stagnancy in innovation experienced in recent decades, raises criticisms of the framework presented in Section 3.3. Moreover, a suite of "Grand Challenges" (as coined by Longley (2007)) which must be overcome in order to achieve the advancements prophesised are documented, each presenting a barrier to potential progress. These challenges require acknowledgement, as a minimum, or addressing, if possible, to raise the standard of the current outputs and help move the field forwards. To date, the potential advancements predicted in the literature throughout the past two decades are mostly yet to experience practical fruition outside of the commercial Geodemographics Industry, at least in a widespread sense. Though there has been practical research conducted, these have not led to the development of a widely used open geodemographic classification, or classification framework, beyond that of the 2011 OAC. Thus, it is clear that many of the documented challenges remain relevant today, unaddressed and still impeding development in the field. A holistic consideration of these challenges is therefore necessary to understand where geodemographics is at, particularly in academia and the public sector, and to consider how to advance the practice to reap the potential rewards which have been proffered, whilst highlighting the possible limitations which continue to restrict the development of new open geodemographic classifications.

Thus, this section draws from Longley's (2007) initial presentation of the perceived Grand Challenges, and from additional concerns presented across the wider body of literature, to summarise the barriers and limitations which are most frequently presented or are posing the greatest threat to future developments. These criticisms are presented here in four broad themes. These are raised, not to warn against the use of geodemographics, but to highlight the areas which require development and attention to enable future innovation. Many are challenges which are broader than geodemographics and are either typical in data analysis and data science more generally, or in other elements of GIS, but others are more specific to the field (Longley, 2012).

Whilst Section 2.4.3 addressed criticisms raised against the blind adoption of commercial classifications, and the limitations of using those developed in black-boxed environments, this section extends beyond these concerns to review criticisms of the practical framework for developing classifications, more explicitly, and the restrictions which have seen its fall from a tradition of developing cutting-edge techniques, to a practice reliant on a long established, but heavily criticised, standard framework. The following discussions will critically review the literature published throughout this period to consider the challenges more closely whilst also reflecting on their continued relevance. In doing so, this section will also explore whether the framework itself remains relevant and appropriate, whether it is meaningful both statistically, geographically, or in a real-world sense, or whether, in an era of unprecedented development in Machine Learning and Artificial Intelligence, the process is outdated, and if there might be more cutting-edge, twenty-first century approaches, or methodologies, which ought to be taken advantage of and which might address, or circumvent, the challenges which have presented such robust barriers in progress until now. Outcomes of these explorations will be used to develop and expand the research agenda for this thesis. Whilst the core aim of generating an open place specific geodemographic classification was set in Chapter 2, this review of the critical literature will inform whether (and which) other improvements will also be sought in the development of this classification generated within this thesis.

3.6.1 Theme 1: Outdated thinking and limited innovation

At the turn of the century, Openshaw (2001), a significant academic contributor to geodemographic development, and an equally strong proponent as a critic of the practice, offered one of the most scathing reviews of geodemographic classifications to date. Speaking in the context of employing geodemographic targeting in commerce, he labelled the framework a "simple, sloppy, sixties system" and a "dumb, old-fashioned, legacy technology". Moreover, he questioned whether the "poorly developed modelling system" ought to be improved, or instead, laid to rest and replaced with a more advanced and more appropriate, twenty-first century technology of "next generation systems".

At this time, investment, interest and thus tangible development had been almost exclusively the realm of the private sector for over twenty years (see Chapter 2). Academic disinterest led to a stagnancy in non-commercial innovation (Singleton and Spielman, 2014), and as such, though open geodemographics did see the generation of the 2001 OAC when aggregate census data was made freely available (and the subsequent update following the 2011 census release), the sophistication of the commercial offerings were unmatched by non-commercial alternatives. Moreover, whilst there have been some developments derived within the commercial Geodemographics Industry since its inception in the 1970s (see Section 3.5), these have been limited in both quantity and scope, and as mentioned earlier, evolutionary rather than revolutionary (Gale et al., 2016).

Though this might imply a level of success in the current classification offering which does not warrant improvement, the wealth of criticisms made in the academic literature suggest that this is unlikely. Instead, it is more likely that potential developments which might benefit the accuracy or relevance of the classification itself might, in a commercial sense, offer limited returns on investment. As is the nature of the private sector, commercial requirements might not align with the need to develop the 'best' or most accurate classification, per se, but instead value a product which satisfies the end user and maintains a level of customer loyalty (discussed further in a consideration of success measures in Section 3.6.2). As such, financial and time investments will likely be made with this objective in mind. Consequently, as per Openshaw's critique, the Standard Framework presented in Section 3.3.1 has experienced limited development, despite substantial technological advancements throughout the interim years.

Openshaw himself conceded in the same critical review that, "even old-fashioned geodemographic targeting is better than no targeting" (Openshaw, 2001). However, it is, again, important to note that this remark was made in relation to commercial applications of the existing geodemographic classifications. It is difficult to know whether he would have similarly concluded in the use of geodemographic targeting in a public sector context, where the stakes are potentially much higher, and where accuracy could be more critical (Longley, 2012; Singleton, 2016a). In the development of public sector classifications, it is important to reconsider the implications of Openshaw's concerns, to ensure that any future developments do not continue to extend the life of legacy systems which, if outdated in 2001, are likely increasingly so today, particularly as other fields leap ahead of geodemographics in their adoption of cutting edge Machine Learning methodologies (Singleton and Longley, 2009b).

Though (Openshaw, 2001) contemplated, almost two decades ago, whether the common practices for deriving geodemographic classifications required an overhaul, highlighting his concerns with the methodology even then, the practice which is still largely unchanged today, continues to age. Consequently, Dalton and Thatcher (2015) express similar concerns, presenting geodemographics as an antiquated approach, and suggesting that the next phase of development might adopt the current practice as a precursor upon which to develop more advanced techniques, integrating the big spatial data now available to more intelligently derive insights about populations and target individuals.

However, whilst Dalton and Thatcher (2015) make loose references to contemporary "tech companies" and their transformation of targeted marketing enabled by their access and use of the spatially aware data reserves, the discussion of what these advancements could practically look like, and how the open geodemographics sector might learn from them to underpin their own transformation, is limited. To take complete and meaningful advantage of the growing availability of data, researchers will likely need to adapt their practices, closely considering practices currently more commonplace in other fields, such as Computer Science, namely Machine Learning (ML) and data visualisation, and make use of the techniques found there (Arribas-Bel, 2014).

Rather than entirely replacing the legacy system with a whole new, modern approach based on opportunities afforded by more recent statistical and technological developments, more commonly, the trend within the literature has been more in-line with Openshaw's iterative improvement approach, retaining the strengths which underpin the popularity of the practice whilst introducing advancements. This is evidenced in the commercial sector throughout the advancements listed in Section 3.5. In academia, studies from Singleton and Longley (2015, 2009a) and Burns et al. (2018) demonstrate recent examples of researchers identifying and exploring adaptations of the Standard Framework to shift to a local-level extent, generate a domain specific classification reflecting groups based on a single issue and develop a classification for individuals, respectively.

This thesis itself presents the next phase of *place-specific* geodemographic classifications with an intended public sector primary end-user base, but with a consideration for transferability into other sectors. The work presented throughout Chapters 4-8 offers further customisations of the Standard Framework, evolving the traditional approaches to several of the steps listed in Section 3.3.1, and in the research and studies which have come before. Whether there is a more appropriate approach available in the technologically advanced times in which we find ourselves today, which does not rely on the legacy framework at all (as per Openshaw's (2001) musings), remains to be seen. However there are definitely advancements proposed in the literature which have yet to be developed and which might offer improved solutions and more appropriate classifications than those currently available. Thus, this work aims to test some of these theories first, including, primarily, shifting to a local-level classification and expanding the data to include non-census variables. This is a more pragmatic and realistic approach in the scope of this project than condemning the practice as it currently stands and starting again, looking to discover an entirely new system.

3.6.2 Theme 2: Pragmatic divorce from theory

Whilst the pre-cursors of geodemographics took a journey through sociological theory in the era of Urban Ecology, introducing a particular consideration of the meaning behind the societal structures identified, the development of modern geodemographics has been somewhat divorced from theoretical underpinnings for many years (Beaumont and Inglis, 1989; Alexiou and Singleton, 2015), instead, employing an approach which has been fundamentally data-led (Singleton and Longley, 2009b).

This echoes the experience of many fields adopting Data Science more broadly, where an increasing dependence on data-led practices has often occurred at the expense of theoretical development (Singleton and Arribas-Bel, 2019). This is particularly promoted in the adoption of unsupervised learning methods, as in geodemographics, which are run without a-priori expectations and which thus behave counter to the tradition of testing a hypothesis, or seeking understanding of causality (Harris et al., 2005). These are, however, not new criticisms of geodemographics. Theoretical concerns plagued development in the era of Factorial Ecology, subsequently resulting in the withdrawal of study from academics who were sceptical of the fundamental assumptions being made (see Section 2.3). In an effort to avoid a similar fate, modern open geodemographics ought to ensure that research in the field is grounded in relevant and appropriate theoretical constructs.

Disregard for theoretical principles of geography

There is a concern that the importance of geographic principles in the development of the methodologies underpinning the Standard Framework might also be relegated in this shift in practice. As spatial data generated from commercial transactions, social media communication and local government activity is made available at unprecedented speeds and volumes, encouraging its adoption by users beyond those with GIS specialisms, there is a risk that considerations for the unique epistemology required to suitably handle such data could be somewhat lost. This is a problem broader than geodemographics and is inherent in the wider practice of spatial Data Science in general, credited to increased dependence on the data and data-led practices coinciding with a decreased emphasis on theory development, which once prevailed (Singleton and Arribas-Bel, 2019).

In response, geodemographic researchers particularly promote the importance of geographical theories in the literature, with Longley (2007) proclaiming spatial literacy as essential in analysis of spatial data, which comes with unique problems requiring a specialised skillset. Whilst spatial analysis in the consumer domain is not seeking to derive geographic understanding, as per some of the Urban Analysis pre-cursors to geodemographics, but instead are looking to offer a means of better targeting consumers (Dalton and Thatcher, 2015), there are concerns that critical geospatial principles are not being afforded the necessary attention.

In some instances there seems to be a disregard for the spatial element of the process entirely. Notably, the clustering methods employed in STEP 4 of the Standard Framework are applied a-spatially. Fundamentally, this is not a geographic procedure until the result is mapped back onto the small-area geographies for which classifications have been derived (Harris, 2001). In so doing, it becomes difficult to be sure whether specific geographical effects, which are known and understood by geographers and GIScientists have been appropriately accounted for, or at least given the necessary considerations (Singleton and Arribas-Bel, 2019).

There are several more tangible concerns with regards to the spatial literacy required in other stages of the framework. For instance, it is essential that geodemographic classification developers are also aware of the potential pitfalls of well-known geographic principles such as the Ecological Fallacy and the MAUP (discussed in Section 3.3.2, STEP 1), particularly in the early stages when defining the geographic boundaries and aggregating the data. These are dangers ever present in the use of spatial data, which are yet to be solved and thus necessary to be understood (Dalton and Thatcher, 2015). An understanding of less widely discussed dangers, such as Ecological Correlation, which cautions against attempting to infer correlations from aggregate to individual level, is also advised (Robinson, 2009).

Effects of Ecological Correlation could be particularly pertinent in the interpretation of geodemographic classification outputs. Inferring characteristics of the classification group back to the small-area geographies in the interpretation stage could re-introduce the potential for falling foul of the Ecological Fallacy, particularly in the development of pen portraits or population profiles. There will be some heterogeneity within clusters, and thus, not all small-area geographies will map directly to the characteristics of the cluster average (Harris et al., 2005), though improved discriminatory power of a classification will naturally reduce the potential impacts (Burns et al., 2018).

Therefore, caution should be taken to account for the effect of these principles at various stages throughout the process, including in the data preparation phase and in the classification group naming stage, where there is a further risk of propagating the Ecological Fallacy (Gale et al., 2016). Thus, these are important considerations for developers and end-users, alike. Particularly in the development of the proprietary commercial classifications, it is difficult to assess whether, or ensure that, developers have appropriately generated the classifications with these urban methodologies in mind (Singleton and Arribas-Bel, 2019; Reibel, 2011).

No standard definition or measure of "success"

Challenges associated with evaluating a geodemographic classification's *success* are well documented. Indeed Openshaw (2001) claimed that it is rarely possible to establish whether the result achieved is better than random. Whilst others are not quite so pessimistic, the literature does highlight a swathe of complexity in this regard. Primarily, difficulties exist in developing a definition of success, creating clear ideas of what it looks like, and developing methods of measurement. Moreover, *validation* of the model and *success* are regularly conflated across the literature, further complicating the discussion. One might consider whether the predominant concern should be in generating the optimal model, in deriving the most accurate result, or in creating an output which will add the most value in future applications. For example, one might argue that is is good enough to develop a classification which is fit-for-purpose in an application, which might not be mathematically optimal, but works to an acceptable level of accuracy for the intended end-use. Whether these intentions are mutually exclusive, and how one might tangibly measure the achievement of each, are also necessary considerations.

These considerations are particularly important in the development of classifications to be used in the public sector, where the notion of good *enough*, might be substantially different to in commercial applications. There is a sense that success is often considered financially in the commercial sector, in terms of customer loyalty and continued licensing (Longley, 2007). End-users, of both paid-for and free classifications, might also develop their own measures of success in terms of the benefits gained in subsequent applications of the output. Both are inherently vague methods of evaluation, dictated by the level of accepted accuracy. Critics suggest that if, as is the case in many applications in the private sector, the intention is to add any amount of additional insight which could result in an uplift in sales, the boundary of success is lower than, for example, in the application of the output for predicting requirement of medical resources. In the latter example, the stakes are potentially higher and might thus demands a higher threshold of success (Beaumont and Inglis, 1989; Longley, 2005). Again, this relates back to the discrepancy in motivations discussed in Section 2.4.3.

Alternatively, many academics propose notions of success in a far more abstract sense. This could be theoretically, in terms of the identification of real-world divisions as per Singleton and Spielman's (2014) definition. Otherwise, it could be mathematically, in terms of the identification of optimal cluster groupings which perform better based on measures of similarity and dissimilarity. This is seen in Singleton and Longley's (2015) development of the LOAC, and underpins the iterative clustering methodology in the development of the 2011 OAC (Gale et al., 2016).

Fundamentally, analysts are looking to capture *similarity* within classification groups. In a real-world sense, these are highlighted through shared characteristics (Voas and Williamson, 2001). Such a notion is rooted in Tobler's first law (quoted in Section 2.2), cited frequently by academics as the founding concept underpinning geodemographics (Harris et al., 2005), which speaks of an association between the similarity of things and their physical closeness. However, this somewhat ambiguous statement seems to be an over-simplification of a complex concept (Voas and Williamson, 2001). The concept of similarity in this context is sparsely defined in either the theoretical or practical sense. With no defined metrics of similarity, either quantitatively in terms of measurement, i.e. *how* similar, or qualitatively, i.e. in *which* dimensions similarity is expected, the tangible interpretations of the notion have manifested in a series of different ways.

For example, it is unclear in how many dimensions one might expect objects (e.g. small-area geographies) in a group to be similar to consider the grouping a success, or what the measure for similarity might be, if there even is a universal objective measure or if it is context dependent. Moreover, whether the shared characteristics identified by a geodemographic classification should relate to the demographics of the individuals in the small-area geography, or to their attitudes, behaviours, preferences or observed activity, is undefined. Such a decision is regularly dictated by the context in which the classification is later applied. This could be a consequence of the unsupervised machine learning methods employed and the lack of hypothesis to test against when attempting to validate the result (Harris et al., 2005).

Despite these challenges, there are a suite of proposed methods for measuring and validating the success of a classification. The validation stage in the Standard Framework (Section 3.3.2, STEP 8) offers tangible techniques for assessing the classification performance, though each differ in their intentions, exposing the inconsistencies in the standard approach. For example, a mathematical evaluation of the closeness of the clusters based on similarity and dissimilarity measures is suggested. This is a clear validation of the statistical process, but pays no heed to the real-world meaning of the groupings. There are also no defined guidelines dictating the necessary measure of closeness which must be achieved to be considered a successful result. The results are thus open to interpretation based more on their utility than any technical standards (Parker et al., 2007). Additionally, it is not possible to observe or participate in this form of validation for the black-boxed proprietary classifications developed in the private sector (Harris, 1998).

Empirical analysis involving the application of ancillary data to the output is also frequently employed to identify whether patterns emerge, indicating shared characteristics within classification groups (Singleton, 2010) (discussed in more detail, supported by examples, in

Chapter 8). Alternatively, Singleton and Longley (2009b) and Vickers and Rees (2011) recommend consultations with residents, which Vickers and Rees refer to as “ground-truthing” the result, evaluating how well the output reflects real-world phenomena from the perspective of the lived experience and granting increased autonomy to the end-user (discussed in Section 3.3.2, STEP 8). In theory, the optimal solution might be one which succeeds in each of these proposed measures, however, there is limited guarantee that the optimal cluster result will identify meaningful real-world divisions.

Whether these techniques identify success is dependent upon the definition, for which there is no clear consensus (see Section 3.6.2). Moreover, each of the proposed solutions are inherently subjective, and even the ‘objective’ statistical measures rely on a subjective interpretation. In seeking to assign meaning to the groupings, it is easy to see how such an approach might be open to identifying and attributing meaning or interpreting phenomena which might not tangibly exist, for instance, in deriving findings which are simply artefact of the statistical processes, or of decisions made by the developer, or which are biased, influenced or led by pre-existing expectations of the population (Singleton and Longley, 2009b; Vickers and Rees, 2007). This was the fundamental criticism of the work undertaken by Shevky and Bell (1955) over 70 years ago, which weakened trust in classifications developed at the time (see Section 2.3.1). Moreover, in consultation with residents, Parker et al. (2007) identified a phenomena of individuals disassociating with output results themselves, but identifying the highlighted attributes in their neighbours, as noted in (Section 3.3.2, STEP 8). Thus a balance of expert and local opinions might preferably be sought.

It thus remains difficult to tangibly and definitively define success in terms of a geodemographic classification, or to assign meaning to the results in a way which supports interpretation and useful application by the end-user. It is evident that there is no standard, objective method of assessing how good a classification is, or even in defining what it means for a classification to even be good. Moreover, it is an aspect of the development process which has received limited attention for decades (Alexiou, 2017). The literature focuses more on how to employ geodemographics, than if and how they work, who for, and how success might tangibly be measured (Webber, 2004). Future research might seek to re-consider these concerns in the development and evaluation of the next generation of classifications (Harris et al., 2005).

Denunciation of traditional general-purpose classifications

Finally, there is a lively debate in the literature regarding both the *appropriateness*, and increasingly, the *necessity* for general-purpose classifications, particularly in comparison to domain specific classifications, the development of which is becoming more accessible (Longley and Singleton, 2009a). This will not be discussed in detail here, since a detailed discussion on this topic is to follow in Chapter 8. However, many of the key arguments of the debate in favour of domain specific classifications, particularly those relating to its appropriateness, criticise the theoretical principles underpinning the core notion of general-purpose classifications, and challenge the contextual and real-world relevance of classifications developed without an a-priori purpose (Voas and Williamson, 2001). Thus the debate warrants at least a brief mention within this general theme relating to the divorce from theory.

3.6.3 Theme 3: Outdated, untested or undefined methodological approaches

As per Openshaw's (2001) critique, elements of the Standard Framework are seemingly outdated and warrant more exploration in light of current technological advancements. Particularly, the literature highlights a need for improvements in the methodologies used for variable selection, increased consideration of the data transformation practices and broader testing of alternative clustering methods, as summarised below.

Requirement for improved variable selection methods

Conscientious preparation of the model upfront, particularly in terms of choosing suitable input parameters, can assist generating better outcomes. Specifically, the considered selection of appropriate attribute variables for input into the model, and associated data, has for a long time been widely acknowledged as important in determining the success of the resulting classification (Rees, 1972), for example, in the development of more meaningful and appropriate classifications, where the societal divisions identified are truly driven by the data and variables employed (Singleton and Spielman, 2014). The structures of most interest in the data might be defined by just a subset of the input variables. In such a case, the other variables might be useless, or worse, introduce harmful noise, detracting from the ability of the method to best decipher the genuine structures (Maugis et al., 2009). Thus careful and intelligent variable selection is essential.

This is arguably the most subjective element of the process, however, clear reasoning behind variable selection is not always present, limiting the potential for external appraisal. This is often a consequence of classification development primarily occurring within the geodemographics industry, where commercial sensitivity has led to vast amounts of the process being hidden from observation in each case (Longley, 2005, 2012). Some objectivity is introduced in the development of the 2011 OAC by the use of sensitivity testing methods (Gale et al., 2016), however, the results of this objective analysis are taken as simply guidance and are overridden where decisions might conflict with priorities (see Section 5.3.2).

Burns et al. (2018) stress the necessity for ensuring that the variable selection process is not arbitrarily dictated, potentially degrading the meaning of the final classification. Arbitrary variables might identify *some* patterns in residential distributions, though these might not align with the core classification objective. The variable selection stage might thus also be used to increase the relevance of the classification in terms of its *purpose*, for instance, in the case of a domain specific classification, this phase should support the selection of relevant attribute variables which are going to generate meaningful classifications, in the context of the specified domain.

A review of the promotional materials in the commercial literature released by the commercial classification vendors suggests a broad generality in the types of data and variable domains incorporated into the generation of general-purpose classifications. These are typically demographic variables, though are increasingly supported with some alternative behavioural and attitudinal data, often drawn from transactional sources (Singleton and Spielman, 2014). Though this is true for general classifications, many suppliers of geodemographics also offer a suite of context specific classifications, as demonstrated in Section 3.2, which are each derived from a more selective set of idiographic variables.

The number of variables selected might also impact the output as much as the context of the variables. The inclusion of too many variables has the potential to reduce the ability to identify similarity between areas. Conversely, the inclusion of too few might limit the ability to derive a rich, holistic picture. Thus a balance must be struck (Burns et al., 2018).

Increasingly, as societies grow ever more complex, and data availability simultaneously introduces seemingly infinite potential, the necessity for sophisticated data and variable selection methodologies is magnified (Reibel, 2011; Longley, 2012). Such methodologies necessarily need to become more discriminatory. Data should be interrogated for its usefulness, not simply adopted because it is available, or easy to incorporate (Harris, 1998). This is particularly true in the development of domain specific classifications, or, as is the focus of this thesis, in the development of place-specific classifications. In the latter, the identification of local populations might benefit from the use of bespoke sets of attribute variables informed by a deeper understanding of the factors which uniquely drive social disparities in the local context. Therefore, the development of such improved variable selection methods will become another focus of the research agenda underpinning this thesis.

Implications of complex data transformation options

The use of different data transformation methods also has the potential to effect, or change, the resulting output. As such, these decisions warrant careful consideration. Again, there is no standard practice adopted across all classifications. The challenge presented by the complexity of transformation methods is evidenced in the considerations made when transforming, or normalising, of the data.

As mentioned in 3.3.2 (STEP 3), skewed data can negatively affect the cluster assignments in the common cluster methods (Gale et al., 2016). However, it is not always desirable to normalise input data to the mean, particularly where to do so could remove genuine features which are worth understanding and including (Singleton and Longley, 2009b), such as those reflecting unique local phenomena (Singleton and Spielman, 2014), or which could be theoretically interesting (Reibel, 2011). These phenomena could be exactly the local distinctions which differentiates communities well and could thus underpin a good geodemographic classification.

As an alternative, some developers opt instead to apply weights to the data to adjust the influence of each variable, instead of transforming the data. However, this could introduce new concerns, particularly regarding the potential for subjectivity or bias in the application of the weights, since these are often empirically chosen (Singleton and Longley, 2009b). It *is* possible to run tests of different weightings, comparing the results against ancillary data to see the impact of the different decisions on the discriminatory power of the result (Webber, 2004) to regain some objectivity, however, this is still not the standard practice.

The developers of the 2011 OAC (Gale et al., 2016) acknowledged the potential impacts of the transformation processes in opting to select the procedures applied to their data based on multi-method testing, analysing the influence of several transformation procedures (detailed in Section 3.4.2). Such an approach might offer some reassurance in supporting the complex decisions made in this stage of the classification development process.

Recommended exploration of alternative clustering methods

As mentioned in Section 3.3.2 (STEP 4), k-means clustering is adopted as standard in the development of most geodemographic classifications (Alexiou, 2017). Although it is difficult to irrefutably verify this claim with regards to the proprietary classifications developed in closed environments in the private sector (Longley, 2007).

There is limited evidence in the literature of the selection of k-means being based on thorough research of the possible options available in terms of clustering algorithms, or multivariate analysis more broadly. One might thus infer, as per Openshaw's (2001) comments, that this decision is one which is guided primarily by ease and tradition, and which has been made somewhat uncritically (Longley, 2012). Despite the assertions that such decisions warrant repeated review, Openshaw (2001) is specifically sceptical that 'better' methods of classifying produce results which are substantially better. However, there is a hint of conjecture in this claim, and specific research to test such a hypothesis might be warranted.

One development which Openshaw (2001) does champion is a consideration of the use of fuzzy clustering methods, which allow for the allocation of small-area geographies to more than one cluster group, taking the uncertainty of the cluster solution (discussed in Section 3.3.2 (STEP 8)) into consideration. A basic, manual demonstration of such an approach was employed by Charles Booth, in his early descriptive map of poverty in London (Harris et al., 2007) (see Section 2.3.1). Despite the potential for more nuance in a fuzzy result which accounts for uncertainty (Slingsby et al., 2011), such a practice is less common today, where simplicity is favoured, and classifications are designed to generate mutually exclusive groups.

Additionally, cluster methods are inherently aspatial. The effect of this is to consider similarity only in the attribute space, which has the potential for reducing local sensitivity (Alexiou, 2017). If Tobler's law is the theoretical underpinning of the methodology, as it is so often cited to be (Harris et al., 2005), one might wonder whether geographical distance might necessitate a role in the methodology. There is thus some scope to develop methods which build some geographic context into the methodology, to develop *location aware geodemographics*. This could involve the application of a spatial weighting to the variables based on neighbourhood relationships, prior to running the cluster analysis (Adnan et al., 2013; Alexiou, 2017). In doing so, the idea of 'close' will need a more tangible definition. The research of Alexiou (2017), in particular, acts as a recommendation for future research focusing on the exploration of alternative cluster methodologies to extend their exploration in the direction of location aware geodemographics.

Some critics dispute more broadly whether clustering methods in general should be the go-to methodology for deriving the classification groupings at all (Reibel, 2011). As outlined in the previous chapter, though the process detailed in the Standard Framework is well established, a range of other methodologies were trialled throughout the long history of research commonly regarded as precursing Modern geodemographics (see Section 2.3.1). The exploration of alternative methods, or even alternative cluster methodologies is not prioritised in this thesis, since a single thesis can not address the impact of all potential decision processes which could be made in the development of a geodemographic classification (Singleton, 2016a), and the exploration of other elements have taken priority. Moreover, many of the decisions made in this thesis extend a consideration to the potential for replication of the output. To this end, the ease of application in the incumbent methodology, namely k-means clustering, does present a particular benefit in this case. However, it remains vital to have an awareness of the criticisms which have been raised against this methodology, and its potential limitations, particularly when it comes to interpreting the classification output.

3.6.4 Theme 4: Practical challenges constraining development

Some of the challenges raised in the literature, which have thus far constrained development, are simply practical. One of the most crucial, in the context of the aims of this thesis, are the challenges presented in the effort to adopt increasingly non-census data in open geodemographics.

Though the literature discusses the broadening availability of open and public sector data, also presented here in Section 2.4.1, in the UK, the public sector data infrastructure and open data landscape are not yet developed enough to offer alternative data sources reliable enough to reduce the dependence on the census (Singleton et al., 2016). There is thus quite some way to go to completely end the use of census data, as has been achieved in the private sector in the development of Acorn by CACI (Sleight, 2014). In comparison to the availability of data in the commercial sector, the public sector availability is extremely limited. Moreover, the presence of the data is not the only concern. Often, even the available data is poorly stored, maintained, documented or embargoed by complex sharing agreements. Examples of good, positive uses of this data might encourage a relaxing of the fears which have necessitated such caution (Longley, 2005), however, it is difficult to produce such examples without first gaining access to the data. Nevertheless, the government initiatives discussed in Section 2.4.1 could play an increasing role in opening up these forms of data going forward and in improving its quality, in addition to organisations such as LIDA and the CDRC, which has played host to the work conducted in this thesis, and collaborations with LAs, as demonstrated with the partnership with LCC here.

Yet thus far, the commercial sector have been in a position to respond more quickly, to take advantage of even publicly available and open data in the development of geodemographics (Singleton and Longley, 2009b), likely benefiting from better infrastructure and processes, more investment and an increased understanding and a more longstanding respect of the value of data. The commercial sector might have also been in a better position, or more ready to accept the trade-off for the less easily quantifiable benefits which come with an investment in data. The public sector, which burdened by its accountability to the tax payer and restricted by budget cuts, might be encumbered by a reluctance to invest as readily in the groundwork needed to support future innovation.

Evidence of these practical constraints will be presented, and discussed in detail, in Chapter 5, where efforts are made to begin to adopt novel, non-census data into the Standard Framework.

3.7 Twenty-first century approaches: Evolving and extending beyond the Standard Framework

In response to Openshaw's (2001) reference to extending the current practices of geodemographic classification development beyond the traditional, legacy systems, and bringing them up to date with more contemporary approaches, the following section considers the recommendations which have been made to extend the field in this direction. The body of literature including this type of discussion is far more limited than has been dedicated to describing and considering adaptations of the traditional approaches. There is thus a trend for continuing the tradition of incremental, evolutionary improvements, rather than offering approaches for a complete revolutionary overhaul of existing practices, as noted by Gale et al. (2016).

Their own study, the development of the 2011 OAC, itself promotes progress in such a manner. Currently, as outlined throughout this chapter, although there is a standard framework upon which the development of most classifications are based, the framework itself typically remains the property of the developer, and the standard practice is to share the outputs produced in Section 3.3.2 (STEP 7). The consequences of this, as discussed at length, include an inability to test and re-produce the result and to make regular amends, for instance, to manually update with additional data, more up-to-date data, or to be able to adapt to produce more bespoke domain specific outputs. The decision of Gale et al. (2016) to openly publish the 2011 OAC, including all source code, input data and supporting materials, were made with these facilities in mind. They were explicitly keen to support exploration and adaptation and increase the autonomy of the end-user in their adoption and application of the classification. Though original outputs are produced, users are encouraged to adapt the material to develop their own extensions.

However, the possible adaptations might be somewhat restricted in practice. For example, given its reliance on static input data which has just a ten year update cycle, data updates would require access to new and appropriate datasets, the availability of which at the required national extent and OA level might be limited. There is, however, scope to extend or develop the methodology in several other ways. For example, it is possible to generate a version of the 2011 OAC at a place-specific extent, as demonstrated in the development of the LOAC and here in the next chapter. This shift in extent could in turn open up access to new appropriate datasets. In this direction of development, the code and documentation of the 2011 OAC offers a usable Standard Framework which can be customised, making the development of new classifications more accessible. Moreover, the outputs offer a good "benchmark" against which outputs resulting from customisations of specific elements, be it those developed with new data, a new geographic extent, new clustering techniques, or new variable selection procedures, might be compared to and validated against.

Similarly, Gale et al. (2012) have extended this idea to develop "GeodemCreator", a tool to support the development of bespoke classifications (introduced in Section 2.4.4). Whilst this tool supports the inclusion of custom data, its focus on an end-user who is not an experienced geodmeographic classification developer means that it is otherwise largely inflexible.

Nevertheless, the acknowledgement that future researchers might want or even need to adapt the 2011 OAC, or have a convenient tool for developing their own classifications, is also an acknowledgement of the evolving notion in geodemographic classifications that there might not be a one-size-fits all solution. Singleton et al. (2016) similarly identify that it might become increasingly appropriate to support a shift towards end-users as developers, enabling them to create and build fit-for-purpose classifications *on-the-fly*, encouraging a problem-centric approach. These discussion of on-the-fly, or *dynamic* geodemographic classification development currently feel the closest to revolutionary developments of the field, despite the notion still being underpinned by a version of the Standard Framework.

The extension of more flexible approaches for developing classifications might also support the generation of more *real-time*, or at least more regularly updated, classifications, as identified in the commercial sector as a priority already (see Section 3.2). Such a development in the open geodemographics sector are particularly attractive, where there is a concern that the accuracy and appropriateness of the current census based classifications decreases over time (Singleton et al., 2016). Such a shift might also introduce a new dimension to the potential of geodemographic classifications, one which enables the consideration of life-stage trajectories, again, a current priority of the commercial sector, particularly TransUnion (see Section 3.2). This kind of practice could evolve to incorporate the consideration of two speeds of turnover in an area, overlaying residential turnover a-top of the more stable and ever-present underlying social structures (Longley, 2012), introducing a more dynamic element into geodemographic classification.

However, a dynamic approach to classification development might not be favourable to all end-users, who might be reluctant participants in the creation phase (Singleton et al., 2016). Naturally, the extensive detail outlined in this chapter regarding the essential considerations to be made by classification developers, particularly the decisions which are crucial in ensuring an accurate, non-biased result, demonstrate that any shift in this direction must support end-user-developers to build classifications carefully, and in a manner considerate of all of the potential pitfalls. As such, one approach might involve a close working partnership between those with geodemographic expertise maintaining the framework and supporting the end user in understanding the development process and making the necessary decisions. It would be incumbent upon the party with geodemographic experience to ensure that the impact of the decisions were understood, and upon everyone ensure that the process is run without bias and as impartially as possible.

There are also technical challenges to on-the-fly classification development which still need to be overcome. For instance, the use of slow-running k-means clustering could present some challenges, unless adapted for with high computational processes (Singleton and Longley, 2009b). Studies by Singleton et al. (2016) and Adnan et al. (2010) have explored the possibility of this idea somewhat. However, issues were identified with the current computing capacity, though both are hopeful that such issues are not insurmountable.

This discussion represents some general aims for the future of geodemographics, beyond those which have been discussed before and which are the specific aims of this thesis, i.e. the shift to place-specific classifications and the intelligent and informed incorporation of a broader range of data in the build of an open geodemographic classification. The advancements which have been discussed in this section are beyond the scope and focus of this study, in a broad sense, but are worth considering nevertheless to capture a holistic picture of the status of current research in the field. Moreover, some of these considerations, particularly in the facilitation of more bespoke, purpose-ed classifications by the end-user, specifically where the end-user might be LCC, are afforded a little attention in Chapter 8 and Chapter 9 of this thesis.

3.8 Next steps and project outline

This chapter opened by asking whether the next evolution of progress in geodemographic classification development should continue to adapt and evolve the traditional framework, or whether it is time to seek a more revolutionary approach. It also questioned, in each case, which the most critical developments are that should be addressed. After documenting the Standard Framework as the template upon which almost all modern geodemographics have been developed, followed by an in-depth summation of the criticisms and "Grand Challenges" facing the existing practice, it is evident that there are many elements which could be targeted for improvement, or update. However, it is also clear that the solutions to most of the concerns raised are non-trivial, given the time that has passed with limited widespread advancements, particularly in the open geodemographics landscape, since Longley (2007) published his comprehensive warning of these challenges.

This thesis takes advantage of its high-level collaboration with LCC to present a series of practical investigations aimed at further extending and evolving the geodemographics literature and practices. As Singleton (2016a) alluded to, it is not feasible to attempt to address all concerns in a single research project, some must be prioritized, and others must be recommended for future research. Consequently, in this thesis, Chapter 2 has already highlighted place-specific exploration as an overarching theme of the research. The scope for broadening the input data source beyond the census is also of interest and will be similarly reviewed. This chapter has additionally highlighted variable selection procedures as a candidate for further exploration. Since variable selection concerns have plagued geodemographics since SAA (see Section 2.3.1), it has been a longstanding critical issue. However, it is also increasing in contemporary importance as new data sources are considered for inclusion in classification development. It therefore seems natural that it should also receive attention.

Explorations will begin with the generation of a benchmark Leeds-specific version of the 2011 OAC in Chapter 4 (named the “LSOAC”), which has received a detailed review in this chapter to support its development. This will act as the baseline for iterative adaptations to follow in the remaining chapters, which will focus on extensions to include novel data and new variable selection procedures. For consistency, and to act as validation of the impact of each change made in each chapter, the remainder of the methodology will remain consistent with the Standard Framework as adopted by the 2011 OAC. Each method will be compared as per the methodology demonstrated in the LOAC/2011 OAC comparison (Singleton and Longley, 2015), to either the OAC, the LSOAC, or the best update which has been generated prior to the current method. This will enable an evaluation of the impact of the change made in each iteration.

Summary

This chapter has outlined the current practical status of geodemographic classification development in the UK alongside the perceived weaknesses of the current practices, situating the research to follow in the existing landscape, whilst also setting out a baseline classification to be used throughout the analysis.

Chapter 4 - Place-specific classification for Leeds

4.1 Introduction

As budgets decrease and service provision and resource allocation is further devolved, LAs are increasingly looking for ways of developing smart decision making to achieve much needed efficiency (Longley, 2005; Ashby and Longley, 2005). Geodemographic classifications purport to offer a level of insight into populations which can support more targeted public sector interventions, with similar success having been achieved in commercial enterprise. However, LCC, who have been seeking to employ such classifications in their own practice, have raised concerns regarding the suitability of those available at present, and their ability to accurately reflect the residents within their city.

The previous chapters have outlined these concerns in more detail, and have offered as an explanation, the limitations associated with the tradition of deriving classifications at a national extent. In summary, there is a concern that such a methodology for deriving classifications could result in outputs which fail to identify, or may even mask, population patterns uniquely present in particular regions (Alexiou, 2017; Singleton and Longley, 2015). Such an effect, it is theorised, could impose restrictions on the ability to derive the level of local context which is required for successful targeted application in the public sector. The most popular geodemographic classifications, both public and commercial, are all currently generated at a national extent. However, these concerns regarding their suitability is leading to growing calls for a systematic shift to a more local context, as discussed in Section 2.4.4.

This chapter extends recent work published in the growing interest area of place-specific geodemographic classifications. The requirements for place-specific classifications in general, and in Leeds specifically, are assessed, and recent place-specific exploration is adapted to develop preliminary Leeds-specific classification methodologies. The concerns of LCC are presented in Section 4.2, alongside an assessment of how well the 2011 OAC currently represent the population of the city. Section 4.3 presents the limited history of place-specific geodemographic development, which Section 4.4 builds from to propose two approaches for developing this methodology to improve the output of the 2011 OAC for Leeds. The results of these approaches are explored in Section 4.5, before Section 4.6 details potential extensions of the work presented, and Section 4.7 concludes with a discussion of the next steps and recommendations for further development.

4.2 Issues with existing classifications of Leeds

Experts within LCC believe that current classifications are not capturing the key characteristics of the population of the city. This could be a consequence of there being features distinct to the complex and somewhat unique geography of Leeds which national classifications are ill-equipped to identify. Similar issues have been identified by LCC in both the commercial classification which they use and in the freely available 2011 OAC outputs. Moreover, Burns et al. (2018) case study, which similarly focused on Leeds, also identified concerns with the applicability of the 2001 OAC in the city. As such, LCC are keen to see an exploration of the scope for, and possible potential of, place-specific classifications for Leeds, to understand whether such a shift might improve the relevance of the result, providing a primary motivation for exploring place-specific classifications for Leeds in particular.

Before such exploration commences, it might be beneficial to investigate national-level outputs in the city, to understand where the weaknesses and limitations exist. LCC's current licence with one commercial classification provider exclusively enables access to the output result and supporting materials, and not a behind-the-scenes look at the build underpinning the classification. As such, it is possible to explore the commercial classification outputs in Leeds, and identify where they differ from LCC's expectations of the city. However, any future work in tweaking the same classification, or re-classifying at a Leeds-specific extent to evaluate for improvements, would not be possible.

Therefore, the work to follow focuses instead on the 2011 OAC (described in Section 3.4). Not only is this a freely available as a national statistic published by the ONS, all of the information relating to its build is also openly and transparently available for use (Gale, 2020). This approach has several benefits. Firstly, in addition to seeking a more appropriate solution with increased relevance in the city, LCC are keen to identify a more cost-effective alternative to the commercial classification currently used, which an improved 2011 OAC may provide. Moreover, a focus on the 2011 OAC enables open and transparent publication of the work conducted here, supporting replication across other cities which may also benefit, and who also have access to the 2011 OAC.

4.2.1 Suitability of the 2011 OAC for Leeds

The LA boundary for LCC covers a diverse geography, comprising a multicultural city-centre, in addition to suburban and exurban districts. The latter encompasses a rural fringe of market towns including Wetherby and Otley. Though geographically small relative to other large cities in the UK, Leeds has a growing population of over 790,000 residents. The population is ethnically diverse, with around 20% of residents identifying as an ethnic minority (Leeds Observatory, 2020a), and hosts a large population of current students, recent graduates and alumni of its five universities. The city has also experienced considerable urban restructuring which has affected the composition of the underlying residential structure. This is reflected in reports of a recent increased popularity of city-centre living (Swinney and Carter, 2018).

Figure 4.1 shows the distribution of the 2011 OAC Supergroups across the city's OAs. The map highlights spatial patterns in the allocation of some of the Supergroups seemingly in a concentric ring structure from the centre of the city, signifying the presence of underlying population structures similar to those theorised by (Burgess., 1925) (see Section 2.3.1). "Cosmopolitans" are almost wholly constrained to the city-centre and nearby OAs stretching to the North-West (see Appendix A.1 for reference). "Ethnicity Central" are almost entirely located in the inner-city OAs, particularly in the South and the East of the city. "Multicultural Metropolitans" and "Constrained City Dwellers" are more prevalent in the surrounding suburbs beyond the outskirts of the centre.

Conversely, as the name would suggest, "Rural Residents" present largely in the rural fringes of the city, in the North and East of the wider city region. However, whilst this Supergroup appears dominant in Figure 4.1, it is attributed to just 1.7% of the city's 2,543 OAs (Table 4.1). This is an under-representation of this Supergroups in comparison with the breakdown of the Supergroups nationally. The perceived prevalence of "Rural Residents" in the city, based on the map, is a consequence of the OA boundaries being derived based not on geographic scale, but on household counts. As such, OAs in rural regions with low population density are naturally geographically larger.

Besides the under-represented "Rural Residents", and the conversely over-represented "Multicultural Metropolitans" and "Urbanites", the distribution of the city's OAs across the su-

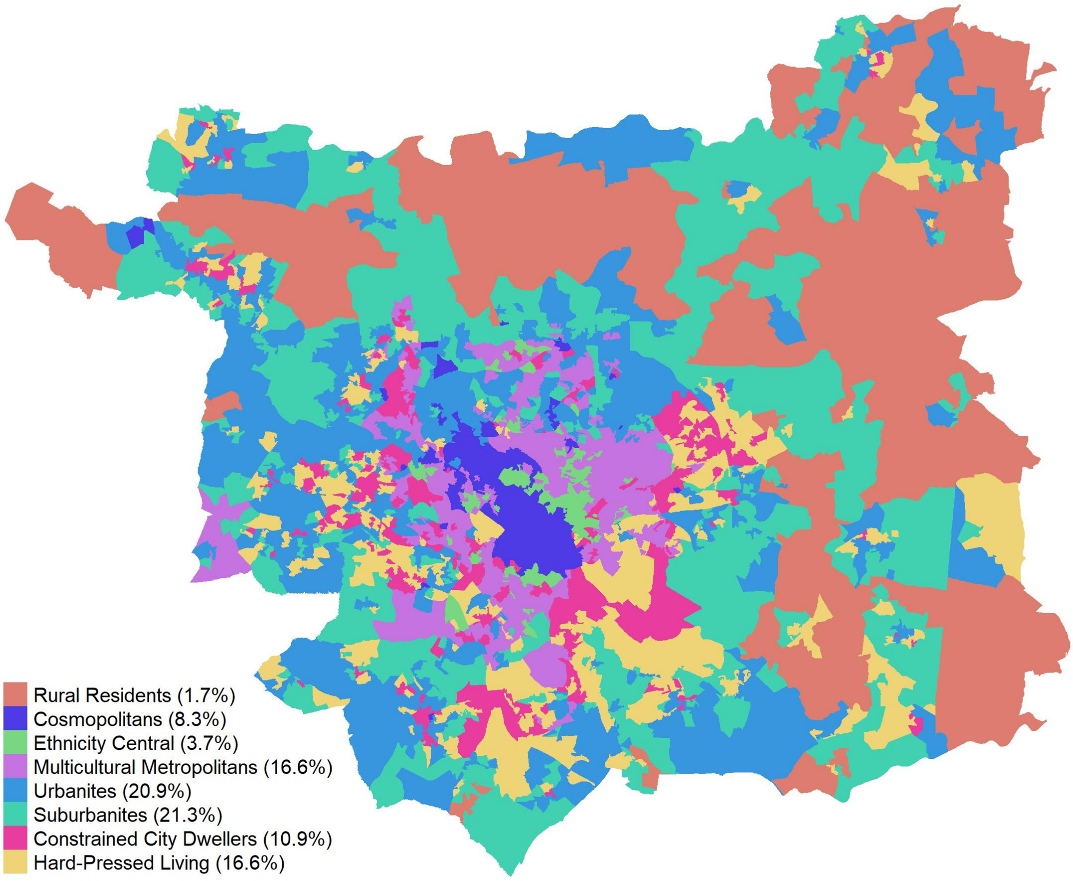


Figure 4.1: Distribution of the 2011 OAC Supergroups across Leeds OAs.

pergroups otherwise reflect a similar pattern to the national-level distribution (Table 4.1). This suggests that, in many ways, Leeds is a very ‘average’ city.

To gain a better understanding of how well the the Leeds population is reflected at the national level, and if and where the attribute characteristics of the population deviate from the national average, Figure 4.2 depicts the variance for each of the 60 input variables of the 2011 OAC at each extent (see Appendix B.1 for the full list). The parallel coordinate plot is, an effective visualisation technique which enables a multivariate comparison. The plot summarises the transformed and standardised 2011 OAC input data to demonstrate the range of values for each variable across the OAs, both in Leeds and nationally, for comparison. A bold line indicates the median value for the OAs at each extent. Shading with linearly decreasing lightness from the median to highlight the first to ninth decile (as per Slingsby et al. (2011)) illustrates the distribution of values for each variable. The original order of the variables (as per the list in Appendix B.1) are maintained in the direction of the

2011 OAC Supergroup	**UK OAs (%)**	**Leeds OAs (%)**	**Increase/Decrease in Leeds (%)**
Rural Residents	11.8%	1.7%	-10.1%
Cosmopolitans	5.6%	8.3%	2.7%
Ethnicity Central	5.1%	3.7%	-1.4%
Multicultural Metropolitans	10.1%	16.6%	6.5%
Urbanites	16.7%	20.9%	4.2%
Suburbanites	20.2%	21.3%	1.1%
Constrained City Dwellers	11.7%	10.9%	-0.8%
Hard-Pressed Living	18.9%	16.6%	-2.3%

Table 4.1: Percentage breakdown of UK and Leeds OAs by 2011 OAC Supergroup.

x axis, and can thus be similarly grouped into the five variable domains outlined in Section 3.4.2.

The overall picture for the two geographies is extremely similar. However, deviations occur particularly across the Housing domain, both in house type and housing tenure, indicating a housing profile in Leeds which is somewhat distinct from the national profile. The Housing domain, therefore, presents a good candidate for development in Chapter 5 and Chapter 6, which will extend beyond the census data and look to increase the local relevance of the input data used.

In both instances, for Leeds and for the UK as a whole, the variation from the mean in values across the OAs is relatively low for many of the variables. By the nature of the classification methodology, the results are driven by the variables exhibiting variance, i.e. variables which change from OA to OA. Consequently, this result suggests that both classifications are likely being driven by a limited number of the total 60 variables. Again, the housing domains are displaying the broadest variance at both scales, in addition to the socio-economic variables and demographic variables relating to age, marital status and ethnicity, suggesting that Leeds also contains a more diverse demographic profile than the national average.

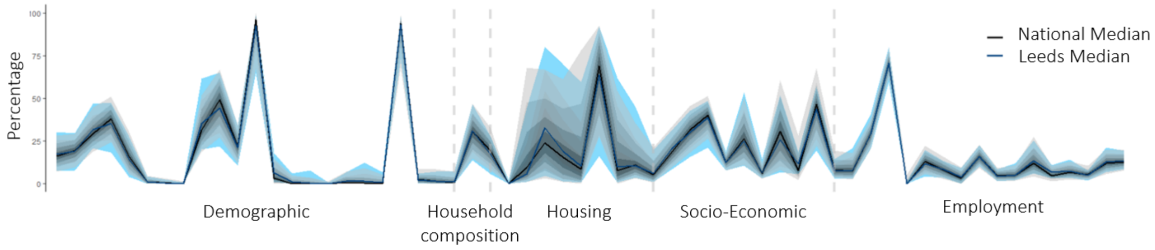


Figure 4.2: Parallel coordinates plot showing a comparison of the median and decile ranges (1 to 9, shaded) of the 60 census variables, for the UK and for Leeds.

A similar representation of the variation within variables split by the 2011 OAC Supergroups provides additional insights regarding how well the Supergroups are able to capture the variance in the city (Figure 4.3). This presents an indication of the fit of the Supergroups. Again, the variance from the median for each variable is demonstrated with the shaded deciles (1-9). Comparisons of the UK result (top) with the result just in the Leeds OAs (bottom), highlight broad increases in the variance of each Supergroup at the local perspective, suggesting that the fit of the classification is better at the national level. Again, the housing and socio-economic indicators show the greatest variance in both cases, which is further propagated at the local level.

"Cosmopolitans" appear to present the worst fit overall, particularly at the Leeds level. The characteristics associated with this Supergroup are predominantly those which relate to students. Reasons why the student population might be the most poorly represented will become clearer throughout this section.

Additional insights can also be gained by considering the relationships between variables, both within Leeds and at the national level. Since geodemographic classifications intend to capture multivariate phenomena in an area, if the relationships between variables differ at the local and the national level, the relevance of the national classification will be affected when transferred to the local level. The heatmap in Figure 4.4 illustrates the pairwise Pearson correlation between each of the variables nationally (above the diagonal) and within Leeds exclusively (below the diagonal). Again, the variables are displayed as per the original order set in the 2011 OAC, and the domain groups are highlighted. The plot illustrates broad symmetry, indicating largely consistent relationships between the majority of variable pairs across the two geographic extents.

Differences which do exist are difficult to pick out here and are better illustrated in Figure 4.5, which shows the difference in the absolute correlation at the national and Leeds levels in instances where the correlation is stronger in Leeds, i.e. where there is more of a relationship between the variables in Leeds than in the UK as a whole. These differences indicate the presence of attribute structures in the city which are not identified nationally. Though most differences of this nature, where they occur, are relatively minor, a few notable exceptions are highlighted.

The "Highest qualification - level 3" indicator is correlating more strongly with "Individuals

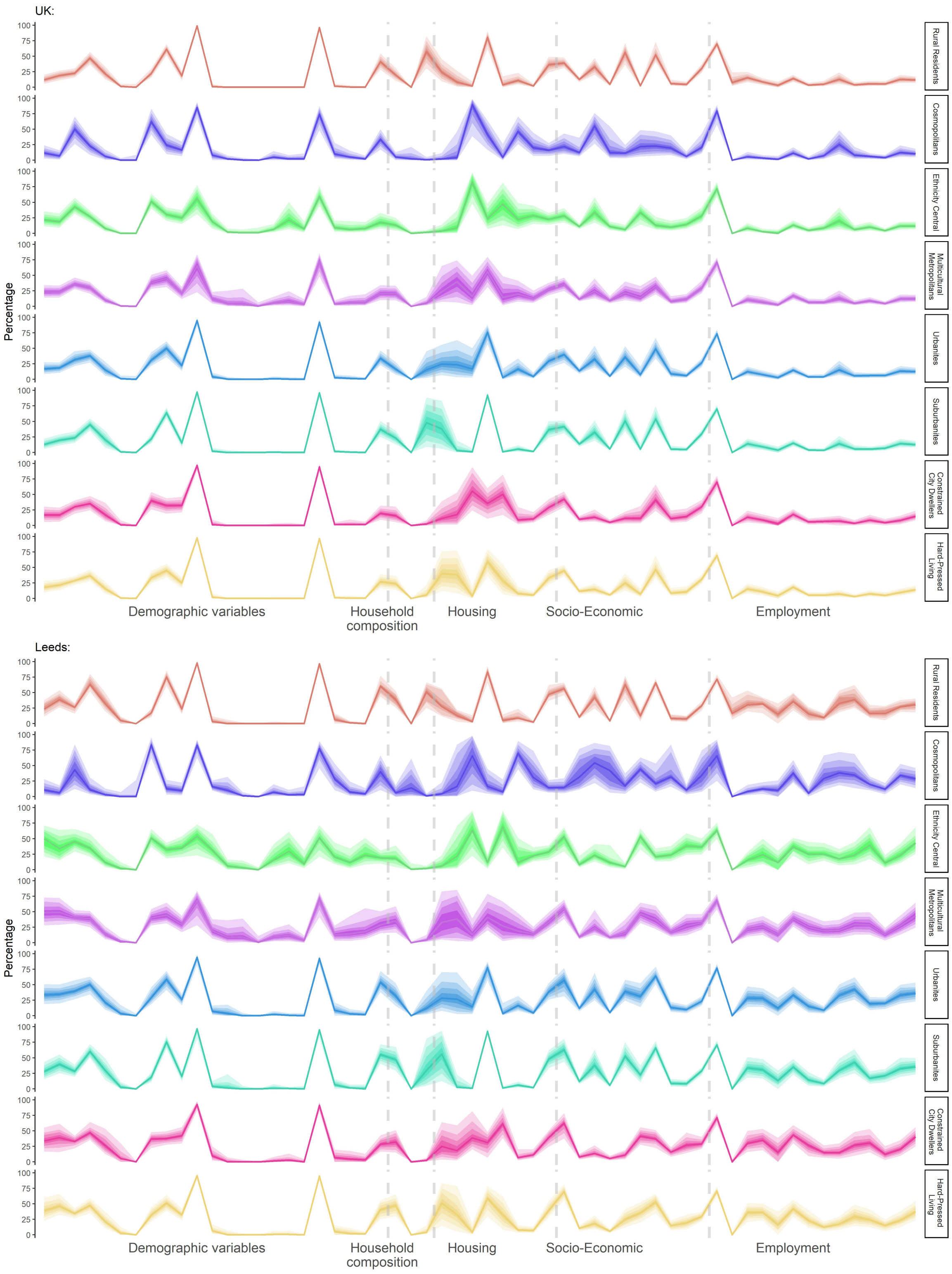


Figure 4.3: Parallel coordinates plot showing a summary of the median and decile ranges (1 to 9, shaded) of the 60 census variables by Supergroup for the UK (top) and for Leeds (bottom).

employed full-time", "Individuals employed part-time" and "Individuals employed in roles in service industries" in Leeds than would be expected based on the national average (Fig-

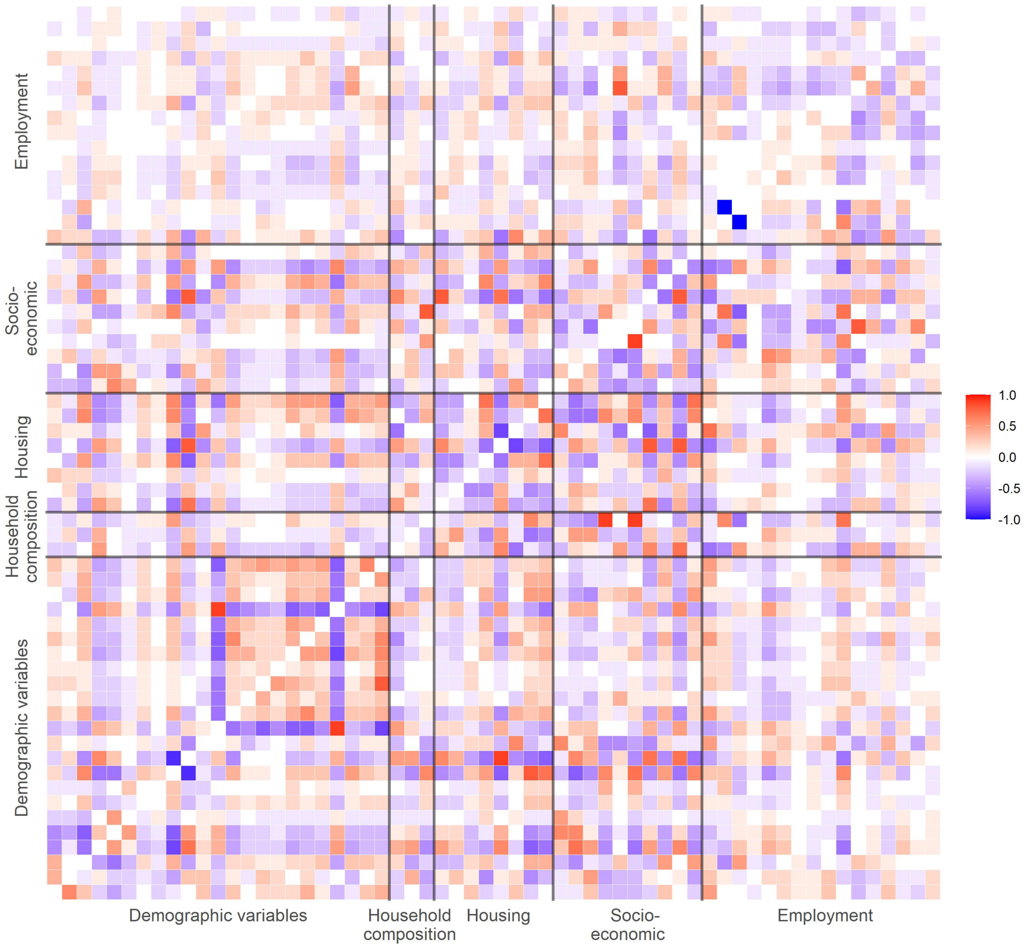


Figure 4.4: Pairwise correlations of the input data).

ure 4.5). As per Table 4.2, these correlations are strongly positive, negative and positive, respectively. The first two indicators represent the opposite of one another, and thus, will necessarily result in complementary correlations.

