Putting "Geo" into Geodemographics: Evaluating the performance of national classification systems within regional contexts

Thesis submitted in accordance with the requirements of the University of Liverpool for the Degree of Doctor in Philosophy by

Alexandros Alexiou

Department of Geography and Planning

School of Environmental Sciences

University of Liverpool

September 2016

Table of Contents

Abstract

Geodemographics is an academic field that engages in identifying socio-spatial patterns through the process of organizing areas, typically referred to as neighbourhoods, into categories or clusters that share similarities across multiple socio-economic attributes (Singleton and Longley, 2009). Geodemographics can thus provide a simplified measure of socio-spatial structure through discrete segmentation of geographic space. In nomothetic terms, the basis of the spatial aggregations is based on societal homophily, the tendency of people to associate themselves with similar people. In this sense, people who live close by are bound to have more in common than a random group of people.

While geodemographic analysis can be viewed as an established methodology, the simplistic nature of the theoretical framework along with the lack of a single global optimization function produces a lot of uncertainty regarding the success of national geodemographic classifications, i.e. whether they can actually provide good representations of socio-spatial patterns. A review of the relevant literature has shown that little has been done within geodemographic research in the last 30 years as a response to issues of classification uncertainty and system-wide accuracy (Openshaw et al., 1980; Twigg et al., 2000; Voas and Williamson, 2001; Petersen et al., 2011; Reibel and Regelson, 2011). Evaluation constrains are further enhanced by the lack of classification transparency, that would otherwise enable replication and modification which are necessary in order to advance the field (Longley, 2007; Fisher and Tate, 2015).

This Thesis focuses on the issue of system-wide accuracy, specifically whether national classification systems can capture spatial variation of socio-spatial patterns at a regional level. Arguably, classification methods are a function of scale; therefore, patterns that are important locally are not necessarily captured in a data-driven national taxonomy. In particular, methodological issues are raised when aggregations into categorical measures sweep away contextual differences between regions, so that final classifications assume that areas within the same cluster have the same underlying characteristics. With this ecological fallacy standard geodemographic classifications fail to incorporate near-geography effectively, and despite the term, geodemographics could be “aspatial”. As a response to this problem, regional classifications are developed in order to adequately accommodate local or regional structures that diverge from national patterns. Such an example is the London Output Area Classification (LOAC) (Singleton and Longley, 2015).

This Thesis tries to elucidate some of the inner workings of Geodemographics by systematically exploring the accuracy of national classification systems for various geographic scales. The main

aim of this research is firstly to compare the classification similarity between national and regional patterns, and secondly introduce a methodological extension to conventional geodemographic analysis that accounts for spatial contexts, thus assuming better correlations between places and social identity.

In order to provide a comprehensive approach to the issue, the Thesis initially provides a more in-depth review of the theoretical and methodological framework of geodemographic analysis. It demonstrates the evolution of geodemographics, from precursor studies aiming to measure socio-spatial segregation, to a contemporary exploratory and analytical tool with many applications. It also demonstrates the evolution of tools and techniques used in Geodemographics, since G.I.Science and the computational power currently being offered has made a variety of methods available. The Thesis provides a review of such methods, with an emphasis on clustering techniques that are typically used with such socio-economic, quantitative data. It also practically demonstrates the methodological framework of geodemographics through a bespoke classification regarding the build environment and morphology of British neighbourhoods.

Based on these methodological frameworks, the analysis explores the issue of accuracy vis-à-vis scale. A number of administrative and functional zones are used in order to delineate various geographic contexts. Comparisons are then carried out between a national classification, which acts as a baseline model, and a series of regional and local classifications at the UK level. The analysis uses arc cosine similarity to evaluate similarity levels between cluster centres and the Rand Index to evaluate a measure of cluster assignment, similar to spatial correspondence (Openshaw et al., 1980). In order to evaluate hundreds of regional contexts simultaneously, an automated process within the R programming language has been developed.

Results indicate considerable divergence from national socio-spatial patterns across the UK on a case by case basis. Exploration results showed that, excluding several large conurbations, middle-sized urban areas perform better, while smaller Local Authorities and rural towns score lower. Outcomes suggest several policy implications regarding the applications of geodemographics; areas that national classification seems to perform worse are the same areas that would benefit the most out of a national geodemographic system, considering they are more likely to lack resources and expertise to carry out classifications at their local level. Furthermore, economically lacking and remote areas are prospective targets of national socio-economic policies, and as such, discrepancies are seriously undermining the usefulness of national classification systems, as spatial identification might actually be misleading in regions where it is needed the most.

A second step of the analysis is the methodological extension to the traditional geodemographic methodology that accounts for spatial context within the clustering process. The methodological framework is based on Webber's (1980) response to national classification critics, suggesting that national classifications do not work locally because they operate on different attribute means and standard deviations. Based on this observation, geographic dependencies are built within attribute values by means of regional standardisation, enabling classifications to be more sensitive to local variation of attributes. In particular, the model introduces a geographic factor g that adjusts the level of impact of contextual geography to attribute values, for various levels of regional geography – Regions, Travel-to-Work Areas and Local Authority Districts. Model results for various level of g show that the intensity and nature of cluster transitions between neighbourhoods is highly cluster-dependant, while also suggesting that Regional classifications seem to outperform other contexts in terms of neighbourhood representation and cluster cohesion.

This research is not developed as a critique to Geodemographics, but rather tries to systematically evaluate certain aspects of classification methodology. Although results are of tentative nature, a model where attribute values are conjoined spatially can help mitigate scale effects. The limitations of the approach are mainly the selection of the extents of near-geography, i.e. the contextual geography used to standardise values, and the value of the g factor, which are both biased parameters and as such should reflect the theoretical rationale and purpose of the classification creator.

Acknowledgements

I would like to thank my supervisor, Prof. Alexander Singleton for his valuable contribution and feedback. I would like to thank him for encouraging my research and for allowing me to grow as a research scientist. I would also like to thank Dr. Paul Williamson and all the members of the Geographical Data Science Lab for their valuable feedback, comments and suggestions during my studies. I am very grateful for the vibrant and friendly environment that you have all offered me during the last three years, which almost made writing this PhD Thesis a pleasant experience.

I would also like to thank the ECRC and the North West Doctoral Training Centre for their financial and academic support they have given me that made this research possible.

A special thanks goes to my family for the huge support and encouragement they have given me, my parents, my brother and particularly my wife, Giota, for the patience she has shown during my studies. This Thesis is dedicated to my daughter, Fay, which I love very much.

List of Figures

List of Tables

Acronyms and Abbreviations

ACS	Angular Cosine Similarity
AZP	Automatic Zoning Procedure
CDRC	Consumer Data Retail Centre
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DZ	Data Zones
ED	Enumeration District
IC	Internal Cohesion
IMD	Index of Multiple Deprivation
IUC	Internet Usage Classification
LAD	Local Authority District
LOAC	London Output Area Classification
LSOA	Lower Super Output Area
MAUP	Modifiable Area Unit Problem
MODUM	Multidimensional Open Data for Urban Morphology
OA	Output Area
OAC	Output Area Classification
ODbL	Open Database License
ONS	Office for National Statistics
OS	Ordnance Survey
PAM	Partitioning Around Medoids
SED	Squared Euclidean Distance
SOA	Super Output Areas
SOM	Self-Organizing Map
SSD	Sum of Squared Distance
SSE	Sum of Squared Error
TTWA	Travel-to-work Area
WCSS	Within-Cluster Sum of Squares

Chapter 1. Introduction

1.1. Introduction to the research problem

Geodemographics is an academic field that engages in identifying socio-spatial patterns through the process of organizing areas, typically referred to as *neighbourhoods*, into categories or clusters that share similarities across multiple attributes of people and places in which they live (Singleton and Longley, 2009). Geodemographics offer an established methodology that can provide a simplified measure of socio-spatial structure through discrete segmentation of geographic space. Such classifications have a legacy in demonstrating their utility over a range of public and private sector applications (Longley, 2005; Longley and Goodchild, 2008; Reibel, 2011; Singleton and Spielman, 2013).

Geodemographic classifications are typically created by applying a clustering algorithm on multidimensional socio-economic variables. This methodological framework can capture a wide set of neighbourhood attributes, taking advantage of a plethora of public or proprietary data to generate neighbourhood profiles (Harris et al., 2005). By aggregating people into a zonal area, there is a notion that there is some degree of homogeneity in characteristics, behaviours or attitudes, such as love for gardening, TV viewing choices, spending habits, hiking, etc. (Webber and Farr, 2001).

This methodological framework however lacks a solid theory; the basis of the spatial aggregations is loosely based on societal homophily, the tendency of people to associate themselves with similar people. People who live close by (i.e. in the same neighbourhood) are bound to have more in common than a random group of people. Geodemographic methodology has been characterized as simplistic and ambiguous (Voas and Williamson, 2001). While it is convenient to use this notion on the basis of within-neighbourhood aggregations, there is no definite magnitude of the spatial effect of homophily; in most cases, the concepts of "*small-area*" or "*neighbourhood*" are defined arbitrarily. Such neighbourhood effects can very well expand much further than the zonal area that is selected for the aggregations. Furthermore, information about which of these attributes are influenced by neighbourhood effects is very limited.

A key issue is that conventional classification methodologies lack any explicit information on the geographic context of an area. The central *"control by aggregation"* concept of geodemographics (Webber and Farr, 2001) has only been applied to attributes within the

clustering process, and not to the geography of areas. Conventional geodemographic classifications have no input regarding the location of neighbourhoods. As such, clustering algorithms account only for similarities in the attribute space; and areas are essentially treated as independent from one another. The traditional "aspatial" approach has a number of implications when generating profiles. Arguably, aggregations into categorical measures sweep away contextual differences between proximal zones; as such, the final classifications assume that areas within the same cluster have the same underlying characteristics. This type of ecological fallacy raises methodological questions regarding the accuracy of geo-classifications, given the inherent loss of within-cluster variation due to the aggregation process.

The shortcomings in the traditional geodemographic methodology can be associated with the issue of scale. Conventional classification methods are a function of scale; therefore, patterns that are important locally are not necessarily captured in a data-driven national taxonomy (Reibel, 2011). Singleton and Longley (2015) claim that there are strong a priori reasons to anticipate that differences between regions will impede the utility of national classifications. Such issues have drawn the attention of various academics (Openshaw et al., 1980; Petersen et al., 2011; Reibel and Regelson, 2011). There is a longstanding debate originating in the earliest of UK classifications about the effects of geographic scale in emergent clusters and whether comparisons between local and national classifications are nonsensical, because classifications are effectively built for different purposes (Openshaw et al., 1980; and Webber, 1980).

One practical example of a national and regional classification comparison is the regional London Output Area Classification (LOAC) (Singleton and Longley, 2015). LOAC was developed due to the criticism towards national classifications that do not adequately accommodate local or regional structures that diverge from national patterns, such as the Output Area Classification (OAC). In this instance, a comparison between London OAC and LOAC showed that although ethnicity patterns might be a defining attribute in OAC, within London, these attributes are far less pronounced. Furthermore, central core areas of London were shown to deviate significantly from national patterns, with a more disaggregate series of clusters both reflecting particular built environment characteristics and also being less marked by concentrations of certain ethnic groups. The authors claim that although the uniqueness of London offers a solid case as to why socio-spatial patterns are so distinctive there, it is not necessary limited to that area. The Scottish countryside for instance may be another example.

The above research highlighted the fact that there is significant merit to consideration of geodemographic structure at a regional level, taking into account that national models can smooth away key characteristics of internal socio-spatial structure. On the other hand, it could be

argued that a regional approach to classifications may diminish their practical benefits associated with operability (Webber and Farr, 2001). From the very beginning, Webber (1980) states that the value of national classification systems is compelling because not all organizations have resources to carry out and interpret classifications. Furthermore, it would be extremely difficult to compare locations between different classifications; it would be difficult for central government, for instance, to compare deprived neighbourhoods and make policy arrangements using individual local classifications when none of which are performed on a common basis. The same would hold true in the private sector, as the capabilities of companies to plan national branch development or conduct national surveys would be seriously compromised.

These theoretical and methodological limitations are undermining the success of geodemographic classifications. For marketing-related applications of geodemographics, a lack of local sensitivity may have fiscal implications, such as a reduced uptake of a product or service. However, in public sector uses, the consequences may be more severe, with mistargeting having potential implications on life chances, health and wellbeing.

Arguably, there is little evidence on whether a national classification can provide a reasonable description of the differences within, as well as between, regions of the country. This Thesis tries to elucidate some of the inner workings of Geodemographics by systematically exploring the similarity between national and regional classifications across the UK. In essence, it tries to expand on the hypotheses made by previous researchers regarding the level of divergence between regional and national classifications systems (Openshaw et al., 1980; Singleton and Longley, 2015). The Thesis makes a unique contribution to the area of geodemographics by introducing an extension to the traditional geodemographic model that incorporates spatial relationships of attribute values within regions, thus assuming better similarity between national and regional classification systems.

1.2. Research Aims

This research sets out to firstly determine how national classifications perform regionally and secondly explore how information about an area's locality can be captured and utilized within a geodemographic framework. The latter aim of this research is to produce a new methodological framework of geodemographics that is more sensitive to local variation of socio-spatial patterns, while also benefiting from the advantages of national classification systems. As such, the objectives of this research are:

- *A systematic evaluation of the level of agreement between national and regional classifications.*
- *The development of an extension to traditional geodemographic methodology that incorporates spatial context and relationships between small area geography.*

The main hypothesis of this approach is that space is infinitely complex, and as such there are unobserved variables unaccounted for in a spatial classification. By assuming some level of spatial dependency between proximal zones, there is a notion of control over conditions that are not necessarily taken into account in the classification process. These conditions will tend to be more similar locally, as suggested by Tobler's first Law of Geography: *"everything is related to everything else, but near things are more related than distant things"* (Tobler, 1970, p. 236). Areas within the same geographic context will therefore tend to be more similar that those further apart.

As a practical example of this approach, consider the attributes of home and car ownership. A family living in London has a much lower propensity to have a car or own a home than the rest of the country, even if its socio-economic profile would suggest otherwise, because of restrictions and limitations that are based on location. In this sense, clustering together a family living in an owned, detached house with two cars in Rural Scotland with a family that owns a detached house and two cars in central London is fundamentally questionable. If more variables regarding e.g. cost of living, housing prices, availability of parking space etc. where to be included in the classification, these two instances would belong to different socio-economic clusters. However, this kind of information may be difficult to acquire. By using some notion of similarity within geographic contexts, there is actually some control over these unobserved variables, similarly to the "*control by aggregation*" notion that is used to aggregate households into neighbourhoods.

The concept that is introduced to describe this aspect of geo-classifications is defined here as "*geographic sensitivity*", and measures the degree of influence of "*near-geography*" to the overall similarity between neighbourhoods. A geodemographic classification with high geographic sensitivity will thus tend to cluster together proximal areas more than a classification with low geographic sensitivity.

Within this context, this Thesis initially establishes how "*near-geography*" or, in this sense, "*geographic context*" can be defined, and explores the extent of classification differences between them. In order to accomplish such a task, the research questions that need to be answered are:

1. *How can the geographic context of an area be delineated?*

2. *How can comparisons be drawn between national and local classifications?*
3. *To what extent does contextual geography impact classification outcomes?*

The impact of "near-geography" is difficult to measure, especially since the extents of contextual geography cannot be comprehensively defined. Regional and local patterns may differentiate as a result of socio-economic conditions, historical backgrounds or specific features of the built and physical environment. The effects of such conditions can be very localized or affect much wider regions. The analysis focuses on the type of impacts that affect wider regions. These are explored through a series of comparisons between regional and national classifications, where regional refers to various geographic scales. The analysis also does not assume any prior reasoning regarding specific geographic zones that could present divergent patterns, as was the case with LOAC, hence the analysis takes place across the UK.

The definitions of regions or geographic contexts draws upon existing research on the Modifiable Area Unit Problem (MAUP), i.e. the effects of aggregation scale and zonal shape in the analysis of spatial variation (Fotheringham and Wong, 1991). Optimally, regions should represent some level of the internal organization of communities. However, due to the nature of data availability, it would be next to impossible to decouple completely any level of administrative or Census zonal geography from the examined contexts. For this purpose, besides the UK classification (which will be henceforth referred to as the *national classification*), three more levels are considered: Regional (formerly known as the Government Office Regions), Local Authority Districts (LADs) and Travel-to-work Areas (TTWAs). Regions, LADs and TTWAs are all non-overlapping contiguous areas covering the whole of the UK. TTWAs are defined to approximate self-contained local labour market areas, where the majority of an area's resident workforce work and live, so they have some degree of territorial cohesion (ONS, 2015a). Their sizes vary, but generally they lie in between the Regional and LAD scale. They also have the advantages of being consistent across the UK and fit existing lower level administration geographies.

The classification methodology is based on the publicly available 2011 Output Area Classification from the Office for National Statistics (ONS), which serves as a baseline model, with a few modifications. The exploration draws upon existing literature in order to construct a methodological framework to compare classification outcomes, with an emphasis on the spatial variation of the level of agreement. In particular, the exploration quantifies the differences across all clustering dimensions, i.e. cluster membership and cluster formation. In order to perform the comparisons among the baseline and numerous regional classifications, an automated process has been programmatically defined in the *R* programming language to produce systematic results.

The second aim of this analysis is to provide a theoretical and methodological framework on how to incorporate geographic context into geodemographic analysis. The extension to the geodemographic model would allow attribute values to be adjusted to reflect underlying conditions of contextual areas, thus incorporating some level of spatial dependency between neighbourhoods within the same geographic context. In so doing no claim is made that this type of methodology is "better" than the conventional, only that one the proposed model would be more sensitive to local spatial patterns, and as such would perform sufficiently well on local policy applications.

1.3. Thesis Structure

As far as an outline of this Thesis is concerned, the next two chapters provide a comprehensive review of the current theoretical and methodological framework of geodemographic analysis. The second Chapter aims to set the stage of geodemographic analysis by looking at the drivers and utility of such analyses inside and outside academic settings. It provides a literature review of socio-spatial segregation studies, urban sociology and spatial analysis that demonstrate the historical reasons behind the development of geodemographics. It provides an account of relevant research since the early 20th century, to factorial ecologies and multivariate analyses in the 1970s. It continues with the development of geodemographics by briefly discussing the geodemographic theory and illustrating its value through some of the work carried out within its various applications; from the initial deprivation studies to the private sector applications that dominated the 1980's and to the public sector renaissance during the last few decades.

Before this Thesis continues forward to addressing the research questions, it is essential to provide a more in-depth analysis of clustering methodologies, which are the main analytical framework of geodemographics. Therefore, the third chapter provides the framework of exploratory analysis and data mining as well as some key definitions and concepts in the arguably fragmented literature on the topic. While emphasis is placed on quantitative unsupervised clustering algorithms, such as the *K*-means, it tries to provides a comprehensive review of quantitative clustering methods, including cluster types, similarity measures and validation and inference procedures. These offer the background in order to understand the complexities associated with classification comparisons and similarity measures in multidimensional space which will be used extensively in the following Chapters.

The fourth Chapter provides a review of the methods and techniques that are used within geodemographic analysis. It draws upon available data structures, such as the decennial Census of the population, local government, public databases and generally Spatial Data Infrastructures in the UK, while discussing the limitations of such data sources within academic research. The second part of the chapter gives a comprehensive summary of the geodemographic analysis, how it is carried out, how it is evaluated and what are the strengths and weaknesses. Finally, it provides, through a case study of Liverpool, an example of the methodological steps needed to construct a conventional geodemographic classification, from data preparation to cluster analysis and interpretation, using the *K*-means algorithm presented in the previous Chapter.

Although the selection of a clustering algorithm is largely intuitive, there are certain limitations of the traditional *K*-means approach. Chapter 5 tries to provide an alternative approach to producing geodemographic classifications, while also provide a practical example of a bespoke classification. It describes the creation of small-area measures of physical and built environment morphology and builds a typology of neighbourhoods using Self-Organizing Maps (Alexiou and Singleton, 2016). The resulting classification, named Multidimensional Open Data for Urban Morphology (MODUM) Classification, is built for England and Wales and is based exclusively on physical and built environment attributes to construct the typology. In order to capture morphological aspects of neighbourhoods, the proposed methodology utilizes adjacent and proximal effects of built environment features in relation to building units. This approach not only demonstrates new ways of capturing data at high granularity through geocomputation, but also ensures that, through an open data approach, results are reproducible and updatable. Finally, an example of a classification comparison is made by looking at the relationships between the MODUM typology and the socio-economic typology produced by the 2011 OAC. The comparison presented acts as the basis for all the comparisons that are carried out between regional and national classifications in the following Chapters.

Chapter 6 sets out to address the first research question by providing a systematic evaluation of the level of agreement between national and regional classifications. It draws upon the methodology described in Chapter 4 and 5 and provides practical evidence about the theoretical rationale, the selection of data and the geographical contexts used in the creation of regional classifications. The geodemographic analysis follows that of the 2011 OAC and the aim of the exploration is to measure, ceteris paribus, the similarity between regional and national scale classifications. The "*national*" and "*regional*" descriptors are used intuitively, with national referring to the UK context and regional to any other subset geographies, i.e. Region, TTWA and LAD scale. Instead of the qualitative way (e.g. Openshaw et al., 1980) initial correspondence

between all classifications is determined through an automated process using the Arc Cosine Similarity measure, described in Chapter 3. Once cluster pairs are determined, the approach described compares regional to national classification both on the basis of mean attribute values of the resulting clusters (defined as *attribute fit*) as well as cluster memberships of neighbourhoods at various geographic scales (defined as *spatial fit*).

Comparison outcomes are illustrated through several examples for every level of regional geography. This includes comparisons of radial plots, cluster vector similarity, and mapping results for visual interpretation. Aggregate results of the spatial fit between national and regional classifications are demonstrated through cross-tabulation tables and the Rand Index. Results indicate considerable spatial variation in the homogeneity of socio-spatial patterns across the UK. A key outcome of the exploration is that the more unique the distribution of attribute vales within a region, the more divergent that area is from national socio-spatial patterns. This relates to how absolute attribute values compared to relative values within a contextual geography represent the nature of the neighbourhood.

Considering the results of the exploration, a methodological extension to the traditional geodemographic methodology that accounts for spatial context within the clustering process is developed and presented in Chapter 7. The analysis carried out tests the hypothesis that the amount of information that can be retrieved from an attribute value at a particular area is dependent on the area's locality. The methodological framework is based on Webber's (1980) response to national classification critics, suggesting that national classifications do not work locally because they operate on different attribute means and standard deviations. Based on this observation, geographic dependencies are built within attribute values by means of regional standardisation, enabling classifications to be more sensitive to local variation of attributes. In particular, the model introduces a geographic factor g that adjusts the level of impact of contextual geography to attribute values, for various levels of regional geography. Model results for various level of g show the intensity and nature of cluster transitions between neighbourhoods. The analysis concludes with a visual illustration of geodemographic models for various levels of g, and an evaluation of their performance against the 2011 OAC using an internal clustering criterion (Chapter 3), which suggest that Regional classifications outperform other contexts in terms of neighbourhood representation and cluster cohesion.

The last chapter consolidates the findings of this research. Chapter 8 critically reviews the contributions that have been made in the field, emphasizing on how certain limitations and shortcomings were addressed. It summarizes the exploration key outputs and outlines the strengths and weaknesses of the model described. Although results are of tentative nature, a

model where attribute values are conjoined spatially can help mitigate scale effects. The limitations of the approach are mainly the selection of the extents of near-geography, i.e. the contextual geography used to standardise values, and the value of the g factor, which are both biased parameters and as such should reflect the theoretical rationale and purpose of the classification creator. The thesis concludes with general discussion about the possible applications and future directions of this model within geodemographic research.

Finally, this Thesis includes two Appendixes. Appendix I contains selected *R* code that has been used in the analysis. It is divided into parts that are referenced throughout the research. The second Appendix II contains a list of publications by the author.

Chapter 2. Geodemographic Theory and Applications

2.1. Introduction

Geodemographics is a field of quantitative geography that engages into the classification of populations into discrete classes based on the socio-economic and built environment characteristics of small-area geography. Simply put, geodemographics is the "*analysis of people by where they live*" (Sleight, 1997, p. 16). A geodemographic analysis is essentially a data reduction methodology that aggregates populations, so that correlations between sub-populations can be drawn upon with ease. It involves the process of producing key statistics of a particular area or household, on the basis of the characteristics of its residents.

The inferential nature of the aggregations relies on the notion of societal homophily, or in other words the "*birds of a feather flock together*" phenomenon. People who live close by (i.e. in the same neighbourhood) are assumed to have more in common than a random group of people. This is a fundamental and generally established axiom in human geography, commonly known as Tobler's first law of geography: *"everything is related to everything else, but near things are more related than distant things"* (Tobler, 1970, p. 236). Although that geodemographics have evolved considerably over the years (Singleton and Spielman, 2013), its conceptual background is still wedded to the principle that people tend to align themselves with the behaviour and aspirations of the local communities in which they live.

These common attributes can relate to a plethora of human aspects, from demographic and housing to health characteristics and consumer behaviour. In contrast to other multivariate analyses, the output classification does not rely on a single measure or index, as for example the Index of Multiple Deprivation (IMD), but rather on a qualitative description that portrays the attributes of each cluster. Geodemographic research can focus on a variety of such features, depending on the purpose of the analysis, and produce befitting classifications.

Such classifications have demonstrated utility over a range of public and private sector applications (Longley, 2005; Longley and Goodchild, 2008; Reibel, 2011; Singleton and Spielman, 2013). Geodemographic applications were initially developed as a strategy to analyse and systematically document socio-spatial segregation. The associated data reduction methods were established in the 1970s (Webber, 1978), although a wider review and interpretation would extend right back to the 'human ecology' studies from the Chicago School of Sociology in the

1920s (Burgess, 1925), social area analysis in the 1950s (Shevky and Bell, 1955) and the factorial ecologies of the 1970s (Janson, 1980).

This Chapter provides the framework of the evolution of geodemographics in more detail, from early geodemographic precursors to multivariate analyses and early classifications. It continues with a detailed review of the first instances of geodemographic research, its main drivers, as well as the strengths and weaknesses of the analytical framework. In order to demonstrate their value, the chapter concludes with the illustration of geodemographic applications across various scientific fields.

2.2. Early Geodemographic precursors

Geodemographics focus on the analysis of distribution patterns of various populations across geographic space. Historically, various terms have been proposed to describe the partitioning of urban space and an equally great number of empirical methods have been introduced to analyse and understand this phenomenon (White, 1987; Wilkinson and Pickett, 2010; Dorling, 2014; Catney, 2014). One of the key concepts of this partitioning is socio-spatial segregation, a social phenomenon where population groups are physically separated based on various demographic factors, primarily considered in terms of ethnicity, religion and income status.

For instance, by analysing segregation at the neighbourhood level, Schelling (1971) identified "tipping points" as part of the natural processes regarding neighbourhood evolution and ethnic diversity. His study models how individual preferences regarding neighbours can lead to ethnic segregation. He introduced a predefined "tolerance threshold" about the fraction of similar residents in a neighbourhood, according to which individuals tend to relocate. Ongoing research postulates that the organisation and emergence of segregation is typically very complex in nature; for example, there is controversy about what are the measurable dimensions of the phenomenon (Massey and Denton 1988; Reardon and O'Sullivan, 2004) and whether there are specific forces regarding the processes behind it (Atkinson and Flint, 2004). These issues still remain difficult to address, as they involve method and interpretation, which have always been problematic in the field (Lloyd, 2014).

From the early 1900s onwards, researchers tried to systematically document spatial segregation and establish a series of general principles about the internal spatial and social structure of cities, commonly motivated by the ill effects of residential segregation of the poor and ethnic minorities. Residential segregation is evident when "neighbourhoods" have different

demographic contexts as a result of historic, cultural or socio-economic factors. There is an abundance of literature on this topic, particularly its ethnic dimension, generating a legacy of applications aimed at the identification of patterns and measurement of socio-spatial differentiation (Nightingale, 2012).

Within the UK specifically, concerns regarding the conditions of residential accommodation were widely seen as the main aspect of inequality, and some of the earliest work regarded mapping housing conditions at the neighbourhood level, for instance, Thomas R. Marr's work on the classification of housing conditions in Manchester and Salford (Fig. 2.1).

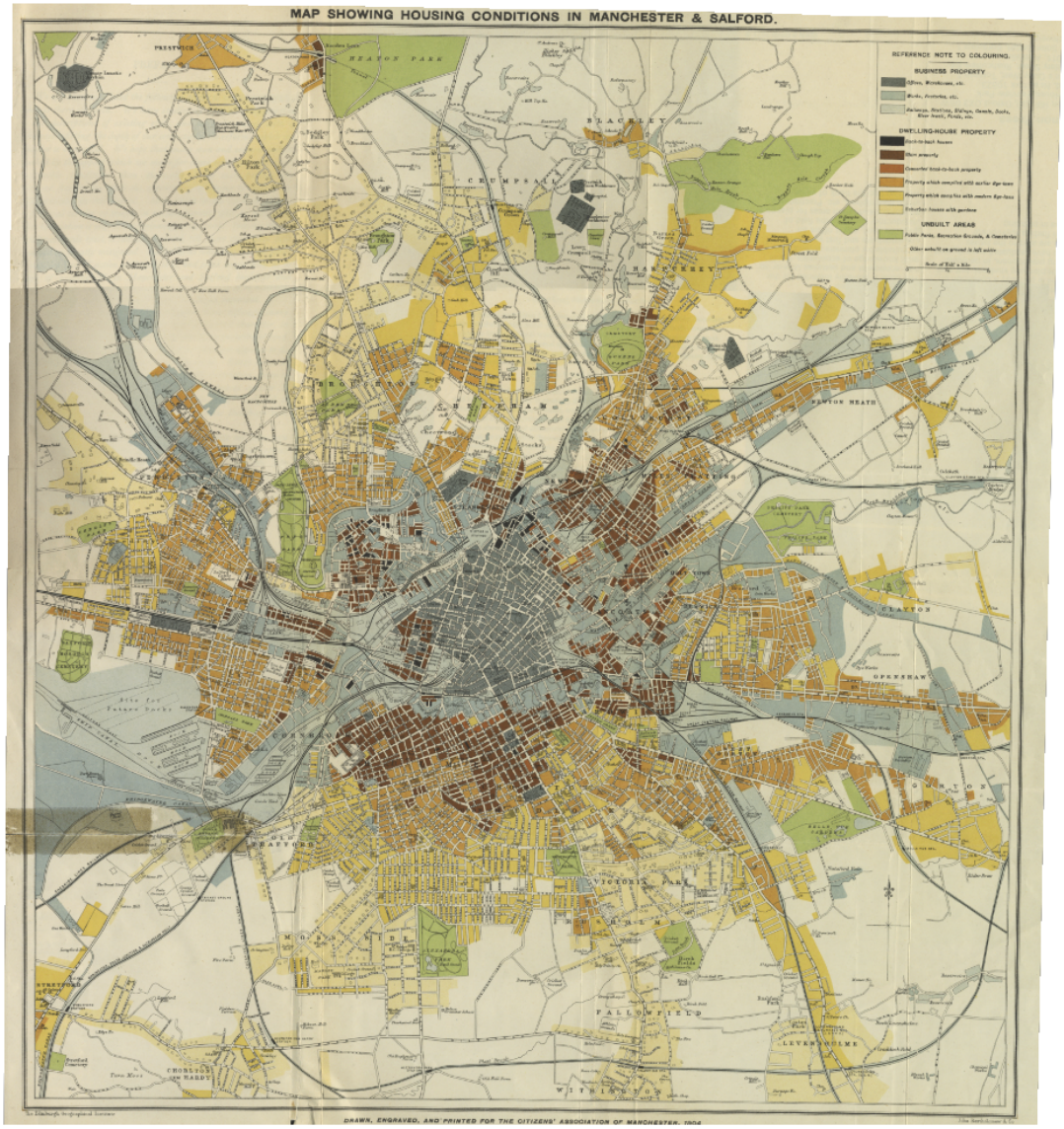

Figure 2.1 Thomas R. Marr's map of housing conditions in Manchester and Salford for the Edinburgh Geographical Institute for the Citizens' Association of Manchester, 1904 (source: <http://manchester.publicprofiler.org/marr/>).

Charles Booth's poverty maps were one of the first large-scale attempts to map the socio-spatial structure of London (Fried and Elman, 1969). His published work *"Life and Labour of the People in London"* between 1889 to 1903 which was compiled together with his associates, is a well-organized collaborative project to analyse and describe the Trades of London associated with poverty through a number of distinct topics, particularly women's occupation and ethnicity. Charles Booth was also one of the first to produce detailed maps of the poverty in London, maps that were colour-coded in accordance to socio-economic class of its inhabitants (O'Day and Englander, 1993) (Table 2. 1).

Table 2.1 Booth's Colour Coding Table for his maps of poverty. The colours represent the general complexion of the street in socio-economic terms. Purple and Pink streets include representatives of several classes (recreated from O'Day and Englander, 1993, p. 47.)

Class*	Description		Map Colour for Streets	
A	The lowest class of occasional labourers, loafers and semi-criminals		Black	
B	Casual earnings: 'very poor' (below 18s. Per week for a moderate family)		Dark Blue	
C	Intermittent earnings	Together 'the poor' between 18s. and 21s. per week for a moderate family	Light blue	Mixed
D	Small Regular earnings			Purple
E	Regular standard earnings	Above the line of poverty	Pink	
F	Higher class, labour	Fairly comfortable good ordinary earnings		
G	Lower middle class	Well-to-do middle class	Red	
H	Upper middle class	Wealthy	Yellow	

* Socio-Economic Class of People

One of his most famous maps, "*Descriptive Map of London Poverty*" (Fig. 2.2), was one of the first attempts to produce geographically referenced yet highly granular classifications of the population of inner London, on a building by building basis. The classification scheme outcomes presented a wide diversity of social patterns across very small geographical areas.

Booth's mapping of socio-spatial patterns influenced a number of sociologists collectively known as the School of Chicago (Kemper, 2006). In the late 1920s, Ernest W. Burgess and Robert

E. Park from the Chicago School of sociology constructed one of the first comprehensive models of urban socio-spatial structure, known as the *concentric zone theory* (Burgess, 1925). The model was based on a human ecology approach through studying the metropolitan city of Chicago, and utilising the then recently introduced Census of the Population, alongside extensive fieldwork and map-making (Burgess, 1964).

Figure 2.2 An example of a mapped area of inner London from the "Maps Descriptive of London Poverty 1898-99" by Charles Booth. A legend explaining the classification colours can be found in Table 2.1. The twelve maps cover an area of London from Hammersmith in the west, to Greenwich in the east, and from Hampstead in the north to Clapham in the south (source: London School of Economics and Political Science, Charles booth Online Archive <available at http://booth.lse.ac.uk/>).

A key feature in this classical ecological study was the assumption of "natural areas" as the basic unit of analysis, and was inspired by the Darwinian evolution hypothesis that affected natural ecosystems (Park, 1936). Their model was specified for the city of Chicago and featured five main concentric zones, the Central business district, the Factory zone, the Zone of transition, the Working class zone, the Residential zone and the Commuter zone (Fig. 2.3). Analogue to natural ecosystems, it was argued that competition for land and resources ultimately led to the spatial differentiation of populations, since more desirable areas would command higher costs of living. According to the model, per capita prosperity increased and density decreased as one

moved away from the central business district. Along these zones a population flow was identified, which was described as the "succession cycle" of residents.

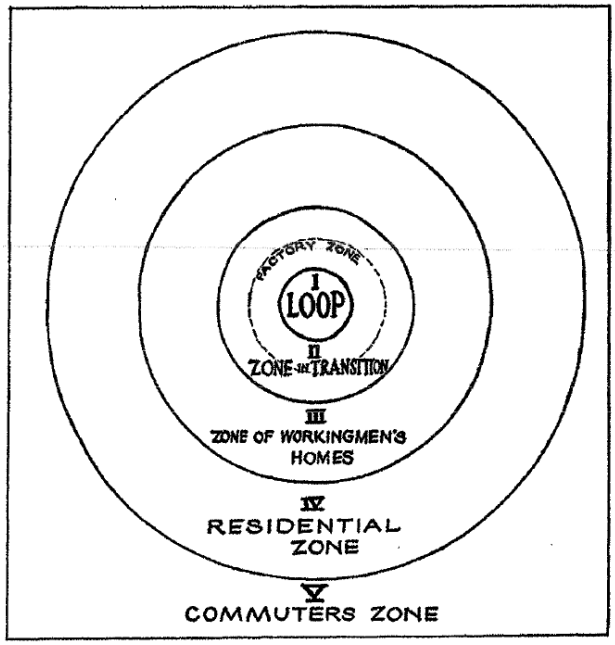

Figure 2.3 Concentric Zone Model Diagram (source: Burgess, 1925, p. 51).

Within urban sociology, Burgess's work was praised for the clear inclusion of space when analysing socio-economic patterns as well as his methodological emphasis on spatially referenced data and mapping techniques (Bulmer, 1984). Although he regarded mapping the spatial physical attributes as "*basic*" but of high importance in identifying "*natural regions*" (Burgess, 1925; Park and Burgess, 1925), he established thematic mapping as an analytic tool in a variety of fields, such as criminology and public policy. The Chicago School and their ecological approach certainly influenced a lot of research thereafter as a framework in urban data analysis and interpretation. Hoyt's (1939) analysis of Chicago extended the model with a new land use - social class schema that focussed on the outward growth of the city taking into account the characteristics of its railway infrastructure.

Nevertheless, a number of scholars refuted the validity of the concentric zone model. The main point of debate over the Burgess zonal hypothesis was the significance of the observed irregularities in spatial structure, and the extent to which these patterns are replicable within different geographic contexts. During the 1920s, Chicago was experiencing large flows of immigrant populations, expanding exponentially in a circular shape. The model worked for some other urban regions (Longmoor and Young, 1936; Bowers, 1939), but others exhibited significant pattern irregularities (Davie, 1937). The Burgess hypothesis was largely focused on the residential

location of income and rental conditions, and didn't take into account other important factors in locational population analysis such as household composition, occupation, ethnicity, etc. Quinn (1940) argued that unless more factors could be taken into account, such as detailed knowledge of existing spatial distributions for all significant social data and the physical environment, including constraints on travel, then an adequate test of the ecological approaches couldn't be made. A relative remark was made by Burgess in one of his latter works (Burgess, 1964), highlighting that little effort had been made to properly address the impacts of the build environment to local socio-economic patterns, such as the effects of physical deterioration and redevelopment. The role of transport is also another significant aspect that has been simplified in the original model. Transport is considered a crucial attribute of historical socio-spatial segregation, and there is extensive research on the role of transport in shaping these patterns. There are strong indications that, historically, connectivity directly affects residential choices and land uses. The importance of connectivity attributes in spatial segregation patterns are also highlighted in early work by McKenzie (1926) as in ecological time-cost distance versus the spatial linear distance.

The correlation of regional context (e.g. a city's infrastructure) to socio-economic phenomena is of particular interest, since it is a key aspect of this research and a foundation to the underlying analysis in the following chapters. Much research during the 1980s and onwards by sociologists and geographers collectively known as the Los Angeles school of urbanism suggests that Burgess's original model of urban ecology could not be universally applied. Dear (2002) for example, argues that the city of Chicago infrastructure was substantially different than other cities at the time. Chicago was mainly serviced by railroad infrastructure connecting the CBD area with transportation lines that radiated outwards in a hub-and-spoke fashion into the suburbs. The city of Los Angeles on the other hand, was built around freeway infrastructure that developed a multi-polar CBD pattern, and thus Los Angeles' socio-spatial patterns could offer a much more comprehensive paradigm of the postmodern metropolitan areas. Relevant studies however do not show strong indications that Los Angeles is indeed the paradigmatic city (Nijman, 2000; Dear and Flusty, 1998).

2.3. Factorial Ecologies and Urban Typologies

During the 1950s detailed census data was becoming more readily available, expanding the range of analysis in urban phenomena. A keystone work in the understanding of socio-spatial patterns

as well as one of the most important geodemographic precursors is "*Social Area Analysis*", which was developed through the work of Eshref Shevky and his associates. Shevky developed social area analysis as a variant of the urban ecology theory that takes into account detailed analysis of demographic, social and economic census data and through a statistical procedure produce salient underlying variables. Instead of a zonal structure, social area analysis aims to assert a typology of urban places measured through aggregated areal data. A comprehensive account of the theory can be found in Duncan Timms's "*The Urban Mosaic"* (1971).

In the initial study by Shevky and Williams (1949) titled "*The Social Areas of Los Angeles: Analysis and Typology*", the authors offer an original formulation about the residential differentiation across the urban region of Los Angeles. The newly introduced methodology measures residential differentiation by constructing three broad yet clearly defined indexes: "*social rank*", "*urbanization*" and "*segregation*". This multivariate analysis uses seven variables that are calculated and standardized based on data acquired for U.S. Census Tracts. The first index, social rank, is based on data about occupation, education and rental levels. Similarly, urbanization is an index based on household composition, specifically a fertility ratio, women employment and single family dwellings. Finally, segregation reflects the ethnic background of the population. The methodology initially assumes a differentiated space within a two-dimensional diagram whose axes are social rank and urbanization. The sub-areas that are defined are later further differentiated by the degree of segregation in that area (the third dimension), thus formulating a typology that could be spatially referenced and mapped. In 1955, Shevky and Bell revised the methodology to omit rent, due to approximation issues for owner-occupied homes within the 1950 census, and also possible miscalculations of rent levels due to market controls.

Social area analysis was considered extremely important at that time, not only for its utility as a methodologically innovative proposal to measure social segregation in urban areas, but also because it offered ways of comparison between cities or longitudinal studies of the same urban region. A large number of studies in the United States during the 1950s and 1960s employed this analytical framework (Tryon, 1955, Arsdol et al., 1958) as an early data driven ideographic form of science used to illustrate urban socio-spatial structure. The methodology quickly expanded over the US and applications were carried out in the UK (Herbert, 1967), Sweden (Sweetser, 1965), Italy (McElrath and Dennis, 1962) and Egypt (Abu-Lughad , 1969). Ecological approaches using these census tracks at the neighbourhood level dominated quantitative geography in the 1960s and so were later collectively labelled as "factorial ecologies", due to the use of factor

analysis to differentiate areal units and wide scope and approach of human ecology aspects used to explain urban phenomena (Sweetser, 1965; Janson, 1980).

Factor analysis is essentially a variable reduction technique, and as such the introduced methodology attempted to create simplified albeit meaningful urban typologies by reducing the complexities of human settlements (Abler et al. 1971). A desire to capture more socio-economic attributes and explore differentiations between urban areas through comparative studies led to a broader methodological framework of analysis that similarly employed factor analysis or principal component analysis to identify the major underlying attributes of spatial structure (Timms, 1971). Batey and Brown (1995) postulate that this stimulation over urban ecological research originated in the increased availability of US census data. During that period, census data was issued by the introduction of "census tracks", the lower geographical scale that aggregated data were available and referenced approximately 4000 residents. While the set of input variables and the final indexes were never fully justified, it could be suggested that a broader range of variables may produce better results. Such an extension is demonstrated by Rees (1972), although one should still demonstrate the theoretical rationale for the selection of the alternative variables.

A lot of similar work was directed towards comparative studies between cities. Many of these studies applied factor analysis on census data to produce results. Such multivariate analyses were used to produce classifications that would describe communities at a city-region or metropolitan level (urbanized areas), albeit some of this early research made distinctions between the central city and its suburbs (Walter and Wirt, 1972). For instance, factor analysis was employed to test the level of homogeneity between suburban regions (Schnore and Winsborough, 1972). Although these taxonomical studies were mainly aiming to find a common typology that would help with the interpretation of fundamental processes by which cities operate, there were various shortcomings creating a wider typology (Berry, 1972).

Factorial ecologies stirred a lot of debate regarding their utility. Throughout these studies there are indications of a number of theory implications between idiographic and nomothetic terms of research, i.e. understanding processes vs. describing them. Most of the critique focuses on the theoretical concepts underlying the methodology and argues that the majority of these studies took Shevky's theoretical synthesis for granted. In general, a lot of the initial criticism derived from the lack of a comprehensive conceptual and theoretical framework (Arsdol et al., 1958; Berry and Kasarda, 1977; Brindley and Raine, 1979). Others argued that the methodology was indeed parsimonious, yet had wide applicability (Bell and Greer, 1962). Berry (1972, p.2) for example argues that social segregation research shifted on the "*methods flowing from*

identification of variations of cities and following from the selection of dimensions relevant to a specific purpose", but due to the very high complexity of urban ecological models, which take into account socioeconomic, political and environmental factors, it may actually be impossible to amalgamate a universal model of residential segregation. At the small scale however (i.e. city specific), these multivariate applications seemed to produce sufficient results with high predictive power, but larger comparison analyses were deemed insufficient, mainly because of the lack of cohesion between the required datasets (Batey and Brown, 1995). Early censuses varied significantly in terms of the sets of observed variables, making longitudinal studies particularly difficult, while variation in the geographic aggregation levels could have a significant impact on outcomes (Openshaw and Taylor, 1979).

In this setting, geodemographics emerged as an evolution of multivariate applications as a way to include diverse socioeconomic variables to produce general population typologies without the need to select specific dimensions. While the usefulness of factorial ecologies started to decline, geodemographic analyses offer a useful way to explore and measure social-spatial segregation through the classification of geographic space. Still, geodemographics offer little chance to identify the drivers, government or market-driven, of socio-spatial patterns. They can be used however to map, explore and interpret correlations between typologies and other geographically referenced variables. A detailed review of geodemographic research during the last 40 years is presented in the following section.

2.4. Geodemographics

2.4.1. Introduction

The previous sections chartered the origins of geodemographic analysis through a selection of cornerstones in the evolution of socio-spatial analyses. Notably, a range of ecological, social area and multivariate analyses played a pivotal role in understanding social structure across geographic space. Key aspects of these studies were the nature of data and the geographical unit of analysis, given that the term "*neighbourhood*" depends largely on interpretation.

Certain shortcoming to these approaches forced later research within the academia to attenuate. Harris, Sleight and Webber (2005) provide a number of plausible reasons about this sharp decline, such as the growing discomfort in applying quasi-scientific, data-led and technique-

driven methods which forced the decoupling of analysis from theory, as well as an emerging political climate that tended to eschew analytical studies in policy making.

Nonetheless, the study of urbanised areas at finer geographic levels gradually offered more fertile ground for exploratory analysis, partly because of the reduced uncertainty of the outcomes (due to certain advances in data availability and data processing), and partly because of the growing indications of the utility of such detailed studies when addressing socio-economic phenomena with a narrow scope and purpose (such as unemployment, deprivation etc.). In this sense, a lot of the above multivariate studies acted as precursors to a more granular and purpose-specific analysis that established itself in the following years with the term "*Geodemographic Analysis*".

Geodemographics emerged mainly in the United States and the United Kingdom during the late 1970s as an extension to these earlier empirically driven models of urban socio-spatial structure. Typically, the central concept of the analysis is the production of a set of nominal classifications of neighbourhoods, in accordance to the socio-economic attributes of that area's residents. This kind of analysis differs from earlier multivariate models in terms that there are no inherent commensurable factors determining spatial discrimination; every cluster has unique properties based on the collection of variables that have been used in the construction of the classification.

In the following years, geodemographic classifications gained wide popularity as their utility was demonstrated across a variety of fields. Areal classifications have been applied in a broad spectrum of academic research, such as health, education and policing studies, policy planning and resource allocation. However, the majority of applications come from the private sector, as geodemographics have been heavily utilized in industries for retail planning, locational analysis and market segmentation (Reibel, 2011). The following section outlines the emergence of geodemographics through the work of pioneering research and applications of small-area multivariate classifications.

2.4.2. Small-area Classification Studies

Geodemographic analyses were initially developed as a "*strategy*" to identify patterns from multi-dimensional census data (Webber, 1978). Similar to the development of other multivariate approaches, the availability of detailed socio-economic data played a key role in the development of geodemographics. The reformation of Britain's Census scheme during the second half of the

20th century was one of such key factors that enable its growth. In contrary to the US regime, Britain was initially slow to respond to the supply of census data at lower geographic scales, but after the 1951 Census adopted a much more granular approach. Aggregated census data was supplied at the level of Enumeration District (ED), which consisted of about 270 households for England and Wales and 150 for Scotland (Cox, 1976, p. 66), nearly four times smaller than the US census tracks. More information on the Census can be found in Chapter 4.

In the UK the initial justification for constructing area classification systems using small-area statistics was to provide central and local government with a tool for targeting policies and allocate resources on a priority-area basis (Webber and Farr, 2001). Early work on the analysis of enumeration district data includes Gittus's (1964) study of the Northwest's conurbations using 1951 ED census data and Robson's (1969) study of Sunderland. Further attempts and experimentation on a multivariate methodological framework of the 1961 Census allowed for more detailed analysis at finer geographic scales. A central figure in the development of these studies was the Centre for Urban Studies at UCL, as their work was very influential in terms of method, scale and application. In fact, the majority of these early attempts at socio-spatial structure analysis were either carried out by the Centre, or by other researchers who adopted their methodology (Batey and Brown, 1995).

Small-area classifications applications begun to increase during the late 1960s, largely as a result of the increased interest for such classifications by British Local Authorities in order to evaluate local policies and the allocation of social service resources. One of these studies, the "Third Survey of London Life and Labour" (Norman, 1969) was particularly inspiring in the sense that it enabled a classification of socio-economic structures of Inner London through a principal component and cluster analysis of 28 census variables, aggregated at the ED level. Another unique quality of this study was that the six clusters of populations that were defined were accordingly mapped and assigned names (e.g. Upper Class, Bed-Sitter, Almost Suburban), which were further illustrated through "pen-portraits" based on the interpretation of the main census variable characteristics (Fig. 2.4). Kelly (1969) undertook a similar study for the Greater London Council Research and Intelligence Unit, creating a series of London Borough classifications. Throughout these studies however, the significance of scientific research did not go beyond the demonstration of what can be achieved, and it was not clear whether such studies had any immediate applications (Batey and Brown, 1995).

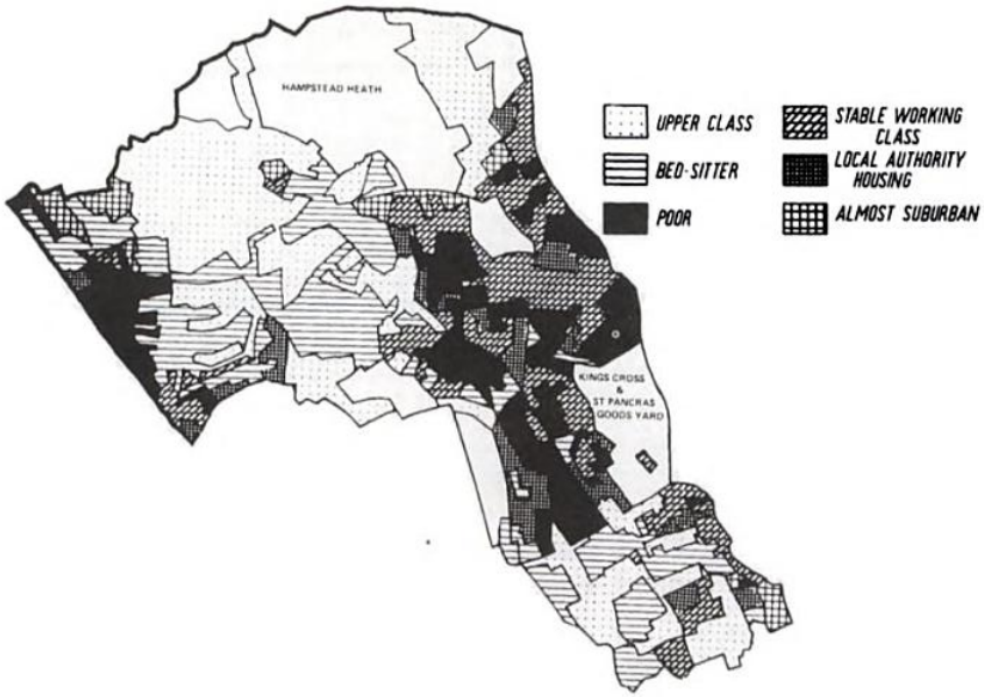

Figure 2.4 An example of the enumeration district classification for the area of Camden from the "Third Survey of London Life and Labour" study (source: Batey and Brown, 1995).

A key study that furthered this work and addressed some of the aforementioned issues was the Inner Area Study for Liverpool or Liverpool Social Area Study as it came to be known, funded by the Department of the Environment in 1969 (Wilson and Womersley, 1976). The Liverpool study is considered pioneering in geodemographic analysis. Its purpose was to classify Liverpool's neighbourhoods based on affluence and identified concentrations of social problems, however it used census data to build the classifications and local government data (social statistics collected by council departments) to examine output clusters (Webber, 1975; Webber and Craig, 1978). Various local authorities made use of this methodological framework exploring deprivation levels, programme targeting and structure plans. Wirral for instance successfully used a similar area-based classification analysis to justify partaking in the central government priority funding budgets (Webber and Farr, 2001).

Shortly afterwards, Richard Webber, as one of the main researchers of the Liverpool Social Area Study, begun to work on national scale classifications. He used census Enumeration District data to group areas at the scale of Wards, Local Authorities and Parliamentary Constituencies (Webber, 1977). Until then, most of these studies were restricted to cities or specific regional areas, while a national-scale classification had the advantage of comparisons between specific

cluster variables to the national mean. Webber's classification studies drew the attention of various industries that sought to exploit areal classifications for commercial purposes. His work was soon acquired by CACI, and the classification was adjusted to a postcode-based geographic unit and rebranded as *ACORN* (A Classification Of Residential Neighbourhoods), in order to link consumer behaviour data with residential socio-spatial patterns (Harris et al., 2005).

National classifications have proven useful for marketing and were widely used as predictors of consumer behaviour (Webber, 2007). In the following years, proprietary classifications gained a lot of popularity, particularly in the UK where the term geodemographics and market analysis became practically synonyms (Birkin, 1995). By the mid-1980's, the classification methodology had already been established, and following the same analytical framework, many other proprietary systems offered similar national classifications as market segmentation systems, such as the *PiN* (Pinpoint) and *MOSAIC* (CCN, now Experian) classification in the UK and the *Lifestyles* classification in the US. The apparent bloom of proprietary classifications was evident when data relating consumer behaviour and media preferences where brought together with socio-economic profiles, as well as the introduction of computerized datasets in conjunction with spatial data management systems, Geographical Information Science and spatial analysis (Birkin and Clarke, 1998).

Despite a common starting point, geodemographics have evolved through different paths between the UK and the US. While the US applications focus on commercial uses, in the UK context there is a long history of free and more recently open classifications available, and they have been more broadly used in public policy and academia (Singleton and Spielman, 2013).

Since the 1980s, academics in the UK tried to produce geodemographic classifications utilizing the decennial census of population, in an effort to provide alternatives that would appeal to the academic community, and lacked the necessary resources to acquire proprietary classification systems. Charlton, Openshaw and Wymer (1985) detailed the creation, as in the techniques and challenges, of such a national classification that could be used similarly to the commercial ACORN, using 1981 census data. The classification is described as a "general-purpose classification" with potentially "novelty" value and undetermined usefulness.

This academic legacy continued in the following years and further attempts were made to produce geodemographic classifications with improved classification performance and updated data. Within the public domain, several efforts had been supported towards that goal, for instance a wide range of applications in health, community safety and higher education of the 1994 Super Profiles (Brown et al., 2000).

Within academia per se, there were several attempts to construct open and transparent classification such as the GB Profiles, a small-area census classification of Britain's residential areas using 1991 census data (Blake and Openshaw, 1994), and later the Output Area Classification for 2001 (Vickers and Rees, 2005) and 2011 which is supplied by the ONS using census data from the 2001 and 2011 Census respectively.

2.4.3. Geodemographic Systems

The outcome of a geodemographic analysis is a classification that represents a categorical summary measures that aims to capture detailed salient multidimensional characteristics of both built environment conditions and socio-economic characteristics of small geographical areas. It is very common however for geodemographic products to have multiple classification levels. These products are collectively known as geodemographic systems.

Within the UK, contemporary classifications are typically of high geographical granularity, constructed at small-area or address level. Cluster units are usually calculated per Output Area, the smallest area that aggregate census variables are currently available, or postcode level. The geographical extent of such classifications can also vary significantly, and unlike their precursors in ecological studies and urban typologies that were limited to urban specific cities or urban conurbations, the span of the classification can range from national level to individual regions or cities. A comprehensive analysis of the methodological steps and the techniques used in the creation of geodemographic classifications is detailed in Chapter 4.

Classifications are typically labelled through the study of a set of attributes attached to a particular cluster of the population. For instance, households with both parents employed in managerial positions, living in detached housing at lower densities and with a higher than average car ratio might possibly be construed as prosperous suburban families. However, geodemographic classifications tend to be highly dimensional, and interpretation is much more complex as they use a plethora of input variables from a variety of sources.

From the outset (Webber, 1977), geodemographic methods have typically employed a pragmatic variable selection strategy, combining the experience of the classification builder (what is deemed to work) with the overarching purpose of a classification (what is required). During the last 30 years, current geodemographic systems have evolved considerably, and may use a variety of data from public (e.g. census data) or private sources (e.g. market surveys, credit check histories) to generate profiles (Birkin, 1995; Singleton and Spielman, 2013). Typical classification

datasets include anywhere from a few dozens to several hundred empirically derived socio-economic and built environment characteristics. However, different sets of variables operate on different scales and there are various ways in which such variables are managed, ranging from simple apportionment from aggregate to disaggregate scales, small area estimation or micro-simulation (Birkin and Clarke, 2011).

The number of outcome clusters also varies significantly among classifications, and can range from several to a few hundreds, depending on purpose (Singleton and Spielman, 2013). Table 2.2 summarizes various geodemographic systems and the profiles they offer, collated from web data sources. Typologies emerging from these clusters are also typically presented as a hierarchy with varying tiers of homogeneity (Table 2.3). Such hierarchy can be created from the top or the bottom. A top-down approach includes the creation of larger groups of cases that are subsequently divided into smaller sub-groups. For example, this method was implemented to produce the 2001 OAC, included 7 Super-Groups, which were respectively split into 21 Groups and further into 52 Sub-Groups. A bottom-up approach on the other hand, more prevalent within the commercial sector, includes the creation of numerous smaller groups which are then aggregated based on their similarities into larger groups (Alexiou and Singleton, 2015a).

Table 2.2 Number of Classification Profiles per hierarchy level for selected public and proprietary geodemographic classification systems.

Classification System	**Number of Profiles (Clusters)**		
	High Aggregation (Coarse tier)	*Moderate Aggregation (Middle Tier)*	*Least Aggregation (Detailed Tier)*
ACORN	6 (Categories)	18 (Groups)	62 (Types)
Mosaic	15 (Groups)		66 (Types)
CAMEO	10 (Marketing Segments)		68 (Types)
P^2 (People and Places)	4 (Categories)	41 (Branches)	157 (Leaves)
PRIZM NE (CLARITAS)	14 (Groups)		66 (Types)
ESRI Tapestry Segmentation	14 (LifeMode groups) 6 (Urbanization groups)		67 (Segments)
ONS Output Area Classification (2001)	7 (Super-Groups)	21 (Groups)	52 (Sub-Groups)
ONS Output Area Classification (2011)	8 (Super-Groups)	26 (Groups)	76 (Sub-Groups)

Table 2.3 An example of the nested hierarchy for one of the Super-Group clusters described as "*Constrained City Dwellers",* from the Office of National Statistics (ONS, 2015b).

Super-Group	**Group**	**Sub-Group**
7: *Constrained City Dwellers*	*7a - Challenged Diversity*	*7a1 - Transitional Eastern European Neighbourhoods*
		7a2 - Hampered Aspiration
		7a3 - Multi-Ethnic Hardship
	7b - Constrained Flat Dwellers	*7b1 - Eastern European Communities*
		7b2 - Deprived Neighbourhoods
		7b3 - Endeavouring Flat Dwellers
	7c - White Communities	*7c1 - Challenged Transitionaries*
		7c2 - Constrained Young Families
		7c3 - Outer City Hardship
	7d - Ageing City Dwellers	*7d1 - Ageing Communities and Families*
		7d2 - Retired Independent City Dwellers
		7d3 - Retired Communal City Dwellers
		7d4 - Retired City Hardship

The interpretation of cluster attributes is very important to the success of the classification. Clusters are represented by naming and describing the resulting clusters with written "pen portraits" that best fit the profile of areas. Classification systems also commonly augment such descriptions with accompanying visual materials such as photographs, housing images and bar graphs or radar charts (Fig. 2.5). Depending on the intended end users, labelling and description must be carefully selected in order to expand the user's understanding of the group, while producing a summary of their key attributes (Alexiou and Singleton, 2015a). A more in-depth approach to cluster interpretation can be found in Chapter 4.

Figure 2.5 Examples of visual materials accompanying cluster representations provided in the ACORN by CACI (London, UK) classification (CACI, 2013).

2.4.4. Strengths and Weakness of the Analytical Framework

The theoretical framework of geodemographics is based on a number of human behavioural phenomena. The main conceptual framework is based on a fundamental notion in social structures, homophily - the principle that people tend to be similar to their friends (Easley and Kleinberg, 2010). This notion manifests spatially as a general tendency that people live in places with similar people, much like the *"birds of a feather flock together"* adage suggests, and is evident in many academic studies. Hitherto, their popularity stems from this upholding validity (Longley, 2007; Sleight, 2004).

The spatial dimension of homophily can be best explained, if it is examined as a variation of spatial autocorrelation across geographic space. Another fundamental and generally established axiom in human geography is Tobler's first law of geography (Tobler, 1970). In general, areal units across a region will have positive autocorrelation with other units that share similar attributes and vice-versa.

Another viewpoint is that people may be driven by their personal circumstances, yet they also tend to also align themselves with the behaviour and aspirations of the local community they are living in. The collection and analysis of such complex behaviour data would be next to impossible. By aggregating people based on the neighbourhoods there is a notion that there is actually control

for these variables (e.g. love for gardening), although information about which of these attributes are influenced by neighbourhood effects is very limited (Webber and Farr, 2001).

An inherent trait of geodemographics, inter alia, is the ability to adjust the diversification of the constituent classes, or specifically the level of hierarchy. Output classifications can range from broader descriptions (e.g. Affluent families, Middle class or Struggling households) to very specific (e.g. Aspirational tech workers, Terraced Pakistani working families). Researchers without the necessary data and/or expertise in statistics and geographical information science may find such systems convenient to use. The strength of geodemographics is the ability to use pre-built classification systems for a number of hierarchies, which can be populated by various attributes making them versatile and easy to update. In other words, geodemographics are very easy to operationalize.

On the other hand, the nature of the classification is very specific to the underlying data and the methodology adopted by the creator. Since there is no single global optimization function during the classification procedure, geodemographics are highly subjective to the operational decisions during the creation process (Openshaw and Gillard, 1978). Critique on geodemographics focuses on the ambiguity of the scientific basis (Goss, 1995; Harris et al, 2005) as well as the fact that more geodemographic classifications lack required transparency in order to test their validity (Longley, 2007). The quantitative evaluation of the relative performance of classifications has been until now very limited (Voas and Williamson, 2001; Webber, 2004; Brunsdon et al, 2011).

Evaluation attempts can be increasingly difficult with absence of classification transparency (Fisher and Tate, 2015). Classification transparency relates to the data, methods and underlying techniques used to construct a classification, so it can be easily replicated, updated or otherwise customized to fit the needs of the creator (Brunsdon, 2015). One such example of a transparent classification is the OAC, which makes available not only the classes but also all the required class centroids which makes evaluation of the classification uncertainly level possible (Fisher and Tate, 2015). This “black box” issue is problematic not only because of the degree of effectiveness these geographically crude measures have but also because they inherently prohibit the development of established tests to their validity (Longley, 2007). There is a greater debate that has recently emerged in quantitative geography, the issue of reproducible research (Brunsdon, 2015); when certain outcomes cannot be easily reproduced, updated or modified, advancements in the field are bound to be limited.

As aforementioned, despite a lineage of use, geodemographic classifications lack a solid theory. In nomothetic terms, geodemographics can be labelled as methodologically unsatisfactory since the underlying theory is "*simplistic*" and "*ambiguous*" (Harris et al., 2005). The analytical weaknesses, often coupled with a lack of any statistical clothing, often make it difficult to assess either the significance of apparent trends found in data or the importance of predictor variables that might explain those (Harris et al., 2007).

Furthermore, geodemographic classifications as currently developed can be considered contradictory to Tobler's statement. The central concept of geodemographics has only been applied to the clustering processes, and not to the geographical context of each area. The methodological implications are based on the fact that aggregations into categorical measures sweep away contextual differences between proximal zones; and as such, the final classifications assume that areas within the same cluster have the same underlying characteristics. Standard geodemographic techniques have failed to incorporate near-geography in a sophisticated way, and despite the term, geodemographics are in fact "aspatial".

This traditional aspatial approach also disregards the issues of scale (Reibel and Regelson, 2011) and has a number of implications when generating profiles. There is a longstanding debate originating in the earliest of UK classifications on the impact of the methodological implications in national geodemographic system's performance, and whether classifications built for national, regional and local extents are effectively built for different purposes, and as such undermine comparison (Openshaw et al., 1980; and Webber, 1980).

These particular methodological shortcomings are the main focus of this research and will be addressed in detail in the following chapters. The Thesis aims to add and extend existing geodemographic research in order to address the methodological weaknesses associated with near-geography and evaluate it vis-à-vis traditional geodemographic models. To do so, the proposed methodology draws upon existing (but limited) research on the topic. Thus far, there have been very few attempts to build a unified framework where the relative benefits of both spatial interaction and geodemographic approaches can be maximised. For example, Singleton and colleagues (2012), expanded the model with a spatial interaction framework that demonstrated the spatial flows between clusters, and Debenham, Clarke, and Stillwell (2003) modelled the classification methodology to include regional differences.

In the private sector, proposed methodologies have used a number of controversial techniques to address these limitations, such as selecting attribute contextual measures or spatial interaction models. Among the most common techniques to address these limitations are the

implementation of radial buffers for zones (i.e. calculating attribute averages within a fixed radius of a zone centroid), and selecting attribute locational contextual measures (Harris et al., 2005). However, these arbitrary zones or administrative boundaries may not represent the organization of actual localities.

Furthermore, underlying techniques that are used in this manner are typically obscured and thus impede reproduction. Classification transparency relates to the data, methods and underlying techniques used to construct a classification, so it can be easily replicated, updated or otherwise customized to fit the needs of the creator (Brunsdon, 2015). One such example of a transparent classification is the OAC, which makes available not only the classes but also all the required class centroids which makes evaluation of the classification uncertainly level possible (Fisher and Tate, 2015). This "black box" issue is problematic not only because of the degree of effectiveness these geographically crude measures have, but also because they inherently prohibit reproduction and any development of established tests to their validity (Longley, 2007).

2.4.5. Geodemographic Applications

Geodemographic classifications have been used in a variety of fields. Population typologies are useful in order to infer behavioural, health or other specific characteristics of a particular population group. A simple analysis would reveal whether the attribute under research is under- or over-represented in areas corresponding to specific clusters. Geodemographic classifications, due to the already consolidated detailed information they contain about a variety of population attributes at high geographical scales, offer huge advantages towards the analysis and recognition of geographical patterns. They can be used as an exploratory technique and help identify important associations or triggering factors (e.g. mortality rates, policing or internet usage) of nearly the whole spectrum of social phenomena. The inherent ease in analysing the patterns that emerge from the variation in representation or penetration rates within individual types of areas is one of most important strengths of geodemographic applications.

Another advantage of geodemographic classifications, encouraging their proprietary uses, is the fact that they are easy to operationalize. As Webber and Farr (2001, p. 55) state, classifications can be *"simple and low-cost but versatile and effective forms of data fusion, whereby analysis undertaken from one database can be operationalised through another"*. The classification is easy to append to customer, prospect or respondent records via their address, which can be usually acquired as public domain data.

Within this framework, geodemographics were initially heavily utilized in the private sector, as the macroeconomic conditions alongside the freedom-of-information tradition created an environment that quickly tried to exploit (census) data commercially (Flowerdew and Goldstein, 1989). The first commercial applications during the early 1980's were mainly advertised for retail analysis and market segmentation, and many companies advertised both general purpose and market specific "*geodemographic discriminators*" (e.g. financial services) depending on the nature of the fused market research data (Beaumont and Inglis, 1989). Their composition differs depending on the scope and probable usage of the intended stakeholders; available geodemographic products include a variety of classification systems, and produce discrete classes primarily designed to describe consumption patterns, without that limiting the potential applications of geodemographic only to the retail sector (Webber, 2007). For instance, Atlas (1989) adopted the geodemographic approach to analyse voting patterns in the U.S. Among the conventional general purpose classification systems are some privately developed classifications such as the MOSAIC (by Experian), ACORN (by CACI), P2 People and Places (by BD), Claritas (by PRiZM) and EuroDirect (by CAMEO).

There is an abundance of literature regarding the utility of geodemographics in retail planning and market analysis, detailing the creation methods, database operationalisation and management of geodemographic systems (Birkin et al., 2002; Harris et al., 2005). Available research also captures a wide set of specific subjects, from consumer marketing (Sivadas et al, 1997) to location planning and catchment areas (Clarke, 1998; Birkin and Clarke, 1998).

Besides the utility of geodemographic segmentation in the commercial sector, geodemographics have a variety of uses in other academic studies regarding public sector management. In general, there is a recent renaissance on geodemographic applications for public sector usage, particularly regional planning and policy analysis, mainly driven by government pressure to demonstrate value for money and the advent of new application areas (Longley, 2005). Applications of geodemographic classifications are observed in a wide range of academic literature, and they are as diverse as school performance screening (Butler et al., 2007) to fire incidences (Corcoran et al., 2013) and natural hazard vulnerability (Willis et al., 2010).

Batey and Brown (2007) developed a method of evaluating the success of area-based initiatives by using a geodemographic classification to produce spatially targeted socio-economic profiles. They assessed the efficiency of urban policies by examining how many of the people they contain are in fact not those for whom the initiative is intended, in which case, it is defined as inefficient or incomplete. Longley and Goodchild (2008) also provide a comprehensive overview of the issues of geodemographic applications in a range of public sector settings.

Among the research topics that geodemographics have been used is health screening, for instance in geographic epidemiology, where detailed geographical information is often unavailable. In these studies, finer geographic granularity is essential in order to produce accurate ecological estimates and infer correlations or interaction effects between health and demographics (Brown et al., 1991; Openshaw and Blake, 1995; Hedges et al., 1997; Tickle et al., 2000; Aveyard et al. 2002). Farr and Evans (2005), for instance, developed a geodemographic model that could help identify people at risk of Type II diabetes, while Dever, Smith and Stamps (2005) used the Claritas PRIZM clusters to assess perinatal health statuses. Identifying such health patterns can also aid in health campaign and neighbourhood targeting (Petersen et al., 2011). Small area aggregates can also be used to increase statistical power, as small area ecological data can alleviate bias due to measurement errors in individual-level data (Jackson et al., 2006). Geodemographic have also been used in survival models through the analysis of mortality patterns (Shelton et al., 2006). Mortality ratios are also widely used in actuarial economics and, in this context, geodemographic models can provide a major boost in explaining risk variation at postcode level (Richards, 2008).

Another major application of geodemographics regards the geography of participation in higher education. Batey, Brown and Corver (1999) explored the prospect of further expansion in student taking into account a number of factors that may determine the scope for expansion in particular regions and sub-regions. Singleton (2010) and Singleton, Wilson and O'Brien (2010) explored the patterns of access to higher education by linking summary measures of local neighbourhood characteristics with individual-level educational data. Through a spatial interaction framework, they demonstrated the size of spatial flows between socio-economically stratified areas and institutions, with the aim that such a tool could be used by key stakeholders to examine potential policy scenarios. Brunsdon, Longley, Singleton and Ashby (2011) extended the analysis of the participation index with an added evaluation of the classification performance using a number of different classification systems and comparing results.

Other notable examples include the application of geodemographic approaches in ecological studies regarding neighbourhood profiling. Ashby and Longley (2005) and Williamson, Ashby, and Webber (2006) used geodemographic analyses of local policing environments, crime profiles, and police performance in order to infer a neighbourhood classification that is produced explicitly to reflect differing policing environments and help allocate policing resources accordingly.

Many geodemographics applications focus on inferring correlations or interaction effects among different demographic groups based on spatially referenced individual level data. For instance, Riddlesden and Singleton (2014) used a similar classification approach to analyse and

infer relationships between internet engagement and demographic profile on a neighbourhood level. Bearman and Singleton (2014) used the same set of techniques to model CO_2 emissions of the home to school commute using individual level data. In general, there is a growing need for geodemographic systems that are open and versatile enough to handle the abundance of big data that are readily available.