In real terms, the “Highest qualification - level 3” variable represents the percentage of individuals aged over 16 in an area who have achieved 2 or more A-levels. Both nationally and in Leeds, this indicator also strongly correlates with “Full-time students” (Table 4.2). These combined results suggest a notable prevalence of students in part-time employment in the service industry in Leeds, further indicating a uniqueness in the student population, and in the working population within the city more broadly, which is not a feature at a national perspective. As such, it is likely that this will not be a feature which is appropriately represented in the 2011 OAC, a result which is supported by the poor fit associated with “Cosmopolitans” in Leeds shown in Figure 4.2.

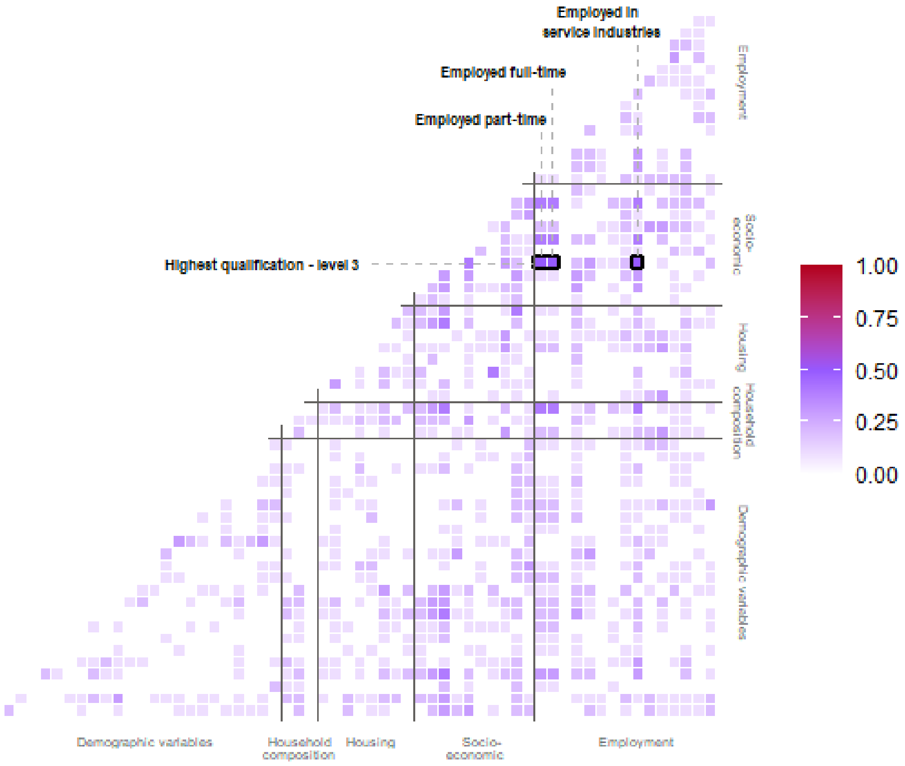


Figure 4.5: The correlation increase for variable pairs which are more highly correlated in Leeds than nationally).

	Correlation with 'Highest qualification - Level 3'	
	in Leeds	in the UK
Households with full-time students	0.9	0.5
Persons who are schoolchildren or full-time students	0.9	0.5
Employed part-time	0.6	0.1
Persons aged over 16 who are single	0.6	0.1
Households who are private renting	0.6	0.1
Employed in Service industries	-0.6	-0.1
Employed full-time	0.6	0.1

Table 4.2: Variables strongly correlating with the "Highest qualification - level 3" variable in Leeds, and the correlation at UK level.

The distinctive characteristics of Leeds which are highlighted throughout this section are enough to suggest that the city could be in a position to benefit from a move to place-specific geodemographic classifications, and to recommend Leeds as a good candidate for further exploration.

4.2.2 Implications for public sector application

It is worth re-iterating that there is a strong desire, and even a need, to encourage the production and adoption of open and transparent geodemographic classifications which can be adopted for public sector use. However, the uncertainties raised regarding the suitability of the 2011 OAC for capturing *some* unique local populations (illustrated in Section 4.2.1) suggest that there is work to be done to achieve open classifications which better represent some unique local populations in some cities, including Leeds. As eluded to, a poor or unreliable representation of the city could not be used to accurately or effectively inform local-specific public sector activity, such as local-level policy making decisions and resource allocation, as is increasingly desirable. It is therefore important that the classifications are validated for their ability to accurately reflect the populations which they represent, and where doubts are raised regarding their suitability, as is the case here, that alternative, purpose-built classifications are developed. It is not appropriate for existing classifications, either commercial or freely available, to just be re-purposed and adopted where there is a concern regarding their suitability.

4.3 Existing place-specific methodologies

Criticisms relating to national-level classifications, and the idea of alternative place-specific classifications, are not new (see Section 2.4.4). However, practical progress towards a methodology which serves to address these criticisms is in its infancy, and has yet to lead to the development of a unified and widely adopted framework for delivering locally relevant, place-specific classifications. Thus national-level classifications are currently the de-facto practice in the commercial geodemographics industry and beyond. Nevertheless, exploration of place-specific alternatives has commenced, particularly from within academia, as detailed in Section 2.4.4. This work has resulted in one primary local-level classification in the UK which is openly available, the LOAC, classifying the OAs in London, which is discussed further in the next section.

4.3.1 Introduction to the LOAC

A consideration of the distribution of Greater London OAs (henceforth referred to as simply 'London') across the 2011 OAC Supergroups highlights the potential weaknesses of the national-level methodology in capturing population characteristics which are uncommon in the wider extent, and which deviate from the national mean. Compositions of population characteristics which are prominent but unique within the city were not identified in the classification. This is a consequence of the methodology, which seeks to identify common group structures across the input data. Labelled "the UK's only global city" by Singleton and Longley (2015), London is very much a unique geography in comparison to the rest of the UK, particularly in terms of its ethnic make-up. As a result, is not well represented within the 2011 OAC result.

The percentage allocation of the London OAs to each of the Supergroups (Table 4.3) shows that almost 70% of the London OAs were allocated to just two Supergroups, "Ethnicity Central" and "Multicultural Metropolitans". Naturally, such a limited allocation does not appropriately reflect the diversity of the population of the capital, and thus does not distinguish well enough to underpin targeted public policy strategies. There is not enough to be learnt from this classification to make its application in the development of policy worthwhile (Petersen et al., 2010). Further analysis of the results, carried out by Singleton and Longley (2015) indicated a particular weakness in their ability to represent the diverse ethnic makeup of the city.

2011 OAC Supergroup	**Percentage of UK OAs**	**Percentage London OAs**
Rural Residents	11.8%	0.1%
Cosmopolitans	5.6%	14.3%
Ethnicity Central	5.1%	37.0%
Multicultural Metropolitans	10.1%	32.9%
Urbanites	16.7%	9.1%
Suburbanites	20.2%	4.6%
Constrained City Dwellers	11.7%	1.1%
Hard-Pressed Living	18.9%	1.0%

Table 4.3: Distribution of Greater London OAs across 2011 OAC Supergroups.

The 2001 OAC had itself faced similar criticisms in its reflection of the population of London, and had also allocated the city's OAs to a limited number of Supergroups. Consequently, questions were already being raised prior to the development of the 2011 OAC regarding the appropriateness of employing national-level classifications and their ability to suitably identify the nuances of particularly unique local populations (Petersen et al., 2010). Despite this, maintaining consistency with the 2001 OAC was regarded with greater priority in the development of the 2011 OAC, and thus, the classification was once again derived at a national extent (Gale and Longley, 2012). Though the national approach had prevailed, the decision was made to undergo a transparent and open release strategy which included publishing the methodology, data and code underpinning the 2011 OAC classification. Not only did this strategy align with the broader trend for transparency in research, Gale and Longley (2012) specifically cite the potential that it afforded for future studies to adapt the materials and derive locally bespoke versions of the classification as an influential factor in this decision.

In response to their criticisms of the 2011 OAC and its representation of London, Singleton and Longley (2015) have taken advantage of the opportunity afforded by the transparent release of the 2011 OAC to derive the LOAC, a sub-national classification specifically classifying the London OAs to re-produce the classification at a local level, in the hopes of drawing out some of the hidden population structures. Again, the results have been made publicly available for free and open use (Singleton, 2016b).

The LOAC primarily adopts the underlying data, methodology and assumptions of the 2011 OAC exclusively for the London OAs, though some small amends have been made to the data to retain its accuracy at a local extent. As described in Section 3.4.2, most of the input data represents the percentage of a given attribute in each OA, with the exceptions being a Population Density Ratio and a Standardised Illness Ratio (SIR) (again, both explained in more detail in Section 3.4.2). Whilst the Population Density Ratio is calculated independent of the focus extent, and translates to a local level, the SIR was calculated based on the 2011 OAC base population, the UK. As such, it was necessary to re-calculated to adjust for the shift in the base population from the entire UK to London. Similarly, all of the data was re-normalised and re-standardised to adjust for the new extent.

The LOAC study proceeds to subsequently evaluate the impact of the new classification, presenting a comparison between the 2011 OAC and the LOAC, for the London OAs, based on a Within Cluster Sum of Squares (WCSS) statistic (introduced in Section 3.3.2, STEP 4). This is adopted as a measure of the 'fit' of each Supergroup. In a comparison, a lower WCSS represents a better fit, demonstrating more homogeneity across the OAs in the Supergroup, and thus a superior performance (Singleton, 2016b). It is returned as part of the standard package of outputs when the 2011 OAC code is run, as per the published assets (Gale, 2020).

The WCSS is presented in the LOAC as a validation of the shift to a place-specific extent, demonstrating a widely improved performance (Singleton and Longley, 2015). Analysis of the LOAC results seemingly demonstrate that the shift to a local extent has achieved an improved representation of the city's population, particularly in OAs with composite characteristics which were not well captured by the former.

4.4 Beginning place-specific experimentation for Leeds

The developers of the LOAC concluded with an endorsement for similar isolated investigations in other cities which likewise exhibit evidence of city-specific demographic compositions diverging from national patterns, and which are thus also ill-represented at a national extent (Singleton and Longley, 2015). Based on the findings in Section 4.2, it seems that Leeds might be one such example.

The plots in Section 4.2.1 demonstrate evidence that some of the input variables which underpin the 2011 OAC show a large amount of variation across the city. This suggests that some benefit might be achieved in a place-specific classification for Leeds. As the predominant academic study in local-level geodemographic classifications at this time, the ideas presented in the LOAC act as a jumping off point in the development of the place-specific classification for Leeds (the primary objective of this thesis). Moreover, many of the evaluation methods employed in the development and exploration of the LOAC are drawn on here to evaluate the results of the exploration. Many of the freely published assets which were used to derive the 2011 OAC (and as a product, the LOAC) are directly employed in this study (adapted where relevant). This methodology underpins this initial phase of the development, which will then be further extended throughout Chapters 5-8. As explained in Chapter 2, the study in this chapter offers an initial base place-specific classification for Leeds, upon which extensions will be tested throughout the subsequent chapters. This Leeds-specific classification is henceforth labelled the "LSOAC".

In addition to creating a Leeds-specific classification by re-applying the LOAC methodology to re-classify the Leeds OAs, a second test is also executed. This test considers whether the uniquenesses which have been identified in London by Singleton and Longley (2015) might in themselves be causing poorer results elsewhere in the UK, specifically in Leeds in this case, by skewing the results. Thus, a second classification is also derived which, again, adopts the 2011 OAC methodology, but in this case is applied to all OAs in the UK *except* those found in London. This generates the complement of the LOAC, and the result for OAs in Leeds can be evaluated for improvements. This second classification is henceforth labelled the “nLOAC”.

The development of both the LSOAC and nLOAC are detailed below, alongside additional supporting information.

4.4.1 Leeds specific OAC - “LSOAC”

As outlined above, the development of the LSOAC directly reflects the LOAC approach, but in this case, re-applied to Leeds. It is likewise developed upon a subset of the original 2011 OAC input data, in this case, limited to the data from the 60 input variables which relates to the 2,543 OAs in Leeds. As in the development of the LOAC, the SIR is again re-calculated for just the Leeds OAs, and all input data is locally normalised and standardised. Each of these processes are executed as described in Section 3.4.2. No additional data preparation is conducted. The remainder of the methodology is maintained as per the 2011 OAC (Gale et al., 2016) (see Section 3.4.2). To support this, the published assets used to derive the original 2011 OAC (Gale, 2020) are employed, including the base dataset and source code. These are adapted where necessary to facilitate the amendments outlined.

As such, a k-means clustering is again employed. As per the preliminary explorations carried out as part of the 2011 OAC and LOAC development, a scree plot is created to estimate the number of groups at which to derive in the k-means cluster analysis at the most aggregate level, the Supergroup level (Figure 4.6). As described in some detail by Singleton and Longley (2015), the plot demonstrates the total WCSS values derived for cluster procedures generated with differing numbers of clusters (represented by *k).* The purpose of the plot is to help identify the value of *k* at which the declining rate of the WCSS begins to stabilise. The authors note, however, that the plot is not conclusive and is often adopted to support a wider, more qualitative decision. This decisions is typically made based on a review of the scree plot alongside a consideration of the broader aims of the clustering process. As such, the LOAC is derived with 8 classes based on the evidence in Figure 4.6, and to maintain a consistency in the number of Supergroups generated by the 2011 OAC.

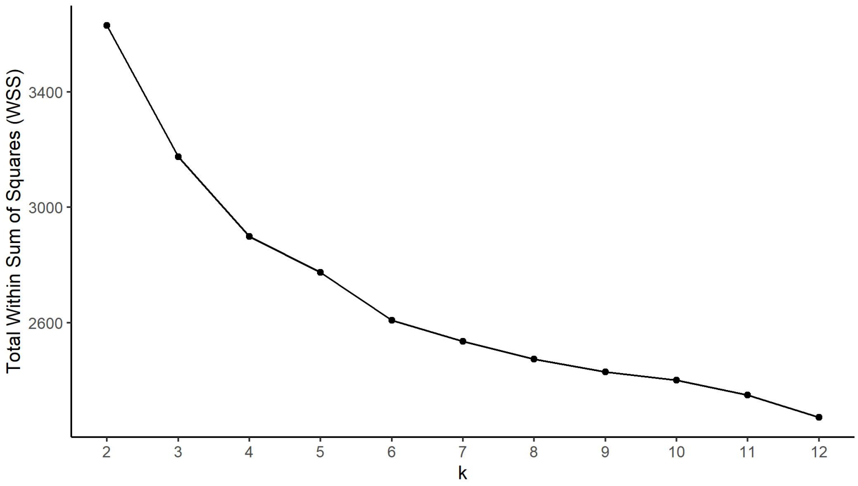


Figure 4.6: Scree plot for LSOAC input data.

Whilst the 2011 OAC employed a hierarchical cluster process in which each of the Supergroups were further divided into Groups and each once again into Subgroups (discussed in Section 3.3.2, STEP 5), a process which was partially replicated in the generation of the LOAC, due to the limited number of OAs in Leeds, the LSOAC development terminates at the Supergroup level. This helps to retain meaning in the results, which might be eroded with a reduction in groups size further down the hierarchical process.

4.4.2 Removing London from the OAC - “nLOAC”

With a total of 25,053 OAs (compared to Leeds’ 2,543 OAs), Greater London contains 10.8% of the OAs in the UK. The allocation of these OAs to the 2011 OAC Supergroups is outlined in Table 4.4. As highlighted by the authors of the LOAC (Singleton and Longley, 2015), the London OAs are extremely over-represented in a limited number of the Supergroups. Approximately 70% of the London OAs are classified as either “Ethnicity Central” or “Multicultural Metropolitans”. This result has led to the suggestion that the 2011 OAC is not sufficiently differentiating the sub-populations within the capital.

On the flip side, the impact of this weighted allocation is such that a large proportion of all of the OAs assigned to these Supergroups are found in London, contributing 78% of all “Ethnicity Central” OAs, 35% of “Multicultural Metropolitans”, and 27% of all “Cosmopolitans” (Table 4.4). If the London OAs are not well represented by this classification, as the LOAC study indicates, then one might hypothesise that the inclusion of so many poorly assigned OAs could give rise to a poor result for the other OAs, specifically in the

Supergroups which contain a substantial London presence. With 28.6% of all of the Leeds OAs falling into one of these three Supergroups (Figure 4.1), if this is an issue, it could be affecting a sizeable amount of the Leeds OAs.

2011 OAC Supergroup	**Total OA Count**	**London OA Count**	**Percentage of Supergroup OAs found in London**
Rural Residents	27,300	15	0.05%
Cosmopolitans	13,125	3,584	27.3%
Ethnicity Central	11,849	9,263	78.2%
Multicultural Metropolitans	23,502	8,233	35.0%
Urbanites	38,697	2,285	5.9%
Suburbanites	46,850	1,141	2.4%
Constrained City Dwellers	27,135	281	1.0%
Hard-Pressed Living	43,838	251	0.6%

Table 4.4: Count and percentage of 2011 OAC Supergroups allocated to London OAs.

The nLOAC is generated to test this hypothesis for the Leeds OAs. The methodology again replicates the 2011 OAC, this time removing just the OAs in London to investigate whether this might itself improve the 'fit' of the classification in Leeds.

Again, the data is re-normalised and re-standardised, and the SIR is re-calculated based on the new geography (as per the methodology outlined in 3.4.2). Likewise, a scree plot (Figure 4.7) is created to assess the WCSS of a range of potential k values (as explained in the previous section). Again, the plot supports the desire to maintain 8 clusters to aid in comparison with the 2011 OAC and the LSOAC, described above. Similarly, the exploration again terminates at the generation of the Supergroups, since additional levels can not be compared to the LSOAC.

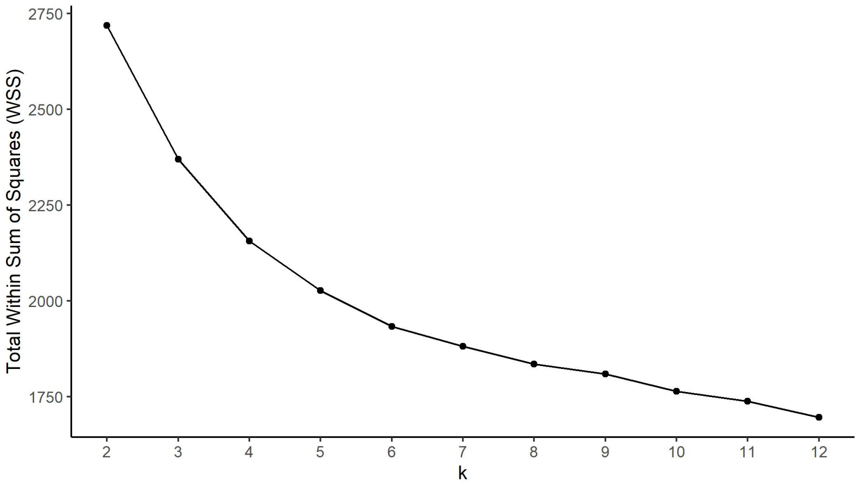


Figure 4.7: Scree plot for nLOAC input data.

4.5 Results

4.5.1 Comparing the LSOAC and the nLOAC with the 2011 OAC

The geographical distributions of the LSOAC and nLOAC Supergroups across the Leeds OAs are depicted in Figure 4.8. To maintain a distinction, the LSOAC and nLOAC Supergroups are labelled A-H and I-P, respectively. Some general city-level patterns appear consistently across each classification (including the 2011 OAC mapped in Figure 4.1), such as a distinction between the Supergroups assigned to the city centre OAs and the OAs located the North and East of the city (see Appendix A.1 for reference), which Oldroyd et al. (2020) distinctly identify as the more *rural* areas of the city. Though this is reduced somewhat in the nLOAC, in which Supergroup "N" is more widely distributed across both aspects of the city than any of the LSOAC Supergroups. It seems, therefore, that the LSOAC is more keenly identifying the rural/urban make-up of Leeds.

As per the methodology of the LOAC (discussed in Section 4.3.1), the WCSS is used to compare the fit of each OA in the LSOAC and the nLOAC with the fit of the 2011 OAC, as a measure of performance success. This is calculated as per the explanation in Section 3.3.2 (STEP 4). Lower WCSS scores indicate that the OA is closer to the cluster mean, and thus indicate a better performance. Figure 4.9 highlights the OAs in which the LSOAC and nLOAC perform better than the 2011 OAC. Whilst improvements in the LSOAC are largely constrained to the OAs in the city centre, and to the South and West of Leeds, there is an improvement in the closeness of the Supergroups derived by the nLOAC methodology

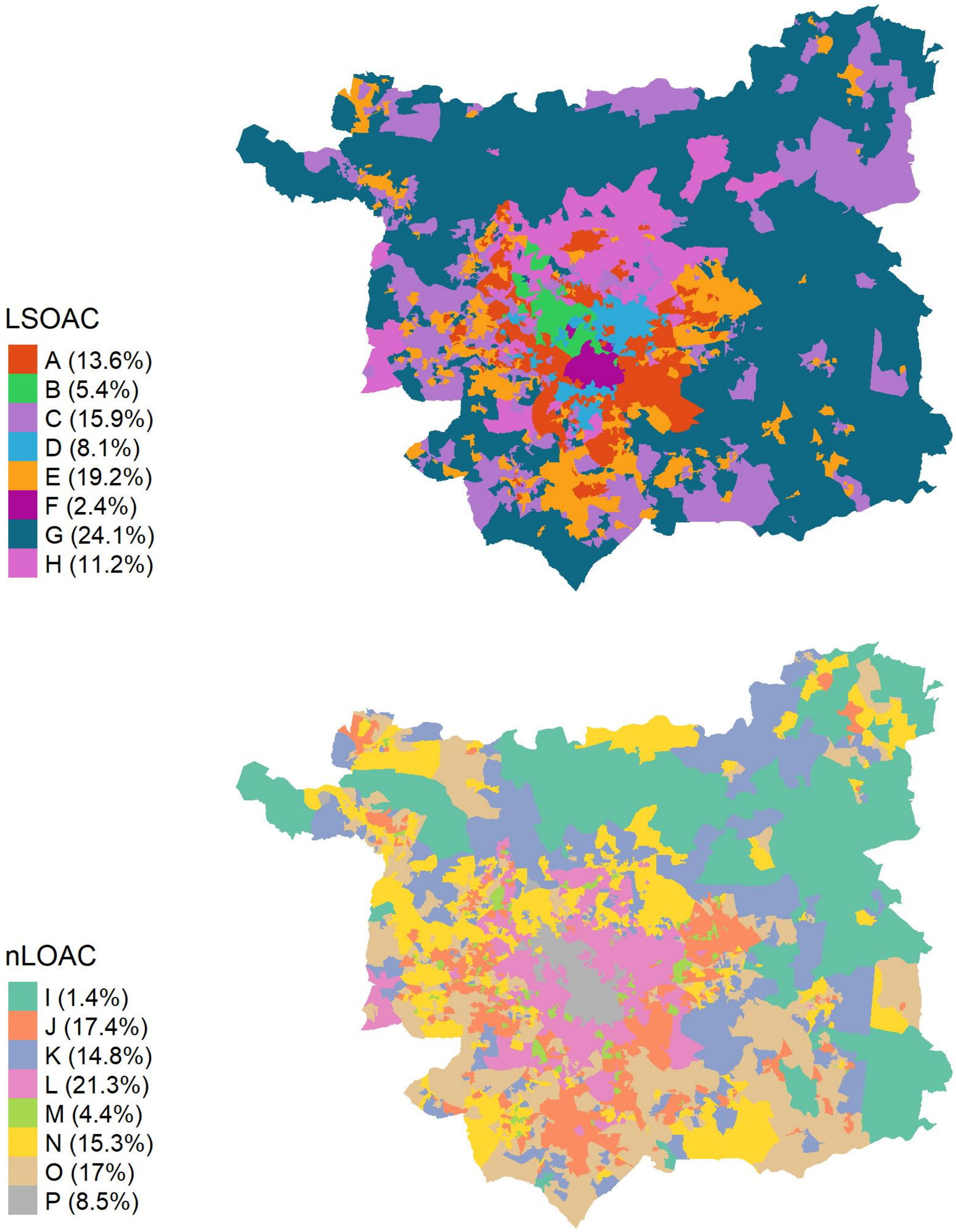


Figure 4.8: Percentage and spatial distribution of LSOAC and nLOAC Supergroups across Leeds OAs.

in many more OAs from right across the city.

This result is echoed in Table 4.5, which shows the percentage of OAs experiencing an improvement in each re-classification, both in total, and based on the 2011 OAC Supergroup to which each OA was originally assigned. The removal of London in the nLOAC improves the performance of the classification for just over half of the OAs in the city. Thus, overall, it performs similarly well to the 2011 OAC. With the exception of “Rural Residents” and “Ethnicity Central” in which the 2011 OAC performs largely better, and of “Cosmopolitans” in which the 2011 OAC performs largely worse, the balance of the performance in the 2011 OAC and the nLOAC is fairly even across the OAs in the other Supergroups.

Since 78% of “Ethnicity Central” is made up of London OAs in the 2011 OAC (Table 4.4), the better performance of the 2011 OAC suggests that the characteristics of some of the inner-city OAs in Leeds, where this Supergroup is most present, are better represented *alongside* these London OAs. This could reflect a diversity in these areas of Leeds which is unique not just to the capital but potentially to other large UK cities, albeit on a smaller geographic scale. Alternatively, just 0.1% of the “Rural Residents” OAs are within London (Table 4.4). This could explain the limited number of OAs experiencing an improvement in the nLOAC performance for this Supergroup.

The LSOAC methodology performs better than the 2011 OAC in 28% of the OAs, a little over half as many as the nLOAC (Table 4.5). Again, the LSOAC performs best in the OAs previously classified as “Cosmopolitans” in the 2011 OAC. Both the LSOAC and the nLOAC generate improvements in this Supergroup (which was also highlighted in Section 4.2.1 as the 2011 OAC Supergroup exhibiting the poorest fit in Leeds). This seems to suggest that there might be a characteristic of the “Cosmopolitans” which is local to Leeds, and which has thus far been masked by classifications at the national extent, as hypothesised.

		% of OAs Improved	
2011 OAC Supergroup	**Total OA Count**	**by the LSOAC**	**by the nLOAC**
All Supergroups (total)	2,543	27.8%	51.9%
Rural Residents	44	2.3%	20.5%
Cosmopolitans	212	69.3%	72.2%
Ethnicity Central	94	23.4%	10.6%
Multicultural Metropolitans	421	46.6%	64.1%
Urbanites	531	19.4%	46.0%
Suburbanites	541	17.4%	49.4%
Constrained City Dwellers	278	26.3%	43.5%
Hard-Pressed Living	422	16.8%	58.5%

Table 4.5: Count of Leeds OAs assigned to each 2011 OAC Supergroup and percentage of each Supergroup improved by the LSOAC/nLOAC.

Though the LSOAC generates a smaller *quantity* of improved performances across the OAs

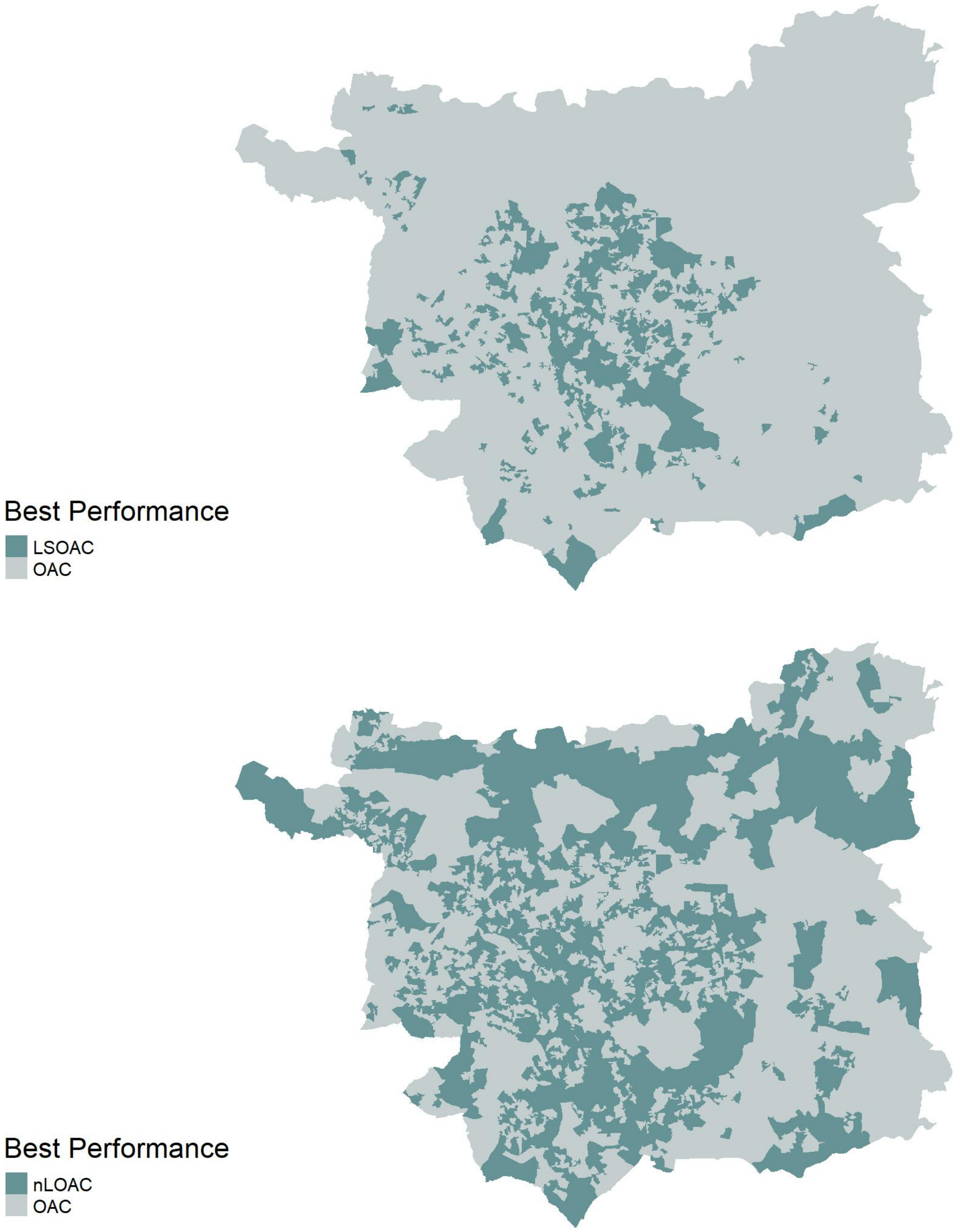


Figure 4.9: Best performance in comparison between the 2011 OAC and LSOAC (top) and the 2011 OAC and nLOAC (bottom).

than the nLOAC, a consideration of the average Squared Euclidean Distance (SED) (introduced in Section 3.3.2, STEP 4) of the OAs in each cluster indicates that the average *size* of the improvements generated by the LSOAC is greater for all but those classified as "Rural Residents" in the 2011 OAC (Table 4.6). Additionally, the improvements generated by the LSOAC are much more discriminant, both geographically (as previously mentioned) and across the 2011 OAC Supergroups. Both of these features are potential indicators of the LSOAC generating a result which is more representative of the underlying structures within the city.

		Average SED Improvement	
2011 OAC Supergroup	**Average SED in 2011 OAC**	**by the LSOAC**	**by the nLOAC**
Rural Residents	0.76	0.029	0.041
Cosmopolitans	1.32	0.354	0.057
Ethnicity Central	1.10	0.080	0.046
Multicultural Metropolitans	0.99	0.114	0.044
Urbanites	0.84	0.099	0.045
Suburbanites	0.80	0.122	0.051
Constrained City Dwellers	0.93	0.084	0.046
Hard-Pressed Living	0.78	0.054	0.048

Table 4.6: Average SED of each 2011 OAC Supergroup and improvement in average SED of the OAs in which the LSOAC/nLOAC performed better.

4.5.2 Exploring the LSOAC Supergroups

To better understand the Supergroups produced by the LSOAC and the city structures which they represent, the prominent variables driving each groups are identified, and each is given a label to describe the residents based on these prominent characteristics (Table 4.7). The results highlight distinct populations in the city which align well with the spatial patterns presented in Figure 4.8, based on local knowledge. OAs in the city centre and inner-city areas are represented as densely populated, young and ethnically diverse, characteristics which are replaced by indicators representing families, increasing wealth and ageing in-line with a move out towards the suburbs, and beyond (see Appendix A.1 for reference). These findings also align with both traditional city structures and contemporary expectations of a modern UK city, as described by Burgess.'s (1925) concentric rings theory (see Section 2.3.1) and Thomas et al.'s (2015) discussion of modern city living in the UK, respectively.

Whilst these results align in the most part with patterns typical of most cities, through its re-classification of OAs classified as “Cosmopolitan” in the 2011 OAC, the LSOAC also introduces a level of nuance which was not present in the nationally derived classification, and which is potentially benefiting from an increased local context. The relationship between the original 2011 OAC Supergroups and the re-classifications which are derived through the LSOAC are depicted in Figure 4.10. Though some of the re-classifications in the LSOAC almost wholly mirror a single, existing 2011 OAC Supergroup, such as the “Urbanites” and

Supergroup	Prominent characteristics	Label
A	Social renting, Unemployed, Mixed ethnicity, Single parent families	Affordable living
B	Students, Single, Private renting, Communal living, Part-time work	Students
C	Full-time employment in knowledge industries, Highly educated, Home owners and private renters, Car owners, No children, Terraced housing	Urbanites
D	High population density, Families, Mixed ethnicity, Asian and Black/Black African/Black Caribbean, Migrants, Unemployed/employed in manual roles	High density multicultural families
E	Social renting, Unemployed/employed in manual roles, Elderly/ageing, Poor health outcomes	Ageing workers
F	Aged 25-44, Highly educated or students, Single, Employed in Tech/Finance, No Children, Privately rented flats	Aspirational, young workers
G	Middle-aged/elderly, White, Married, UK born, Non-dependent children, 2 or more cars, Home owners, (Semi-)detached/bungalows	Settled, ageing families
H	Asian, Semi-detached/bungalows, Home owners, 2 or more cars, Married, Employed in Education and knowledge industries	Stable professionals

Table 4.7: LSOAC Supergroup key characteristics and labels.

"Suburbanites" (re-labelled as "Urbanites" and "Settled Ageing Families" in the LSOAC), the OAs previously classified as "Cosmopolitans" are seemingly split into two entirely new and distinct population structures in the LSOAC; "Students" and "Aspiring young workers".

This occurrence, supported by the improvements identified in Table 4.5 and Table 4.6, again, suggests a weakness in the ability of the national-level classification to sufficiently capture the uniqueness of the student population in Leeds.

The OAs assigned to the 2011 OAC "Cosmopolitans" Supergroup largely extend North-West from the city centre through the Hyde Park and Headingley areas of Leeds (see Appendix A.1 for reference). These OAs almost entirely map to the LSOAC "Aspirational Young Workers" and "Students" Supergroups derived by the LSOAC, as per Figure 4.10, and as demonstrated spatially in Figure 4.11, with the city-centre OAs mapping to the former, and almost all of the others mapping to the latter.

Local knowledge of Leeds proposes as a potential explanation for the diverging of the "Cosmopolitans", that the city's current student population are resident in and reflected by the

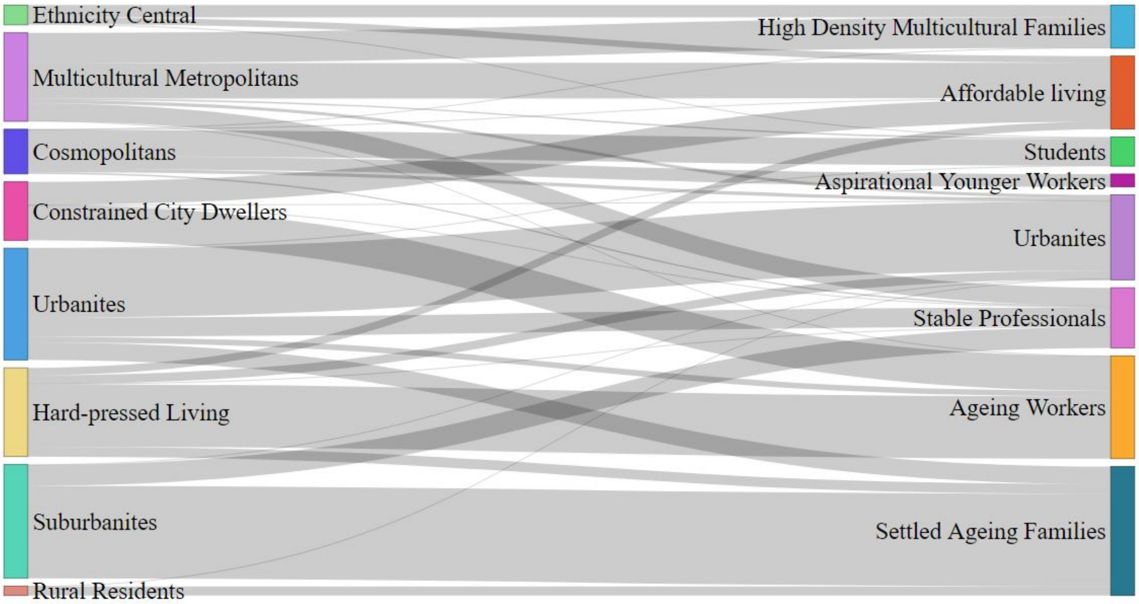


Figure 4.10: OA re-classification between the 2011 OAC and the LSOAC.

"Students" OAs, whilst the "Aspirational Young Workers" is associated with areas more popular with recent graduates and individuals in early aspirational careers. This explanation also aligns with the description of these Supergroups as per the prominent characteristics driving the classification groupings (Table 4.7). This result signifies that there are local-level characteristics in the Leeds population at the Supergroup level which are not immediately identified in the national-level classification, as anticipated, thus supporting the necessity to develop a place-specific classification for Leeds.

Figure 4.10 also reveals that the OAs classified as "Rural Residents" in the 2011 OAC do not form a similarly distinct group in the LSOAC. Instead, these OAs comprise a new Supergroup, "Settled Ageing Families", together with the majority of previously "Suburbanites" OAs and a minority of "Urbanites" and "Hard-Pressed Living". Over 13% of the UK OAs were classified as "Rural Residents" in the 2011 OAC, compared to just under 2% of the OAs in Leeds, demonstrating an under-representation in the city.

Such circumstances present the antithesis to the limitations of the national-level classification considered thus-far. In this case, in contradiction to Tobler's First Law of Geography (discussed in Section 2.2), the minority of "Rural Residents" OAs in Leeds exhibit characteristics which are more akin to other rural areas than to the nearby urban OAs. These similarities are irrespective of geographical distance. The limited quantity of these OAs restricts their ability to form a distinct group at the local extent, and as such, unlike the

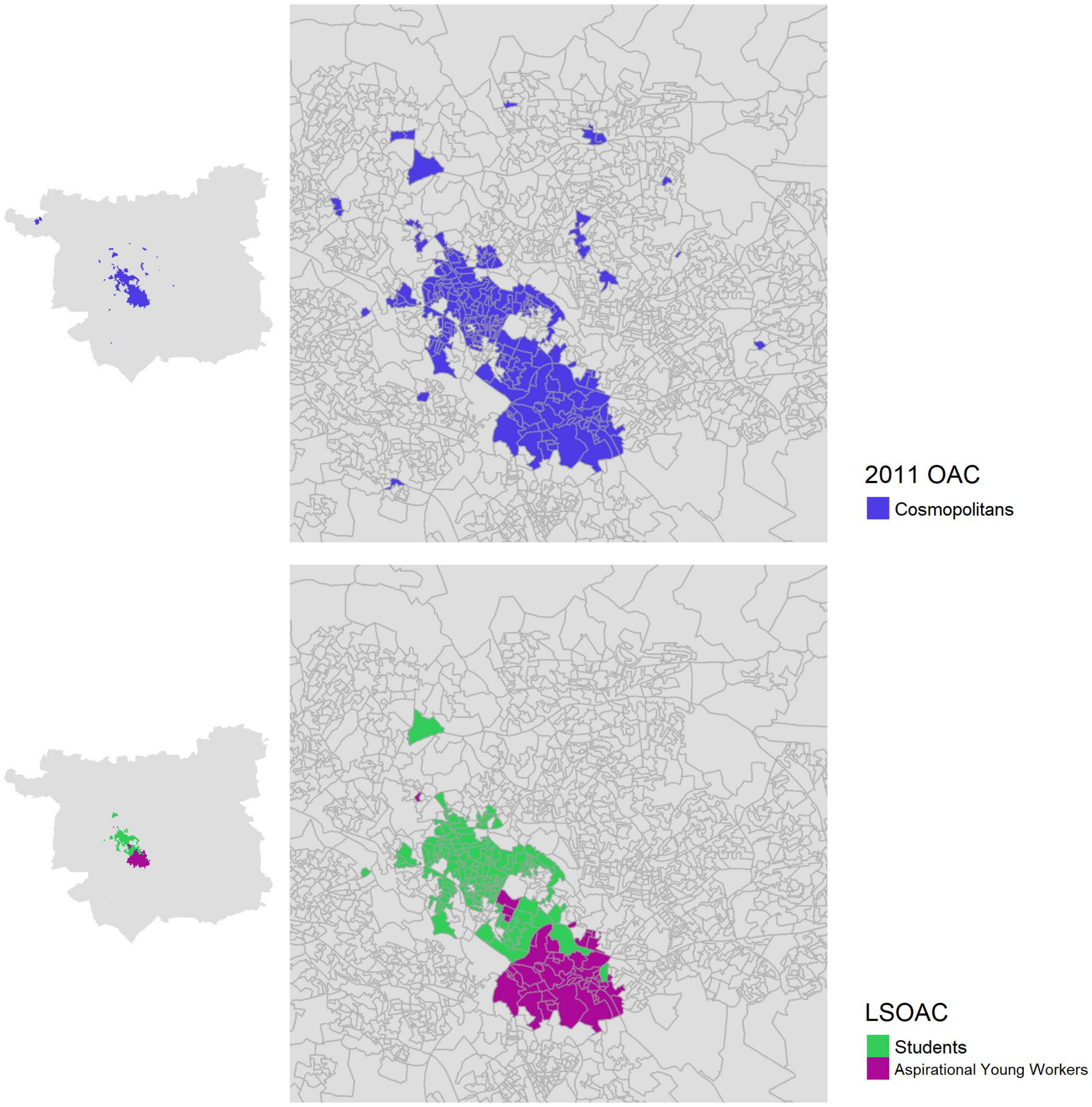


Figure 4.11: Leeds OAs classified as "Cosmopolitans" in the 2011 OAC, and their reassignment to the LSOAC Supergroups "Students" and "Aspirational Young Workers".

inner-city OAs which host uniquely local attribute combinations and which benefited from the sift in extent offered by the LSOAC, these rural OAs are poorly classified at the local extent (see Table 4.5). A similar phenomena was identified in the original LOAC study (Singleton and Longley, 2015).

4.6 Potential for follow-up tests

The application of several follow-up tests which might further validate the LSOAC results presented above, and explore the scaling issues outlined, have been considered. Each involves the adaptation of elements of the methodology to assess the impact of the changes. These have primarily focused on attempting to extend the LSOAC to derive an improved outcome for the rural OAs, for example, including increasing the number of Supergroups from 8, to see if this results in these OAs re-emerging in a distinct Supergroup, and removing the OAs classified as "Rural Residents" in the 2011 OAC, effectively re-drawing the city boundaries to see if this improves the 'fit' of the OAs classified as "Settled Ageing Families" in the LSOAC.

However, the decision is made not to explore these avenues further at this time based on a balance of the time constraints against their potential value. The LSOAC and the nLOAC present preliminary investigations into the necessity and potential gains associated with employing more local-level classifications in Leeds. As anticipated, the results encouragingly indicate that there is scope for the city to benefit from such a shift.

However, as outlined in Chapter 2 and Chapter 3, the broader focus of this thesis extends beyond a basic re-imagining and re-scaling of the 2011 OAC, as presented in the LOAC, instead the ultimate aim of this work is to take advantage of the opportunities that a change in the geographic extent affords, to develop a new framework for local-level classifications which prioritises the employment of novel and potentially more insightful local level data. Whilst it is important to be aware of the structures which the LSOAC in particular has identified, the follow-up tests proposed, which would continue to exclusively interrogate the limited input data of the 2011 OAC and seek to find the 'best' fit for these census variables, at this stage were not deemed worthwhile, and were considered to add little value. Instead, lessons will be learned from these results and used to inform the next phase of the development, which will begin to look towards the introduction of new data. Where necessary, the potential adaptations to the methodology might be tested in the future phases, if deemed appropriate or useful.

4.7 Conclusion and next steps

This chapter has generated a place-specific version of the 2011 OAC for Leeds based on the same data and methodology, which has been named the "LSOAC". In comparison with the 2011 OAC, the LSAOC has successfully identified some unique population phenomena which was not identified in the national-level classification, demonstrating the weaknesses of a national-level, one-size-fits-all approach. However, LSOAC outputs relating to the OAs classified as “Rural Residents” in the 2011 OAC demonstrate that generating the LSOAC based on an exact replica of the 2011 OAC methodology at a local level has not benefited all local populations uniformly. If applied without caution, it seems that local-level classifications could potentially be equally vulnerable to generating results as an artefact of the geographic extent, which are not necessarily meaningful and representative. Implementation of any place-specific approach must therefore be adopted with caution, ensuring a thorough understanding of the area and of the potential for these kinds of issues. Nevertheless, the broader findings do present further evidence for the necessity of place-specific classifications in the city, and in general, and as such, these recommended cautions should not prevent such development.

Analysis of the LSOAC indicates that there is considerable scope to apply the 2011 OAC variables and methodology to generate locally specific classifications which better reflect the unique characteristics of specific localities. However, restricting the classification by exactly mirroring the 2011 OAC input variables may not fully capture the nuanced nature of the diversity between Leeds OAs. Chapters 5 will therefore progress a next iteration of the LSOAC. In collaboration with LCC, additional local level data will be identified, to update and extend the 60 census variables adopted in the study presented here. This collaboration will enable the identification and use of novel small-area data from within LCC’s own data repositories, including population data which is routinely collected as part of local government activity in the city, alongside other openly available datasets. In Chapter 6, this novel data will be used to update the LSOAC, which will henceforth act as a benchmark for place-specific classifications in Leeds, and to explore the impacts on the Leeds OAs.

Summary

This chapter initiated the practical study of place-specific classifications, introducing and adopting the methodologies and outputs found within the existing literature to develop a first-attempt place-specific classification for Leeds. Although some improvements were seen, a positive endorsement for shifting towards a place-specific approach, there is evidence that the resulting classification could be improved further still. The next chapter will continue to explore this same methodology, but will look to extend the Leeds-specific classification developed here by introducing alternative, novel input data.

Chapter 5 - Introducing novel local data: Sourcing novel administrative data

5.1 Introduction

The output of any statistical process is dependent on the input data. This remains true for classification processes (Harris, 1998). Moreover, the relative success of a classification is judged by its ability to meaningfully differentiate the population. As such, the importance of the data input into a classification cannot be overstated (Rees, 1972). Relevant and meaningful input data is a fundamental component of achieving a relevant and meaningful output. Yet, despite over a decade of academic literature identifying the potential and benefits of adopting a broader source of input data for open geodemographic classification development (Savage and Burrows, 2007; Singleton and Longley, 2009b; Singleton and Spielman, 2014; Singleton and Longley, 2019), and the increasingly mainstream employment of open and public sector data in research and policy-development in a range of other contexts (discussed in Section 2.4.1 and Section 5.2.1), the most popular and widely used free and open classification, the 2011 OAC (introduced in previous chapters), remains firmly based on input data from the decennial census.

Consequently, the local level LSOAC produced in Chapter 4, which adapted the data and methodology of the 2011 OAC, was also developed entirely from the same census data. This chapter seeks to develop local, public sector specific classifications further by exploring the viability and scope for introducing new and novel data to extend the development of the LSOAC, emulating the trend seen in the commercial classifications, and building upon the recommendations in the preceding literature.

A more detailed consideration of the contemporary landscape supporting the inclusion of new data into the development of open geodemographic classifications is presented in Section 5.2. This considers the barriers which have prevented such progress to-date, and the potential that increasingly available open and public sector data now affords, followed by a discussion of the practicalities of adopting non-census data. A practical case study is presented in Section 5.3 in which administrative and open data is obtained and evaluated for its potential to derive alternative input variables to replace or extend census variables in a new geodemographic classification, generated in Chapter 6.

5.2 Background

5.2.1 Contemporary data practices in geodemographics and beyond

The commercial geodemographics industry has evolved with the current times, maximizing on the advantages afforded by increasingly available data and advancements in technology to produce more sophisticated and widely adopted outputs (Harris et al., 2005). As a result, many LAs, including LCC, elect to licence commercial geodemographic classifications developed for the public sector instead of employing the 2011 OAC, despite it being a freely available national metric (see Section 2.4.2).

However, the decennial census still acts as a foundation dataset for almost all commercial offerings, which is enriched ad-hoc with supplementary data from a range of other sources to represent additional characteristics of the population. One Experian brochure, for example, reveals that 38% of the input data for their Mosaic classification comes from the census. Though commercial sensitivities prevent the commercial providers listing the precise input data used, a review of marketing literature identifies a combination of publicly available and proprietary datasets (see Section 3.2). Acorn developed by CACI, however, claims to be the exception to this rule, positioning itself as independent of the census, having completely shifted to a reliance on alternative datasets (CACI, 2020). Whilst this is celebrated in their marketing literature, and might offer some protection in the advent of an anticipated end to the decennial census (discussed in Section 5.2.2), Section 5.2.4 presents a raft of widely accepted benefits associated with the use of census data, including coverage, completeness and accuracy, which a complete departure from the census may lose if not explicitly replaced in the alternative data.

The employment of administrative data as an alternative to traditional census variables to support better understanding of populations for specific purposes, has also become increasingly commonplace in academic research. Notable examples include studies by Webber (2007) and Lansley (2016), explicitly reviewing the potential for adopting family names and vehicle registration information obtained from the Driver and Vehicle Licensing Agency (DVLA) as surrogates for the traditional census statistics relating to ethnicity and car ownership, respectively. In both instances the use of administrative data was endorsed. Specifically, Webber (2007) noted an improvement to the capacity for distinguishing ethnicity at an increased subtlety afforded by the administrative data.

A study by Singleton and Longley (2009a) also demonstrates an example of non-census input data within geodemographics itself, in this case presenting a domain specific geodemographic classification for use in the Higher Education sector, improved by an application of relevant administrative data, primarily sourced directly from the Higher Education sector. However, the broad shift away from the census in favour of routinely incorporating non-census data in the development of a non-commercial geodemographic classification, or classification framework, with which to support public sector decision-making has not yet been made.

Just as success in the commercial sector has been driven by access to data, access is similarly important for the success of public sector developers. Any public or academic sector developed geodemographic classification capable of rivalling the commercial outputs will necessarily be built upon an equally broad set of input data. Though proprietary data underpinning commercial classifications, such as consumer transactions and credit information (Harris et al., 2005), is not openly available, the open data employed in the commercial development naturally is, thus providing an initial avenue for exploration. Examples of the alternative datasets adopted across these core providers include County Court Judgement (CCJ), electoral register data and residential property sales data from the Land Registry (Harris et al., 2005; Tate, 2018; Experian, 2018). Moreover, the government initiatives discussed in Chapter 2, which the commercial classifications have themselves already been beneficiaries of (Tate, 2018), might offer new data prospects.

5.2.2 Shifting from a reliance on the census

The momentum to shift the development of geodemographic classifications away from a reliance on decennial census data, whilst primarily grounded in the potential benefits discussed above, is also stimulated by an uncertainty surrounding the future of the census itself, at least in its current form. Reflecting the internal and external trend for re-purposing administrative data, the government has itself been considering whether the census still serves a purpose in this climate, and if so, in what capacity.

To decide, the Office for National Statistics (ONS) undertook three years of research and public consultation with key stakeholders in a project titled “Beyond 2011” (ONS, 2020b), assessing the contemporary necessity and relevance of the census. This project considered several potential futures, including a proposed discontinuation of the census from 2021 onward to be succeeded instead by reduced-scale yearly surveys supported by additional data drawn from a suite of alternative administrative sources. At this time, all proposed alternatives to the census considered in the Beyond 2011 project faced criticism. Notably, the concluding report relating to the public consultations (ONS, 2014) cited the current immaturity of the government administrative data infrastructure and methodology as a significant barrier to proposals to deviate from the current structure of the decennial census. As such, the 2021 census is set to continue unchanged, but for the format of collection methods. This will be the first “digital first” census, where participants are encouraged to complete their form online (ONS, 2021). Many of the apprehensions raised in the public consultation report, including the immaturity of the government data infrastructure, are echoed in similar discussions within related academic literature, examined further in Section 5.2.4.

Despite the outcome of the Beyond 2011 project, investment into the development of additional inter-census data from administrative sources was welcomed in the final report, based on an understanding that better data could support better decisions, and contained calls for the ONS to develop administrative data. However, it is worth noting that often data which is produced more regularly sacrifices the granularity afforded by the census, as demonstrated by the mid-year population estimates for which the lowest available geography is Lower Super Output Area, a step coarser in the census geography hierarchy. In response to the positive reception, a strategic plan (ONS, 2014) was published, promoting the development of supporting administrative data alongside the next census, and a successor project, the Census Transformation Programme (CTP) (ONS, 2020c), has been established to continue this exploration. This includes the introduction of the Administrative Data Census Project (ONS, 2020a) which, since its subsequent formation, has been explicitly producing potential alternatives and extensions to many of the census variables from administrative data available within the government. This work is documented in detail in the public domain.

A review of the outputs highlights a particular focus on developing improved population characteristics, including models of housing characteristics derived from non-census data.

All of this work is discussed at a minimum granularity of LA level, which is not detailed enough to be directly utilised here, in the development of geodemographic classifications at OA level. Moreover, the study does not look at potential issues encountered in the piecing together of disparate local data, instead, data employed in the work completed to now has already primarily been available at a consistent national extent and within repositories available to central government departments, such as the Department for Education (ONS, 2019), the Department for Work and Pensions and the NHS Patient Register (ONS, 2018b).

The work does, however, support the need for additional work to aid in progressing beyond the census, further extending the discussion and highlighting many of the practicalities of doing so. It also further indicates that the indefinite presence of the census is no longer inevitable. The consequences of such a discontinuation are potentially enormous. As things stand, should the census cease to continue, open and free classifications would also cease, unless a complete set of alternative data could be identified and gathered from new sources as demonstrated by CACI with Acorn. Though, the commercial geodemographics industry would also face substantial disruption.

As such, the future of geodemographics could depend on the identification of appropriate alternative solutions, and therefore a consideration of the alternatives to the census is becoming ever more timely. Consequently, precautionary action is recommended in preparation for the event of a post-census scenario (Singleton and Spielman, 2014). Where CACI claim to have already completed this shift, citing the Beyond 2011 project (ONS, 2020b) as its inspiration (CACI, 2020), the majority of others will find themselves with this work still to do.

The case study below demonstrates, in practice, the potential scale of this task from the perspective of developing public sector focused geodemographic classifications outside of the commercial sector. Considering a single shift from census data in just one of the five domains which underpin the 2011 OAC, the Housing domain, the work illustrates the depth and breadth of considerations which will need to be made to support the achievement of any shift from the traditional reliance on the census in geodemographics and beyond.

5.2.3 A new landscape for local and public sector geodemographics

Though the government have announced initiatives to promote the open data landscape in the UK (discussed in Chapter 2), this has not yet translated into tangible advancements of the free geodemographic classifications. This could, at least in part, be explained by the complicated data economy of the UK public sector. The Administrative Data Census Project demonstrates the difficulties in producing usable data at the geographic granularity required.

However, the Administrative Data Census Project is seeking to derive data consistent at a national extent. Much of the data collected and stored within the public sector occurs in practice at a local level, largely within LAs. As local devolution continues to progress (UK Government, 2020), this is likely to continue. The disparate structure resulting from such a system has led to a lack of consistency between regions, both in the collection processes, content and availability of data, which is not conducive to harnessing and employing data at a national level and thus could have prohibited development towards the inclusion of administrative data in national level geodemographics (Gale et al., 2016).

Naturally, this limitation is relieved in the development of a LA specific geodemographic classification, particularly when developing a standalone local classification for a single geography such as Leeds, as is the focus of this study. Harnessing the unprecedented levels of local population data available within individual LAs is a more achievable task than that faced by national-level classification developers, and consequently, the identification and use of relevant administrative data could be more achievable in the development of a local specific classification.

Nevertheless, should the case study methodology which is developed throughout this thesis be extended to develop bespoke local geodemographic classifications for multiple local regions, replicability issues might be introduced when seeking to employ any dataset which is unique to specific regions. This is not an issue which requires addressing at this stage in the exploration process, though the work carried out here, and the data recommended post-exploration for inclusion in the development of the geodemographic classification, might act as a guide for LAs who do not already gather such data to begin to do so, to support any desired extension of the methodology into additional LAs.

5.2.4 The census vs. open and administrative data debate

Although a reduced reliance on the census in geodemographics may be an attractive proposition theoretically, and might soon be necessary if the census ends (as per Section 5.2.2), the Beyond 2011 (ONS, 2014) project demonstrates the continued appeal of the census in practice, highlighting many of the factors which still need to be addressed in order to confidently embrace administrative data as a complete replacement. Many of these factors mentioned echo the enduring debate in the academic literature (Singleton and Longley, 2009b; Lansley, 2016; Longley, 2012).