Chapter 3. The Analytical Framework of Geodemographics

3.1. Overview

In order to understand the methodological background of Geodemographics it is important to address a few definitions and concepts about the methodological and technical framework commonly used this context. As discussed previously, central to geodemographic systems are socio-economic classifications of neighbourhoods typically based on night-time population attributes. Early geodemographic precursors relied on other multivariate data reduction techniques in order to identify structure, such as factor and principal component analysis. Geodemographic systems rely now on a plethora of attributes to generate profiles, and exploratory techniques are essential in order to examine and understand underlying patterns. Classification systems are easy to use and are becoming more accurate, such as bespoke classifications (Riddlesden and Singleton, 2014; Alexiou et al., 2016), and assume better correlations between places and social identity (Longley et al., 2008).

The classification process is usually specific to the underlying data and methodology. A larger part of geodemographic analysis is thus exploratory; the analytical steps include a combination of statistical techniques, descriptions and visualizations that could lead to the discovery of the best data organization. A data organization technique is thus crucial in order to describe the hidden structure from unlabelled data. For such multivariate observations, it must take into account the similarity between all their attributes and assign them into homogenous groups.

One of the most common data organization techniques used in geodemographics is clustering. Clustering refers to various techniques used in order to classify objects based on their observed characteristics. Methodologically, geodemographics rely now on a cluster analysis of multidimensional and geographically referenced data to deliver categorical descriptors of small area geography (Singleton and Spielman, 2013). Data clustering is used in order to identify the geographical socio-economic patterns and their characteristics, assign neighbourhoods to distinct and concisely labelled categories, and quantify the amount of heterogeneity between groups.

Due to the sheer complexity and inherent uncertainty of clustering problems, there are many decisions that need to be taken before and during the classification process. Harris, Sleight and Webber accurately point out that geodemographics are "*born as a cross-breed between art and science*" (Harris et al., 2005, p. 181) and that the experience of the creator is vital to the analysis.

From the outset, there was an overarching principle that geodemographics rely on a pragmatic approach, and some degree of empirical evaluation prior to creating a classification (Webber, 1977). Depending on the scope of the analysis, the creator must take into account variable selection and data availability, measurement scales, weighting schemes and clustering parameters. The classification builder has a number of choices to make at all these stages of the analysis.

As discussed in Chapter 1, the aim of this research is to produce a classification methodology for geodemographic research that incorporates geographic sensitivity of small-area attributes. However, in order to define geographic sensitivity and explore the possible methodological extensions, it is important to initially provide the framework of such methodologies, what exactly is their objective, what are the preconditions that need to be met, i.e. the data structure assumptions, clustering methods and how their results should be construed.

Clustering methods are certainly used in a variety of applications, as demonstrated by the relevant literature. Scientific interest about clustering grew exponentially after the 1970's (Blashfield and Aldenderfer, 1978; Millligan and Cooper, 1987). The abundance of methods reflects the level of complexity of clustering algorithms depending on discipline and the nature of data. Various algorithms have been designed and modified to address specific problems. Because of the breadth of the disciplinarily, the lack of cross-reference and communication between disciplines is imposing disproportionate effort to the inexperienced researcher, as to what method to select and how to apply it (Jain and Dubes, 1988).

Depending on the field, terms and definitions also vary quite significantly. Besides clustering, umbrella terms used for these problems are, among others, taxonomy analysis, data reduction, unsupervised learning or statistical density estimation. It is important to address the terminology issues as concise as possible, and try to illuminate some aspects of cluster analysis through a pragmatic perspective. In the following sections of this Chapter we provide some useful definitions and concepts regarding classification problems, present summaries of current clustering techniques popular within geodemographics and social sciences in general, as well as address a few issues regarding clustering validation and inference.

3.2. Definitions and Concepts

3.2.1. Exploratory Analyses and Data Mining

Within geographic research, there is an increased interest in exploratory analyses over confirmatory ones, mainly as a result of the combination of the "*extreme complexity of geographic processes and the availability of large databases and sophisticated software"*, such as G.I.S. (Rogerson, 2015, p.4). The latter reason has recently rendered clustering a fashionable topic, due to the need to summarize and identify patterns in very large and often chaotic datasets. Such data, often called "*big data*", are now collected ubiquitously; their size has increased exponentially in recent years, and many researchers now focus on how potential insights can be harnessed from them (Graham and Shelton, 2013).

There is an abundance of such datasets offered "*raw*" or "*in bulk*", the majority of which have been made available through the developments in computer science, such as the automation of governmental and business activities (Frawley et al., 1991) and the World Wide Web (Liu, 2006). Such a process is now called "*knowledge discovery*" or more frequently "*data mining*", as in the ''*the extraction of implicit, previously unknown, and potentially useful information from data*" (Witten and Frank, 2005, p. xxiii). Data mining is effectively the process of turning data into information, and lies at the intersection between statistics, artificial intelligence, machine learning and database systems. It entails the exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns (Tan et al., 2008, Chapter 1). Essentially, it is a type of data analysis without any *a-priori* hypotheses.

Data mining has gained many proponents during the last decade, mainly because of the advances in computational power that made processing extremely large datasets feasible. In this framework cluster analysis has been used extensively as a form of exploratory analysis on multivariate data and across academic domains. For instance, data mining is becoming increasingly important in marketing applications. Dividing customers into homogenous groups was one of the basic strategies of marketing, often in order to identify population types and their correlation to a product uptake (neighbourhood targeting) (Frank and Green, 1968). The advantages of such approaches were identified very early on, for example the analysis carried out by Green and colleagues (1967) about the relationships between newspaper circulation and city type. More recently, geodemographic classifications imbued with a variety of data are used widely as one of the most powerful discriminators of consumer behaviour (Graham, 2005).

3.2.2. Classification vs. Clustering

Historically, terms and concepts about clustering vary, depending on the discipline, and terms such as (mathematical) taxonomy (Sokal and Sneath, 1963; Jardine and Sibson, 1971), clustering (Tryon and Bailey, 1970; Hartigan, 1975), classification (Good, 1965) or data segmentation (Hastie et al., 2009) have been used interchangeably. This also applies to the terms cluster, group and class, which hitherto are being used intuitively.

While initially the word clustering was synonymous with classification (Hartigan, 1975; Milligan and Cooper, 1987) there is now a general consent, at least within computer science and machine learning, about the critical differences between clustering and classification methodologies. The objective of cluster analysis is to simply find a suitable and valid organization of the data into clusters, and not to establish rules for separating data into categories (Jain and Dubes, 1988). A classification analysis, also known as *discriminant* analysis, is the task of assigning objects to one of several predefined categories, i.e. detecting spam email, categorizing plant species or classifying galaxies. Decision tree classifiers for example are simple yet broadly used types of modelling techniques (Tan et al., 2008, Chapter 4).

In contrast to clustering, a discriminant analysis involves the collection of a labelled, pre-classified data, either by identifying or theorizing their classes (depending on a probabilistic or descriptive approach). The aim of the analysis is to label newly encountered yet unclassified data based on the known (also known as training) dataset. In the case of clustering, the problem is to group an unlabelled set of data into meaningful clusters. In a sense, labels are also associated with clusters, but resulting category labels are data driven (Jain et al., 1999).

The classification approaches are more widely known within the field of *machine learning* as *supervised learning*, originating from the iterative natures of the algorithms that can be seen as a learning process. Clustering on the other hand is a form of *unsupervised learning*, since it involves grouping attributes to understand the underlying patterns in data (Kohavi and Provost, 1998). In general, clustering data suggests an exploratory dimension within the analysis.

Although geodemographics involve a classification of geographic space as an output, these classifications are a result of spatial clustering used to identify and analyse these patterns; therefore, the conventional process does not entail an inherent and predefined classification that each neighbourhood is assigned into. In this sense, the term geodemographic classification is misleading; the classification refers to the output of the analysis and not the underlying process.

Geodemographic analysis could be better described as the process of creating a classification via clustering of neighbourhoods.

3.3. Cluster Analysis

3.3.1. Overview

Cluster analysis is essentially a set of tools and techniques regarding effective ways to sort heterogeneous data into homogenous blocks (Hartigan, 1972). More recently, possibly to distinguish clustering and *discriminant* analysis, Kaufman and Rousseeuw (1990) defined cluster analysis as the classification of similar objects into groups, in which the number of groups and classes is initially unknown.

Perhaps the most familiar form of clustering is the taxonomy of animals and plants, which groups various species into broader classes. Clustering has been used extensively in many academic studies and in a wide range of disciplines, such as biology, medicine, psychology, geography and computer science. From a geodemographic perspective, clustering is used to cluster neighbourhoods based on their common attributes, i.e. the socio-economic characteristics of their residents. In a similar manner, cluster analysis has been recently used extensively within the private sector, and particularly within market research (often referred to as data segmentation). For example, cluster analysis is used in order to group customers based on their demographics and identify "niche" markets (Mooi and Sarstedt, 2011, Chapter 9).

The scientific extent of cluster analysis is currently very broad, and it can range from classifying genome characteristics to document and image analysis. Due to the sheer number of applications, an abundance of clustering techniques have been developed over the years, tailored to one's specific needs and nature of data under investigation. Within this framework, a number of articles have been published to provide a review of methods and techniques used in cluster analysis (Blashfield and Aldenderfer, 1978; Milligan and Cooper, 1987; Jain et al., 1999).

Everitt et al. (2011), argue that the utility of clustering is twofold. At one level, a classification can be used to identify underlying aetiological relationships between observations (for instance, by grouping various health symptoms of individuals to identify those suffering from particular diseases). However, a classification scheme has also value in itself; clustering can be viewed as a summarization technique, where large data is divided into smaller but similar groups. This

property of cluster analysis rather than finding "natural" clusters is often called *dissection* (Everitt, 1980).

Clustering is a good exploratory technique, since it can enable the researcher to understand the data, make concise labels about what each group represents, and draw relations between groups and other phenomena with ease. Geodemographic analysis might also be used as a first iteration before variable selection in some other kind of model, where such approaches have been demonstrated with creditable performance (Brunsdon et al., 2011).

Computational clustering also offers several advantages over manually "looking" at a data matrix to detect clusters. While it is easy to identify clusters in a two- or three-dimensional plane and visualizing the data points, a clustering algorithm can apply a specific objective criterion objectively and across multiple dimensions. Jain and Dubes (1988) argue that various individuals will often see different clusters, depending on their educational and cultural background, and that the speed of organization of a clustering algorithm is overwhelming compared to that of human beings. Secondly, they argue that clustering algorithms can also be used to reduce the complexity of decision-making algorithms in pattern recognition, as a data reduction strategy.

It is worth noting that clustering can refer both to clustering objects (i.e. cases, observations) and clustering variables (parameters). One of the most common ways of grouping variables is factor analysis and principal component analysis. Factor analysis has greatly influenced the analytical framework within socio-economic analyses (see Chapter 2 for some examples), and it has been used extensively in order to understand the impact of one particular dimension of variability to the observations. Nevertheless, for the scope of this research focus is placed on object clustering methods and their variations.

3.3.2. Clustering Methods

According to Milligan (1996), a clustering method is essentially the technique through which clusters are formed, whereas cluster analysis refers to a broader set of tools and techniques that are essential in order to carry out the analysis. Clustering methods aim to give a solution to the problem of sorting data into homogenous groups. The problem under consideration is, given a set of data points, how can they be grouped into clusters so that points within each cluster are similar to each other, and points from different clusters are dissimilar. Typically, such problems involve data points that are in a high-dimensional space, so similarity (or dissimilarity according

to some authors) is defined using some kind of appropriate distance metric, e.g. Euclidean distance, Jaccard, taxicab or graph distance.

Milligan and Cooper (1987) provide a useful analysis of clustering methods with emphasis on algorithm performance in applied research. According to them, a conventional cluster methodology can be described by a seven step process:

1. Selection of objects to be clustered.
2. Selection of the attributes of the objects that provide scientific information about their similarities.
3. Information about the standardization procedure, if any, to the clustering objects.
4. Selection of a similarity or dissimilarity measure as an objective function to quantify similarity.
5. A clustering method should be selected, with emphasis on what type of clusters structures are expected to emerge in the analysis.
6. The number of clusters that need to be determined.
7. Obtain the clustering outcomes and try to interpret and evaluate the resulting clustering structure.

The majority of clustering techniques are based on iterative optimization algorithms, which have some form of statistical clothing such as significance tests based on assumed distributions, probability models, and error functions, for instance the *K*-means algorithm and its variations thereof (MacQueen 1967; Lloyd, 1982; Hardigan and Wong, 1979). Another group of clustering algorithms is based on spatial relationships and topology, such as the nearest neighbour analysis (Clark and Evans, 1954) and density-based clustering (Ester et al., 1996).

Depending on the underlying philosophy of the algorithm, cluster analysis methods can be divided into two broad categories, partitioning (sometimes referred to as bottom-up or unnested) and hierarchical (also referred to as top-down or nested) methods (Struyf et al., 1997):

- Partitioning methods:

  This category includes algorithms that divide the observations of a dataset into *k* number of clusters, based on some form of objective function that the algorithm aims to minimize (e.g. variance). These methods force the user to specify the *k* value in advance. Typically, the algorithm also supplies some kind of quality index that allows the user to select the value of *k*, after running the algorithm for a range of *k* values. One of the most common partitioning algorithms is *K*-means.

- Hierarchical methods:

  The methods in this category yield an entire hierarchy of clustering of the observation dataset. Hierarchical clustering methods can be further divided into two types, agglomerative and divisive. Agglomerative algorithms generally start at a state where each object in the dataset forms its own cluster, and then successively merge clusters based on some form of similarity criterion until only one large cluster containing all observations remains. Divisive algorithms work in reverse, i.e. they start by considering the whole set as one cluster, and then split up clusters into two until no cluster has more than one object.

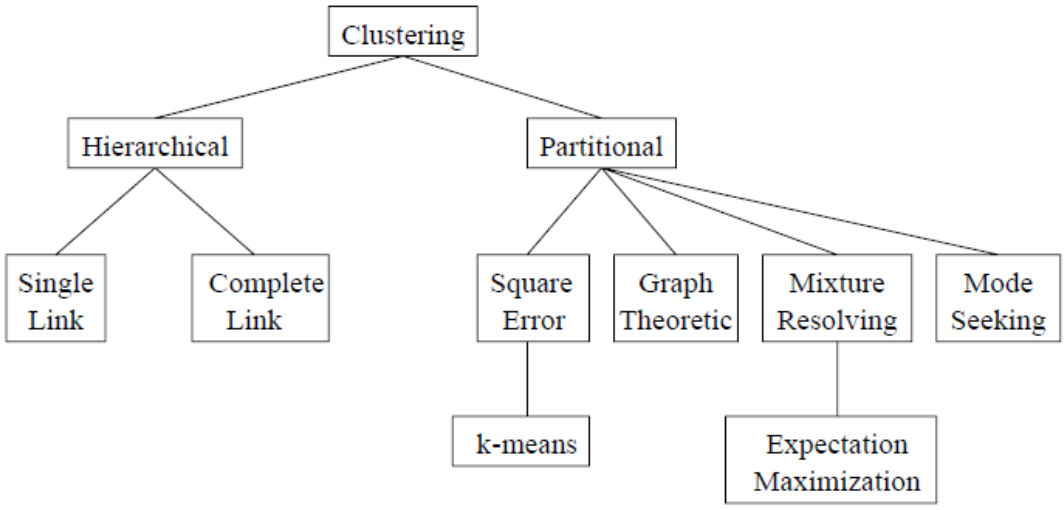

Figure 3.1. The taxonomy of clustering techniques (adopted from Jain et al., 1999).

Generally speaking, partitioning algorithms are the most straightforward and least computationally expensive compared to hierarchical methods. The latter have inherent large computational and storage requirements (since they require a dissimilarity matrix). Moreover, the fact that all merges are final can be problematic for noisy, high-dimensional data, such as document data (Tan et al., 2008). Hierarchical methods however work more comprehensively. They produce a tree of the cluster hierarchy, from the complete dataset to every object individually, where the researcher can explore meaningful data divisions and decide the optimal cut-off point.

There are also other means of differentiating clustering methods (Fig. 3.1), for instance in terms of cluster formation: clustering algorithms can either be *exclusive* or *overlapping*. In exclusive algorithms every object has membership to exactly one cluster. A non-exclusive or overlapping clustering algorithm reflects situations where an object can belong to more than one

group simultaneously. Fuzzy clustering on the other hand is also a type of non-exclusive clustering where an object is given a membership "weight" based on the probability of that object belonging to that particular class (Bezdek, 1981). Due to the probabilistic nature of fuzzy clustering, the sum of all the probabilities for every cluster should amount to 1. These types of algorithms have been employed when the underlying nature of the objects is more likely to be arbitrary (hence the term), although fuzzy classifications are often converted to exclusive by assigning each object to the cluster in which probability is the highest (Tan et al., 2008).

While the above methodological variations assign all objects to one or more groups, there are also situations where such an approach should be avoided. *Partial* or *non-exhaustive* clustering happens when not all objects are assigned into clusters. The motivation behind this approach is that sometimes data points may not represent actual objects, for instance noise. In a similar manner, some objects may not belong to any particular cluster and much like outliers they would skew results. In an analysis of retail centres for example, clusters could be represented by the focal group of venues around a retail centre. There could be other retailers scattered in between centres, but if they are beyond a critical distance, they should not be assigned to any retail cluster. Most density-based algorithms are considered to be partial clustering techniques, such as ISODATA (Ball and Hall, 1965) and DBSCAN (Ester et al., 1996), although these techniques could be converted to complete, for example the OPTICS algorithm (Ankerst et al, 1999). More details on these clustering techniques are presented in the following sections of this Chapter.

3.3.3. Cluster types

Everitt et al. (2011) state that the term cluster has been commonly defined in terms of internal cohesion and isolation of data points, although he also suggests that no single definition can be sufficient to describe clusters in all possible situations. Tan, Steinbach and Kumar (2008) however provide a useful categorization of cluster types defined as data structure notions:

- Well-Separated: The set of objects inside a cluster are more similar to themselves than they are from any other objects.
- Prototype-Based: The cluster objects are more similar to a prototype object that defines the cluster, than to any other prototype. For example, the *K*-means algorithm produce prototype-based cluster as cluster means, or medoids (*K*-medoid algorithm).

- Graph-Based: When data are represented as a graph where the nodes are the objects and the links represent the network connections to any similar object, much like the contiguity-based clusters. An example of such a cluster output is produced by the OPTICS algorithm.
- Density-Based: A cluster is a dense region of objects that is surrounded by a region of no or low density objects (i.e. the DBSCAN algorithm).
- Shared-Property: Often called conceptual clusters, this type of clusters is generally defined as a set of objects that share a common property.

It should be noted that these types are not distinctive of one another, and a cluster can belong to multiple types simultaneously. However, they can give a good yet simple approximation of the notion of cluster presented in an analysis.

In the geodemographics context, neighbourhoods are grouped together based on a collection of attributes to which the population share commonalities with. In this sense, output clusters can be considered prototype-based. For example, the cluster *2: Cosmopolitans* in the 2011 Output Area Classification is defined by a population living in densely populated urban areas, living in flats, with a high ethnic integration etc. The population of such a neighbourhood is very similar to that neighbourhood type (prototype) than any other neighbourhood type. This cluster notion is also an inherent outcome of *K*-means algorithm used in the cluster analysis to produce the OAC classification.

3.3.4. Similarity measures

Before a clustering method is applied, some sort of measurement of similarity or dissimilarity (depending on the scope of the methodology) between objects should be established. Two objects are similar when their dissimilarity or distance is small or their similarity large. Dissimilarity is more often used, measured as the distance between two objects on the attribute space. Numerous types of distances have been introduced, depending not only on variable nature but also on the nature of the phenomenon under analysis.

The most popular metric for continuous features is the Euclidean distance of two data objects in multidimensional space. Euclidean distance is found to work well whilst a data set has cohesive clusters, however its drawback being its tendency of large valued attributes to dominate cluster formation (more so if the squared distance is used). To tackle this, some form of normalization of the attributes is important before applying the clustering algorithm. Linear correlation among

features can also distort distance measures, although this can also be dealt with using different weights on the attributes based on their variances and pair-wise linear correlations (Jain et al., 1999).

In general, there can be several types of distance or proximity measures that are appropriate for a given type of data. For continuous data, one other type of popular distance in Euclidean space is the Manhattan distance (also known as city block distance or taxicab, because it measures distances travelled in street configuration). For categorical data, these are usually expressed as binary data (where 0 and 1 represent the absent or presence of a category). Distances used for binary data are the Jaccard measure or the Tanimoto distance (Everitt et al., 2011), whereas the angular separation metric (cosine similarity) is often used for high-dimensional or non-Euclidean space, for example in order to cluster documents.

Generally speaking, there are many cues in data mining research that suggest that Euclidean distance does not behave well in high-dimensional data (Domingos, 2012). High-dimensional data in Euclidean space challenging dataset for clustering algorithms which are often subsumed under the term "curse of dimensionality". In higher dimensions, as space increasing exponentially data becomes sparse, statistical significance becomes problematic and clustering algorithm efficiency is very low (Zimek, 2012).

In the geodemographic context, most public systems have used Euclidean distance to derive prototype clusters. However, it should be noted that as geographic datasets and computational power become more available, geodemographic cluster analysis will all the more shift towards higher dimensional spaces. It is entirely possible that future systems cannot rely on Euclidean distance as a dissimilarity measure. For the proprietary sector this could already be a reality, as private systems use hundreds of variables to generate profiles (Harris et al., 2005). For instance, one should consider the arc cosine similarity distance measure, which is a variation of the cosine similarity. It measures the arc of the normalized angle between vectors, thus forming a proper distance metric according to the triangular inequality notation. To this point however, very little research has been carried out towards exploring other ways of which large multidimensional data can be handled robustly by the clustering process, at least within academia.

3.3.5. Validation and Inference Procedures

Clustering is within of the exploratory data analyses framework, and as such it depends heavily on the number of a person's decisions that were made during the application; significant

classificatory activity is carried out at a subjective level (Openshaw et al., 1980). Along with cluster analysis, inference procedures have been developed that help decision-making during the various procedural steps and parameterization of the clustering algorithms.

One of the most frequent questions in cluster analysis is how to check whether the clustering of objects actually represents the true nature of the data. This analytical step occurs after the clustering algorithm has been implemented and the solution has been obtained, and arises from the possibility that one will arrive at some reasonable clustering solution even if there are no groups in the data; all clustering methods are heuristic in nature and virtually all clustering algorithms always give a solution regardless. Since there is not one universally accepted method to test this, validation and inference is a crucial part of cluster analysis (Steinley, 2006). These tests are usually carried out on steps (6) and (7) described above, mainly to find the optimum number of clusters and verify that the given method can recover the cluster structure of the dataset.

Validation techniques can be developed either by using an external criterion (external criterion analysis), for instance by using data or information not used in the cluster analysis or an internal one (internal criterion analysis), where validation is carried out using information obtained from within the clustering process (Milligan and Cooper, 1987). Many of the formulaic approaches that have been suggested for choosing the number of clusters, were developed ad hoc for a specific problem and require strong parametric assumptions, or are computationally intensive or both (Sugar and James, 2003).

A variety of validation techniques have been introduced to identify the optimum clusters, from mathematical derivations to empirical datasets and Monte Carlo simulations. Within social science, techniques are usually graph-based, and aid in the decision process rather than command it. After all cluster analysis is an exploratory technique and some empirical evaluation is necessary at every step of the methodology.

A scree plot is a popular way to visualize clustering results, and help decide the number of clusters, as it can be interpreted much like a scree plot used in factor analysis. A scree plot is basically a graph representing the sum of squared error (SSE) or squared distance (SSD) from the cluster centroid (mean) for a number of cluster solutions – often called *scatter*. The total Within-cluster Sum of Squares (WCSS) can be seen as a global measure of error, since as the number of clusters increases, the SSE decreases by definition. By plotting the WCSS against sequential cluster levels can provide a useful graphical way to choose an appropriate cluster amount. An appropriate

cluster solution should be the point at which the reduction in WCSS slows dramatically. This produces an "elbow" in the sequence of the WCSS values (Peeples, 2011).

Some authors advocate plotting the objective criterion and search for the 'flattening' of the curve that indicates the correct amount of clusters K (Thorndike, 1953; Gierl and Schwanenberg, 1998), however, the method is highly subjective and prone to the same criticism as the scree plot in factor analysis (Steinley, 2006). In social sciences, a scree plot is generally acceptable, although the "elbow" of the plot is not always clearly visible, as shown in Figure 3.2 (Alexiou and Singleton, 2015a):

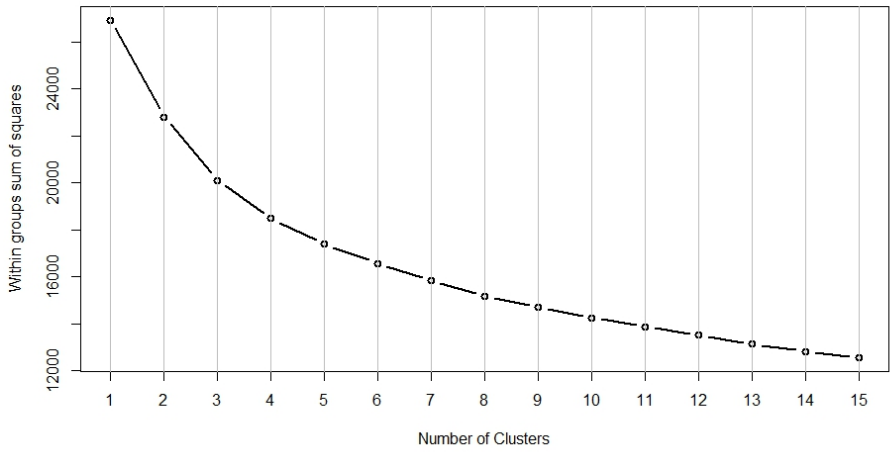

Figure 3.2 K-means: distance from mean by cluster amount, from the Liverpool Classification (Alexiou and Singleton, 2015a). In this case the authors selected 5 clusters for their geodemographic analysis.

The GAP-statistic has been proposed in the statistical literature as a measure that formalizes the notion of elbow in an SSE plot. Tibshirani and colleagues (2001) suggested a formula that calculates a statistic, called GAP, which can point to the optimal number of clusters. Essentially, the optimal number of cluster happens when the GAP statistic is the largest.

There are also more sophisticated graph approaches such as the "silhouette" method introduced by Rousseeuw (1987). This cluster evaluation method is proposed for partitioning techniques using a graphical display of the cluster's cohesiveness. Each cluster is represented by a silhouette, which is based on how similar the cluster is in terms of its own members (tightness) and other clusters (separation). The entire clustering can be combined into a graph, where silhouette width average provides an evaluation of clustering validity. For the purposes of hierarchical clustering, another graph-based evaluation technique is the clustergram (Schonlau,

2002), a variation of the traditional dendrogram for large datasets that explores how cluster members are assigned to clusters as the number of clusters increases.

Milligan and Cooper (1985), provide a review of statistical methods using over 30 decision rules. In general, most of these techniques use a combination of various forms of sum of squares within or between clusters, or both. The Calinski-Harabasz (VRC) criterion for clustering evaluation for instance measures the ratio of the between cluster variance to the overall within cluster variance, and has been found to work well with *K*-means using Euclidean distances (Calinski and Harabasz, 1974). Similar indices are the index of Krzanowski and Lai (Krzanowski and Lai, 1985) and Hartigan's rule (Hartigan, 1975).

Various procedures also exist to test whether there is a significant cluster structure underlying the data, or test whether the output clustering structure is cohesive enough. For example, Bock (1985) provides a series of test statistics for cluster results in terms of average similarity, minimum within-cluster sum of squares, largest gap between observations and the resulting maximum *F* statistic. Milligan (1980) gives a useful Monte Carlo exploration on the performance of various clustering algorithms when the cluster structure is hidden by some sort of error. Specifically, he looks at error terms regarding random perturbation of interpoint distances, outlying data points and a variable insertion that is irrelevant to the cluster structure. He then tests the performance of 15 algorithms in terms of internal cohesion and external isolation of clustering outcomes, and concludes that some algorithms are more prone to particular types of errors than other. For instance, *K*-means had poor recovery performance based on the initial random seed selected.

For non-Gaussian clustering and mixture models, some Bayesian methods have been introduced (Banfield and Raftery, 1993) in order to tackle some methodological weakness of conventional maximum likelihood algorithms. Most clustering techniques are based on a maximum likelihood approach, typically using the sum of squares criterion, which may give problematic results due to overfitting (i.e., the phenomenon when algorithms find highly tuned models that fit only part of the data perfectly).

In the geodemographic context, most cluster evaluation methods rely on the WCSS of data points or similar types of internal criteria to evaluate cluster cohesiveness. Some studies also use external validations to explore uncertainty, for instance using another classification to evaluate cluster consistency (Riddlesden and Singleton, 2014).

Within retail analysis, common market segmentation evaluation techniques involve the use of the Lorenz curve and Gini Coefficient as a way of assessing the level of the market penetration of

clusters (Novak et al., 1992). Similar techniques have also been employed to evaluate health group penetration potential of the 2001 OAC clusters (Petersen et al., 2007).

3.4. Clustering algorithms

One of the main issues when selecting a method is the type of attributes used to define objects. Such data types can be 1) continuous, for instance a measurement of income, 2) discrete, for instance type of housing or 3) binary, for instance the existence of central heating. Ordinal variables are usually represented as continuous or nominal (Everitt, 2011), although there are similarity measures that can handle ordinal data efficiently, like Gower's general dissimilarity coefficient of (Gower, 1971). In order to conduct any type of cluster analysis, one must take into account a) the type of variables under consideration and b) the measurement of similarity that is going to be used to identify clusters, as discussed in a previous section.

Data types and similarity are very closely related, and various algorithms have been developed in order to best handle situation with specific data types. Data distributions have also a significant role to play. Some algorithms assume or work best when underlying data distributions are identified. Typical partitioning algorithms like the *K*-means work best when the distribution of data is Gaussian (Steinley, 2006). There are also "non-parametric" algorithms that have no significant underlying assumptions constraining them, like the Self-Organising Maps (Kohonen, 2001). However, since a metric of distance is involved in almost every algorithm, data points need to be reconfigured to a unified scale prior to clustering.

In this context, the following sections present a small set of algorithms that are popular within the social science domain, with a focus on continuous data.

K-means

One of the most common techniques used to identify geodemographic clusters is the iterative allocation – reallocation algorithm, known as K-means. The K-means clustering method aims to produce externally isolated and internally cohesive clusters. Historically, many researchers attempted to describe this process and independently developed the K-means method as a strategy that attempts to find optimal partitions (MacQueen 1967; Lloyd, 1982; Hardigan and Wong, 1979). Since this development, K-means has become extremely popular, earning a place in several textbooks on multivariate methods (Steinley, 2006).

K-means assigns N observations into K clusters in such a way that within each cluster, the average distance of the variable values from the cluster mean is minimized. Taking into account that for any set of observations S there is an argument that describes the minimum squared distance defined as:

$$\bar{x}_s = \underset{m}{\operatorname{argmin}} \sum_{i \in S} \|x_i - m\|^2 \qquad (3.1)$$

Then for the aggregate of the total clusters there is a set of arguments that minimize the total within cluster variation of the multidimensional data points:

$$WCSS = \min_{c} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2 \qquad (3.2)$$

where WCSS or SSE is the within-cluster sum of squares for a cluster distribution C with K seeds, $x_i \in N$ is the data observations and $\bar{x}_k$ is the *k* cluster mean. Briefly, the group of *K*-means partitioning methods work in the following fashion (Aldenderfer and Blashfield, 1984):

Step 1. Select K points as initial centroids.

Step 2. Allocate each data point to the cluster that has the nearest centroid.

Step 3. Compute the new centroids of the clusters after all data points have been allocated.

Step 4. Repeat step 2 and 3 until no data points change clusters or centroids do not change, at which point the algorithm converges.

K-means is typically initiated with a random set of initial seeds, and then the algorithm assigns every observation to a seed based on the least squared distance. New data means based on the assignments are then calculated, and observations reassigned to their nearest cluster mean, again based on the least squared distances. The algorithm converges when the when the cluster means and cluster assignments no longer change, i.e. when the algorithm finds a local optimum for which the within-cluster sum of squares is minimized.

This technique is one of the least computationally expensive and most straightforward method used to classify multidimensional inputs. *K*-means clustering uses squared distance as a dissimilarity function, but it can handle a variety of variable types, provided an appropriated distance measure is used. There are a number of choices that are guaranteed to provide convergence, for example Euclidean distance, which can be used for continuous variables, or cosine distance for document types.

However, the algorithm needs a specific predetermined number of clusters (K), and furthermore, classification results differ based on the initial k centres that are selected. Research has shown that *K*-means provides a local optimum that does not always represents the global one, and to avoid local optima, it is typical to run *K*-means multiple times for an analysis (Hartigan, 1975), extracting the results for each converged cluster set, and evaluating them on the basis of some metric – most commonly, as an effort to minimise the SSE, thus creating more compact and cohesive clusters).

It should be noted however that just minimizing SSE does not necessarily guarantee a higher quality cluster solution. Research with Monte Carlo analysis on known datasets showed that sometimes K-means fails to provide a solution that represents the cluster structure, even when the cluster solution offers very little variance (Steinley, 2006). As far as other weaknesses are concerned, the algorithm seems to have difficulty handling non-globular clusters and data that contains outliers. In general, K-means is tuned for data that have a notion of centre in individual data clusters but there are variations such as the *K*-medoid algorithm presented below that bypass these restrictions (Tan et al., 2008).

The K-means algorithm is one of the most popular methods in geodemographic research. K-means has been used to produce the Output Area Classification of 2001 (Vickers and Rees, 2007) and 2011 (ONS, 2015b). According to the OAC methodologies, *K*-means also seems to work best in order to produce the neighbourhood type hierarchies (Group and Sub-Group levels), since hierarchical clustering techniques proved to be either computationally very expensive or produce clusters with very few cases.

Many variations of the *K*-means clustering have been developed as it can easily be adjusted for diverse objective functions; for instance, to provide an optimum solution other than cluster compactness. Theiler and Gisler (1997) suggest a variation of the algorithm that incorporates spatial dependencies of data points, a property very useful in some geographical analyses. In particular, under the hypothesis that data points which are spatially contiguous are more likely to be in the same class than are random points, they defined a new objective function that accounts for both spatial contiguity and attribute compactness. This criterion is defined as:

$$E = \lambda D + (1 - \lambda)V^* \qquad (3.3)$$

where *D* is sum of all the (dis)contiguity of data points, expressed as a ratio (percentage) of neighbours that have the same class to the total neighbours of a data point, *V** is the average WCSS of the clustering solution provided by equation (3.2), and λ is a biased parameter taking

values between 0 and 1 and indicating the relative importance of the 2 properties in the classification.

The performance of the algorithm has been illustrated using LANDSAT image data, which seems to perform well for small values of λ, as there are considerable gains to the contiguity of the cluster with virtually no loss in compactness (Theiler and Gisler, 1997). However, the method still uses two biased parameters to produce optimal results which may not always be intuitive. The first is the selection of neighbours; for imagery data, this would be the 8 adjacent pixels, but for other data points, such as geographical areas, adjacent polygons may not always be representative of the locality. Unless polygons are presented in gridded form, some measure of distance to centroid should be incorporated. Secondly, the user must also specify the λ parameter which depends highly on the nature of data clustered.

In general, while contiguity properties may work well with the physical environment (e.g. in order to track points along a road), they may not work well with socio-economic data. Adjacent neighbourhoods are not always similar and a contiguity *K*-means could produce ecological fallacies about the properties of an area, especially since physical barriers between neighbourhoods (such as major roads, railways, waterfronts etc.) are not taken into account when defining neighbours.

K-medoid

The *K*-medoid clustering problem is very similar to *K*-means with the differentiation that it tries to minimize absolute distances rather than squares between points and that it chooses data points as centres through the allocation – reallocation process. The Partitioning Around Medoids (PAM) algorithm introduced by Kaufman and Rousseeuw (1990) is one of the most commonly used algorithms. It aims to find K representative objects, called medoids, among the objects of the dataset. These medoids are computed such that the total dissimilarity of all objects to their nearest medoid is minimal. This differentiation is critical to the notion of cluster presented, as *K*-means delineates clusters as centroids while PAM delineates them as "*exemplar*" points, so a different cluster configuration is produced (Everitt et al., 2011).

Hierarchical Clustering Algorithms

Hierarchical Clustering methods have existed for a relatively long time; they are conceptually simple to understand and therefore still enjoy widespread use. Essentially the method employs

a similarity matrix to sequentially merge (or split) the most similar (dissimilar) cases. Other distinctive properties include that the sequence of clusters formation at every pair of cases can be represented visually by a tree diagram, often called a dendrogram (Fig. 3.3). By default, these methods are nested, in the sense that cluster is a subset of a larger, more inclusive cluster at a higher level of similarity, but also very computationally expensive; a similarity matrix must be computed for the clustering to be applied, which makes hierarchical clustering very ineffective for large, multidimensional data (Aldenderfer and Blashfield, 1984).

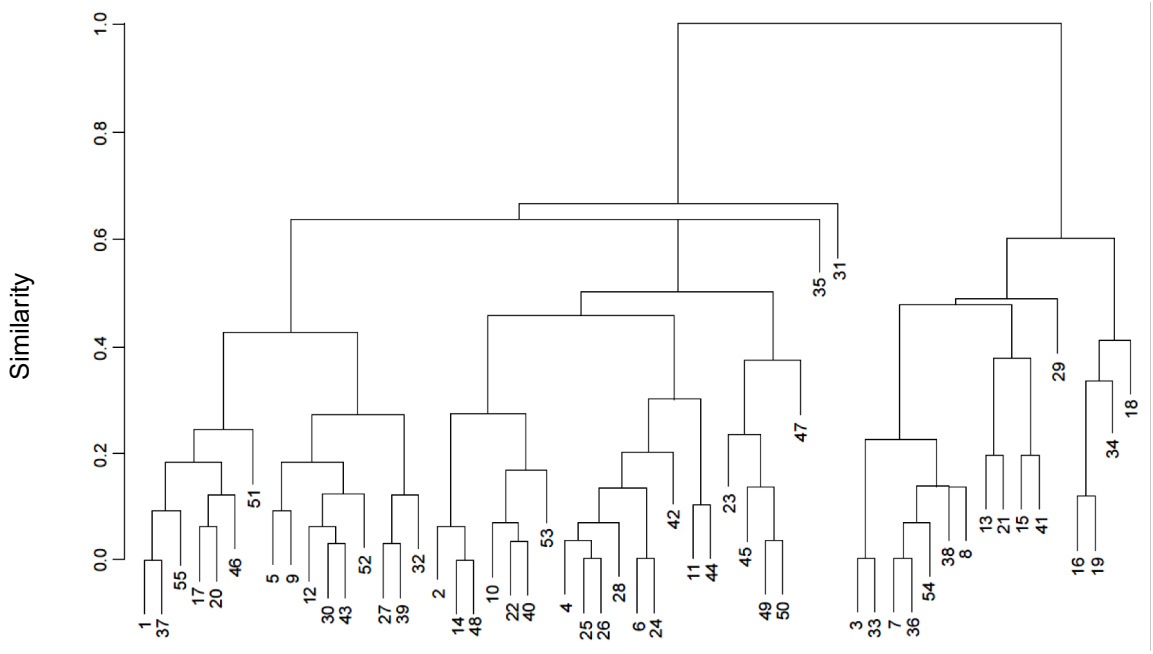

Figure 3.3 An example of a dendrogram showing cluster formation using a dataset describing global cities (source: Everitt et al., 2011).

Hierarchical methods include two major families of algorithms, agglomerative and divisive, however agglomerative hierarchical clustering techniques are by far the most common. Agglomerative methods generally start at a state where each object in the dataset forms its own cluster, and then successively merge clusters based on some form of similarity criterion until only one large cluster containing all observations remains. Some of their variations include the a) *Single linkage,* where objects are joined to clusters if at least one of the existing members of the cluster has the same similarity level (i.e. based solely upon single links between cases and clusters, also referred to as *min distance*), b) *Complete linkage*, where linkage rules dictate that similarity to all cluster member (max distance) is essential for an object to be added in that cluster and c) *Average linkage*, which is a combination of the two above. Divisive algorithms on the other hand work in

reverse, starting by considering the whole set as one cluster, and then split up clusters into two until no cluster has more than one object.

Within agglomerative hierarchical cluster analysis, Ward's method or Ward's clustering criterion is designed to optimize the minimum variance within clusters, or specifically the SSE. At the initiation clustering process, each case is its own cluster and the SSE is 0. Ward's clustering criterion can be applied to merge clusters with the least amount of between-cluster variance, thus producing the minimum increase in total within-cluster variance after merging (Ward, 1963). Ward's method has been used in social sciences and geodemographics as a bottom up approach to produce higher hierarchy clusters, while maintaining the loss of variance to a minimum, e.g. from Sub-Groups to Groups and Super-Group respectively.

Density Based Algorithms

Assuming the notion of a cluster types as data points in metric space, the concept of clustering would be that clusters should present regions where data points are dense, separated by regions where data points are scarce. Several approaches have been suggested that can functionalize this notion, such as mode analysis and nearest neighbour clustering (Everitt et al., 2011).

A similar approach that has been popular within the Geography field is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996). This algorithm classifies objects as clusters, if they exist in a dense region, or noise, if they exist in a low-density region. Every cluster is defined by a core and a border (the edge of the cluster) which define a *neighbourhood*. When neighbourhoods are close together, they merge into one - anything outside the neighbourhoods is considered noise. Euclidean distance is typically used as the similarity measure, although other suitable measures could be used without inherent restrictions.

The basic algorithm of DBSCAN requires two parameters to be predetermined that will help identify core points, 1) *Eps*, a distance around the point (radius) and 2) *MinPts*, the threshold of data points in a neighbourhood. A data point is considered core point when there are equal or more than *MinPts* data points within a distance Eps from the data point.

The DBSCAN algorithm is easy to operate, and can be used not only for clustering but also for outlier detection. A disadvantage of the algorithm is that it does not cope well with clusters of varying density, or in high multidimensional space; data points in higher dimensions become increasingly sparse – an issue related to the *"curse of dimensionality*" described in the previous section (Tan et al., 2008, Chapter 8). For the lattermost issue, the algorithm CLIQUE (Agrawal et

al., 1998), proposes dividing the multidimensional space into a grid, where each cell is classified based on density.

Self-Organizing Maps

There are clustering methods, particularly within pattern recognition, that try to imitate neural networks and their computational capabilities to classify cases. A Self-Organizing Map (SOM) is an unsupervised classifier that uses artificial neural networks to classify multidimensional observations in two-dimensional space based on their similarities (Kohonen, 2001). The technique also has the advantage of not assuming any hypotheses regarding the nature or distribution of the data, and responds well to geographic sensitivity.

A further advantage of using a SOM is the capacity to visualise the structure of data values aiding initial data exploration. A SOM typically organize observations by projecting them onto a plane, and through consecutive iterations finds the best configuration of observations so that every observation is most similar to the others closest to them. It effectively uses an artificial neural network to classify space, based on the configuration of attributes that "fit" each neuron.

Typically, the SOM mapping process employs a lattice of squares or hexagons as the output layer called nodes, and the results of the organization can easily be mapped retaining their topology. Figure 3.3 shows an example of a visualization of the distance between each node (representing a group of similar OAs) and its neighbours. Areas of low neighbour distance (in standard deviations) indicate groups of OAs that are similar, while greater distances indicate OA groups that are much more dissimilar to others. The sum of distances to all immediate neighbours is used as data input to classify the nodes through e.g. a conventional K-means or a regionalization algorithm.

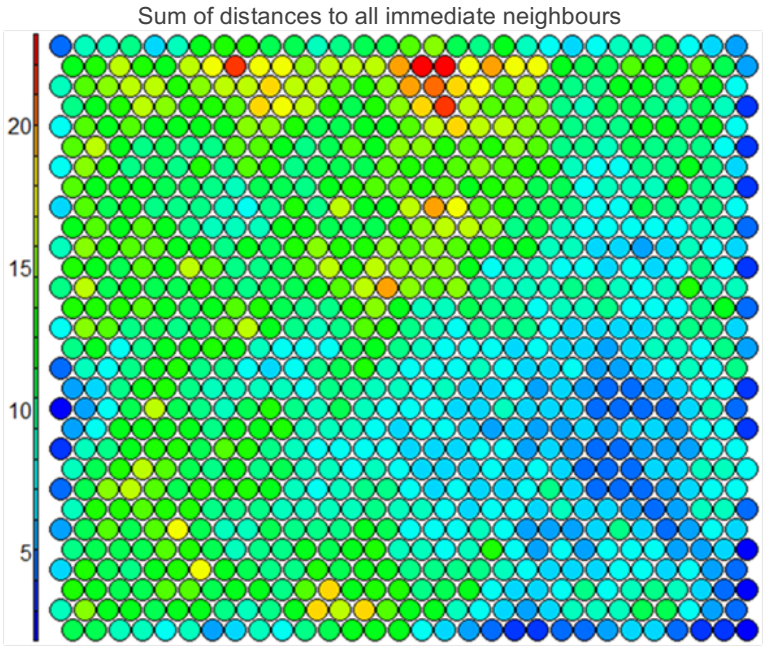

Figure 3.4 An example of an SOM U-Matrix plot on a 30x30 hexagonal grid, using 9 variables from the Built Environment Dataset used in the MODUM classification (Alexiou et al., 2016).

Self-Organizing Maps have drawn attention in the clustering literature because they address the issue of cluster formation and cluster similarity at the same time. One of the problems of classifications is that the groups are discrete; it is not clear just by looking at the classification results how similar or dissimilar data groups are. Clusters are usually described by their means values hence overall similarity is difficult to comprehend, particularly in the multivariate scale. SOMs however produce topological relationships of these groups on a map, giving some information about how similar or dissimilar groups of cases are based on the distance between them on the projected space.

SOMs have many applications in a broad range of fields, from medicine and biology to image analysis and computer science. SOMs have been used in geodemographics as early as the mid-1990's, most notably as an alternative classifier for the creation of the GB Profiles (the other being a conventional *K*-means algorithm). GB Profiles was one of the first, open, geodemographic system for accessing general residential classifications for Great Britain, developed at the University of Leeds (Openshaw and Blake, 1995).

SOMs have also been tested as an alternative classifier of census data (Spielman and Thill, 2008; Arribas-Bel and Schmidt, 2013) where they seem to perform well for socioeconomic data at the US Census tract scale. Arribas-Bel, Nijkamp and Scholten (2011) have also demonstrated the algorithm capabilities to measure urban sprawl in Europe using a similar attribute set,

specifically six variables: connectivity, decentralization, density, scattering, availability of open space and land-use mix. SOMs have also been tested as an alternative classifier of census data in the UK (Openshaw and Wymer, 1995). In this framework, we use an SOM to produce a classification of neighbourhood morphology, illustrated in the next Chapter. SOMs can be very useful when analyzing datasets such as built environment measures, where there are little to non a-priori hypotheses on the underlying distribution and patterns of neighbourhood morphology.

Fuzzy Clustering Methods

Another type of clustering method that has been used to classify cases in a less deterministic fashion is based on fuzzy logic and probability analysis. These clustering methods are known as fuzzy methods or soft classifiers. In fuzzy clustering, objects are not assigned to one cluster but have a membership value indicating the strength of membership to each cluster. Fuzzy clustering algorithms have been developed very early, such as the Gustafson –Kessel algorithm (Gustafson and Kessel, 1979) and Fuzzy C-Means (FCM) algorithm (Bezdek et al., 1984).

Typically, the fuzzy clustering has a probabilistic approach, where 0 represents no possibility of a particular object belonging to a cluster and 1 absolute membership (and by this extent all cluster possibilities should amount to 1). The degree of average membership among cases can also reflect the similarities or dissimilarities between groups.

Similar to fuzzy clustering methods are overlapping, non-exclusive or clumping methods. The difference is that unlike fuzzy clustering, cases are permitted to be part of more than one cluster by nature, so there is no membership function. Development in this research was stimulated by clustering documents in linguistic research since words (cases) could have multiple meanings, hence they should part of more than one cluster (Aldenderfer and Blashfield, 1984, Chapter 3).

The inherent ability of fuzzy classifications to assign cases to more than one clusters with varying membership values have attracted some attention in geodemographics. The advantages of such an approach are that spatial units can be "*ex post facto*" adjusted depending on the value of membership to account i.e. for ecological fallacies and "*neighbourhood effects*" (Feng and Flowerdew 1998). See and Openshaw (2001) also claim that a fuzzy approach to geodemographic classifications could deal with the vagueness of the geodemographic systems methodology, incorporate neighbourhood effects and provide a modelling framework that can be calibrated to the fuzzy targeting mechanism. Most studies regarding geodemographic analysis that use fuzzy classification employ the Fuzzy C-Means algorithm (Feng and Flowerdew 1998; Mason and

Jacobson, 2007; Wijayanto et al., 2015) or the Gustafson-Kessel algorithm (Grekousis and Hatzichristos, 2012).

Regionalization Methods

Regionalization methods are a special case of clustering methods used predominantly in geographic analyses. The aim of these methods is to produce analytical (or functional) regions, which are homogenous yet continuous regions formed by smaller areas grouped together using some geographical criteria. A regionalization algorithm is essentially an aggregation method with an extra constrain on the spatial attribute of the clustering elements (known as spatial contiguity constraint). The idea, also known as "*zone deign*", was first introduced by Openshaw (1977) as a way to geographically cluster multidimensional socioeconomic datasets.

There are various types of algorithms suggested, but depending on purpose but the majority of this algorithms follow some common set of rules (Duque et al., 2007):

1. The number of regions is defined by the user.
2. Spatial units are aggregated based on optimizing an aggregation criterion.
3. The areas within a region must be geographically connected.
4. Each area must be assigned to one and only one region.
5. Each region must contain at least one area.

Regionalisation algorithms were developed because analysis by administrative or otherwise arbitrary, non-homogenous regions (also called normative regions) may induce weak statistical inference and can produce ecological fallacies and aggregation bias (Opensaw, 1977; Fotheringham and Wong 1991; Paelinck, 2000). These issues are generally related to the Modifiable Area Unit Problem (MAUP). The MAUP arises from the inherent analytical limitations when data are collected on a spatial aggregation basis, for instance census data, where population counts are aggregated for zones. The basic premise is that spatial variation is dependent on the aggregation *scale* and zonal *shape* of the zones, thus results for any type of analysis is dependent on these two factors (Lloyd, 2014, Chapter 3). The impact of the MAUP as well as geographic scale in general within geodemographics analysis is further discussed in a following section.

There are several approaches in regionalization methods (Assuncao et al. 2006). One type of which can be described as a two-step process: at the first step a conventional clustering algorithm is applied to identify spatial units with the same underlying characteristics. At the next step, a

spatial contiguity matrix is created which is used to aggregate similar spatial units that are connected geographically (Openshaw 1973; Openshaw and Wymer, 1995).

A second type of regionalization algorithms is those that use spatial relations between units as an optimization process, for instance the popular AZP algorithm (Openshaw and Rao, 1995), which has been used to partition space based on socioeconomic census data. The AZP (Automatic Zoning Procedure) allocates areas into partitions based on the objective function, subject to proximal constraints. Typically, the algorithm works through iterations until it finds a good allocation organization. A similar approach utilizes the geographical coordinates of spatial units as attributes in the clustering process, which can capture spatial patterns in terms of homogeneity, compactness and equality measures (Haining et al., 2001). The latter attribute has been used by Martin (1998) in terms of population, to produce Output Areas for the 2001 and 2011 UK Censuses.

Others variations of regionalization algorithms include the ZDES automatic zoning system (Alvanides et al., 2002), which is a modification of AZP, the Spatial K-cluster Analysis by Tree Edge Removal - SKATER algorithm which uses minimum spanning trees (Lage et. al., 2001) and the max-p algorithm (Duque et al., 2011).

Chapter 4. Geodemographic Analysis

4.1. Introduction

Building a successful classification may seem fairly straightforward but it can be a difficult and highly time-consuming process. Alexiou and Singleton (2015a) outline the methodological steps carried out during a geodemographic analysis and also provide a case study on how to create a geodemographic classification for the Liverpool local authority. Harris, Sleight and Webber (2005) provide more thorough descriptions of the techniques typically used to build geodemographic classifications, particularly methods used in the private sector, and also provide some examples within the UK context. Vickers and Rees (2007) also provide a detailed step-by-step analysis of the process of building the 2001 OAC classification, which was built upon previous work on clustering methodologies by Milligan (1996) and Everitt, Landau and Leese (2011), paraphrased to fit areal classifications.

As aforementioned, it is important that a classification addresses the needs of the classification's stakeholders, but, one must also consider data availability, coverage and weighting. In general, it is vital that the user recognizes the critical decisions that need to be made and their impact on the outcomes of the cluster analysis. From the outset (Webber, 1977), geodemographic methods have typically employed a pragmatic variable selection strategy; combining the experience of the classification builder (what is deemed to work) with the overarching purpose of a classification (what is required), alongside some degree of empirical evaluation.

Aside from academia, little is known about how geodemographic classifications are built within the private sector; commercial geodemographic classifications have an inherent commercial confidentiality, and as such, most of their methodologies remain a "black box", which some have argued impairs not only reproduction, but also scientific questioning of the ways in which the clusters emerged from the underlying data (Longley, 2007; Singleton and Longley, 2009).

Building on the existing methodology presented in the various academic works mentioned above, a geodemographic analysis can be divided into 4 steps:

a) Data Sources, Geographic Scale and Variable Selection,
b) Data Preparation and Evaluation,

c) Clustering Method and

d) Cluster Labelling and Evaluation.

It is important to note that the steps outlined here are not entirely comprehensive. The selection of available data sources, variables and geographic scale for the classification are interconnected, particularly in terms of which variables should be selected that best represent the typology of neighbourhood that the analysis aims. Variable selection is obviously dependent on the data sources, but geographic scale may pose limitations on the availability of such measurements at a small-area level. Furthermore, variables can later be discarded due to high correlations, unfit distributions or missing data values. In general, geodemographic analysis is not a linear process, and it is typical to revisit previous steps for modifications, when the classification outcomes are not comprehensive enough.

Before this analysis proceeds with the core methodological steps, it is important to address a few key issues about data sources and the spatial data infrastructure available in the UK and how that impacts the selection of input variables.

4.2. Data Sources, Geographic Scale and Variable Selection

4.2.1. Data sources

Geodemographics use a variety of data to generate profiles. Information collected on the population characteristics can derive from various data sources, public or private. Geographic data contains information about the socio-economic and build environment attributes at any given geographic space. The basic sources of information are censuses and other population registrations, surveys and remote sensing techniques, such as night-time light imagery (Rhind, 1991). Naturally, data used in small-area classifications must have a reference to a specific geography, meaning they should be attached to some kind of spatial data structure. Most commonly this is directly embedded through the use of coordinates attached to data (e.g. points) or a code that corresponds to a specific feature that already referenced, such as current administrative structures.

Census data in the UK and the majority of government-issued data are spatially attached to various areal administrative units, such as Super-Output Areas, Wards, Postcodes or Output areas. OA or Postcodes are currently the finest areal unit, and it is preferred for detailed analyses

since it provides the highest granularity. Data offered in the private domain is usually more granular as it is collected at the individual or household level, using from various sources, e.g. credit card histories, product registrations and private surveys (Singleton and Spielman, 2013).

Population data in the public domain are usually offered aggregated to some level of administrative geography. Censuses for example are traditionally carried out by taking onto account individual enumerators that the population characteristics are known, and subsequently produce much more detailed geographical subdivisions, which usually nest in a hierarchy. Such a nested model is for instance the Super Output, Middle Super Output, Lower Super Output (LSOA) and Output Area (OA) spatial structure respectively. An alternative approach is to produce census output aggregated to grid cells. This approach has been used for the 1971 Census in the UK, and while it offers advantages of stability over time, some cells are likely to be very sparsely populated and cannot be published due to confidentiality reasons and has been abandoned since (Martin, 2000).

The confidentiality issues are the main reason why census datasets are published in aggregated format. Aggregated attributes are usually offered as population counts per characteristic (e.g. *Age, Education Level, Number of Cars*) while the description denotes the appropriate denominator (e.g. *Total population*, *Population of people over 16 years* and *Total number of households* respectively), which can be easily converted to percentages, with a few exceptions such as population density.

Spatial aggregation however imposes difficulties in producing accurate ecological inferences, as described by the Modifiable Areal Unit Problem. Furthermore, the spatial borders of enumeration districts among censuses have changed over the years. For the 2001 Census, Output Areas were introduced as a level of geography resulting for a population normalization methodology, i.e. to contain approximately the same amount of population and households (Martin et al., 2001). This methodology was reapplied for the 2011 Census, however, since OAs should hold a minimum amount of population or households due to confidentiality, differences in population distribution forced the Office of National Statistics to make changes to 2.6% of those. Furthermore, the number of OAs has increased to account for population growth, from 175,434 (2001) to 181,408 (2011) for England and Wales, imposing further difficulties into the interoperability of spatial data infrastructure (ONS, 2012).

It is also important to note that not all government-issued data is offered in all available geographic scales. For instance, while census data are typically available at the OA level, data on Jobseeker's allowance is available on the LSOA level and data on house prices on the postcode

unit level. Despite the fact that there is a plethora of central or regional data sources openly available (see Table 4.1 for the majority of socio-economic non-census datasets available in the UK), the differences in the spatial structure that they are offered poses difficulties in the creation of one cohesive input dataset. Furthermore, these datasets are offered as snapshots for various time periods and with irregular updating intervals.

Table 4.1 Public open data sources in the UK and available spatial geography.

Data Sources	**Available Geography**
Pensions credit data	LSOA
Attendance Allowance (AA)	LSOA
Disability Living Allowance (DLA)	LSOA
Employment and Support Allowance (ESA)	LSOA
Incapacity Benefit / Severe Disablement Allowance (IB/SDA)	LSOA
Income Support (IS)	LSOA
Jobseekers Allowance (JSA)	LSOA
Pension Credit (PC)	LSOA
State Pension (SP)	LSOA
Children in Low-Income Families	LSOA
Fuel Poverty	LSOA
Child benefit data	LSOA
Accessibility of key services	LSOA
Council Tax Band Data	OA
Workless benefit claimants	OA
House Price	Postcode
Mortgage Lending	Postcode
POLAR – Participation of Local Areas	Ward
School Performance	School
Police Data	LSOA
Electricity and Gas Consumption	LSOA

It is true nonetheless that one of the most valuable dataset in geodemographic analysis is still the decennial Census of the population. Data input used in the cluster analysis of the 2001 OAC comprised of 41 census variables (Vickers and Rees, 2007), while the 2011 OAC used 60 variables, across the domains of demographic structure, household composition, housing, socio-economic character and employment (ONS, 2015b).

In the private sector, variables can also be non-census in nature. Harris, Sleight and Webber (2005) suggest that non-census variable offer some advantages compared to the Census, as the latter generally focuses on demographic and socio-economic disadvantages, and not the advantages of the population. This can be crucial in order to identify those more privileged members of the community. Another advantage is that, as previously mentioned, privately collected data is available at a finer granularity such as household or individual level, which makes it very easy to collate to any level of geography. Lastly, they suggest that non-census data can also be used to provide intercensal estimates, and update or maintain classification systems easily and robustly.

Intercensal estimates can be very useful in research as well, as the 10-year gap between censuses may be too large for an analysis to be useful (for instance, internet usage and online shopping has changed significantly over the last decade). Within geodemographics various indexes have been suggested to examine the temporal stability of the clusters and whether other secondary data sources and internal measures might usefully indicate local high level of uncertainty (Singleton et al., 2016a). Finally, there is a growing trend in the usage of Open Data sources in geographic applications, particularly in academic research. Open Data sources offer a level of transparency in Geographical Information Science that can enhance reproducibility, replication and extension of applied geographic models (Singleton et al, 2016b).

4.2.2. Geographic Scale

An important stage in building a geodemographic classification is to assemble a database of inputs that are deemed important for differentiating areas. The geographical unit of reference used to collate such data will depend on the purposes of the classification, and also pragmatically on those data available to the classification builder for different scales; in this framework, geographic scale and variable selection are intertwined. For example, most open (and some commercial) geodemographic systems in the UK are based on data aggregated at the OA level, which represents an average population of approximately 300 people, and is the smallest scale at which census data is provided. However, different sets of variables can have different scales and there are various ways in which such are managed, ranging from simple apportionment from aggregate to disaggregate scales, small area estimation or micro simulation (Birkin and Clarke, 2011).

Consideration of scale in geodemographics is typically aligned with attribute selection, and the availability of data (including licensing constraints) to the classification builder. Selecting a

"neighbourhood" scale can generally be dictated by the scope of the analysis, the availability of data or even the definition of neighbourhood that the creator has in mind. Clearly, there is no one single universal definition of neighbourhood; in sociological terms, the concept can be expressed as the interrelationships between people and places (Harris, 2003). Various definitions have been proposed, usually in terms of the context in which the term is used. In quantitative analyses for instance, it is defined in terms of a zonal/spatial extent (i.e. city blocks, postcodes) or units enclosed (i.e. population).

Regarding the spatial extents of neighbourhoods, Martin (1998) identifies three different approaches of the neighbourhood definition: informal, formal, and analytical. Informal context suggests a neighbourhood with indeterminable boundaries, often overlapping and with multiple names. Formal context suggests clear-cut boundaries, imposed by an administration or agency for a specific process. Analytical contexts are those used by social geographers to describe an area with discrete boundaries which is based on an objective function, for instance an area that has some degree of homogeneity in terms of social or physical characteristics. Nevertheless, analytical neighbourhoods are frequently reconstructed from formal contexts in order to fit some subsistent spatial data infrastructure.

There is a longstanding debate about the importance of spatial ontologies in geographic analysis and the epistemology of Geographical Information Systems (see for example Pickles, 1995; and Openshaw, 1984). Geographic space is infinitely complex, and spatial scale is consequently also a complex concept with multiple definitions, where information loss in inherent (Goodchild, 2001). Any type of analysis needs to sample and make assumptions on the spatial properties of geographic space in order to make the handling of the data representing those properties manageable (Lloyd, 2014, Chapter 1). The spatial variation of a phenomenon is linked to the spatial scale used to observe to variation (scale dependence), where spatial scale relates to the spatial extents that are used as to provide measurements (Atkinson and Tate, 2000). Spatial scale, spatial measurement and spatial variation are thus interlinked in a geographic analysis.

In spatial analyses applications, it is generally recommended that a range of spatial scales are used in order to explore how statistics change, however in geodemographics this issue has not been systematically addressed. The neighbourhood term is used within geodemographics to describe small-area geography that can be classified into specific types. Since the type of neighbourhood is based on aggregated values of that geographic extent, the type of neighbourhood may change significantly if the aggregation level was to change, either in scale or shape. This issue, known in spatial analysis as the Modifiable Area Unit Problem, imposes several

limitations on the robustness of the classification results (Lloyd, 2014, Chapter 3).

Geodemographics however are heavily data-dependant, and as such, the geographic scale in geodemographics and similar small-area applications are based on the availability of data rather some objective rationale. The spatial extents of neighbourhoods cannot therefore be based only on an analytical point of view; data availability (i.e. formal data structures) is the main discerning factor when considering scale in geodemographic analyses.

For example, the OAC for 2001 and 2011 were created at the Output Area level, the smallest geographic level that census data are available. Another bespoke classification, the Internet Usage Classification (IUC), has been created at the LSOA level as the survey data on internet usage used at input offered very little differentiation or significance at the OA level, and some OAs were significantly under-represented in the survey sample (Riddlesden and Singleton, 2014). On the other hand, even if there is data availability with high granularity, a classification creator could select another geographic scale in order to aid inclusion or correlations with data offered at another spatial level. For example, the input dataset for the Multidimensional Open Data for Urban Morphology Classification (MODUM), comprised entirely from built environment attributes, was first assembled at the building unit level, but then aggregated to the OA level since many other socio-economic classifications are offered at that level, thus making comparisons possible (Alexiou et al., 2016). In general, the aggregation on Census geography allows the incorporation of Census data which are distributed for these units.

4.2.3. Data formation and Variable Evaluation

Assuming the relative datasets used for the analysis have been selected, the next step in the analysis is preparing the data to a format that can be easily handled, and apply a set of exploratory techniques to the variables as part of an evaluation process. In common with practice when creating inputs to multidimensional classifications, preference should be for those attributes which in addition to theoretical rationale also provide useful differentiation between areas (Spielman and Singleton, 2015). Similar to a data exploration approach, datasets selected as input in the classification must be checked not only for value-related inconsistencies (i.e. missing or incorrect values), but also in terms of the impact they could have on the final classification due to cross-correlation, unfit distributions or small sample size.

Attributes can be collected and compiled with a variety of measurement types including percentages, index scores, ratios or composite measures (e.g. principal components, weighting,

etc.), so it is important to prepare them on an equal measurement basis and collate them in a data table format, where lines represent areas and columns represent attributes. This process, defined here as data preparation, is also known in data mining and information theory as *data munging* or *data wrangling*, which are general terms used to describe the conversation of data, often through scripting languages, from "*raw*" format to a format that is easier to work with.

There is often need to restructure, for example, census variables by typically aggregating counts to wider classes prior to the analysis. In cases such as age, ethnicity or car availability, the classification creator should avoid having classes with too few observations, or variables that would not be very meaningful. For instance, a variable with the amount of children aged one and another for children aged two should be group together as a variable representing families with small children (typically 0-4 years). The following table (Table 4.2) gives an example of the variables used in a bespoken classification for Liverpool (Alexiou and Singleton, 2015a), and how census variables are aggregated into wider classes (*data binning*).

Table 4.2 Initial Dataset used for the Liverpool Classification (Alexiou and Singleton, 2015a).

Variables	Variable Definition
Demographic	
V1: Age 0–4	Percentage of resident population aged 0–4 years
V2: Age 5–14	Percentage of resident population aged 5–14 years
V3: Age 15-24	Percentage of resident population aged 15-24 years
V4: Age 25–44	Percentage of resident population aged 25–44 years
V5: Age 45–64	Percentage of resident population aged 45–64 years
V6: Age 65+	Percentage of resident population aged 65 or more years
V7: Ethnic Group, White	Percentage of people identifying as white
V8: Ethnic Group, Black	Percentage of people identifying as black African, black Caribbean or other black
V9: Ethnic Group, Asian	Percentage of people identifying as Indian, Pakistani, Bangladeshi, Chinese or Other Asian
V10: Population Density	Number of people per hectare
Housing	
V11: Privately Owned	Percentages of households that are privately owned
V12: Rent (Private):	Percentage of households that are private sector rented accommodation

V13: Rent (Public):	Percentage of households that are public sector rented accommodation
V14: Detached	Percentage of all household spaces that are detached
V15: Semi-Detached	Percentage of all household spaces that are semi-detached
V16: Terraced	Percentage of all household spaces that are terraced
V17: Flats	Percentage of households which are flats
V18: Central heating	Percentage of occupied household spaces with central heating
V19: No central heating	Percentage of occupied household spaces without central heating
Economic Activity	
V20: Working full-time	Percentage of household representatives who are working full-time
V21: Working part-time	Percentage of household representatives who are working part-time
V22: Unemployed	Percentage of household representatives who are unemployed
V23: Retired	Percentage of household representatives who are retired
V24: Student	Percentage of household representatives who are full-time students
V25: No Qualifications	Percentage of people over 16 years without further education qualifications
V26: Higher Education	Percentage of people over 16 years for which the highest level of qualification is level 4 qualifications and above
V27: No car household	Percentage of households with no cars
V28: 1 Car household	Percentage of households with 1 car
V29: 2+ Car household	Percentage of households with 2 or more cars

This process can be very time-consuming. When managing quantitative data, in most cases variables will not seem appropriate to use in their current format. When preparing data observations, it is important to remember that sometimes variables have varying propensities among different groups of people, typically by age or sex (Table 4.3). For instance, long–term illnesses indexes frequently have higher values between groups of older people. An area that has a higher ratio of older to younger people will ceteris paribus tend to have higher rates of illnesses as well. In these cases, age standardization is recommended since it can scale values in accordance to age structure; scaled ratios are calculated as the sum of the age-specific rates multiplied by the area population per age group. If area specific rates are not provided, they can be obtained from

the national or regional average.

Table 4.3 Common techniques used to format data points.

Obtaining ratios per areal unit		
Percentages	$x'_{a,i} = \frac{x_{a,i}}{P_a}$	where $x_{a,i}$ is the attribute value *i* of area *a* and P_a is the population of reference (denominator) of area *a*, i.e. total population, number of households, etc.
Standardized by group	$x'_{a,i} = \frac{x_{a,i}}{\sum_g r_{N,g} P_{a,g}}$	where $x_{a,i}$ is the attribute value *i* of area *a*, $r_{N,g}$ is the observed national ratio *N* for group *g* and $P_{a,i}$ is the population of group *g* in area *a*.
Ratios	$x'_{a,i} = \frac{x_a}{y_a}$	where x_a is the attribute value of area *a* and y_a is another value of area *a*, i.e. density (population / area).

The next step of the analysis is the variable evaluation. Available data can have skewed distributions, contain a high rate of missing values or originate from small sample sizes smaller than desired, thus generating uncertainty. In general, a detailed assessment of each variable is typical prior to the clustering process in order to identify "unfit" data.

Evaluation typically includes mapping, distribution plots (such as density plots and histograms) and correlation analysis. During this step, some of the previously selected variables may be excluded from the rest of the analysis for a number of reasons. It is customary to start with a larger pool of variables when carrying out an analysis and then progressively removing those that seem problematic or are likely to skew results. To illustrate, the 2011 OAC initially considered over 167 variables but only 60 made it to the final classification (ONS, 2015b).

It is suggested that attributes with very high correlation between them (cross-correlation) should be avoided, as they effectively measure the same phenomenon (Harris et al., 2005). For instance, consider that the UK Census provides two variables regarding households and central heating: a count of households with central heating, and a count of households without one. Including both of these variables will effectively duplicate the impact of the central heating attribute across the classification. However, there is no definitive rule to a clear cut-off correlation point. On the other hand, some of the highly correlated variables can be retained since they could capture behavioural variation across areas, which could be interpreted as pairs of variables having

significant descriptive and predictive power (Voas and Williamson, 2001).

Assuming the set of variables has been selected, some data transformation could be carried out prior to feeding the variables into the clustering procedure. In order for a clustering algorithm to work efficiently, some conditions regarding the data structure must be first met. These conditions may vary depending on the algorithm, but for this research the conventional *K*-means algorithm will be used to create the classification, so the data transformation procedure will be discussed having a *K*-means algorithm in mind.

Data transformations that are applied, if any, will be applied to each of the variables in order to meet certain conditions regarding the measure of distance that will be used as an objective dissimilarity function. The distance measure must have the same measuring scale across all variables and not violate the triangle inequality property. Furthermore, algorithms such as *K*-means work best when data point distributions are normal, so sometimes transformation of variables to approach the normal distribution is recommended in order to better recover the cluster structure. These two actions are described here as *Variable Standardization* and *Variable Transformation* respectively.

Variable standardization refers to the universal scale of measurement that should be applied to every observation prior to clustering, such as range standardization or standardized *z*-scores. Variable transformation refers to various functions that can be used to replace the variable measurements of cases to said function of values. For instance, power transformations are a group of functions that replace values to a power of that value, for example *x* to *log(x)*.

Variable standardization is directly related to the distance or dissimilarity measure applied in the clustering process. Standardization would seem necessary in those cases when a dissimilarity measure such as Euclidean distance is selected which is sensitive to differences in the magnitudes or the variability of scales of the input variables (Milligan and Cooper, 1988).

Standardization makes sure distances are measured in the same units, so every variable can contribute to the sum of distances under the same conditions. Since variables can be collated while originating from different contexts (ratios, counts or sum of values), disproportionate measurements will frequently affect the dissimilarity function of the clustering technique towards variables with higher values. For instance, a variable that ranges between 0 and 1 will outweigh, in terms of impact contribution, a variable that ranges between 0 and 100.

The standardization process is applied in order to transform the data in order to equalize range and/or variance. Two of the most common functions applied are *z*-scores and range standardization. The *z*-scores function is transforms the data points so that every variable has the

same mean (0) and equal standard deviation (1), but produces different ranges. Range standardization techniques on the other hand produce values with equal ranges (e.g. 0 to 100) but with different means and variances.

The selection of an appropriate standardization technique is associated on how outlier values will be handled in the dataset. Range standardization will produce variables that will have less impact when there are extreme outliers present, while *z*-scores will produce variables that will have more impact when there are extreme outliers present. Assuming a normal distribution, in the first case, the majority of cases will have a very low value (e.g. close to 0), adding very little the overall distance measure, while in the second case the majority of variables will be between 0 and 1, as expected, but the outliers will gain very high values (possibly in the dozens of standard deviations) outweighing the impact factor of that variable, as the sum of its distances becomes greater.

Range standardization was used as the best choice during the creation of the OAC 2001 and OAC 2011 classifications as well as the ONS 1991 classification of local and health authorities (Wallace and Denham, 1996). Indeed, a study by Milligan and Cooper (1988) found that, through a series of simulation of artificial data configurations, range standardization actually performs better compared to *z*-scores in terms of cluster recovery, for a range of error terms. However, the study used hierarchical agglomerative methods to produce results and *K*-means that was used in these instances.