The census data is comprehensive, covering almost the entire population, it is also reliable, with a well-documented, clean and transparent collection and preparation process, resulting in accurate, quality data with good provenance. This data is openly available at a relatively granular geographic aggregation, ideal for supporting reproducibility in research (Gale et al., 2016; Lansley, 2016). Conversely there is often nervousness surrounding the quality and completeness of open and administrative data, and an insecurity regarding its uncertain continuance. Whilst recent government policy recommends the development of public sector data standards to promote consistency and the effective use of data, no such standards yet exist in practice (Public Accounts Committee, 2019). Consequently, in many cases the primary purpose for the collection of an administrative dataset is linked to current policy or LA activity which is liable to change. This could result in consequential amends to how the data is collected, affecting the temporal or geographic consistency of the resulting data, or even a complete discontinuation, which could render it unreliable in longitudinal research (ONS, 2014).

There is also an increased potential for the introduction of bias from the use of alternative data sources. For example, the potential for instances of *missing* data are increased in the collection of administrative data, where individuals who are “off-the-grid”, who have limited interaction with the public sector, are unaccounted for within the data collected. Alternatively, individuals with many points of interaction may be over-represented if caution is not taken. Additionally, when employing multiple administrative datasets in place of the census, the representation of each individual unit within (be it person, household or OA) is constructed by the combination of fragmented viewpoints, which in itself could offer an incomplete picture of the individual. Moreover, this is often achieved in the linking of distinct datasets, which could itself further propagate any bias in the individual datasets, and make the bias difficult to keep track of (Longley, 2012). However, this technique could promote the use of different datasets to validate one another, or to address missing data in one dataset through the use of other datasets. As such, it is evident that there is some complexity in the employment of administrative data, and these complexities should be thoroughly understood, highlighted and mitigated for, where possible, to ensure its appropriate use (Savage and Burrows, 2007).

Nevertheless, census data also has its limitations. Administrative data is often naturally more up-to-date than decennial census data, offering a potentially more relevant representation of its subject. Longley (2012) specifically postulates that the relevance of the census data in geodemographics in particular is decreasing as populations become increasingly complex and nuanced. Instead, he suggests that lifestyle characteristics which represent attributes of the population beyond those captured within the census outputs are increasingly required to more appropriately reflect the factors which drive decisions in modern society. The evolution of the commercial classification products and the behavioural data included within their construction further supports this theory. Moreover, in many cases, administrative data offers more flexible geographic and temporal aggregations, and the restrictions created by the finitude of the census questions are lifted by the new possibility of sourcing administrative datasets containing information not traditionally collected, affording a particularly appealing increased richness and depth unmatched by the census (Longley, 2012).

5.2.5 Potential sources of administrative data

Good solutions for extending geodemographic classifications beyond the census will benefit from the adoption of a broad range of alternative novel data (Singleton and Longley, 2019). In the interest of promoting reproducibility within the public sector, much of this data should be sourced from the public sector and Open Data landscape. In the development of geodemographic classifications, careful and considered data sourcing is essential (Singleton and Longley, 2009b). This will become progressively truer in the shift to adopt increasingly more non-census data in their development, and as developers therefore seek reliable sources of new, novel datasets with which to build classifications capable of competing with the commercial offerings. Reliable adoption of administrative data, from whichever source, depends on clear provenance and good documentation. It is also essential to consider the expected longevity of the dataset going forwards. The commitment from the UK government to publish the census every ten years, prioritising consistency and comparability between releases, has contributed to its continued adoption in research, including geodemographics. Similar reassurances would thus also be desirable in seeking future alternative datasets, where possible.

As mentioned, the public sector itself has increasing availability and access to data and technology capable of supporting a similar expansion into the use of non-census data, such as data routinely collected and stored by individual LAs and central government as a by-product of other services and activity, including a variety of population data. Increasing appreciation for the value of this data and its re-employment potential have promoted a raft of local-level initiatives and the mainstream adoption of data-led policy making, mirroring the national level philosophies outlined in Section 2.4.1.

Within LCC this is exemplified in improved data collection and sharing, both internally and externally through open data platforms. These include Data Mill North (2020) (see Section 5.3.3), Leeds Observatory (2020b), a council-led website providing key data about the Leeds population and economy, and Smart Leeds (2020), a council-led programme created in partnership with the Leeds Open Data Institute (2020) to deliver new technologies and innovative solutions in the city, supported by a focus on open data and analytics. In addition, LCC collaborates closely with academics, sharing the data which underpins a range of research (Carroll and Crawford, 2020). These efforts have promoted transparency and propagated the use of administrative data, though have also highlighted some practical challenges (discussed further in the Section 5.2.6).

The research in this chapter and, Chapter 6, itself benefits from these sources of administrative data, both from the direct sharing of data from LCC partners and from the use of the platforms listed above, namely Data Mill North. Data relating to many dimensions of the resident population of Leeds are available openly on these platforms, from top-level population and demographic information on the Leeds Observatory to detailed data on communities, travel and transport, health, wellbeing and housing data on Data Mill North. Many of the datasets publicly shared represent summarised snapshots of extended datasets housed within LCC itself. However, open data sources face confidentiality restrictions which can limit the geographic granularity of the data, for example. On such occasions, repositories such as this might instead act as resources for identifying potential available datsets, which can instead be sourced directly from the original data owners, in this case, LCC themselves.

Moreover, though this data can be combined with open data published by central government departments which relate to many of the same topics, such as housing data from the Land Registry (HM Land Registry, 2016) which has already been identified as a source utilised by the commercial geodemographics developers, many of the initiatives of LCC are entirely locally focused and locally specific. Though the data combined could offer a rich tapestry of information with which to develop a profile of the residents within the city of Leeds, the wealth of this administrative data is as yet untapped in the development of open geodemographic classifications, which have thus far focused primarily on developing classifications at a national extent, as discussed in previous sections. Similarly, no evidence has been found of data such as this being incorporated into place-specific classifications developed anywhere in the UK. Unencumbered by the limitation of generating this data at national level, this study can seek to take advantage of these local-level administrative data repositories to develop a place-specific classification for the city which extends beyond the traditional data practices.

5.2.6 Overcoming the limitations of administrative data

In the adoption of non-census data within the development of open geodemographic classifications, particularly developed for and by the public sector, it is evident that the infrastructure within LAs, housing much of the potential administrative data, is increasingly important. However, many of the practical concerns raised in Section 5.2.4 which currently limit the routine use of administrative data still need to be considered and addressed. These are commonly summarised by, but not restricted to, the following themes:

- Immature public sector data infrastructure
- Lengthy data sharing processes restricting speedy use of data
- Uncertain data provenance and quality
- Inconsistent documentation of data
- Unsupported linking of data across disparate datasets

- Difficulties identifying available data across departments
- Confidentiality issues introduced by individual or household level data

Despite LCC having adopted a culture of data-led policy development, and as such being, in the main, a forward-thinking LA, keen to support the work of academics in exploring their data to derive insights for improving their practices, in practice, there remain limitations and barriers. Though there is an increasing recognition of the value of data within LCC specifically, which has led to a growing emphasis on data-led policy making and the adoption of data-centred initiatives as laid out in Section 5.2.5, many of the issues listed above remain, and could take some time to develop and overcome. Whilst some simply add complexity to the development of innovative approaches, which might be frustrating but not insurmountable, others have the potential to derail or stifle such innovation, though individuals and teams within LCC are working extremely hard to ensure that this does not happen. However, should the census cease in the near future, the scale of the task to ready the internal data infrastructure to compete with the data quality and provenance expected by census data users, in particular, would necessitate much more support for LA teams. This would be particularly true if those expectations were increased to include the use of more frequent and regular releases of data at an equally granular geographic level. In the construction of a multivariate geodemographic classification which is rooted in generating a holistic picture of populations from data across domains, and thus across government and local government departments, the difficulties faced are potentially exacerbated.

In order to progress in this direction, current legislative and technical hurdles relating to the themes listed above are going to need to be overcome. Though this presents a number of challenges, it also offers an opportunity for the public sector to make necessary technological progression. Specifically, data sharing practices which can differ from department to department and which currently present a barrier to more fluid collaboration with academia would need to be addressed (Carroll and Crawford, 2020). Additionally, an emphasis on developing a uniformity in the methods used to collect, format and adequately document data would also require improvement to support its use. Though much of the data might be collected as a by-product of other council activity, LCC-wide recommendations could be made to ensure that potential secondary activity is considered upfront in the collection of all data to encourage the prioritisation of good practices required to ensure quality data with strong provenance. Similarly, national standardisation practices which are not yet in place (ONS, 2014) would also be crucial to support the development of national classifications, or of compatible local-specific classifications. However, each of these recommendations would add increasing workloads and pressures to LAs who are in many cases already stretched delivering existing expectations to budget. Though the desire and passion for using data to innovate within LCC is evident, wide-scale innovations as would realistically be required to achieve a complete shift in reliance from census data would require greater commitment and funding from central government to support and extend existing local level initiatives. Further, many LAs may be less progressed than LCC, so would require even more interest and investment.

The scope of the potential barriers limiting the easy adoption of administrative data in open geodemographics are thus not insignificant. The specifics of these limitations, and their potential practical impact, are best demonstrated and understood in practice. These will therefore be outlined and discussed as they are encountered throughout the following case study, highlighting the real-world difficulties which have thus far prevented progress in shifting from theoretical recommendations into a reality, and which might need to be addressed to facilitate and propagate widespread adoption of the approach.

5.3 Case Study: Practically sourcing administrative data to extend the Housing domain in a next iteration of the LSOAC

5.3.1 Introduction

The discussions in Section 5.1 and Section 5.2 have highlighted the compelling policy and operational needs for extending beyond the current 2011 OAC to include richer and more timely data into the development of an open public sector focused geodemographic classification. As referenced many times, the capacity for such a development is currently greater than ever due to the increasing wealth of administrative and open data now available within and to the public sector. The following case study represents the steps involved to acquire novel administrative data which could be used as input data in an improved classification at a local level. The data identified and gathered in this chapter will be used in Chapter 6 to extend the work in Chapter 4, recreating the LSOAC, this time including variables from these additional data sources.

The 2011 OAC (and LSOAC) is currently derived from 60 census variables relating to one of five domains, outlined in Section 3.4.2. Due to the extent of the work and the time required to practically implement such an broadening of the data beyond the census, this case study extends just one domain, the Housing domain (listed as "Housing type" in Section 3.4.2), to provide a real-world demonstration of the scope for such an extension, highlighting the practical realities at each stage, and illustrating the impact on the resulting classification.

Section 5.3.2 presents the reasoning behind the selection of the Housing domain, in particular, as the basis of this work. Section 5.3.3 focuses on the identification of appropriate and available housing data from within LCC and across the Open Data landscape, and details the datasets obtained.

5.3.2 Background

The 2011 OAC Housing domain

The Housing domain is a component in the development of both the 2001 and 2011 OACs, accounting for 8 of the 60 input variables in the build of the 2011 OAC (Gale et al., 2016): Households who live in a detached house or bungalow; Households who live in a semi-detached house or bungalow; Households who live in a terrace or end-terrace house; Households who live in a flat; Households who own or have shared ownership of a property; Households who are social renting; Households who are private renting; Occupancy room rating -1 or less.

These variables pertain exclusively to *property type* and *property tenure* (and is thus listed as "Housing type" in the list of domains in Section 3.4.2). Though statistical methods were employed in the variable selection process of the 2011 OAC, as noted in Section 3.4.2, the result of these were used as a guide to support context-led, empirical decision making. The housing variables listed above were thus chosen based on a combination of their perceived ability to reflect the income or wealth of the resident population and the built environment found in each OA, two indicators which were previously deemed important and representative characteristics (Vickers and Rees, 2007; Gale et al., 2016). Despite indications from the statistical methods employed in the variable selection process of the 2011 OAC that many of these housing variables might not aid in the identification of homogeneous populations as per the aim of the classification, they were still selected as final input variables. This decision was based on the justification that they "represented a key facet" of the 2001 OAC census domains and to maintain continuity with the 2001 OAC (Gale et al., 2016, p. 11). The latter was an uncompromised priority underpinning many of the decisions made through the build.

One might consider such a justification for including variables as unsatisfactory if they have been shown, as they were in the sensitivity testing, to negatively impact the output of the classification, particularly in the development of a classification which is to be used to understand the needs of the population and inform service provision. Yet there are many potential benefits of retaining *some* housing indicators in the development of classifications, particularly for the end-user. In the commercial sector, there is an understanding that housing indicators can support a more informed understanding of consumer behaviour (CACI, 2020). Moreover, since LAs themselves are responsible for a high degree of housing needs, an understanding of the local housing infrastructure and distribution of living situations could be extremely useful information with which to inform policy and LA activity.

Concerns regarding contemporary relevance of traditional housing indicators

As suggested above, there is some debate as to whether property type and property tenure information alone are enough to differentiate individuals, particularly in relation to their public service needs. At a national level, Webber and Burrows (2018) suggest that there has been a reduction in homogeneity in populations by house type and housing tenure over recent decades and describe a range of housing circumstances which are too nuanced to be understood based on the consideration of housing type and housing tenure alone. For example, the authors note the impact of the "right to buy" social housing scheme, which promoted home ownership among the most "better-off" council tenants, in contrast with the more recent phenomenon of young people facing greater difficulty in affording homes and often being priced out of the property market by the increase of private landlords. Additionally, they note further complexity introduced by the trend for students to rent shared houses in streets alongside young professionals.

Moreover, housing decisions are also becoming more complex at a local level in cities such as Leeds, where there has also been a recent surge in city-centre living, particularly among students and young professionals choosing to live in rented, city-centre apartments (Thomas et al., 2015; Swinney and Carter, 2018). A review of new-build property data shared by the Land Registry (HM Land Registry, 2016) supports these findings, illustrating a boom in the development of expensive city-centre apartment buildings over the past decade and introducing a new dimension to the traditional profile of people who live in "flats". Similarly, the data also indicates that both some of the least and most expensive housing in Leeds can both be categorised as "terraced" housing, further highlighting the weaknesses of focusing on property type and property tenure alone. This highlights the difficulties for LAs to generate targeted strategies and allocating services and resources based on increasingly heterogeneous population indicators.

The inclusion of additional housing information routinely collected by LCC and other data vendors has the potential to generate a richer classification result which is more relevant for the public sector use. For example, alternative housing indicators could derive a more holistic picture of populations to distinguish residential properties inhabited by students from those inhabited by families or young professionals. Alternatively, it could be used to gain an idea of the turnover of residents in an area to derive a broader context, from which a more nuanced understanding of public sector needs could be inferred. Yet, if house type alone is decreasingly representative of the residents, then it is necessary to consider other potential housing characteristics which might be more relevant if the Housing domain is to remain in the generation of future classifications.

Justification for an extension of the Housing domain

Based on the above discussions, the Housing domain seems an ideal candidate for improvement. Although the extension of all domains to include additional data is desirable, concerns regarding the the appropriateness of the variables currently comprising the Housing domain support its prioritisation above the other four domains. Moreover, extensions of this domain have already been prioritised in the commercial sector, which have benefited from the government's open data initiatives in their inclusion of open housing related data, primarily house price data published by the Land Registry. Debenham (2002) also found property transaction data to be a useful addition in the development of his classification, which likewise explored the potential for including non-census variables into geodemographics, highlighting the transaction data as important drivers of the clustering process. Although both the commercial classifications and Debenham's (2002) study have singled out the housing domain for such extension, the data considered has been limited to property transaction data, and has not explored the wealth of housing related data available within local government.

In a practical sense, administrative data relating to households is more readily available and linkable than data relating to individuals, supporting the construction of a tapestry of datasets. Additionally, not only does LCC have access to a variety of housing related data, but by its nature, it is related to households and not at an individual level, thus reducing its sensitivity and the work required to account for many potential issues of confidentiality. Additionally, many of these datasets already contain a consistent Unique Property Reference Number (UPRN), enabling the linking of this data on a house-by-house basis. The Beyond 2011 consultation report (ONS, 2014, p. 20) flagged the inability to link data as a barrier to the more frequent use of administrative data. Specifically, the report highlighted that countries where the use of administrative data was more mainstream, it was typically enabled by a "population register" which helped link the data, similar to the housing UPRN. Where the data is aggregated to OA already, this necessity is reduced, as the linking can be done on the OA code, unless there is a desire to develop composite variables from information held in different datasets, which is a reasonable thing to want to do and has been done here within a single dataset (i.e. grouping house types). This could also be important in identifying duplicate data in the same dataset, facilitating the identification of the same UPRN in a single dataset, for example. Moreover, the work carried out based on this methodology can be reproducible in other LAs where similar data is available, since UPRNs are adopted nationally.

Furthermore, the near one hundred percent completeness of census data (Lansley, 2016) which appeals to many users (see Section 5.2.4) is similarly available in much of LCC's housing related data, which contains records of every house in the city. This domain thus seems to present a good opportunity to test the inclusion of LCC's existing datasets, demonstrating tangible potential for the secondary use of administrative data.

5.3.3 Sourcing and selection of alternative data

Data collection: Methodology and sources

In order to develop a classification which moves beyond data solely sourced from the census, alternative datasets including variables with which to directly replace, or enrich and extend, the current census data must be identified and obtained. Here, these are sought from across the UK's Open Data landscape and from the datasets held internally within LCC. To maintain consistency with the data adopted from the 2011 OAC, and supported by the discussions presented in Section 3.3.2 (STEP 1) regarding the appropriate geographic granularity for a public sector classification, variables are derived from these datasets at OA level.

Though it is possible to infer OA level data from a coarser geography, in doing so, the accuracy of the data is reduced. Moreover, such preliminary preparation of the data could introduce errors, which might then be further propagated through the classification process, undermining the reliability of the final result (Gale and Longley, 2012). Additionally, converting previously aggregated data back to a more granular geography through inference methods poses the risk of introducing the MAUP, discussed in 3.3.2 (STEP 1). Consequently, the gathering of appropriate and available housing datasets is carried out with geographic granularity as a priority, alongside a desire for data as recent and up-to-date as possible. Thus, data is sought at OA level or below, to support the potential for aggregating up to OA level.

Since there does not currently exist a known comprehensive repository, or list, of LCC data, datasets are identified through a largely two-fold approach: first, exploring and obtaining datasets already known to the partners and existing contacts within LCC; and secondly, reviewing open data platforms, specifically Data Mill North (2020), from which datasets are either directly obtained, where openly available, up-to-date and at the appropriate geographical granularity, or otherwise requested in the required format from the listed LCC data owners.

Data Mill North (2020) is a collaborative open data website originally developed by LCC with backing from the Cabinet Office's Release of Data Fund. Its purpose is to provide a platform for openly sharing data for the North of England, enabling individuals and organisations to combine and explore datasets and seek to gain a deeper understanding of the region and the problems it faces. Whilst LCC seek to keep Data Mill North updated with a range of information collected as a by-product of their activities and interactions with residents, the site does not hold data collected within all departments. Moreover, in many cases of data which are uploaded to the platform, often the data is not the most up-to-date version or is only available at an aggregate level to adhere to confidentiality procedures. Additionally, many of the datasets are not supported by relevant metadata or supplementary background information to ensure their appropriate use. As such, in this study, the platform is largely used as a resource to aid in the identification and compilation of a list of potential datasets, and for gaining the details of the data owners (as described above). The relevant individuals at LCC are subsequently contacted in relation to obtaining the data, as required. Each are able and kind enough to answer questions relating to all aspects of the data. Where necessary, elements of these discussions are presented alongside the details of the data in the sections to follow.

Other specific means used to identify particular datasets are outlined alongside the data.

Summary of datasets obtained

The extensive exploration of open data repositories and discussions with LCC partners and representatives in relevant council departments result in the identification of 5 appropriate and available datasets of interest (Table 5.1). Each dataset represents a snapshot at the time of collection.

These datasets are selected for their potential to either replace or extend the census variables which currently comprise the Housing domain in the 2011 OAC. Improved data on house type and housing tenure is already successfully adopted in some commercial classifications, in addition to bedroom count, age and price of property and sales trends data (Tate, 2018). The data gathering stage here seeks to replicate the inclusion of many of these variables, where possible, alongside more novel data, which is exclusively available within the public sector. Each dataset introduces additional context into the classification to offer a more nuanced differentiation of the population, arming the end-user with a richer picture of the housing circumstances of the residents in each area with which to support more targeted and bespoke servicing of the unique needs across the city. The potential benefits of each

Data description	Details included	Date of data	Source
Property transaction records at household level	Property address Property sale price Date of sale Property type Property age Transaction type	Jan-1995 to Dec-2019	Land Registry
Council Tax data at household level	UPRN Property address Property type Council Tax band	Nov-2018	LCC
Council Tax exemption records at household level	UPRN Property address Exemption type	May-2018	LCC
Council managed social housing records at household level	UPRN Property address Property type	Oct-2018	LCC
Council managed social housing rental information at OA level	Count of properties Average tenancy duration Average rent Count of properties advertised for rent Average bids for advertised properties Count of long-term empty properties Average days empty property is empty	Oct-2018 (Except advertisement and bids data collected Mar-13 to Oct-17 and Mar-13 to Oct-18, respectively)	LCC

Table 5.1: Details of the datasets included in this study.

dataset are discussed throughout this case study.

In summary, the Council Tax and council managed social housing information provide a more up-to-date source of information relating to the city's overall housing stock and social renting stock, respectively. This is enhanced with the additional context in terms of the band assigned to each property and the each social property's value and turnover, offering a way of distinguishing more relevant sub-groups *within* each of the main property types and the social housing across the city. The Land Registry property transaction data offers the potential for a more accurate picture of property value than achieved through the sole use of house type as a proxy indicator, whilst the Council Tax exemption data highlights properties with unique household compositions. In some cases the latter could be more reflective of the lifestyle of the residents than an awareness of the built environment itself, and thus, could be a more reliable proxy for their needs. Notice, whilst all of the household level LCC datasets in 5.1 include a UPRN attached to each record, this is not the case for the external property transaction records obtained from the Land Registry (HM Land Registry, 2016).

Additional datasets considered

The following additional datasets are obtained and explored but are not included in further analysis for the reasons outlined.

- **House in Multiple Occupancy (HMO) licences**, issued in Leeds between November 2013 and November 2018 (supplied by LCC), comprising: UPRN; Address of each house with HMO licence; Licence issue date; Property description.

  A licence from LCC is required to let a large property as a HMO in Leeds (see Gov.uk (2020)). This dataset contains a record of the 2,632 houses actively HMO licensed in the time period outlined. These are located in 280 of the 2,543 OAs in the city. Understanding the distribution of HMOs is an attractive prospect, since these properties represent a specific household composition which is not captured within the existing Housing domain. However, properties with an active licence issued prior to November 2013 are not included in the records, thus introducing the potential for missing data. Conversely, since no end-date is included in the data, properties which have ceased to be HMOs will still be incorrectly included in the data. The data is therefore judged unreliable and the decision is taken not to include it in the development of the classification. Moreover, experts within LCC suggest that the Council Tax data might offer a more reliable record of HMOs in the city.

- **Council Tax band charges**, as of 2018 (supplied by LCC)

  The fees for each of the Council Tax bands are standard across the city. However, an additional "parish charge" is applied to some properties, at the discretion of the parish in which the address is located. This is rolled into the Council Tax charge for the property, rendering slightly different fees for some properties within the same Council Tax band which are not within the same parish. Since this difference in the final Council Tax fee is not necessarily dictated by the characteristics of the population in the area, this discretionary uplift is disregarded, wherever it applies. Therefore, each Council Tax band is considered as standard across the city. As such, the charges represent no additional information beyond the breakdown of the properties by band, which is already included (see Section 5.1).

- **Energy Performance Certificates (EPC)**, issued 2008-2015 (supplied by LCC)

  An EPC is required for all residential properties when newly constructed, sold or let. The records for these certificates hold potentially useful information, including the number of habitable rooms, the total floor area, the floor level (if the property is a flat), the count of extensions to the property and type of conservatory, if applicable. Although these would be desirable characteristics to incorporate into this study, since the data is only available for the properties meeting the requirements for a certificate (53% of residential properties in Leeds), its use is ruled out at this stage on the grounds of incomplete coverage. Alternatively this dataset could be used to identify newly constructed properties, or those which are sold or let, through the existence of the EPC. However, the associated records do not state the criteria requiring the EPC. It is thus not possible to split the data into such categories. Since the Land Registry data already provides information on all sales, including explicitly listing new build properties, it might be possible to remove the sales and new builds identified in the Land Registry data from the list of EPCs issued to identify just the properties obtaining an EPC on the basis of a letting, however, once issued EPCs are valid for 10 years and do not need updating upon each re-letting of the property (Ministry of Housing, Communities and Local Government, 2019). As such, the granularity of the information is not detailed enough to accurately derive useful inferences.

Additional datasets for future exploration

Additionally, there are several other desirable datasets which are also sought for exploration, but which prove difficult to obtain within the scope of this study:

- **Sold and demolished LCC managed social housing data**

  Any adjustment to the social housing provision in an area will likely occur concurrently with a shift in the population and their needs. Understanding that an area has recently undergone such a change could offer a context not captured by a simple count of the existing social housing stock. Obtaining this data for inclusion in future iterations of the classification is therefore recommended, if possible.

- **Socially rented private properties data**

  The LCC managed social housing information (in Table 5.1) includes properties which are rented by LCC directly, but does not include social housing provided by private landlords. This is a gap in the context provided by the existing dataset. Whilst LCC does not routinely record this information, a list of the main Housing Associations in Leeds is available. This could be used alongside the Council Tax data to identify the properties assigned to these associations. The completeness of this data could not be guaranteed, since independent social landlords and other organisations which are not included in the list would be missed, and the complete dataset could take some time to compile. However, this data is coveted.

- **Zoopla and postal data**

  Additionally, the LCC managed social housing rental data obtained does not directly contain information pertaining to private rentals, which would be a valuable addition. This is a shortcoming which is not easily addressed using data available in the open or public data space. Other studies researching the private rental market have successfully sought data from vendors such as the property website Zoopla (Clark and Lomax, 2018). In some cases, this data has also been enhanced with data from the Royal Mail redirection service, providing details of the forwarding postcode when a resident moves from the property. Again, this could offer another potential source of information relating to population change, which might add further valuable context. However, in the interest of maintaining reproducibility (Lomax, 2020), this option is not pursued here at this time.

- **Help to Buy data**

  Whilst traditionally the housing stock indicators adopted from the census have been considered as proxies for the wealth of the residents in an area, supplementing this with data relating to uptake of the government's "Help to Buy" scheme, which offers support to help first-time-buyers on to the property ladder, could offer a richer picture of the circumstances of some home-owners. This information could potentially highlight areas which are popular with first-time-buyers of this kind, offering a more nuanced indicator reflecting the life-stage or the financial status of these home-owning residents. Statistics pertaining to the uptake of this initiative are identified in the public domain (Ministry of Housing, Communities and Local Government, 2020), however, not at the geographic granularity required.

5.4 Discussion

The opening sections of this chapter included a broad discussion of the potential of administrative data in terms of its usefulness, usability and relevance in geodemographics. The theory presented in these sections was subsequently extended and explored in practice in the case study which followed. The decision was made to prioritise the extension and enhancement of the variables of the 2011 OAC Housing domain based on strong justifications which included the potential and availability of administrative housing and property data, and weaknesses of the existing 2011 OAC housing variables.

Consequently, a thorough identification and review of potential alternatives was undertaken. This included extremely detailed appraisals of the open and public sector data landscape within Leeds and LCC, specifically, and of particular datasets identified. Many of these datasets are introduced in finite detail. This groundwork was judged to be essential to guarantee confidence both in the data and in the classification result derived from it in Chapter 6, where the data is used to enhance the Housing Domain in a re-classification of the LSOAC (developed in Chapter 4).

However, the work involved, including the identification and subsequent gathering and preparation of the data to support its use, consumed many months and much careful consideration. Each phase, from the lengthy research and consultation process taken to identify potentially useful data owned and managed by LCC, to cultivating relationships with the various data managers, agreeing the structure of each dataset, confirming licence agreements, acquiring the data, becoming an expert in the intricacies of each dataset to understand any potential weaknesses inherent within, and finally cleaning and preparing the data, cost valuable research time. This is despite the emphasis which LCC places on data and data-led approaches, its respect for the potential of public sector data (which is evidenced in the strong data networks and initiatives listed in Section 2.4.1 and Section 5.2.5), and the importance that LCC places on collaborative research with the University of Leeds (Carroll and Crawford, 2020).

Indeed, the partnership with LCC which underpins this thesis was vital throughout, in addition to the time granted by the nature of doctoral research which enabled dedicated attention and the scope to build and maintain crucial relationships. These factors may be infeasible in other research settings and thus the outcomes presented here might likewise be unachievable. Moreover, whilst elements of the time burden would be reduced when replicating this activity in-house, namely time consumed agreeing the terms of licence agreements, much would not. The helpfulness, hard work and commitment of LCC to support this activity was hampered in many ways by the data infrastructure and disparate storage of data across the siloed departments, and the limited capacity of individuals and their ability to divert attention from their primary roles to provide support. These are factors which would similarly impede internal activity, and which have been identified as broader issues in local government across the UK (ONS, 2014). However, certain practical steps could be taken to alleviate some of these constraints, which will be discussed further in Chapter 6. Nevertheless, much success has been made in this chapter in identifying and selecting several candidate datasets. Chapter 6 will continue and extend this case study, candidate input variables will be derived from these datasets, variable selection procedures will be applied, and the LSOAC will be re-classified based on this extended and amended set of input variables as a test of including novel administrative data into the build of a geodemographic classification.

Summary

This chapter has demonstrated both the practical potential and challenges of extending the input data employed in geodemographic classification development to incorporate novel administrative data sourced from the open and public sector data economy. Chapter 6 will extend the work presented here to generate a new place-specific classification for Leeds which incorporates variables derived from the datasets identified here.

Chapter 6 - Introducing novel local data: Incorporating novel admin data into the Leeds-specific classification

6.1 Introduction

This chapter extends the case study in Chapter 5, taking the datasets identified and deriving relevant variables with which to update and extend the original set of census variables used in the 2011 OAC, before once again adopting the 2011 OAC methodology to generate new classifications for the OAs in Leeds based on the updated variable sets.

Section 6.2 outlines the necessary preparations to be made to each dataset, and the process of identifying and extracting the relevant variables. Section 6.3 details the process of converting each dataset to the same geographic scale, highlighting several issues preventing the replication of the household to OA aggregations made in the generation of the 2011 census statistics in the household level administrative data, and the associated implications. In Section 6.4, the 2011 OAC variable selection methodology is conducted on the candidate variable set curated, to identify variables with which to replace and extend the census variables currently comprising the Housing domain of the 2011 OAC. Finally, two new geodemographic classifications for the Leeds OAs are generated, first based on the updated and then the extended variable sets, the results of which are presented in Section 6.5 in a comparison with the results of the LSOAC derived in Chapter 4. In doing so, this chapter presents the novel development of an open classification for the city which is extended with alternative data in a transparent and reproducible manner.

6.2 Data preparation and identifying candidate variables

This section details the process of data preparation and exploration employed to identify and extract a set of candidate variables for input into the re-classification. A lengthy process of cleaning and preparing the data is initially carried out to ensure accuracy, quality and a thorough understanding before potential variables are identified in each of the raw datasets listed in Table 5.1. This section includes a thorough discussion of each of the datasets, outlining what is and is not included in each, and any weaknesses and limitations associated with their use. It also details the most significant issues which are faced in the preparation of each of the datasets during this phase, outlining the problems, explanations and solutions identified in each case. This work is explicitly documented, including details of thoughts and decisions made, to highlight the practicalities of understanding such an endeavour and to support confidence in the subsequent use of the variables derived. It also contains a record of recommendations to assist LCC, or any other LA, to apply similar approaches to their data, and to strategically collect data in ways which enable and support this form of analysis.

6.2.1 Property transaction records

Property transaction records from the Land Registry, a non-ministerial government department which deals with the registration of property and land ownership, are employed in the development of several of the commercial classifications (Harris et al., 2005). This data (described in Table 5.1) is openly available and is extracted directly from the Land Registry website (HM Land Registry, 2016). The dataset extracted contains a record of property sales in the LA boundary of Leeds for all years between 1995 and 2019, inclusive [1], a total of 333,842 records relating to 186,653 distinct properties. As detailed in Table 5.1, the data includes the full address of the property, including the postcode, the property sale price, the date of the sale, a property type indicator of either “detached”, “semi-detached”, “terraced”, “flat/maisonette” or “other”, and an indicator of whether or not the property was a new build. A transaction type of either “A” or “B” is also assigned to each record, as per the following definition (HM Land Registry, 2016):

- A - *Standard Price Paid entry, includes single residential property sold for value.*
- B - *Additional Price Paid entry, including transfers under a power of sale/repossessions, buy-to-lets (where they can be identified by a Mortgage) and transfers to non-private individuals.*

As such, transactions assigned to category “B” include both sales of commercial properties and sales of residential properties purchased specifically for private rental.

Though there have been examples of other classifications which have incorporated data

[1] There are several circumstances under which a transaction has been excluded from the data, for example, in instances of re-mortgaging or transactions mandated under a court order. These are listed in the guidance (HM Land Registry, 2016). Since the data is not complete for all properties anyway, this is not considered to be an issue.

relating to non-residential land use, for example the London Workplace Zones Classification (Census Information Scheme, 2017), the intention here is to identify variables specifically relating to residential housing. As such, property transaction data relating to the sale of commercial properties is of no interest here. However, the records relating to sales of properties for the purposes of residential renting are of interest, since they still represent residential housing stock in the area. Yet, there is no easy differentiation within this dataset of these two distinct types of sales recorded as category “B” transactions. It is therefore undesirable to simply remove all category “B” properties, as data relating to residential properties purchased for the purpose of renting could be lost.

However, an inspection of the category “A” sales data reveals that there are no properties categorised at type “A” transactions with property type “other”. Instead, all residential properties are assigned to one of the traditional property types listed. Consequently, one might assume that the “other” category is reserved for the description of non-residential properties. Nevertheless, although this is neither confirmed or disputed in an extensive review of the published documentation supporting the release of this dataset. Since the properties can not be confirmed to be residential, the properties recorded as category “B” with house type “other” are universally treated as commercial properties, which are not of interest, and thus these records are discarded.

Consequently, since they do have descriptions which suggest that they may be residential, the remaining category “B” records are assumed to represent the properties bought to let and thus are treated similarly to the type “A” properties [2]. These decisions could lead to *some* data loss, but will retain as much data as possible while increasing the reliability of the resulting dataset.

An alternative approach could be taken. The transactions data are appended to the Council Tax data (also listed in Table 5.1, detailed in the next section), which contains a comprehensive record of all residential properties in the city, which would identify all residential properties in the Land Registry data. However, such an approach would rely on linking the data on the address fields in each dataset. Since there is some discrepancy in how addresses

[2] If this data represented a comprehensive record of *all* buy-to-let properties, using this information to derive a count of private rental properties, which is highlighted as a missing information set (in Section 5.3.3), it would be an interesting additional variable. However, since this data does not have complete coverage, it cannot be used in this way. Moreover, the data also does not contain information regarding the subsequent fate of the property, and whether it continues to be rented, or whether the owner is now living in the property themselves, nor does it contain information for properties purchased prior to 2014.

for the same property are recorded between the two datasets, this is is also not a reliable solution, and could be a lengthy process, so is disregarded in favour of the initial approach outlined above.

Whilst this dataset contain records of sales occurring in 1995-2019, inclusive, a review of the documentation accompanying the data also reveals that type “B” transactions have only been “identified” since October 2013 (HM Land Registry, 2016). It is not clear whether residential type “B” transactions prior to this date have been recorded as type “A”, or whether type “B” transactions were simply not recorded at all. To avoid confusion and the potential misuse of the data, all analysis is carried out on transactions occurring in 2014 or later. This leaves a total of 80,266 transactions relating to 71,420 distinct properties (24% of the privately owned properties in the city). All transactions in this period are summarised by property type in Table 6.1.

	Category “A”	**Category “B”**
Detached	12,638	516
Semi-detached	26,574	1,689
Terraced	21,461	3,840
Flats	11,566	1,982

Table 6.1: Property types in the property transaction data for 2014-2019, inclusive, by transaction category.

Variables representing the following area characteristics are derived from this dataset: (1) Average property value; (2) Turnover of homeowners; and (3) A measure of increased population driven by housing development.

Although property type information is included in this data set, this is not used. Since the data is not comprehensive, counts of this data by property type are not appropriate to reflect counts of property types across the city.

The *average property value* can be calculated from this data for each OA based on the mean sale prices for all transactions in each OA. However, since this data is a snapshot of property values at each sale, and since property value fluctuates year-on-year (typically increasing due to inflation and other factors), it is first necessary to adjust the price paid in a given year to generate an estimate of the expected price of the property in 2019 (to match the most recent data). This ensures that the comparisons of property prices across the dataset are like-for-like.

Since this study is only modelling for properties in Leeds, a city-level percentage change in the average house price is calculated year-on-year (shown in Figure 6.1), which is applied to transactions occurring between 2014 and 2018, inclusive, to adjust for the change in property values in the city. Specifically adjusting the house prices based on property type is also considered, however, this is judged to be inappropriate based on the variance of properties captured within the same property type. An average property price for the OAs based on these adjusted house prices can now be calculated.

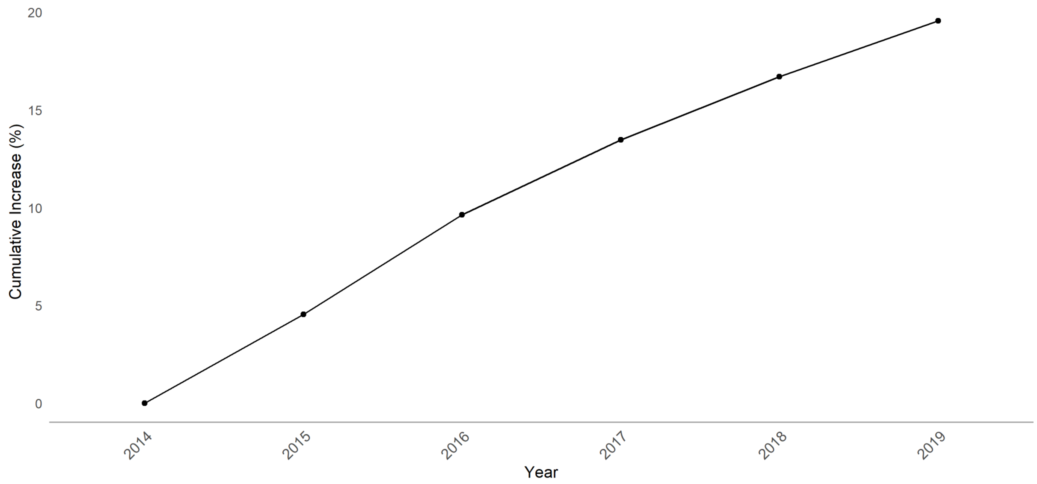


Figure 6.1: Cumulative percentage increase in Leeds house prices year-on-year.

There are 45 OAs in which no properties were sold during this period. As such, an average property price for the OA cannot be derived in this way. Instead, the property value for each postcode sector which is updated and published monthly by property website Zoopla is used as a surrogate. The average property value for each OA is taken from the published monthly values for May 2020 (the extraction date) and adjusted to derive an estimate of the property value in December 2019, for consistency. Supporting documentation indicates that the published value is generated based on a range of data, including previous sales prices, changes in market value and characteristics of the property and the local area (Stanford-Tuck, 2020), however, since the calculation is made in a black-boxed environment, and the specifics are not shared, this method is only adopted in the 45 OAs for which no average property value has been generated through the use of the Land Registry data in the primary method, above.

A *measure of turnover* can be inferred by calculating a ratio of property sales in an area to the total properties in the area, taken from the Council Tax data (listed in Table 5.1) (after removing the socially rented properties which cannot be sold, taken from the Council Managed Social Housing rental data, also in Table 5.1). Since the data is most reliable for 2014-2019 (inclusive) this is not taken as a measure of the average length of time individuals live in the area, but a measure of the sales over this period. Whilst this provides a sense of both the level of new residents in an area and the number of properties which have been inhabited for over 5 years, it does not provide information regarding of the number of years a property has been lived in beyond this period.

A sense of *housing expansion* in an area can be derived from the recorded "new-build" field in the property transactions data. This can be used to indicate areas of expansion to aid in the delivery of the additional public services driven by an increased population. New housing introduced in an area can substantially affect the local population structure (Debenham et al., 2003), this could thus be crucial for a public sector end-user with the provision of services and resources to meet the unique needs of residents. Whilst the introduction of new properties through housing development is not the only indicator of an increasing population, which could also occur in the conversion of a single property into flats or through home extensions, it is likely the most significant factor which directly indicates an increase in population and a subsequent stretching of existing resources. Since the data is readily available in this dataset, it makes sense to also capture this dimension for further exploration.

Again, this measure is derived from the transactions occurring between 2014-2019 (inclusive). Although this is somewhat arbitrarily led by the data, in future iterations it could be led by literature or LCC experience of the length of time that it takes to develop the infrastructure in an area, both physical and for the resources needed to service a sudden population increase.

Therefore, the 3 variables derived from this dataset are: (1) *Average property value of each OA*; (2) *Turnover rate for homeowners in each OA*; and (3) *Rate of residential housing expansion in each OA*. Each of these variables represent new insights not captured by the census variables used in the development of the 2011 OAC.

6.2.2 Council Tax data

This dataset contains information for all 353,860 properties in Leeds, compiled in the process of Council Tax collection. The data contains property types and Council Tax bands, summarised in Table 6.2. There are discrepancies when comparing the total count of properties in this data to the total households counts in the census data. Some of these discrepancies can be attributed to the building, demolishing and adapting of properties in the intervening period. Experts within LCC believe the Council Tax data to be a more accurate reflection of the housing stock in the city , thus this dataset is already employed internally to derive insights for policy development and recommend it as as a reliable alternative source of housing data (highlighted in discussions with LCC data managers).

		Council Tax Band							
Property type	**Total**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Semi-detached	120,020	29,018	31,127	42,739	11,437	3,924	1,296	442	37
Terraced	101,566	56,581	25,698	11,624	4,922	1,882	635	216	17
Self contained flat	80,708	49,369	15,467	8,416	5,547	1,422	363	114	10
Detached	45,898	432	561	4,828	11,850	13,923	7,633	6,131	540
HMO Parent	3,632	932	2,310	256	78	34	13	5	4
HMO not further divided	1,284	574	371	186	90	40	15	6	2
Care/Nursing homes	263	10	19	7	8	32	44	56	87
Caravans	160	160	0	0	0	0	0	0	0
HMO bedsits	65	65	0	0	0	0	0	0	0
Residential education	61	6	15	9	6	7	3	8	7
House boats	8	7	1	0	0	0	0	0	0
Hospitals and Hospices	3	0	0	0	0	0	1	0	2
Prisons	3	0	0	0	0	0	0	0	3
Sheltered accommodation	2	2	0	0	0	0	0	0	0
N/A	2	0	0	1	0	1	0	0	0
Privately owned holiday caravans/ chalets	1	1	0	0	0	0	0	0	0

Table 6.2: Summary of properties by house type and Council Tax band in Leeds.

As per the guidance from the government (Local Government, 2016), by default, bands in England are assigned to a property based on its estimated sale value in 1991. Bands for new builds and properties which have undergone adaptions affecting their size are automatically re-assessed based on additional characteristics, including the property size, layout, character, location and change in use. It is possible for any property owner to dispute their default Tax

Band and to request a similar re-assessment. This methodology is vulnerable to bias where particular individuals, properties or neighbourhoods might be more aware of this option, or more likely to appeal. Moreover, the assigned bands may be inappropriate in areas in which the property value has not moved in line with the overall trend, and thus may be less representative. Future work might seek to identify, measure and mitigate for this bias. For now, the potential value of this data is considered to outweigh the concerns raised.

The *property type information* could be used to replace and extend beyond the four census property types currently used in the 2011 OAC Housing domain, (1) detached properties/bungalows, (2) semi-detached properties/bungalows, (3) terrace/end-terrace properties and (4) flats.

Property types with low counts

Replicating the 2011 OAC methodology as outlined by Gale et al. (2016), it is necessary to remove property types which are affected by Statistical Disclosure Controls (SDC) (introduced in Section 3.4.2), i.e. property types represented by small counts. These include "Houseboats", "Privately owned caravans and chalets", "Sheltered accommodation", "Caravans" and "Care/Nursing homes".

Combining these into an "Other" category is considered, however, since the different property types are reflective of different residents, such a category would not represent any specific population characteristic and thus would not add anything to the analysis. Such an "Other" category might therefore skew the analysis by generating a cluster based on very disparate households. This option is therefore decided against, and the five property types listed above are removed from the data.

Communal living and Housing in Multiple Occupation

Four additional property types are judged as irrelevant as they do not reflect characteristics associated with clearly identified residential dwellings: "Hospitals and hospices", "Prisons", "Communal Residences (e.g. hostels, Refuge Centres, Convents and Monasteries)", and "Unknown (where house types are unknown)". The latter also represents a very small number of properties.

Additionally, manual inspection reveals that the property type "Residential Education (e.g.

Boarding school accommodation and Halls of Residence)" does not appear to include University halls of residence, as the name might suggest. This indicates a quirk in the management of halls of residences by universities in the city and private providers and the system by which LCC record or license these types of properties. It also flags a further area of caution to be undertaken when using administrative data which employ common terminology, particularly in instances where secondary or external users might benefit from additional insights to avoid misunderstandings or inconsistent expectations. This property type is therefore also removed from the final list of variables drawn from this dataset.

There are also several house types in the Council Tax data listed as Houses in Multiple Occupation (HMO). These are rented properties, typically short-term rentals, shared by multiple individuals with individual rental contracts. This is not a *property type* in the traditional sense as the description does not relate to the physical infrastructure, but to the composition of the residents. This data is provided in lieu of a traditional property type. In the case of a HMO, the property itself is being used in a different way to a non-HMO property. The living situations of residents in a HMO in any property type will more likely align with one another than with residents in non-HMOs of the same property type. For example, residents in a HMO which happens to be a semi-detached will not use the property in the same way as a resident in a non-HMO semi-detached property. As such, understanding that the property is a HMO will help infer more about the circumstances and situation of the residents than it will to assign the traditional property type to the HMO. Moreover, if property type is to be taken as an indicator for wealth, or even infrastructure, it is potentially misleading to treat non-HMO and HMO properties in the same way.

HMOs are therefore treated as a separate property type of their own. Replicating the methodology in the 2011 OAC (Gale et al., 2016), a new 'combined' HMO property type variable is generated by summing three HMO property types found within the Council Tax data: "HMO Parent"; "HMO not further divided (where bedsits not recorded)" and "HMO Bedsits/non self-contained accommodation". Since every property is listed in this dataset exactly once, these property types are mutually exclusive, thus making it possible to combine these types into a single type.

However, consultations with LCC have indicated that the classification of a property as a HMO or otherwise can be complex. It is possible for individual rooms within a HMO to be classified instead as a “Self-contained flat” in the Council Tax data if the room contains access to a private toilet in addition to even limited cooking facilities, for example, a private microwave. This can be true even if the property containing the room includes a communal kitchen. Moreover, these classifications are regularly retrospectively applied and made ad-hoc, often as a by-product of unrelated interactions with the council. As such, the definitions of a HMO are applied loosely, and the classification itself can be somewhat inconsistent.

Moreover, manual inspection of the data reveals information relating to university student halls contained within the HMO data. However, it is not possible to exclusively identify and extract the student halls to remove from the HMO information, or to use as an individual variable in itself. This should be noted, however, as it could assist in the interpretation of the results to understand that the HMOs identified might reflect student halls. To support the use of this data in the future, LCC might benefit from comprising a separate list of student halls including UPRNs to assist in this extraction and to add further nuance to the HMO data.

Nevertheless, though the data may not capture all possible HMOs in the city, and whilst this data combines traditional HMOs with student halls, there is potentially valuable information to be gained in including this data in the geodemographic classification.

Council Tax bands

The *Council Tax band* information could also add extra information above and beyond the property type. Council Tax bands are assigned to residential properties based on a review of several property features, including size, layout, character, location and property value. The bands are on a scale from A to H, where H is typically assigned to the largest, most valuable properties (Local Government, 2016). Each band indicates the amount of Council Tax the tax payer(s) who are responsible must pay. The fees are consistent for all properties assigned to the same band, regardless of geography, but for a separate “Parish charge” which is added in some OAs (see Section 5.3.3). Since this additional charge does not reflect the characteristics of the resident population, bands are considered to reflect all of the properties allocated to them in the same way in this study. This information is thus of potential value in the development of a geodemographic classification since it indicates homogeneity between properties in the same bands based on a more holistic mix of characteristics which are not necessarily reflected solely in the consideration of property type.

As demonstrated in Table 6.2, it is possible to also derive a variable based on the combination of each property type with each Council Tax band (e.g. "Band A Detached property", etc.). However, this level of detail is judged to be unnecessary and potentially unhelpful. The numbers in each of the resulting categories would be small, and since both the property types and Council Tax bands are to be included in the analysis already, this extra level of detail would likely introduce redundancy.

Therefore, the 13 variables to be derived from this dataset are: (1) *"% of detached properties in each OA"*; (2) *"% of semi-detached properties in each OA"*; (3) *"% of terraced properties in each OA"*; (4) *"% of flats in each OA"*; (5) *"% of HMO properties in each OA"*; and (6-13) *"% of properties in each Council Tax band in each OA (A-H)"*.

In addition to the overall discrepancy in total household counts, further discrepancies are introduced when the above variables, aggregated at OA level, are compared to the household statistics for each OA as per the census data. These discrepancies are discussed at length in Section 6.3. Though not ideal, the issues which have resulted in the differences are unavoidable and all action is taken to mitigate the impact, thus no further action can or is to be taken in an to attempt to improve the aggregation.

6.2.3 Council Tax exemption records

This dataset contains both a list of addresses in the Leeds LA boundary which are HMOs, and for which the owner is liable for the payment, and a list of the addresses that are subject to a student exemption in Council Tax. Though datasets relating to other Council Tax exemptions and discounts exist, this is the only such dataset which is easily interpretable and is not deemed too sensitive to obtain, such as is a barrier to the use of disability discount, for example. The single resident discount is considered, however, in addition to properties containing individual residents, this exemption can also be applied to households containing a lone parent, or containing only one individual who is not eligible for any other exemptions, for example, in the case of co-habiting couples where one of the parties is a student. These instances reflect very different living situations to households occupied by a single individual alone, and as such, limit the usefulness of the single resident discount.

The primary purpose of the compilation of this data by LCC was to inform planning use and was thus not intended to represent a definitive list of all HMOs in the city (Data Mill North, 2020). Each record contains a UPRN for the property, a complete address including the postcode, and the exemption code under which the property has been recorded as exempt from Council Tax based on the circumstances of the residents, from one of three possible codes:

N = Property solely occupied by full time-students

M = Purpose built student accommodation solely occupied by full-time students

OL = (Owner Liable) A dwelling inhabited by persons who do not constitute a single household where the owner is liable for Council Tax (a HMO property)

The "M" and "N" codes relate to the *residents*. Properties with "M" and "N" exemptions are mutually exclusive and thus can be combined to represent all properties which are inhabited solely by students. It must be noted, however, that some accuracy in this data could be lost if eligible students are not claiming this exemption, as in many cases, this relies on action from the eligible student themselves. This is also a count of *properties* and not *individuals*, so properties with many students living within would be counted equal to a property containing a student who lives alone. Moreover, the data also only contains information relating to properties solely occupied by students, and not residences which are shared by students and non-students, for example, students with a non-student partner, or students living in their family homes. However, it has been decided that this data could still prove useful in highlighting 'traditional' student areas, which is important in Leeds in particular. Student populations can live very differently to other resident populations cohabiting within close proximity (Tate, 2018). Students are also likely to live in HMOs, but again, in distinct circumstances to other HMO compositions, such as those inhabited by young professionals. An ability to thus identify areas based on presence of HMOs in addition to a presence or absence of students could be extremely informative. Moreover, local knowledge suggests that there has been some migration of student populations in the intervening decade since the 2011 census. Thus, this could be a useful variable for inclusion in the development of the classification, particularly in providing a more up-to-date picture of student residents beyond the census variables.

The "OL" code relates to the *property*. This code is assigned to HMO[3] properties such

[3] This is the third dataset considered here which makes reference to "HMOs", though there is inconsistency in the use of this terminology, and conflicting definitions have emerged through further consultations

that the occupants have separate tenancy agreements. Thus, under Council Tax legislation, the *owner* is liable for the Council Tax, rather than the tenants. The “OL” exemption is therefore not mutually exclusive to the “M” or the “N” exemptions, which are applied in relation to the circumstances of the *occupants*, since a property can be both owner liable *and* solely occupied by students. Thus, there are properties duplicated in the data where the property is a HMO solely occupied by full-time students. However, these duplicates are retained. Their removal would be considered if it might add information, for example in the identification of HMOs which contained no students, however, this would not account for mixed student and non-student households, and thus would not have provided an additional source of information.

Since this dataset is unlikely to contain a comprehensive record of HMOs in the city, and since this information is already captured by the Council Tax HMO property type information derived in the previous section, the “OL” exemption is not employed directly. However, this dataset does contain 1,088 records relating to properties which are not classified as HMOs in the Council Tax data. Further inspection suggests that these records might pertain to some of the HMOs in which the individual rooms are recorded as self-contained flats in the Council Tax data due to their access to qualifying private facilities, and are thus missing from the Council Tax HMO records. As noted, these properties are still essentially lived in as HMOs, and as such, it is desirable to include a variable which might be able to differentiate these properties from self-contained flats in the traditional sense. Whilst it is, again, not possible to be entirely sure that this dataset captures all properties of this type, the “OL” records which fall into this category are retained as a new “HMO Self-contained” variable, as a best-fit proxy variable. Recommendations are made to seek to improve upon the data used for this variable in future studies, be it by evaluating the comprehensiveness of this method, or by identifying properties exhibiting this phenomena via any other method.

Therefore, the 2 variables to be derived from this dataset are: (1) *“% of properties with either an ‘N’ or ‘M’ student exemption code in each OA (i.e. % of properties solely occupied by students)”* ; and (2) *“% of properties with an Owner Liable (‘OL’) exemption code in each OA which are not classed as HMOs in Council Tax data (i.e. the ‘HMO Self-contained’*

with LCC. Though these discrepancies are thoroughly researched and addressed in this study, mitigating against misunderstanding, the potential for misunderstanding in the use of administrative data is present wherever conflicting definitions widely used terminology occur. Thus caution is recommended when adopting terminology.

variable)".

6.2.4 LCC Managed Social Housing data and rental information

This dataset relates to LCC owned, socially rented properties. This data comprises two individual parts, (1) a list of all 55,628 socially rented properties under LCC management in the city and (2) an aggregation, by OA, of council property values and other rental information e.g. count of bids [4] and length of empty properties (see Table 5.1).

The *list of council properties* contains a UPRN for each property and the full address, including postcode. It also includes information on whether the property is categorised as "Extra Care" or "Sheltered" housing. For those which are not, this dataset also includes information relating to the size of the property, by bedroom count.

This data could therefore be linked by the UPRN to the Council Tax data to get property types. However, this would not add very much beyond the variables already derived. Instead, pulling out the bedroom count could be used to add further context and potentially act as a proxy for household composition for the social renters, which is not represented yet in any of the other new datasets described in this section. This is also true of the information on "Sheltered" and "Extra Care" social housing. Counts of these categories in this dataset can be seen in Table 6.3.

Property category	**Total count**
Bedsits and 1 and 2 bedrooms	33,751
3 bedrooms	15,603
4 plus bedrooms	1,932
Sheltered and Extra Care housing	4,342

Table 6.3: Summary of LCC managed social housing by category.

To account for SDC, the bedsits are aggregated with the 1 and 2 bedrooms. These were already aggregated in the original data. Properties with 3 bedrooms are retained as a category of their own as they might represent a potentially different household composition to 1/2 bedroom properties, and SDC does not necessitate their aggregation. Any property with 4 or more bedrooms are again aggregated, to account for SDC. Moreover, these represent large house with more than the average number of bedrooms (Local Authority Building Control, 2018).

[4] Prospective tenants can see a list of the properties which are advertised for rent, and subsequently 'bid' on properties, to express their interest in renting the property.

"Sheltered" and "Extra Care" housing is also of interest as it represents another attribute of the resident population. Though these properties do not list the number of beds, this is a different type of social property altogether which adds another interesting dimension to the data. Currently the Housing domain in the 2011 OAC contains a variable representing the percentage of households self-reporting as socially renting in each OA, thus this data might act as a more timely replacement.

As mentioned in Section 5.3.3, LCC also hold information on LCC managed social housing properties which have been sold and demolished. Access to this information could support the existing social housing data by demonstrating a changing population in a given area. It could also help explain any discrepancies between the levels of social renting recorded in the census and the LCC data. However, interdepartmental difficulties prevent this data from being obtained. This is not helped by the disruption caused by the onset of the Covid-19 pandemic. This data is therefore not included in this iteration.

The *aggregated rental information* relating to the LCC social housing properties contains the following information, by OA: Count of properties; Average tenancy duration; Average rent; Count of properties advertised for rent; Average bids for advertised properties; Count of long-term empty properties; Average days that an empty property remains empty.

The average duration of the tenancy per OA indicates a turnover rate of the socially renting residents in OA, which is an interesting and useful social characteristic to understand about an area. The average rent of the properties in an OA is a useful indicator which is similar to, and could extend, the house price indicators adopted by many of the commercial classifications and discussed in Section 5.3.3. The number of properties advertised in each OA does not show anything particularly useful since it does not differentiate between the same properties being re-advertised multiple times and different properties being advertised just once. The average bids received for the properties advertised for rent in each OA is initially considered as a proxy for the popularity of an area, however, it is deemed an unreliable proxy. Whilst high bids might represent popularity in an area, it might also reflect the popularity of the properties themselves, which could be based on features other than the location of the property. Moreover, low bids might incorrectly suggest unpopularity in areas where properties were simply not advertised due to low turnover, i.e. where the properties did not go on the market, which could actually be an indication of a well-liked area with low turnover. There are also a number of other potential concerns relating to the inference of popularity from this data, and as such, this variable is not used going forwards.