It is important to note that there are also other variations of standardization functions like the interdecile range and rank standardization. A list of functions related to standardization is presented in Table 4.4. Lastly, when outliers are being very problematic, capped range standardization is suggested (e.g. when handling densities), where variables after a specific quantile (e.g. 99^{th}) are capped to that value.

Table 4.4 Variable transformations used for standardization / scaling (based on Milligan and Cooper, 1988).

Variable Scaling		
z-scores	$z_i = \frac{x_i - \mu}{\sigma}$	Where μ is the mean, and σ is the standard deviation of the variable.
Equal Variance	$z_i = \frac{x_i}{\sigma}$	Where σ is the standard deviation of the variable.

Range standardization	$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$	Where x_{min} is the minimum value and x_{max} the maximum of the variable.
Interquartile range standardization	$z_i = \frac{x_i - x_{Q2}}{x_{Q3} - x_{Q1}}$	Where x_{Q2} is the quantile at 50% of the values, x_{Q3} at 75% and x_{Q1} at 25%.
Interdecile range standardization	$z_i = \frac{x_i - x_{Q2}}{x_{90} - x_{10}}$	Where x_{Q2} is the quantile at 50% of the values, x_{90} at 90% and x_{Q1} at 10%.
Rank standardization	$z_i = Rank(x_i)$	Transformed variable with mean *(n+1)/2*, range *n-1*, and variance *(n+1) [((2n+1)/6) - ((n+1) /4)]*.

Another issue lies within the distribution of the variables themselves. It is important to note that standardization does not alter the shape of the distribution; only the measurement units. Variable transformations are used for a number of reasons, most notably in order to reduce skewness, produce equal variance spreads (the problem of *heteroscedasticity*) or simply convenience (for instance using the log of values in order to look at linear relationships than curved).

For instance, by transforming the data points to a logarithmic scale, the absolute distances between values at the highest extremities of the dataset are reduced more that the distances between smaller value. Therefore, the problem of outliers dominating the cluster formation is reduced considerably. To illustrate, the majority of Census variables are measured in counts and when transformed to percentages usually have a right skew. For example, when looking at the 2011 Census variable *KS404: Number of cars or vans per household*, the number of classes are *None*, *One*, *Two*, *Three* and *Four or more*. The case of *Four or more* is likely to be a lot cases with zero or very small percentages, some cases with moderate percentages and a few cases with very high percentages as shown in Figure 4.1. When addressing socio-economic dimensions some phenomena that are rare are typically concentrated in a few number of cases. This produces highly right skewed data distributions.

A popular method to mitigate outlier impact when dealing with percentages is the inverse hyperbolic sine, calculated as:

$$\sinh^{-1} \mathrm{x} = \ln\left(x + \sqrt{x^2 + 1}\right) \qquad (4.1)$$

The inverse hyperbolic sine, unlike the logarithmic transformation, does not hold the place of zero and it responds very well to positive and negative values. The OAC of 2001 used logarithmic scale

to transform the variables while the 2011 OAC used the inverse hyperbolic sine function (ONS, 2015b).

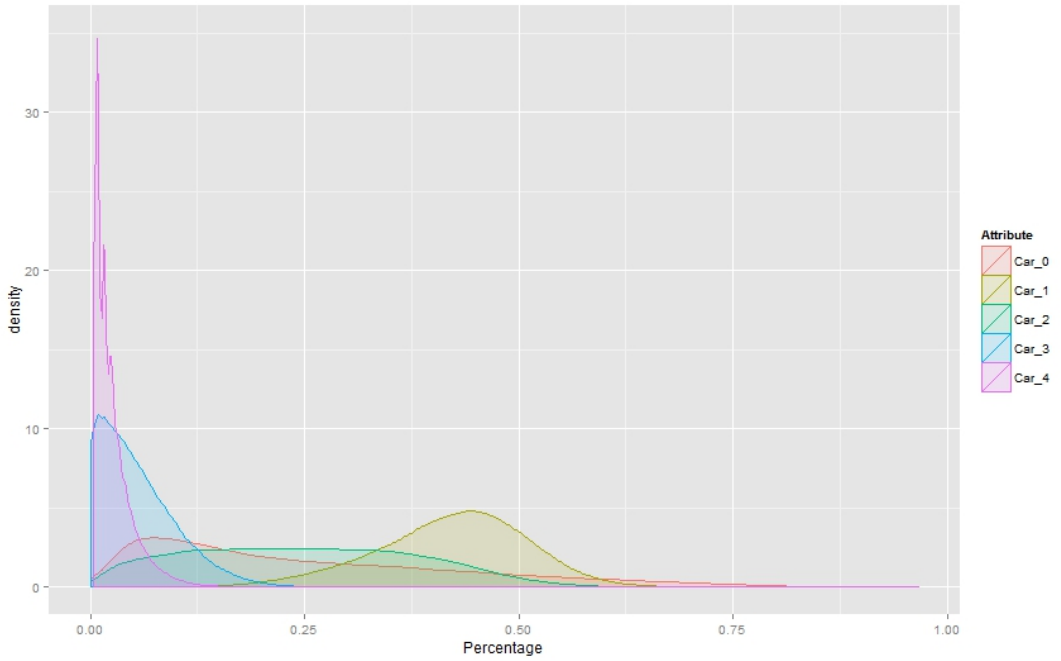

Figure 4.1 Density plot showing the variable distributions of car availability (%) for England and Wales at the OA level (kernel = 512) (Data Source: ONS, Census 2011).

Another issue that has been identified as problematic for effective cluster formation, particularly with the commonly used *K*-means clustering algorithm, is the non-normality of the variable distribution. A vital step in the data preparation process is the transformation of the variable values to approximate normal distributions, usually through various power transformations. Normalization is also required in order to apply any parametric tests, like the Pearson correlation.

While there are statistical tests for checking the normality of a distribution, researchers tend to report favouring visual inspection, i.e. "*eyeballing the data*", of the variable or the error term distribution (Orr et al., 1991). This can be achieved through either plotting a variable histogram or a *q-q plot*, a plot that provides visual comparison of the sample quantiles to the corresponding theoretical quantiles.

A popular method of normalising variable distributions systematically is the Box-Cox transformation (Box and Cox, 1964). The Box-Cox transformation is an iterative technique that is applied in order to find the best λ power transformation for x, in order for $\frac{x_i^{\lambda}-1}{\lambda}$ to approximate

the normal distribution. For instance, for λ=1 there is no transformation applied. Typically, the user predefines λ on a range of values (i.e. from -1 to +3), and with an appropriate step (i.e. 0.2). Since λ can take any number of values, it is possible to find the best λ that suits the data transformation (Osborne, 2010).

Table 4.5 summarizes the transformations used in geodemographics to deal with skewed data observations that are commonly associated with the Census of the population.

Table 4.5 Variable transformations used for normalization.

Normalization Transformations		
Box – Cox	$x'_i = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda}, & if\ \lambda \neq 0 \\ \log x_i, & if\ \lambda = 0 \end{cases}$	The power λ achieves the best normalization and can be estimated algorithmically
Square root transformation	$x'_i = \sqrt{x_i}$	When distribution is slightly right-skewed
Log transformation* *(holds the place of zero)	$x'_i = \log x_i$	When distribution is very right-skewed
Inverse hyperbolic sine	$x'_i = \sinh^{-1} x_i$	When distribution is very right-skewed but does not hold the place of zero
Reciprocal transformation	$x'_i = x_i^{-1} = \frac{1}{x_i}$	When distribution is extremely right-skewed
Square transformation	$x'_i = x_i^2$	When distribution is left-skewed

Some authors however argue that variable transformations obscure differential information about clusters and should be avoided, and suggest a differential weighting algorithm on variables could better recover cluster formation (De Soete et al., 1985; Hohenegger, 1986). In the private sector, when normalization of the variables is not applicable, a similar technique is applied that adjusts the weighting of variables to adjust their influence on the final classification. These treatments include simple weighting carried out on the basis of what is deemed appropriate by the classification builder, or using principal component analysis to identify common vectors of variables that help reduce data complexity and noise (Harris et al., 2005).

4.3. Clustering Approaches and Techniques

Clustering approaches and techniques can differ quite radically, depending on the purpose, but also on the nature of the data to be clustered. For instance, *K*-means seems to perform well with socio-economic data, but not, for instance, with built environment data where data points are generally sparse. Built environment features are generally more concentrated spatially (a relatively few neighbourhoods have access to surface water, parks, major roads or railway stations). A classification of retail places on the other hand should not only be based on the retail site attributes but also on the overall site density to create a typology; as such a density based clustering algorithm should be more appropriate. Furthermore, a geodemographic typology should usually be presented as a hierarchy; with different clusters produced for varying tiers of aggregated areas (Table 4.5). In order to archive this, there can be multiple clustering methods applied to create the geodemographic system.

Table 4.6 An example of the nested hierarchy for one of the Super-Group cluster "*Blue Collar Communities"* from the 2001 Output Area Classification.

Super-Group	*Group*	*Sub-Group*
1: *Blue Collar Communities*	1a: *Terraced Blue Collar*	1a1: *Terraced Blue Collar (1)*
		1a2: *Terraced Blue Collar (2)*
		1a3: *Terraced Blue Collar (3)*
	1b: *Younger Blue Collar*	1b1: *Younger Blue Collar (1)*
		1b2: *Younger Blue Collar (2)*
	1c: *Older Blue Collar*	1c1: *Older Blue Collar (1)*
		1c2: *Older Blue Collar (2)*
		1c3: *Older Blue Collar (3)*

Such hierarchies can be created from the top or the bottom. A top-down approach includes the creation of larger groups of cases that are subsequently divided into smaller sub-groups. This method is typically implemented with hierarchical agglomerative clustering or *K*-means clustering algorithms. *K*-means for instance was used to produce the 2001 OAC, specifically 7 Super-Groups, which were respectively split into 21 Groups and further into 52 Sub-Groups. The same approach was carried out to create the 2011 OAC, only with a different amount of clusters.

As aforementioned in Chapter 3, *K*-means is typically initiated with a random set of initial

seeds, and then the algorithm assigns every observation to a seed based on the least squared distance. New means based on the assignments and then calculated, and observations reassigned to their nearest cluster mean, again based on the least squared distances. The algorithm converges when the within-cluster sum of squares is minimized, i.e. when the cluster assignments no longer change. This technique is the easiest and most straightforward method used to classify multidimensional inputs; however, the algorithm needs a specific predetermined number of clusters (K), and furthermore, research has shown that classification results differ based on the initial K centres that are selected. As such, it is typical to run *K*-means multiple times for an analysis, extracting the results for each converged cluster set, and evaluating them on the basis of some metric – most commonly, an effort to minimise the within sum of squares (i.e. a more compact, and therefore homogeneous clusters).

A bottom-up approach to clustering methods is more prevalent within the commercial sector. It includes the creation of numerous smaller groups using a *K*-means algorithm, which are then aggregated based on their similarities into larger groups, typically with hierarchical algorithms such as Ward's clustering criterion (Alexiou and Singleton, 2015a). Ward's clustering criterion can be used in geodemographics as a bottom up approach to produce higher hierarchy clusters, while maintaining the loss of variance to a minimum, e.g. from Sub-Groups to Groups and Super-Groups respectively.

There are other clustering methods used in geodemographic classifications, most notably the SOM approach which was used as an alternative classifier of census data in the UK to produce the GB Profiles (Openshaw and Wymer, 1995) and fuzzy classification studies that employ the Fuzzy C-Means algorithm or the Gustafson-Kessel algorithm (Feng and Flowerdew 1998; Grekousis and Hatzichristos, 2012) to produce clusters (detailed in Chapter 3.2), as well as multinomial logistic regression models, also known as m-logit models (Jackson et al., 2006). A logit model has the advantages of using continuous, binary or categorical data to generate clusters, plus, these can also be considered as a soft classifier as they output the probability of each spatial area belonging to each cluster category. Such models have been used in health geodemographics and epidemiology, where detailed geographical information is often unavailable so small-area aggregate data can be utilized to increase descriptive power.

4.4. Cluster analysis and interpretation

The final step in building a geodemographic classification includes the review and testing of the

cluster results, alongside descriptions of the typology. For example, checking the size of clusters is one of the basic steps in the optimization procedure. Clusters with relatively low representation of cases should generally be avoided, by either adjusting the number of clusters or by re-evaluating the data input (Alexiou and Singleton, 2015a). For instance, outliers within the data points may sometimes form distinct clusters, in which case, these spatial units might be weighted accordingly to reduce impact. Furthermore, if, measured in terms of variance, two or more of the output clusters look very similar merging should be considered, or inversely split if the clusters are too large. Harris et al. (2005) provides "a rule of thumb" for merging similar clusters, if the loss of variance within the dataset is less than 0.22%. Other ways to test an output classification is to correlate it with existing classification systems, or via sampling, such as cross-tabulation with geocoded survey data.

Regarding the analysis of each cluster per se, a useful way of obtaining information specifically on how variables load onto each cluster is through a radar plot. Figure 4.2 shows a summary of the average distribution of values in standard deviations within a cluster from the Liverpool Classification (Alexiou and Singleton, 2015a).

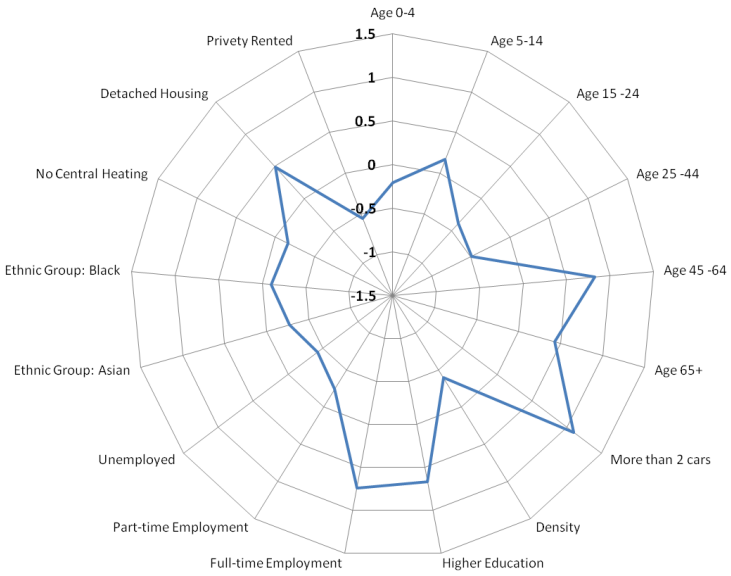

Figure 4.2 An example of within-cluster variable analysis of a cluster using a radar plot, as adopted from the Liverpool classification (Alexiou and Singleton, 2015a).

The radar plot is useful since it is easier to interpret a multidimensional set of attributes where axes represent the attributes of the classification. Since values are transformed in *z*-scores, the

circle with a value of 0 represents the mean attributes for Liverpool, so values above that are over-represented in the cluster and values below are under-represented. In this framework, this particular cluster consists mainly of neighbourhoods of middle-aged families, the majority which are full-time workers with higher education degrees. Families are more prevalently living in low density, detached houses while the high ratio of car ownership indicates these areas may be more affluent. This cluster was thus named "*White Collar Families*".

Once the cluster interpretation appears successful, a map showing the cluster membership of Output Areas and their attributed names can be drawn. An example of such a map of Liverpool is seen in Figure 4.3.

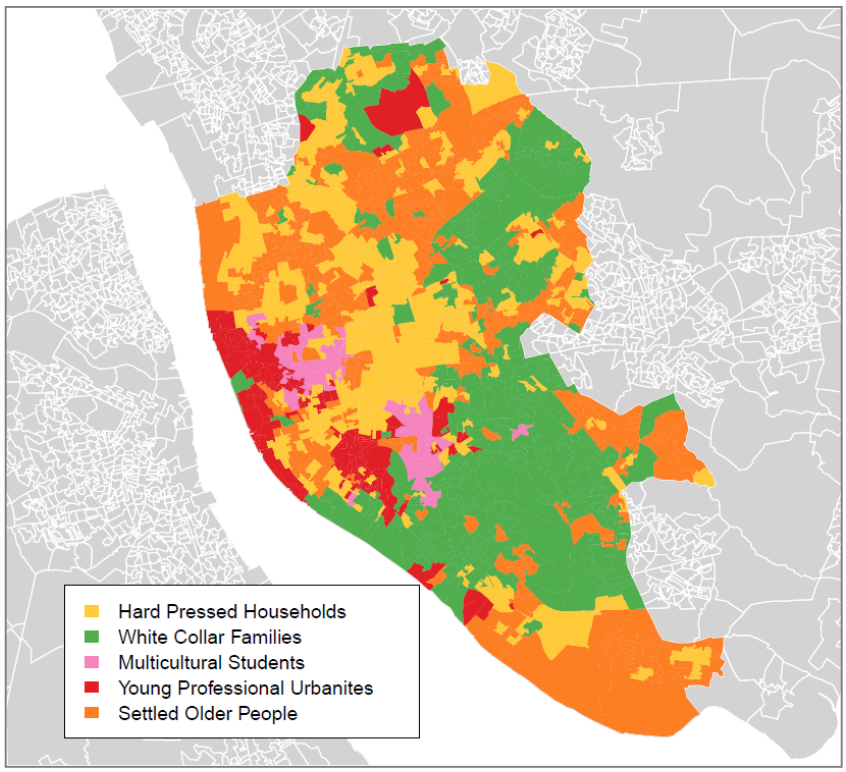

Figure 4.3 A map showing the results of the Liverpool classification which groups the 1,584 OAs of the Liverpool Local Authority District into 5 clusters, along with their interpretations. Note that there is considerable degree of spatial autocorrelation among OA cluster membership (Alexiou and Singleton, 2015a).

A final step in interpretation is naming and describing the resulting clusters with written "*pen portraits*" that best fit the profile of areas represented by the clusters. The process of creating such descriptions can be quite difficult, especially in lower hierarchies, where the cluster dissimilarities are subtle (Vickers and Rees, 2007). An extract of the profile for the cluster *"Affluent Achievers"* from the commercial classification Acorn by CACI:

"These are some of the most financially successful people in the UK. They live in wealthy, high status rural., semi-rural and suburban areas of the country. Middle aged or older people, the 'baby-boomer' generation, predominate with many empty nesters and wealthy retired. Some neighbourhoods contain large numbers of well-off families with school age children, particularly the more suburban locations. These people live in large houses, which are usually detached with four or more bedrooms…." (CACI, 2013).

Classification systems also commonly augment such descriptions with other visual materials such as photographs, maps and bar graphs or radar charts. Depending on the intended end users, labelling and description must be selected appropriately in order to expand the user's understanding of the group, while taking into account that the end user might not be accustomed to geodemographic classifications.

4.5. Concluding Remarks

This Chapter outlines the methodological steps, advantages and implications related to geodemographic analysis. Geodemographic classifications are relatively easy to use and can provide a system that is open and versatile enough in order to handle the abundance of big data that are currently available. Their popularity stems from their practicality and an upholding validity. It is also true that geodemographics can be more of an art than a science (Harris et al., 2005, p. 181); as Webber and Craig (1978, p. 3) explained for the outset,

"No claim for infallibility is made. So far as the methods are concerned the crucial test will not be a methodological debate but whether the classifications are actually found to be useful."

Many argue that this statement is moot, claiming that it is unsatisfactory to justify geodemographic classifications on pragmatic grounds, simply because of the lack of theory behind them and a systematic validation of their system-wide accuracy (Voas and Williamson, 2001). As Openshaw, Cullingford and Gillard (1980, p. 423) state:

"[..] plausibility and usefulness can only be determined in relation to purpose and that while necessary, these are not, by themselves, sufficient conditions for a "good" classification".

Some authors argue that the critical stages are highly subjective and operational decisions made there may substantially determine the utility of the results. It is true that the classification

process described here can be very specific to the underlying data and methodology adopted by the creator. Webber (1980), in a response to these claims, highlights that the alleged methodological implications are unfounded because the classification methodology is based on a long history of expertise and analysis of Census variables, specifically designed to meet the needs of policy-orientated researchers. This expertise is embedded in the classification methodologies that have adopted ever since, same as the proprietary classification creators that have built around their own data and expertise to refine theirs.

It is generally acknowledged however that an inherent disadvantage of all geodemographic classifications is that lack of a single global optimization function during the classification procedure, making them highly susceptible to the operational decisions during the creation process (Openshaw and Gillard, 1978). This uncertainty is further enhanced by the lack of classification transparency; transparency relates to the data, methods and underlying techniques used to construct a classification, so it can be easily replicated, updated or otherwise customized to fit the needs (scope) of the creator (Fisher and Tate, 2015).

Despite these issues, there has been a renaissance of interest in geodemographics from the public sector, mainly driven by government pressure to demonstrate value for money and the advent of new application areas (Longley, 2005). Part of that renaissance is the recent development of application-specific classifications, i.e. classifications refined for a specific purpose through the augmentation of sector specific data to predict these phenomena on a local scale (Singleton and Longley, 2009). These types of bespoken classifications, as opposed to general purpose classifications like the OAC, have been very popular in recent years since they are primarily created to explore specific spatial phenomena with increased accuracy.

Assuming geodemographic applications can produce reliable results by taking advantage on the abundance of data that is currently available and the experience gained through the years in the analysis of Census datasets, geodemographic research may face substantial challenges in the near future. Many geodemographics have historically relied on the analysis of the decennial census of the population, but institutional shifts in both the U.S. and UK are already changing the nature and availability of such data, given the growing costs associated with their collection (Singleton and Spielman, 2013). As such, the granularity currently offered by census data might not be readily available in the future; and as such, more research is needed into how the linkage of non-census attributes (both commercial and non-commercial) can be both validated and made more accessible.

Secondly, geographic classifications as currently construed, do not account for spatial relations

between proximal zones. In their standard form, clustering algorithms are optimised only on the basis of this attribute space, and as such, do not account for spatial associations between the small areas. For instance, some argue that the relationship between areal typology and behaviour might not be spatially constant (Twigg et al., 2000). Openshaw (1998) during a lecture at the Institute of Direct Marketing in Leeds, illustrated this point by drawing attention to the fact that geodemographics, as is, incorrectly assumes that residential areas of i.e., type 27, behave similarly regardless of where they are located. He concludes that areas of the same typology may respond differently depending on their location. He also added that other variables such as proximity to major roads, urban centrality, altitude and general land uses should also be taken into account. Openshaw raised an important issue regarding the *spatial context* of geodemographic classifications that, to this day, have received little attention, at least within the academic works. It regards the contextual differentiations of socio-spatial patterns at a regional scale as a result of location characteristics, regional economies and public policies.

In the following Chapters this Thesis tries to address these issues from an analytical and methodological point of view. While the issue of geographic context is the main focus of this research, Chapter 5 addresses the first issue of data availability by exploring the creation of secondary, non-census data, extracted from physical and built environment attributes. This type of data can be used to shift the historical focus on Census data from public geodemographic applications, and help alleviate any effects on the availability of Census data in the future. The classification application also a) demonstrates how information on built environment features can be extracted at a neighbourhood level and complied into a database on a national scale, b) provides a clustering method that can successfully create bespoke classifications from sparsely-populated data and c) test the association of such measures to current socio-economic patterns. Moreover, the detailed description of the methodological steps required to create the classification serves as a practical example of a geodemographic analysis, although on a more advanced level due to analytical and geocomputational challenges involved.

Chapters 6 and 7 try to address the second issue in more detail. Chapter 6 explores the regional contextual differences of socio-economic patterns, by measuring the degree of influence of the near-geography to the overall similarity between neighbourhoods. This part of the research addresses the issue of geographic context by analysing and evaluating various local, regional and national extents that can be used as attribute contextual weights. It provides the theoretical and practical rationale of how attribute values can be adjusted, and what impact do these adjustments have on cluster formation. Results are demonstrated across the UK, and a model that incorporates such measures is presented in Chapter 7. The evaluation of outcomes is carried out

by contrasting emergent clusters to a conventional geodemographic system that serves here as a base model to the analysis, as well as by using measures of internal cohesion of clusters.

While both issues seem to affect the success of geodemographic classifications simultaneously, building a unified model would require extensive exploration. These dimensions have not been incorporated or evaluated in conventional geodemographic system, nor have the interactions between them, so the amount of operational decisions that would be required would simply be too great for any robust generalization. The following three Chapters try to address these issues by providing a required framework for future research on the subject.

Chapter 5. Building a Geodemographic Classification using Physical and Built Environment Attributes

5.1. Introduction

While most geodemographic classification systems include a plethora of socio-economic attributes, there is arguably little to no input regarding attributes of the built environment or physical space. Furthermore, their relationships to socio-economic profiles have not been evaluated in any systematic way. The aim of this research is to capture through the multidimensionality of the data both microscopic and macroscopic identifiers of urban morphology.

The value of this research is twofold; firstly, the creation of small-area summary measures of built environment and physical space can be used within geodemographic classifications as open and versatile non-census attributes. Secondly, classifications from these attributes can be used as input or evaluation to more complex socio-economic models, increasing robustness. For instance, areas with a strong prevalence of specific built environment and land use features could be impacted by economic deprivation differently compared to others.

The rationale for this research draws from strong evidence that residential preference is a significant part related to the form of the built environment, suggesting that there is an important dimension to residential decisions beyond homophily. For instance, even at the most expensive of neighbourhoods, one would expect house prices to drop significantly very close to railway tracks, making them more affordable to certain demographic groups. Other households prefer to reside within reach of an urban park. However, these localized phenomena are aggregated in the general context of the area, and thus patterns get "*smoothed away*", raising some issues about the accuracy of geo-classifications. While gathering this type of behavioural data would be next to impossible, their outcomes can be observed through local neighbourhood morphology.

This Chapter is drawn from the work regarding a bespoken classification of Multidimensional Open Data of Urban Morphology, presented in Alexiou, Singleton and Longley (2016). The following sections outline how summary measures of built environment characteristics can be collected at the small area level and what are the methodological implications during of such an analysis. The generation of neighbourhood characteristics and other attributes is carried out using a geocomputational approach, taking advantage of the increasing availability of spatial data from

open data sources. A Self-Organizing Map is used as a clustering method, mainly due to the unique nature of the data. The research concludes with a comparison of the MODUM Classification to the 2011 OAC, which tests whether and to what extent built environment patterns systematically follow conventional socio-economic profiles.

A geodemographic approach is adopted in order to identify neighbourhood patterns. Results are presented similarly to a geodemographic application, illustrating cluster attributes, pen portraits as well as a few maps to aid the spatial interpretation of the classification. Besides identifying physical and built environment patterns on national level, this research also tests whether specific and multidimensional urban morphologies systematically correspond with socio-economic characteristics at the neighbourhood level by comparing the resulting classification to that of OAC 2011. The comparison methodology will also serve as a model for classification comparisons described in following Chapters.

The value of this bespoken classification is unique in the sense that the geodemographic model presented can provide a summarized and simplified structure of the physical properties of geographic space at the small-area level based on within and proximal features; it can be useful in representing a simplified structure of the physical properties of geographic space. The resulting typology can also be used to explore correlations with other spatial phenomena, potentially in a variety of applications, from real estate and house prices to health and wellbeing.

5.2. Data Sources

Although geodemographic frameworks can capture a wide set of input attributes, current classification systems typically include little to no input of explicitly spatial attributes regarding the built and physical attributes of neighbourhoods. There is, however, an abundance of variables that might be collected on the built forms and relative locations that underpin neighbourhood differentiation. For instance, proximity to certain amenities is important to residential decisions such as transport nodes, parks, retail and healthcare-facilities. There has, for example, been extensive research into the topic of analysing relationships between accessibility and urban development patterns, (e.g. land use - transportation interaction models); and connectivity has been advanced as a key feature in shaping urban residential dynamics and socio-spatial segregation (Dear, 2002).

Research on residential decisions has also attracted a lot of attention over the years, particularly through hedonic modelling. While most of the relevant research focuses on the

importance of work location (Van Ommeren et al., 1999; Renkow and Hoover, 2000), there is strong evidence that certain demographic groups favour some relative locations over others, and that the nature and configuration of the local built environment and land-use characteristics is also relevant (Hui et al., 2007). For instance, individuals with children often favour greenspace and recreational opportunities nearby, while those without children prefer smaller residences that offer closer proximity to central services (Colwell et al., 2002). Other characteristics may impact the area as unfavourable, negative externalities, such as high-speed roads or railway tracks within the vicinity of the neighbourhood (Parkes et al., 2002). It is unclear how exactly how such characteristics impact upon residential decisions as there are many synergies involved across life cycles (Kim et al., 2005). For instance, moderate proximity (200m to 300m) to a green space may mitigate negative effects of noise pollution (Gidlof-Gunnarsson and Ohrstrom, 2007).

Some census variables reflect limited built environment characteristics, for instance housing type and population densities. For classification systems that have been developed entirely from census variables, such as the publicly open ONS Output Area Classification for 2011, attributes such as density can be misleading; the arbitrary nature of the geographic extents of the administrative areas for which population measurements are offered renders comparisons between the physical features ineffective. Other proprietary geodemographic classifications, such as Mosaic by Experian (Nottingham, UK) and Acorn by CACI (London, UK) include some measures of relative location (CACI, 2013; Experian, 2014). However, to what precisely these attributes pertain, how they are used in the clustering process and the weight they are assigned in the final classification remains obscure, because of the commercial sensitivities that are inherent in 'black box' commercial solutions (Singleton and Longley, 2009).

This research aims to capture a variety of physical attributes collected for a small-area geography, and in order to enhance reproducibility, replication and extension these inputs are assembled from Open Data sources (Singleton et al., 2016). The classification is produced at the 2011 UK Census Output Area level for the 181,408 Output Areas that make up England and Wales. One of the main providers of geographical data for England and Wales is the national mapping agency Ordnance Survey (OS), and there are many datasets available within their repository, with varying degrees of granularity, depending on whether they are publicly accessible or available for purchase. As this research focuses on Open Data sources, considerations were made into a variety of open vector data sources which can be used directly or supplementary, such as OpenStreetMap (www.openstreetmap.org). Nevertheless, in order to maintain a consistent level of accuracy, the OS Open Map - Local product is used, which is the most recent and detailed open OS vector data product currently available (Ordnance Survey, 2015). The OS vector data product provides a

variety of information including outlines of buildings, street network with hierarchy, railways, woodland areas, surface water and important functional sites.

While the OS Open Map – Local provides the main source of these data, there were a few other sources within England and Wales deemed of utility. These included data about listed buildings and historic parks and gardens supplied by the *Historic England Archive* (*https://services.historicengland.org.uk/NMRDataDownload/*) which is regularly updated (November 2015 update used here) and also under Open Data License. For Wales, the corresponding provider is the *Cadw* heritage organisation, (available through the UK data Service, *https://data.gov.uk/dataset/listed-buildings-in-wales-gis-point-dataset*), although the data are slightly outdated (September 2011). Commercial buildings for local retail centres were identified using data from the Local Data Company, an Open version of which is available through the ESRC Consumer Data Retail Centre (CDRC). Finally, the dataset includes aggregated data on housing type from the 2011 Census supplied by the Office for National Statistics. Unfortunately, there are currently no Open Data available on building age or height. The UK Environmental Agency now provides raw LIDAR data that may offer a few possibilities to that end, but they have only been available very recently (<*https://data.gov.uk/publisher/ environment-agency*>), and they still do not offer complete coverage.

Table 5.1 below summarizes the range of inputs used to derive measures featured in this analysis.

Table 5.1. Description of the spatial dataset compiled.

Dataset Name	*Dataset Description*
D1: OA Boundaries	181,408 Output Area boundaries, as defined by the 2011 Census. All other data were spatially joined with respective OAs that they fall into (data features were split when falling into more than one OA).
D1: Buildings	12,878,666 Building objects represented as polygons. Note that these areas do not represent individual households.
D2: Road Network	Road network is represented as line segments, approximate to the road centre. The categories include 'Motorway', 'Primary Road', 'A Road', 'B Road', 'Minor Road', 'Pedestrianised Street', 'Local Street' and 'Private Road Publicly Accessible', as well as their 'Collapsed Dual Carriageway' counterparts.

D3: Woodland	Areas of trees represented as polygons, described as coniferous and non-coniferous.
D4: Functional Sites / Important Buildings	120,677 Building polygons that can be can be found within functional sites. They are categorised into themes such as Air Transport, Education, Medical Care, Road Transport and Water Transport, which are further classified into numerous more discrete classes.
D5: Railway Stations and Tracks	Railway tracks and tunnels represented as lines (in this instance we used tracks only in the analysis) and Railway Station defined as points.
D6: Surface water	Polygons of surface water. Small rivers and streams are represented as lines and are not included in the dataset. The dataset was also supplemented with 'sea water', derived from the country's coastline.
D7: Registered Historic Buildings	406,496 listed historic buildings defined as points, which were geolocated.
D8: Registered Parks and Gardens	2,007 Polygon features with extents of the parks / gardens, classified as I, II*, or II, from most to least important. For Wales, the 372 sites were identified from points from a *"Named Places*" dataset and given an approximate 200m radius.
D9: Retail Centres	1,312 Retail Centres across the England and Wales. There is no recent update for this dataset which dates back to 2004. The centres are only depicted as points and have no typology attached. We assumed an average radius of 200m to convert them to areas.
D10:Housing Type	Percentage of households that are classified by the Census as Detached, Semi-detached, Terraced or Flat.
D11: Population	Population of total persons per OA.

The selection of the OA zonal level offers advantage over other administrative units in England and Wales since many other socio-economic classifications are offered at the OA level, such as the 2011 OAC, thus making comparisons possible. Additionally, such geography also allows the incorporation of Census data which are distributed for these units. However, for the range of the derived measures that are described in the remainder of this section, there are problems with this approach. OA borders were designed to minimise within zone homogeneity in population characteristics (population normalization), without regard to the geographical features of the area (Martin et al., 2001; see Figure 5.1). As such, for proximity based inputs there were challenges about how such measures might be calculated, and to which area they should be

attributed.

A similar attempt to create such a dataset was made by the Department for Communities and Local Government in 2005, within the framework of the ONS Neighbourhood Statistics, described as *Land Use Statistics*. The dataset was described as a generalised land use database aggregated into OAs. The dataset contained estimates of built environment attributes, such as roads, paths, domestic and non-domestic buildings, domestic gardens, water, rail etc. Despite the fact that the proprietary *OS Enhanced Basemap* was used to create this resource, ONS classified it as experimental, as there were accuracy issues because only the centroids of features were taken into account to produce summary measures of characteristics.

To facilitate these methodological shortcomings three different types of attribute measures are introduced for each OA that related to either two types of proximity measures including *adjacency effects* or *intermediate effects*; and additionally *direct measures.* The lattermost of these are simply attributes captured at the OA level, while the first two assume buildings as the initial unit of analysis which are then later assigned to OAs. Building polygon features serve as observations in this input dataset, and represent homogenous built-up areas which can include one or more households. A graphical representation of the model is described in Figure 5.1.

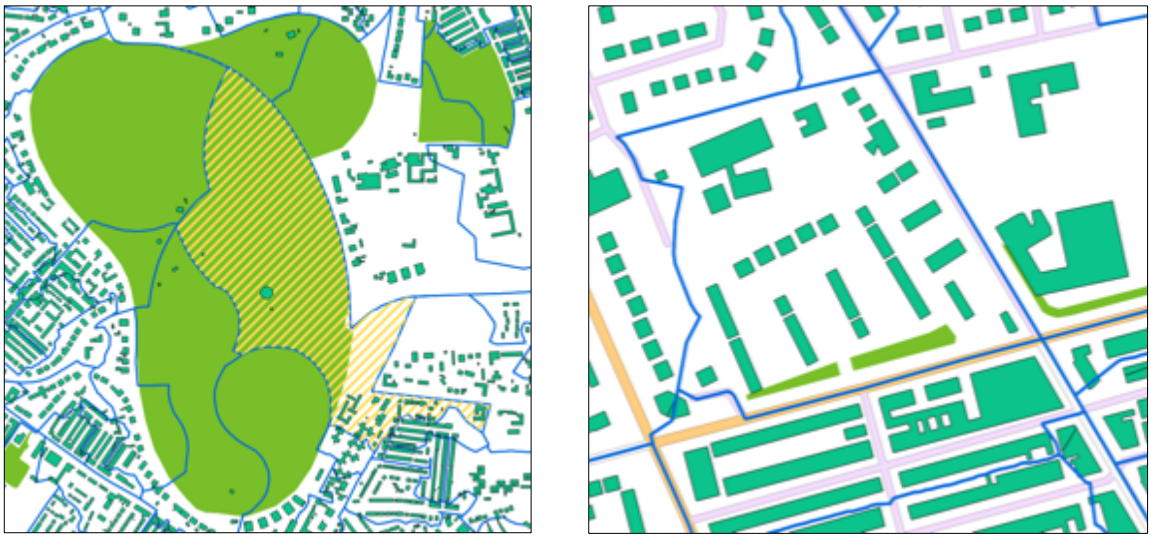

Figure 5.1 Maps looking at the un-generalised Output Area borders (blue lines). Left: Sefton Park, Liverpool. Notice how the area of the park is divided arbitrarily between proximal OAs (yellow hashed line pattern). Right: Output Area borders usually coincide with the street network, making simple street network–to-area assignments impracticable.

For both types of proximity measure, a series of spatial queries were used that identified buildings that fulfil certain criteria, for instance, *"Which buildings are within a set distance of a major street?"*. The buildings that met each criterion were then assigned to OA aggregations with weights determined by their building surface. Thus, within each OA, a ratio of the area of buildings meeting the criteria relative to the total built area was calculated for each of area attributes considered in the analysis. The necessity to differentiate between adjacency and intermediate proximity effects follows the logic that not all built environment characteristics have the same effect, and these effects may vary in scale. For example, when considering the location of a residential property, being adjacent to a very major road might be perceived as having a negative impact, given noise / pollution associated with increased traffic volumes, whereas being near, but not adjacent to a busy road might be perceived as advantageous, given the enhanced connectivity this might facilitate.

This research defined *adjacency effects* to features measured within 100m linear distance, as commonly used in the literature on negative externality effects of built environment features, such as noise or pollution from roads (Rijnders et al., 2001). For *intermediate effects* a distance of 600m was used, on the basis of various western international definitions of "within walking distance". The distance figure generally varies depending on the context of analysis, but distances between 300m and 900m are considered appropriate for urban features (Hui et al., 2007; Barbosa et al., 2007; Villeneuve et al., 2012; Vale, 2015).

Outside of these distances, it is assumed there are no adjacency or intermediate effects. The delineation of *adjacency effects* or *intermediate effects* brings additional practical considerations which relate to the overall density of the built environment features being considered. In common with practice when creating inputs to multidimensional classifications, preference should be for those attributes which in addition to theoretical rationale, also provide useful differentiation between areas (Spielman and Singleton, 2015). For example, in this application, when 600m buffers were used for major roads, this resulted in more than 50% of buildings meeting this criterion, providing a weak differentiation. These tasks were computationally expensive, as the complete dataset contains more than 12.8 million observations (building polygons). Therefore, the database was pre-processed into regional datasets which were then computed separately within the *R* coding language.

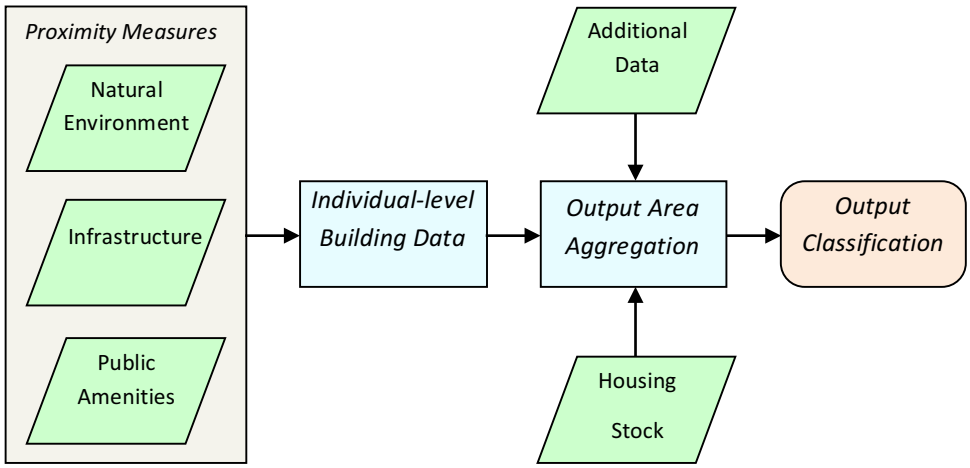

Figure 5.2 The spatial data model used to process data and produce Output Area inputs to the classification.

Finally, there were two further types of *direct* measures: those which were derived from geographic features, and those which were simple inputs from secondary data. The derived direct measures included listed buildings and cul-de-sacs (dangling segments in the road network). The later of these was defined geocomputationally as the end of a line segment that did not intersect with any other such segment. A sensitivity of 10m was applied to this criterion in order to avoid topological errors and intermittent street segments. Results show that such measures can capture specific urban morphologies even at the small-area level, as illustrated in Figure 5.3.

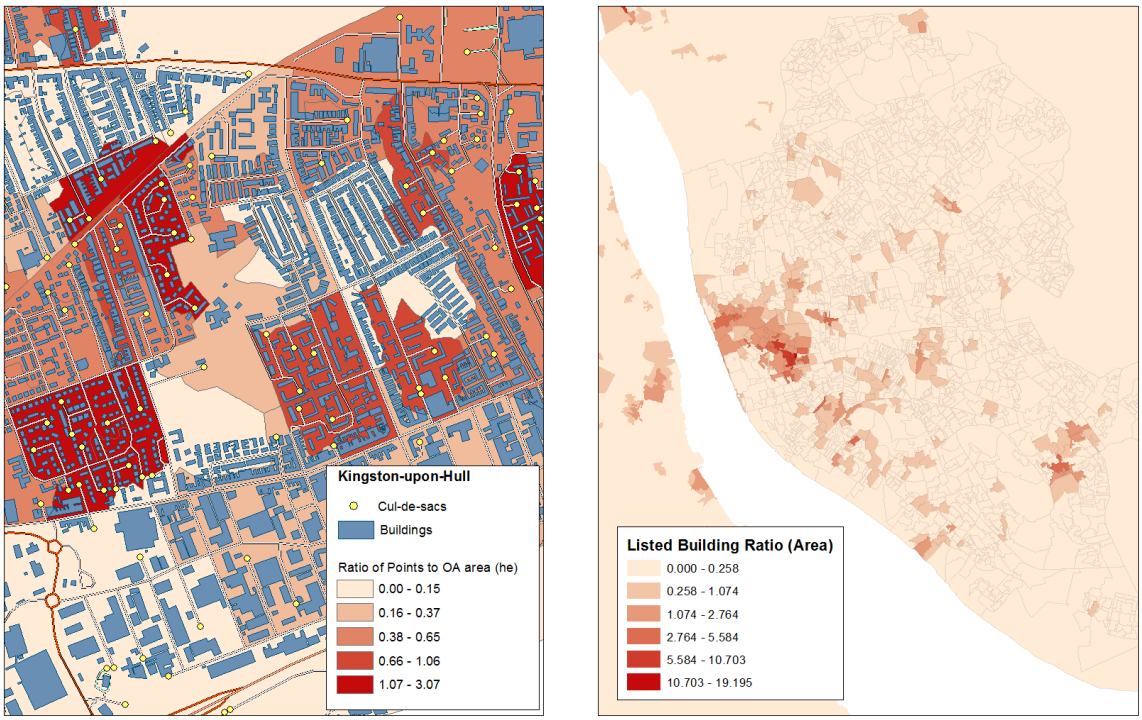

Figure 5.3 Left: Cul-de-sac ratio per OA area (ha) at Kingston-upon-Hull, Yorkshire. Right: The sum of listed (registered) building surface (ha) per OA in the area of Liverpool.

For the other non-derived direct measures, the variables were simply aggregated directly at the OA level, such as housing type. Population density was calculated using a ratio of persons per total building area, which potentially would give more accurate results regarding housing conditions. The final OA attributes along with their descriptions are provided in Table 5.2.

Table 5.2. Built environment attributes used in the classification.

Variables	Variable Description, Aggregated per OA Code
Adjacent effects	
1. Major Roads	Percentage of the area of buildings that the centroid is within 100m of a major road to the total building area. We defined major as those of type "Motorway", "A Road" and "Primary Road".
2. Arterial Roads	Percentage of the area of buildings that their centroid is within 100m of an arterial road to the total building area. We defined Arterial roads as those with type "B Road".
3. Pedestrian Roads	Percentage of the area of buildings that their centroid is within 100m of a pedestrian road or footway to the total building area.
4. Railway Tracks	Percentage of the area of building units that their centroid is within 100m of railway tracks, excluding tunnels to the total building area.

5. Woodland Areas	Percentage of the area of building units that their centroid is within 100m of woodland features to the total building area.
6. Surface Water	Percentage of the area of building units that their centroid is within 100m of surface water (inland) and seafront (calculated by the distance from the coastal line), but excluding small rivers and streams, to the total building area.
Intermediate effects	
7. Railway Stations	Percentage of the area of building units that their centroid is within 600m from the centroid of a railway station to the total building area.
8. Parks and Gardens	Percentage of the area of building units that their centroid is within 600m from the registered site extents to the total building area.
9. Retail Centres	Percentage of the area of building units that their centroid is within 600m from the retail centre centroid plus 200m to the total building area.
10. Schools	Percentage of the area of building units that their centroid is within 600m from the sites that are identified as primary through secondary education to the total building area.
11. Higher Education	Percentage of the area of building units that their centroid is within 600m from the sites that are identified as further and higher education to the total building area.
Direct measures	
12. Detached Ratio	Percentage of unshared households that are classified by the 2011 Census as detached housing to the total building area.
13. Semi-Detached Ratio	Percentage of unshared households that are classified by the 2011 Census as semi-detached housing to the total building area.
14. Terraced Ratio	Percentage of unshared households that are classified by the 2011 Census as terraced housing to the total building area.
15. Flat Ratio	Percentage of unshared households that are classified by the 2011 Census as Flats to the total building area.
16. Density	Ratio of persons to total building area (people/ha).
17. Cul-de-sac	Ratio of cul-de-sacs or dead-end road points to the total OA area (points/he).
18. Registered Buildings	Ratio of listed buildings to the total OA area (points/ha)

5.3. A multidimensional classification of the built environment

Methodologically, the cluster analysis follows a conventional approach, as detailed in Harris et al. (2005); however, only physical and built environment data are used to create the typology. A common clustering technique used in geodemographic analyses is the iterative allocation – reallocation algorithm, known as *K*-means. Although this algorithm has been used in a variety of

geodemographic applications, this dataset is characterised by very sparsely populated attribute values. Essentially, the majority of values are zero, indicating the absence of the particular built environment or physical characteristic from that area. In this case, *K*-means did not to respond well to sparse and non-Gaussian distributions that characterise this dataset (see Chapter 3, section 4).

Due to these shortcomings, an alternative technique is used as a clustering algorithm, the Self-Organizing Map (SOM). A SOM is an unsupervised classifier that uses artificial neural networks to classify multidimensional observations in two-dimensional space based on their similarities (Kohonen, 2001). A SOM typically organize observations by projecting them onto a plane, and through consecutive iterations finds the best configuration of observations so that every observation is most similar to the others closest to them (see also Chapter 3, section 4). Typically, the SOM mapping process employs a lattice of squares or hexagons as the output layer, and the results are therefore easily mapped as they retain their topology. SOMs have many applications in a broad range of fields, from medicine and biology to image analysis and computer science. SOMs have also been tested as an alternative classifier of Census data (Spielman and Thill, 2008; Arribas-Bel and Schmidt, 2013) where they seem to perform well for socioeconomic data at the US Census tract scale. Arribas-Bel, Nijkamp and Scholten (2011) have also demonstrated the algorithm capabilities to measure urban sprawl in Europe using a similar attribute set, specifically six variables: connectivity, decentralization, density, scattering, availability of open space and land-use mix.

The technique also has the advantage of not assuming any hypotheses regarding the nature or distribution of the data, and responds well to geographic sensitivity. A further advantage of using a SOM is the capacity to visualise the structure of data values aiding initial data exploration. This feature can be very useful when analyzing datasets such as built environment measures, where there are little to non a-priori hypotheses on their underlying distributions.

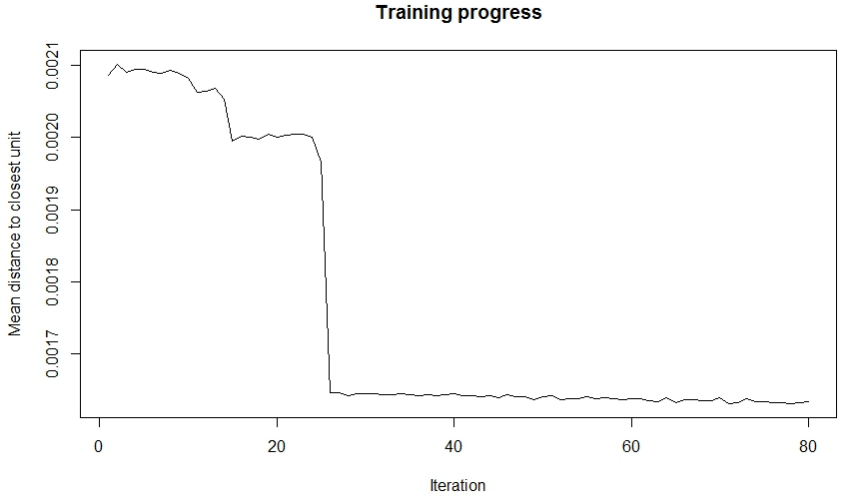

Figure 5.4 Plot of the SOM training progress. The algorithm seems to have converged at ~25 iterations, with no significant changes thereafter.

The input dataset consists of 18 variables, summarized in Table 5.2, which were transformed into *z*-scores in order to standardise measurement scales. The majority of the analysis and output production was performed in the R programming language using the "*Kohonen*" library (Wehrens and Buydens, 2007). The specifics of the SOM clustering approach were carried out following the methodology described by Spielman and Folch (2015).

A relatively unexplored built environment classification with too many clusters would be difficult to interpret, so a selection of a 4-by-2 hexagonal grid was made, which produces 8 distinct clusters. The cluster analysis implements a hexagonal geodesic grid to project results. The hexagonal representation offers increased spatial interactions between cells and the geodesic plane forces the cells' relations to "loop" around the edges, and so this configuration benefits from every cell having six immediate neighbours.

The other main parameters of the SOM algorithm are the learning rate *alpha*, which is defined to progress linearly from 0.05 to 0.01 over fifty reconfigurations (*updates*), and the initial size of the neighbourhood, in this instance a distance chosen in such a way that two-thirds of all distances of the map units fall within the topological extents. The neighbourhood decreases linearly during training until the algorithm reaches equilibrium. The algorithm has achieved equilibrium at ~25 iterations (Fig. 5.4), meaning that no more changes to the observations' configuration were required, with the mean distance to the closest unit in the map at 11.34. Once areas were assigned to clusters, mean attribute values are assigned to radar plots in order to map

cluster characteristics and label them accordingly, as seen in Figure 5.5.

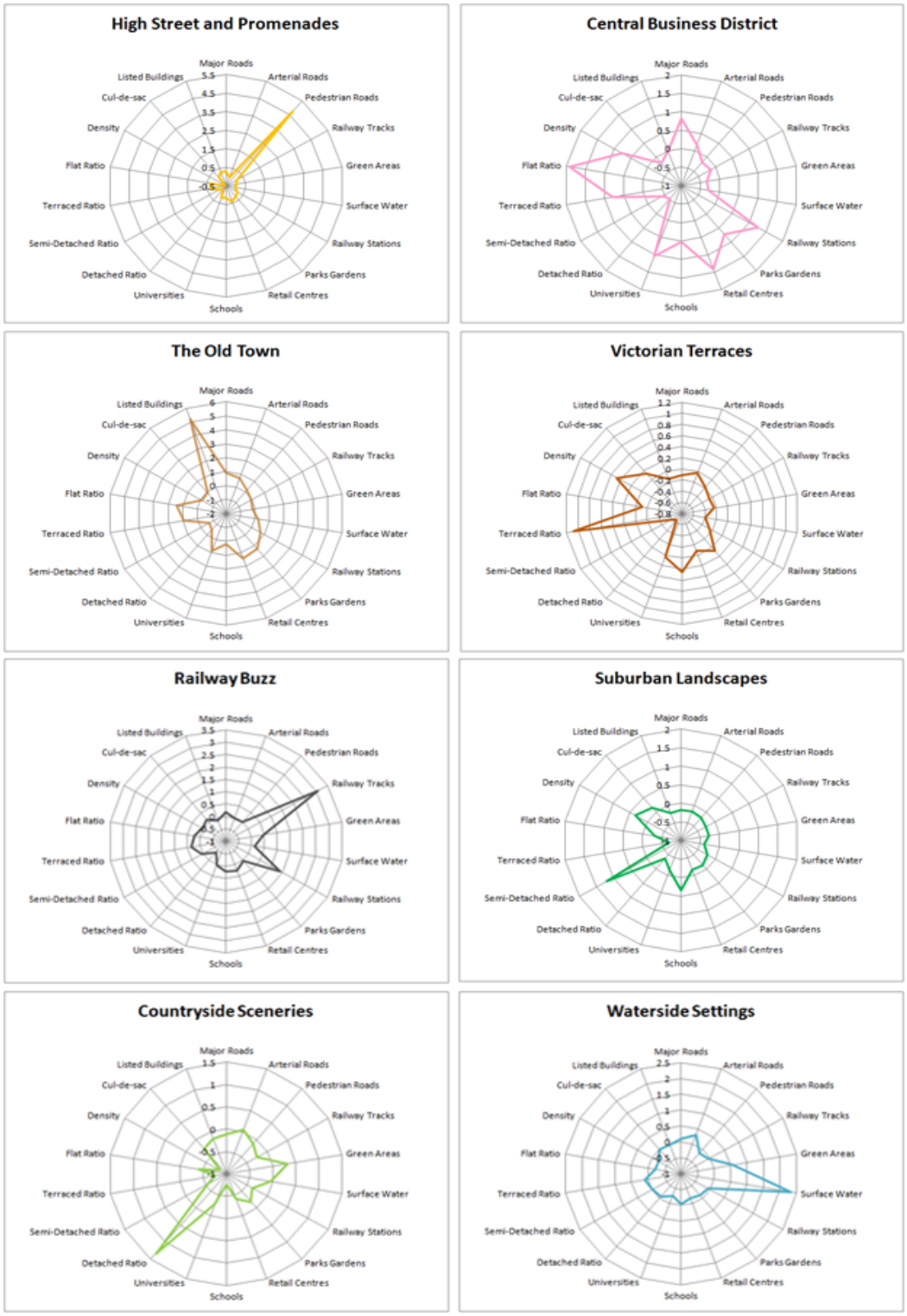

Figure 5.5 Final cluster results produced by the SOM, with mean attribute centres per cluster.

Radial plots are used extensively in Geodemographics as they are very intuitive in identifying the nature of formed clusters. A radial plot essentially depicts the cluster centre; it is a vector representing each attribute mean (in this case for 18 variables) within the cluster. Each attribute mean can be traced along every radial axis at their intersection, forming a unique pattern for every cluster. Since values were standardized to *z*-scores, values of zero suggest that the cluster attribute mean is equal to the national mean, while values above or below zero suggest that cluster attribute means are above or below national average respectively. It also suggests that the values shown are measured in standard deviations.

To illustrate with an example, assume that *Cluster C: The Old Town*, is under consideration. The radial plot shows that Cluster C has an above average prevalence of major roads (1.0), pedestrian streets (0.4), parks and gardens (1.4) and retail sites (1.5). It has below average values of detached and semi-detached housing ratios (-1.6 and -1.7), but a high concentration of flats and terraced housing (1.4 and 1). The defining aspect of this cluster however is the listed buildings attribute, which has an average value of 5.1 within the cluster. From the mean values of attributes of Cluster C, it is suggested that these neighbourhoods are in the periphery of the city centre, proximal to some major roads and retail activities. The amount of historical buildings and the presence of flats and semi-detached housing suggest neighbourhoods that have been historically affluent, potentially with a strong presence of churches or administrative buildings, which have been repurposed to housing (e.g. flats) or recreational facilities (e.g. pubs and restaurants).

An example of such a cluster typology is the *Georgian Quarter* in Liverpool, a historic affluent housing neighbourhood built in the 1800's. Figure 5.6 and 5.7 show the MODUM classification superimposed over an Openstreetmap basemap layer (source: Openstreetmap.org), along with the aerial photo of the neighbourhood (source: <*http://www.visitliverpool.com/explore-the-city/neighbourhoods/georgian-quarter*>). In general, visual evaluation of clusters aids greatly in interpretation.

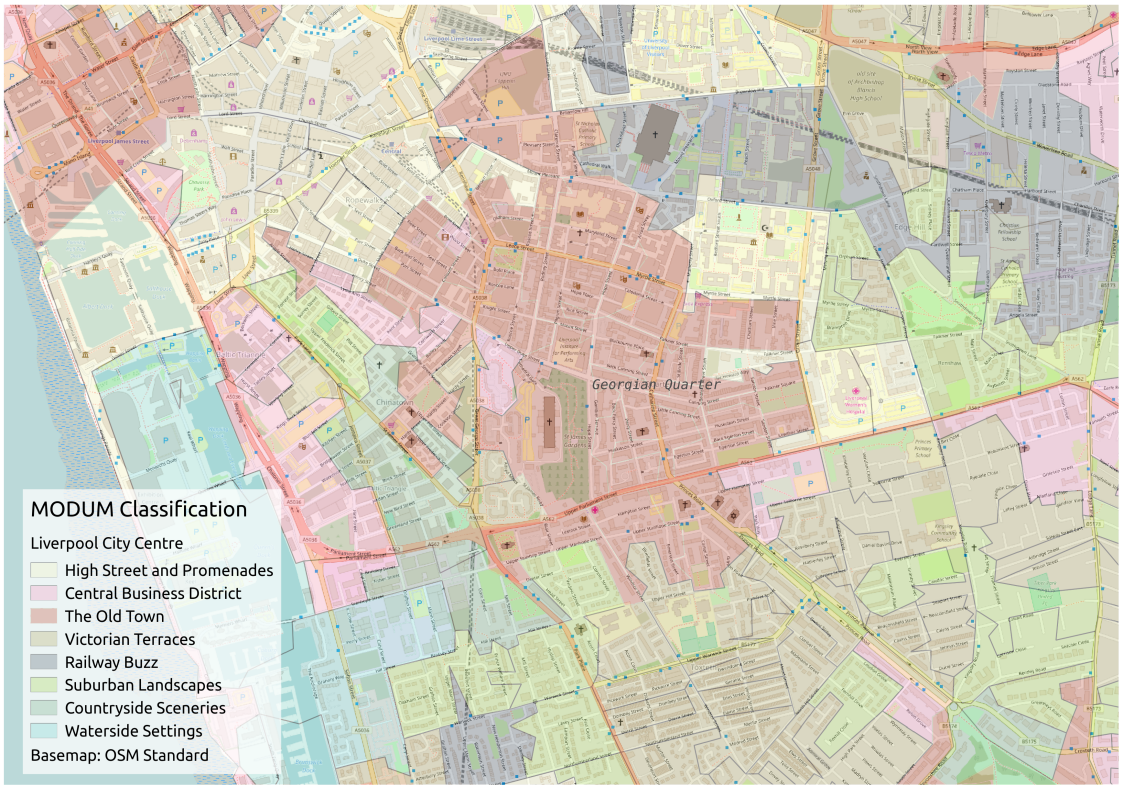

Figure 5.6 Output Classification with cluster types superimposed over an Openstreetmap basemap (source: Openstreetmap Contributors; CC-BY-SA), showing the historical Georgian Quarter in Liverpool City Centre.

Figure 5.7 The Liverpool Georgian Quarter, aerial photo (source: http://www.visitliverpool.com/explore-the-city/neighbourhoods/georgian-quarter).

In a similar manner, the rest of the clusters were examined in order to identify defining characteristics. This enabled cluster types to be labelled and the following short descriptions to be created:

A. *High Street and Promenades*

These clearly depicted areas represent the main retail centres of urban regions located along the main commercial streets. This cluster also includes areas with significant pedestrianised street network, especially along seafronts, where a lot of recreational and leisure venues can be found.

B. *Central Business District*

The area often called city centre. Typically, high-rise buildings with a lot of commercial and office spaces, hence the relatively low net population density. These areas have proximity to the majority of public amenities, and have plenty of access vial major roads and railways. For moderate-size cities the title holds true, but in areas such as London they tend to be too expansive to be labelled as central (Fig. 5.9).

C. *The Old Town*

The traditional town centre, usually close by the main high street. It is strongly defined by the amount of registered buildings. Typically, a lot of recreational facilities can be found there, like pubs and restaurants, along with many administrative buildings and some historical major roads. Although it does have a considerable amount of flats, densities remain low, potentially due to refurbishments and change of usage.

D. *Victorian Terraces*

These are typical neighbourhoods with terraced housing, average densities and some access to amenities. It is one of the few morphologies that can be found anywhere.

E. *Railway Buzz*

These areas are dominated by railway tracks and railway stations. They have no other major distinguishing attributes which may suggest that they are actually rather heterogeneous in physical structure.

F. *Suburban Landscapes*

These areas are typically of semi-detached houses, with good access to parks. They tend to be quite distant from town centres. They are primarily residential areas, and tend to be

close to schools. Cul-de-sacs are relatively common, probably because of organized developments and gated communities.

G. Countryside Sceneries

These areas are dotted with detached houses, and are located either near or within open countryside. Most rural villages fall into this category, along with some city fringe developments that lie beyond the classic suburbs.

H. Waterside Settings

The principal defining attribute of these neighbourhoods is their proximity to surface water such as rivers, canals or sea (these are very distinctive at the East of England, as illustrated by Figure 5.8). Some of these areas are ports, industrial or post-industrial sites. Distinctive infrastructure is arterial roads, i.e. roads wide enough to be used by lorries for the distribution of goods.

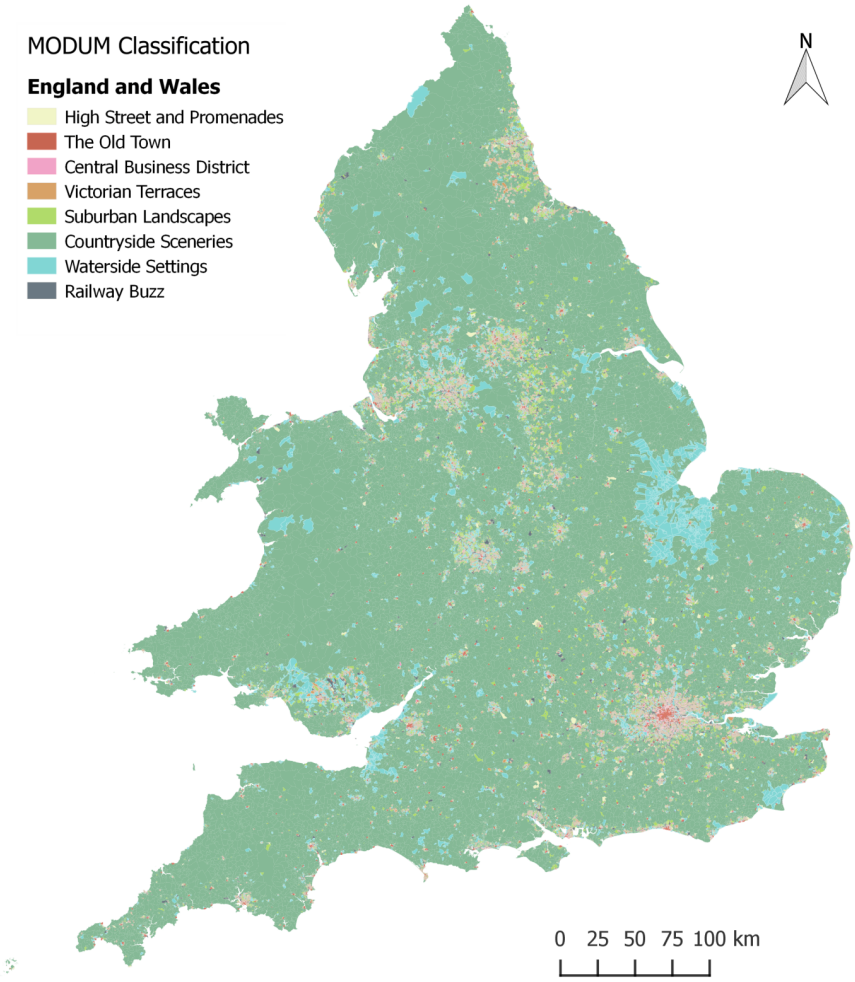

Figure 5.8 Mapping the MODUM classification for England and Wales.

5.4. A Comparison of MODUM and OAC

In order to test whether the Multidimensional Open Data Urban Morphology (MODUM) classification systematically follows the conventional OAC geodemographic classification, this research correlates the two sets of output classes via a contingency table. Table 3 shows the frequency distribution of MODUM within OAC 2011.

Table 5.3 Contingency table showing frequencies of OAC 2011 classes within MODUM.

	Output Area Classification 2011 - Super-Group Level								
MODUM Cluster Description	1 – Rural residents	2 – Cosmop olitans	3 – Ethnicity central	4 – Multi-cultural me/tans	5 – Urbani tes	6 – Suburba nites	7 – Constrai ned city dwellers	8 – Hard-pressed living	OA Amount
1 - Suburban Landscapes	5.53%	2.83%	3.38%	24.82%	23.77%	38.97%	22.12%	43.33%	46,788
2 - Railway Buzz	0.99%	10.61%	13.50%	10.09%	8.31%	3.08%	7.31%	5.33%	12,186
3 – The Old Town	0.25%	17.87%	5.35%	0.58%	4.05%	0.05%	4.76%	0.30%	2,812
4 – Victorian erraces	1.20%	14.43%	16.56%	43.93%	24.59%	1.79%	39.38%	34.98%	49,860
5 – Waterside Settings	8.43%	5.03%	3.56%	6.98%	12.08%	6.73%	8.04%	8.82%	12,468
6 – Countryside Sceneries	82.45%	2.05%	0.43%	2.91%	18.89%	47.79%	2.14%	3.90%	3,172
7 – High Street and Promenades	1.07%	6.20%	4.28%	3.00%	4.03%	1.50%	4.98%	2.47%	1,299
8 – Central Business District	0.08%	40.99%	52.94%	7.68%	4.26%	0.09%	11.27%	0.88%	52,823
Sum (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	181,408

Super-Group 6 – Rural residents seems to be identified fairly well by the morphological features, with a correlation of more than 82%, followed by a small percentage *of Waterside Settings* and *Suburban Landscapes*. About half the areas categorized as suburban also fall into this category, which is to be expected taken into account that typologies tend to blend out at the urban edges.

The expansive central areas seem to be mainly populated by *Super-Group 2 - Cosmopolitans* and *Super-Group 3 – Ethnicity Central*. Moving out of the centre, *Victorian Terraces* seem to scattered across three classes, *Super-Group 4 - Multicultural metropolitans*, *Super-Group 7 –*

Constrained city dwellers and *Super-Group 8 – Hard-pressed living*. The analysis of the suburban class makes an interesting case of why physical morphology is not always on par with socio-economic characteristics. While there is a ~40% match between the two classifications (classes 1-*Suburban Landscapes* and 6–*Suburbanites*), another 43% of the areas classified as *Suburban Landscapes* are populated by areas identified as *Hard-pressed living*.

Generally speaking, unique classes in the MODUM classification such as the old city centre and railway-heavy areas seem to be equally dispersed among classes. Some further analysis could provide better insight as to why, and even reveal interesting patterns. Figure 5.6 provides two different sets of maps of the area of Bristol and Leeds, in order to demonstrate the overall pattern relationships between MODUM and OAC.

This could show a dichotomy between the traditional affluent suburbs, and semi-suburban areas, in the sense of areas that have some suburban elements but are far less attractive, i.e. with neglected green spaces and badly maintained housing. Note that the neighbourhood morphology classification as presented here does not account for the quality of the physical characteristics; a green space for instance can either be a treed garden or picnic area or just a grassy field.

A more quantitative analysis of the correlation between the MODUM and OAC classifications is carried out by treating their frequencies as categorical values. In this instance the chi-square statistic x^2 *(49, 181408) = 136280, p < .001* of the two categorical values shows that the two classifications are not independent and have a significant relationship between them. The strength of the association can be measured by calculating the Cramer's V value:

$$V_c = \sqrt{\frac{x^2/n}{\min\,(r-1, c-1)}} \qquad (5.1)$$

where x^2 is the chi-square statistic, n is the total observations and r or c the number of rows or columns in the table respectively (whichever is smaller). The table above gives as a $V_c = 0.328$**,** which indicates a moderate level of association, given that V_c can take values between 0 (no association) and 1 (complete association).

A visual interpretation of the two classifications is always meaningful in evaluating emergent clusters, as illustrated by the following maps (Figure 5.9). The two classifications, MODUM and OAC 2011, share many common locations, especially towards the city centre. In general, axial zones exhibit much more strongly in the morphological classification (Figure 5.10), while OAC seems to have a more "regionalized" patterning, at least within local extents.

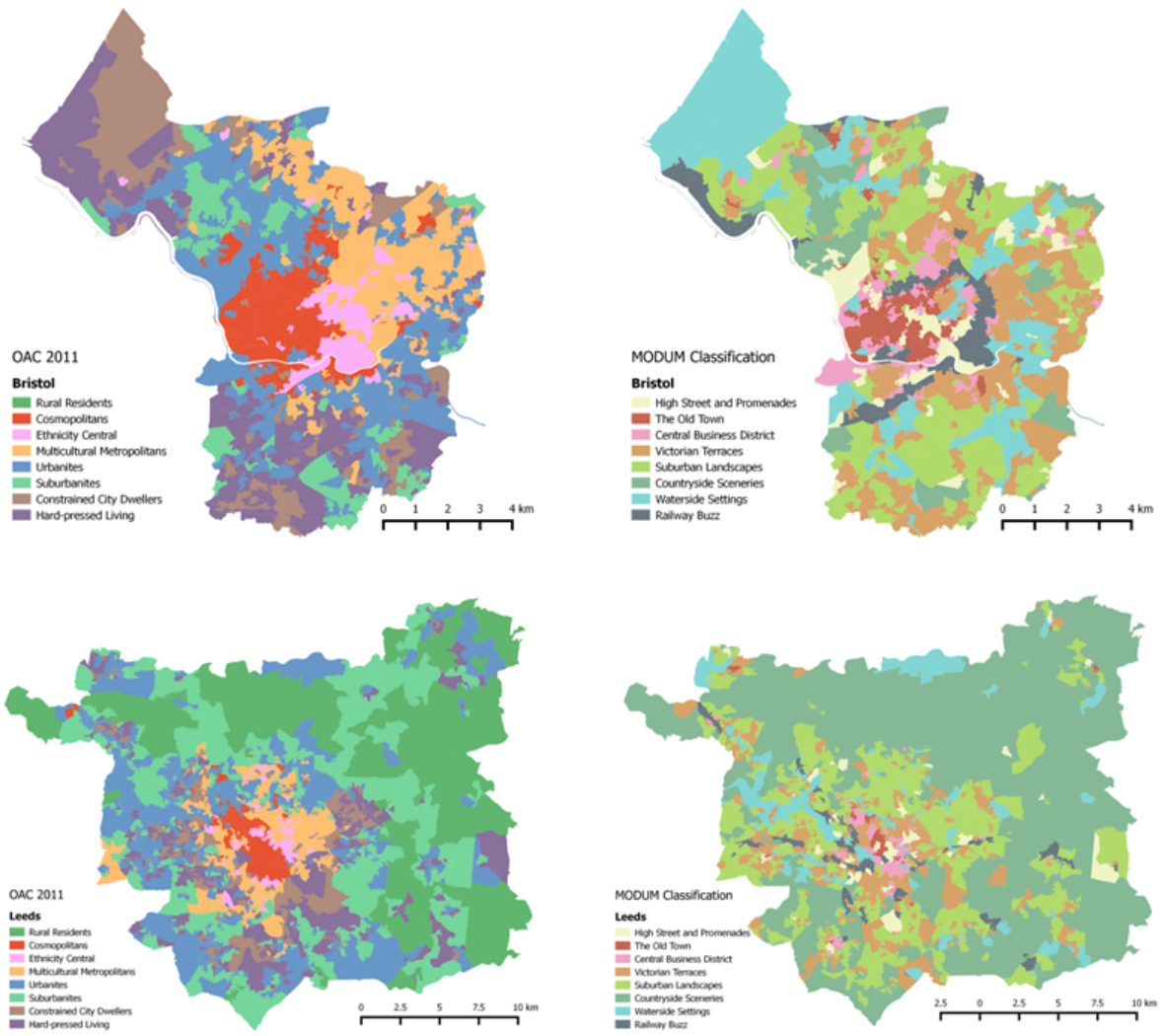

Figure 5.9 Built environment and socio-spatial patterns for the cities of Bristol (top) and Leeds (below).