Likewise, the variables relating to empty properties (count of long-term empty properties and average days that an empty property remains empty) are also not used for this study.

Therefore, the 7 variables to be derived from across both the LCC managed social housing individual and aggregated records are: (1) *"% of social housing in each OA"*; (2) *"% of social housing which are classified as Sheltered Housing or Extra Care properties in each OA"*; (3) *"% of social housing which are bedsits or 1/2 bedrooms in each OA"*; (4) *"% of social housing which are 3 bedrooms in each OA"*; (5) *"% of social housing which are 4 or more bedrooms in each OA"*; (6) *"Average tenancy duration of social housing in each OA"*; and (7) *"Average rent of social housing in each OA"*.

These complete a list of 25 total candidate variables for inclusion in the build of the geodemographic classification (see Appendix C.1, for the complete list).

6.2.5 Additional data caveats

As identified in Section 6.2.2, the classification of HMOs are open to interpretation and thus potential inconsistency. Notably, the Council Tax data contains records of "Self-contained flats" which, for all intents and purposes, are simply rooms in a HMO with access to private facilities. In such instances, the property is lived in as a HMO, and most likely, upon sale, the entire property will be sold as a single unit. However, the Council Tax records issue a UPRN to each flat, and lacks the information required to re-group as a single HMO. As such, in an area with many HMOs of this kind, though the Council Tax data might reflect an area with many self-contained flats, the Land Registry data could indicate this as an area with high property value, reflecting the property values of the entire HMO. This could be important to note for interpretation of the results, since the composite property profile of such an area could closely resemble an area with many genuine high-value self-contained flats, though the resident profiles of the two areas are likely to be very different. Further consultations with LCC conclude that the Council Tax bands in this case might also not assist in the differentiation of these two areas as there is often insufficient nuance in their allocation. Local knowledge could therefore be useful in generating careful interpretations where this phenomena might actualise.

This issue tangibly illustrates the challenge and complexity of working with administrative data. Whilst a deep understanding of any such data employed is required to facilitate confidence in both the input data and the final result, it is becoming increasingly evident that gaining such an understanding can be a lengthy and intricate process.

6.3 Generating variables: Geographic conversion issues

6.3.1 Introduction to the issues

The novel variables listed in Appendix C.1 are now derived from the associated datasets. To support the use of these new, novel variables alongside the traditional census variables in developing the classification, all variables must be derived at, or set to, the same geographic scale. Since the census variables are already aggregated to OA level, the data underpinning the new variables, many of which are gathered at individual household level, are also aggregated at OA level. However, an in-depth review of the documentation which supports the publication of the census statistics, and discussions with David Martin (2020) have highlighted that, as per the design of the census statistics aggregation, there are no means available to reproduce the actual allocation of households to OAs as per the 2011 census aggregate statistics. As such, there is no way to aggregate the households in the new household level data so as to match the aggregation of the same households in the census statistics.

As summarised by Martin (2020), immediately prior to the 2011 census in England and Wales, the ONS collaborated with address data vendors to produce a census specific address register to support the mail-out of the census forms. However, a series of confidentiality and sharing restrictions prevented its release outside of the ONS. Instead, postcode to OA directories derived from the register were published to address the gap, providing a matching between all postcodes and an OA (discussed further Section 6.3.3).

Subsequently, a joint venture between central government and Ordnance Survey saw the creation of a new organisation "GeoPlace", who oversaw the development of a high quality and definitive, licensable address register named "AddressBase" (GeoPlace, 2021) to which access is granted for LAs and central government departments. In extension of the census-specific address register and the published postcode to OA directory, AddressBase is linked specifically to household UPRNs. Consequently, continuously updated UPRN to

OA directories are now produced and published by the ONS, assisting in the more granular allocation of specific households, rather than entire postcodes, to OAs. However, since the AddressBase was derived subsequent to the publication of the 2011 census statistics, and since there is no available snapshot of the exact links between households and OAs at the time, only postcodes to OAs, there is no accurate means of retrofitting the households back to OAs exactly as per the 2011 census aggregations.

6.3.2 Potential implications of imperfect geographic matching

These discrepancies compromise the integrity of any classification built on a combination of census and administrative data, be it at OA, postcode or household level, since all geographic conversions rely on imperfect matching of census geographies to other administrative geographies. This is particularly pertinent in areas with diverse populations living in very close proximity, where even the slightest shift in geographic boundaries between datasets may result in a distinctly different population captured and represented as the same OA in each of the different datasets. Though this is unavoidable, these issues should be mitigated against, where possible, for instance in seeking to at least apply a consistent aggregation across all of the non-census data. Moreover, improvements should be implemented to reduce, or remove, this challenge in future censuses.

Since the AddressBase now exists, and will form the basis of the 2021 census address register, this should make such a mapping between households and OAs much easier in the future. However, Martin (2020) predicts that these developments are still unlikely to support a perfect reallocation of households to OAs in this or any future censuses going forwards, commenting that the complex and confidential post-census processes are never going to be replicable.

The immediate implications of the challenges with the 2011 census household to OA allocations in the context of this thesis will be illustrated in the sections to follow. The potential wider implications of these challenges remaining post-2021, as researchers and policy makers continue to put data increasingly at the heart of their decision-making processes, will be discussed in more detail in Section 6.6.

6.3.3 Introduction to available conversion methods

Based on the postcode to OA and the UPRN to OA directories published by the ONS (introduced in Section 6.3.1), there are three main potential methodologies which are commonly employed to aggregate household level data to OAs in research practices.

Method 1 involves the use of an ONS postcode to OA directory published by the ONS Open Geography Portal (2020a), or lookup table, which when filtered for the LA boundary of Leeds, matches each postcode in the city to a single OA. In this lookup table, several postcodes are often assigned to the same OA in a many-to-one relationship. Since this lookup has been established the longest, this is taught to students of GIS to support their use of census statistics, and is typically the default approach adopted by researchers. For example, it is explicitly documented as the methodology employed by Singleton and Longley (2009a) to link from households to OA in their development of a Higher Education geodemographic classification which incorporates non-census education data.

Method 2 alternatively employs the UPRN to OA directory also published by the ONS Open Geography Portal (2020b). This method may be somewhat more limited since it requires a UPRN linked to each household, however as demonstrated in Table 5.1, this variable can be routinely appended to public sector administrative data.

Method 3 involves an application of GIS methods, mapping each household atop of the OA boundaries (provided by the UK Data Service (2018)), and a spatial aggregation made of the households based on the OA polygon in which they fall.

Due to the commercial value of methodological practices employed, it is not possible to know which method is used in the development of any proprietary classifications.

6.3.4 Aggregation of the Case Study data: Challenges

Of the data in Table 5.1, the LCC managed social housing rental data is provided already pre-aggregated based on a LCC adapted version of the postcode to OA lookup. The property transaction records do not contain a grid reference or a UPRN, so can only be aggregated to OAs based on the postcode of each property, ruling out Method 2 and Method 3. The remaining datasets each contain both a postcode and a UPRN, which can also be used to link to the X,Y grid reference of each property, which is stored in the Council Tax records.

As such, it is possible to employ any of the three methods outlined in their aggregation.

However, for each method there are limitations preventing the generation of an aggregation, which perfectly matches the census aggregation. Each postcode in the ONS postcode to OA directory, employed in Method 1, is linked to a single OA in a many-to-one relationship, i.e. all households in a given postcode are assigned to a single OA (though many postcodes are often assigned to the same OA). However, in reality there does not exist such a many-to-one relationship in all cases, since postcodes were not drawn to nest perfectly within OAs (Debenham et al., 2003). Consequently, the simplifications made by this postcode to OA lookup are such that there is a credible risk of some households being mis-allocated, simply to prioritise consistency across the postcode. In these instances, it is not clear how the assignment is made, whether the OA in which the majority of the households in the postcode explicitly fall is the one to which all households in the postcode are assigned, or whether the focus is on the OA containing the postcode centroid, or any one of several alternative approaches.

There is a secondary lookup available which identifies the postcodes which straddle OA boundaries (ONS, 2020d), which is purported to support a more accurate assignment of postcodes to OAs with a little additional work, however, this table does not contain a record of which *specific houses* in each postcode are assigned to the wrong OA in the simplification, and thus cannot facilitate the re-assignment of these households to the correct OA. As such, the household to OA match provided by the postcode to OA directory will necessarily have households mis-allocated to the wrong OA.

Methods 2 and 3 offer a more granular, and thus potentially more accurate approach than the postcode lookup employed in Method 1. However, even adjusting for new build properties in each OA (based on records in the HM Land Registry (2016) in Table 5.1), there are substantial discrepancies in the use of *each* of the three approaches to match the census aggregations in Leeds. Of course, there is a risk of introducing some inaccuracies, since the new build properties must be allocated to OAs based on the postcode lookup, as the Land Registry data does not include a UPRN or grid-reference, and thus some may be mis-allocated. Moreover, the data does not contain demolitions or property adaptations which might have resulted in the creation of new households from an existing property. However, these considerations are unlikely to account for the scale of the discrepancies identified (discussed in Section 6.3.9).

6.3.5 Selection of the aggregation methods

The property transactions data (in Table 5.1) is aggregated using Method 1, due to the absence of a UPRN or X,Y grid coordinate in the data to support Method 2 or Method 3, respectively. Whilst it is desirable to maintain consistency in the aggregations of the novel data wherever possible, the scale of the discrepancy generated by Method 1 is such that the use of this method is avoided in the aggregation of the novel data where an aggregation by Method 2 or Method 3 are possible. As a pragmatic solution, Method 2 is adopted in the aggregation of the other datasets. This approach is more accessible than Method 3 since it does not require the use of specialist GIS software or skills, and thus, supports more widespread future reproducibility within LAs.

6.3.6 Closer evaluation of Method 2 outputs in Leeds OAs

As mentioned in Section 6.3.5, Method 2 is the preferred aggregation method to aggregate the household level novel data to OAs, where permissible. However, there are some issues which arise when applying Method 2 to this data, thus it needs to be employed with some adjustments.

As per their design, each OA in England and Wales should contain between 40 and 250 households, though the target is around 125 households, with the majority of OAs comprising less than 140 (ONS, 2016a; Harris et al., 2005). However, applying Method 2 to aggregate the residential properties in Leeds based on the ONS UPRN to OA lookup (where the Council Tax records detailed in Table 5.1 are taken to represent the housing stock in the city) identifies 49 OAs containing more than 250 households. An exploration of the properties aggregated to these OAs in the property transactions data and in the Council Tax exemptions data indicates that each of these OAs contain either (or in some cases both) high quantities of student halls, where rooms were not enumerated individually in the census but are recorded individually in the Council Tax records, or apartment complexes built in the years post-census. These OAs consequently contain a high quantity of households captured within the administrative datasets which are not present in the census data. However, these are genuine households located in these OAs, not quirks of the aggregation method. As such, no further action is taken to address these discrepancies.

Potentially more concerning are the four OAs which are assigned fewer than 40 households (the lower limit for households in an OA) in the Method 2 aggregation. One is assigned 20 households, another is assigned a single household and two are assigned no households at all. Since the OAs were designed explicitly around households, there should be *no* instances of such few household counts in each OA. To understand and have the required confidence in the aggregation, it is necessary to investigate these four results, and take action where required to ensure that the results can be meaningfully appended to the census data, and other administrative data, to support their input into the classification.

1. *"E00056774":*

This OA, which contains a total of 98 households as per the census counts but just 20 households in the Method 2 aggregation of the Council Tax data falls on the Leeds/Bradford boundary. Due to the black-boxed nature of the census aggregation, there is not enough openly available information to discern whether there has been a boundary related adjustment which has resulted in some Bradford households included in Leeds OAs in the census aggregation, making up the 20 or more household discrepancy. However, this seems likely. Since the focus of this case study is just the households in Leeds, no further action is to be taken to address this issue.

2. *"E00170034":*

Despite 177 households being assigned to this OA in the census total household count, just a single household has been assigned to this OA in the aggregation of the Council Tax data as per Method 2. As Figure 6.2 illustrates, the OA boundary runs directly through the Bridgewater Place apartment complex. Each apartment within the complex are geo-located to a single grid reference in the Council Tax data, and thus are similarly assigned to the same OA in the Method 2 aggregation. However, it seems that the allocation of the individual households within the complex was applied with more nuance in the census aggregation, somehow distributing the 195 apartments across the two output areas. Again, since this detail is not available, and since an OA containing just a single household is not very useful, data relating to these two OAs are therefore merged, and a single classification is derived for the two OAs.

3. *"E00169799":*

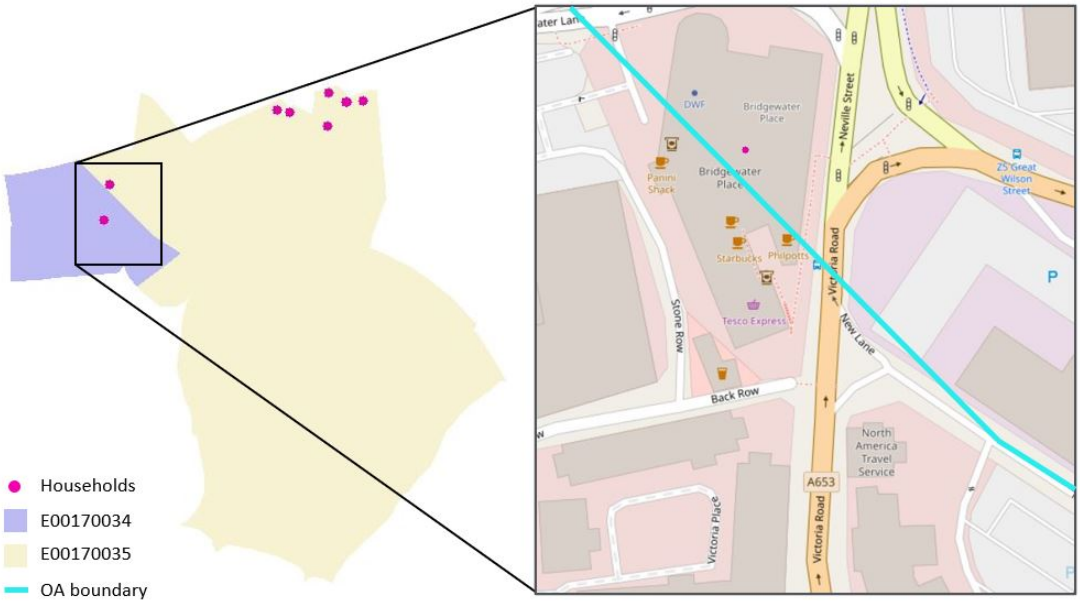


Figure 6.2: Boundary issue in OA “E00170034”

The census data reports a total of 104 households in this OA, whilst the aggregation of the Council Tax data by Method 2 identifies no houses in this area. Figure 6.3 again indicates another example of a communal housing block located across the boundary of the OA. In this case, the City Island apartments straddles 3 distinct OAs. To resolve this issue, the data for *E00169799* is merged with the data for *E00169791*, since the households in *E00169791* are closer to *E00169799* than the households located in the third OA. The merged OAs are, again, treated as a single OA in the classification.

4. *“E00169816”:*

Again, the aggregated Council Tax data derived by Method 2 has failed to allocate any households to this OA, despite the census household counts listing 173. Figure 6.4 indicates another example of a property sitting over the boundary of the OA. In this example, two apartment complexes, Riverside West and Whitehall Waterfront sit in part within the OA and in part in two neighbouring OAs. Here, the data for this OA (*E00169816*) is merged with the data from the neighbouring OA, *E00169817*, which contains fewer and geographically closer other households than the alternative neighbouring OA *E00170261*.

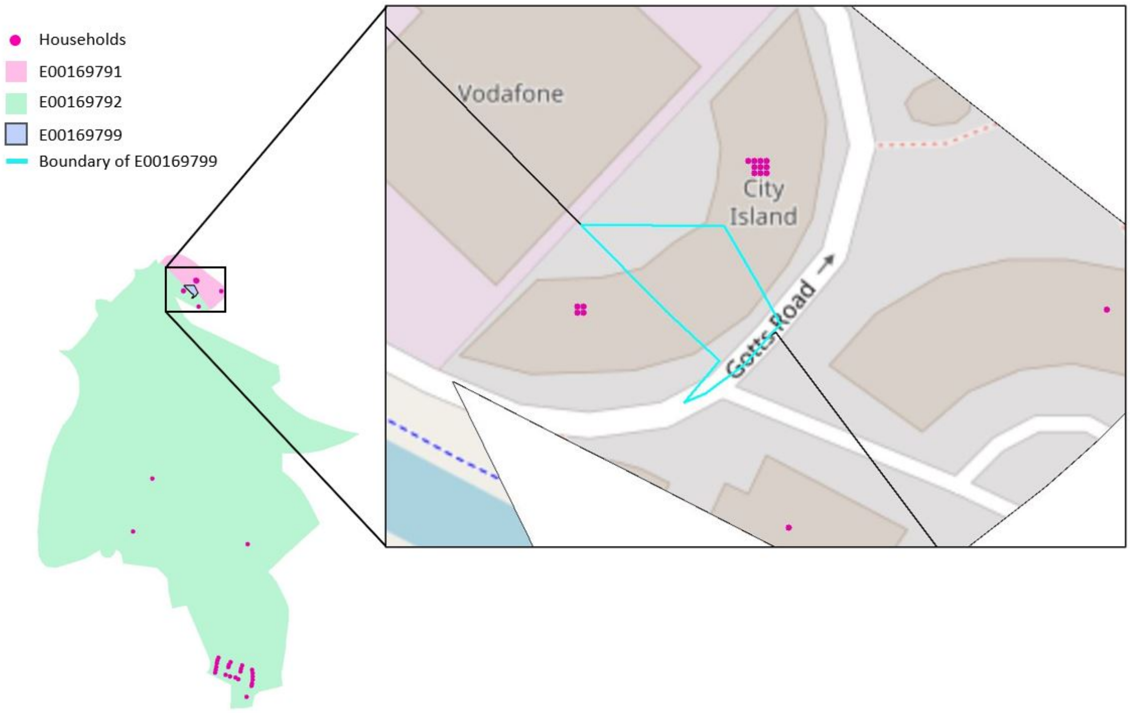


Figure 6.3: Boundary issue in OA “E00169799”

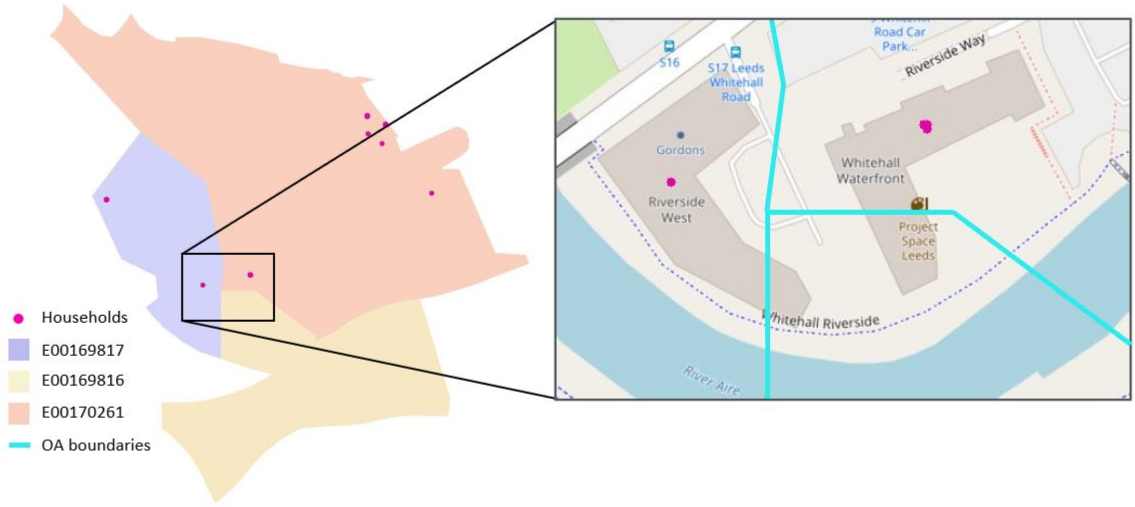


Figure 6.4: Boundary issue in OA “E00169816”

6.3.7 Closer evaluation of Method 1 outputs in Leeds OAs

Although Method 2 is the preferred method in this study, and all data is aggregated by this approach where possible, the property transaction data requires a Method 1 aggregation, since it does not contain the appropriate data to support any other approach (as explained in Section 6.3.5). When all households are aggregated to OAs based on the postcode to OA directory, a review of the results highlights an additional problem OA in this aggregation (*E00169604*). No properties are assigned to this OA, despite it containing 40 households as per the census household counts. Again, in order to meaningfully use this dataset, this needs to be explored and addressed.

By contrast, the aggregation of the households in the Council Tax data by Method 2 assigns 40 households to this OA, matching the total household count of the census[5]. However, the property transactions dataset must be aggregated based on Method 1, since there is no way of liking this data to the Council Tax data to use Method 2, as noted above. A review of the UPRN to OA lookup employed by Method 2 reveals that the 40 households identified in this OA using Method 2 are in postcodes *LS6 3EP* an *LS6 3ES*. However, in the postcode to OA lookup (used in Method 1), these postcodes are instead assigned to a neighbouring OA *E00169603*. Since it is not possible to identify these exact households in the transaction data to re-assign them back to *E00169604*, in order to meaningfully employ the property transaction data, the decision is made to merge all data relating to these two OAs and to treat the two as a single OA.

This is the only OA with a *noticeable* issue using Method 2, however, that is not to say that the other OAs are perfectly assigned, but this is the only one where the issue is *noticeable* (see the discussion in Section 6.3.9).

6.3.8 Merging the "problem" OAs

To execute the merging of the "problem" OAs with the neighbouring OAs, as identified in the previous section, any data relating to the "problem" OAs in all datasets are re-labelled as relating to the relevant neighbouring OA. Updating the census data, in this way involves summing the counts of each variable for the two OAs and re-calculating the percentages based on the combined total households, as per the census total household counts. The two ratio variables *"k007 - Number of persons per hectare"* and *"k035 - SIR"* must also be re-calculated. Since the averages in the LCC managed social housing rental were provided pre-calculated, these variables can not be re-calculated for the merged OAs. Instead, the averages relating to the neighbouring OA merging on to the problem OAs are used. Although not ideal, this is judged to be the best pragmatic solution. In each merge, both OAs already share the same OAC and LSOAC result, so this will not cause an issue in the comparison of the new classification result with the LSOAC in these merged OAs. This also further supports the decision to merge these OAs, since the residents in the two OAs have been

[5] It should be noted that these are not necessarily the *same* 40 households, but the count does match.

identified as as similar in the 2011 OAC, seemingly maintaining homogeneity in the newly derived OAs. The final count of OAs to be classified in the city is therefore 2,539.

6.3.9 Additional remarks: Caveats, recommendations and warnings

These examples have highlighted the real-world challenges of converting between geographies and of trying to replicate a black-boxed allocation of the households to OAs.

The decision to merge OAs is made to remove the existence of OAs which contain very few, or no households at all in any dataset, which is not meaningful and certainly would not be comparing like-for-like across the datasets. However, this intervention is only possible due to the small count of OAs affected. The investigation of these instances and decision of which OAs to merge has been a manual process which would not be feasible if there were more OAs to consider. Additionally, re-calculation of the census percentages, and specifically the census ratios, is a non-trivial task.

Moreover, this solution only considers the OAs experiencing *obvious* discrepancies in the allocation of households in the census vs. in the use of Method 1 or 2, i.e. OAs which were allocated few or no OAs by Method 2. This investigation does not address any OAs to which Methods 1 and 2 might have allocated a set of households which are actually quite distinct from the households assigned to the same OA in the census aggregation, but for which similar total counts mask the discrepancy.

To quantify the potential scale of such a problem, the secondary lookup table provided by the ONS (ONS, 2020d) which lists all postcodes in England and Wales which are split across OAs (referenced in Section 6.3.4) is used to identify any Leeds postcode where a split occurs. This highlights a total of 1,108 postcodes in Leeds which are affected by being split across one or more OAs, 3.7% of all of the postcodes in the city. These postcodes are split across a total of 1,269 OAs in the city (49.9% of all OAs) which share one or more postcodes with another OA, and contain 58,788 households (as per the Council Tax records). These results indicate that up to 16.6% of households in the city could be allocated differently in the use of the postcode to OA classification (Method 1) compared with the black-boxed allocation method used by the census.

Of course, not all of the households in the affected postcodes are allocated differently between the two methods, however, it is not possible to identify which exact households are affected, or even to calculate how many per OA. The secondary table employed here does offer a *percentage of the population* in each postcode which is assigned to each of the distinct OAs which the postcode is split across, though it does not offer a similar indicator of the *distribution of households*. Nevertheless, it is evident that the scale could be large enough to undermine the consistency in the allocation of the households to the OAs using the different methods, and as such, to weaken any classification generated on the basis of data employing two or more of the different aggregation methods.

Thus action taken to support consistency in the aggregation of data within public sector is recommended. Though there is now no way to address the issues caused by the lack of an exact household to OA lookup as per the specific aggregations made in the 2011 census, standards could be implemented to ensure a consistent aggregation employed across all other published and shared public sector administrative data. Moreover, the weaknesses and limitations of the postcode to OA and UPRN to OA lookups should be discussed more widely, and emphasised more clearly, in the accompanying descriptions of the lookups. It is understandable that many researchers might make the assumption that these lookups provide an exact match, without further checks or investigation. Had the discrepancies not generated some obvious errors in Leeds, with some OAs assigned no households, these assumptions might have continued untested, which is likely not an uncommon occurrence in the research generated by the users of these lookups.

Furthermore, in anticipation of the upcoming 2021 census, the ONS should prioritise the re-consideration of the approach taken in the past to obscure the precise allocation of households to OAs employed in the aggregation of the published census statistics, even if this information is published with restrictions, or shared with the support of sharing agreements to safeguard against misuse. As it is becoming more commonplace for researchers and public sector analysts to want to combine census data with alternative administrative datasets, it is difficult to understand the justification of orchestrated obstructions presenting barriers to such progress in the future.

6.4 Variable selection

Now that the novel variables have been derived, the next step of the Standard Framework is to execute a variable selection process (see Section 3.3.1).

In the development of the OAC in both 2001 and 2011, both correlation and sensitivity analysis were carried out to reduce all of the possible census variables to a subset of just 41 and 60 final variables, respectively (Vickers and Rees, 2007; Gale et al., 2016). Such methods supported the developers' two key objectives of capturing the variances in the population whilst achieving parsimony. Here, parsimony is not a main objective, since computer power is not restricted and the software used (statistical computing software "R", which is used throughout this thesis) could capably handle the entire size of the candidate data set. However, stripping out variables which are either introducing redundancy and/or any which are negatively impacting the potential of the cluster analysis to distinguish individual population groupings is still considered crucial. Therefore, correlation and sensitivity analysis for variable reduction remain appropriate methods for achieving this objective. These tests are thus applied to the variables, to decide which to retain. As per the 2011 OAC methodology (Gale et al., 2016), these tests are used to guide the variable selection process in conjunction with other contextual considerations, not as final decisions on whether or not to include each variable.

Correlation analysis:

As per the 2011 OAC methodology (Gale et al., 2016), a Pearson correlation analysis is created for the combined variable set, including both the 60 census variables used to develop the 2011 OAC and the 25 new candidate variables derived in the previous section. Although the census variables were already subjected to it in the 2011 OAC, this analysis is carried out for all 85 variables, in case of correlation between the census variables and the novel variables, or between any of the novel variables themselves. In line with Gale et al.'s (2016) parameter, the new candidate variables with greater than 0.6 or less than -0.6 correlation with any other variable are examined.

There are 7 new candidate variables correlating above this threshold with one or more other variable, most commonly, with the census variables representing home and car ownership. However, many of these relationships are deemed of interest for their predictive and descriptive power (again, as per the 2011 OAC variable selection justifications detailed in Gale (2014)). Therefore, only the variables duplicating the dimension represented by another variable are removed. These have correlation coefficients in the matrix of greater than 0.85:

- *"k027 - Households who live in a semi-detached house or bungalow"* removed in favour of Council Tax property type "Detached".
- *"k028 - Households who live in a semi-detached house or bungalow"* removed in favour of Council Tax property type "Semi-detached".
- *"k029 - Households who live in a terrace or end-terrace house"* removed in favour of Council Tax property type "Terraced".
- *"k030 - Households who live in a flat"* removed in favour of Council Tax property type "Self-contained flat".
- *"k032 - Households who are social renting"* removed in favour of Council property data total of social housing properties.
- *"k040 - Persons aged over 16 who are schoolchildren or full-time students"* removed in favour of Council Tax exemption "M/N" (Properties solely inhabited by students).

Sensitivity analysis:

Sensitivity tests are also run on the new candidate variables, excluding those which are now explicitly included in replacement of census variables (as per the correlation analysis above), since the 2011 OAC variable selection had already retained the original variables. A series of k-means classifications are run, each time holding one of the remaining candidate variables back. The resulting WCSS and BCSS (introduced in Section 3.3.2, STEP 4) in each test are then compared to the WCSS and BCSS of a k-means analysis run on the entire variable set, to evaluate the impact of withholding each variable (Gale, 2014). Results which maximise the ratio between the BCSS and WCSS are considered to be indications of a variable having a positive impact on the cluster result, whilst variables which decrease the WCSS and increase the BCSS suggest a negative impact. Thus the latter are highlighted for omission (Gale et al., 2016; Liu et al., 2019). A ratio BCSS/WCSS, as per Liu et al. (2019), is calculated to assess the overall quality of each result and to identify the best solution, i.e. the set of selected variables generating the best clustering result. Maximising the ratio between BCSS and WCSS is the target (Liu et al., 2019), favourably indicating relative between-cluster heterogeneity and within-cluster homogeneity and representing a better cluster result. The BCSS/WCSS ratio for each test is illustrated in Figure 6.5.

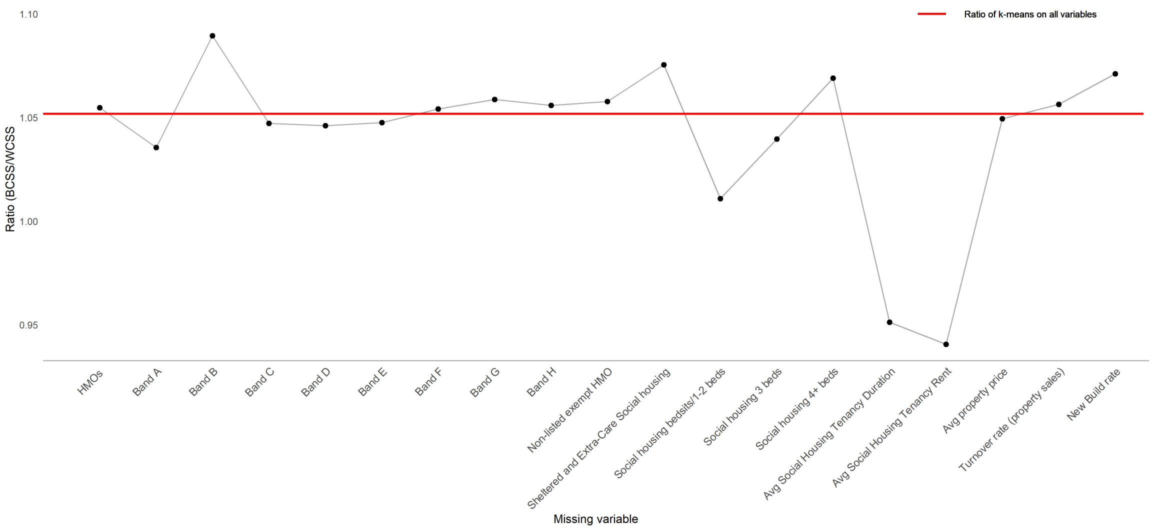


Figure 6.5: Sensitivity analysis of candidate new data variables.

In total, 9 variables are highlighted for omission, but just two are removed from the candidate set:

- “Band A”, “Band C”, “Band D” and “Band E” are retained for consistency, since the other bands are retained.
- Social housing Bedsits, 1-2 bedroom and 3 bedroom properties are retained for descriptive potential, since total count and 4+ bedrooms are retained.
- “Average property price” is retained due to small impact (albeit negative).
- “Average social housing tenancy duration” is removed.
- “Average social housing rent” is removed.

Following the correlation and sensitivity tests, a total of 77 variables, are retained for inclusion in the classification. This comprises the census variables listed in Appendix B.1, excluding the 6 removed following the correlation analysis above, and the 23 novel variables listed in Appendix C.1.

6.5 Re-classification results

The 2011 OAC development methodology is applied to the final set of 77 variables from across the novel and census data, adopting the data transformation and scaling procedures, the variable selection process, and the cluster analysis detailed in Singleton et al. (2016) and outlined in detail in Chapter 3. Similarly, the parameters of the clustering analysis are retained as per the decisions made in the 2011 OAC (Gale et al., 2016), including a cluster count of 8 to assist in comparisons with the LSOAC, and all parameters used in the optimisation process.

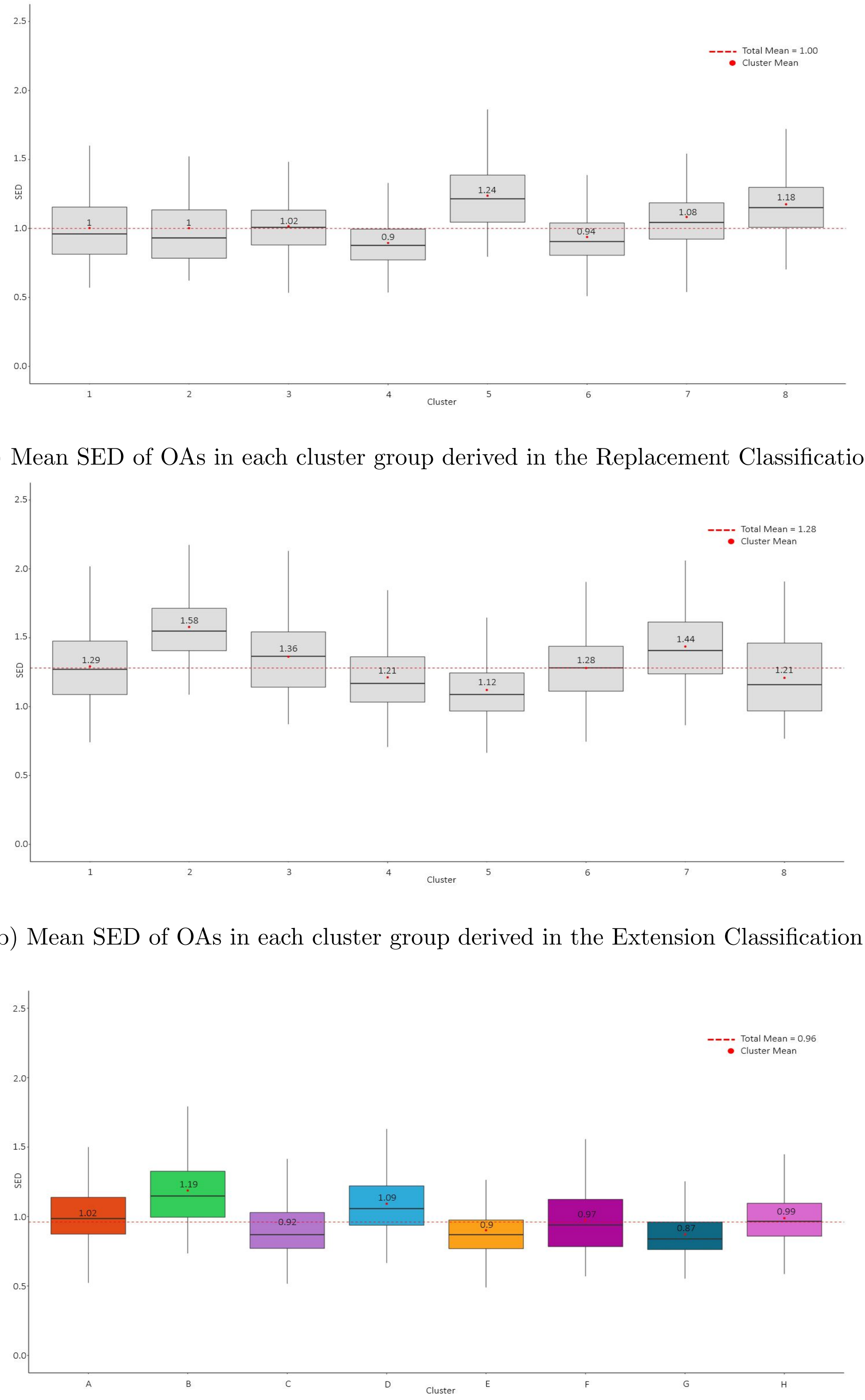
This analysis takes a two-fold approach. A first re-classification is generated based on just the remaining 54 census variables and the 6 novel variables identified as replacements for census variables in the correlation analysis in Section 6.4 (listed in Appendix C.1 as "Replacement"). This is henceforth referred to as the "Replacement Classification". A second re-classification is also produced containing all 77 of the variables selected in Section 6.4. This is henceforth referred to as the "Extension Classification".

Boxplots in Figure 6.6 represent the SED (introduced in Section 3.3.2, STEP 4) by the clusters in (a) the Replacement Classification, (b) the Extension Classification and (c) the LSOAC. The ordering of the clusters is arbitrary, as such, like-for-like comparisons should not be made between individual clusters.

The mean SED across the three classifications show 1.00 and 1.28 for the new classifications, respectively, compared to 0.96 for the LSOAC. Since a lower SED represents a better fit, these results indicate that there has not been an improved fit overall on the LSOAC, or across the individual clusters, by either classification, although the Replacement Classification has performed better than the Extension Classification, and only marginally poorer than the LSOAC.

Of the 2,539 OAs, the Replacement Classification performs better than the LSOAC in 686 (27.02%), compared to 36 (1.42%) OAs improved by the Extension Classification. These improvements are displayed, split by the LSOAC groups, in Table 6.4.

These results suggest that there are some benefits to be gained from improving the census data by replacing with administrative data, where possible, but in terms of extending beyond with new data, it seems that it might not always help. This might be a genuine consequence of the inability of novel data which has been added to differentiate structures in the population well, or it might be a natural consequence of more dimensions in the data leading to a greater SED (Debenham, 2002). Nevertheless, either circumstance might suggest that there needs to be a process of more intelligently determining which variables should be added, and being more discriminative in identifying which variables are going to



(a) Mean SED of OAs in each cluster group derived in the Replacement Classification.

(b) Mean SED of OAs in each cluster group derived in the Extension Classification.

(c) Mean SED of OAs in each cluster group derived in the LSOAC (derived in Chapter 4).

Figure 6.6: Comparison of the mean SED for OAs in each cluster group derived in the new classifications including novel housing data and the LSOAC derived in Chapter 4.

LSOAC cluster	Total OA count	Average SED	% of OAs Improved by Replacement	% of OAs Improved by Extention
A - Affordable living	347	1.02	20.46%	0.58%
B - Students	138	1.19	13.04%	0.72%
C - Urbanites	404	0.92	24.01%	1.73%
D - High density multicultural families	205	1.09	12.68%	0.49%
E - Ageing workers	489	0.90	21.47%	2.25%
F - Aspirational young workers	61	0.98	59.02%	6.56%
G - Settled, ageing families	613	0.87	41.6%	2.45%
H - Stable professionals	286	0.99	28.67%	1.75%

Table 6.4: Summary of LSOAC clusters (derived in Chapter 4) and OAs with SED improved by each of the classifications derived using novel housing data.

add value.

Whilst the inclusion of alternative, non-census data does not automatically generate a better fit, there is scope to improve the fit in some OAs, as demonstrated here. These results are encouraging in terms of illustrating that administrative data sources do provide a viable alternative source for many of these variables, with five of the eight variables in the Housing domain being substituted here, which could make these classifications more timely and/or future proof if census taking in revised.

6.6 Discussion

The case study presented across this chapter and Chapter 5, has raised a number of key issues and future considerations. These are important to reflect on when understanding and interpreting the classifications which have been derived, and also in seeking to extend and repeat similar developments in the future, either employing other novel administrative datasets or for other LAs. These issues are outlined here, supported by some recommendations for mitigating their impact.

6.6.1 Census aggregation issue

The first issue which must be re-iterated due to its potentially substantial impact on the accuracy of geodemographic classification development are the problems introduced when combining variables from both census and non-census data. This practice has been widely discussed and encouraged for many years in geodemographics (Gale et al., 2016; Debenham et al., 2003), and is a technique broadly adopted across the commercial Geodemographics Industry. However, the work conducted here has highlighted an issue about which there has been no discussion and thus no solution found in the research.

Since it is impossible to aggregate household level non-census data in the exact same way as the census statistics have been aggregated, which was intentionally orchestrated in the design of the 2011 census (Martin, 2020), there is never going to be guaranteed consistency in the populations captured in an OA across all data sources where both census and non-census data is used. The significance of this in geodemographics is such that, whilst the classification is understood to present a description of the population resident within each OA, this description is not derived based on a single, consistent population. This is a concern as it potentially undermines the premise of geodemographic classifications. It is thus important that this weakness is understood and, as a minimum, presented as a caveat to any classifications developed in this manner. Here, manual mitigations have been implemented, however, these have only been possible due to the low number of obvious "problem" OAs identified. This might not always be possible in other case study areas. Moreover, attention has only been paid to OAs to which the aggregation method used has allocated noticeably low, or no, households. As mentioned in Section 6.2.5, a larger, unquantifiable number of errors whereby households are being moved between OAs in different aggregation methods could be being masked by the enforced focus on total household counts.

However, there is also no mention of this potential inconsistency in the documentation published alongside the widely used ONS postcode to OA and UPRN to OA directories. As such, it is likely that many researchers employ these directories with limited, if any, awareness or understanding of the likelihood of generating discrepancies in their geographic aggregations when using these directories to support the linking of census and non-census data. Thus, this thesis recommends that every effort be made to attempt to rectify this issue in the execution and publication of future censuses. This is going to become particularly crucial as researchers and policy makers continue to increasingly use both census and non-census data at the heart of their decision-making processes, and in doing so, risk undermining the accuracy of the results, insights and subsequent decisions generated if this issue continues to go unaddressed. In the meantime, all efforts must be made by the developers of geodemographics not to further exacerbate the issue. This could be achieved by employing a consistent conversion method, where possible, across all of the non-census data.

6.6.2 Barriers restricting the use of consistent aggregation methods

The recommendation made in Section 6.6.1 to mitigate against aggregation consistency issues by adopting consistent approaches across all non-census data was not achievable in this study for two reasons. The first was the absence of consistent location features across the datasets. Notably, the absence of a UPRN in the Land Registry data, with which to support the use of Method 2 in this dataset, to match the others. The second was that some data was provided by LCC pre-aggregated due to confidentiality concerns, based on a method which was judged to be inferior to the method identified as favourable. As such, the use of the favourable method on the other datasets was prioritised over the use of consistent methods here. However, the ability to employ a consistent, reliable method across all datasets would be the ideal goal. Since some LCC data must be pre-aggregated, to maintain confidentiality, this target must start within local government itself. Any aggregated data provided by the public sector should employ appropriate and transparent aggregation methods as standard. This is already true for LCC, who consistently adopt the same postcode to OA directory across their aggregations, supporting confidence in the use of the data. However, in light of the census aggregation issues identified, this standard practice might warrant some revision.

6.6.3 Twenty-first century improvements to public sector data infrastructure

One novel idea which might help ensure consistent aggregation practices in the public sector (even beyond LCC), but which might additionally have broader benefits, such as in speeding up and improving the quality of public sector data use, is the development of "Feature Stores". This is a concept which is becomingly increasingly recommended and routinely demonstrated in the commercial sector, much more broadly than commercial geodemographics (Li et al., 2017; Lécuyer et al., 2019). Feature Stores act as a repository for a set of standard variables rather than the raw data (where feature is simply another term for variable), enabling data-users to select and use pre-defined variables confidently in their individual analysis (Patel, 2020).

In the development of Feature Stores the data preparation and documentation is conducted application-agnostic. Such practice would replicate much of the work demonstrated in

Chapter 5, however, this would be executed in-house (within local governments primarily), reducing the burden of a geodemographic developer to go through the lengthy and cumbersome activity documented in the previous chapter. Instead, each prepared variable would have clear and complete documentation, including a description of the raw data and the aggregation processes which underpin its creation, and any critical weaknesses that it might have that should be understood to support its use. Patel (2020) describes this idea as a "democratising" of data within an organisation, supporting its use beyond a handful of employees, and suggests that the practice could make organisations more "productive", "agile", and "competitive".

Whilst this is not common practice in local governments, and may seem like a big ask given the current status of the data infrastructure which has already been discussed, and the existing time pressures of local governments, the potential benefits are undeniable. For example, with respect to time pressures, standard preparation of raw data could cut the potential for duplicate work, eliminating the burden of data preparation which is required each time the raw data is used. This could be crucial in geodemographic classification development, where time can be wasted preparing data which might be discarded in the variable selection procedure. Moreover, pre-processing could be carried out by individuals who are very close to and familiar with the data. Additionally, all data would be guaranteed a consistent geographic aggregation, and confidentiality concerns, which lead to lengthy data sharing discussions, could be alleviated since all data would be aggregated. Additionally, issues with different definitions of the same terminology, which risks end-user misinterpretation of data and which could critically undermine its use or lead to inconsistent expectations, could be addressed at the source. Standard practices, such as the inclusion of a consistent UPRN in all of the LCC data gathered in this chapter have already demonstrably supported the use of this data, yet more standard practices could offer further assistance. Moreover, the idea is not entirely new in the public sector. The aggregate census statistics are themselves a Feature Store containing variables which are aggregated from census responses and are ready for use. The aggregation and release of census data has encouraged and supported substantial research.

There are repeated calls for public sector data practices to be re-addressed to alleviate the concerns outlined here, with sharing agreements and the standardisation of practices at central and local government level (Carroll and Crawford, 2020; ONS, 2014). If the government are as serious about the potential of their data as the discussions in Section 2.4.1 suggests, the development of Feature Stores could be one initiative which might warrant the time and financial investments. Not only could such a practice standardise the use of public sector data, but it would necessitate, and thus ensure, good maintenance, regular updates, strong documentation, remove duplicates and guarantee that the same statistic is not calculated differently (i.e. in the generation of averages and percentages from consistent denominators), be easy to use for non-technical end-users and offer transparency in highlighting the data that is available to support research. Each of these benefits have the added potential for relieving analysts and researchers of the time currently attributed to data preparation, freeing up their time which could be reinvested in developing further innovation. However, as in the release of the census data, the success of this approach is still dependent on good documentation and a reliance on end-users ensuring that they understand the variables they are working with.

6.6.4 Subjective variable selection

All elements of the 2011 OAC methodology except the data and geographic extent have been adopted unchanged in this chapter. However, to further support the inclusion of a broader set of input variables, the variable selection procedures might warrant further attention. Although statistical methods have been employed for variable selection here and in the 2011 OAC, the results have been vetoed where conflicts occur with other priorities, such as retaining a broad range of data, and consistency with the 2001 OAC development (Gale et al., 2016), introducing subjective decision making into the procedures. This is a feature common to many variable selection processes adopted in geodemographics, which has provoked much criticism (see Section 3.6.3). Chapter 7 will expand upon these criticisms and the related discussion, and begin to explore alternative variable selection procedures.

6.6.5 Extending to other domains

This chapter has focused on extending the Housing domain of the 2011 OAC. Future work might consider extending some of the other domains, or introducing new types of data altogether as per the commercial classifications. For example, previous discussions have highlighted the potential for incorporating behavioural, lifestyle or attitudinal data (see Section

5.2.4). Data such as these are often appended to the classification post-development, to generate richer, more insightful descriptions of each classification group (discussed further in Chapter 8). However, there could be scope to incorporate data of this type into the classification development, if the data is necessarily comprehensive. However, caution is advised in the subsequent application of a classification derived from such data, since the classification may no longer constitute a general-purpose output Harris (1998). If the development of a domain specific classification is intended, however, this type of data could become even more valuable in its inclusion.

6.7 Conclusions and next steps

Although the classifications generated in this chapter, which included the use of census data alongside novel administrative data, has not been as effective as anticipated, it has validated the claims that such a development is now possible (discussed in Chapter 2), particularly with open and public sector data. It has also added value in demonstrating the process in practice, and in highlighting many issues which such a progression raises, as per the discussions in Section 6.6. In both cases, this chapter has extended the current geodemographics literature. It also invites other studies to extend this work further, either by exploring alternative datasets, extending the variables in the other non-housing domains, considering other types of data such as more behavioural or attitudinal data, or by supporting the inclusion of non-census data with improved variable selection processes, in theory, more robustly testing the suitability of each candidate variable before including it as an input variable.

It is this last suggestion which underpins the remaining analysis chapters in this thesis (Chapter 7 and Chapter 8). These chapters will present explorations of more sophisticated methods of variable selection, seeking to reduce the subjectivity which currently underpins popular variable selection methods (see Section 3.6.3), and suggest new ways of deciding which are the most suitable, appropriate and relevant variables for inclusion. Since the potential candidate variables which might be included in geodemographic classifications are only set to expand if the inclusion of non-census variables becomes routine, filtering and selection techniques will similarly increase in importance. This underpins the prioritisation of their exploration in the coming chapters.

Summary

This chapter has presented a practical application of the novel, administrative housing-related data, gathered in Chapter 5, into the development of a general-purpose local level classification for Leeds, extending such activity from hypothesis into practice. This activity has uncovered several issues which, if considered and addressed, could help propagate such a shift into more routine practice. In extension, the next chapters (Chapter 7 and Chapter 8) consider the necessity and scope for improving beyond common variable selection procedures to further support such routine practice.

Chapter 7 - Exploring variable selection methods: Feature Extraction through Factor Analysis

7.1 Introduction

As data availability improves, and thus the potential candidate variables expand, an objective methodology for selecting the input variables underpinning geodemographic classifications could become increasingly desirable. Moreover, the development of more appropriate and relevant locally specific classifications might benefit in particular from a bespoke, place-specific selection of input variables identified based on the specific drivers of diversity in the local area of interest. However, as outlined in Chapter 3, the variable selection methodologies currently employed in geodemographic classification development are typically pragmatic and subjective, based on a combination of convention, data availability and developer preference, rather than any single objective methodology (Brunsdon and Singleton, 2015).

Reviewing the criticisms previously presented and technical difficulties in its application to spatial data, this chapter considers whether Feature Extraction methods, in particular Latent Variable Models (LVM), such as Principal Component Analysis (PCA) or Factor Analysis (FA), which were once commonplace in geodemographics, have a new place in this space to improve variable selection, particularly in the development of place-specific classifications.

Section 7.2 includes a discussion of the history of LVMs in geodemographics alongside a specific technical summary of FA which presents an overview of the mathematical principles upon which the method is based. This is underpinned by a review of the practicalities of applying FA in practice, including the conflicting guidance and best practice advice which have historically weakened trust in the method and limited its employment in this context. Section 7.3 details a FA framework which could act as an alternative methodology for variable selection in geodemographic classification development. Finally, Section 7.4 demonstrates a practical case study employing the framework generated in the prior section to generate a new place-specific geodemographic classification for Leeds, illustrating the potential of using FA for variable selection in this context.

7.2 Background

The nature of cluster methods is such that they will always output cluster groups, irrespective of the input data or parameters used. It is thus important to ensure through the build of the classification that the drivers of the clustering, and the results consequently produced, are as appropriate as possible (Vickers and Rees, 2011). Since the result is often strongly dependent on the relevance and quality of the input variables (Blake and Openshaw, 2005) (discussed at length throughout this thesis), refining the methodology used to select input variables is seemingly crucial in refining the result. The target therefore is to identify and employ a methodology which seeks to select variables capable of genuinely representing and distinguishing the social identities which are present within the population.

There have been efforts made to identify such an approach in the past, frequently relying on the use of LVMs, particularly FA and PCA targeted for their perceived potential to bring objectivity to a previously empirical process. For example, such an endeavour in the mid-twentieth century brought about a new phase in the evolution of geodemographics in Factorial Ecology. However, mounting criticism of the approaches taken by Factorial Ecologists, coinciding with a shift in geodemographic developments from academia to the commercial sector, saw such practices fall out of favour around four decades ago (see Section 2.3.1). As academic interest in exploring variable selection procedures once more experiences a resurgence, both here and in other recent research such as in Liu et al. (2019), it could be time for FA to be reconsidered, particularly in a local-level context.

7.2.1 Introduction to Latent Variable Models

Prior to its application in practice, this chapter will introduce the theory of LVMs, in a broad sense, and the practice of adopting the methods in practice for variable selection. This will particularly focus on its uses in geodemographics, in an extension of the existing literature.

Theoretical principles

LVMs are multivariate statistical models used for feature extraction, offering a methodology for reducing an initial candidate set of variables to a smaller set of selected variables. The procedure aims at removing redundancy whilst preserving the information captured in the original variable set (Dean, 2018). The process seeks to identify unobservable constructs (or *latent variables*), for which there is otherwise no means of direct measurement, but which can be identified through the relationships of the measurable candidate variables (or *observable variables*) (Everitt and Hothorn, 2011).

Whilst LVMs efficiently simplify the intricacy of the relationships which exist between the input variables, even for a substantial number of variables, the theories which underpin the models are mathematically complicated, and their practical application is technical and complex (Rummel, 1967). Though the latter has been relieved somewhat by the development of advanced statistical software, rendering the methods increasingly accessible (Tabachnick and Fidell, 2013).

Common LVMs include PCA and FA, in which the latent variables are referred to as *components* and *common factors* (or simply *factors*), respectively. Though closely aligned and regularly discussed alongside one another, there are some key differences between the two which can be used to promote the use of one or the other in specific circumstances (Dean, 2018). Though PCA has been regularly adopted in geodemographics (see Section 7.2.2), this study focuses on FA in particular, for reasons which will be outlined throughout.

Moreover, all mentions of FA throughout this chapter will refer specifically to Exploratory Factor Analysis (EFA). This method is often used early in research processes to generate theoretical hypotheses of underlying phenomena within a given dataset, based on the structures identified in the analysis (Henson and Roberts, 2006). As an unsupervised method, EFA is applied without assumptions and a-priori expectations (Tabachnick and Fidell, 2013), and can be useful in the absence of theory (Rees, 1971). A second FA methodology is available, Confirmatory Factor Analysis, which, conversely, provides a more explicit framework for confirming prior notions about the structures, testing models developed based on prior analysis (Dean, 2018; Everitt and Hothorn, 2011). Confirmatory Factor Analysis is not considered here.

Defining the mathematical concepts and terminology of Factor Analysis

Though it is beyond the scope of this thesis to outline the underlying mathematical equations at play in FA in fine detail, it is of some benefit to present an overview of some of the significant concepts to aid in the application and interpretation of the analysis, including the introduction of key definitions for the specific terminology associated with the approach. Several are summarised upfront here, with more introduced where appropriate throughout the chapter.

Each variable in the candidate dataset can be considered as a dimension in vector space. This could be conceptualised algebraically or geometrically (Rummel, 1967; Yong and Pearce, 2013) where the angles between the dimensions indicate the associated correlations, i.e. the size of the relationship between the variables. This is indicated in Figure 7.1 for an example candidate set of 4 observed variables across 3 observations, for demonstrative purposes only. When the variable set increases beyond 3 variables, exceeding the graphical limit, it becomes difficult to physically plot the vectors in this way (Rummel, 1967). However, these relationships can also be represented in a matrix of associations containing a quantitative measure of the relationships between the variables. This matrix is used to compute the calculations of the FA (Yong and Pearce, 2013).

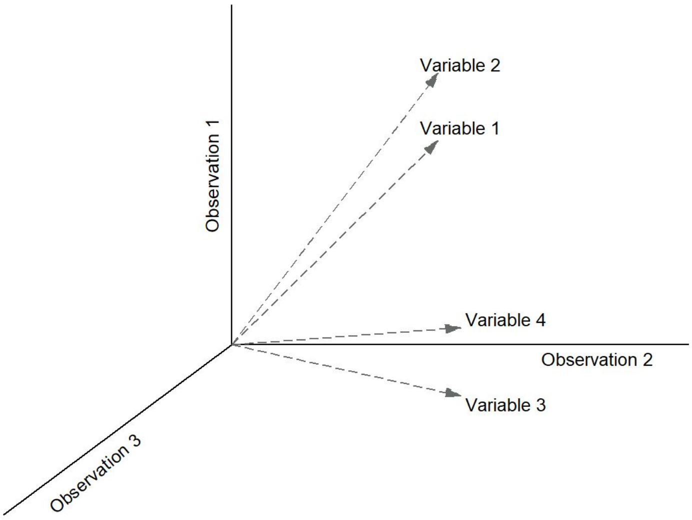


Figure 7.1: Demonstration of graphing four variables over three observations as vectors.

The ***factors*** (the latent concepts sought) are essentially drawn from distinct groupings of correlated vectors, with the vector points being projected to distinct factor axes defined by the groupings (demonstrated in Figure 7.2). In practice, in the context of population variables, this action identifies social constructs (labelled "factors") which are present within the variable set and which are represented by different groupings of related variables.

The correlation between each the observed variables and each factor axis represents the

(factor) loading of the variable for the given factor. The greater the value, the more the variable contributes to the factor (Yong and Pearce, 2013), i.e. the more the variable drives the social construct. It is therefore desirable to identify variables exhibiting greater factor loadings (Kline, 2014). Factor loadings are on a scale of ±1 and can be interpreted like correlation coefficients (Rees, 1971; Rummel, 1967). Geometrically, the variables which "load" more highly on a factor will be in closer proximity to the factor axis (Yong and Pearce, 2013). This enables the practice of taking the identified factors (each representing a social construct) as an input variable itself into the development of a subsequent geodemographic classification, instead of each of the attribute variables which underpin these constructs. Each factor therefore represents an unobservable construct characterising the relationships of the observable candidate variables. Unlike in PCA, in FA there may be some correlation between the factors identified (Dean, 2018), i.e. there may be some relationship between the social constructs identified.

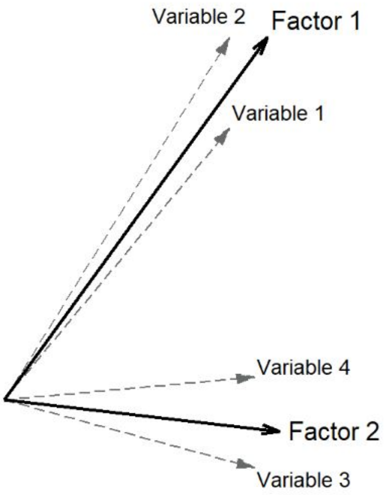


Figure 7.2: Demonstration of projecting the four variables from Figure 7.1 on to two factors.

In some applications of FA the final objective is to reduce the dimensionality of the data by generating factors and using these as variables in replacement of the original dataset in future analysis (Yong and Pearce, 2013). This was an important application of these techniques when computational power was more limited (Vickers, 2006).

There are several methods of feature extraction which can be employed in FA (see Section 7.3.2), however, most seek to assist the analyst[1] in identifying the amount of ***variance***

[1] The term 'analyst' is used here in relation to the application of an FA in general, 'developer' is also synonymously used in this chapter, specifically when referring to the the application of FA within the

which can be explained by the inclusion of each of the candidate variables to support the removal of variables which add limited unique information into the model (Child, 2006), thus retaining the maximum information from the remaining variables (Dean, 2018). To achieve this, each variable is assessed for its potential to *predict* the other variables based on the shared (or common) variance of variable pairs, identifying how well one variable could be predicted given knowledge of the other variable (Rummel, 1967).

A ***communality score*** is also generated for each variable based on the variance observed (Dean, 2018). The communality represents the amount of the variable's total variation which is involved in generating each of the factors (Rummel, 1967), or considered another way, the amount of the variable that can be predicted given the factors (Yong and Pearce, 2013). High communality reflects a more informative solution. This is the opposite of ***uniqueness***, a measure of the variable's unrelatedness to the other variables, which is also often calculated (Dean, 2018).

The ***(percentage of) variance explained*** for a given derived factor indicates the amount of the data in the original association matrix which could be predicted by the factor, i.e. the accuracy with which the social constructs identified could predict the underlying attribute variables upon which they have been derived. Thus, the greater the variance explained by the factor, the more useful, or important, the factor (Dean, 2018). The amount of variance explained by the individual factor is represented by its ***eigenvalue***.

The result of FA is typically returned as a table comprising the factor loadings alongside additional summary statistics including, but not limited to, the variance explained by each factor and the associated eigenvalues.

General concerns and criticisms of Factor Analysis in research

Though there is a consensus across the literature that LVMs can be useful, with the potential to derive critical hidden insights if properly understood and appropriately applied (Clark et al., 1974), resources presenting practical "best-practice" advice to aid in executing a reliable and meaningful application of FA is conflicting and often contradictory, as will be highlighted throughout this chapter.

Moreover, in addition to difficulties surrounding the sound execution of FA, many raise con-

development of a geodemographic classification.

cerns regarding an inability to quantitatively validate the accuracy of the result (Tabachnick and Fidell, 2013; Rees, 1971). There is also no guarantee that the mathematical result will be meaningful in the real world (Dean, 2018). A good solution is recognised broadly as one which is interpretable and "makes sense" (Tabachnick and Fidell, 2013; Dean, 2018). However, it is often easy to conjure some explanation of the result that makes sense, particularly when the accuracy of a contrived explanation cannot be quantitatively measured (Rees, 1971).

As a consequence, Tabachnick and Fidell (2013) claim PCA and FA have a "somewhat tarnished reputation" as being used in "sloppy research", with Everitt and Hothorn (2011) suggesting that FA has probably attracted more criticism than any other statistical technique. Nevertheless, FA continues to enjoy a longstanding history of use in many fields, most notably in sociology and health related subjects, predominantly psychology (Williams et al., 2010).

7.2.2 Latent Variable Models in Geodemographics

LVMs have appeared in a number of guises throughout the history of geodemographic classifications. Whilst most dominant in the development of its precursors in Factorial Ecology, examples of their use still persist in the construction of recent classifications, albeit more rarely and in a much reduced capacity. Nevertheless, somewhat of a revival has been made to assist in the harnessing of the increasingly vast quantities of available data, and in parsimoniously selecting relevant variables, though their employment remains beyond the norm (Longley, 2005). The impetus behind this revival of interest echoes many of the justifications for the initial introduction of these methods and their early applications in the context of Social Area Analysis (SAA) in the 1950s, where they were adopted to evolve the field against the backdrop of increasing data availability and rapidly improving multivariate statistical methods (Singleton, 2014) (see Chapter 2).

Historical significance

In response to criticisms of early, empirically grounded research in SAA by Shevky, Bell and Williams (Shevky and Bell, 1955; Shevky and Williams, 1949), which was perceived to support subjective variable selection justified by post-facto rationalisations (Robson and Robson, 1969), Factorial Ecology emerged as the next generation of Urban Analysis, underpinned by the novel use of FA to provide a seemingly more objective foundation, particularly at the point of variable selection (Batey and Brown, 1995). In their seminal research which maintained the focus on Los Angeles, which had provided the case-study area of much prior investigation, Anderson and Bean (1961) made use of FA to make original observations of the city, drawing out nuance in the spatial composition of family characteristics which had been previously overlooked in the more empirical studies.

The success of work such as this further emphasised both the potential and the necessity of using methods such as FA in this context. The use of PCA and FA seemingly offered some insight into the relationships at play between the variables employed, highlighting latent social constructs which were driving the formation of population structures (Voas and Williamson, 2001). Whilst some contemporary reviewers welcomed this potential, endorsing the use of FA in Urban Analysis, albeit under the non-trivial caveat of the methods being thoroughly understood and appropriately applied (Clark et al., 1974), such a caveat has proved largely unachievable.

The Factorial Ecology approach was widely challenged with criticism relating to its technical application of FA and PCA, and caution was emphasised in the use of such approaches. The main perceived potential weaknesses of the methods included the influential role of the developer in conducting the analysis, and the potential limits to the real-world, contextual meaning of the mathematical constructs identified in the statistical analysis (Hunter, 1972; Rees, 1971; Lebowitz, 1977). The resulting widespread distrust of the approaches contributed not insubstantially to the subsequent demise of Factorial Ecology and the halting of exploration into the methods in the late 1970s. This saw the approach fall out of fashion as modern-look geodemographics began to increasingly attract interest (Harris et al., 2005), leading to new traditions being developed in the creation of geodemographics which no longer relied on approaches from Factorial Ecology and the routine use of LVMs (see Section 2.3.2).

Recent revival of interest

The black-boxed nature of commercial classifications impede the ability to assert whether LVMs have been, or are currently, employed in any capacity in the development of proprietary offerings. However, discussions have been possible with TransUnion, thanks to their partnership with this thesis. These discussions confirm that the standard methods for variable selection, in particular, pertains more to subjectively-led selection informed by expert knowledge, experience and empirical sensitivity testing of candidate variables.

Despite this, some employment of LVMs have continued to occur in the recent academic literature. However, again, these instances are atypical in the development of standard classifications, especially for informing variable selection. These are largely restricted to the use of LVMs as a means to a theoretical or analytical end (Reibel, 2011), most popularly as a technique for dimensionality reduction (Vickers, 2006; Voas and Williamson, 2001), generating a reduced number of factors with which to replace the many input variables (Yong and Pearce, 2013). However, as computational power has increased, even the necessity for reducing the dimensionality of the data has also become seemingly redundant (Vickers, 2006). Though Brunsdon et al. (2018) highlight alternative benefits of LVMs, which go beyond dimensionality reduction to reduce computational overheads, in particular, highlighting the ability of PCA to aid in mitigating against the potential weighting impacts of highly correlating variables.

Moreover, Liu et al. (2019) have re-introduced the notion of employing LVMs for variable selection into the contemporary discussion, in a study more reminiscent of early Factorial Ecology. Testing the scope for improving variable selection processes through the development of an automated application of PCA, the work re-discovers the previously hailed benefits of the process in the context of better informing the development of the 2011 OAC. However, there is little focus throughout the studies of Brunsdon et al. (2018) and Liu et al. (2019) on the technical criticisms of LVMs which were rife in the Factorial Ecology era, which ultimately degraded confidence in their use. Nor is there much of a focus on efforts to mitigate against similar criticisms. Despite this, both studies report positive results from the applications of the approaches.

Specifically, in concluding their tests a success, Liu et al. (2019) promote and encourage the uptake of these methods for improved variable selection in future geodemographic classification development beyond their single example, explicitly commenting on the potential for integrating local and national level considerations in the development of more nuanced classifications.

Locally specific potential

Whilst Liu et al.'s (2019) study made a passing reference to the potential benefit afforded to place-specific classifications through the use of LVMs, much stronger assertions were made in many of the research outputs emerging from Factorial Ecology in its heyday. As highlighted in Chapter 2, in early precursors to geodemographics the geographic focus was typically at an individual city level. As the early half of the twentieth century saw some US cities uniquely develop, driven by circumstances linked in no small part to migration, locally-specific socio-spatial patterns began to emerge. Many of the early precursors developed in the US were thus particularly focused on understanding these emerging city structures at the local level (see Chapter 2.3.1).

PCA and FA offered the means to begin to computationally understand how and why residential areas in different cities differed from one another by unearthing which structures in the city defined the unique distribution of the population (Rees, 1971). Further, Sweetser (1965) professed that such methods might even offer an explanation of the relationships between local population characteristics and behaviour. Importantly, Van Arsdol et al.'s (1958a; 1958b) studies using LVMs to compare the structures of several US cities identified distinct ethnic make-up and associated social constructs which were uniquely present in some cities and not in others, indicating the importance of local-level considerations. Consequently, Palm and Caruso (1972) even suggested from these findings that one might be cautious when inferring that the presence of any social construct can be assumed to represent a general model of society elsewhere, as is the basis of the national-level classifications today.

Moreover, further findings in Van Arsdol et al.'s (1958a,b) research highlighted instances of particular social traits, which were identified in several cities and which had previously been considered universally representative of a particular phenomena, to actually represent different phenomena between geographies depending upon local circumstance. Thus, one might consider this an endorsement for employing caution when adopting characteristics to act as a proxy for social attributes at a national level, as is typical practice in most geodemographic classifications, and moreover, thus supporting the adoption of techniques such as the use of LVMs in the selection of variables for both local and national level classifications to capture such instances.

Empirically, UK cities remain vulnerable to the potential for geographic discrepancies in the representation of a single variable, for example, in the use of the census variables "Access to a vehicle" and "Owns own home", which in some cities might represent wealth, yet are less common in areas of London with unique property markets and where cycling is more prevalent in the city's culture, irrespective of the wealth of the area. Such considerations are a fundamental driver of the place-specific focus of this thesis in general, and in the exploration of LVM methods specifically. Since LVMs have demonstrably discovered locally specific phenomena such as the geographic variance in meaning of the same characteristics in the past, these methodologies might offer a good starting place in the search for similar levels of locally specific differentiation in the future. Since objective variable selection procedures could be even more important in local specific classifications, the use of LVMs for variable selection at a local level might offer a more informed input variable selection for generating place-specific classifications which does not rely on assumptions in the use of typical proxy variables, thus supporting their exploration in this chapter.

7.2.3 Proposed revival of Factor Analysis models in Geodemographics

The purpose of this study is therefore to build on this earlier body of research to explore the potential for using FA to improve the selection of input variables in the development of place-specific classifications. The objective is to assess whether the employment of FA can derive a classification with a good 'fit', not to develop the best possible FA methodology or solution, though efforts are made to ensure the resulting methodology is statistically reliable and reproducible.

There is currently no single source, or template, to guide the use of FA for variable selection in geodemographic classification development. There is also very limited technical discussion of the methodologies used or justifications of the decisions made in the application of FA in the Factorial Ecology studies published, although some of the procedures and parameters used may well be outdated today anyway. The methodology developed throughout Section 7.3 is thus described and documented in thorough detail to enable the necessary scrutiny of the method, and to offer a review of the decisions to be made throughout the process and a template which can be adapted and utilised in future geodemographic classification development.

As is typical in many statistical processes, there is no single correct answer to be derived through FA, and the decisions of the analyst can impact the resulting output (Davies, 1978a; Henson and Roberts, 2006). Practical completion of FA is complex with few concrete rules and many potential options (Costello and Osborne, 2005). The perceived subjectivity which this introduces has previously contributed to the weakened trust in the method with regards to Urban Analysis, and consequently reduced its appeal. Similarly, criticisms have been raised in other fields where FA is more widely accepted, such as psychology. Such criticisms in these fields typically concern the decisions made by the analysts employing the FA (Hogarty et al., 2004).

Simultaneously, as the statistical packages have become increasingly easy to use, it has become all the more critical to ensure that they are being applied with a good understanding of the underlying processes and their capabilities (Tabachnick and Fidell, 2013). Poor decisions have the potential to reduce the accuracy of the results and the usefulness of the FA (Hogarty et al., 2004). Naturally, a reliable result is desired, particularly where the output could be used to inform social policy decisions.

Since the reputation and statistical robustness of the methods have previously been brought into question (Tabachnick and Fidell, 2013; Everitt and Hothorn, 2011), a particular emphasis is placed on curating the methodology in Section 7.3 through a process of careful decision making and clear justifications, based on the available literature and technical and contextual considerations. In doing so, the potential weaknesses of the approach, as have been thoroughly documented in previous uses of the methods in Urban Analysis, are highlighted, and pragmatic, workable solutions are presented. To support this study, the underlying mathematical principles and concepts of FA are also outlined.

7.3 Methodology: Researching and building a Factor Analysis model

Various solutions to alleviate accuracy concerns are proposed across the literature. “Multi-method testing” is frequently recommended, running the FA a series of times based on different decisions and comparing the similarity of the results to test for stability (Hunter, 1972; Davies, 1978a; Costello and Osborne, 2005). Alternatively, ensuring strict justifications underpin all decisions made, to enable an end-user to trace back through the process if necessary, is also proposed (Davies, 1978a). This study aims to better enable the latter approach, seeking to ensure that objectivity is prioritised throughout the decision making process, particularly seeking not to play too much with the model to fit a pre-defined narrative or mine for patterns (Dean, 2018), but to select a framework upfront and accept the results. The potential benefits of multi-method testing will be considered further in Section 7.5.

In order to justify each decision made, a thorough review of related literature is carried out. This review revealed a substantial discussion relating to the process of completing FA, including a raft of recommendations, guidance and "rules" considered to enable the development of a meaningful FA framework. As this section will highlight, the recommendations presented throughout this literature are complex, often contradictory and in some cases, context specific, revealing the difficulties of not only running FA but in being confident in the accuracy and meaningfulness of the results.

Though there exists debate regarding the details, there is a consistent core technical framework which emerges structured around three commonly identified phases, (1) data preparation, (2) set-up of the model, and (3) iterative running of the FA itself. These discussions from the literature are consolidated and a technical framework is derived here. This is summarised in Table 7.1. The order of the steps, particularly in the data preparation phase, is significant, as each can affect the next (Tabachnick and Fidell, 2013). The remainder of this section will present the specific technical advice relating to each of the steps identified in Table 7.1 in more detail, outlining the key considerations to be made, and indicating in each case where experts advice differs. This presents a basis upon which the decisions concerning specific parameters selected in the case study to follow will be made.

7.3.1 Data and data preparation

Checking appropriateness of data

FA works better with all data of the same type, e.g. all continuous data. It is less easy to combine continuous and categorical data in the same FA (Dean, 2018). In the geodemographic classifications developed so far, all data has been continuous. As such, this would not present a challenge if applied in these circumstances.

Framework phase	Action
Data preparation	Check the appropriateness of the input data
	Standardise the input data
	Check the input data for non-normality/Linearity/Variability
	Check the input data for outliers
	Check the input data for 'multicollinearity' and 'singularity'
	Confirm 'factorability' of data
Model set-up	Select extraction method
	Select rotation method
	Select appropriate number of factors
Analysis	Check for variables with 'low communality'
	Check for variables which do not 'load highly' on any factor
	Check for factors with limited variables loading 'highly'
	Check for variables loading across factors
	Evaluate the Eigenvalues of the factors identified
	Evaluate the total percentage of variance explained

Table 7.1: Summary of framework steps.

Data Transformation/Standardising the data

Though standardising the data is not essential in FA, it is common to set the data to a consistent scale for comparability (Yong and Pearce, 2013). However, similar to the discussions presented regarding the transformation of data in Section 3.3.2 (STEP 3), FA experts for a long time have cautioned that transformation could make interpretation harder (Clark et al., 1974). Nevertheless, just as illustrated by Gale et al. (2012) in the development of the 2011 OAC, transformations might also address some non-normality in the data. It is thus still commonly executed, and is recommended that this is completed step *before* checking the normality of the variables (Tabachnick and Fidell, 2013). Since it is common practice to standardise input data in geodemographics (as discussed in 3.3.2 (STEP 3)), this does not constitute additional work, or cause any conflict, in the process of using FA in geodemographic classification development.

Since the aim of FA is fundamentally correlational, it is typical to apply the FA to the matrix of associations describing the relationships between the variables, rather than the raw data (as explained in Section 7.2.1). In order to do this, the analyst must decide which matrix of associations to use, the correlation matrix, the variance matrix or the covariance matrix (Henson and Roberts, 2006). Whilst some mathematical justifications upon which to base this decision have been published Dziuban and Shirkey (1974), none of the example applications discussed in this chapter appear to have employed these techniques. Instead, the decision appears to be led by preference and ease. For example, the use of the covariance matrix nullifies the necessity to standardise the data prior to the analysis, so might be preferred to avoid transforming the data, but it is still less commonplace than the use of the correlation matrix which is typically preferred (Dean, 2018).

Normality/Linearity/Variability in Data/Outliers

The appropriateness of using non-normally distributed data in FA is widely discussed across the literature with limited consensus, both in terms of univariate and multivariate normality. Some experts suggest that non-normally distributed data should never be employed in FA (Yong and Pearce, 2013), whilst others allow for its inclusion supported by claims that stable structures will still emerge in the solution regardless (Davies, 1978a).

Tabachnick and Fidell (2013) propose a pragmatic approach. Whilst they acknowledge that the inclusion of non-normally distributed variables in FA has the potential to degrade the final result, particularly if the direction of skew differs among the variables, and affirm that normality is preferable, they suggest that non-normally distributed variables may still be included, particularly in the analysis of datasets with large sample sizes, which could mitigate some of the impact. However, they do recommend looking at the distributions to be aware of the circumstances under which the FA is being applied, though this recommendation is limited to a review of the skewdness and kurtosis of single variables based on the "over sensitivity" of difficult to apply multivariate normality tests. This discussion has not led to generalised rules of thumb across the literature regarding hard boundaries dictating when a variable should be removed, if variables are to be considered on a case to case basis.

This is as per the standardised practice of geodemographic classification development. The normality checks made on the variables in the development of the 2011 OAC (Gale et al., 2016) are simply not led by hard rules, but rely on the developers' own interpretations and decisions. Thus, this element of subjectivity in the variable selection process remains. However, in the existing variable selection processes adopted in the creation of the 2011 OAC and other classifications, the normality checks are a significant element of the variable filtering process itself, and limited additional checks are made of the variables that remain. In the execution of FA, this is simply a preliminary data preparation stage prior to the main selection procedure.

Checking for multicollinearity and singularity

It is also necessary to remove multicollinearity and singularity from the input data prior to FA. Multicollinearity highlights variables which are too highly correlated to include in FA, effectively identifying variables which are representing the same phenomena and as such, are not both required (Williams et al., 2010). Conversely, singularity identifies redundancy in the case of a single variable representing the same phenomena as a combination of two or more other variables in the dataset (Tabachnick and Fidell, 2013). Including variables exhibiting either multicollinearity or singularity runs the risk of adding redundant information, inflating the prominence of a single dimension by inappropriately promoting (or weighting) such a dimension in the statistics in a way which does not reflect the real-world circumstance. This is one additional test of the input variables which is not commonplace in standard variable selection procedures in geodemographics.

In order to mitigate circumstances of multicollinearity, and since bivariate correlation is easily rectified with the removal of one or both of the variables, tests involving Squared Multiple Correlation (SMC) are recommended to identify problem variables to remove. Variables with SMC close to 1 indicates multicollinearity (Yong and Pearce, 2013). Tabachnick and Fidell (2013) explain that FA will not run where the input data includes perfectly correlated variables with SMC = 1, and statistical problems will be caused where the SMC is very close to 1. Consequently, thresholds of 0.01 to 0.0001 are typically used as a default tolerance (i.e. SMC $>$ 0.99 to SMC $>$ 0.9999) to identify variables which should be removed, with Tabachnick and Fidell (2013) suggesting the removal of variables from the analysis based on a hard cut-off of SMC $>$ 0.99. A threshold close to 1 is recommended since FA still relies on the existence of good correlation amongst the variables.

In the case of singularity, recommendations suggest the removal of variables which exhibit SMC close to 0, for example $<$ 0.1 (Yong and Pearce, 2013; Tabachnick and Fidell, 2013). However, in the case of multivariate correlation, it becomes more difficult to identify which variable should be removed. However, if there is singularity present in the data, FA will fail to complete, so it is possible to run tests to identify which variable may be breaking the analysis (Tabachnick and Fidell, 2013). The removal of problem variables will again rectify the issue, both mathematically and contextually in terms of removing unnecessary and redundant information.

Whilst correlation is regularly employed in variable selection for geodemographics, the interpretation is often far more subjective, and as in the 2011 OAC development, can be ignored if the results conflict with other priorities for including particular variables (Gale et al., 2016). In FA, this step is a more fundamental element of the process itself, and although there is some debate as to the necessary cut-off thresholds to adopt, the decisions to remove variables which are causing multicollinearity or singularity is, and should be, taken without regard for the context of the variable or the developer's preference for retaining it.

Confirming "Factorability"

For a dataset of variables to be "factorable" the correlation matrix (discussed in Section 7.3.1) must contain several considerable correlations of at least greater than ±0.3, a standard rule of thumb repeated throughout the literature (Tabachnick and Fidell, 2013). If this is not the case, then the use of FA should be reconsidered as the presence of latent constructs within the data is unlikely. When using FA to develop a geodemographic classification, this might offer an early indication that the candidate input variables do not reflect population structures, as hoped, and that other variables might need to be sought.

However, the presence of bivariate correlations is still not evidence of the presence of underlying factors (or social constructs, when applied to population attribute data) in and of itself, this could simply reflect relationships which simply exist between pairs of variables. Tabachnick and Fidell (2013) suggests that one might consider examining the matrix of partial correlations where pairwise correlations are adjusted for the effects of all other variables. Potential measures include the consideration of Bartlett sphericity test, or by use of the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Dziuban and Shirkey, 1974). Both are well established methods for assessing the factorability of data prior to FA with a rich academic history (Dziuban and Shirkey, 1974; Knapp and Swoyer, 1967; Kaiser, 1970). For the former, a result of 0.05 is deemed statistically significant (Dziuban and Shirkey, 1974; Tabachnick and Fidell, 2013). The rules are less clear for the latter, with Tabachnick and Fidell (2013) recommending a KMO > 0.6 as desirable, whilst one of the founders of the method Kaiser (1970) suggested a KMO > 0.8 might be necessary to indicate "good" factor-analytic data, or even > 0.9 to indicate "excellent" data. The use of the KMO measure is recommended by Tabachnick and Fidell (2013) due to a perceived over-sensitivity on Bartlett's test.

7.3.2 Setting up a model

Select extraction method

There is more than one way to extract factors in FA. In a practical sense, FA requires an "extraction method" as an input parameter to instruct the analysis on how to mathematically identify the factors within the data. This must be selected upfront by the analyst. The extraction methods seek to remove variance which is common to sets of variables. Each time a common variance for a set of variables is found, it is identified as a factor. Subsequent factors are then found to explain as much of the remaining variance as possible, until no more factors can explain the remaining variance (Henson and Roberts, 2006). Each extraction method employs a different approach to this process. Costello and Osborne (2005) noted the limited discussion regarding the strengths and weaknesses of the possible methods was scarce over 15 years ago, and not much seems to have changed in the intervening period. Nevertheless, Tabachnick and Fidell (2013) do provide some technical summaries of many of the available extraction methods and their differences.

No single method is universally recommended, the decision can vary based on the specific context of individual studies (Costello and Osborne, 2005). This process is often guided more by pragmatic considerations than theoretical (Yong and Pearce, 2013). Though the selection requires attention, some suggest that there should be limited difference in results derived from different extraction methods, particularly in FA applied to datasets with many variables, large sample size, and with variables with similar communality estimates (Tabachnick and Fidell, 2013; Davies, 1978a). However, several experts do indicate a preference for Principal Axis Factoring, particularly if there is potential for the data to be non-normally distributed, as was demonstrated of much of the input data underpinning the 2011 OAC (Gale et al., 2016), since it seems to handle this kind of data better than other methods (Tabachnick and Fidell, 2013; Costello and Osborne, 2005).

Select rotation method

A method of "rotation" is also a standard input parameter into FA. Rotation in this context seeks to explain the correlations identified in the analysis, and as a result, improves the interpretability of the factors identified. Again, Tabachnick and Fidell (2013) provide a detailed explanation of the underlying mathematical processes involved. Following the extraction, there is an infinite number of possible rotations which could be applied, each of which accounts for an equal amount of variance, but which differ slightly in their definition of the factors (Tabachnick and Fidell, 2013). Several methods of rotation could be applied. These can be divided into two main categories, *oblique rotation* and *orthogonal rotation*. See Rummel (1967) for a comprehensive overview of the mathematical principles of each. In practice, an oblique rotation may be necessary if there is correlation between the factors, otherwise an orthogonal rotation should suffice. The FA solution must have at least two factors to be considered for rotation (Dean, 2018).

Again, the advice conflicts in the use and the selection of a rotation method. Voas and Williamson (2001) recommend use of a rotation method only when it improves the interpretability of the result. Conversely, Yong and Pearce (2013) recommend use regardless, irrespective of extraction technique, to reduce the ambiguity of the raw factors. Whilst some always recommend the default use of orthogonal rotations for their simplicity (Tabachnick and Fidell, 2013), others suggest the default use of an oblique method, proposing that an orthogonal solution will be produced anyway if correlation does not exist (Costello and Osborne, 2005; Hunter, 1972). These debates demonstrate just one of the numerous examples of conflicting advice that can make the confident execution of FA difficult which, as mentioned, has weakened trust in the use of geodemographic classifications in the past.

The *varimax* rotation, an orthogonal rotation method, has been the longstanding common default rotation recommended across the literature (Tabachnick and Fidell, 2013; Hunter, 1972). This was also the default rotation in practice in the *fa* package for carrying out FA in the statistical computing language R (which will be used to compute the solution in the Case Study in Section 7.4). However, it has since been replaced (in 2009) with the oblique *oblimin* rotation (Revelle, 2020), though it is possible to manually override the default option.

Select number of factors

When factoring the variables, the total number of possible factors is equal to the number of variables. However, some of these factors may not be interpretable, or may contribute minimally to the solution. These are not useful to the analysis and could introduce unnecessary noise or error. In this case, it is advised to retain only the contributing factors. Since the aim of the analysis is to derive the minimum number of factors whilst capturing the maximum variance, selecting the optimum number of factors to achieve this is an important decision to be made (Henson and Roberts, 2006).

There is, again, no single method consistently recommended for identifying the number of factors which should be retained, and there is some debate regarding the success of each of the options. The most common preference is to adopt *Kaiser's criterion*, evaluating the eigenvalue of each factor and retaining just those with an eigenvalue > 1, where the eigenvalue represents the amount of variance exhibited in the given direction represented by the factor (Tabachnick and Fidell, 2013). This is the default in most statistical software packages. However, though commonplace, there is evidence cited across the academic literature that this may be an inaccurate method (Costello and Osborne, 2005). In citing two separate studies which have conflictingly found that this rule could both substantially overestimate the number of factors to retain and might actually underestimate the number of factors, Henson and Roberts (2006) highlight the difficulties in employing this criterion with complete confidence. Alternatively *Joliffe's criterion* suggests 0.7 as a more appropriate threshold (Jolliffe, 1972; Tabachnick and Fidell, 2013). However, this too should be used with caution (Yong and Pearce, 2013). Alternatively Bartlett's Chi-Squared can be employed to decide the number of factors which should be kept, although this is also regarded as inconsistent and particularly influenced by sample size and non-normal distributions in the data (Henson and Roberts, 2006; Dean, 2018). Since non-normality is an issue in the 2011 OAC candidate variables (Gale et al., 2016), this would not offer the most suitable solutions. In practice, a decision is often achieved by running PCA on the data and identifying the number of meaningful components based on one of the thresholds proposed (Dean, 2018).

However, regarded as the most popular test, the scree test is most regularly recommended for informing the number of required factors (Tabachnick and Fidell, 2013; Liu et al., 2019), as is similarly commonly employed in deciding the number of clusters to derive in a geodemographic classification development (see Section 3.3.2). Such a test involves examining the graph of the eigenvalues, looking for the natural bend in the data, the inflection point, indicating a flattening of the curve (Yong and Pearce, 2013). The recommended factor count is drawn from the number of datapoints prior to the bend identified (Costello and Osborne, 2005). Although this test is once again not without its criticisms, focused largely around its subjective nature and potential for producing unclear results (Henson and Roberts, 2006;

Costello and Osborne, 2005).

Finally, Parallel Analysis offers another alternative approach. In summary, the process involves averaging the eigenvalues of each component derived in repeated runs of PCA on a randomly generated simulated dataset and comparing to the eigenvalues of the components from a PCA of the real data, retaining only those greater than the average results (Tabachnick and Fidell, 2013). Though its superior accuracy beyond the other tests presented here has been proposed for some time, the complexity of historically calculating it by hand has limited uptake. Since its inclusion in modern statistical software packages, however, its popularity has seemingly increased somewhat (Henson and Roberts, 2006).

Due to the many caveats documented, analysts are often recommended to make an informed decision based on the outcome of more than one of these techniques (Tabachnick and Fidell, 2013; Henson and Roberts, 2006; Yong and Pearce, 2013). Alternatively, an iterative approach could be taken, completing several FA runs, initially guided by the rules and tests, with the final factor count dictated by the outcome which generates the most desirable solution, defined by the desirable characteristics outlined in the next section (Yong and Pearce, 2013). Again, this phase of the FA process introduces several key decisions to be made by the developer, which could affect the outcome of the variable selection.

7.3.3 Setting model parameters and rules to support interpretation

The aim of the FA is to seek a “strong” dataset with uniformly high communalities, without cross-loadings and with several variables loading on each factor (Costello and Osborne, 2005). Variables are discarded throughout the process until this objective is achieved. The dataset which remains thus presents the selected variable set. As such, this is the core variable selection process of the FA. To achieve this, it is necessary to define the parameters at which these objectives will be judged to have been met, in the geodemographics context. These are the parameters which dictate the candidate variables which should be retained for inclusion in the subsequent clustering process, and which should be discarded. The idea, as mentioned at the outset of this chapter, is to set these parameters upfront, to introduce much desired objectivity in to this variable selection process, wherever possible.

The following discussions of each of the parameters to set provide the context and justification for the decisions made in the case study in Section 7.4. It also further demonstrates the difficulties involved with confident applications of FA for variable selection, in practice. Moreover, many are not considerations required in the execution of FA where the objective is not to discard candidate variables for filtering purposes, but where the factors will be used as input variables in the geodemographic classification, which has been the predominant use of FA, as discussed. As such, the explicit discussion of these decisions, again, are limited in the existing geodemographics focused literature.

Communality thresholds

The first parameter to set is the "communality threshold". Since communality represents the variance of a variable explained through the factors (see Section 7.2.1), and the aim is to explain the maximum variance through the factors (Yong and Pearce, 2013), it is advisable to seek to retain variables with high communality scores. As a common rule of thumb, any variables with communality scores ≤ 0.2 (representing 80% unique variance) should be removed, and the FA should be run again on the remaining variables (Tabachnick and Fidell, 2013; Yong and Pearce, 2013). This process can be iteratively repeated until no variables fall below the threshold.

Loading thresholds

The "loading threshold" is the next parameter to set. FA seeks *high* factor loadings for variable removal/retention, since variables with higher loading scores on a given factor contribute more to the factor (see Section 7.2.1), and thus factors with higher associated loadings are considered more meaningful. To generate the best final result, some recommend the removal of any variables which do not load high on any factor (Yong and Pearce, 2013). Automatically, it is possible to simply ignore these variables in the interpretation of the result (Kline, 2014; Tabachnick and Fidell, 2013), however, if the aim is for variable selection, seeking to reduce the original variable set and to identify the useful variables, it makes more sense to *remove* these variables and re-run the analysis until there is a strong final output (Williams et al., 2010). As a counter argument, Yong and Pearce (2013) contemplate whether low loadings might in themselves offer useful information, in reflecting an absence in influence of the variable to a factor. Though this is not discussed elsewhere across the literature, and the standard rules remain to remove or ignore these variables. In employment in the variable selection phase in a geodemographic classification development, the former, removal, is the most appropriate option.

In order to identify the variables for removal, it is necessary to set a threshold above which any loading score is recognised as high. However, several rules of thumb are suggested across the literature. Studies investigating the significance of loadings have demonstrated 0.3 as a useful, if conservative, such threshold (Davies, 1978a; Yong and Pearce, 2013). Though this threshold is most popular (Hogarty et al., 2004), alternative thresholds are regularly suggested. For example, Tabachnick and Fidell (2013) propose 0.45 as high and Costello and Osborne (2005) suggest 0.5. Others offer several categorical thresholds, such as Kline (2014) who cites 0.3 as moderately high but prefers the use of 0.6 as a high loading score, and Comrey and Lee (1992) who present a scale from an excellent loading (0.71) through to very good (0.63), good (0.55), fair (0.45) and poor (0.32).

Additionally, the outset of this section mentioned a "strong" result as containing several variables loading on each factor, recommending that several variables should load highly on each factor for the factor to be considered useful (Costello and Osborne, 2005). Most mentions of this in the literature suggest that a factor should only be retained in the analysis if 3 or more variables load highly on the factor, labelling factors with fewer high loading variables as "weak", "unstable" and potentially meaningless and recommending their removal from the analysis (reducing the analysis by the number of factors displaying this characteristic) (Hogarty et al., 2004; Tabachnick and Fidell, 2013; Costello and Osborne, 2005). Although Yong and Pearce (2013) make a case for considering any with just 2 variables and deciding subjectively, based on the context, on an individual basis. When a factor is removed, the FA should be re-run with a reduced factor count, adjusting for the number of factors removed.

Dealing with variables loading across factors

Another suggestion which appears across the literature proposes that a strong solution does not contain any variables which load highly on multiple factors (Costello and Osborne, 2005). These are interchangeably referred to as *complex variables*, *cross-loading variables* and *split loadings*. However, discussion surrounding such a phenomenon again fail to consensus, both in the significance of the threat posed, and in how to handle such variables. Some earlier studies were not as concerned by cross-loading variables, opting to retain them in the analysis without further consideration (Anderson and Bean, 1961). However, discussion has increased in later years to consider the removal of any such variables (Williams et al., 2010; Yong and Pearce, 2013). Whilst with some recommending their immediate removal, followed by a re-running of the analysis, others suggest a more nuanced approach, supporting their decisions by considering the contextual relevance of the cross-loading, and the implications of retaining the variables on a case-by-case basis (Tabachnick and Fidell, 2013), arguing that some variables may legitimately contribute to more than one latent construct at play in the structure of the data (Costello and Osborne, 2005). This argument seemingly makes sense in the identification of urban structures, where a single character attribute might very well be common to multiple societal constructs. For example, a student population might be characterised by a high proportion of 16 to 25-year-olds, but so too might areas containing settled families with older children still living at home.

A compromised approach which might allow for such occurrences, but which also could retain some objectivity is the idea of a 'hard cut-off', choosing a cross-loading threshold beyond which the variables are removed (Yong and Pearce, 2013). However, as with the overall loadings threshold, there is again no consensus on what that should be. Again, whichever approach is to be taken, and any related thresholds, should be decided upfront.

Evaluation of eigenvalues and explained variance

finally, as outlined in Section 7.2.1, the greater the variance of a factor, the more useful and important the factor. Solutions exclusively containing factors exhibiting high variance are therefore desirable and represent a stronger result. Dean (2018) suggests considering the *percentage of variance explained* by each factor. He specifically suggests identifying any with variance $< 10\%$, and re-running the analysis with this many fewer factors, since the removal of any factor affects the structure of the result and thus necessitates a re-calculation of the result. However, this threshold is not widely recommended across the literature. More commonly an evaluation of the eigenvalues of the factors is recommended, removing factors based either on Kaiser's criterion or Joliffe's criterion (as discussed above).

Yong and Pearce (2013) make a further suggestion that analysts should not just consider the variance of a single factor, but should aim for factors to *cumulatively* explain approximately 75-85% of the variance of the original associations matrix. However, whilst agreeing on the approach, Williams et al. (2010) explains that such thresholds are much debated and can depend strongly on the context of the data, suggesting varying expectations between the natural sciences, which expects as high as 95%, and the humanities, in which the variance explained is often far lower. A fixed threshold is therefore rarely employed. Typically, a value as high as possible is sought. Nevertheless, it is still often a useful indication of the strength of the result (Tabachnick and Fidell, 2013).

7.3.4 Summary

The difficulty of selecting the most appropriate model set-up is evident. Based on the conflicting advice outlined, Hogarty et al. (2004) question whether simple rules of thumb (such as those suggested throughout this section) should be relied upon to guide an application of FA, and whether contextual factors should instead be considered when making all decisions. However, such an approach could introduce undesirable subjectivity into the model and risks the potential of overfitting the model to the single dataset upon which it has been built, which again should be avoided, if possible.

This section has sought to summarise the literature relating to FA, presenting several sides of the lively debates which surround the decisions to be made in the process into a single source. The following section presents a case study employing FA as a method for variable selection in a Leeds-specific geodemographic classification, as a further extension of the LSOAC developed in Chapter 4. The absence of such a single source outlining and supporting the practical decisions which will need to be made in the execution of the case study, specifically in the context of geodemographics, initially presented a barrier to the confident completion of the case study. As such, this summary offers a much needed clarity and a framework which can be used to facilitate the step-by-step process and support the decisions made throughout.

The next section thus builds on this work, presenting the case study as a demonstrative example of using FA for variable selection within the Standard Framework for geodemographic classification as an alternative to the variable selection procedures demonstrated thus far.

7.4 Case study application

7.4.1 Introduction

The following section outlines one example application of FA for variable selection in the development of a geodemographic classification for Leeds, for which the remainder of the adopted framework replicates the methodology adopted in Chapter 4, adapted from the 2011 OAC methodology (detailed in full in Section 3.4). Decisions in the FA process are made throughout based on the discussions presented in Section 7.3. Due to the existence of much conflicting advice, this is not presented as the single, final solution for the factors in the city, but simply offers a demonstration of the process in action to consider the potential of FA for improving upon the existing variable selection processes employed in other studies.

The FA model developed is used to select a set of input variables, for a subsequent geodemographic classification, from a larger set of candidate variables. The final selection is then run through the clustering methodology of the 2011 OAC to derive a new classification for comparison. This work focuses just on the records relating to the 2,543 OAs in the Leeds LA boundary, as per the Case Study in Chapter 4. As such, the final result is compared to the LSOAC developed in that chapter, rather than the 2011 OAC itself, to ensure that conclusions made are based on the alternative variable selection process used here and are not also affected by the shift in geographic extent.

The three phases of FA described in Section 7.3 (data preparation, model set-up and execution of the FA method on the prepared candidate data with the parameters outlined) are each conducted in the case study (Section 7.4.4), followed by a subsequent k-means clustering of the resulting variable set to derive the new classification for comparison (Section 7.4.5), supporting a review of the impact of the proposed variable selection process.

While decisions made throughout the FA phase are based either on the recommendations made in Section 7.3, other decisions in the remainder of the Standard Framework are made to maintain consistency and promote comparability, mirroring those made in the development of the 2011 OAC to promote comparability with the LSOAC derived in Chapter 4. The latter is true particularly in the data preparation phase, and certainly in the k-means analysis. Such decisions and their basis are highlighted.

7.4.2 Background

As described in Chapter 3, two empirical methods were employed in the variable selection methodology adopted in the 2011 OAC (Gale et al., 2016). First, pairwise correlations between each of the variables were considered, and several variables were removed based on a guide of correlations ≥ 0.6 representing redundancy. Next, a cluster based sensitivity analysis is conducted to test the "impact" of each of the remaining variables in the formation of clusters. This is based on a measure of the total WCSS (see Section 3.4). These tests were used subjectively to guide the final selection of the input variables for the cluster analysis (as discussed in Section 6.4).This process identified a final set of 60 variables to be employed in the cluster analysis.

The correlation analysis is employed to mitigate for the inclusion of variables with shared dimensions which could add a superficial prominence in this direction by inadvertently weighting such variables (Gale et al., 2016). Though the FA tested in this chapter is primarily adopted in place of the sensitivity analysis, built-in procedures should also mitigate against the inclusion of redundancy relating to variables correlating too highly.

7.4.3 Proposed framework

Figure 7.3 illustrates the complete FA process adopted in the case study application, developed as a result of the discussions presented in Section 7.3. The process is iterative and runs until all criteria are met and thus an optimal solution is reached, in doing so, reducing the candidate variable set and executing a variable selection.

7.4.4 Conducting the Factor Analysis

Data and data preparation

To enable a comparison between the final input variable selection made for the 2011 OAC and the output of this study, the initial set of candidate variables are the same 167 census variables which Gale et al. (2016) considered for inclusion in the 2011 OAC (listed in full in (Gale, 2014, p.475)). However, as per the study in Chapter 4, the data used here is filtered just for the records relating to the 2,453 OAs in the Leeds LA boundary.

These candidate variables were initially selected by Gale et al. (2016) for their consistency across each of the different censuses of England and Wales, Scotland and Northern Ireland and for representing the five domains of interest (see Section 3.4.2). Though this criteria represents some pre-filtering in the initial identification of the candidate variables, some of which is based on practical limitations no longer present when considering a single city represented by just one census, this has been overlooked in favour of consistency and comparability, and the resulting 167 variables are adopted nevertheless.

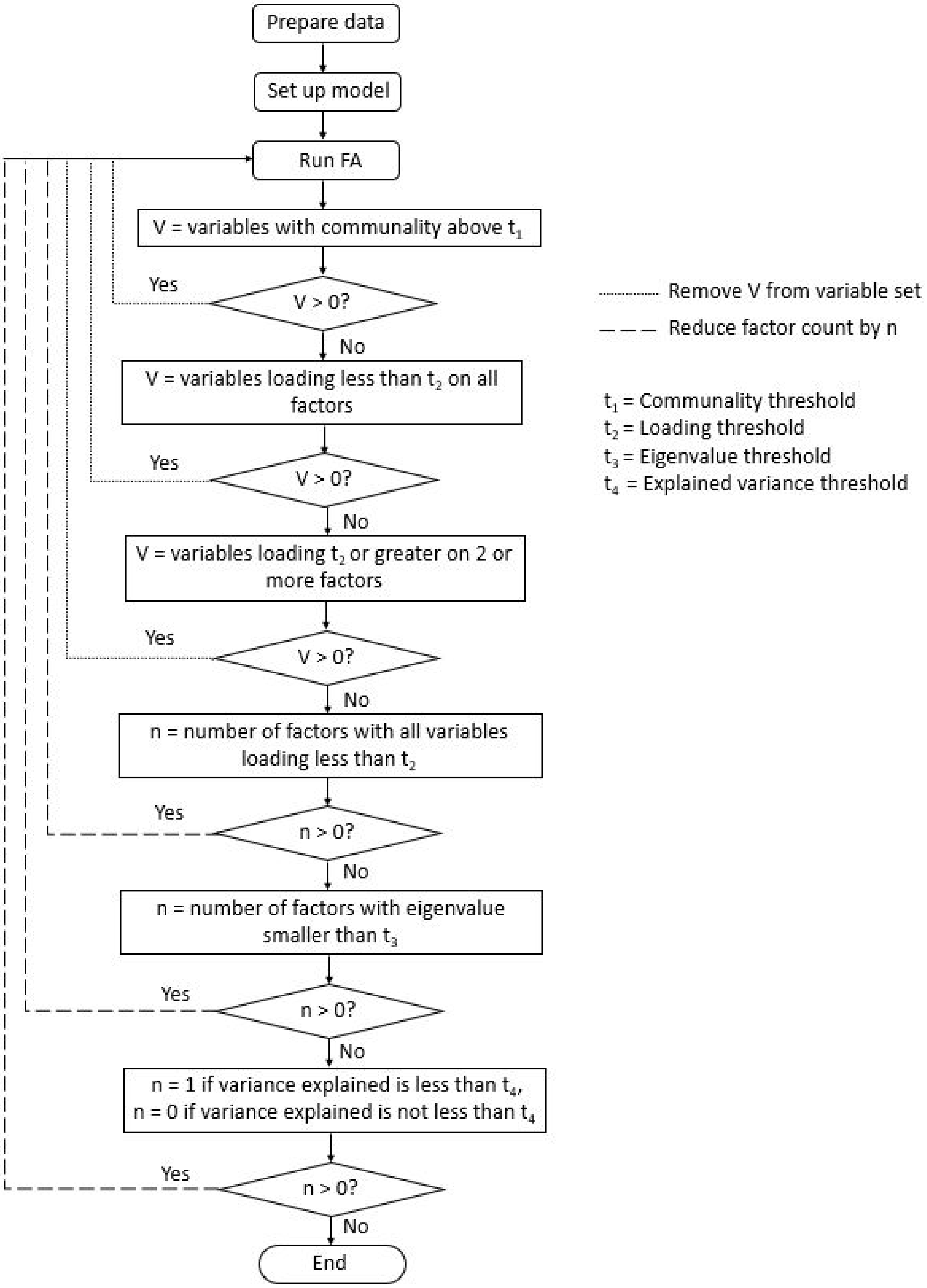

Prepare data
Set up model
Run FA
V = variables with communality above t_1
V > 0?
Yes
No
V = variables loading less than t_2 on all factors
V > 0?
Yes
No
V = variables loading t_2 or greater on 2 or more factors
V > 0?
Yes
No
n = number of factors with all variables loading less than t_2
n > 0?
Yes
No
n = number of factors with eigenvalue smaller than t_3
n > 0?
Yes
No
n = 1 if variance explained is less than t_4, n = 0 if variance explained is not less than t_4
n > 0?
Yes
No
End
Remove V from variable set
Reduce factor count by n
t_1 = Communality threshold
t_2 = Loading threshold
t_3 = Eigenvalue threshold
t_4 = Explained variance threshold

Figure 7.3: Iterative decision process of the FA leading to optimum solution.

Though there is limited data cleaning required, since data from the census is already clean, complete and comprehensive (see Section 5.2.4), the following sections prepare the data further, including as per some of the decisions made by Gale (2014) and Gale et al. (2016), in addition to the data preparation required in FA.

Pre-variable selection preparation (as per 2011 OAC)

As per the 2011 OAC methodology, variables which share a denominator are combined to generate composite variables upfront. As per Gale et al. (2016) rationale, the separate inclusion of each of these variables could add a weighting through the promotion of the shared dimension which, rather than being an interesting representation of a shared relationship, simply highlights that these variables are representative of the same phenomena, potentially adding unhelpful noise and redundancy. For example, individuals who are separated and divorced are combined into a shared composite variable with limited loss of information.

This could also benefit variables which individually represent low counts of the population, but naturally group with other variables to represent a larger share, for instance, in the combining of some ethnic minority groups. Again, this process is fairly subjective and could have been omitted, but is included to maintain consistency with the 2011 OAC and to ensure that the variable selection process is being applied, as close as possible, to the same candidate data set in both instances. Table 7.2 lists the composite variables made, as per the 2011 OAC justifications (see Gale (2014, p.475) for the original rationale).

Though most of the variables which correlate too highly and thus introduce redundancy is identified and removed within the FA itself (specifically in the check of multicollinearity and singularity), some of the candidate variable rejections made in the 2011 OAC methodology are to be upheld upfront. This is the case where the generation of the above composite variables have introduced new variables which are directly represented in other, existing variables, or where Gale (2014) has identified particular variables as having limited descriptive power, and thus being of limited interest. The variables which are discarded based on these criteria are listed in Table 7.3.

Data transformation

Transformation processes were employed by Gale et al. (2016) in the 2011 OAC development process both to standardise the data to a consistent scale, and to account for the impact of varying degrees of skew and outliers found in the data, which if left unaddressed could have compromised the quality of the clustering process (Gale et al., 2016; Vickers et al., 2005). Both objectives are similarly requirements of FA (Section 7.3). Thus, the

New variable description	Original candidate variables combined
Age 5-14	“u008”: Age 5-9, “u009”: Age 10-14
Age 25-44	“u012”: Age 25-29, “u013”: Age 30-44
Age 45-64	“u014”: Age 45-59, “u015”: Age 60-64
Age 65-89	“u016”: Age 65-74, “u017”: Age 75-84, “u018”: Age 85-89
Married or in a registered same-sex civil partnership	“u023”: Married, “u024”: In a registered same-sex civil partnership
Separated or divorced	“u025”: Separated, “u026”: Divorced
White	“u028”: White: British and Irish, “u029”: White: Other
Asian/Asian British	“u034”: Asian/Asian British: Chinese, “u035”: Asian/Asian British: Other
Region of birth: UK/Ireland	“u042”: Region of Birth: UK, “u043”: Region of Birth: Ireland
Main language not English/ No English	“u049”: Main language is not English: Cannot speak English well, “u050”: Main language is not English: Cannot speak English
One family only: Married, same-sex civil partnership or cohabiting couple: No children	“u062”: One family only: Married or same-sex civil partnership couple: No children, “u065”: One family only: Cohabiting couple: No children
One family only: Married, same-sex civil partnership or cohabiting couple, or lone parent: All children non-dependent	“u064”: One family only: Married or same-sex civil partnership couple: All children non-dependent, “u067”: One family only: Cohabiting couple: All children non-dependent, “u069”: One family only: Lone parent: All children non-dependent
Occupancy rating (rooms) of -1 or less	“u098”: Occupancy rating (rooms) of -1, “u099”: Occupancy rating (rooms) of -2 or less
Part-time work	“u137”: Part-time: 15 hours or less worked, “u138”: Part-time: 16 to 30 hours worked
Full-time work	“u139”: Full-time: 31 to 48 hours worked, “u140”: Full-time: 49 or more hours worked
Work in mining, quarrying or construction industries	“u142”: Mining and quarrying, “u146”: Construction
Work in energy, water or air conditioning supply industries	“u144”: Electricity, gas, steam and air conditioning supply, “u145”: Water supply; sewerage, waste management and remediation activities
Information and communication or professional, scientific and technical activities industries	“u150”: Information and communication, “u153”: Professional, scientific and technical activities
Financial, insurance or real estate industries	“u151”: Financial and insurance activities, “u152”: Real estate activities

Table 7.2: Summary of new composite variables made.

Variable removed	Justification for removal
"u097": Occupancy rating (rooms) of 0	Represented by the composite variable of "u098" and "u099": Occupancy rating (rooms) of -1 or less.
"u105": Very good health	"u104": SIR used as a preferred indicator of health.
"u106": Good health	"u104": SIR used as a preferred indicator of health.
"u107": Fair health	"u104": SIR used as a preferred indicator of health.
"u108": Bad health	"u104": SIR used as a preferred indicator of health.
"u109": Very bad health	"u104": SIR used as a preferred indicator of health.
"u123": Economically active: Part-time	Represented by the composite variable of "u137" and "u138": Part-time work.
"u124": Economically active: Full-time	Represented by the composite variable of "u139" and "u140": Full-time work.
"u133": Unemployed: Age 16 to 24	Represented by "u126": Economically active: Unemployed.
"u134": Unemployed: Age 50 to 74	Represented by "u126": Economically active: Unemployed.

Table 7.3: Summary of variables to be removed based on redundancy due to new composite variables, or the limited descriptive power of the variable.

same transformation procedures are repeated here to maintain consistency. However, transformation procedures should always be undertaken with caution (Singleton and Spielman, 2014).

The variables are first converted to percentages to reflect the proportion of the population represented by each attribute in each OA. This is a straightforward process for 129 of the variables, where the original unit of measurement represents a count of the population, however, 5 variables need separate consideration. As per the 2011 OAC methodology, variables "u006: Density (number of persons per hectare)" and "u104: Day-to-day activities limited a lot or a little Standardised Illness Ratio" are adopted as ratios, re-calculated for the population of Leeds, as per Section 4.4.1. However, the latter is first adjusted (as per Section 3.4.2) to account for age distribution across the city. The final 3 non-count variables are removed upfront for a combination of inappropriate unit measurements and redundancy, see Table 7.4.

Appendix C.2 contains the updated list of remaining 131 final candidate variables to be input into the FA, and their updated associated variable codes which is referred to going forwards.

Variable removed	Justification for removal
"u005": Area (in hectares)	Incomparable unit measurement. Attribute appropriately represented by "u006": Density (number of persons per hectare).
"u020": Mean age	Incomparable unit measurement. Attribute appropriately represented by age indicators "u007"-"u019".
"u021": Median age	Incomparable unit measurement. Attribute appropriately represented by age indicators "u007"-"u019".

Table 7.4: Summary of variables to be removed based on redundancy due to new composite variables, or the limited descriptive power of the variable.

An Inverse Hyperbolic Sine (IHS) is next employed to normalise this final dataset, before the data is range standardised on a scale of 0-1, as per the transformation process adopted in the development of the 2011 OAC. The rationale behind the use of these specific transformation processes, which were carefully selected based on tests involving a series of potential alternatives, is outlined in detail in Gale et al. (2016).

Checking for normality and outliers

Though the data has already been transformed, Figure 7.4 indicates that some non-normality in terms of skew and some outliers remains within the dataset. Each of these characteristics have the potential to weaken the output of both the FA and clustering processes, though both processes will still run with data of this kind included (see Section 7.3.1).

In the development of the 2011 OAC, Gale et al. (2016) empirically considered the skew of each variable individually before deciding on the inclusion or exclusion of the variable. This decision was informed by, but not completely determined by, variables with skew values beyond a threshold of ±1. Instead, contextual considerations were also applied to those beyond this threshold, and those which were identified as containing the capacity for area differentiation were retained (Gale, 2014). Moreover, since the skew is directly linked to the real-world presence or absence of each variable, it was noted that an extremely low or high presence of a given variable could offer information of value (Gale et al., 2016). In the development of a classification of commuting flows for England and Wales based on the 2011 OAC, Hincks et al. (2018) also removed variables on a case-by-case basis dependent on a consideration of outliers, skewness, kurtosis and correlation.

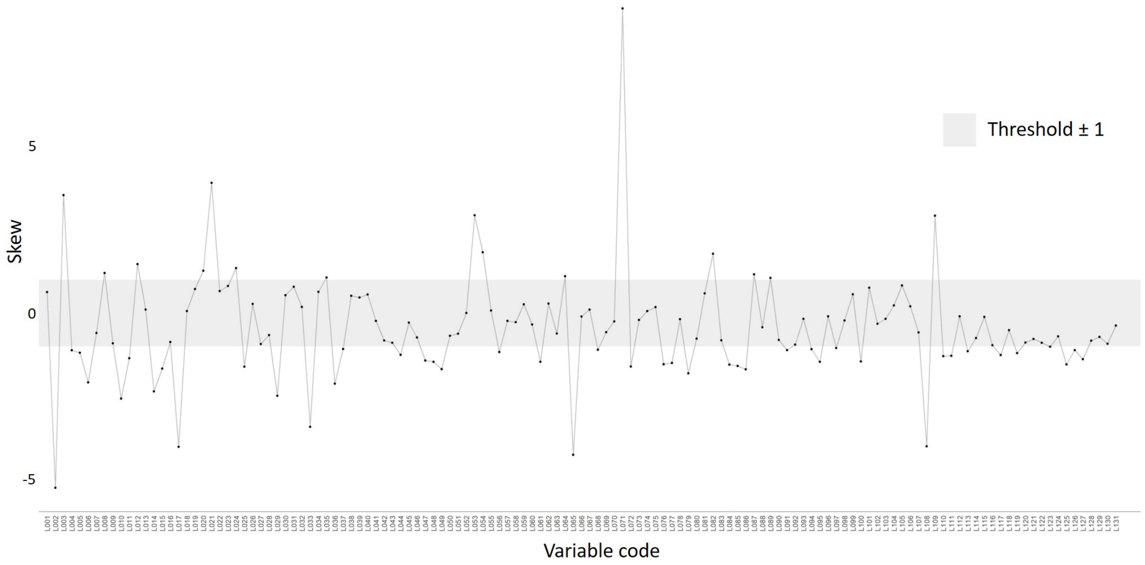


Figure 7.4: Plot of skew for each of the 131 variables in the candidate variable set.

Of the 131 variables here, 59 fall beyond the threshold proposed by Gale et al. (2016) (Figure 7.4). In an effort to reduce the introduction of unnecessary subjectivity, and based on the combination of Gale's contextual concerns and the recommendations in Section 7.3.1, variables are not removed in this analysis subject to non-normal distributions.

Similarly, no action is taken to remove any outliers which may exist. The outliers do not appear to be causing association which do not exist, nor are they the result of missing data. Moreover, whilst there are areas with a far higher or lower than typical presence of one particular attribute present in this data, these are legitimate results and some important information could be lost in their removal.