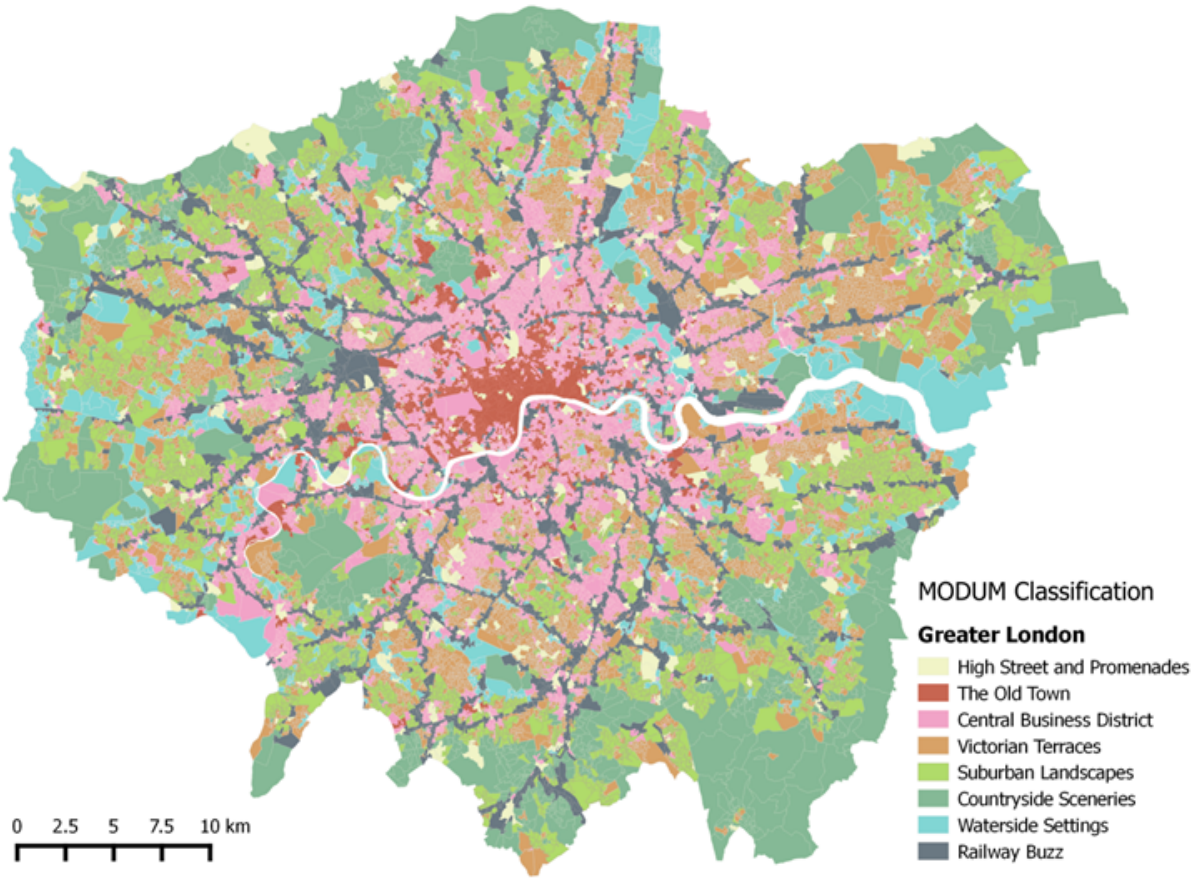

Figure 5.10 Mapping the MODUM classification for the Greater London Area.

5.5. Conclusions

The development of the MODUM classification illustrates that the production and analysis of a classification of the built environment using Big and Open Data can offer unique insights into some aspects of geodemographic structure of urban areas. The results capture through the multidimensionality of the data both microscopic and mesoscopic identifiers of urban morphology. A further step in this research would be to test whether socio-economic homogeneity within neighbourhoods is independent or not of built environment features. Such an analysis could offer cues on which areas are more prone to entail high uncertainty and how it can be handled.

Furthermore, the MODUM classification cannot only enhance socio-economic classifications by take into account microscopic variation, but also it can also prove useful in itself; it can provide

a simplified structure of the physical properties of geographic space that can be used to explore correlations with other spatial phenomena, potentially in a variety of applications, from real estate and house prices to health and wellbeing. In a dynamic sense, it can be used by urban planners and investors in the built environment to identify the areas in which the physical preconditions exist for neighbourhood renewal or upscaling.

On the other hand, the classification process described here is very specific to the underlying data and methodology. An inherent disadvantage of all geodemographic classifications is that lack of a single global optimization function during the classification procedure, making them highly susceptible to the operational decisions during the creation process (Openshaw and Gillard, 1978). However, this type of classification can be valuable in many circumstances. The classification is easy to use, and offers the ability to append and update data as they become available, while keeping the same model infrastructure intact. In general, it meets the growing need for geodemographic systems that are open and versatile enough to handle the abundance of big data that are currently available.

Chapter 6. A Systematic Evaluation of National Classification Performance Within Regional Contexts

6.1. Introduction to the Research Problem

As discussed in the second Chapter, the theoretical framework of geodemographics is the notion of spatial homophily, people's tendency to align themselves with the behaviour and aspirations of the local community they are living in. Since the collection and analysis of complex behaviour data would be next to impossible, by aggregating people based on a zonal area there is a premise that there is actually control for these variables, such as love for gardening, hiking, TV viewing choices, spending habits, etc. (Webber and Farr, 2001).

While it is convenient to use this notion while using within neighbourhood aggregations, there is no definite magnitude of the spatial effect of homophily; in most cases, the concepts of "*small-area*" or "*neighbourhood*" are arguably arbitrary. Such neighbourhood effects can very well expand much further than the zonal area that is selected for the aggregations. Furthermore, information about which of these attributes are influenced by neighbourhood effects is very limited.

This phenomenon can be observed in geodemographic classifications when neighbourhood types display a high degree of spatial autocorrelation. In theory, areal units across a region will have positive autocorrelation with other units that share similar attributes and vice-versa. If there are significant underlying attributes that are not accounted for, but affect much wider areas than those defined by, for instance, OAs, it could be suggested that by including a level of attribute similarity based on the geographic context of an area, these attributes can be controlled for. Areas that are proximal will thus tend to be more similar than those further away. In this approach classifications will be supportive of Tobler's First Law of Geography (Tobler, 1970). This property of geodemographic classifications has not been incorporated in conventional geo-classifications, while the methodological issues that arise have not been explored systematically in any way.

The issue under investigation is the "*aspatial*" nature of the classification methodology. Conventional geodemographic classifications have no input regarding the location or geography of neighbourhoods. As such, clustering algorithms like the *K*-means, account only for similarities in the attribute space; and areas are essentially treated as independent from one another. The traditional aspatial approach has a number of implications when generating profiles. Arguably,

national aggregations could sweep away contextual differences between proximal zones, reducing the local sensitivity of classifications and obscuring potentially important patterns (Openshaw, 1984). This type of ecological fallacy raises methodological questions regarding the accuracy of geo-classifications, given the inherent loss of within-cluster variation due to the aggregation process (Voas and Williamson, 2001).

To illustrate, Harris, Sleight and Webber (2005) argue that although some areas could have similar Census profiles, there could be underlying processes and dynamics associated with their location and geographic context that are not captured by the Census. Two postcode areas with the same behaviour in terms of employment composition may differ quite radically in those instances where areas are in transitioning phase between types.

For marketing related applications of geodemographics, a lack of local sensitivity may have fiscal implications, such as a reduced uptake of a product or service. However, in public sector uses, the consequences may be more severe, with mistargeting having potential implications on life chances, health and wellbeing.

Counter to this argument is that classifications constructed at the national, regional and local extent are effectively built for different purposes, and as such undermine comparison. This is a longstanding debate, originating in the earliest of UK classifications (Openshaw et al., 1980; and Webber, 1980). In particular, Openshaw, Cullingford and Gillard (1980) raised a number of methodological and philosophical questions regarding the success of national classification systems, specifically whether a national classification can provide a reasonable description of the differences within, as well as between, regions of the country.

To do so they tested both the new at the time National Ward/Parish Classification and Census Enumeration District (ED) Classification, by comparing them to a local classification for the region of Tyne and Wear, specifically built with the same data. In order to evaluate the level of similarity, they introduced three different types of evaluation procedures, a) a qualitative approach that looks at the resulting sets of cluster means to see if they are in agreement, b) a comparison of the SSE retained in the classifications and c) "*spatial correspondence*", i.e. a measure of how many spatial differences exist between sets of clusters that were previously identified to be broadly similar.

Their analysis showed that although the SSE differences remains relatively low (4.3%), the impact on ED cluster assignment is significantly greater, as shown by the mean spatial correspondence (39%). Based on their analysis, they argued that a lot of shortcomings in the national classification systems are evident. In particular, they found that most of the

disagreement between the classifications occurs between clusters that might be used to identify some kind of priority or transition area, as some of these differences are due to local conditions which it might be unreasonable for any national classification to identify. It is suggested that the sort of differences in cluster formation that occur between the national and local classifications are of considerable practical significance, and that the picture offered by a national classification can be misleading, especially in areas where it matters the most, such as in areas with high deprivation or semi-rural areas. As such, while national classifications provide a plausible general description of socio-economic conditions they cannot be used "willy-nilly" for planning application at the local level.

On the other hand, Webber (1980) responded to this critique by pointing out that classification differences are inevitable costs of working at a larger scale and not methodological deficiencies. He argued that a classification should be assessed based on its discriminatory power, and as such, differences in cluster formation are not misleading, at least to the experienced user, as they just show different dimensions of areal discrimination. He concluded that no only one representation of the data could be truthful, and that "*different representations are both inevitable and equally truthful*" (p. 445). Nevertheless, he recognised that a national classification of thousands of areas collapsed into a few dozen types will have a greater reduction rate that the same amount of types in a local classification, and in this framework, the efficiency of the classification performance should only be evaluated on the same level of data reduction.

He also questions the validity of the discrepancies shown by results in Tyne and Wear, not because of shortcomings in the classification methodology but because classifications operate on different means and standard deviations. A national compared to a regional/local scale classification would have different proportions of their population in different types of areas. For instance, he argued that the amount of immigrants, households in shared dwellings, or agriculture workers in Tyne and Wear, is so low that the standard deviation of each of these variables would be too low by national standards in order to influence a national classification, while on the other hand, these differences are inevitably exaggerated in the clustering process of the local classification. He further argued that when different approaches produce different results, these do not invalidate these approaches, nor does that inherently mean that "better" approaches could be elucidated by further research.

To illustrate, consider the following distribution of values regarding the Census variable *K41: Percentage of households with two or more cars or vans,* between the regions of South West, London and the UK (Fig. 6.1).

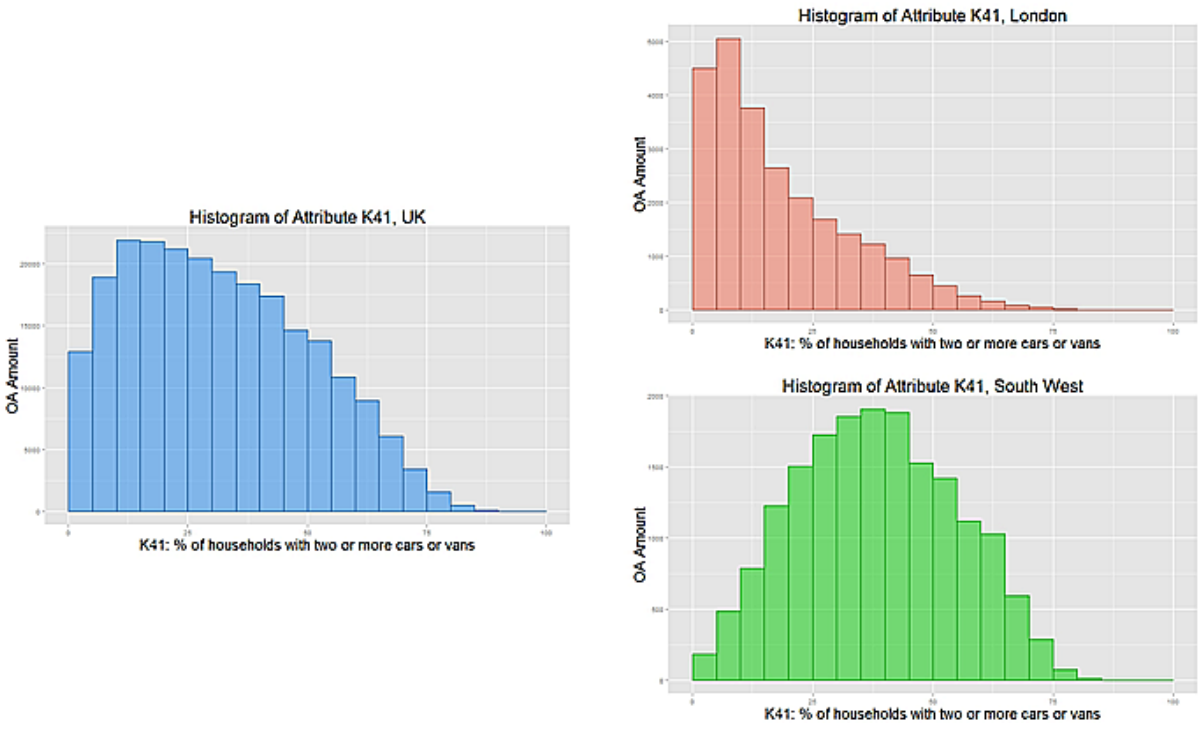

Figure 6.1 Number of areas per value of variable *K41: Percentage of households with two or more cars or vans,* for the Regions of London, South West and the UK (data source: Census 2011, ONS).

It is clear that the underlying conditions regarding car ownership differs quite radically between Regions. London has a steep curve implying that households with two or more cars are relatively scarce. On the other hand, the region of South West presents a completely different picture, as on average more than 40% of households demonstrate high levels of car ownership. Of course, London is by no means "poorer" than the South West. The issue here is not affordability; there are potentially other underlying reasons (e.g. availability of public transport, parking costs, etc.) relative to the London Region that impact car ownership that may not have been accounted for. If these conditions have not been controlled for, comparing values at a national scale may produce significant inaccuracies about the underlying socio-economic conditions.

Underlying conditions may differ considerably more at higher scales. Even if two areas share the same distribution shape, they may have a different distribution base or different local minima or maxima, indicating local variation across OAs that under a national perspective would be negligible. Consider the Census variable "*K45: Percentage of persons aged 16–74 who are unemployed"*. Figure 6.2 shows the distribution of values for the UK context and for the Liverpool LAD per se. The distribution of people who are unemployed in Liverpool is not only much more diverse, it is also wider with significantly more outliers, while there is no significant amount of

areas with zero unemployment as seen on the national scale. The differences in the distribution are also reflected in the averages and standard deviation values between the two geographies; the UK has an average of 4.54% per OA and a standard deviation of 3.38, while Liverpool has an average of 7.05% and 4.02 respectively.

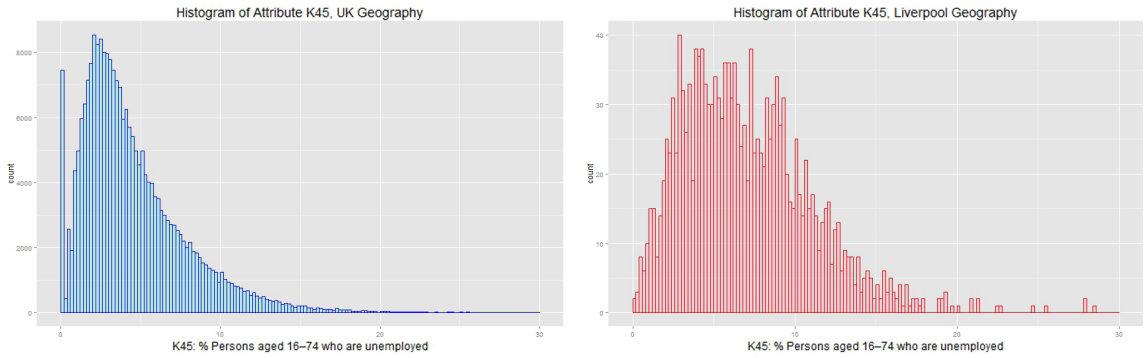

Figure 6.2 Number of areas per value of variable *K45: Percentage of people aged 16–74 who are unemployed* for the UK and for the Liverpool LAD (data source: Census 2011, ONS).

The postulations regarding the concerns of system-wide accuracy presented above can also be investigated by evaluating clustering outcomes of UK classifications, specifically the 2011 OAC. Figure 6.3 maps the total Squared Euclidean Distance (SED) from cluster means of the 2011 OAC Super-Groups at the OA level. The SED shows how well a cluster centre represents an area's attribute values. Low values suggest good representation while high values suggest poor representation. The positive spatial autocorrelation of areas with very high SED suggests that there is a spatial pattern of the "effectiveness" of the clustering process that may need to be addressed. SED values are clearly higher for Scotland, while Northern Ireland and London also appear to have high values. In general, places that have a more unique nature either because of historical or socio-ecomic conditions appear to be poorly represented by the OAC.

It is also worth noting that one would expect London to have considerably higher SED scores (given the shortcomings OAC faced and the subsequent development of LOAC), which could be an indication that London residents have indeed affected cluster means at a national/UK scale, although any further analysis on that particular impact in the OAC is outside the scope of this research.

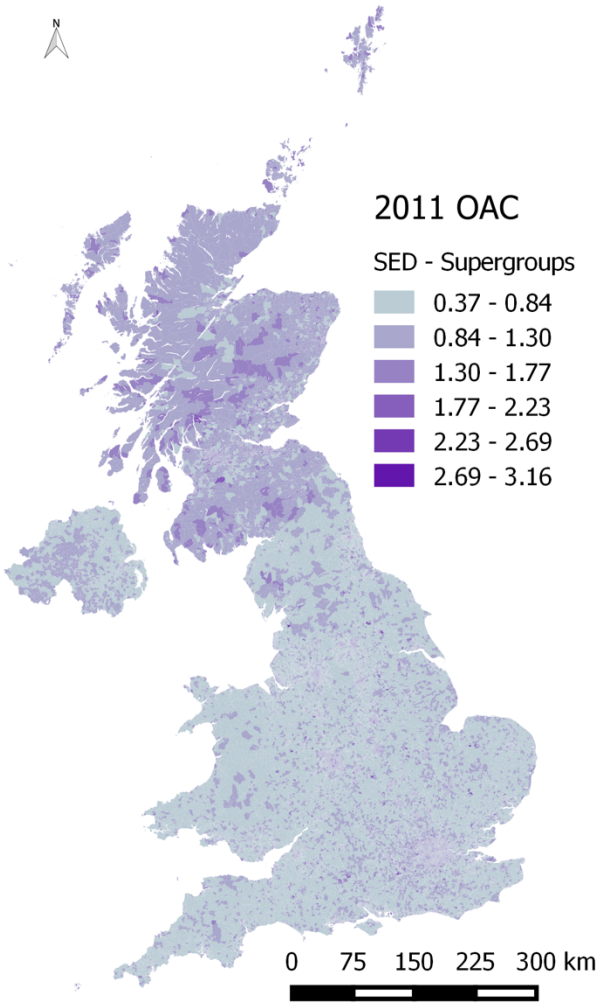

Figure 6.3 Square Euclidean Distance from cluster means, Super-Group level, as outputted from the 2011 OAC (Data source: <geogale.github.io/2011OAC/>).

Another issue for national geodemographic classifications is that conventional methodologies such as those used by the 2011 OAC are prone to weigh more absolute outliers in the data. This is partly because of the way the *K*-means algorithm works (see Chapter 3, section 4), and partly because of the range standardization technique that has been used in both previous OACs. Sometimes however the purpose of the classification is exactly that: to be able to highlight outliers and extremes in the data. Classifications such as the OAC are not in any way inadequate or inaccurate per se, they are just not built to attune to local variation, as it is defined here in this research. They do perform better under other assumptions or circumstances and doing so fit other purposes.

In this sense and as Webber (1980) states, the value of national classification systems is still compelling. Not all Local Authorities have the resources to undertake and to interpret their own

classifications, and even if they did, there is no way in which central government can compare the incidence of i.e. deprived neighbourhoods and make policy arrangements using individual local classifications when none of which are performed on a common basis. The same would hold true in the private sector, as the capabilities of companies to plan national branch development or conduct national surveys would be seriously compromised.

With all this in mind, there seem to be some shortcomings in conventional geodemographic systems with regards to national classification performance the impact of which has not been fully explored. Unfortunately, very little has been done within geodemographic research as a response to this problem. One such example is the London Output Area Classification (LOAC), which was developed precisely because national classifications such as OAC may not adequately accommodate local or regional structures that diverge from national patterns (Singleton and Longley, 2015). Some research is also targeted towards improving uncertainty levels within geodemographic classifications by inserting locational information through multilevel modelling (Harris and Feng, 2016). In general, there have been very few instances of new geodemographic approaches on how local contexts could be handled within geodemographic classifications systems.

In the academia, fuzzy clustering has been implemented to account on an ex post facto basis for geographic context, defined then as "*neighbourhood effects*". The modelling framework provided by Feng and Flowerdew (1998), uses an extension to a fuzzy clustering algorithm that adjusts cluster membership values based on neighbouring units. Although there is merit in this approach, the method suffers from a very localized definition of near-geography. Neighbourhood effects account for only the immediate neighbours and not the general geographic context of the area. Effects between proximal zones are weighted by the length of their common boundary, so if no common boundary is present the model does not assume any spatial interaction. The technique uses two biased parameters *a* and *b*, which represent the weight of the initial cluster membership and neighbouring cluster membership respectively, which can also significantly affect results.

Others have addressed this issue by measuring attribute values in terms of spatial autocorrelation. Adnan, Singleton and Longley (2012) provide a framework for "*Spatially Weighted Geodemographics*" by adjusting attribute values to *z*-scores derived from *Getis-Ord Gi** statistic (Getis and Ord, 1992), prior to clustering. A contiguity structure for the 4 closest neighbours is calculated at the Ward level for the London region and then two Census variables, "*Rent (Public)*" and "*2+ car household*" are used in the calculations. The result of this process is the conversion of attribute values to indicators of spatial association. Although results were

preliminary, their analysis provide a new perspective on geodemographic methodology as they are using of the spatial dependence of attributes (e.g. whether these belong to "*hot*" or "*cold*" spots) and not the attributes themselves.

This approach generally produces a "smoothing" effect on the distribution of variables as areas with high/low values relative to their neighbours are given lower/higher values. The statistic is essentially a *z*-score testing the significance of the area's locality being a "hot" or "cold" spot within the study area. In this framework, a clustering algorithm using these transformed attributes would introduce some level of contiguity among neighbourhoods. However, a geodemographic built with this approach would differ considerably from the traditional geodemographic approach, and the scope of such a classification would be very uncommon. Interpreting this geodemographic would be quite difficult; for instance, clusters with very high values in one attribute would suggest they are comprised of areas in which this attribute's values are both high and clustered together, low values would suggest that the attribute's values are both low and clustered together, while values near zero would indicates no apparent spatial clustering (without any information on the original attribute values). Furthermore, the *Getis-Ord Gi** statistic is extremely dependant on the number of *K* nearest neighbours, the selection of which is not very intuitive.

While such attempts within the academia are very limited, there are cues that some effort has been put forth to include broad spatial interactions within the private sector. Information about how exactly measures are incorporated however is very scarce. It appears most of these introduce a number of attribute weighting techniques, typically through the implementation of radial buffers or zones. Harris, Sleight and Webber (2005) report that proprietary classification providers such as Experian account for geographic context by using a series of concentric circles drawn from the small-area zone and expanding outwards. Geography is incorporated by adjusting attribute values using the using the respective sets of circular context. Allegedly, it has proven useful in differentiating within-city and outer suburban areas and crime risk. The method however is described very obscurely; there is no detailed information about the radius of these circles, or how the original values are adjusted, or if the context is incorporated into the clustering process or through other means.

To conclude, without any systematic evaluation, methods used to take into account near-geography could be geographically crude. They account for spatial context through either an arbitrary zonal distance or through spatial dependencies of adjacent spatial units. Still, near geography should not be evaluated in terms of a fixed distance but correspond with some level of organisation of actual communities (Alexiou and Singleton, 2015b).

In order to put the "*geo*" into geodemographics, this research set outs to a) construct an analytical framework to systematically assess the magnitude of discrepancies in clustering outcomes between local and national classifications and b) propose a theoretical model that can potentially be useful in incorporating spatial dependencies in the classification procedure. The first part of the research involves an exhaustive comparison between national and regional/local classifications. The "national" and "regional" descriptors are used here intuitively, with national referring to the UK context and regional to any other higher-scale geography. The second part of the research regarding the methodological framework for the model extension will be presented in Chapter 7.

The main concept is defined here as "*geographic sensitivity*", and measures the degree of influence of the near-geography to the overall similarity between neighbourhoods. A geodemographic classification with high geographic sensitivity will therefore tend to cluster together proximal areas more than a classification with low geographic sensitivity. A key step in the analysis would be to first define *near-geography*, in terms of spatial proximity or geographic context. Although results are inherently of tentative nature, they can provide a better understanding of the effects and extent of the problem.

Within this Chapter, in addition to the theoretical framework, an exploratory analysis is carried out in order to evaluate how scale variation affects, ceteris paribus, cluster formation. The three main research questions that the exploration tries to answer are:

1. *How can the geographic context of an area be defined?*
2. *How can comparisons be drawn between national and local classifications?*
3. *To what extent does contextual geography impact overall classification outcomes?*

The following sections of this Chapter present the geographic contexts, data and the clustering techniques used to create national and regional geodemographic classifications for the UK. A set of comparisons is carried out between classification, both in terms of cluster composition and cluster assignment. Summary results of comparisons are presented for each geographical context, in addition to a visual analysis of the differences in clustering outcomes for a selection of regions.

6.2. Selection of Geographic Contexts

A critical decision in the analysis is the selection of suitable geographic contexts that spatial dependencies will be based on, thus fulfilling the definition of near-geography. Since there is no

prior research to this problem, a set of contextual zones will be selected and tested independently. Options regarding the selection of geographic contexts are unfortunately very limited. Due to the nature of data availability, it would be difficult to completely decouple any level of administrative zonal geography from examined contexts. Moreover, there is no prior research or any a priori hypothesis on the extents of these areas, nor if these extents are consistent nationally. The theoretical framework suggests however that these areas must be internally cohesive in terms of the socio-spatial patterns identified in order to maximize accuracy, similar to a MAUP approach. Taking all these factors into account, three tiers of contextual geography, in addition to a UK one, was taken into account for the analysis (Fig. 6.1):

1. *Regional Level (Regions)*
2. *Travel-to-Work Areas (TTWAs)*
3. *Local Authority Districts (LADs)*

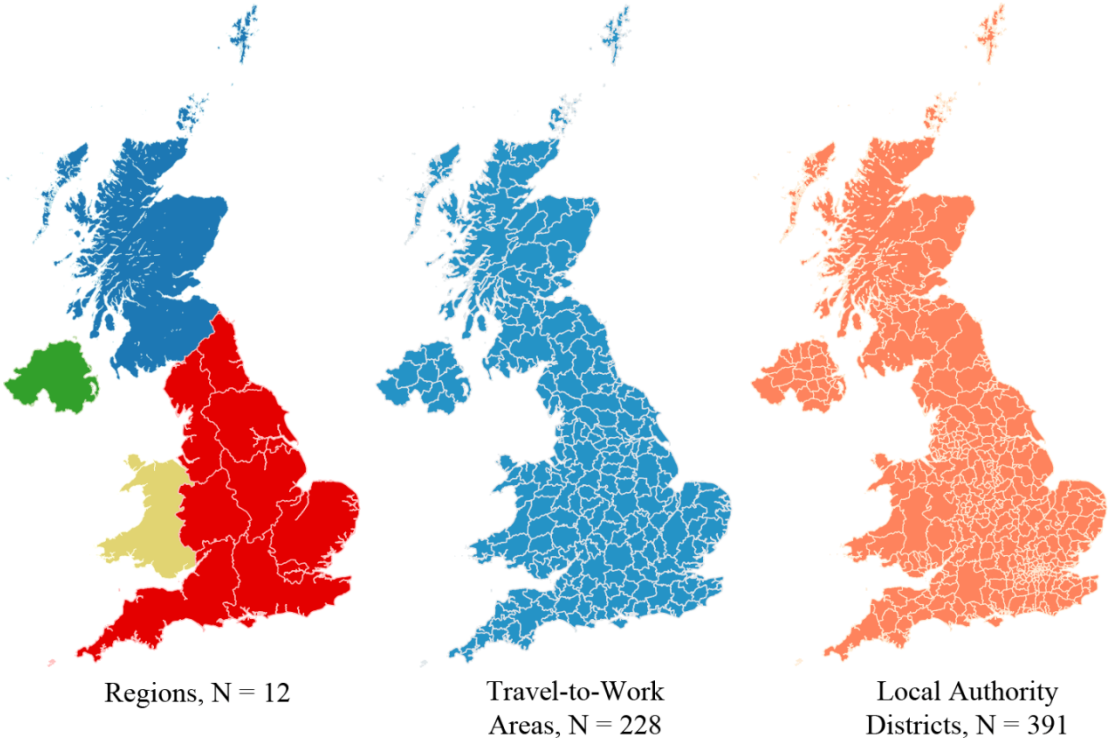

Figure 6.4 The set of geographic contexts that will be used in the analysis (Data Source: ONS).

All three geographies are consistent with our data availability needs and offer complete and non-overlapping coverage across the UK. The Regional and Local Authority levels are purely administrative, however TTWAs are defined to approximate self-contained local labour market areas, where the majority of an area's resident workforce work and live, so they have some

degree of territorial cohesion (ONS, 2015a). Their sizes vary, but generally they lie in between the regional and local scale. They also have the advantages of being consistent across the UK and fit existing lower level administration geographies. Considerations were made to include a sub-regional level of geography larger than TTWAs, which would amount to 2-3 sub-regions per region. The administrative level of NUTSII would actually fulfil these requirements (e.g. there are 40 NUTSII areas in the UK as of 2015), however NUTSII areas are traditionally highly volatile in the UK context; they do not appear to abide to any consistent methodology regarding their creation and they have been constantly redefined almost every year for the past decade.

6.3. Data

The next step of the analysis is assembling a dataset that can be fed into the classifications. In order to maximize comparison efficiency to the 2011 OAC, the same set of variables will be used. The dataset comprises of 60 Census variables, and includes attributes across demographic, household composition, housing, socio-economic, travel behaviour and employment domains. It represents a complete and comprehensive dataset for England, Wales, Scotland and Northern Ireland. The data is assembled in its entirety with 2011 Census variables, provided by the Office for National Statistics and aggregated at the Output Area level. Table 6.1 provides a short description of each variable.

Table 6.1 Selection of input variables from the 2011 Census.

Variable	Variable Description
	Demographic structure
K1	% Persons aged 0– 4
K2	% Persons aged 5–14
K3	% Persons aged 25–44
K4	% Persons aged 45–64
K5	% Persons aged 65–89
K6	% Persons aged 90+
K7	Number of persons per hectare
K8	% Persons living in a communal establishment
K9	% Persons aged over 16 who are single
K10	% Persons aged over 16 who are married or in a registered same-sex civil partnership
K11	% Persons aged over 16 who are divorced or separated

K12	% Persons who are white
K13	% Persons who have mixed ethnicity or are from multiple ethnic groups
K14	% Persons who are Asian/Asian British: Indian
K15	% Persons who are Asian/Asian British: Pakistani
K16	% Persons who are Asian/Asian British: Bangladeshi
K17	% Persons who are Asian/Asian British: Chinese and Other
K18	% Persons who are Black/African/Caribbean/Black British
K19	% Persons who are Arab or from other ethnic groups
K20	% Persons whose country of birth is the United Kingdom or Ireland
K21	% Persons whose country of birth is in the old EU (pre 2004 accession countries)
K22	% Persons whose country of birth is in the new EU (post 2004 accession countries)
K23	% Persons whose main language is not English and they cannot speak English well or at all
	Household composition
K24	% Households with no children
K25	% Households with non-dependent children
K26	% Households with full-time students
	Housing
K27	% Households who live in a detached house or bungalow
K28	% Households who live in a semi-detached house or bungalow
K29	% Households who live in a terrace or end-terrace house
K30	% Households who live in a flat
K31	% Households who own or have shared ownership of property
K32	% Households who are social renting
K33	% Households who are private renting
K34	% Households who have one fewer or less rooms than required
	Socio-Economic
K35	Individuals day-to-day activities limited a lot or a little (Standardised Illness Ratio)
K36	% Persons providing unpaid care
K37	% Persons aged over 16 whose highest level of qualification is Level 1, Level 2 or Apprenticeship
K38	% Persons aged over 16 whose highest level of qualification is Level 3 qualifications
K39	% Persons aged over 16 whose highest level of qualification is Level 4 qualifications and above
K40	% Persons aged over 16 who are schoolchildren or full-time students
	Travel Behaviour
K41	% Households with two or more cars or vans
K42	% Persons aged 16–74 who use public transport to get to work
K43	% Persons aged 16–74 who use private transport to get to work
K44	% Persons aged 16–74 who walk, cycle or use an alternative method to get to work
	Employment
K45	% Persons aged 16–74 who are unemployed

K46	% Employed persons aged 16–74 who work part-time
K47	% Employed persons aged 16–74 who work full-time
K48	% Employed persons aged 16–74 who work in the agriculture, forestry or fishing industries
K49	% Employed persons aged 16–74 who work in the mining, quarrying or construction industries
K50	% Employed persons aged 16–74 who work in the manufacturing industry
K51	% Employed persons aged 16–74 who work in the energy, water or air conditioning supply industries
K52	% Employed persons aged 16–74 who work in the wholesale and retail trade; repair of motor vehicles and motor cycles industries
K53	% Employed persons aged 16–74 who work in the transport or storage industries
K54	% Employed persons aged 16–74 who work in the accommodation or food service activities industries
K55	% Employed persons aged 16–74 who work in the information and communication or professional, scientific and technical activities industries
K56	% Employed persons aged 16–74 who work in the financial, insurance or real estate industries
K57	% Employed persons aged 16–74 who work in the administrative or support service activities industries
K58	% Employed persons aged 16–74 who work in the in public administration or defence; compulsory social security industries
K59	% Employed persons aged 16–74 who work in the education sector
K60	% Employed persons aged 16–74 who work in the human health and social work activities industries

Values were converted into percentages in accordance to their respective denominator, with the exception of variable *K7: Population Density* and variable *K35: Standardized Illness Ratio (SIR)* which were converted to ratios. The 2011 OAC suggests that an inverse hyperbolic sine transformation (see Table 4.5) of values offers the best classification results (ONS, 2015b). The same inverse hyperbolic sine transformation is used here to adjust the distribution of values to approximate normality, which is important in order for the *K*-means algorithm to work properly. The issue of standardization however, poses some difficulties when constructing local classifications. Standardization is dependent on the classification scale, and as such should be performed per contextual zone, prior to feeding data values in the clustering algorithm. The analysis will go into more detail about the standardization effects in the following sections.

As far as the contextual zones are considered, these were obtained as lookup tables per OA Code. For the administrative zones of Regions and Local Authority Districts these tables were distributed by ONS (<http://www.ons.gov.uk/ons/guide-method/ geography/products/census/ lookup/2011/index.html>). The 2011 TTWAs were distributed by data.gov.uk (<https://data. gov.uk/dataset/travel-to-worK-areas-2011-to-travel-to-worK-areas-2011_revised-uK-aug-2015-lookup>), as revised in August 2015, per LSOA Code for England and Wales, Data Zones (DZ) for Scotland and Super Output Areas (SOA) for Northern Ireland (an OA to LSOA lookup table was used as a medium). For convenience, information from the lookup tables was compiled with the

Census data into one complete dataset that assigns each and every OA to all the aforementioned geographic contexts (see *Appendix I: A. Data Input*).

6.4. Classification Methodology

While adjusting classification scale through different geographic extents is a critical step of the analysis, the basic model infrastructure of the *K*-means algorithm which has been already tested extensively and has been known to work well with this type of datasets is kept intact. As discussed in the first part of this thesis, there are many decisions that need to be considered when creating a classification. These parameters, such as the standardization, transformation and clustering method, affect considerably the classification results.

The vast amount of parameterization available in a classification process is too great to draw upon every possible combination within the classification model and its impact on classification comparisons. Without a single operational objective to optimize, identifying the best classification model through all variations would be exponentially complex, as it is next to impossible to consider all the combinations of all the methodological options in every step of the analysis.