Checking for multicollinearity and singularity

The SMC has been calculated for each variable as per the recommendations in Section 7.3.1. The highest and lowest SMC are 0.989 (for "L014") and 0.170 (for "L028"), which are respectively smaller and larger than the recommended thresholds outlined. As such, no variables are removed based on concerns relating to multicollinearity or singularity.

Confirming Factorability

There are good sized correlations in the data, indicative of data appropriate for FA. An additional run of KMO returns a measure of 0.94, greater then the guidance threshold of 0.7, at which a dataset is considered to have good factorability (see Section 7.3.1).

Set-up of final model

Select extraction and rotation methods

As per guidance from Tabachnick and Fidell (2013), a Principal Axis Factoring feature extraction method is used in the FA with a varimax rotation. This is favourable for this dataset, which contains non-normally distributed data (see Section 7.3.2). The rotation method selection is considered trivial since the interpretability of the factors is of reduced importance, based on the objective of the FA being variable selection for subsequent k-means clustering and not as a means to its own end.

A multi-method approach is considered, conducting several FAs based on a variety of extraction techniques to test for the best result, however, since the aim of the study is simply to demonstrate the potential of FA for variable selection in this context and not to seek and present the best result, this approach is judged to be outside of the scope of this study (but will be discussed further in Section 7.5).

Select number of factors

As recommended in Section 7.3.2, several methods for selecting the number of factors are considered. The first method utilises PCA. Figure 7.5 shows the scree plot of a preliminary PCA run on the data, with markers indicating the number of components adhering to Kaiser's criterion (eigenvalue ≥ 1) in red, and Joliffe's criterion (eigenvalue ≥ 0.7) in blue. The former recommends a solution with 17 factors, and the latter recommends a solution with 29 factors. Evaluating the scree plot itself, considering the natural bend in the data, one might recommend a solution with 15 factors, although such a recommendation is subjective and another developer might observe a different natural break.

The results of a Parallel Analysis run on the data (shown in Figure 7.6, the resulting default plot returned from conducting Parallel Analysis in statistical software R using the "fa.parallel" function in the "psych" package) alternatively suggests a solution with 14 factors.

Since the FA process employed here will iteratively remove unnecessary factors (thus reducing the factor count) until an optimal solution is found, as per Figure 7.3, the largest recommended number of factors, 29, is chosen as an initial starting parameter for the case-

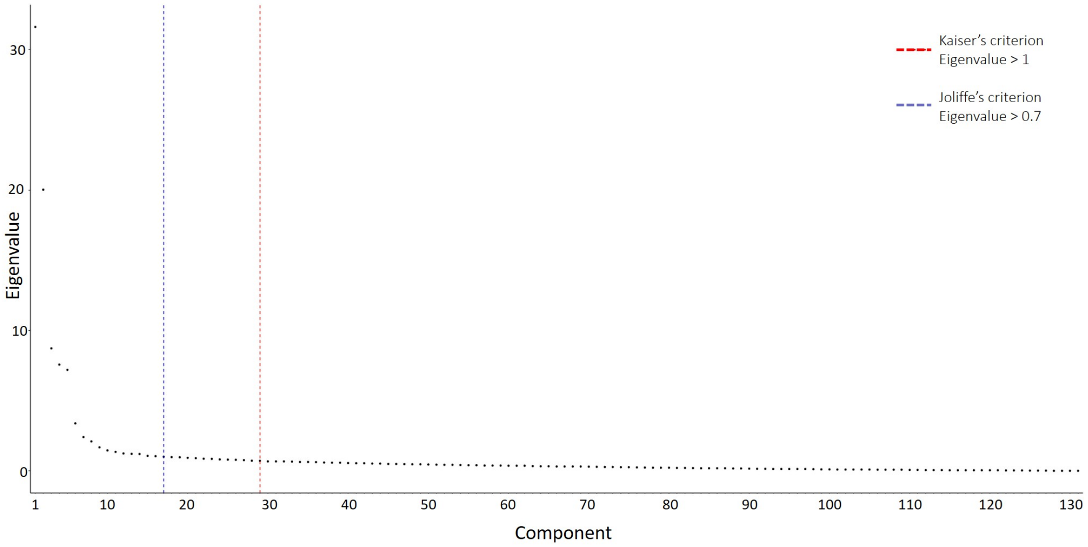


Figure 7.5: Scree plot of PCA applied to the 131 candidate variables.

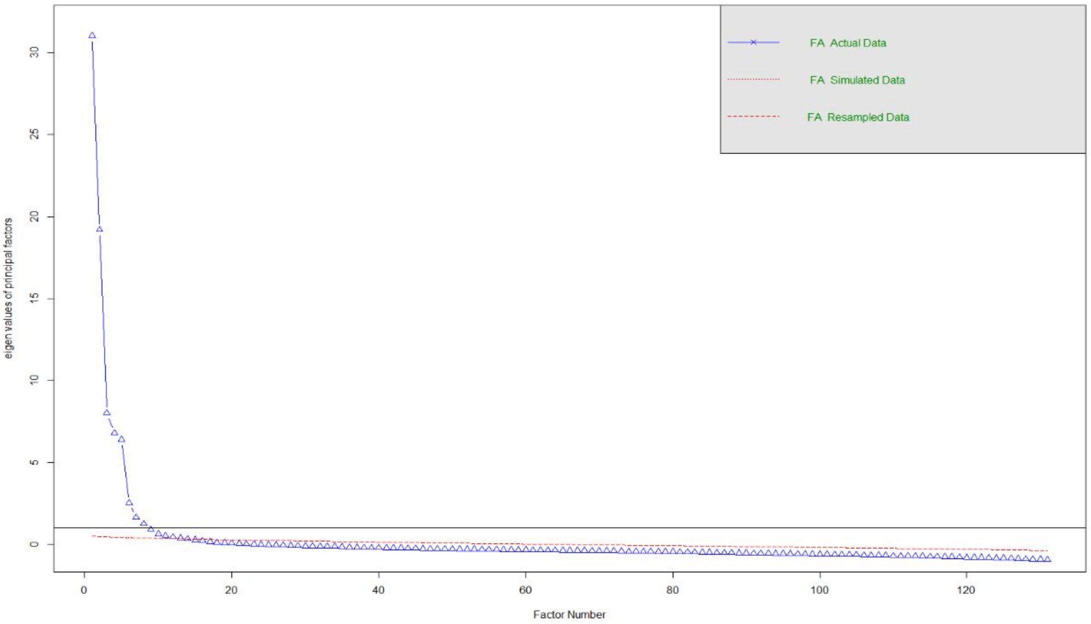


Figure 7.6: Scree plot of parallel analysis.

study application. The process will further reduce and optimise this final count.

Define parameters, thresholds and rules governing the FA

All of the decisions outlined in this section are made as per the discussions in Section 7.3.3.

High communality is set at a communality threshold of 0.2.

To account for the lack of consensus in the threshold determining 'high' loading, the FA process is run seven times, varying the loading threshold through 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9, to test the breadth of thresholds recommended within the literature.

Cross-loading, complex variables are retained based on their potential for contextual relevance. Each case could be considered individually, but in the interest of reducing the introduction of subjectivity, all examples of complex variables are universally retained.

Factors on which fewer than 3 variables load highly (as per the threshold in the test run) are removed (i.e. the factor count is reduced by the number of factors meeting this criteria and the FA is re-run). Again, cases with 2 highly loading variables are not considered for their individual merits to avoid the introduction of subjectivity.

A result for which the eigenvalue of all factors meets Joliffe's criterion (and by default Kaiser's criterion) is sought. Additionally, a result which explains greater then 75% variance is desired, however, this is not built into the model, instead, this criteria is used to evaluate the result of the model upon completion.

The model is iteratively run through, removing variables and reducing the factor count as necessary, until an optimal solution is achieved, whereby the above criteria are satisfied. The variables which remain following the completion of each test represent the variables selected in each case.

Final model with parameters

Figure 7.7 shows the final FA process which is detailed throughout this section, including the selected parameters, which is used to filter the candidate variable set in this case-study application. As noted, this is run iteratively throughout the listed loading thresholds (t_1).

7.4.5 K-means clustering

The resulting variable selections from each test are subsequently clustered using a k-means algorithm, as per the methodology used to derive the Supergroups of the 2011 OAC (Gale et al., 2016). All parameters of the 2011 OAC clustering methodology are maintained, including a cluster count of 8, and the optimisation process by which the best solution of 10,000 initial runs is selected as the final result, based on the lowest total WCSS (see Section 3.4.2). Two statistics, the WCSS and the BCSS (introduced in Section 3.3.2, STEP 4), are extracted from the cluster result, from which a BCSS/WCSS ratio is calculated (as described in Section 6.4), where a higher ratio represents a better result.

This result is compared to the LSOAC (generated in Chapter 4) to evaluate whether the

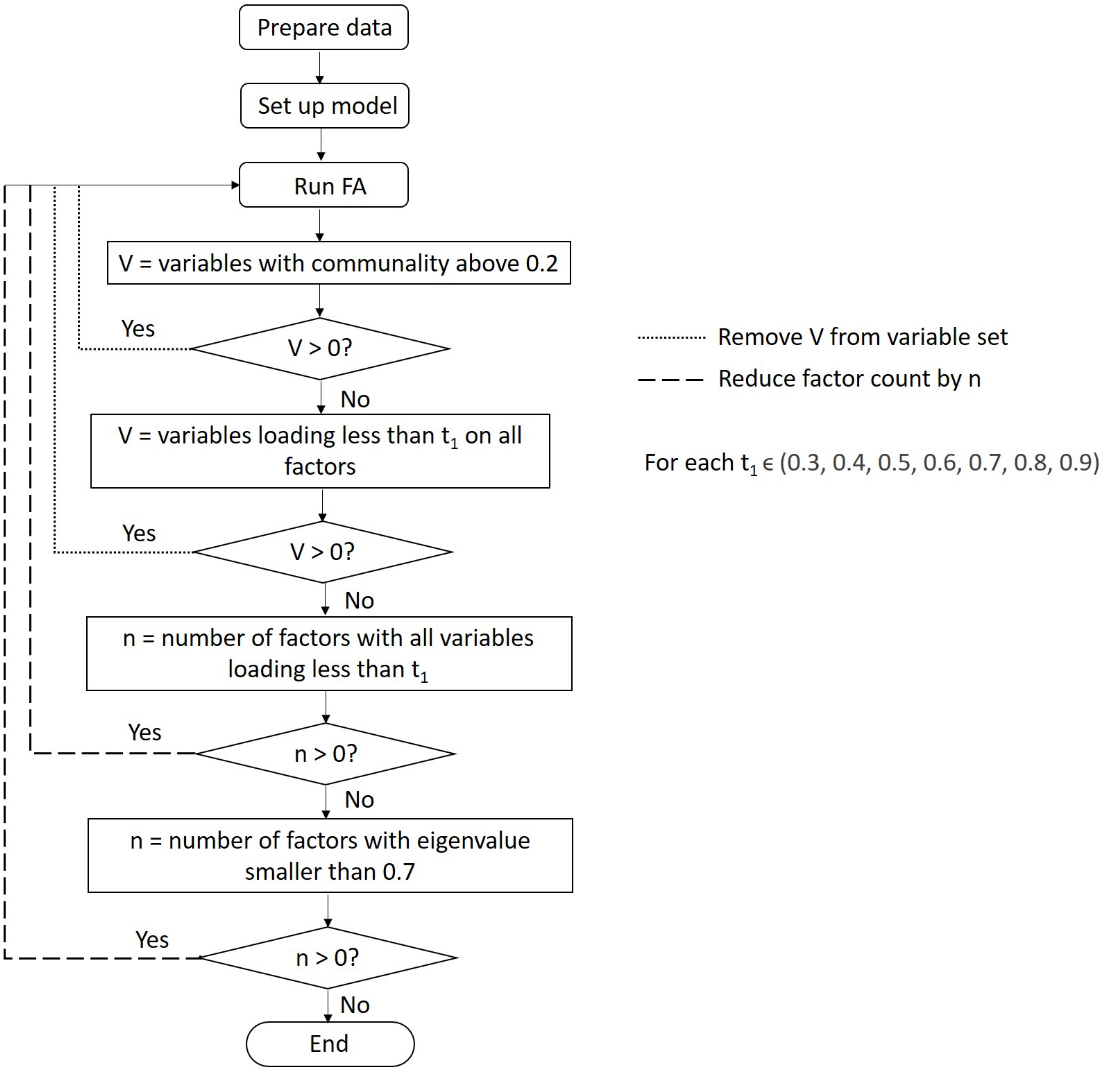

Prepare data
Set up model
Run FA
V = variables with communality above 0.2
Yes
V > 0?
No
V = variables loading less than t_1 on all factors
Yes
V > 0?
No
n = number of factors with all variables loading less than t_1
Yes
n > 0?
No
n = number of factors with eigenvalue smaller than 0.7
Yes
n > 0?
No
End
Remove V from variable set
Reduce factor count by n
For each $t_1 \in$ (0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)

Figure 7.7: Final model for the case-study application of FA, including the parameter selections detailed though Section 7.4.4 (based on the base model in Figure 7.3).

alternative variable selection process proposed has led to a better overall classification of the OAs in Leeds.

7.4.6 Results

Comparing the FA solutions

Loading thresholds 0.8 and 0.9 returned just 1 factor, thus did not result in a usable solution. The result of the FA is therefore five sets of selected variable sets. A summary of the outcomes from the FA for each of the iteration tests based on each of the other loadings thresholds can be seen in Figure 7.5. The total variance explained by each test indicates the "0.7" loading threshold has achieved the best result, explaining 78.5% variance, the only solution explaining more than the target total variance of 75%.

This solution has reduced the 131 candidate variables input into the FA to just 36 variables, 40% fewer variables than the 60 input variables of the 2011 OAC. This reduction indicates that many of the variables adopted in the 2011 OAC might be redundant, or irrelevant, in the analysis of Leeds specifically, and may have been adding noise into the local classification which the place-specific approach using FA for variable selection is able to identify and remove.

Loading threshold	Factor count	Variable count	Total variance explained
0.3	8	120	65.6%
0.4	5	104	65.3%
0.5	5	92	67.8%
0.6	5	59	73.5%
0.7	5	36	78.5%

Table 7.5: Summary of FA result for each iterative test.

Comparing the clustering performances of each test

Table 7.6 presents the WCSS, BCSS and ratio (BCSS/WCSS) of the cluster results relating to each iteration test. The table once again highlights the test with loading threshold “0.7” as the best solution, achieving the maximum ratio between the two cluster measures. Thus, this is selected as the “best” solution which is renamed “FALSOAC” going forwards and compared to the LSOAC result, below.

Loading threshold	WCSS	BCSS	Ratio
0.3	4844.00	4280.62	0.88
0.4	4075.57	4024.47	0.99
0.5	3590.40	3710.56	1.03
0.6	2165.26	2591.93	1.20
0.7	1110.62	1542.83	1.39

Table 7.6: WCSS, BCSS and Ratio (BCSS/WCSS) of cluster results for each set of derived variables.

Comparing the clustering performances of the FALSOAC and LSOAC

Table 7.7 presents the WCSS, BCSS and their ratio for the LSOAC and the FALSOAC. The comparison indicates a higher ratio associated with the new classification, thus indicating a better fit.

As per Section 4.5.1, the SED of the cluster solutions is used to evaluate the ‘fit’ of each of

	LSOAC	**FALSOAC**
WCSS	2450.30	1110.62
BCSS	2098.23	1542.83
Ratio	0.86	1.39

Table 7.7: WCSS, BCSS and Ratio (BCSS/WCSS) of LSOAC and FALSOAC cluster results.

the clusters in the two classifications, illustrated in Figure 7.8. The ordering of the clusters in Figure 4.5.1 is arbitrary, and thus like-for-like comparisons should not be made between individual clusters. However, the mean SED across the two classifications show 0.63 for the FALSOAC and 0.96 for the LSOAC, indicating a better fit in the former. Moreover, overall, almost all clusters in the FALSOAC classification display lower SED than the minimum SED of any cluster derived in the LSOAC.

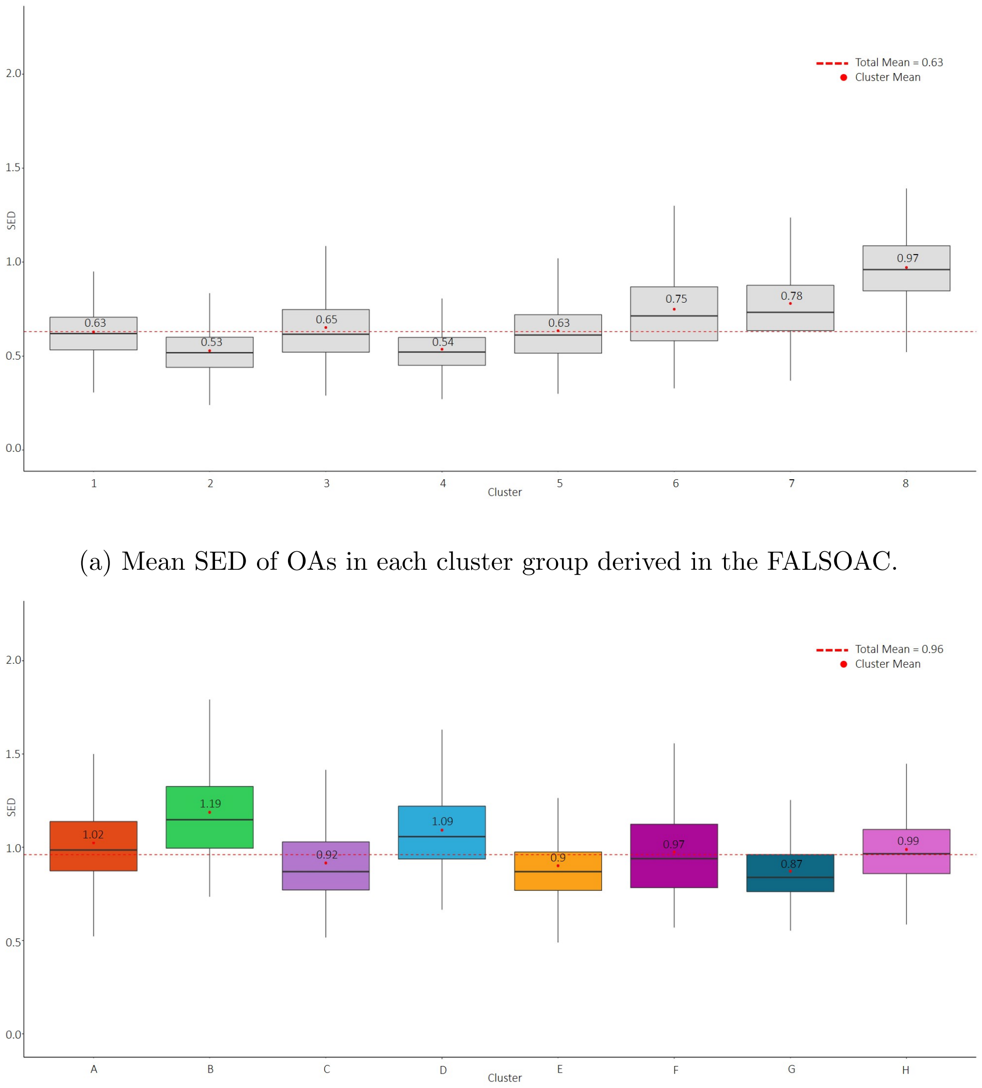
This improvement is further highlighted in Table 7.8, which shows the percentage of OAs in each of the LSOAC clusters which perform better in the FALSOAC. In all but two clusters, almost 100% of the OAs experience an improvement in the fit of the cluster that they are now assigned to. The 2.99% of OAs which do not perform better in the FALSOAC are mainly in the LSOAC “Aspirational young workers” cluster and the “Students” cluster. Again, these are largely the OAs populated with students and recent graduates in the centre of the city and stretching to the North West (see Figure 7.9).

LSOAC cluster	**Total OA count**	**Average SED**	**% of OAs Improved by FALSOAC**
A - Affordable living	347	1.02	99.42%
B - Students	138	1.19	84.06%
C - Urbanites	404	0.92	100.00%
D - High density multicultural families	205	1.09	99.02%
E - Ageing workers	489	0.90	99.59%
F - Aspirational young workers	61	0.98	34.43%
G - Settled, ageing families	613	0.87	98.86%
H - Stable professionals	286	0.99	99.65%

Table 7.8: Summary of LSOAC clusters and OAs with improved SED.

Analysis of final variable selection

A consideration of the final variable selection generated from the FA with loading threshold “0.7” (Table 7.9), which has been used to derive the FALSOAC, offers some explanation of the reduced improvement in the areas of the city largely populated by students.



(a) Mean SED of OAs in each cluster group derived in the FALSOAC.

(b) Mean SED of OAs in each cluster group derived in the LSOAC.

Figure 7.8: Mean SED of OAs in each cluster group derived in the FALSOAC and the LSOAC.

Listing the 36 variables selected, as in Table 7.9, highlights that whilst there remains a broad mix of variables in this updated input variable set, there is a strong emphasis on variables relating to both household structure and ethnicity (and nationality). Several variables which typically represent elderly populations, including the higher-end age variables, a high SIR, widowed, house with single aging occupant, and retired, remain in the selection. This indicates a strong latent structure of an ageing population in the city.

Conversely, the variables directly representative of students are not retained. This is notable in relation to the above findings which indicate a degradation in the classification of the

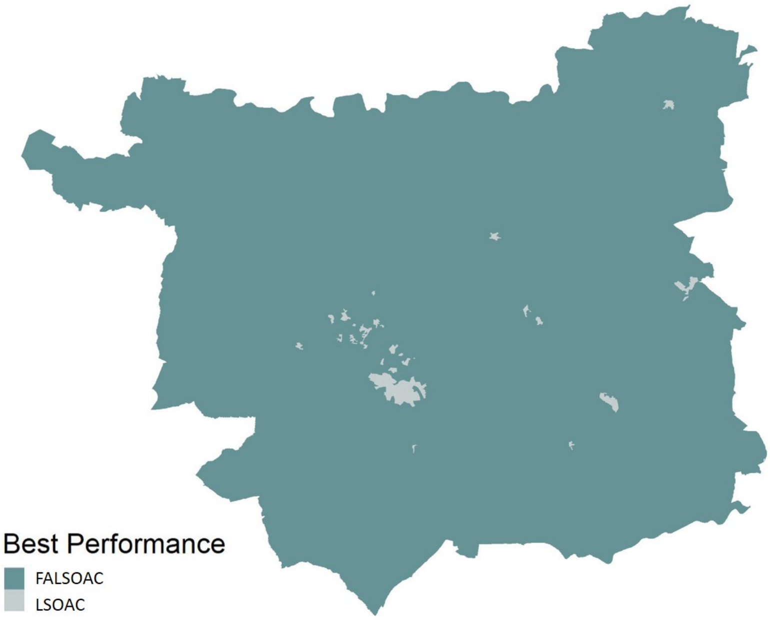


Figure 7.9: Best performance in comparison between LSOAC and FALSOAC.

typically student areas within the city. There is a single variable which remains which is often linked to areas containing high populations of students and recent graduates, "L088: Highest level of qualification: Level 4 qualifications and above". Upon extraction of the factor from the FA solution on which this variable is loading ("Factor 2"), one can see that this variable is loading *negatively* on the factor (Table 7.10). As such, rather than representing a *presence* of individuals with high levels of academic qualifications in the latent construct (represented by the factor), the result indicates the opposite, a *lack* of a such characteristic[2].

In terms of the student related variables discarded throughout the FA process, it seems that the limited number of variables which directly characterise the student populations have not been enough to form a distinct factor with enough variables loading highly enough to adhere to the model criteria and the selected thresholds (outlined in Figure 7.7). This is a potential weakness of employing FA in this way, removing contextual nuance from the decision making, for instance, in considering factors with just two highly loading variables, or in the use of lower thresholds. Nevertheless, since the employment of the latter has led to a weaker final cluster result, and the FA method employed generates a final classification

[2] The results of the FA in full, including similar tables of high loading variables for the other factors, are not presented in this case study since the focus of the research is not on the resulting factors themselves. An exception is made here to investigate a result in the variable selection, which *is* the focus of the research. The factor number (2) has been assigned arbitrarily and is not relevant to the analysis.

Variable code	Variable description
L006	Age 5 to 14
L009	Age 25 to 44
L011	Age 65 to 89
L014	Married or in a registered same-sex civil partnership
L016	Widowed or surviving partner from a same-sex civil partnership
L017	White
L020	Asian/Asian British: Pakistani
L022	Asian/Asian British: Chinese and Other
L026	Other religion
L029	Region of birth: UK/Ireland
L032	Region of birth: Other countries
L033	Main language is English or can speak English very well
L034	Main language is not English: Can speak English well
L035	Main language is not English and cannot speak English well or at all
L036	Living in a couple: Married
L043	Not living in a couple: Widowed or surviving partner from a same-sex civil partnership
L044	One person household: Aged 65 and over
L048	One family only: Married or same-sex civil partnership couple: Dependent children
L051	One family only: Lone parent: Dependent children
L055	No adults in employment in household: With dependent children
L056	No adults in employment in household: No dependent children
L059	Lone parent not in employment
L060	One person ethnic household
L061	All household members have the same ethnic group
L072	Owned and Shared Ownership of home
L076	Occupancy rating (rooms) of +2 or more
L083	Day-to-day activities limited a lot or a little Standardised Illness Ratio (SIR)
L088	Highest level of qualification: Level 4 qualifications and above
L100	Economically inactive: Retired
L103	Economically inactive: Long-term sick or disabled
L107	Part-time work
L108	Full-time work
L116	Work in information and communication or professional, scientific and technical activities industries
L124	Work in professional occupations
L130	Work in process, plant and machine operatives
L131	Work in elementary occupations

Table 7.9: Summary of variables retained by the FA with loadings threshold “0.7”, used to generate the FALSOAC.

with a better fit than the LSOAC in comparison, it seems to be a successful approach. Thus, mathematically at least, the removal of the student variables have improved the outcome.

Variable Code	Variable description	Loading score
L088	Highest level of qualification: Level 4 qualifications and above	-0.83
L055	No adults in employment in household: With dependent children	0.82
L059	Lone parent not in employment	0.82
L051	One family only: Lone parent: Dependent children	0.79
L124	Work in professional occupations	-0.79
L103	Economically inactive: Long-term sick or disabled	0.72
L130	Work in process, plant and machine operatives	0.72
L131	Work in elementary occupations	0.71
L116	Work in information and communication or professional, scientific and technical activities industries	-0.71

Table 7.10: Summary of the variables loading highly on “Factor 2” from the FA solution based on the “0.7” loadings threshold.

That being said, there is certainly a dimension of the population characteristics lost in this final result. However, this result might suggest that the use of a multidimensional approach to identify population characterised easily by a single dimension (student or not) is not necessary, with the approach more useful and applicable in the identification of more complex population structures characterised by a less easily observed mix of composite variables, as is the traditional objective and benefit of deriving geodemographic classifications.

Analysis of the derived classification

The focus of this research is to test the scope for using FA in variable selection for developing better geodemographics, using the method to select an input set of variables from a broader set of candidate variables, where a better fit has been defined mathematically based on an improved WCSS, BCSS, BCSS/WCSS ratio and SED. The focus is not on the development of a classification for its own sake. It is therefore beyond the interest of this research to consider the classification derived in too much detail, particularly in labelling or describing the resulting classes. However, it could be of interest to consider the re-allocation of the OAs between the LSOAC and the FALSOAC, and a comparison of the geographical distributions of each of the classifications, to better understand the spatial impact that a re-classification of the OAs based on an amended variable set could have, and does have here.

The re-allocation of the OAs between the two classifications is represented in Figure 7.10. As expected from the discussions above, the OAs classified in the “Student” and “Aspirational young workers" classes in the LSOAC are impacted, almost entirely re-classified together as per the original 2011 OAC (see Chapter 4).

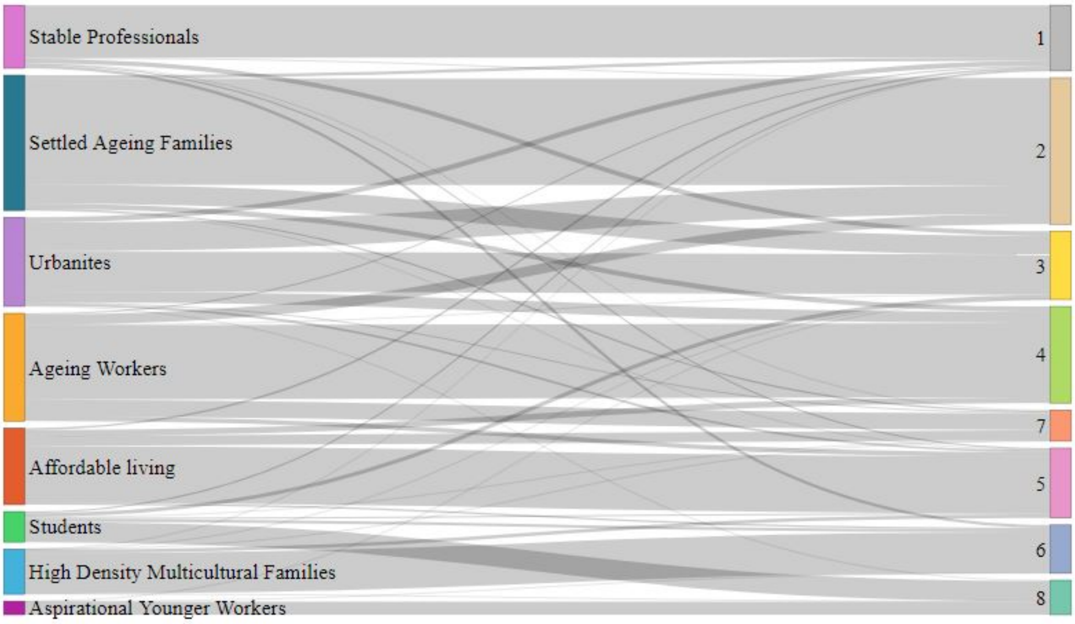


Figure 7.10: OA re-classifications between the LSOAC and FALSOAC.

Conversely, a new classification (represented by FALSOAC "Cluster 7") has emerged from a portion of the OAs classified in both the "Ageing workers" and "Affordable living" classes in the LSOAC. A summary of the most important variables driving this cluster reveals the population of this new class to be predominantly characterised by the elderly variables discussed above, including an aged population, indicators of poor health (long-term sickness, disability and a high SIR), widowers and retirees (Table 7.11).

Variable Code	**Variable description**
L103	Economically inactive: Long-term sick or disabled
L043	Not living in a couple: Widowed or surviving partner from a same-sex civil partnership
L016	Widowed or surviving partner from a same-sex civil partnership
L044	One person household: Aged 65 and over
L060	One person ethnic household
L083	Day-to-day activities limited a lot or a little Standardised Illness Ratio (SIR)
L011	Age 65 to 89
L056	No adults in employment in household: No dependent children
L100	Economically inactive: Retired

Table 7.11: Summary of the variables driving "Cluster 7" in the FALSOAC .

The spatial distribution of the FALSOAC and LSOAC also indicate differences (Figure 7.11). There are similarities, which are to be expected since the bulk of each of the LSOAC classes map to a single FALSOAC class, for the most part. Clear spatial patterns emerge in both instances and, similar to the distributions seen in the LSOAC (and the 2011 OAC, see Chapter 4), there is a clear distinction in the FALSOAC between the city centre and the more rural outer regions of the city. However, the class made up largely of OAs representing "Stable professionals" in the LSOAC is much more geographically constrained in the FALSOAC to the North, extending much further North than in the previous classification, suggesting a stronger geographical influence to this class than was previously identified.

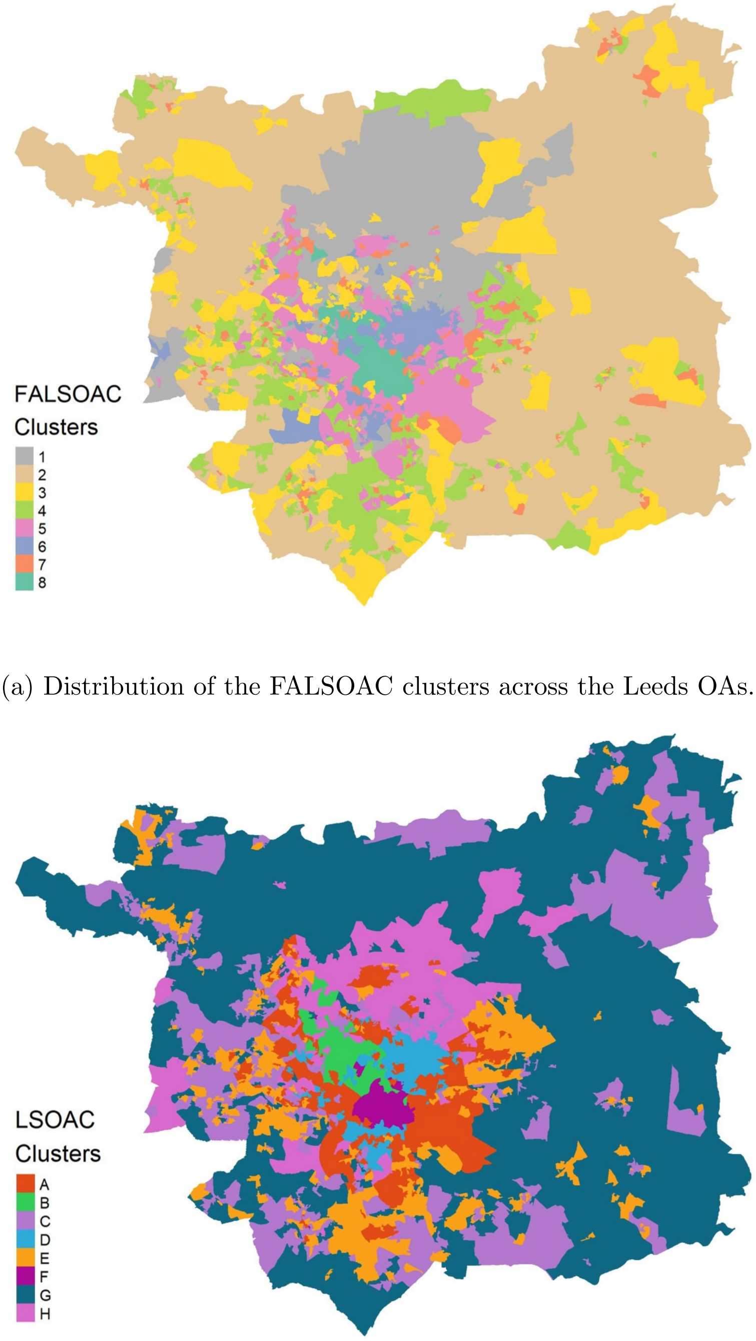
7.5 Discussion

This chapter has re-iterated, and practically demonstrated, many of the challenges which face an application of FA within the variable selection phase of geodemographic classification development. In doing so, it has highlighted many of the weaknesses and challenges to the practice which saw it fall from favour, signalling the end of research in Factorial Ecology.

7.5.1 Making real-world sense of the results

One of these primary challenges has been that there is no guarantee that the use of a statistical process, especially FA, will generate a result which will make real-world sense. However, the importance of generating a real-world meaning from the FA result might be somewhat reduced, so long as the final classification output is meaningfully interpretable. This is the case here, where the FA is used simply to identify which variables to include in the clustering. In this instance, the FA has been successful, since the resulting classification output (FALSOAC) seems to make sense based on knowledge of the city. Moreover, this output even highlights some inputs which were not captured by the comparable LSOAC derived in Chapter 4. The FA selected a set of just 36 variables, compared to the 60 variables adopted in the LSOAC, which were chosen using much less complex variable selection procedures in the development of the 2011 OAC (Gale et al., 2016). This reduction indicates that many of the variables adopted in the 2011 OAC might be redundant, or irrelevant, in the analysis of Leeds specifically.

Additionally, whilst the FALSOAC output highlighted population groups which were largely consistent with the LSOAC output, supporting a confidence in the results of each classifica-



(a) Distribution of the FALSOAC clusters across the Leeds OAs.

(b) Distribution of the LSOAC clusters across the Leeds OAs.

Figure 7.11: Cluster distributions across the Leeds OAs.

tion, there were some discrepancies in the spatial distributions of the comparative groups. For instance, the LSOAC "Stable Professionals" were found to have more in common with the OAs stretching further North in the FALSOAC, generating a new Supergroup in the FALSOAC largely gathered in this area of the city. This new Supergroup ("Cluster 1") exhibits a much strong geographical influence than was exhibited by the Supergroups to which the OAs were assigned in the LSOAC. These results thus demonstrate there is benefit to be gained in not only selecting local variables at a place-specific level, but by the use of more sophisticated methodologies.

However, if there is an alternative aim in the use of FA, to understand the social constructs which underpin the city (as represented by the FA factors), it might be desirable to assign some meaning to the factors identified. This use of FA is not considered here, but might be employed by a developer in seeking to extend the variables considered, by using the FA of an initial candidate variable set to inform the identification of similar, alternative variables which might also be relevant in the geography. This extension might be a useful application as the potential source of candidate variables increase, to offer some guidance for gathering an initial candidate variable set (iteratively), however, as indicated above, this could be hampered by an inability to meaningfully interpret the variables selected in the output of the statistical procedure.

7.5.2 "Generalisability"

A secondary concern which has been highlighted in the use of FA in Urban Analysis in the past is the notion that the results of the FA in this context might not be *generalisable.* These criticisms focus on two core themes. First, research conducted in the era of Factorial Ecology indicated that the results were only appropriate for the city for which they had been derived, and could not be transferred to other cities (Van Arsdol et al., 1958a,b). Secondly, other critics raise the question of whether the results are simply reflective of the input data upon which the results were generated, rather than a reflection of the actual structure of the city, since the data offers just a snapshot of the city from a limited perspective at a single given time (Hunter, 1972; Davies, 1978a).

The generalisability of the result from region-to-region is less of a concern to this study, which is specifically interested in identifying local populations. However, the template presented here enables repeats of the case study for other UK sub-geographies (i.e. other cities) in future work, somewhat relieving this concern. In relation to the data, methods are available to mitigate such concerns. For example, it is possible to alleviate concerns that the results might simply be an artefact of the data by running several FA processes on different data snapshots, for instance, snapshots at different times (Hunter, 1972). This would much easier on administrative data which is being continually collected than on decennial census data, however, a simulation could be generated in a sensitivity analysis using a bootstrapping of the census data, repeatedly selecting subsets of the data and comparing the results generated by each. Although this has not been tested here since it extends beyond the objective of demonstrating the practicality of employing FA in geodemographic classification development, it could be considered in a future extension of this work.

7.5.3 Developer influence in the procedure

Similar concerns often cite the risk of the FA outputs simply reflecting the decisions of the developer, or being a consequence of the model employed, rather than representing real societal structures. Although the exploration of FA was initially intended to offer a variable selection process with less subjectivity than the incumbent methods, the process outlined clearly involves many decisions still to be made by the developer. However, wherever possible, these could be made prior to the execution of the procedure, so as to not be influenced by the results derived. However, an element of subjectivity and decision making remains nonetheless. Again, mitigation methods are available. These include a proposed "multi-method" strategy, similarly iterative as in the bootstrapping method outlined above. In this case, this involves executing several FA procedures based on altered decisions and comparing the results of each to identify consistently emerging factors and gain confidence in the genuine existence of these constructs. Such suggestions were echoed across the Factorial Ecology literature and are still mentioned in more recent technical discussions (Hunter, 1972; Davies, 1978b; Costello and Osborne, 2005). Automating this process to reduce developer input has been explored by (Liu et al., 2019) but it is a difficult approach to manage because of the complexities of the decision making and the specific nuances required to execute FA (Hogarty et al., 2004), decisions would still need to be made upfront by the developer. In this case study, a multi-method approach might have tested the impact of various extraction and rotation methods, or in the use of different threshold parameters. However, the aim here was simply to demonstrate an example of the method to demonstrate that improvements to the fit of the classification could be achieved through the use of FA in the variable selection phase, not to identify the best solution for doing so. Such an alternative objective could underpin future work.

Many of the criticisms raised here are common in the execution of statistical procedures.

Although mitigation methods have been presented, in terms of ensuring that the outcome of an FA is not an artefact of the developer's decisions, or the specific data upon which it has been developed, such concerns may not be possible to ever truly overcome. In processes which rely on input decisions, there is always the chance that such decisions may affect the outcome, however, the target is therefore to ensure that decisions are made carefully, in a justified and informed manner, and that the impact of these decisions are tested, wherever possible.

7.5.4 Attributing meaning to the results

A final concern relates again to the meaningfulness of the classification output produced via the methods demonstrated in this chapter. Although the above discussion concluded that the results of the FALSOAC were meaningful based on local knowledge of the area, this was a rather loose use of the term. Though the method has targeted the strongest mathematical result, and whilst the results reflect many general expectations of the city (discussed in Section 7.5.1), this method has still derived a general-purpose classification. One might question what the meaning of the result is in an applied sense, asking what the classes derived are representative of, whether the classes have distinguished different behaviours or attitudes across the groups, and if so, with respect to what? One might also wonder whether it is likely that the classes derived are predictive of activity in a universal way which could inform all policy decisions or the allocation of resources and services.

These are questions raised by proponents of "Domain Specific" classifications (see Section 3.6.2), who instead suggest that classifications might be more genuinely meaningful when developed with a more targeted objective than the general-purpose classifications which have been derived throughout this thesis. However, the development of a good classification will always be dependent on the initial set of candidate variables being useful and meaningful in the first place. This is even truer of a domain specific application. FA may not be best placed to support such a selection of variables, since the decisions made throughout the FA process are context-agnostic. There is a manual process of selecting the candidate variables which precedes the FA, which could be led by empirical knowledge and context to develop domain specific alternatives. However, it could be interesting to consider whether there is a process which might combine the contextual element into the mathematical variable selection procedure. This underpins the exploration of more sophisticated variable selection processes in Chapter 8.

7.6 Conclusions and next steps

This chapter has considered the use of FA as a method of variable selection within the geodemographic classification development Standard Framework. The complex methodology and often conflicting advice for the implementation of FA has been drawn from the relevant literature into a clear summary, which has been used to develop a framework to support its adoption. A case study application has been presented which has adopted this framework in an update of the LSOAC (derived in Chapter 4) with apparent success. In a comparison of the results, the new classification has, overall, performed better than the original LSOAC, producing an improved fit in many of the OAs. The summary of the process, and the framework derived, extend the existing geodemographics literature. This offers the potential for the case study to be repeated in other cities, or based on alternative candidate input data, to derive similar benefits outside of this thesis. This framework also offers a much more streamline and standard procedure for variable selection, which could be crucial as the quantity of possible input variables grows with the considerable increase in data availability demonstrated in Chapter 5.

However, whilst the classification developed in the case study has derived a mathematically improved solution, the output still has several weaknesses. Notably, the classification again offers a general-purpose result, which fails to address increased scepticism of general-purpose classifications. It could be manually adapted to take as input candidate data a list of variables which have been pre-selected to relate to a specific domain. However, an extension to the exploration contained here might consider whether there are alternative mathematical methods available which might be capable of selecting relevant variables as part of the variable selection procedure itself, in doing so, reducing the reliance on the developer's knowledge and contextual expertise. Chapter 8 will therefore extend this work by considering alternative approaches which might be capable of such a feat.

Summary

This chapter has explored the scope and potential for employing LVMs, particularly FA, to support variable selection in the development of a place-specific geodemographic classification. Early sections outline the theory underpinning the method, before summarising the *complexities of accurately applying the methodology, presenting a comprehensive record of the decisions made throughout the process, and the recommendations to support each, as discussed in the detailed and often contradictory FA literature. This work offers clarity to support the development of a case study application, implementing FA in the selection of variables for a classification for Leeds, before a comparison is made between the resulting classification and the output of the LSOAC developed in Chapter 4. Whilst the results are demonstrably successful, challenges are raised regarding the meaningfulness of the variable selection approach. Consequently, Chapter 8 will look to consider alternative, statistically based variable selection methodologies which might introduce more contextual meaning into the resulting classification.*

Chapter 8 - Next evolution of "Application Specific" classifications

8.1 Introduction

Disillusionment with general purpose geodemographic classifications has been gaining momentum for a number of decades (see Section 3.6.2), encouraging the development of domain specific alternatives, both in the commercial sector and in academia alike. This shift towards developing bespoke classifications with a more precise purpose is being led by a perceived potential for developing more meaningful outputs (Voas and Williamson, 2001). It is also being facilitated by a new era of availability in data and free and open software capable of handling complex statistical procedures(Singleton and Longley, 2009b; Dalton and Thatcher, 2015). Currently in the academic examples, general-purpose and domain specific classification development procedures are typically differentiated by the focus of their input data, with the latter undergoing a more context specific selection process. However, the examples discussed in this chapter reveal that such a selection remains empirical, relying on expert domain knowledge and experience to identify the variables which are contextually relevant.

This chapter seeks to develop a more objective methodology for selecting relevant input variables, specifically, by adopting supervised Feature Selection techniques, offering a new approach to the important step of variable selection which might produce more meaningful classifications for better informing public sector decisions. Section 8.2 presents a brief discussion of some of the applications of general purpose geodemographic classifications, considers how well they work, and outlines the theory and practical methodologies underpinning domain specific classifications in the current landscape. Section 8.3 presents a discussion of the utility and scope for taking a "problem-first" approach to the development of bespoke classifications, adopting supervised Machine Learning (ML) methods as part of the variable selection procedure to inform the development of targeted and bespoke "Application specific" classifications. Section 8.4 presents a case study application of these ideas, developing a classification of the OAs in Leeds specifically focused on differentiating the population based on a propensity for library use. The results are compared to the FALSOAC derived in Chapter 7 to evaluate the ability of the new methodology used to better differentiate library use in the population.

8.2 Background

The focus of this thesis is on evolving the process for developing geodemographic classifications and has therefore, thus far, included limited discussion of their applications beyond referencing example domain areas. However, Chapter 7 has made recommendations for including more contextual consideration in updated variable selection processes to support the generation of more meaningful results. Such an endeavour would necessitate a consideration of the end-user, and the likely applications to which any classification is to be applied, in order to surmise what a "meaningful" classification might look like. This section will therefore look a little more closely at some common applications of geodemographic classifications. However it is worth noting that, again, the focus here is on understanding applications of geodemographic classifications from the perspective of gaining insights to continue to improve their development, and thus this objective will dictate which research is discussed here, and the focus of the insights that are drawn from the research discussed.

8.2.1 Common application methods

Traditionally, geodemographic classifications are derived as general-purpose, standalone descriptors of small-area geographies (Voas and Williamson, 2001). Applications of these general-purpose classifications typically adopt one of two main approaches.

The first takes the classification of each area as a base profile for that area, to which ancillary data is appended to develop a richer profile based on a broader set of characteristics (Singleton, 2010). This approach was first demonstrated in the early days of the commercial classification industry, where it continues to be widely employed to produce comprehensive pen portraits of each classification group based on lifestyle and consumption behaviours. This is supported by the assumption that such characteristics are likely to be shared amongst people living in the same area (Moon et al., 2019), an adoption of Tobler's first law of geography (discussed in Chapter 2). In commercially focused discussions of classifications the term "birds of a feather flock together" is repeatedly echoed to analogise these sentiments (Harris et al., 2005). It has also been subsequently adopted within academia and the public sector (Longley, 2005), where domain focused profiles are often derived in such a way. For example, health indicators can be combined with geodemographic classifications to generate mappable health profiles, a useful resource providing indirect insight into spatial patterns of public phenomena such as health behaviours and associated outcomes (Samarasundera et al., 2010).

The second approach is to use geodemographic classifications as independent variables in subsequent analysis. In this sense it can aid in modelling and predicting outcomes. The classification is typically appended to an outcome indicator in an ancillary dataset to split the ancillary dataset by the classification groupings, or included as an independent variable in a prediction model alongside other variables to derive an association between the classification groups and a propensity for the outcome in each of the groups. Extensions have also seen the classifications included within, or in combination with, other spatial interaction techniques specifically (Singleton et al., 2012; Birkin and Clarke, 2012). In this way, the classification can be used as a proxy for the outcome in areas where actual data representing the outcome does not exist (Moon et al., 2019), providing additional intelligence and even substituting missing context (Williamson et al., 2006) to support specific decision making processes. Examples of this approach include the use of geodemographic classifications in targeted marketing campaigns or public sector resource and service allocation. This approach has been adopted in public sector focused research for several decades, seeking to illustrate and understand propensities for social phenomena across a geography and to identify areas of greatest need (Batey and Brown, 2007). Examples include spatial analysis of crime (Williamson et al., 2006; Ashby and Longley, 2005), poor health outcomes or disease risk (Moon et al., 2019; Powell et al., 2007; Farr and Evans, 2005; Aveyard et al., 2002), public transport use (Liu and Cheng, 2018), road traffic collisions (Anderson, 2010), participation in Higher Education and educational attainment (Batey and Brown, 2007; Singleton, 2010), fire safety (Corcoran et al., 2013) and the use of public sector services (Batey and Brown, 2007; Samarasundera et al., 2010).

8.2.2 Do applications of geodemographic classifications “work”?

The early chapters of this thesis, and many of the results presented throughout, indicate that the concept of geodemographic classifications holds a great deal of potential. This has warranted the in-depth focus granted by this thesis and the efforts made to continue their use. However, this thesis has sought to pursue improvements to the standard methodology employed in the development of such classifications, and has highlighted several weaknesses which could potentially undermine their accuracy, and ultimately their usefulness (particularly noted in Chapter 3). Chapter 3 also introduced the difficulties present in evaluating the success of geodemographic classifications and in conclusively determining whether they "work".

As noted, the lucrative nature of the commercial Geodemographics Industry suggests a clear measure of their success. However, Harris et al. (2005) notes that the extent to which the commercial classifications "work", and in which contexts they perform best, is undocumented. Particularly, the authors remark that publications of the specifics of any successful applications, or more crucially, any less successful applications, are understandably limited due to commercial sensitivities. The competitive nature of the industry and the desire to increase license sales restrict such transparency.

Academic and public sector research is not bound by such concerns, and conversely, good practice encourages transparent and open research. Therefore, several studies have sought to test the applicability of existing classifications within specific applications, and have openly reported their results (Harris et al., 2005; Ashby and Longley, 2005; Harris et al., 2007; Brunsdon et al., 2011; Moon et al., 2019). In each instance, benefits have regularly been achieved through the use of geodemographic classifications, which have offered encouragement and supported their continued application. However, in each case these are offset by apprehensions indicating that the classifications employed might not have provided the definitive 'best' solution. This has repeatedly raised the question of whether better results might have been achieved through the use of improved classifications, or even a different approach altogether. Thus, the question of whether classifications "work" still remains, alongside a desire to continue to seek improved alternatives to the available offerings.

Addressing this question more generally, Longley and Singleton (2009a, p.761) suggest that it is unlikely that it will ever be answered "unequivocally" and "universally". Moreover, they claim that "few generalisations about socio-economic distributions are founded on incontestable facts" and that, in practice, issues of trust in the results are key. As such, efforts to seek such might be misguided, and time might be better spent validating and developing trust in the insights that are derived to understand both the potential and limitations of geodemographics, rather than seeking a perfect solution that objectively "works".

8.2.3 Debated meaningfulness of general-purpose classifications

Trust (highlighted as crucial in the previous section) has been particularly reducing with regards to the relevance of general-purpose classifications. Notably, many of the discussions repeatedly circle back to the applicability and meaningfulness of the classification in the domain or application to which it is being employed. For example, despite gaining useful intelligence to support police activity in an application of MOSAIC, Ashby and Longley (2005) raise concerns regarding the high level of heterogeneity within classification groups, and the potential implications of such. Similarly, Moon et al. (2019) ponder whether their positive results would be maintained in an extension to alternative geographies or domains. Harris et al. (2007) are more critical of the use of *general-purpose* classifications specifically. Despite conceding that the use of geodemographic classifications have had a proven value in such decisions, the authors question the statistical robustness of evidence presented in their favour and the successes accredited to them. The ability of general-purpose classifications to discern social patterns is debated in this research which asks whether their use offers the necessary nuance required to underpin important policy decision and action.

These discussions have instigated an important methodological debate. Debenham (2002) demonstrates an approach which employs regression to derive a measure of the contribution of each variable to the segmentation of the data into cluster groups, facilitating an identification of the variables which drive the clustering. However, this approach still fails to attribute any contextual importance to these variables prior to the clustering. Though input variables can mathematically drive a clustering based on their relationship with other input variables, and can thus be perceived as important indicators, these variables might have no relationship with the outcome phenomena to which they are subsequently appended, in an application of the classification. In such an instance, the variables used have not generated a meaningful clustering output which is relevant in determining any insights about the outcome phenomena (Maugis et al., 2009). All applications of general-purpose classifications are predicated by an assumption that such a relationship exists, but there is no guarantee. Thus, if the cluster outcome is nevertheless used in such a way, the potential for generating misleading insights could be introduced.

Both approaches adopted in the application of geodemographics (detailed in Section 8.2.1) widely rely on a technique of indexing the ancillary data or outcome across the groups. This process compares the local mean to the global mean to conclude that the characteristic is more likely in one group than another, to add the characteristic to the profile of the groups (approach 1), or to infer that there is a higher or lower than average propensity for the outcome (approach 2). However, this inferred likelihood is therefore relative. To infer that there is a higher than average propensity for an outcome in one area than another cannot demonstrate that there is actually a high likelihood of that outcome, just that it is higher than average. For example, in a city widely populated by individuals on low income, if one area has a slightly higher income than the others, it will have a higher income than average, yet still might be a low income in absolute terms. There is therefore a concern that if there does not exist a real relationship between the allocation of the classification groups and the outcome, then any such inference will be incorrectly drawn.

There is thus no reason why geodemographic classifications which are derived from input variables which have been selected *without* a specific purpose should generate universally meaningful clusters, or are necessarily discriminative in a way which can differentiate the population to reflect the spatial patterns of any given phenomena (Harris et al., 2005; Longley, 2007). As such, there is debate as to whether it makes sense that general-purpose classifications are transferable to all situations, and will offer meaningful insights in each circumstance (Singleton and Longley, 2009a). To this end, Longley (2005) mused whether some classifications might be more reliable in some domains, and in evaluating particular behaviours over others. However, he noted that the limited investigations available in the literature restrict the definitive drawing of such a conclusion. Brunsdon et al.'s (2011) study made attempts to compare the performance of geodemographic classifications in predicting participation in Higher Education specifically. This study concluded in an endorsement of employing classifications, but alluded to there being further benefits which could be gained from using more appropriately derived classifications, suggesting that a more reliable approach might be to shift to the development and use of "domain specific" classifications. These ideas were not novel. Specifically, Voas and Williamson (2001) asserted a decade earlier that using general-purpose classifications rather than classifications developed for a specific purpose might result in an inferior solution.

Possibly of more concern are the cases in which classifications derived with a specific purpose in mind are used as general-purpose classifications and applied in other domains. This is an accusation being increasingly made against many of the commercial classifications. Many of these classifications have been derived with an underlying focus on indicators of income (see Chapter 3). These are of importance when segmenting consumers, as is useful in the development of classifications intended to support applications in marketing, but are arguably less so when differentiating populations on demand for public services and resources. These concerns underpin a lively debate regarding the transferability of proprietary commercial classifications to address public sector problems (Singleton and Longley, 2009a; Webber, 2004). From this debate, the consensus has been to recommend caution when applying any geodemographic classification without complete clarity of the initial purpose for which it was developed (Twigg and Moon, 2002; Samarasundera et al., 2010).

The potential negative impact of using classifications in the public sector which were not designed for the specific purpose in which they are being used, or worse, were designed purposefully with another intention in mind, could therefore present a risk which should not be underestimated (Singleton, 2010). The stakes associated with analysis in the public sector, which could have real-world consequences in the lives of the population, require confidence in decisions made and actions taken. Particularly, the application of inappropriate classifications increases the potential for the development of *inefficient* or *incomplete* policy initiatives (targeting those who should not be targeted, or alternatively missing those who should). Though there is some inherent inevitability of both in the use of geodemographics derived at a small-area geography to target households or individuals (Batey and Brown, 2007), any effort that can be made to mitigate the magnitude of such an occurrence should be made.

8.2.4 "Domain Specific" geodemographic classifications: Methods, benefits and limitations

The discussions in the previous section underpin a trend in the development of geodemographic classifications which has seen a slow evolution from a broad reliance on general-purpose classifications towards a promotion and subsequent acceptance of bespoke, *domain specific* alternatives (Brunsdon et al., 2011). This has particularly picked up pace since technical advancements and data availability are such that there is no longer a need to rely on a small number of general purpose classifications (Longley and Singleton, 2009a). In the development of such alternatives, classifications are derived to be more discriminant with regards to a specific context or domain, and are thus more likely to reflect underlying dimensions of relevance to the domain of interest. This is typically achieved through the inclusion of a more context specific, or more "relevant" selection of input variables (Singleton and Longley, 2009a). These are selected for their importance as indicators specifically of the domain (Beaumont and Inglis, 1989).

Several of the main commercial geodemographic classification vendors offer bespoke, domain specific classifications. These include the "Public Sector" alternatives of their classification products. Some also offer separate classifications for other sectors, as detailed in Section 3.2. The availability of such products demonstrate an acknowledgement of the concerns raised in the previous section regarding the perceived consumer focus of their original offerings. However, it is still not possible to see the input data underpinning their development, to evaluate how they differ from the general-purpose alternatives, or to understand how the variables have been specifically selected for input. Consequently, questions regarding their applicability remain.

Several examples have also been developed in the open and academic geodemographics landscape in the UK, including the classification of residential geographies of workplace commuters (Hincks et al., 2018), classifications of digital behaviours (Longley et al., 2008; Longley and Singleton, 2009b) and an "educationally weighted classification" (Singleton and Longley, 2009a) designed to encourage the identification of classes which more appropriately discriminate on propensity to participate in Higher Education. Alternative domain specific classifications have also been developed to classify non-residential small-area geographies, including Singleton and Longley's (2019) classification of London workplaces, which offers an extension of the national Classification of Workplace Zones (COWZ) (ONS, 2018a).

Such bespoke classifications are less reliant upon implied inferences, which general-purpose classifications suffer from in their employment of more arbitrary variable sets (Longley and Singleton, 2009b). However, there is little evidence of the use of sophisticated variable selection procedures in the identification of important and relevant variables, beyond a subjective selection of variables based on a review of domain specific theory. In the development of their Higher Education classification, Singleton and Longley (2009a) supplemented the correlation and sensitivity analysis traditionally adopted in the variable selection phase of the 2011 OAC development (Hincks et al., 2018; Singleton and Longley, 2019) with analysis of associations between variables and the domain outcome (Higher Education participation) and discussions with stakeholders. Whilst these approaches offer some justification for the inclusion, or exclusion, of particular variables, they do not constitute sophisticated methods for measuring the importance of the variables in the desired context, to facilitate the selection of only the important variables.

Moreover, the academic examples listed above relate to very specific contexts. Conversely, the commercial sector "Public Sector" classifications are a far broader example. Although these are still more targeted than the general-purpose alternatives, the decision of the academic researchers to be even more targeted in their approach suggests that an overarching public sector offering might still not be capturing the necessary nuance. The variety of public sector issues are extremely broad, and the indicators which drive the consumption of public sector services and resources can be equally variable across different contexts (Singleton, 2014). In respect of this, the commercial sector has already developed some more nuanced classifications, such as CACI's "Health" version of Acorn (referenced in Section 3.2). Similarly, benefits might be gained from the consideration of additional specific public sector classifications targeting more focused domains, such as Adult Social Care, or the use of individual public services, e.g. libraries or recreational sports facilities, to support service and resource allocation, priorities of LCC (Otley et al., 2018; Smart Leeds, 2021). However, the development of meaningful and relevant sub-domain level, or even application specific, classifications developed via the existing methodology would be incumbent on the developers possessing or gaining a level of expertise in each sub-domain, or application, which is good enough to support an appropriately informed selection of the important input variables in each case. Such an expectation might not be completely realistic, or at the very least, could add further complication or cost additional development time.

8.3 The next evolution of geodemographic classifications

8.3.1 A proposal for Machine Learning led "Application Specific" geodemographic classifications

The core purpose for the development of Singleton and Longley's (2009a) Higher Education classification was to derive groups which better discriminate between areas characterised by extremely low and extremely high Higher Education participation rates. Although this has been categorised as a “domain specific” example of geodemographic classifications in other educational contexts, one might instead consider it an application specific classification. Whilst there is scope to re-use the output classification, the specific purpose of its development was on understanding the patterns of spatial distributions in a single *observable outcome*, in this case, participation in Higher Education.

It is therefore not a huge leap to extend this example, and the thinking behind it, to consider a more routine shift from “domain specific” and towards “application specific” classifications, as discussed in the previous section. However, as highlighted, more targeted classifications require a more targeted variable selection process, extending beyond the common practices which have to date been largely reliant on the expert knowledge of the classification developer and their team. To support a shift towards the routine development of application specific classifications, a more sustainable, intelligence led approach would be beneficial.

The core objective of this chapter is to propose and explore the benefits of a re-framing of the thinking behind the development of targeted classifications by considering the application outcome (e.g. participation in Higher Education, or any other observable outcome such as the propensity for consuming specific public sector resources or services) upfront in the development of a related classification. Such a shift might thus generate a classification output which more intentionally differentiates on the observable outcome of interest, and in turn, flip the role of geodemographic applications from a ‘one-size-fits-all’ approach applied as a solution to all problems, to take a problem centred approach in the development of a problem-focused solution. Thus the aim is not to derive a set of conclusive labels, or define areas definitively based on a single pen portrait of the average characteristics representing the area, but to identify areas to target in specific applied scenarios.

8.3.2 The potential of Machine Learning variable selection methods in geodemographic classification development

In practice, the proposal is, instead of appending a geodemographic classification to an *observable outcome* as has been a traditional approach to the use of geodemographics, to consider the outcome of interest as a dependent variable, and to employ it in the selection of input variables for the classification. Such a shift would support an intelligent selection of variables which are identified as important indicators directly relating to the observable outcome itself, and which might hold a greater potential for identifying more relevant cluster groups. This approach is theoretically similar to the current practice, however, instead of the existing reliance on domain knowledge to support a subjective selection, the manual selection would be replaced by an application of a more sophisticated supervised Machine Learning (ML) method.

The boundaries of which practices are considered to be "Machine Learning", and thus the definition of the term itself, can raise lively debate. However, the use of the term in this thesis is simply as an umbrella term to refer to any process which can *learn* without being explicitly programmed. The importance and scope of this definition in itself is of limited significance, since this chapter will focus on explicit examples, which will be clearly defined. However, the definition of *unsupervised* and *supervised* ML methods warrant a little more attention. Supervised ML methods (as introduced in Chapter 1) refer to such processes which, given an input, are trained to generate outputs based on patterns learned from observations of known input and output pairs. This process can enable predictions of an outcome of interest to be made from a input dataset, based on previously observed outcomes and their relationships with variables in the input data. Conversely, unsupervised ML methods have no such a-priori knowledge and thus self-learn based on the input to generate an output.

The use of ML methods are not entirely new in geodemographics. K-means classification has as already been highlighted in this thesis, described as an unsupervised ML method, since it identifies clusters in the data without a-priori knowledge of the clusters or cluster structures. Singleton et al. (2020) have also employed supervised ML techniques for predicting missing input data. However, no examples have been identified in the use of supervised ML methods as a method of variable selection.

To derive more discriminant geodemographic classifications, the introduction of supervised ML techniques into the development process could be useful (Guyon and Elisseeff, 2003; Gregorutti et al., 2016). The previous chapter explored the use of the FA (an unsupervised Feature Extraction ML method) to seek to introduce a subjective element to the variable selection. However, this approach did not consider the variables in terms of their contextual significance. Whilst this was noted as a limitation, it is an even more fundamental omission when applied in the development of a non-general-purpose classification, where the context of the variables holds an increased significance. Consequently, it is necessary to seek a methodology which prioritises the selection of variables which are contextually meaningful.

In the present era of using ML methods for drawing insights from "big data", this chapter seeks inspiration from other fields beyond geodemographic classification development. In other scenarios which apply predictive methods, the observable outcome provides the starting point, data permitting, and insights are derived from these observations to underpin the prediction. The developments which have occurred in multivariate statistics, particularly in ML techniques, combined with advancements in software and computational capacity which have been made since the last big evolution of geodemographic classification development have opened up new scope for adopting increasingly twenty-first century approaches, seeking to adopt popular contemporary methods, as has been the tradition throughout the history of the field (see Chapter 2). Longley (2005) outlined the desires of the public sector to develop more advanced techniques for anticipating and targeting public service demand, particularly at a local level. This approach is seeking to better support this activity.

Though there are unsupervised variable selection methods available which attempt to imitate the abilities of supervised methods (Xing and Karp, 2001; Karegowda et al., 2010), in a setting where the subsequent use of the classification is known, i.e. where the developer is in a position to build a classification with a specific application in mind, more favourable supervised methods are supported. Thus, this chapter will therefore consider a supervised approach.

8.3.3 Introduction to Feature Selection methods

Supervised ML methods in variable selection offer the potential to derive a more targeted approach. Taking an observable outcome as a dependent input, the selection of variables can be made by prioritising variables which have the strongest associations with the observable outcome. This presents an opportunity to incorporate domain knowledge and generate a final variable set which is extremely application specific (Guyon and Elisseeff, 2003). In ML, such variable selection techniques are often categorised as methods of Feature Selection (FS), where "feature" is simply a synonym for variable.

Unlike the *unsupervised* Feature Extraction (FE) methods considered in Chapter 7, FS methods can employ either supervised or unsupervised ML techniques (Karegowda et al., 2010). Moreover, whilst both FE and FS methods seek to reduce the number of variables from an initial candidate set, FE traditionally achieves such a feat by forming new variables from the original variable set (Brunsdon et al., 2018), whereas FS methods are designed to include or exclude “important” or “unimportant” variables which are left unchanged, retaining the original meaning of the variables. The adoption of Factor Analysis (FA) in Chapter 7 (a FE technique) did not ultimately use the new variables (*factors*) as new variables in the subsequent analysis, as is traditional. Instead it focused on reducing the set of input variables until an optimised FA model was achieved, and then took the remaining variables as the final variable set (as per a similar technique adopted by Liu et al. (2019) in their application of Principal Components Analysis (PCA), another FE method). However, criticism can still be levelled at the approach in terms of the contextual relevance of the remaining variables (discussed in Section 7.5). At no point did the procedure consider the contextual relevance of the variables included or excluded.

Conversely, methods designed for FS seek to identify a relevant variable set which simultaneously reduces redundancy (Kohavi and John, 1997). Since FS is regularly adopted as a preliminary technique in prediction analysis, the contextual relevance of the variable set selected is a key priority. In prediction analysis, the aim is to select the least number of variables which are important to an outcome to enable the development of a model which will have the greatest prediction accuracy, thus, the main aim of FS is to select as few variables which are important drivers of the observable outcome as possible, whilst introducing parsimony (Guyon and Elisseeff, 2003).

These key objectives mirror the core aims of the variable selection phase in most geodemographic classification development. FS therefore might also offer a new approach to the variable selection phase in application specific geodemographics, where the intention is to derive a classification which generates groupings linked to propensity of an observable outcome. In such a case, rather than selecting variables to underpin the development of a subsequent prediction model, the variables selected could be used to underpin a subsequent clustering to derive bespoke, application relevant classifications. These could demonstrate combined geographic distributions of relationships between the input variables which have a relationship with the outcome. Though no evidence of such an approach has been identified in the literature for deriving geodemographic classifications, especially with respect to local-level application specific classifications, the use of FS for clustering has been adopted in other sectors, notably in medical domains (Karegowda et al., 2010).

FS is an established yet evolving field with many techniques. Whilst it is beyond the scope of this thesis to discuss the methods in deep mathematical detail, the three main categories of FS methods; *filter*, *embedded* and *wrapper* methods (discussed in detail by Guyon and Elisseeff (2003)) can be summarised as follows.

Filter methods are developed based on a specified performance metric, for example, correlation or chi-square (Karegowda et al., 2010). Thus, correlation analysis, which often underpins variable selection procedures in the development of geodemographic classifications (including in the 2011 OAC, see Chapter 3), is itself a foundational FS technique. However, filter methods do not employ a learning method, simply evaluating the importance of variables based on their inherent characteristics. This chapter seeks to identify more advanced methods.

Embedded methods employs algorithms with built-in FS methods, to perform a FS within the execution of a broader objective, for example, as a part of a classification or regression model (Guyon and Elisseeff, 2003). Popular methods include LASSO linear regression and Decision Trees (Gregorutti et al., 2016). Since these are not stand-alone FS methods, as are sought here to act as a basis for the variable selection phase within the broader geodemographic classification framework, embedded methods are not considered further here.

Finally, *wrapper methods* consider the selection of features as a search problem, treating the features as the inputs and seeking to optimise model performance. These methods can either use *forward selection*, *backwards elimination* or a *bidirectional search.* Forward selection starts with an empty candidate set and adds variables in one-by-one if they are deemed important. Conversely, backwards elimination starts with the whole candidate set, removing the least important variable one at a time. A bidirectional search offers a combination of the two (Guyon and Elisseeff, 2003). Each are thus iterative methods which build many models based on the observed outcome (as a dependent variable), each with different subsets of candidate features, to identify the attributes which best support the development of an accurate model to predict the observed outcome. In doing so, these approaches evaluate the importance of each of the candidate variables with respect to the observed outcome of interest.

When used for prediction, the target of wrapper methods is to identify a variable set which maximises the accuracy of a final model used to predict the dependent variable, the observable outcome of interest. Such applications have been shown to improve predictor performance beyond the capabilities of simpler methods, such as correlation methods (Karegowda et al., 2010; Gregorutti et al., 2016), the incumbent favoured method of variable selection in geodemographic classification development. An alternative use is to identify a variable set which maximises the relevance of the variables selected, with relation to the observable outcome (Kohavi and John, 1997). Such an approach might offer the potential for selecting a variable set for application specific geodemographic classification which might result in more homogeneity in the final cluster groupings with respect to this outcome.

8.3.4 Summary and next steps

These discussions have presented justifications for exploring the use of FS methods to support variable selection within the development of application specific geodemographic classifications. The intention of such an approach would be to increase their predictive power, differentiating areas in a way which is more appropriate to the application at hand. The case study below (Section 8.4) demonstrates the potential of these proposals in practice and illustrates the scope for this theory to generate improvements in the homogeneity of the resulting classification groups with respect to the application of interest.

The end objective is to develop increased spatial intelligence related to an observable outcome. This will necessitate not just the development of a single new classification, but a framework which will support the development of problem bespoke geodemographic classifications. This will thus not be a framework for generating a traditional general-purpose, nor will it be for developing increasingly popular domain specific classifications, rather, it will support the on-the-fly development of more targeted application specific classifications. This could encourage the development of classifications based only on variables of relevance, thus reducing the chance of missing appropriate and relevant variables and including unnecessary and unhelpful variables. Moreover, it could also introduce a new approach to the use of classifications in public sector development, shifting the reliance from static one-size-fits all products towards a culture of developing potentially more relevant situational alternatives with which to inform specific decision making processes. Moreover, as LCC improve their storage and use of their database systems and their data, this framework would also support the routine development of classifications based on new, more timely candidate data.

The next section presents a demonstrable example of these ideas in a case study focusing on the propensity for using public libraries, illustrating the practical scope and offering a methodological template for such a development.

8.4 Case study application: Employing Feature Selection methods for variable selection

8.4.1 Introduction and context

Libraries are recognised as a vital community resource offering a range of important services (Department for Digital, Culture, Media and Sport, 2018). However, as local government budgets decrease, libraries have increasingly been at risk of closure. This has led to unsuccessful petitions to ringfence UK government funding for libraries and protect library services, an action that was rejected by central government. In response, the government reinforced that it is the responsibility of local governments to make spending decisions relating to public libraries based on local need (UK Government and Parliament, 2018). Additionally, as a result of the subsequent coronovirus pandemic and the further budget strains induced, all Leeds libraries are now at imminent risk of closure (Drury, 2020). However, Roumpani et al. (2020) highlight a fragmentation in the public library sector which has led to poor national data availability, resulting in a lack of intelligence regarding the provision of library services in the public libraries sector, and in its value in different communities. This is substantially hampering evidence based decision making capabilities at a time where there is both a need and desire to gain a better understanding of this provision and to use these insights to advise future planning. Moreover, Roumpani et al.'s (2020) research particularly points to understanding the differentiation of library use by different societal groups as a key priority area.

Despite the national level data issues, LCC hold local level library use data which could offer some helpful local insights. As such, an opportunity is presented to test the potential for developing a bespoke, application and place-specific geodemographic classification with the intention of better understanding the spatial distribution of propensity for library use in the city, which could generate valuable future insights for LCC. This section thus presents a case study focusing on this data to develop such a classification.

Once again, the classification developed here is Leeds-specific, alleviating the issues of national data availability. This case study incorporates data relating to library use in the city, employing FS methods in the variable selection phase of the classification development, testing the potential for deriving application-relevant classification groups with increased within-group homogeneity. In this case, these tests are executed with respect to library activity, which could be used by LCC to generate more informed decisions regarding library services and resources. In doing so, this case study also offers a template to facilitate more routine development of local, application specific classifications for use in the public sector more broadly in the future.

8.4.2 Data

Overview

Data relating to library use has been sourced from contacts at LCC, as per previous examples in Chapter 5. From this, the *dependent variable* for use in the FS method is derived. The independent, *predictor variables* from which the FS is to identify the important drivers of propensity for library use are drawn from an initial candidate set of census variables.

Library data

A snapshot of data relating to loans made from the libraries which fall within the Leeds LA boundary has been supplied by LCC. This data contains a record of any loans taken out across the 33 LCC run libraries between January 2017 and February 2018, inclusive. This includes loans of books, predominantly, but also CDs, DVDs, pamphlets, periodicals, sheet music, maps and some other miscellaneous items. Each loan of an item is recorded as a single record. Each record contains a timestamp of the loan, the library from which the loan took place, a “borrower type” containing a standard description of the borrower, typically “adult”, “child” or “pensioner”, a full postcode, OA, and year of birth of the borrower, and some information relating to the item loaned, including the title, author and genre. It is understood that the pre-assigned OA included in the data has been allocated as per the internal LCC postcode to OA lookup (see Section 6.3.4). This location information enables an analysis of the loans with relation to the residence of the borrower themselves.

Census data

For consistency, the 131 candidate census variables adopted in Chapter 7 are once again used here as the candidate predictor variables. The FS is applied to these predictors to identify the combined attributes in the candidate set which seemingly drive library use, and can thus best predict propensity for it. The use of this data supports a comparison of the classification derived here, where a novel FS method is instead applied to the variables in the variable selection phase of its development, with the classification derived in Chapter 7, where a FA was applied. This variable set has already undergone an initial filtering process and been transformed in Section 7.4.4 as per the transformation processes adopted in the development of the 2011 OAC (Gale et al., 2016). A complete list of these 131 variables can be found in Appendix C.2.

Data preparation

The library data is filtered to remove any records relating to borrowers living outside of the 2,543 Leeds OAs, and any records relating to loans made by organisations and not individuals, including community groups and schools, based on a review of the "borrower type". Records relating to book renewals are also removed under the advice of experts within LCC, since these records are identified as containing "spurious" information. Finally, the data is checked for duplicate records, which are also removed, where identified. These procedures discarded a total of 58% of the records which has been initially supplied.

Although it would be desirable from an analytical perspective, it is not possible to identify individuals in the data, based on the information available. Since the limited personal information is restricted to just a year of birth and a postcode, it is possible for multiple individuals to share matching identifying information. Nevertheless, it is possible to infer distinct *visits* from the data, grouping the records by timestamp, loan library, and the personal identifiers listed. Whilst it is theoretically possible that individuals sharing these personal identifiers *could* have made loans at the same library at the exact same time, this is deemed highly unlikely, particularly since the timestamp is recorded to the second. Thus, a record of unique visits is derived from the raw data, including a count of the items loaned at each visit.

There are some limitations to this methodology. For example, visits to libraries which did not result in the loan of an item, i.e. visits spent browsing or for any other purpose such as computer use, are not contained in the raw data, and as such, are not recorded as visits in the visit counts derived. Additionally, if multiple loans are made within the same visit but are processed in individual transactions, this is identified as two distinct visits. It is not possible to identify such an occurrence in the data. However, these limitations are judged acceptable, provided that they are understood when interpreting the result.

The final data set contains a record of 696,117 loans from 257,489 distinct visits. Since there is a disparity in the number of households in each OA, a ratio of loans and visits is calculated to develop a measure of loans and visits per household. An index has also been derived from this ratio, standardised to 100 (as explained in Section 3.3.2, STEP 7), to get an idea of the spatial distribution of library visits. Figure 8.1 shows the OAs which are "over-indexing" (have an index of 100 or higher), i.e. the areas which have higher than average library use. The public libraries in the city are also shown. Although there is seemingly some relationship between library location and library use, the patterns suggest that distance to the nearest library is not the only driver of library use.

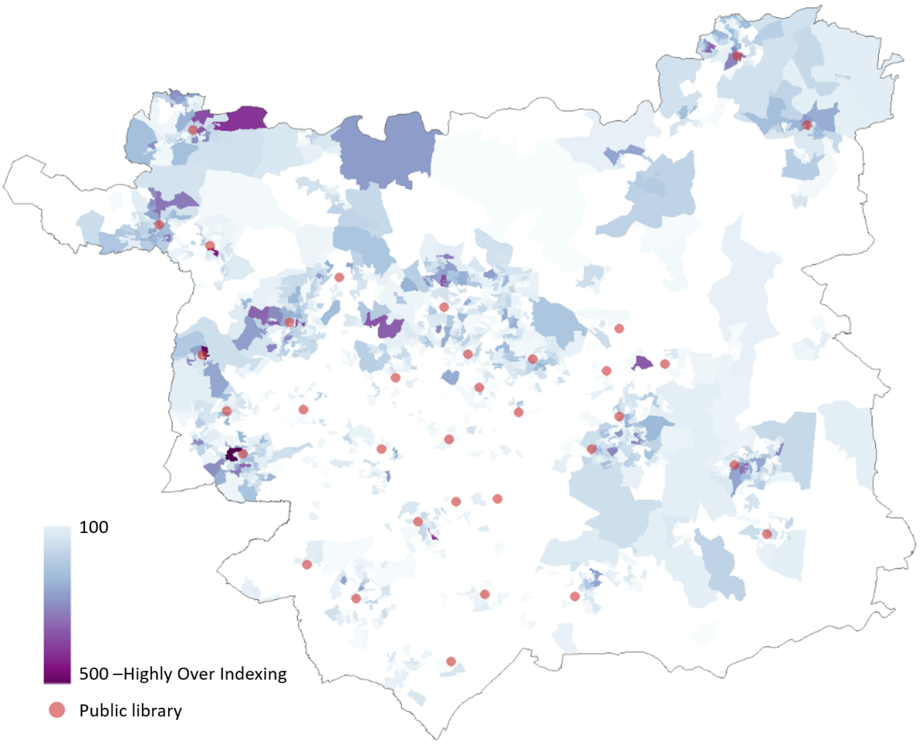


Figure 8.1: Spatial distribution of Leeds OAs with higher than average library use, and locations of public libraries in the city.

Deriving the dependent variable

Figure 8.2 demonstrates a strong positive correlation between library visits in an OA and the count of items loaned. It is therefore deemed unnecessary to include both in the analysis.

Consequently, the ratio of the visits per household in each OA acts as the dependent variable in the FS, below. This represents the *observable outcome* of interest, the propensity to borrow items from LCC libraries.

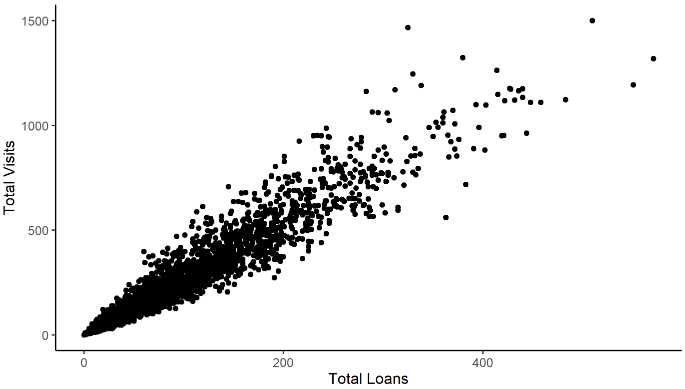


Figure 8.2: Correlation of library visits and library items loaned in each OA.

8.4.3 Methodology

This case study begins by applying a FS method to the 131 predictor variables (in Appendix C.2) (the census variables) to identify a subset of the most important ones, with respect to the dependent variable (the ratio of library visits). A k-means classification is subsequently applied to the variable subset identified, generating a new classification for the OAs. Finally, the discriminative power of the resulting classification in relation to the dependent variable is measured in comparison to the discriminatory power of the FALSOAC classification (derived in Chapter 7 based on the same set of candidate input variables). This comparison is used to evaluate the impact of adopting the FS method to derive cluster groups which are more internally homogeneous with respect to library usage.

Recursive Feature Elimination

A single FS method, a Recursive Feature Elimination (RFE) based on a Random Forest (RF) regression, is presented in this case study to offer a demonstration of the potential for adopting such methods in this context. Random Forest is a popular ensemble decision tree method with an in-built mechanism for measuring and determining the “importance” of each input variable based on its contribution to the decision tree. It versatile and can handle continuous data.

RFE is a popular wrapper method based on backwards elimination, first proposed by Guyon et al. (2002). The broad procedure can be summarised as follows (Gregorutti et al., 2016):

1. Train the RF.
2. Compute the importance of the predictor variables.
3. Eliminate the least important variables.
4. Repeat steps 1-3 until there remain no further variables to remove.

The process detailed above demonstrates the iterative nature of the RFE wrapper method, and the dependence on an *importance measure*. Since it is a backwards elimination method, the process begins by including all variables and removing the least important on each run through until an optimal subset is identified.

Whilst there are several alternative FS methods which could have been considered instead, there is evidence to suggest that RFE capably handles correlated variables, particularly when compared to a standard RF approach (Gregorutti et al., 2016). Since correlations are present among the candidate census variables used here as the predictor variables (see Section 7.4.4), an RFE seems to be an appropriate choice. Future work beyond this thesis might seek to expand on this study by considering alternative methods which might improve upon the performance of the results derived here.

Statistical computing language R has in-built packages which enable the easy running of RFE based on an input of custom parameters. This package is published alongside comprehensive, continually maintained documentation which provides an overview description of RFE in application supported by practical examples (Kuhn, 2019). This documentation is used as a model template to support the application of RFE in this case study.

As is common in ML, the data is split randomly into training and test data subsets. This is based on a standard ratio of 2:1. The important predictor variables are first identified in the training set. However, there is an increased risk of overfitting in a wrapper method, training the result based on particular nuances unique to the data in the training set, which might not be reflective of other subsets of the data. Consequently, a single training set may be insufficient to derive reliable results, and as such, it can be beneficial to run an RFE within an iterative process of *re-sampling*, training the data on a series of subsets of the training data. Though the evidence suggests benefit in increased performance, such a process can be computationally costly (Kuhn, 2019). It is nevertheless recommended, and thus is implemented here in the form of *cross-validation*. As per the default setting in the *rfFuncs* package used, 10-fold cross-validation is executed here. To further reduce the risk of error, this is repeated 5 times with the mean result taken as final.

Traditionally, the configuration of the algorithm underpinning RFE is such that all possible subsets of the candidate predictor variables are explored, however, such a process can be computationally expensive and slow (Guyon and Elisseeff, 2003). Consequently, it is possible to reduce the burden by restricting the tests to a predefined list of subset sizes. Here 46 tests are run in each re-sample, these are of sizes 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130 and 131. This demonstrates complete enumeration of the smaller subsets but groups of five for the larger ones, based on a likelihood of the importance added by each new variable being greater with smaller numbers of subsets, so warranting a more granular evaluation.

Thus, the iterative process of using the cross-validated RFE for variable selection, which is used in this case study, is summarised as follows:

1. Generate the training data sample.
2. Train the RF.
3. Compute the importance of the predictor variables.
4. Set subset size.
5. Eliminate the least important variables up to subset size.
6. Repeat steps 4-5 for all subset sizes.
7. Repeat steps 1-6 for each re-sampling iteration.
8. Calculate the performance profile of the outputs.
9. Determine the appropriate number of predictors.
10. Identify the final list of important predictors.

K-means clustering

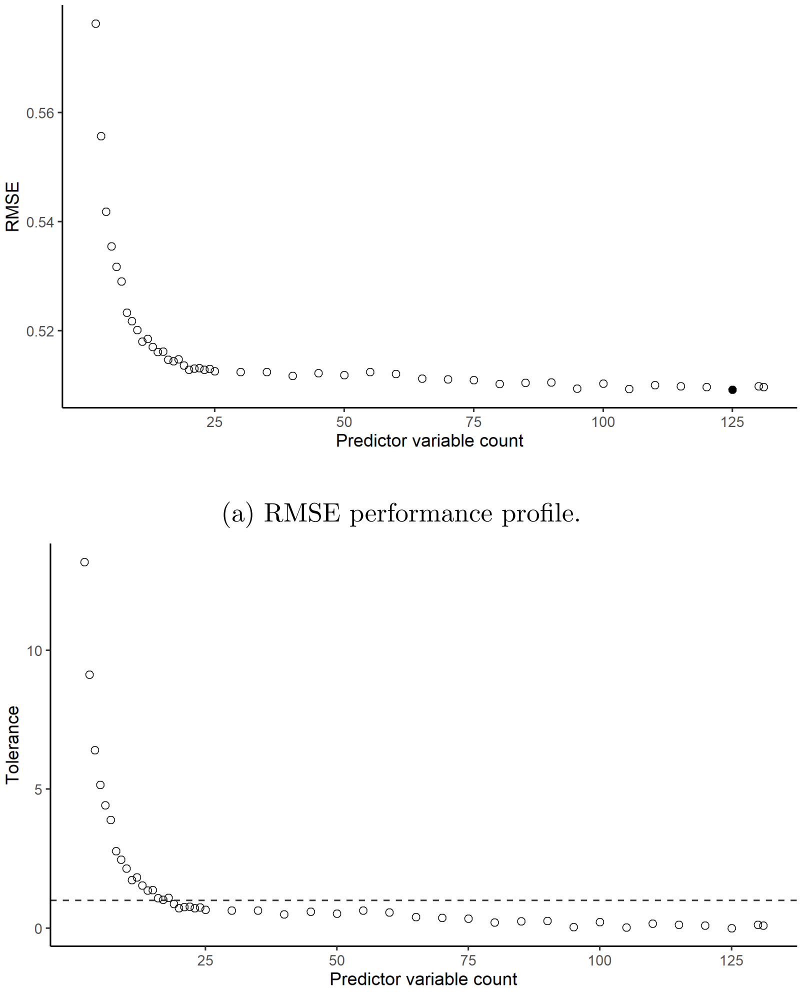
A k-means is now run with the final list of predictors (identified via the methodology laid out in the above section) as the input variables. The k-means is again run with the same parameters as each of the examples in previous chapters, and as per the methodology of the 2011 OAC, including retaining a cluster count of 8 to support comparisons with results from Chapter 4, Chapter 6 and Chapter 7.

8.4.4 Results

RFE

Figure 8.3 (a) illustrates the performance profile of the predictor variable subset sizes tested, based on the Root Mean Square Error (RMSE) as a measure of performance. The plot indicates a best result based on 125 predictor variables. However, the long, flat tail of the plot also indicates that many smaller subsets may produce very similar results, a common characteristic of RF results (Kuhn, 2019). In pursuit of parsimony, to benefit from a reduced number of predictor variables whilst sacrificing minimal performance, an error of 1% is accepted, i.e. a 1% reduction on the best performance is tolerated to benefit from a reduction in the size of the variable set. Figure 8.3 (b) illustrates that such a 1% tolerance reduces the minimal acceptable predictor variable number to 19. This is a substantial reduction in variables for a minimal loss, suggesting that few of these variables have much of an impact on library visits and are thus adding noise into the model.

The 19 candidate predictor variables identified as *most important* are listed in Table 8.1 in order of decreasing importance. These variables predominantly reflect traditional wealth and stability indicators. Several education and employment indicators are also included. Ethnicity and age indicators are noticeably absent, suggesting a limited relationship between such traditional demographic attributes and the propensity for library use.



(a) RMSE performance profile.

(b) 1% tolerance of best RMSE.

Figure 8.3: Performance profile of the predictor variable subset sizes.

Comparing the results of the FALSOAC and FSLSOAC classifications

Performance indicators

The performance of the new classification derived from the variables listed in Table 8.1 (which is henceforth labelled “FSLSOAC”) can be evaluated in several ways.

The objective of the test is to identify whether the adoption of supervised ML methods in the variable selection phase of a place and application specific classification can result in a more targeted output which is more relevant to the application. Therefore, a successful output is considered to be one which has generated groups which are more internally homogeneous, with respect to the application. Likewise, increased heterogeneity between the classification groups is also considered a success.

Importance Rank	**Variable Code**	**Variable Domain**	**Variable Description**
1	L072	Housing	Owned and Shared Ownership
2	L076	Housing	Occupancy rating (rooms) of +2 or more
3	L036	Household Composition	Living in a couple: Married
4	L088	Socio-Economic	Highest level of qualification: Level 4 qualifications and above
5	L124	Employment	Professional occupations
6	L096	Socio-Economic	On foot, Bicycle or Other
7	L013	Demographic	Single (never married or never registered a same-sex civil partnership)
8	L095	Socio-Economic	Private Transport
9	L039	Demographic	Not living in a couple: Single (never married or never registered a same-sex civil partnership)
10	L097	Employment	Economically active: Self-employed
11	L120	Employment	Education
12	L131	Employment	Elementary occupations
13	L094	Socio-Economic	Public Transport
14	L014	Demographic	Married or in a registered same-sex civil partnership
15	L085	Socio-Economic	No qualifications
16	L090	Socio-Economic	No cars or vans in household
17	L042	Household Composition	Not living in a couple: Divorced or formerly in a same-sex civil partnership which is now legally dissolved
18	L123	Employment	Managers, directors and senior officials
19	L092	Socio-Economic	2 or more cars or vans in household

Table 8.1: Candidate variables identified as *important* in the RFE.

The mean differences in the library use ratio for each OA compared to the local classification group mean, for each group, is calculated as a measure of *dissimilarity* for both the FSLSOAC derived here, and the FALSOAC derived in Chapter 7, thus providing a measure of within-cluster homogeneity. This is effectively a one-dimensional replication of the SED which has been adopted for evaluation in the previous chapters, where a lower dissimilarity measure represents greater within-cluster homogeneity. The *Gini coefficient* weighted by the split of OAs in each classification group is adopted as a measure of between-cluster heterogeneity for each of the two classifications. This is a common metric employed in the comparison and validation of heterogeneity in geodemographic classifications both in academia and in the commercial Geodemographics Industry (Petersen et al., 2010; CACI, 2019). A higher Gini coefficient represents greater between-cluster heterogeneity (Petersen et al., 2010). Both of these metrics for the two classifications are shown in Table 8.2.

	FALSOAC	**FSLSOAC**
Dissimilarity	3.13	2.92
Gini coefficient	0.216	0.232

Table 8.2: Comparison of within-cluster homogeneity and between-cluster heterogeneity of FALSOAC and FSLSOAC.

The results of both metrics indicate that the methodology adopted here has performed successfully in terms of generating a classification which has both an increased within-cluster homogeneity and between-cluster heterogeneity, with respect to library use, thus indicating that the use of FS methods in the variable selection phase of a geodemographic classification has derived classification groups which differentiate with a higher degree of relevance to library use, and thus presents a more meaningful result.

Analysis of clusters

The violin plots in Figure 8.4 show the distribution of library use for each of the cluster groups in both classifications. The clusters are ordered left to right from the highest to the lowest mean library use, where the red dots signify the cluster mean. These plots are similar to box-plots, however, they also depict the distribution of the data within each cluster.

Although there is an overall similarity between the two plots, the FSLSOAC does appear to demonstrate a more distinctive split in the library use across the clusters. The clusters with the lowest mean library use in the FSLSOAC (clusters 5, 7, and 8) are more bottom-heavy, with a distribution weighted more lower than the three clusters with the lowest mean library use in the FALSOAC (clusters 6, 5 and 8), indicating that the FSLSOAC has derived clusters with more defined low library use. Likewise, the distribution of the FSLSOAC clusters with the highest mean library use (clusters 1, 2 and 4) are weighted more towards higher library use than their counterparts in the FALSOAC (clusters 2, 1 and 3). Generally, there is a greater distinction across the cluster distributions in the FSLSOAC than in the FALSOAC.

The plot does indicate that there are still OAs in the clusters identifying higher library use which exhibit low library use. This is to be expected in a real-world setting, since all similar OAs will not behave the same. Moreover, these OAs with a high propensity of library use, but who are exhibiting low use, are useful to identify and understand in terms of policy development. Overall, these plots are encouraging and demonstrate the potential of the methodology.

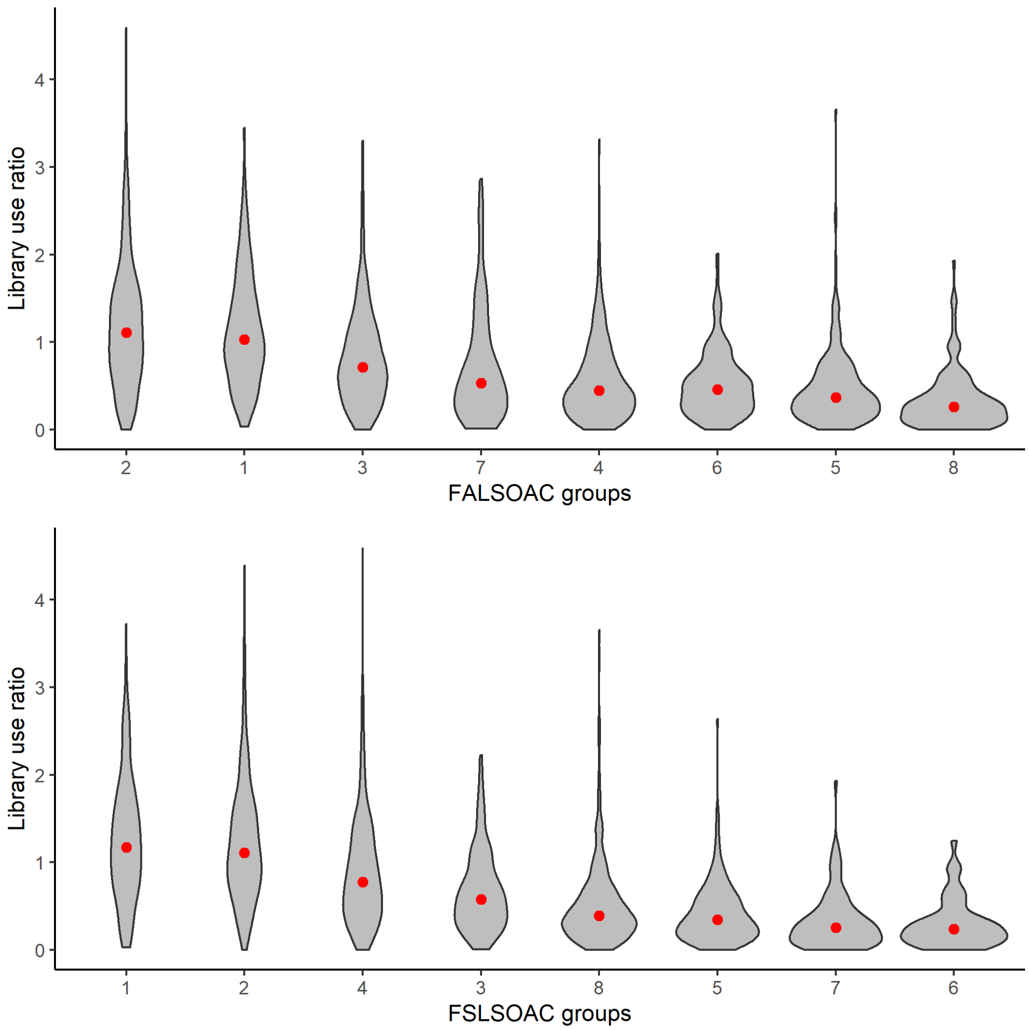


Figure 8.4: Mean and distribution of library use in each classification group in the FALSOAC and FSLSOAC.

The radial plot in Figure 8.5 shows the index of each of the input attribute variables, standardised to 100 (explained in Section 3.3.2 and as adopted in the development of pen portraits in the 2011 OAC (Gale et al., 2016)), for the groups with the highest mean library use in each of the classifications, the "high library use" cluster. Whilst the profile of the OAs in each group demonstrate the same patterns with relation to these attribute variables, the OAs in the FSLSOAC cluster are much more exaggerated. The OAs in this cluster are much further from the mean in almost all of variables which drive library use, as identified by the FS methodology. As such, this plot indicates that the FSLSOAC has been more discriminant in identifying OAs which exhibit attributes associated with library use (both positively and negatively) in the "high library use" cluster, again, demonstrating an improved performance.

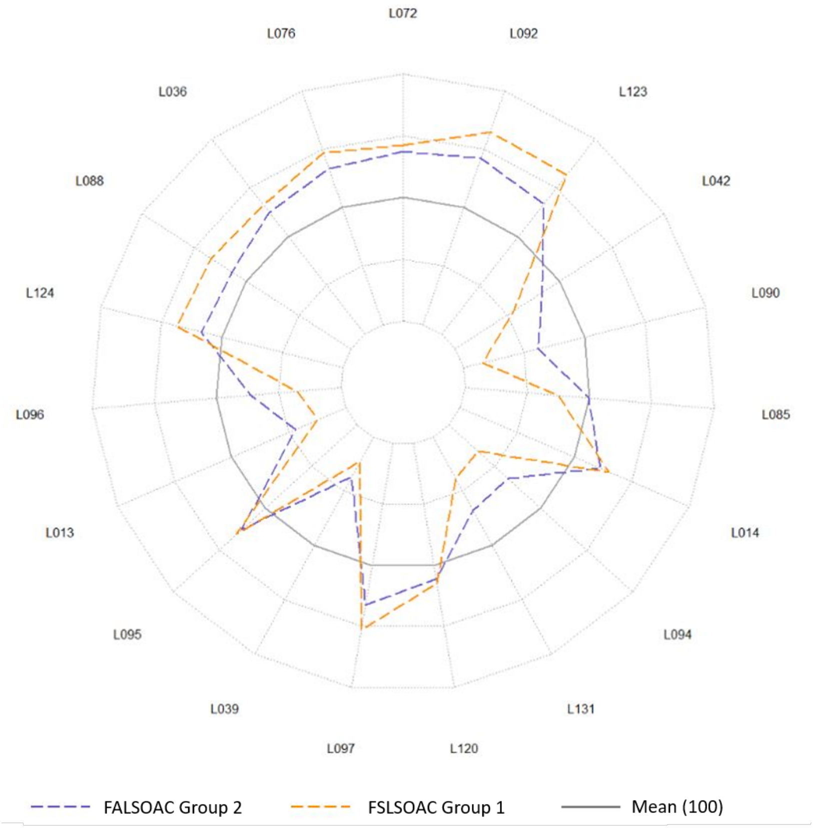


Figure 8.5: Index of census attributes for the groups with the highest mean library use in the FALSOAC and FSLSOAC.

8.5 Discussion

The objectives of this case study were twofold. The first was to demonstrate an application of FS methods for variable selection in the geodemographic classification Standard Framework. The second, to identify the potential for it to introduce an improved fit, with respect to an observable outcome, in this case, the propensity for library visits. As re-iterated throughout, the identification of the best solution was not an objective, nor was the generation of a final classification output. As such, simplicity in this initial demonstration has been prioritised throughout.

8.5.1 Simplicity in the methodology

Consequently, many decisions have been taken here which could be reconsidered in future research to further improve the outcome. For example, visits for any purpose which did not result in book loans were not included in the analysis. The data also focused on a specific snapshot in time, and the ratio of visits per household which was calculated led to small counts in some OAs, both of which could have impacted the analysis and result.

Moreover, the independent input variables, again, were sourced solely from the census, somewhat arbitrarily (to maintain consistency with Chapter 7). Whilst the method proved successful at filtering out unnecessary or irrelevant variables presenting "noise", there might be alternative, non-census data could have a closer relationship to library use and thus should have been included but were missed. Examples might also include indicators which are descriptive of the libraries rather than the population, for example, the distance to the closest library, or even attractions at the library, by distance. Roumpani et al.'s (2020) paper indicates an uptake in library use by parents where children's "story time" is available, for instance.

8.5.2 Alternative FS methods

In terms of the methodology itself, this case study has employed RFE with RF, since it is recommended for variables which exhibit non-normality, and is accessibly applied through the statistical package R, thus supporting replications of the analysis either in future academic research, or within LCC. However, there is therefore scope in research which extends this thesis, or additional work prior to practical implementation, to conduct tests of the impacts or potential improvements afforded by other FS methods. These improvements could affect the process or the result. For example, RFE, as employed here, is computationally expensive and can take some time to execute, it is thus not necessarily suited to a high number of variables (Guyon and Elisseeff, 2003; Karegowda et al., 2010). An alternative approach might therefore be preferred in practice. Similarly, tests could also be conducted to optimise for the best cluster count, in the clustering phase of the classification development. A count of 8 has been selected as default in this case study, again, to maintain consistency with the classification derived in Chapter 7, but this might not generate the best result.

8.5.3 Successes

Despite these caveats, the indicative use-case presented still demonstrates a positive result. The hypothesis that an employment of a supervised ML element in the variable selection could facilitate a more targeted, and contextually relevant classification appears to have been verified in this case. As mentioned, the FS seemed to handle the arbitrarily selected input variable set well, filtering the "noisy" variables, and identifying those which led to a solution that generated groups better able to discern library use, as hoped. In application, the results derived from this case study provide a more relevant idea of who and where the population more inclined to use libraries are. It is also still possible to use the resulting classification in a traditional way, appending ancillary data on to the result, to gain an even richer picture of the library users and non-users more accurately than a similar approach applied to general-purpose classification outputs. This will be able to better support the development of more informed planning strategies, or even marketing initiatives targeted at households identified as having a high propensity for library use, but not yet exhibiting such behaviour.

8.5.4 Future extensions

Moreover, evidence of an ability to employ the FS methodology iteratively to more contextually hone the candidate variables was also detected. The finding that ethnicity variables seem to bare limited relation to library use when combined with the other variables employed might direct the developer to cease seeking similar variables, or even non-census alternatives of ethnicity variables, in extensions which look to combine the activity of Chapter 5 with this case study. Alternatively, the results could also support more targeted sourcing of variables which seemingly are closely related to library use. In both instances, an iterative use of FA could cut wasted time and facilitate the developer in a more targeted approach to data identification. When opening development up to a world of potential data, it is increasingly necessary to find a way of cutting through the noise and identifying the data and variables of value.

However, whilst simplicity has been pursued wherever possible, the application specific framework proposed here does not match the level of simplicity achieved in the established methodologies, which have popularised geodemographics. Although, the load has been added to the development process, and has not translated in more complicated outputs. Whilst this methodology moves the notion of classifications further away from a one-size-fits all approach, the results generated by the case study itself, for the end-user, are extremely familiar and can be employed in the traditional way. In reference to discussions in Section 2.3.2, the advancement proposed here is, again, evolutionary rather than revolutionary. Yet, it is an advancement, which has generated demonstrable improvements in the case study presented.

As alluded to, keeping many of the decisions simple in this case study generates an additional benefit in terms of replicability within LCC (and in the public sector in general). The main outcome of this chapter has not been a single classification, but the development of a template to facilitate recreations of the case study in different applications. However, it is worth considering whether it meets the needs of local government. Although the potential might be to offer more superior results, further research, evaluating whether it is realistic to expect local governments to develop "on-the-fly" classifications in this way, could be needed. This is a discussion which is likely best held within local government, and even in regard to particular applications. However, the broad conclusion of this chapter is that the framework demonstrated offers the potential for deriving more relevant classifications, which would be worth pursuing if practicalities allow.

If the opportunities offered by a template such as this to develop more on-the-fly classifications could be taken, this might also promote a culture of generating classifications more regularly. This could permeate the application-specific focus and see more frequent updating of all existing classifications with more timely data, or with additional variables where new data becomes available. An additional benefit of more fluid classification development could be a shift away from the convention of naming the classifications and deriving fixed pen portraits of areas. Not only would the omission of this step save time, it would also relieve the pressures associated with assigning names which are not only meaningful, but will not lead to misleading interpretations, or offend resident populations. Moreover, in terms of application specific classifications, involving problem owners and their expertise in the development phase, if possible, would eliminate the reliance on developers to be experts in the variable selection process. These ideas will be considered more fully in Chapter 9.

8.6 Conclusions and next steps

The hypothesis proposed in this chapter challenges the established methodology of classification development to ask whether there is alternative way of considering the development of more targeted geodemographic application-specific classifications, and whether such an approach would generate more relevant and meaningful results in practice. The case study presented, which focuses on deriving classifications which better discern library use, seemingly verifies this hypothesis. The results are encouraging, exhibiting demonstrable evidence that the use of supervised FS methodologies for variable selection can lead to an enhanced relevance in the classification derived. Overall, the application specific approach seems to provide a necessary, updated alternative to the existing classification approaches. Next steps would be to consider how these approaches would be received and adopted in practice, particularly within a LA setting. This will be discussed further in Chapter 9.

Summary

This chapter presents a proposal for shifting from a traditional approach for generating general-purpose one-size fits all geodemographic classifications to application-specific classifications (extending the notion of domain specific classifications), followed by a discussion of the practicalities. A case study is presented successfully demonstrating the increased relevance that such a shift can impose on the results based on a specific example focused on developing a classification which is better suited to discerning the propensity for library use across the OAs in Leeds.

Chapter 9 - Discussion, conclusions and future research agenda

9.1 Research outputs

This research has sought to extend and enrich development practices within the open geodemographics landscape in the UK, particularly for development within, and for the use of, the public sector. The research has focused primarily on the development of local, place-specific classifications, based on the case study of Leeds. The work has explored the improvements in classification performance to be gained from shifting to a local extent, and the scope for practical advancements which it facilitates, in terms of opening up the potential for inclusion of new and novel input variables from the open and public sector landscape. Later chapters extend upon these developments, exploring the necessity and potential of improved variable selection procedures, the importance of which are magnified by such an introduction of novel variables.

This thesis has derived several key outputs, important findings and recommendations:

- Shifting to a local extent in the development of a geodemographic classification for a city such as Leeds has the potential to unearth city-specific phenomena. For example, in Leeds, nuances in the student population and the recently graduated population are best captured locally. These are large and important populations in the city, thus it is essential to capture such nuance.
- National and place-specific classifications have the potential to complement one another. This is evidenced the limited rural small-area geographies in Leeds, the populations of which appear to more closely resemble other rural populations found more commonly within the national extent than the local extent.
- The sophistication of the 2011 OAC methodology could be improved, particularly with regards to the variable selection procedures employed and the input data adopted. As demonstrated, there is scope to extend to include novel administrative data, and/or employ ML techniques in the selection of variables.
- Administrative data with the potential for inclusion in geodemographic classification de-

velopment is available within local government, specifically highlighted within LCC, as has been discussed for many years. However, substantial work is required to convert the data into usable variables which are capable of updating and extending open geodemographic classifications beyond the census variables adopted in the 2011 OAC. Nevertheless, the work required is somewhat reduced in the development of place-specific, rather than national level classifications.

- Feature Stores (variable repositories) could offer many benefits in the future use of public sector data, both within and beyond open geodemographics. Critically, their introduction could establish much needed consistency, increasing confidence in any analysis produced from the variables, and free up the time of analysts and developers to concentrate on insight development.

- There is scope for LCC to build their own place-specific geodemographic classifications in replacement of an existing reliance on proprietary commercial products.

- There are substantial boundary and data compatibility issues introduced when linking census and non-census data which must be addressed. The implications of these issues should be understood and reported in all research affected and alongside the ONS resources which facilitate the inconsistent linking of the data. The 2021 census should be more considerate of these issues and mitigate where possible, particularly as the linking of such spatial data becomes increasingly commonplace.

- The incorporation of novel data into the development of geodemographic classifications will not necessarily reap improved performance. The inclusion of new data increases the importance of more discriminatory variable selection procedures.

- The use of unsupervised variable selection methods, such as Factor Analysis, have the potential to offer improved performance, however, these will not address or alleviate increasing concerns regarding general-purpose classifications.

- Supervised learning techniques have shown some success in deriving place-specific classifications which more relevantly discern population structures based on a specific application of interest. This is an indication that they could offer a useful evolution to the development of more relevant classifications which might better support policy development and strategic planning decisions, including in the public sector.

- The methodology presented, which supports the inclusion of supervised learning methods for variable selection, would facilitate the development of useful and meaningful classifications on-the-fly by experts with local or domain knowledge.

This thesis has therefore generated a considerable number of findings which support the development of place-specific classifications from both a practical and theoretical perspective, many of which might also have far wider implications for local government data practices and infrastructures.

9.2 Summary and critique of research findings

Several challenges relating to the research findings listed in Section 9.1 have also been identified. The main themes of these concerns are summarised below.

Firstly, there are concerns that geodemographic classification outputs might have been generated by a consequence of the decisions made by the developer, of the reliability of the data, or may be an artefact of the statistical procedures employed, rather than reflective of genuine divisions in population. These concerns, which predate the work in this thesis, however, are not entirely admonished here, and pose a substantial threat to the confidence of geodemographic classification outputs. Moreover, similar concerns are further raised where additional statistical processes are introduced into the development of classifications, such as for variable selection, as has been explored here. Although it might never be truly possible to eradicate these concerns, many of the explorations contained within this thesis seek to reduce these risks by removing some of the decisions and taking away some of the reliance on the developer, or by discussing and practically demonstrating the level of transparency and exploration of the input data required to re-introduce confidence in the result.

Secondly, the primary output of this thesis is not a final classification of Leeds which is supported by pen portraits for each OA, as is traditional in classification development. Whilst these kind of outputs are familiar and easy to understand and re-use, there are inherent weaknesses, particularly in their misunderstanding and misuse, for example, in the concept of local and global mean, where the idea of more likely is interpreted as likely. Such a simplified output has been avoided here. Instead, transparency in the development process has been prioritised. Having a clearer understanding of the inside of the build and being made to consider the outputs, as presented here, rather than a pre-set profile, could reduce potential misuse.

The main tangible outputs here, therefore, are template frameworks which can be adapted to develop future place-specific geodemographic classifications with novel data and sophisticated variable selection procedures. These templates are supported by rich documentation and a swathe of recommendations. This approach could raise several concerns. Primarily, the target end-user here is principally LCC, and other LAs who wish to replicate similar bespoke classification development. The first concern with this is that, historically, the simplicity of geodemographics has been heralded as the feature which has underpinned its popularity and widespread use in LAs. This element has been removed in this approach. Secondly, the expectation is that LAs are willing and able to produce such bespoke classifications in-house, which might not be realistic. The same can be said for the infrastructure development recommended in Chapter 6, particularly in terms of the suggested Feature Stores.

However, there has already been research published suggesting that it might be an over-simplification in itself to claim that end-users universally value a single and simple output. Slingsby et al. (2011) found that end-users in LAs often felt better equipped employing classifications in practice when supplied with more supporting information, rather than just given the top-level classification and description. Many are experienced, intelligent analysts capable of handling such information, who are working in environments which are becoming increasingly data and insight led. Moreover, LAs might look to partner with research institutions or other organisations with resources and expertise to reduce the burden in-house, if possible, and necessary. Moreover, the work itself need not be over complicated. Chapter 8 illustrates a sophisticated yet achievable Feature Selection methodology which has already demonstrated improved performance. This has been executed in this way to illustrate the potential transferability to LAs. Much of the work here has been purposefully demonstrative to indicate the scope for extension of the traditional practices, to explore the direction which advancements might take, and to illustrate the potential benefits which could be gained.

Finally, a more specific concern is with the concept of place-specific classifications. One perceived benefit of national level classifications, as mentioned, is the comparability afforded, for example, between cities. Place-specific classifications cannot support direct city-to-city comparisons. However, this is somewhat of a moot concern, since the place-specific classifications have been proposed as a complimentary addition to the geodemographics landscape. There may still be applications where the national level classifications are the appropriate choice. However, this thesis has successfully demonstrated instances where the place-specific classifications perform much better at revealing local population structures, an ability which affords substantial benefit for practitioners with a primarily local focus, such as in the allocation of LA resource and services. This thesis has also highlighted, however, that place-specific classifications need not simply be a scaled-down replica of the national level alternatives. The development of place-specific classifications has a greater opportunity to be increasingly bespoke, for example, including novel data which is more readily available at the local level, or being close enough to the detail to generate an application-specific classification, and thus should be purposefully developed.

9.3 Public sector relevance and recommendations

The work contained in this thesis has all been conducted with LCC (and the public sector) as a primary end-user, closely considering LCC's needs and capabilities. As such, the potential for reproducibility by LCC, or other LAs with access to similar data, has been a high priority in all of the work conducted. The research is therefore extremely relevant to the public sector in many ways. Several are highlighted below.

Chapters 4, 6, 7 and 8 have each demonstrated an example of generating a place-specific classification for Leeds. Each has derived insights which were not available through the use of a national level classification, and which could prove particularly useful for informing policy development and decision making within LCC. This thesis has thus presented evidence to support the hypothesis that more targeted geodemographic classification outputs could lead to more targeted policy and intervention development.

Moreover, Chapters 5 and 6 demonstrated the scope for incorporating public sector administrative data into the development of such a place-specific classification. This not only illustrated the positive potential of the data and its possible inclusion in the creation of the classification, importantly, it also highlighted several key challenges and limitations of the data infrastructure within the LA. It evidenced the level of preparation required to convert the data into the necessary format, and documented the time and energy consumed. Valuable recommendations have been proposed to indicate potential improvements to practices which might better facilitate the use of LCC data, both in geodemographic development and beyond. Particularly, Chapter 5 demonstrated the importance of speeding up existing data-sharing practices. This is not specific to this thesis and has been repeatedly discussed before (Carroll and Crawford, 2020), however, is worth re-iterating.

Furthermore, these discussions have also highlighted several areas where LCC are ahead of the curve. Examples include their employment of data, open data initiatives, analytical processes, data and GIS skills, close partnerships with research institutions, and the culture and respect for data which runs throughout the LA. Each of these characteristics have been critical in facilitating much of the work documented in this thesis, and moreover, can act as an example to other LAs seeking to derive value from their own data and data practices. This strong existing platform will also help support the implementation of many of the recommendations made in this thesis.

Chapter 5 has demonstrated the critical necessity to address incompatibility and geographic boundary issues which have been identified, associated with linking census and non-census (including administrative) data. This is common practice in the public sector, where data is regularly aggregated to align with census boundary issues. This thesis has both identified this as an issue and made strong recommendations that these issues be addressed, but in the meantime, that they are better publicised to draw attention to the potential for misuse of public sector directories facilitating geographic conversions. The importance of this will only grow as researchers and policy makers continue to increasingly centre their decision making with both census and non-census data, risking undermining the accuracy of the results, insights and subsequent decisions generated if this issue continues to go unaddressed.

9.4 Further development and future work

This thesis has primarily focused on moving geodemographics to a local level, adopting novel open and public sector data, and developing increased sophistication in the variable selection processes. Underpinning this work, there has also been a focus on supporting a move to increased transparency and reproducibility within geodemographic classification development. As documented in detail in Chapter 3, a raft of other challenges to the established practices of developing geodemographic classifications have not been considered. Examples include, the scope for improving or adopting alternative clustering methods (beyond k-means), addressing Uncertainty within geodemographics and the routine development of on-the-fly classifications. Whilst this work has touched on the latter, these have not been the priority of this work, and thus, might be considered in future work. This thesis has also not extended beyond the development of place-specific classifications to consider their application, which might also drive future work, particularly in the context of the public sector, developing policy and planning initiatives, as have been repeatedly referenced as possible throughout.

A range of extensions and future development of the work which has been conducted are also possible. These include the consideration of alternative novel data, including data relating to the other non-housing domains considered in the 2011 OAC, or entirely alternative data, such as behavioural, attitudinal and lifestyle data. These might also consider alternative Feature Extraction or Feature Selection methods, or to carry out the tests which have been suggested to improve the models presented here, such as combining these techniques with novel data. These might also consider the development of the practices demonstrated here in-house, within LCC, to explore the scope for implementing the recommendations made, particularly with regards to the application-specific classifications proposed, or disseminating the recommendations made here to other LAs.

Finally, future work should be carried out to address and mitigate against the incompatibility and boundary issues introduced in linking census and non-census data, which could have wide and unknown consequences. This should be considered in future work with relation to future census taking and the production of population statistics, and work should be conducted to raise more awareness alongside the publication of the UPRN to OA and postcode to OA directories (and other relevant directories).

9.5 Concluding remarks

The elements of geodemographic classification development which have been addressed here have been highlighted for progress for many decades, notably the shift to a place-specific extent and the routine inclusion of non-census data in open geodemographics. The practical progress, however, is in its infancy, and has thus far been limited. Considering both objectives in combination has supported the exploration of each here. The place-specific focus has increased data availability, substantially improving the prospect of including more novel data within the development process.

However, several longstanding barriers to progress have continued to pose challenges to the work conducted within. In many instances where these challenges have been overcome, the issues have still compromised the speed of the progress. These experiences offer further explanation for the superior sophistication of data and practices adopted in the commercial geodemographics sector, who are less incumbered by many of the challenges faced. Notably, commercial classifications development is supported by large teams with excellent access to data, substantial investment and dedicated attention. Nevertheless, the findings presented in this thesis should not act as a deterrent, but as encouragement, for the future development of open geodemographics in the public sector and academia. Particularly so for place-specific classification development, the benefits of which have been repeatedly professed and evidenced.

The rich and valuable data available within the public sector has been widely discussed, particularly data from within LCC, in addition to the wealth of expertise and the desire to progress the field. This thesis has transparently demonstrated the much easier implementation of statistical processes today, which have previously been a barrier to development. Combined, these factors have supported advancements in the development of a place-specific geodemographic classification for Leeds, here. Likewise, the findings presented should act as encouragement for further development elsewhere, to support in the delivery of the best, most appropriate, and most meaningful geodemographic classifications for local needs. These will present much needed alternatives to the incumbent off-the-shelf, national-level classifications with the potential to offer more bespoke and targeted insights, supporting the creation and implementation of better policy development, service and resource delivery, and local initiatives, leading to tangible benefits all round.

References

ADF Software. 2021. *Censation.* [Online]. [Accessed 20 January 2021]. Available from: https://www.afd.co.uk/data-sets/censation/.

Adnan, M., Longley, P. A., Singleton, A. D. and Brunsdon, C. 2010. Towards real-time geodemographics: Clustering algorithm performance for large multidimensional spatial databases. *Transactions in GIS.* **14**(3), pp. 283–297.

Adnan, M., Singleton, A. D. and Longley, P. 2013. Spatially weighted geodemographics. *GIS Research UK 21st Annual Conference, Liverpool University.*

Alexiou, A. 2017. *Putting 'Geo' into Geodemographics: evaluating the performance of national classification systems within regional contexts.* PhD thesis. University of Liverpool.

Alexiou, A. and Singleton, A. D. 2015. Geodemographic analysis. In *Geocomputation: A Practical Primer.* London: SAGE. pp.137-151.

Anderson, T. K. 2010. Using geodemographics to measure and explain social and environment differences in road traffic accident risk. *Environment and Planning A.* **42**(9), pp. 2186–2200.

Anderson, T. R. and Bean, L. L. 1961. The shevky-bell social areas: confirmation of results and a reinterpretation. *Social Forces.* **40**(2), pp. 119–124.

Arribas-Bel, D. 2014. Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography.* **49**, pp. 45–53.

Ashby, D. I. and Longley, P. A. 2005. Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS.* **9**(1), pp. 53–72.

Atlas, M. 1981. Gambling with elections: The problems of geodemographics. *Campaigns & Elections,(Fall).* pp. 4–12.

Aveyard, P., Manaseki, S. and Chambers, J. 2002. The relationship between mean birth weight and poverty using the townsend deprivation score and the super profile classification system. *Public Health.* **116**(6), pp. 308–314.

Baker, K. 1991. Using geodemographics in market research surveys. *Journal of the Royal Statistical Society: Series D (The Statistician).* **40**(2), pp. 203–207.

Batey, P. and Brown, P. 1995. From human ecology to customer targeting: the evolution of geodemographics. In: Longley, P. and Clarke, G. ed(s). *GIS for business and service planning.* Glasgow: Bell and Bain, pp. 77-103.

Batey, P. and Brown, P. 2007. The spatial targeting of urban policy initiatives: a geodemographic assessment tool. *Environment and Planning A.* **39**(11), pp. 2774–2793.

Batey, P., Brown, P. and Pemberton, S. 2008. Methods for the spatial targeting of urban policy in the uk: A comparative analysis. *Applied Spatial Analysis and Policy.* **1**(2), pp. 117–132.

Beacon Dodsworth. 2021a. *Making P^2 People and Places.* [Online]. [Accessed 20 January 2021]. Available from: https://beacon-dodsworth.co.uk/tools/geodemographic-software/making-p2/.

Beacon Dodsworth. 2021b. *P^2 People and Places.* [Online]. [Accessed 20 January 2021]. Available from: https://beacon-dodsworth.co.uk/landing-pages/p2-people-and-places/.

Beaumont, J. R. and Inglis, K. 1989. Geodemographics in practice: developments in britain and europe. *Environment and Planning A.* **21**(5), pp. 587–604.

Benneworth, P., Bakker, I. and Velderman, W.-J. 2018. Beyond big data, the open data revolution for research. In: *Knowledge, Policymaking and Learning for European Cities and Regions.* Edward Elgar Publishing. pp. 193–205.

Berry, B. J. L. and Kasarda, J. D. 1977. *Contemporary urban ecology.* New York: Macmillan.

Berry, B. J. and Smith, K. 1971. *City classification handbook.* New York: Wiley-Interscience.

Birkin, M. and Clarke, G. 2012. The enhancement of spatial microsimulation models using geodemographics. *The Annals of Regional Science.* **49**(2), pp. 515–532.

Blake, M. and Openshaw, S. 2005. *Selecting variables for small area classifications of 1991 UK census data.* School of Geography, University of Leeds.

Blakemore, M. and Masser, I. 1991. *Handling geographical information: methodology and potential applications.* Longman Scientific & Technical.

Brown, N. 2011. Robert Park and Ernest Burgess: Urban ecology studies, 1925. *Center for Spatially Integrated Social Science .*

Brown, P. J. 1991. Exploring geodemographics. *Handling Geographical Information.* pp. 221–258.

Brown, P. J., Hirschfield, A. F. and Batey, P. W. 2000. Adding value to census data: public sector applications of the super profiles geodemographic typology. *Journal of Cities and Regions.* **10**(19), pp. 19–31.

Brunsdon, C., Charlton, M. and Rigby, J. E. 2018. An open source geodemographic classification of small areas in the republic of ireland. *Applied Spatial Analysis and Policy.* **11**(2), pp. 183–204.

Brunsdon, C., Longley, P., Singleton, A. D. and Ashby, D. 2011. Predicting participation in higher education: A comparative evaluation of the performance of geodemographic classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* **174**(1), pp. 17–30.

Brunsdon, C. and Singleton, A. D. 2015. *Geocomputation: a practical primer.* London: SAGE.

Burgess., E. W. 1925. The growth of city: An introduction to a research project. In *The City.* Chicago: University of Chicago Press, pp.47-62.

Burns, L., See, L., Heppenstall, A. and Birkin, M. 2018. Developing an individual-level geodemographic classification. *Applied Spatial Analysis and Policy.* **11**(3), pp. 417–437.

Burrows, R. and Gane, N. 2006. Geodemographics, software and class. *Sociology.* **40**(5), pp. 793–812.

Cabinet Office. 2012. *New funding to accelerate benefits of open data.* [Online]. [Accessed 17 August 2020]. Available from: https://www.gov.uk/government/news/new-funding-to-accelerate-benefits-of-open-data.

Cabinet Office. 2017. *Government Transformation Strategy: better use of data.* [Online]. [Accessed 3 June 2020]. Available from: https://www.gov.uk/government/publications/government-transformation-strategy-2017-to-2020/government-transformation-strategy-better-use-of-data.

CACI. 2019. *ACORN Technical Guide.* [Online]. [Accessed 21 January 2021]. Available from: https://www.caci.co.uk/sites/default/files/resources/Acorn_technical_guide.pdf.

CACI. 2020. *What is Acorn?* [Online]. [Accessed 3 June 2020]. Available from: https://acorn.caci.co.uk/what-is-acorn.

Caldicott, F. 2020. *Data sharing during this public health emergency.* [Online]. [Accessed 8 January 2021]. Available from: https://www.gov.uk/government/speeches/data-sharing-during-this-public-health-emergency.

Carroll, N. and Crawford, A. 2020. *Unlocking the Potential of Civic Collaboration: A review of research-policy engagement between the University of Leeds and Leeds City Council.* [Online]. [Accessed 18 November 2020]. Available from: https://lssi.leeds.ac.uk/partnerships/review-of-collaborative-working/.

Census Information Scheme. 2017. *London Workplace Zone Classification.* [Online]. [Accessed 12 June 2020]. https://data.london.gov.uk/dataset/london-workplace-zone-classification.

Champion, T. 2014. People in cities: the numbers. *Future of Cities Project, Working Paper.* **3**.

Charlton, M., Openshaw, S. and Wymer, C. 1985. Some new classifications of census enumeration districts in Britain: a poor mans ACORN. *Journal of Economic and Social Measurement.* **13**(1), pp. 69–96.

Child, D. 2006. *The essentials of factor analysis.* 3^{rd} ed. New York, NY: Continuum International.

Clark, D., Davies, W. and Johnston, R. 1974. The application of factor analysis in human geography. *Journal of the Royal Statistical Society: Series D (The Statistician).* **23**(3-4), pp. 259–281.

Clark, S. D. and Lomax, N. 2018. A mass-market appraisal of the english housing rental market using a diverse range of modelling techniques. *Journal of big data.* **5**(1), pp. 1–21.

Comrey, A. and Lee, H. 1992. *A First Course in Factor Analysis.* 2^{nd} ed. New Jersey: Psychology press.

Conservatives. 2015. *The Conservative Party Manifesto 2015.* [Online]. [Accessed 17 August 2020]. Available from: http://ucrel.lancs.ac.uk/wmatrix/ukmanifestos2015/localpdf/Conservatives.pdf.

Corcoran, J., Higgs, G. and Anderson, T. 2013. Examining the use of a geodemographic classification in an exploratory analysis of variations in fire incidence in south wales, uk. *Fire safety journal.* **62**, pp. 37–48.

Costello, A. B. and Osborne, J. 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation.* **10**(7), pp. 1–97.

Dalton, C. M. and Thatcher, J. 2015. Inflated granularity: Spatial "big data" and geodemographics. *Big Data & Society.* **2**(2), pp. 1–15.

Data Mill North. 2020. *Data Mill North: Datasets.* [Online]. [Accessed 11 June 2020]. Available from: https://datamillnorth.org/dataset.

Davies, W. K. 1978a. Alternative factorial solutions and urban social structure: a data analysis exploration of calgary in 1971. *Canadian Geographer/Le Géographe canadien.* **22**(4), pp. 273–297.

Davies, W. K. 1978b. Charles booth and the measurement of urban social character. *Area.* pp. 290–296.

Dean, N. 2018. Factor analysis lab. Workshop notes distributed in An Introduction to Latent Variable Modelling. 2 July, White Rose Doctoral Training Partnership, University of Sheffield.

Debenham, J. 2002. Understanding geodemographic classification: Creating the building blocks for an extension. *Working paper.* School of Geography, University of Leeds. [Online]. [Accessed 26 January 2021]. Available from: http://eprints.whiterose.ac.uk/5014/1/02-1.pdf.

Debenham, J., Clarke, G. and Stillwell, J. 2003. Extending geodemographic classification: a new regional prototype. *Environment and Planning A.* **35**(6), pp. 1025–1050.

Department for Digital, Culture, Media Sport. 2019. *UK National Action Plan for Open Government 2019-2021.* [Online]. [Accessed 17 August 2020]. Available from: https://www.gov.uk/government/publications/uk-national-action-plan-for-open-government-2019-2021.

Department for Digital, Culture, Media and Sport. 2018. *Libraries Deliver: Ambition for Public Libraries in England 2016 to 2021.* [Online]. [Accessed 27 January 2021]. Available from: https://www.gov.uk/government/publications/libraries-deliver-ambition-for-public-libraries-in-england-2016-to-2021/libraries-deliver-ambition-for-public-libraries-in-england-2016-to-2021.

Department of Health. 2012. *Caring for our future: reforming care and support.* [Online]. [Accessed 18 July 2020]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/136422/White-Paper-Caring-for-our-future-reforming-care-and-support-PDF-1580K.pdf.

Drury, C. 2020. All libraries, museums and galleries in leeds at risk of closure as local councils count cost of coronavirus. *Independent.* [Online]. [Accessed 27 January 2021]. Available from: https://www.independent.co.uk/news/uk/home-news/leeds-council-libraries-museums-galleries-close-coronavirus-a9572746.html.

Dziuban, C. D. and Shirkey, E. C. 1974. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin.* **81**(6), pp. 358–361.

Everitt, B. and Hothorn, T. 2011. *An introduction to applied multivariate analysis with R* New York: Springer Science & Business Media.

Everitt, B. S. and Dunn, G. 1991. *Applied multivariate data analysis.* 2^{nd} ed. London: Hodder Arnold.

Experian. 2009. *Mosaic United Kingdom: The consumer classification of the United Kingdom.* [Online]. [Accessed 10 June 2020]. Available from: https://www.experian.co.uk/assets/business-strategies/brochures/Mosaic%2520UK%25202009%2520brochure%5B1%5D.pdf.

Experian. 2018. *Data Sets Guide: Unlocking accuracy and enhancing insight.* [Online]. [Accessed 3 June 2020]. Available from: https://www.experian.co.uk/content/dam/marketing/emea/soafrica/za/assets/data-quality/data-sets-guide.pdf.

Experian. 2021. *Mosaic: Consumer classification for consistent cross-*

channel marketing. [Online]. [Accessed 20 January 2021]. Available from: https://www.experian.co.uk/business/marketing/segmentation-targeting/mosaic/.

Farr, M. and Evans, A. 2005. Identifying 'unknown diabetics' using geodemographics and social marketing. *Journal of Direct, Data and Digital Marketing Practice.* **7**(1), pp. 47–58.

Freeguard, G., Shepheard, M. and Davies, O. 2020. *Digital government during the coronavirus crisis.* [Online]. London: Institute for Government, pp 1-64. [Accessed 8 January 2021]. Available from: https://www.instituteforgovernment.org.uk/sites/default/files/publications/digital-government-coronavirus.pdf.

Gale, C. 2020. *The 2011 Area Classification for Output Areas.* [Online]. [Accessed 29 June 2020]. Available from: http://geogale.github.io/2011OAC/.

Gale, C., Adnan, M. and Longley, P. 2012. *Open Geodemographics: Open Tools and the 2011 OAC.* [Online]. [Accessed 11 January 2021]. Available from: https://www.geos.ed.ac.uk/ gisteac/proceedingsonline/GISRUK2012/Papers/presentation-17.pdf.

Gale, C. G. 2014. *Creating an open geodemographic classification using the UK Census of the Population.* PhD thesis. UCL (University College London).

Gale, C. G., Singleton, A. D., Bates, A. G. and Longley, P. A. 2016. Creating the 2011 area classification for output areas (2011 oac). *Journal of Spatial Information Science.* **2016**(12), pp.1–27.

Gale, C. and Longley, P. 2012. Geodemographic output area classification for london, 2001–2011. Edinburgh: GISRUK.

GeoPlace. 2021. *Our story: bringing location to life.* [Online]. [Accessed 22 January 2021]. Available from: https://www.geoplace.co.uk/about-us/who-we-are/our-story.

Ghosh, P. 2019. *AAAS: Machine learning 'causing science crisis'.* [Online]. [Accessed 17 August 2020]. Available from: https://www.bbc.co.uk/news/science-environment-47267081.

Gov.uk. 2020. *House in multiple occupation licence.* [Online]. [Accessed 18 November 2020]. Available from: https://www.gov.uk/house-in-multiple-occupation-licence.

Gregorutti, B., Michel, B. and Saint-Pierre, P. 2016. Correlation and variable importance in random forests. *Statistics and Computing.* **27**(3), pp. 659–678.

Griffiths, D. 2020. *The Sonar Geo-Demographic System.* [Online]. [Accessed 20 January 2021]. Available from: http://www.tracconsultancy.co.uk/data/uploads/sonar-report-2020.pdf.

Guardian Government Computing. 2012. *Government commits £10m to Open Data Institute.* [Online]. [Accessed 17 August 2020]. Available from: https://www.theguardian.com/government-computing-network/2012/may/23/open-data-institute-plans-published-cabinet-officemaincontent.

Guyon, I. and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of machine learning research.* **3**(Mar), pp. 1157–1182.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine learning.* **46**(1), pp. 389–422.

Harris, R. 1998. Considering (mis-) representation in geodemographics and lifestyles. *3rd International Conference on GeoComputation.*

Harris, R. 2001. The diversity of diversity: Is there still a place for small area classifications?[with response]. *Area.* **33**(3), pp. 329–336.

Harris, R., Johnston, R. and Burgess, S. 2007. Neighborhoods, ethnicity and school choice: developing a statistical framework for geodemographic analysis. *Population Research and Policy Review.* **26**(5-6), pp. 553–579.

Harris, R., Sleight, P. and Webber, R. 2005. *Geodemographics, Gis and Neighbourhood Targeting.* England: John Wiley & Sons.

Henson, R. K. and Roberts, J. K. 2006. Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological measurement.* **66**(3), pp. 393–416.

Hincks, S., Kingston, R., Webb, B. and Wong, C. 2018. A new geodemographic classification of commuting flows for england and wales. *International Journal of Geographical Information Science.* **32**(4), pp. 663–684.

HM Land Registry. 2016. *Guidance: How to access HM Land Registry Price Paid Data.* [Online]. [Accessed 12 June 2020]. Available from: https://www.gov.uk/guidance/about-the-price-paid-data.

Hogarty, K. Y., Kromrey, J. D., Ferron, J. M. and Hines, C. V. 2004. Selection of variables in exploratory factor analysis: An empirical comparison of a stepwise and traditional approach. *Psychometrika.* **69**(4), pp. 593–611.

Hunter, A. A. 1972. Factorial ecology: A critique and some suggestions. *Demography.* **9**(1), pp. 107–117.

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

Janson, C. G. 1980. Factorial social ecology: An attempt at summary and evaluation. *Annual Review of Sociology.* **6**(1), pp. 433–456.

Jolliffe, I. T. 1972. Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics).* **21**(2), pp. 160–173.

Kaiser, H. F. 1970. A second generation little jiffy. *Psychometrika.* **35**(4), pp. 401–415.

Karegowda, A. G., Jayaram, M. and Manjunath, A. 2010. Feature subset selection problem using wrapper approach in supervised learning. *International journal of Computer applications.* **1**(7), pp. 13–17.

Karimi, H. A. 2014. *Big Data Techniques and Technologies in Geoinformatics.* Florida: CRC Press.

Kline, P. 2014. *An easy guide to factor analysis.* New York: Routledge.

Knapp, T. R. and Swoyer, V. H. 1967. Some empirical results concerning the power of bartlett's test of the significance of a correlation matrix. *American Educational Research Journal.* **4**(1), pp. 13–17.

Kohavi, R. and John, G. H. 1997. Wrappers for feature subset selection. *Artificial intelligence.* **97**(1-2), pp. 273–324.

Kuhn, M. 2019. *The caret Package:20 Recursive Feature Elimination.* [Online]. [Accessed 4 January 2021]. Available from: http://topepo.github.io/caret/recursive-feature-elimination.html.

Langford, M. and Unwin, D. J. 1994. Generating and mapping population density surfaces within a geographical information system. *The Cartographic Journal.* **31**(1), pp. 21–26.

Lansley, G. 2016. Cars and socio-economics: understanding neighbourhood variations in car characteristics from administrative data. *Regional Studies, Regional Science.* **3**(1), pp. 264–285.

Lebowitz, M. D. 1977. A critical examination of factorial ecology and social area analysis for epidemiological research. *Journal of the Arizona Academy of Science.* **12**(2), pp. 86–90.

Lécuyer, M., Spahn, R., Vodrahalli, K., Geambasu, R. and Hsu, D. 2019. Privacy accounting and quality control in the sage differentially private ml platform. *Proceedings of the 27th ACM Symposium on Operating Systems Principles.* pp. 181–195.

Leeds Observatory. 2020a. *Population of Leeds.* [Online]. [Accessed 22 June 2020]. Available from: https://observatory.leeds.gov.uk/population/.

Leeds Observatory. 2020b. *Welcome to the Leeds Observatory.* [Online]. [Accessed 18 November 2020]. Available from: https://observatory.leeds.gov.uk/.

Leventhal, B. 2016. *Geodemographics for Marketers: Using Location Analysis for Research and Marketing.* London: Kogan Page Publishers.

Li, L. E., Chen, E., Hermann, J., Zhang, P. and Wang, L. 2017. Scaling machine learning as a service. *International Conference on Predictive Applications and APIs.* PMLR.

Liu, Y. and Cheng, T. 2018. Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science.* **16**(1), pp. 76–103.

Liu, Y., Singleton, A. D. and Arribas-Bel, D. 2019. A principal component analysis (pca)-based framework for automated variable selection in geodemographic classification. *Geospatial Information Science* **22**(4), pp. 251–264.

Local Authority Building Control. 2018. *What is the average house size in the UK?* [Online]. [Accessed 16 June 2020]. Available from: https://www.labc.co.uk/news/what-average-house-size-uk.

Local Government. 2016. *How domestic properties are assessed for Council Tax bands.* [Online]. [Accessed 12 June 2020]. Available from: https://www.gov.uk/guidance/understand-how-council-tax-bands-are-assessed.

Local Government Association. 2013. *Developing a customer classification tool: Guidance document for local authorities.* [Online]. [Accessed 11 January 2021]. Available from: https://www.local.gov.uk/sites/default/files/documents/hull-city-council-develop-4b8.pdf type=Web Page.

Lohr, S. 2015. *Data-ism: Inside the big data revolution* Simon and Schuster.

Lomax, N. 2020. *Household Mobility – Where and how far do we move?* [Online]. [Accessed 12 June 2020]. Available from: https://www.cdrc.ac.uk/household-mobility-where-and-how-far-do-we-move/.

Longley, P. 2005. Geographical information systems: A renaissance of geodemographics for public service delivery. *Progress in Human Geography.* **29**(1), pp. 57–63.

Longley, P. A. 2007. Some challenges to geodemographic analysis and their wider implications for the practice of giscience. *COMPUT ENVIRON URBAN.* **31**(6), pp. 617–622.

Longley, P. A. 2012. Geodemographics and the practices of geographic information science. *International Journal of Geographical Information Science.* **26**(12), pp. 2227–2237.

Longley, P. A. and Singleton, A. D. 2009a. Classification through consultation: Public views of the geography of the e-society. *International Journal of Geographical Information Science.* **23**(6), pp. 737–763.

Longley, P. A. and Singleton, A. D. 2009b. Linking social deprivation and digital exclusion in england. *Urban Studies.* **46**(7), pp. 1275–1298.

Longley, P. A., Webber, R. and Li, C. 2008. The uk geography of the e-society: a national classification. *Environment and Planning A.* **40**(2), pp. 362–382.

Malloy, L. 2016. *Geospatial data and the Internet of things.* [Online]. [Accessed 29 November 2016]. Available from: http://www.intermap.com/the-spatialist/2016/09/geospatial-data-and-the-internet-of-things.

Martin, D.. 2020. *Email conversation with Amanda Otley*, 12 november 2020.

Martin, D. and Bracken, I. 1991. Techniques for modelling population-related raster databases. *Environment and Planning A.* **23**(7), pp. 1069–1075.

Mathers, M. 2020. *Ministers using 'following the science' defence to justify decision-making during pandemic, says Prof Brian Cox.* [Online]. [Accessed 8 January 2021]. Available from: https://www.independent.co.uk/news/uk/politics/coronavirus-brian-cox-minister-follow-science-comments-a9520041.html.

Maugis, C., Celeux, G. and Martin-Magniette, M. L. 2009. Variable selection for clustering with gaussian mixture models. *Biometrics.* **65**(3), pp. 701–709.

Mayer-Schönberger, V. and Cukier, K. 2013. *Big data: A revolution that will transform how we live, work, and think.* New York: Houghton Mifflin Harcourt.

Ministry of Housing, Communities and Local Government. 2019. *Energy Performance of Buildings Certificates Statistical Release: Q1 2019: England and Wales.* [Online]. [Accessed 11 June 2020]. Available from: https://www.gov.uk/government/statistics/energy-performance-of-buildings-certificates-in-england-and-wales-2008-to-march-2019.

Ministry of Housing, Communities and Local Government. 2020. *Help to Buy Equity Loan Scheme -Total Equity Loans Equity Loans First Time Buyers.* [Online]. [Accessed 12 June 2020]. Available from: https://opendatacommunities.org/data/housing-market/help-to-buy/num-loans/loan-type.

Moon, G., Twigg, L., Jones, K., Aitken, G. and Taylor, J. 2019. The utility of geodemographic indicators in small area estimates of limiting long-term illness. *Social Science & Medicine.* **227**, pp. 47–55.

Murphy, S. and Smith, M. 2014. Geodemographic model variable selection spacial data mining of the 2011 irish census. In: *2014 IEEE International Advance Computing Conference (IACC)* pp. 613-622.

Oldroyd, R. A., Morris, M. A. and Birkin, M. 2020. Food safety vulnerability: Neighbourhood determinants of non-compliant establishments in england and wales. *Health & place.* **63**, pp. 1–11.

ONS. 2014. *The Census and Future Provision of Population Statistics in England and Wales: Report on the Public Consultation.* [Online]. [Accessed 9 June 2020]. Available from: https://www.ons.gov.uk/census/censustransformationprogramme/beyond2011censustran sformationprogramme/reportsandpublications.

ONS. 2016a. *Output areas: Introduction to Output Areas - the building block of Census geography.* [Online]. [Accessed 28 July 2020]. Available from: https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas.

ONS. 2016b. *Protecting confidentiality with statistical disclosure control.* [Online]. [Accessed 28 July 2020]. Available from: https://www.ons.gov.uk/census/2011census/howourcensusworks/howwetookthe2011census/howweplannedfordatadelivery/protectingconfidentialitywith-statisticaldisclosurecontrol.

ONS. 2018a. *Classification of Workplace Zones for the UK datasets.* [Online]. [Accessed 30 December 2020]. Available from: https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011workplacebasedareaclassification/classificationofworkplacezonesfortheukdatasets.

ONS. 2018b. *Research Outputs: An update on developing household statistics for an Administrative Data Census.* [Online]. [Accessed 10 June 2020]. Available from: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/householdandfamilies/ researchoutputsanupdateondevelopinghouseholdstatisticsforanadministrativedatacensus.

ONS. 2019. *Admin-based qualification statistics, feasibility research: England.* [Online]. [Accessed 10 June 2020]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/articles/adminbasedqualificationstatisticsfeasibilityresearchengland/2019-10-22.

ONS. 2020a. *Administrative Data Census Project.* [Online]. [Accessed 9 June 2020]. Available from: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject.

ONS. 2020b. *Beyond 2011.* [Online]. [Accessed 9 June 2020]. Available from: https://www.ons.gov.uk/census/censustran sformationprogramme/beyond2011censustransformationprogramme.

ONS. 2020c. *Census transformation.* [Online]. [Accessed 9 June 2020]. Available from: https://www.ons.gov.uk/census/censustransformationprogramme.

ONS. 2020d. *Postcode Headcounts and Household Estimates - 2011 Census - "Table 2". Nomisweb.* [Online]. [Accessed 22 June 2020]. Available from: https://www.nomisweb.co.uk/census/2011/postcode_headcounts_and_household_estimate.

ONS. 2021. *The census during the coronavirus pandemic.* [Online]. [Accessed 22 January 2021]. Available from: https://census.gov.uk/about-the-census/the-census-during-the-coronavirus-pandemic/.

ONS Open Geography Portal. 2020a. *ONS Postcode Directory (May 2020).* [Online]. [Accessed 22 January 2021]. Available from: https://geoportal.statistics.gov.uk/datasets/ons-postcode-directory-may-2020.

ONS Open Geography Portal. 2020b. *ONS UPRN Directory (May 2020).* [Online]. [Accessed 22 January 2021]. Available from: https://geoportal.statistics.gov.uk/datasets/68879b4d8da545a395a8bc8b95572e7d.

Open Data Institute. 2020. *About the ODI.* [Online]. [Accessed 17 August 2020]. Available from: https://theodi.org/about-the-odi/.

Openshaw, S. 2001. *Geodemographics.* [Online]. [Accessed 28 August 2020]. Available from: http://www.geog.leeds.ac.uk/presentations/98-3/sld001.htm.

Openshaw, S., Cullingford, D. and Gillard, A. 1980. A critique of the national classifications of opcs/prag. *The Town Planning Review.* **51**(4), pp. 421–439.

Otley, A., Newing, A., Addison, R. and Ridge, W. 2018. Community resilience: Identifying spatial variation to inform policies for elderly residents in leeds. GISRUK, 17 April, Leicester.

Palm, R. and Caruso, D. 1972. Factor labelling in factorial ecology. *Annals of the Association of American Geographers.* **62**(1), pp. 122–133.

Park, R. E. 1925. The city: Suggestions for the investigation of human behavior in the urban environment. In: Park, R. E., Burgess, E. W. and McKenzie, R. D. ed(s). *The City: Suggestions for the Investigation of Human Behavior in the Urban Environment.* Chicago: University of Chicago Press, pp. 1-46.

Park, R. E., Burgess, E. W. and McKenzie, R. D. 1925. *The City Chicago.* Chicago: University of Chicago Press.

Parker, S., Uprichard, E. and Burrows, R. 2007. Class places and place classes geodemographics and the spatialization of class. *Information, Communication & Society.* **10**(6), pp. 902–921.

Patel, J. 2020. The democratization of machine learning features. *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI).* pp. 136–141.

Petersen, J., Gibin, M., Longley, P., Mateos, P., Atkinson, P. and Ashby, D. 2010. Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *Journal of Geographical Systems.* **13**(2), pp. 173–192.

Powell, J., Tapp, A., Orme, J. and Farr, M. 2007. Primary care professionals and social marketing of health in neighbourhoods: a case study approach to identify, target and communicate with 'at risk' populations. *Primary Health Care Research & Development.* **8**(1), pp.

Public Accounts Committee. 2019. *Challenges in using data across government: Conclusions and recommendations.* [Online]. [Accessed 18 November 2020]. Available from: https://publications.parliament.uk/pa/cm201719/cmselect/cmpubacc/2492/249205.htm.

Quinn, J. A. 1940. The burgess zonal hypothesis and its critics. *American Sociological Review.* **5**(2), pp. 210–218.

Rees, P. H. 1971. Factorial ecology: an extended definition, survey, and critique of the field. *Economic Geography.* **47**(sup1), pp. 220–233.

Rees, P. H. 1972. Problems of classifying subareas within cities. *The City Classification Handbook.* pp. 256–330.

Reibel, M. 2011. Classification approaches in neighborhood research: Introduction and review. *Urban Geography.* **32**(3), pp. 305–316.

Revelle, W. 2020. *RDocumentation: fa.* [Online]. [Accessed 26 October 2020]. Available from: https://www.rdocumentation.org/packages/psych/versions/2.0.9/topics/fa.

Robinson, C. and Franklin, R. S. 2020. The sensor desert quandary: What does it mean (not) to count in the smart city? *Transactions of the Institute of British Geographers.* pp. pp. 1–17.

Robinson, W. S. 2009. Ecological correlations and the behavior of individuals. *International journal of epidemiology.* **38**(2), pp. 337–341.

Robson, B. T. and Robson, P. 1969. *Urban analysis: a study of city structure with special reference to Sunderland.* London: Cambridge University Press.

Roumpani, F., Maricevic, M. and Wilson, A. 2020. Data-driven modelling of public library infrastructure and usage in the united kingdom. pp. 285–308. In: Baker, D. and Ellis, L. ed(s). *Future Directions in Digital Information.*

Rummel, R. J. 1967. Understanding factor analysis. *Journal of conflict resolution.* **11**(4), pp. 444–480.

Samarasundera, E., Martin, D., Saxena, S. and Majeed, A. 2010. Socio-demographic data sources for monitoring locality health profiles and geographical planning of primary health care in the uk. *Primary Health Care Research & Development.* **11**(4), pp. 287–300.

Savage, M. and Burrows, R. 2007. The coming crisis of empirical sociology. *Sociology.* **41**(5), pp. 885–899.

Shevky, E. and Bell, W. 1955. *Social area analysis; theory, illustrative application and computational procedures.* California: Stanford University Press.

Shevky, E. and Williams, M.. 1949. The social areas of Los Angeles, analysis and typology. Berkeley, Pub. for the John Randolph Haynes and Dora Haynes Foundation by the Univ.

Singleton, A., Alexiou, A. and Savani, R. 2020. Mapping the geodemographics of digital inequality in great britain: An integration of machine learning into small area estimation. *Computers, Environment and Urban Systems.* **82**, pp. 2–20.

Singleton, A. D. 2010. The geodemographics of educational progression and their implications for widening participation in higher education. *Environment and Planning A.* **42**(11), pp. 2560–2580.

Singleton, A. D. 2014. *Geodemographics and the Internal Structure of Cities.* [Online]. [Accessed 27 June 2018]. Available from: https://www.youtube.com/watch?v=lslLujtqGlw.

Singleton, A. D. 2016a. Cities and context: The codification of small areas through geodemographic classification. In: Kitchin, R. and Perng, S.Y. ed(s). *Code and the City.* London: Routledge, pp. 215-235.

Singleton, A. D. 2016b. *London Output Area Classification.* [Online]. [Accessed 29 June 2020]. Available from: http://www.opengeodemographics.com/LOAC-section.

Singleton, A. D. and Arribas-Bel, D. 2019. Geographic data science *Geographical Analysis* pp. 1–15.

Singleton, A. D. and Longley, P. 2015. The internal structure of greater london: a comparison of national and regional geodemographic models. *Geo: Geography and Environment.* **2**(1), pp. 69–87.

Singleton, A. D. and Longley, P. A. 2009a. Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Papers in Regional Science.* **88**(3), pp. 643–666.

Singleton, A. D. and Longley, P. A. 2009b. Geodemographics, visualisation, and social networks in applied geography. *Applied Geography.* **29**(3), pp. 289–298.

Singleton, A. D. and Longley, P. A. 2019. Data infrastructure requirements for new geodemographic classifications: The example of london's workplace zones. *Applied Geography.* **109**, pp 1–9.

Singleton, A. D., Pavlis, M. and Longley, P. A. 2016. The stability of geodemographic cluster assignments over an intercensal period. *Journal of Geographical Systems.* **18**(2), pp. 97–123.

Singleton, A. D. and Spielman, S. E. 2014. The past, present and future of geodemographic research in the united states and united kingdom. *The Professional Geographer.* **66**(4), pp. 558–567.

Singleton, A. D., Wilson, A. and O'Brien, O. 2012. Geodemographics and spatial interaction: an integrated model for higher education. *Journal of Geographical Systems.* **14**(2), pp. 223–241.

Sleight, P. 2014. *Geodemographic Classification Systems - The New Breed.* [Online]. [Accessed 20 January 2021]. Available from: https://www.mrs.org.uk/blog/gkb/peter-sleight-geodemographic-classification-systems-the-new-breed.

Slingsby, A., Dykes, J. and Wood, J. 2011. Exploring uncertainty in geodemographics with interactive graphics. *IEEE Transactions on Visualization and Computer Graphics.* **17**(12), pp. 2545–2554.

Smart Leeds. 2020. *Smart Leeds.* [Online]. [Accessed 18 November 2020]. Available from: https://datamillnorth.org/smart-leeds/.

Smart Leeds 2021. *Smart Leeds commitments.* [Online]. [Accessed 27 January 2021]. Available from: https://datamillnorth.org/smart-leeds/our-commitments/.

Stanford-Tuck, V. 2020. *Where does the Zoopla Estimate data come from?* [Online]. [Accessed 16 November 2020]. Available from: https://help.zoopla.co.uk/hc/en-gb/articles/360005677897-Where-does-the-Zoopla-Estimate-data-come-from- type=Web Page.

Sweetser, F. L. 1965. Factorial ecology: Helsinki, 1960. *Demography.* **2**(1), pp. 372–385.

Swinney, P. and Carter, A. 2018. *The UK's rapid return to city centre living.* [Online]. [Accessed 22 June 2020]. Available from: https://www.bbc.co.uk/news/uk-44482291.

Tabachnick, B. G. and Fidell, L. S. 2013. *Using multivariate statistics.* [Online]. 6^{th} ed. Harlow: Pearson Education Limited. [Accessed 20 September 2020]. Available from: https://ebookcentral.proquest.com/lib/leeds/reader.action?docID=5173686.

Tate, P. 2018. *Acorn explained.* [Online]. [Accessed 3 June 2020]. Available from: https://www.caci.co.uk/blog/acorn-explained.

The National Archives. 2013. *The United Kingdom Report on the Re-use of Public Sector Information 2013.* [Online]. [Accessed 17 August 2020]. Available from: https://www.nationalarchives.gov.uk/documents/information-management/psi-report-2013.pdf.

Thomas, E., Serwicks, I. and Swinney, P. 2015. *Urban demographics: Why people live where they do.* Centre for Cities. [Online]. [Accessed 28 January 21]. Available from: https://www.centreforcities.org/reader/urban-demographics-2/why-do-people-live-where-they-do/.

Tobler, W. R. 1970. A computer movie simulating urban growth in the detroit region. *Economic geography* **46**(sup1), pp. 234–240.

Torun, A. 2016. *Integrating Geospatial Technologies: Reflections on Intergeo 2016.* [Online]. [Accessed 29 November 2016]. Available from: http://earsc.org/news/integrating-geospatial-technologies-reflections-on-intergeo-2016.

TransUnion. 2018. *Why CAMEO?* [Online]. [Accessed 19 August 2018]. Available from: https://cameodynamic.com/why-cameo.

TransUnion. 2020. *CAMEO: It's the next generation of segmentation.* [Online]. [Accessed 20 January 2021]. Available from: https://www.transunion.co.uk/resources/tu-uk/doc/products/resources/product-cameo-uk-as.pdf.

TransUnion. 2021. *CAMEO: Take audience segmentation to the next level for more targeted campaigns.* [Online]. [Accessed 20 January 2021]. Available from: https://www.transunion.co.uk/product/cameo.

Twa, M. D. 2019. Scientific integrity and the reproducibility crisis. *Optometry and Vision Science.* **96**(1), pp. 1–2.

Twigg, L. and Moon, G. 2002. Predicting small area health-related behaviour: a comparison of multilevel synthetic estimation and local survey data. *Social Science & Medicine.* **54**(6), pp. 931–937.

UK Data Service. 2018. *Boundary Data Selector.* [Online]. [Accessed 6 April 2018]. Available from: https://borders.ukdataservice.ac.uk/bds.html.

UK Government. 2020. *West Yorkshire devolution deal.* [Online]. [Accessed 8 June 2020]. Available from: https://www.gov.uk/government/publications/west-yorkshire-devolution-deal.

UK Government and Parliament. 2018. *Petition: Protect library services by ringfencing government funding for libraries.* [Online]. [Accessed 27 January 2021]. Available from: https://petition.parliament.uk/archived/petitions/228742.

Van Arsdol, M. D., Camilleri, S. F. and Schmid, C. F. 1958a. An application of the shevky social area indexes to a model of urban society. *Social Forces.* pp. 26–32.

Van Arsdol, M. D., Camilleri, S. F. and Schmid, C. F. 1958b. The generality of urban social area indexes. *American Sociological Review.* **23**(3), pp. 277–284.

Van Zyl, T. 2014. Machine learning on geospatial big data. In: Karimi, H.A. ed. *Big Data Techniques and Technologies in Geoinformatics.* Florida: CRC Press, pp. 133-148.

Vaughan, L. 2018. *Mapping society: the spatial dimensions of social cartography.* London: UCL Press.

Vickers, D. and Rees, P. 2007. Creating the uk national statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* **170**(2), pp. 379–403.

Vickers, D. and Rees, P. 2011. Ground-truthing geodemographics. *Applied Spatial Analysis and Policy.* **4**(1), pp. 3–21.

Vickers, D., Rees, P. and Birkin, M. 2005. Creating the national classification of census output areas: data, methods and results. Working Paper. School of Geography, University of Leeds.

Vickers, D. W. 2006. *Multi-level integrated classifications based on the 2001 census.* PhD thesis. University of Leeds.

Voas, D. and Williamson, P. 2001. The diversity of diversity: a critique of geodemographic classification. *Area.* **33**(1), pp. 63–76.

Webber, R. 1978. Making the most of the census for strategic analysis. *The Town Planning Review.* **49**(3), pp. 274–284.

Webber, R. 2004. Designing geodemographic classifications to meet contemporary business needs. *Interactive Marketing.* **5**(3), pp. 219–237.

Webber, R. 2007. Using names to segment customers by cultural, ethnic or religious origin. *Journal of Direct, Data and Digital Marketing Practice.* **8**(3), pp. 226–242.

Webber, R. and Burrows, R. 2018. *The predictive postcode: The geodemographic classification of British society.* London: Sage.

Williams, B., Onsman, A. and Brown, T. 2010. Exploratory factor analysis: A five-step guide for novices. *Australasian journal of paramedicine.* **8**(3), pp. 1–13.

Williamson, T., Ashby, D. I. and Webber, R. 2006. Classifying neighbourhoods for reassurance policing. *Policing & Society.* **16**(02), pp. 189–218.

Xing, E. P. and Karp, R. M. 2001. Cliff: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics.* **17**(suppl_1), pp. S306–S315.

Yong, A. G. and Pearce, S. 2013. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology.* **9**(2), pp. 79–94.

Appendix A - Leeds maps

A.1 Map of Leeds Local Authority boundary

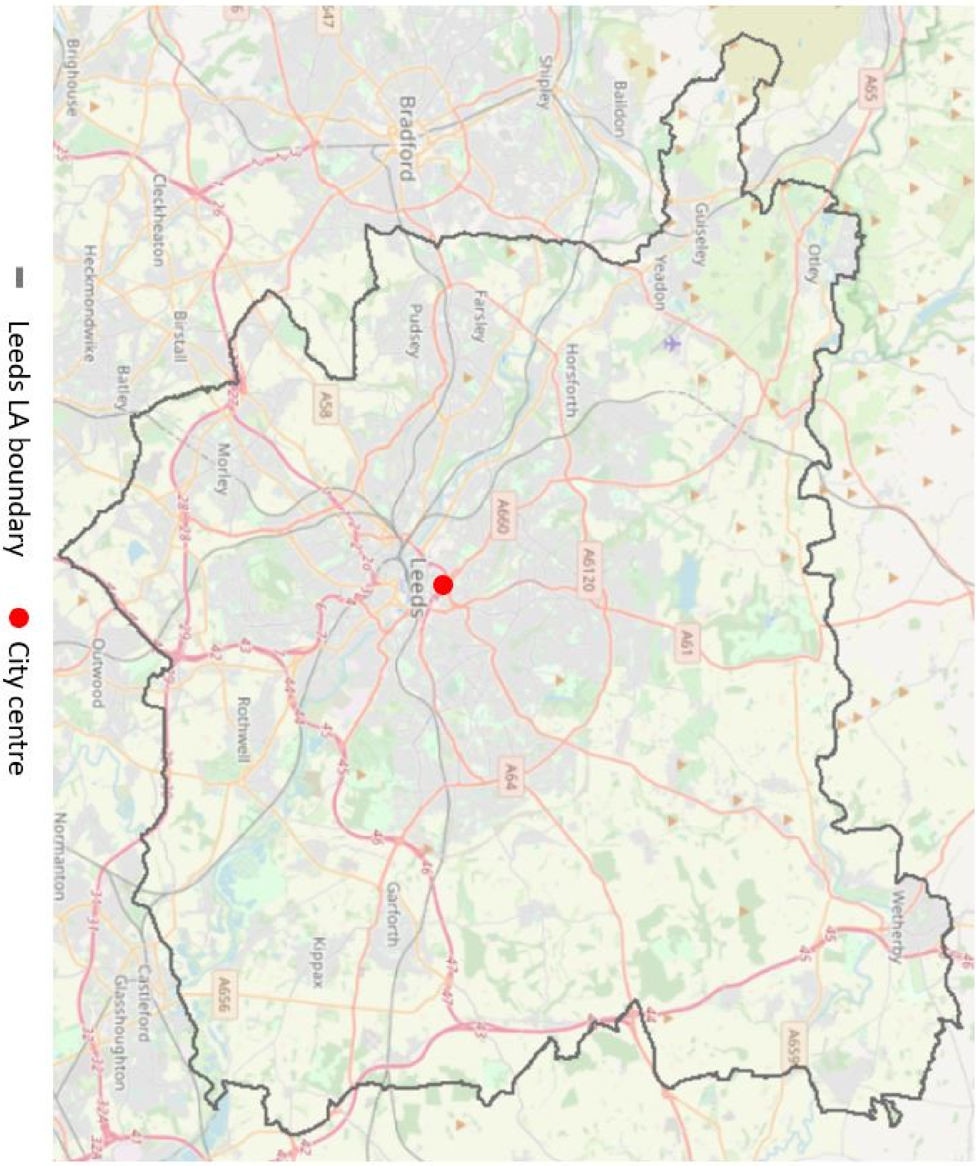


Figure A.1: Map of Leeds Local Authority boundary.

Appendix B - 2011 OAC input data

B.1 Final 60 input variables for 2011 OAC

The following table contains the final selection of 60 input census variables used in the development of the 2011 OAC (Gale et al., 2016).

Variable Code	**Variable Description**
k001	Persons aged 0 to 4
k002	Persons aged 5 to 14
k003	Persons aged 25 to 44
k004	Persons aged 45 to 64
k005	Persons aged 65 to 89
k006	Persons aged 90 and over
k007	Number of persons per hectare
k008	Persons living in a communal establishment
k009	Persons aged over 16 who are single
k010	Persons aged over 16 who are married or in a registered same-sex civil partnership
k011	Persons aged over 16 who are divorced or separated
k012	Persons who are white
k013	Persons who have mixed ethnicity or are from multiple ethnic groups
k014	Persons who are Asian/Asian British: Indian
k015	Persons who are Asian/Asian British: Pakistani
k016	Persons who are Asian/Asian British: Bangladeshi
k017	Persons who are Asian/Asian British: Chinese and Other
k018	Persons who are Black/African/Caribbean/Black British
k019	Persons who are Arab or from other ethnic groups
k020	Persons whose country of birth is the United Kingdom/Ireland
k021	Persons whose country of birth is in the old EU (pre 2004 accession countries)
k022	Persons whose country of birth is in the new EU (post 2004 accession countries)
k023	Main language is not English and cannot speak English well or at all
k024	Households with no children
k025	Households with non-dependant children
k026	Households with full-time students
k027	Households who live in a detached house/bungalow
k028	Households who live in a semi-detached house/bungalow
k029	Households who live in a terrace/end-terrace house
k030	Households who live in a flat
k031	Households who own/have shared ownership of property
k032	Households who are social renting
k033	Households who are private renting
k034	Occupancy room rating -1 or less
k035	Individuals day-to-day activities limited a lot or a little (Standardised Illness Ratio)
k036	Persons providing unpaid care
k037	Persons aged over 16 whose highest level of qualification is Level 1, Level 2 or Apprenticeship
k038	Persons aged over 16 whose highest level of qualification is Level 3 qualifications

k039	Persons aged over 16 whose highest level of qualification is Level 4 qualifications and above
k040	Persons aged over 16 who are schoolchildren or full-time students
k041	Households with two or more cars or vans
k042	Persons aged between 16-74 who use public transport to get to work
k043	Persons aged between 16-74 who use private transport to get to work
k044	Persons aged between 16-74 who walk, cycle or use an alternative method to get to work
k045	Persons aged between 16-74 who are unemployed
k046	Employed persons aged between 16-74 who work part-time
k047	Employed persons aged between 16-74 who work full-time
k048	Employed persons aged between 16-74 who work in the agriculture, forestry or fishing industries
k049	Employed persons aged between 16-74 who work in the mining, quarrying or construction industries
k050	Employed persons aged between 16-74 who work in the manufacturing industry
k051	Employed persons aged between 16-74 who work in the energy, water or air conditioning supply industries
k052	Employed persons aged between 16-74 who work in the wholesale and retail trade; repair of motor vehicles and motor cycles industries
k053	Employed persons aged between 16-74 who work in the transport or storage industries
k054	Employed persons aged between 16-74 who work in the accommodation or food service activities industries
k055	Employed persons aged between 16-74 who work in the information and communication or professional, scientific and technical activities industries
k056	Employed persons aged between 16-74 who work in the financial, insurance or real estate industries
k057	Employed persons aged between 16-74 who work in the administrative or support service activities industries
k058	Employed persons aged between 16-74 who work in the in public administration or defence; compulsory social security industries
k059	Employed persons aged between 16-74 who work in the education sector
k060	Employed persons aged between 16-74 who work in the human health and social work activities industries

Table B.1: 60 input census variables used to build the 2011 OAC.

Appendix C - Chapter 6 variable selection sets

C.1 Final input variables selected in Chapter 6 Case Study

The following table lists the 25 novel variables which are derived from the administrative data in Chapter 6. The table also includes a *variable type*, indicating whether the variable has been selected as a replacement for an existing census variable in the original set of census variables employed by the 2011 OAC listed in (Gale, 2014, p.475) ("Replacement"), to extend the census variables ("Extension"), or has been entirely discarded ("Discarded") (see Section 6.4 for the decision process).

Variable Description	**Variable Type**
Average property value of each OA	Extension
Turnover rate for homeowners in each OA	Extension
Rate of residential housing expansion in each OA	Extension
% of detached properties in each OA	Replacement
% of semi-detached properties in each OA	Replacement
% of terraced properties in each OA	Replacement
% of flats in each OA	Replacement
% of HMO properties in each OA	Extension
% of properties in each Council Tax band "A"	Extension
% of properties in each Council Tax band "B"	Extension
% of properties in each Council Tax band "C"	Extension
% of properties in each Council Tax band "D"	Extension
% of properties in each Council Tax band "E"	Extension
% of properties in each Council Tax band "F"	Extension
% of properties in each Council Tax band "G"	Extension
% of properties in each Council Tax band "H"	Extension
% of properties with either an "N" or "M" student exemption code in each OA (i.e. % of properties solely occupied by students)	Replacement
% of properties with an Owner Liable ("OL") exemption code in each OA which are not classed as HMOs in Council Tax data	Extension
% of social housing in each OA	Replacement
% of social housing which are classified as Sheltered Housing or Extra Care properties in each OA	Extension
% of social housing which are bedsits or 1/2 bedrooms in each OA	Extension
% of social housing which are 3 bedrooms in each OA	Extension
% of social housing which are 4 or more bedrooms in each OA	Extension
Average tenancy duration of social housing in each OA	Discarded
Average rent of social housing in each OA	Discarded

Table C.1: Final selected novel variables from the housing administrative data.

C.2 Final 131 input variables

The following table contains the final selection of 131 input census variables used in the development of the FALSOAC in Chapter 7 and the the FSLSOAC in Chapter 8.

Variable Code	Variable Description
L001	Males
L002	Females
L003	Lives in a communal establishment
L004	Density (number of persons per hectare)
L005	Age 0 to 4
L006	Age 5 to 14
L007	Age 15 to 19
L008	Age 20 to 24
L009	Age 25 to 44
L010	Age 45 to 64
L011	Age 65 to 89
L012	Age 90 and over
L013	Single (never married or never registered a same-sex civil partnership)
L014	Married or in a registered same-sex civil partnership
L015	Separated or Divorced
L016	Widowed or surviving partner from a same-sex civil partnership
L017	White
L018	Mixed/multiple ethnic group: Other Mixed
L019	Asian/Asian British: Indian
L020	Asian/Asian British: Pakistani
L021	Asian/Asian British: Bangladeshi
L022	Asian/Asian British: Chinese and Other
L023	Black/African/Caribbean/Black British: African
L024	Arab or Other Ethnic Groups
L025	Christian
L026	Other religion
L027	No religion
L028	Religion not stated
L029	United Kingdom or Ireland
L030	Other EU: Member countries in March 2001
L031	Other EU: Accession countries April 2001 to March 2011
L032	Other countries
L033	Main language is English or Main language not English: Can speak English very well
L034	Main language is not English: Can speak English well
L035	Main language is not English and cannot speak English well or at all
L036	Living in a couple: Married
L037	Living in a couple: Cohabiting (opposite-sex)
L038	Living in a couple: In a registered same-sex civil partnership or cohabiting (same-sex)
L039	Not living in a couple: Single (never married or never registered a same-sex civil partnership)
L040	Not living in a couple: Married or in a registered same-sex civil partnership
L041	Not living in a couple: Separated (but still legally married or still legally in a same-sex civil partnership)

L042	Not living in a couple: Divorced or formerly in a same-sex civil partnership which is now legally dissolved
L043	Not living in a couple: Widowed or surviving partner from a same-sex civil partnership
L044	One person household: Aged 65 and over
L045	One person household: Other
L046	One family only: All aged 65 and over
L047	One family only: Married, same-sex civil partnership or cohabiting couple: No children
L048	One family only: Married or same-sex civil partnership couple: Dependent children
L049	One family only: Married, same-sex civil partnership or cohabiting couple, or lone parent: All children non-dependent
L050	One family only: Cohabiting couple: Dependent children
L051	One family only: Lone parent: Dependent children
L052	Other household types: With dependent children
L053	Other household types: All full-time students
L054	Other household types: All aged 65 and over
L055	No adults in employment in household: With dependent children
L056	No adults in employment in household: No dependent children
L057	Lone parent in part-time employment: Total
L058	Lone parent in full-time employment: Total
L059	Lone parent not in employment: Total
L060	One person ethnic household
L061	All household members have the same ethnic group
L062	Different ethnic groups between the generations only
L063	Different ethnic groups within partnerships (whether or not different ethnic groups between generations)
L064	Any other combination of multiple ethnic groups
L065	Household spaces with at least one usual resident
L066	Household spaces with no usual residents
L067	Whole house or bungalow: Detached
L068	Whole house or bungalow: Semi-detached
L069	Whole house or bungalow: Terraced (including end-terrace)
L070	Flats
L071	Caravan or other mobile or temporary structure
L072	Owned and Shared Ownership
L073	Social rented
L074	Private rented
L075	Living rent free
L076	Occupancy rating (rooms) of +2 or more
L077	Occupancy rating (rooms) of +1
L078	Occupancy rating (rooms) of -1 or less
L079	Up to 0.5 persons per room
L080	Over 0.5 and up to 1.0 persons per room
L081	Over 1.0 and up to 1.5 persons per room
L082	Over 1.5 persons per room
L083	Day-to-day activities limited a lot or a little Standardised Illness Ratio
L084	Provides unpaid care
L085	No qualifications
L086	Highest level of qualification: Level 1, Level 2 or Apprenticeship

L087	Highest level of qualification: Level 3 qualifications
L088	Highest level of qualification: Level 4 qualifications and above
L089	Schoolchildren and full-time students: Age 16 and over
L090	No cars or vans in household
L091	1 car or van in household
L092	2 or more cars or vans in household
L093	Work mainly at or from home
L094	Public Transport
L095	Private Transport
L096	On foot, Bicycle or Other
L097	Economically active: Self-employed
L098	Economically active: Unemployed
L099	Economically active: Full-time student
L100	Economically inactive: Retired
L101	Economically inactive: Student (including full-time students)
L102	Economically inactive: Looking after home or family
L103	Economically inactive: Long-term sick or disabled
L104	Economically inactive: Other
L105	Unemployed: Never worked
L106	Long-term unemployed
L107	Part-time: 30 hours or less worked
L108	Full-time: 31 or more hours worked
L109	Agriculture, forestry and fishing
L110	Mining, quarrying or construction industries
L111	Manufacturing
L112	Energy, water or air conditioning supply industries
L113	Wholesale and retail trade; repair of motor vehicles and motor cycles
L114	Transport and storage
L115	Accommodation and food service activities
L116	Information and communication or professional, scientific and technical activities industries
L117	Financial, insurance or real estate industries
L118	Administrative and support service activities
L119	Public administration and defence; compulsory social security
L120	Education
L121	Human health and social work activities
L122	Other industry
L123	Managers, directors and senior officials
L124	Professional occupations
L125	Associate professional and technical occupations
L126	Administrative and secretarial occupations
L127	Skilled trades occupations
L128	Caring, leisure and other service occupations
L129	Sales and customer service occupations
L130	Process, plant and machine operatives
L131	Elementary occupations

Table C.2: 131 input census variables filtered from the 2011 OAC candidate variables in Chapter 7