In order to keep complexity to a minimum, parameters for the classification model will be inherited, where appropriate, from a baseline model, one that has been tested for its precision and effectiveness a priori. The most suitable model is the 2011 Output Area Classification. The 2011 OAC can be used as a base model as it is an established classification, taking from and updating the 2001 OAC. It has been thoroughly tested and evaluated, but more importantly it is an open classification, meaning that it transparent enough in order to be successfully reproduced. The initial datasets, methodology and outcomes have been documented and exist in the public domain. It was replicated using the *R* scripts provided at <*https://github.com/geogale/2011OAC*>, which provide detailed information on the model parameters and clustering methodology.

In this instance however, *z*-score standardization will be used, which takes into account the mean and standard deviation of the population (see Chapter 4, Section 4.3). While the baseline model uses range standardization to standardize variables, there are a number of implications regarding the transformation of variables only on the basis of local minima and maxima.

The main reason is that range standardization only takes into account variable extents, and not the distribution of the variable. This measure cannot be used to reflect local socio-spatial variations as does not integrate the mean and standard deviation of values and contradicts our

theoretical framework. Arguably, in the majority of cases the minimum value for Census variables is 0; distributions are right-skewed and a power transformation is generally used to "correct" this. In this sense, range standardization is mainly affected by maximum values. If extreme outliers exist on a far right end of the distribution, range standardization will impact considerably the remainder of values, and values are going to get "cramped up" in a small region of the distribution.

A standardization technique such as the *z*-scores on the other hand is based on subtracting every attribute value from the mean and dividing by the standard deviation and it is much better suited for an analytical framework of comparing regions. This is the only modification we enforce upon our baseline model. A more comprehensive look at the impact of standardization in regional classifications is presented in Chapter 7.

6.5. Exploration of Geographic Contexts

Before presenting the model for a geographically sensitive geodemographic analysis, an important question is whether a national classification can represent local socio-spatial patterns effectively. This is important in order to understand the impact of scale in creating neighbourhood typologies and assess the level of similarity among different scales.

The methodological shortcomings described and the examples provided offer some support to the initial research hypotheses. Further solidifying the model would require however a more in-depth exploration of the selection and impact of contextual geography to cluster types. With the theoretical framework established, it is important to assess the classification differences among the predefined geographic contexts and address research questions (2) and (3).

The aim of the exploration is to measure, ceteris paribus, the similarity between "regional" and "national" scale classifications. The "national" and "regional" descriptors are used here intuitively, with national referring to the UK context and regional to any other subset geographies, in this case at the Region, TTWA and LAD scale. The classifications are designed to be exactly the same in terms of parameterization, except for standardization of attribute values. The datasets are compiled from the 60 Census variables that were summarized in Table 6.1.

The exploration broadly follows Openshaw's, Cullingford's and Gillard's (1980) methodology and Alexiou and Singleton's (2015b) evaluation technique in order to test the level of "*fitness*" between the national and regional classifications. Specifically, it takes a closer look at how the

Regional, TTWA and LAD context affect a) cluster attributes means and b) cluster membership of OAs. The first measure of similarity is defined here as "*attribute fit*", which measures the degree of similarity between cluster attribute means, and the second as "*spatial fit*", which measures the OAs that their clustered membership remains unchanged between two classifications, similar to Openshaw's measure of spatial correspondence.

Results are demonstrated through a series of comparisons between a UK classification and classifications specifically made for every Region, TTWA and LAD in the UK, and calculate aggregate attribute similarity results per zone. The analysis includes the creation of a series of datasets which contain OA observations across 60 variables. Each dataset contains a subset of observations for every Region (12), TTWA (228) and LAD (391) in the UK. Data points are transformed using an inverse hyperbolic sine function similar to 2011 OAC, but each dataset is then standardized individually based on the Regional, TTWA and LAD contexts using *z*-scores:

$$z_{i,\alpha} = \frac{x_{i,a} - \mu_{S_a}}{\sigma_{S_a}}, \quad S_a = \{S_N\} \qquad (6.1)$$

where $x_{a,i}$ is the attribute value *i* of area *a* and $\mu_{S,a}$ is the mean and $\sigma_{S,a}$ is the standard deviation of the observations in area *a* of the national dataset S_N. In order to measure the contextual differences between the four geographical levels, the mean and standard deviation of the OA observations for area *a* is calculated and *z*-scores acquired according to equation (6.1).

Each dataset produced is used in a *K*-means algorithm to produce a classification (see *Appendix I: C. K-means per Spatial Context*). Taking into account that the same 60 attributes are used as the baseline model, the analysis assumes that the broad nature of clusters remains unchanged; and as such the cluster amount *K* remains unchanged. The comparisons are performed at the Super-Group level, which suggests that *K* = 8 at the national level. However, the *K* amount cannot be carried on to higher geographic scales. For instance, most urban LADs do not exhibit a rural typology, and some TTWA zones might not exhibit a student typology. In this framework, and under the assumption that the nature of clusters remains unchanged, the amount of clusters *K* for the various geographic contexts will be the same as the amount of cluster types present in the baseline classification (*see Appendix I: B. Find k per Geography*). The amount of *K* for every geography is summarized in the following table.

Table 6.2 Percentages of OAs based on the amount of clusters they exhibit in the 2011 OAC, for every geographic context.

K Amount	8	7	6	5	4	3	2	1
UK	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Regional	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TTWA	33.77%	20.17%	20.17%	23.68%	1.75%	0.43%	0.00%	0.00%
LAD	27.62%	29.15%	21.22%	18.67%	2.30%	0.76%	0.25%	0.00%

The exploration will be initially based on the Super-Group hierarchy of the cluster analysis. The OAC clustering methodology is top-down, which means that each Super-Group cluster is re-partitioned to 2-4 clusters by applying another *K*-means algorithm. The algorithm assumes that for every cluster, attributes are closer to assigned cluster centre than any other cluster. If substantive differences occur between classifications at the Super-Group level, then by default these differences will be carried on to some or all clusters at Group level. In this sense, measuring similarity at lower hierarchies would provide more accurate results, but not alter the level of disparity between classifications, which is the main scope of this research.

6.5.1. Attribute Fit

Once the optimized sets of *K* cluster assignments are calculated for each geographic context, clusters within each set are matched in order to determine which cluster ID from the outcome local classification "fits" better to the national one. Contrary to the typical qualitative way, i.e. cross-tabulation of the within-cluster distribution, the selection of matching clusters is based on the attribute fit between vector pairs. The difference to this approach is that instead of looking at one local classification and performing a qualitative assessment of cluster labels to draw relations between clusters, an algorithm has been developed to match clusters computationally at the UK level, which normally would take excessive time and effort.

To do so a metric of similarity should be initially defined. The highest level of similarity will match which set of cluster means fits best another. Cluster means are essentially a set of *k* vectors for *n* variable dimensions (attributes). There are various measures to compare vector similarity; the most typical is Euclidean distance. In this research however the Angular Cosine Similarity (ACS) is considered, which is given by:

$$ACS(A,B) = \frac{1 - cos^{-1}(cos\Theta)}{\pi}, \qquad cos\Theta = \frac{\sum_i^n A_i * B_i}{\sqrt{\sum_i^n A_i^2} * \sqrt{\sum_i^n B_i^2}} \qquad (6.2)$$

Both ACS and Euclidean measures can be used as distance metrics since they both qualify the triangle inequality property (as the angular cosine similarity is measured in radians), however ACS is selected for a number of reasons. One advantage of ACS compared to Euclidean is firstly informatory. Euclidean distance is calculated as the sum of differences between pairs of vector values. As such, the final value assessment poses difficulties when comparing different classifications, unless results are standardized to a common scale. The ACS on the other hand takes values between 0 and 1, where 0 means the vectors are exactly at opposite directions and 1 when they are exactly the same. Secondly, ACS has a more "*sentimental*" nature when evaluating similarity: if vector similarity is measured by the similarity of the direction of the vectors and the similarity of their magnitude, ACS has a tendency to weigh more the direction of vectors than their magnitude, while Euclidean distance weighs magnitude more.

This property of ACS is useful in drawing relations between cluster means. Since different areas operate on different means and standard deviations, ACS is not influenced by small variation of values; instead it weighs significantly more variation between high, average and low values in attribute pairs (directional change). One the other hand, Euclidean distance is calculated by the sum of all distances between value pairs, so it is susceptible to small variation of values, especially in a n=60 dimensional space. High variation in one or more pairs could be concealed by the total small variation in all other variables, which could potentially be very important in identifying the nature of the cluster. In general, Euclidean distance does not seem to work well in higher dimensions, as evident in data mining applications (Aggarwal et al., 2001).

With the distance metrics established, the analysis can evaluate similarity levels between classifications. The UK classification will be used as the basis that the similarity is tested against.

If $k_i^a = \begin{pmatrix} \mu_1^a \\ \dots \\ \mu_n^a \end{pmatrix} \in C_a$, represents a vector with the average attribute values μ_i^a of cluster k_i^a of the classification assignment C_a of area a, then a cluster k_i^a is more similar to another cluster $\mathrm{k_j^N} \in C_N$ derived from the same set of observations *N* when the $ACS\left(k_j^N, k_i^a\right)$ is closer to 1. Taking this into account, it is possible to find the combination of pairs for which:

$$C_a \cong C_N$$

If $C_a = \{k_1^a, k_2^a, \dots, k_K^a\}$ represents the vectors of the attributes means for K clusters, then there is one permutation P_a for which the similarity between the two classifications can be maximized:

$$\sum_{i=1}^{n} ACS(C_a, C_N) = \max \qquad (6.3)$$

To calculate this complicated and computationally expensive procedure, an algorithm is developed that performs the following steps in order to find optimal cluster pairs (see *Appendix I: D. Cluster Comparison*):

1. For area a, take the UK and contextual classification and extract the set of vectors of attribute means $\{k_1^N, k_2^N, \dots, k_K^N\}$ and $\{k_1^a, k_2^a, \dots, k_K^a\}$.
2. Obtain the ACS of k_i^a to k_j^N, for every *i* and *j,* and input value to a similarity (distance) matrix.
3. Obtain all the permutations P based on *n* taking *r* at a time, where *n* equals the cluster amount of the UK classification and *r* the cluster amount of the area a.
4. For every permutation P_i calculate the average ACS between attribute means based on the similarity matrix.
5. If C_{a,P_i} has greater ACS than the maximum, make this ACS value the maximum and make P_i the best combination (sequence) of cluster pairs.
6. Return to step 5 until all permutations have been calculated, and assume $C_{a,P_i} \cong C_N$.
7. Return to step 1 until all individual areas α have been processed.

As stated in step 2, the algorithm assesses similarity levels pairwise and outputs values in a similarity matrix (or distance matrix). An example of such a matrix for the case of West Midlands can be found in Table 6.3 below.

Table 6.3 Distance matrix showing the similarity levels between the UK and the West Midlands classification. In this case, the permutation of West Midlands the that best fits the UK classification is the *3,2,1,4,8,7,5,6* sequence, as noted by the highest value per column.

	UK1	UK2	UK3	UK4	UK5	UK6	UK7	UK8
WM1	0.650	0.272	0.911	0.422	0.598	0.648	0.575	0.277
WM2	0.369	0.921	0.271	0.453	0.337	0.390	0.363	0.705
WM3	0.859	0.309	0.677	0.323	0.572	0.831	0.676	0.278
WM4	0.281	0.602	0.248	0.807	0.550	0.320	0.395	0.570
WM5	0.699	0.316	0.655	0.337	0.409	0.577	0.873	0.438
WM6	0.340	0.736	0.271	0.444	0.249	0.352	0.408	0.937
WM7	0.654	0.400	0.626	0.388	0.559	0.871	0.515	0.366
WM8	0.540	0.289	0.627	0.686	0.852	0.542	0.479	0.218

Values in blue cells indicate a good similarity while values in red a low similarity. For instance, cluster UK1 has a very high similarity (0.859) to cluster WM4, UK2 to WM2, etc. Sometimes a cluster can be very similar to more than one clusters, such as in case of UK6, so the algorithm uses every possible permutation in order to maximize total fit. In this framework, the optimization criterion of the algorithm is the *maximum total fit*, as stated in equation (6.3). This is particularly important for smaller geographies such as the LADs, where cluster types are more exclusive. Note that the similarity matrix is not always square; the number of rows equals the number of clusters presented in the area under consideration, as listed in Table 6.3.

In order to better understand how the ACS compares cluster means, Figure 6.5 overlays the radial plots of the values between the West Midlands and UK classifications based on the algorithm matches. The UK classification is consistent across all comparisons, and as such it would be useful to define labels to clusters. A closer inspection of the attribute means shows that the clusters are very similar to the 2011 OAC (which is to be expected given they derive from exactly the same dataset) so the same cluster labelling is adopted.

Values range from 0.85 to 0.94 and the overlays show how the ACS metric responds to different levels of similarity. In general, rural areas, student and suburbanites remain similar; other clusters such as deprived neighbourhoods (*Hard-pressed Living*) and the transitional areas (*Cosmopolitans, Multicultural Metropolitans*) show a lot more diversity among regions. The latter for example, depends heavily on the nature of urban areas that are found predominantly in the heart of the city, as an amalgam of different types of single professionals and students. In this instance, the *Multicultural Metropolitans* Super-Group in the West Midlands Region are generally aged 16-25, less educated and working more part-time than full time. This type of class shows a pattern of a mixture of early-career professionals and part-time students, possibly working on technical jobs and outside higher education.

Figure 6.5 Radial plots comparing cluster attribute means and ACS levels for the UK and West Midlands classifications.

Table 6.4 Output table looking at cluster similarity of the UK regions compared to a UK classification at the Super-Group level.

			A.	B.	C.	D.	E.	F.	G.	H.
Region	Cluster Amount	Average ACS	Ethnicity Central	Rural Residents	Student Living	Hard-Pressed Living	Urbanites	Suburbanites	Cosmopoli-tans	Constrained City Dwellers
London	8	0.7318	0.78	0.76	0.65	0.82	0.61	0.86	0.67	0.71
Northern Ireland	8	0.7695	0.65	0.83	0.77	0.82	0.91	0.81	0.71	0.66
Scotland	8	0.7974	0.65	0.88	0.87	0.46	0.91	0.87	0.83	0.91
East	8	0.8029	0.86	0.91	0.60	0.82	0.65	0.83	0.90	0.85
Wales	8	0.8143	0.91	0.90	0.90	0.84	0.46	0.85	0.82	0.84
North East	8	0.818	0.91	0.52	0.90	0.80	0.81	0.92	0.82	0.87
North West	8	0.8187	0.79	0.50	0.73	0.91	0.92	0.91	0.85	0.94
South West	8	0.8237	0.90	0.85	0.72	0.64	0.88	0.88	0.81	0.90
South East	8	0.8254	0.90	0.85	0.88	0.91	0.74	0.81	0.82	0.69
Yorkshire & The Humber	8	0.8343	0.92	0.85	0.91	0.92	0.94	0.76	0.58	0.80
East Midlands	8	0.8546	0.78	0.83	0.89	0.86	0.78	0.92	0.92	0.85
West Midlands	8	0.8789	0.91	0.92	0.86	0.81	0.87	0.94	0.87	0.85

The information provided by Table 6.4 is exactly how the algorithm output results. This table can be useful in interpreting differences in socio-spatial structure at the local level. Clusters with low similarity suggest that the specific socio-spatial pattern would be very much different from a national classification perspective. At higher scales, neighbourhoods in these particular clusters will not be represented accurately in a geodemographic classification such as the OAC. This is particularly the case for clusters such as *Ethnicity Central* in Northern Ireland (0.65) and Scotland (0.65), the *Hard-Pressed Living* in Scotland (0.46) and the South West (0.64), the *Multicultural Metropolitans* in Yorkshire (0.58) and London (0.65) and the *Cosmopolitans* in the East of England (0.60), to name a few. Differences in similarity regarding rural areas that are visible also demonstrate the dichotomy between the South and the North of England, as seen from the very low scores of the *Rural Residents* cluster in the North East and the North West (0.52 and 0.50) compared to the high scores in the South East (0.85), the South West (0.85), Wales (0.91) and East of England (0.91).

At the regional level, results imply that geographic centrality has an impact in average scores. In general, central regions such as West and East Midlands respond better to the UK classification, while London, Northern Ireland and Scotland appear to score the least (Fig. 6.6). Considering the inherent differences in populations in these regions, the results appear to be consistent with expectations.

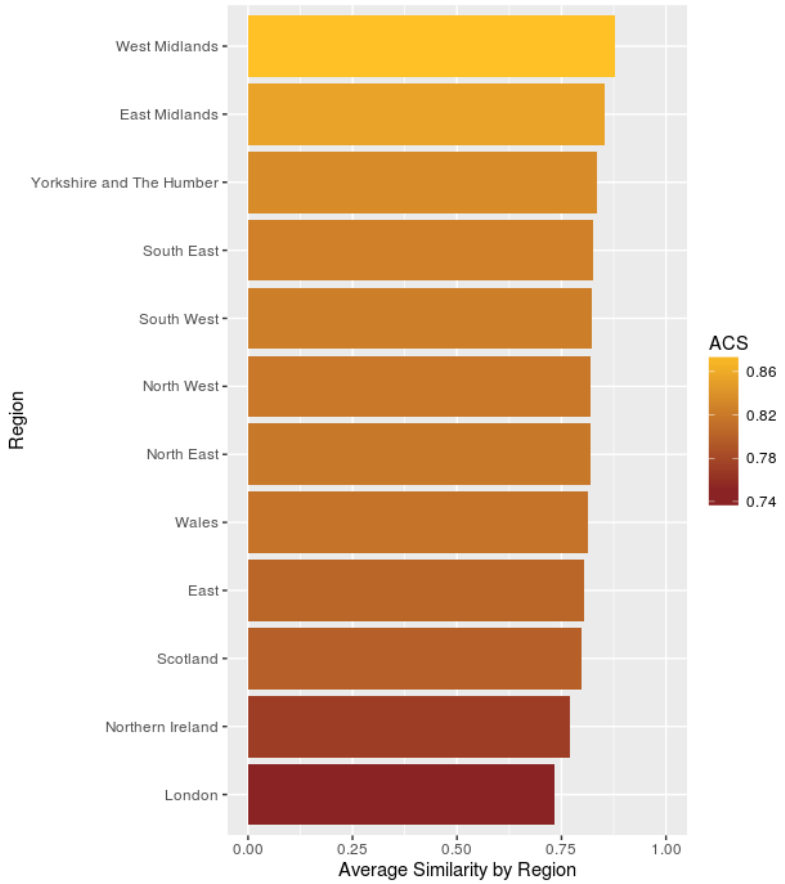

Figure 6.6 Average ACS score by Region.

The best fit belongs to West Midlands with 0.879 and the least to London with 0.731. The reasons why London scores so low are quite obvious, considering the London Region is essentially one very large metropolitan area and a "global" city. West Midlands on the other hand may score high in the ranking because it is a very diverse region. It has the largest conurbation besides London, Birmingham, but in contrast to that, the region also includes areas of remote countryside. It is also true nonetheless that the region has one of the highest proportion of economic inactivity, the highest proportion of people with no qualifications and the highest proportion of households living in relative poverty after London (ONS, 2012). This raises some questions as to why West Midlands appears to score so high. It could either be because of the heavy impact of London in the classification, which shifts cluster means towards London's socio-spatial patterns and skewing

results, or it could be due to the particular selection of input data of the 2011 OAC.

For instance, the percentage of people with non-white ethnicity. in West Midlands is the second largest with 17.3%, but it closer to the average national amount of 14% than other regions because London, with a 40.2% people of non-white ethnic background, shifts average values considerably (data source: *Census 2011, ONS*). Adding this to the fact that the selected dataset has very few socio-economic variables included, cluster formation is heavily weighed on demographics and household composition (26 out of 60 variables). Without any weighting scheme on the variable domains, some very important socio-economic attributes such as affluence, education, etc., may not have the desired impact needed to pick out underlying socio-economic disparities.

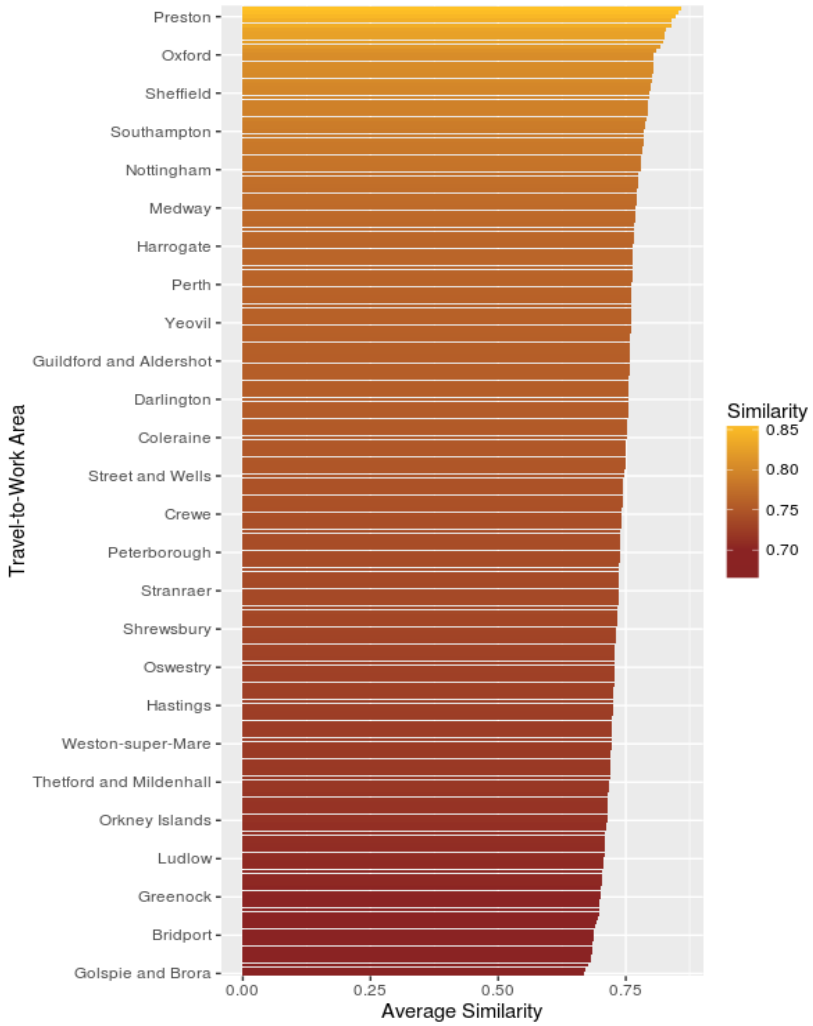

Figure 6.7 Average ACS scores by Travel-to-Work Area.

Travel-to-Work Areas similarity levels (Fig. 6.7) range from 0.668 (Golspie and Brora) to 0.868 (Bristol). On the upper scale of the similarity ranking are areas such as Preston, Plymouth,

Cheltenham and Cardiff (>0.85), followed by areas such as Reeding, Chester, Oxford and Durham (>0.80). On the antipode lie predominantly rural areas on relative remote regions, specifically Golspie and Brora, Fraserburgh, Bude, Thurso, Broadford, Kyle of Lochalsh and other areas found predominantly in North Scotland. Similarly to regions, there is a visible pattern of an *S* curve in the ranking of TTWAs, where one group of areas seems to behave very well in response to the UK classification, while another group seems to diverge significantly from national patterns.

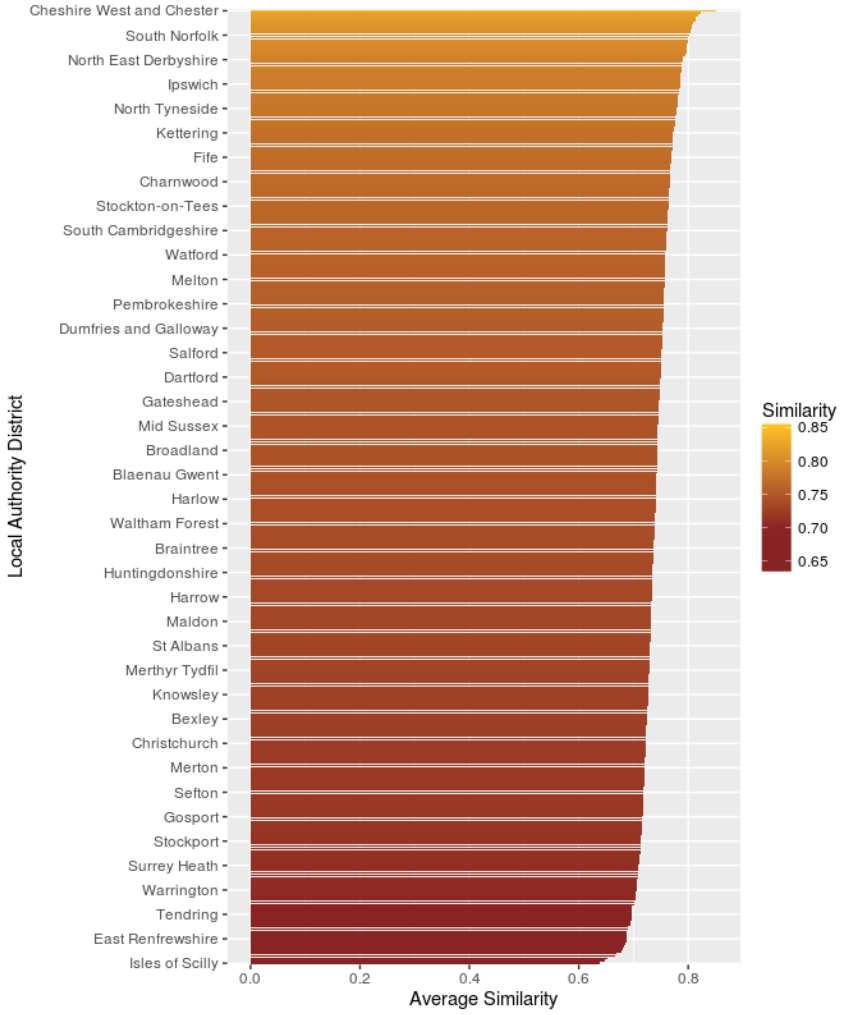

Figure 6.8 Average ACS scores by Local Authority District.

The same behaviour can be seen in the average ACS levels regarding the LAD context (Fig. 6.8), although similarity levels in this case are substantially lower. The main difference between LADs and TTWAs is that the list of bottom-ranking LADs is dominated by London Boroughs. Out of the 391 LADs, the 5 lowest scoring LADs are, with the exception of the Isles of Scilly (a special case as it only has 9 OAs) the Tower Hamlets (0.647), the City of London (0.652), Newham (0.667) and Hackney (0.669).

The level of divergence of London areas compared to national socio-spatial patterns are even

more evident given the few amount of clusters present in the OAC. The City of London LAD only has 2 cluster types (2011 OAC), which indicates that out of all 8 UK Super-Groups, no better fit was possible other than a 0.69 and 0.61. The attribute means of these two clusters are illustrated in Figure 6.9. The main differences in cluster formation seem to lie within the demographics, ethnic background, housing and industry domains. Newham and the Tower Hamlets also have very low scores with only 3 cluster types present. This suggests that the socio-spatial patterns produced by the *K*-means are highly divergent from national patterns. The results confirm the criticism to the OAC about regional accuracy that spurred the need for local-scope classifications such as the LOAC.

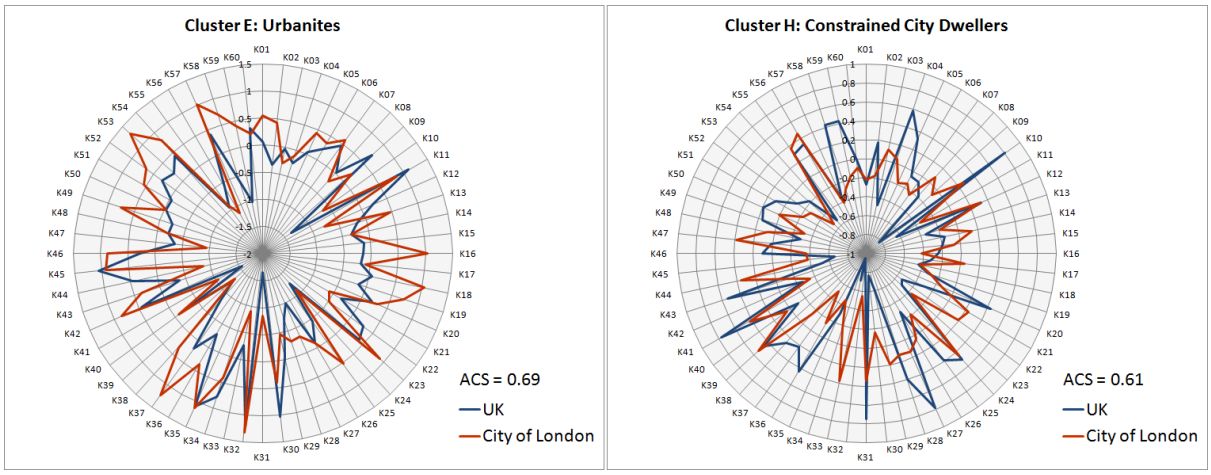

Figure 6.9 Radial plots comparing attribute cluster means and ACS levels for the UK and the City of London LAD classifications.

6.5.2. Spatial Fit

Attribute fit is looking only at one of the aspects of cluster formation. There are other factors that can impact overall similarity levels such as cluster size and algorithm performance. In order to solidify results and assess the impact of geographic scale in geodemographic classifications, as outlined in research question (3), the second measure, spatial fit will also be explored.

Spatial fit is defined by the amount of OAs that remains "unchanged" between two classifications; "unchanged" refers here to the "best fit" cluster pairs that were unidentified at the attribute fit stage. It is possible to correlate the classifications via a contingency table, as previously shown in Chapter 5. Table 6.5 summarizes the correlation between the Regional and UK classification. The labelling of the regional cluster has been done based on the cluster matches

to the UK classification, so the total OAs that remain unchanged is given by the diagonal of the Table 6.5.

Table 6.5 Spatial fit of the Regional Classification to the UK classification.

Regional Classification / **UK Classification**	Multicultural Metropolitans	Rural Residents	Ethnicity Central	Hard-Pressed Living	Constrained City Dwellers	Cosmopolitans	Urbanites	Suburbanites
Multicultural Metropolitans	**5,023**	2,633	273	22	49	4,512	988	3
Rural Residents	1,105	**21,507**	499	2,750	57	3	306	6,141
Ethnicity Central	359	89	**8,951**	1,885	3,400	712	2,807	143
Hard-Pressed Living	1,464	460	546	**32,980**	12,903	6	4,405	203
Constrained City Dwellers	2,509	108	1,604	3177	**12,547**	144	270	12
Cosmopolitans	422	4	135	2	28	**3,871**	82	3
Urbanites	6,603	1,051	1,258	5,097	2,572	322	**16,911**	2,504
Suburbanites	199	1,829	86	7,569	0	1	7,300	**36,892**
OA Sum	17,684	27,681	13,352	53,482	31,556	9,571	33,069	45,901
Cluster Spatial Fit (%)	28.40%	77.70%	67.04%	61.67%	39.76%	40.45%	51.14%	80.37%
Total Spatial Fit (%)								**59.70%**
Rand Index (RI)								**0.8244**

At the regional geographic context, the spatial fit is 59.70%, suggesting that by performing a regional standardization of values prior to classification, 4 out of 10 OAs would change their type. The majority of change would be within the *Multicultural Metropolitans* (28.40%), *Constrained City Dwellers* (39.76%) and *Cosmopolitans* (40.45%) Super-Groups, while the Rural Residents (77.70%), Suburbanites (80.37%) and Ethnicity Central (67.04%) seem to match substantially more. Results agree with the conclusions made from the attribute fit analysis, indicating a significant positive relationship between attribute fit and spatial fit.

Table 6.6 Spatial fit of the TTWA Classification to the UK classification.

TTWA Classification / **UK Classification**	Multicultural Metropolitans	Rural Residents	Ethnicity Central	Hard-Pressed Living	Constrained City Dwellers	Cosmopolitans	Urbanites	Suburbanites
Multicultural Metropolitans	**5185**	121	701	5	180	3692	3311	308
Rural Residents	811	**18450**	172	2560	447	430	597	8901
Ethnicity Central	328	13	**9324**	995	4288	1219	751	1428
Hard-Pressed Living	1623	3686	1437	**29806**	9488	1121	3495	2311
Constrained City Dwellers	1922	1033	2734	3896	**9880**	495	397	14
Cosmopolitans	1375	6	319	2	89	**2608**	144	4
Urbanites	4456	3017	2467	6243	1919	825	**12554**	4837
Suburbanites	472	11134	291	5140	32	381	5534	**30892**
OA Sum	16172	37460	17445	48647	26323	10771	26783	48695
Cluster Spatial Fit (%)	32.06%	49.25%	53.45%	61.27%	37.53%	24.21%	46.87%	63.44%
Total Spatial Fit (%)								**51.09%**
Rand Index (RI)								**0.7862**

Similar contingency tables can be calculated for the TTWA and LAD contexts. Unsurprisingly, there is a steady decline in spatial fit as the analysis progresses to finer geographies. The TTWA spatial fit is 51.09%, while the LAD spatial fit is slightly lower, at 48.63%. Results are summarized in Tables 6.6 and 6.7 respectively.

In this instance, comparison results are provided by fitting a regional classification into the UK classification, and not vice-versa. Provided cluster labels have been identified, a useful way to compare partitions both ways is by using the Rand Index (Rand, 1971), which measures how much pair-wise agreement there is in a set X to set Y, and how much agreement there is in set Y to X. Rand Index (RI) scores take values between 0 (complete disagreement) and 1 (complete agreement) and are also provided in the above tables. RI scores show that all classifications have an average agreement of about 80%. Unexpectedly, LAD scores marginally higher than TTWAs, which shows that TTWAs also produce very localised spatial patterns, despite being twice the size of LADs on average.

Table 6.7 Spatial fit of the LAD Classification to the UK classification.

LAD Classification / UK Classification	Multicultural Metropolitans	Rural Residents	Ethnicity Central	Hard-Pressed Living	Constrained City Dwellers	Cosmopolitans	Urbanites	Suburbanites
Multicultural Metropolitans	**5771**	943	799	279	772	690	2593	1656
Rural Residents	698	**17379**	174	2420	638	572	991	9496
Ethnicity Central	921	277	**7523**	3159	3541	528	1007	1390
Hard-Pressed Living	1448	2103	2828	**27629**	10546	883	4774	2756
Constrained City Dwellers	2617	583	2587	4296	**9157**	726	393	12
Cosmopolitans	1467	55	306	10	137	**2382**	181	9
Urbanites	4714	2443	3761	5749	1469	908	**12754**	4520
Suburbanites	534	10321	648	4953	66	481	6496	**30377**
OA Sum	18170	34104	18626	48495	26326	7170	29189	50216
Cluster Spatial Fit (%)	31.76%	50.96%	40.39%	56.97%	34.78%	33.22%	43.69%	60.49%
Total Spatial Fit (%)								**48.63%**
Rand Index (RI)								**0.7873**

In general, much of the disagreement seems to stem from the inconsistencies at the identification of several key classes. The individual cluster fits can give more insight into cluster relationships. From the cross-tabulation, it seems that differentiating between rural and suburban areas at the regional scale is much more difficult. Both at TTWA and LAD scales have approximately 50% - 60% match between classifications at the rural and suburban classes, with a remainder 30% - 40% clustered as suburban and rural respectively. It is also clear that another source of differentiation is a result of the correlations between the *Multicultural Metropolitans, Constrained City Dwellers* and *Cosmopolitans* Super-Groups. The *Cosmopolitan* type seems to be problematic as it tries to capture mainly students, a typology which is very diverse across the UK. *Hard-pressed Living* and *Constrained City Dwellers* also appear to be heavily influenced by geographical context.

It is also worth noting that proportional cluster sizes do not change dramatically among classifications. While some further research may be needed in order to draw definite conclusions, one interpretation could be that although cluster attributes means “move” between the various classification scales, they do so rather consistently for every area, so the distance between

clusters in attribute space remains relatively constant.

6.5.3. Visual Analysis

While these two metrics provide a good basis for a quantitative evaluation of the impact of geography in geoclassification systems, they can be difficult to interpret. The most effective way to relay such information is by mapping results. A visual analysis of how classification differences take place spatially can offer much more insight about the type and magnitude of the effects laid out in this analysis. A key step in this analysis is to assert similarity levels at a neighbourhood level. Figure 6.10 provides a series of maps comparing the UK to local classifications at the Liverpool area. Similarity scores are calculated per individual OA as provided by the ACS scores of their respective cluster attribute means.

Results show that each geographic context provides diverse results. At the Regional level, dissimilarity seems to be concentrated toward urban areas and near the core of the city. The TTWA level seems to respond differently, as most dissimilar neighbourhoods are located in the city fringes, where neighbourhoods are more likely to be more transitional. Finally, at the LAD level, distinct spatial patterns are less evident, and dissimilarity is scattered across neighbourhoods.

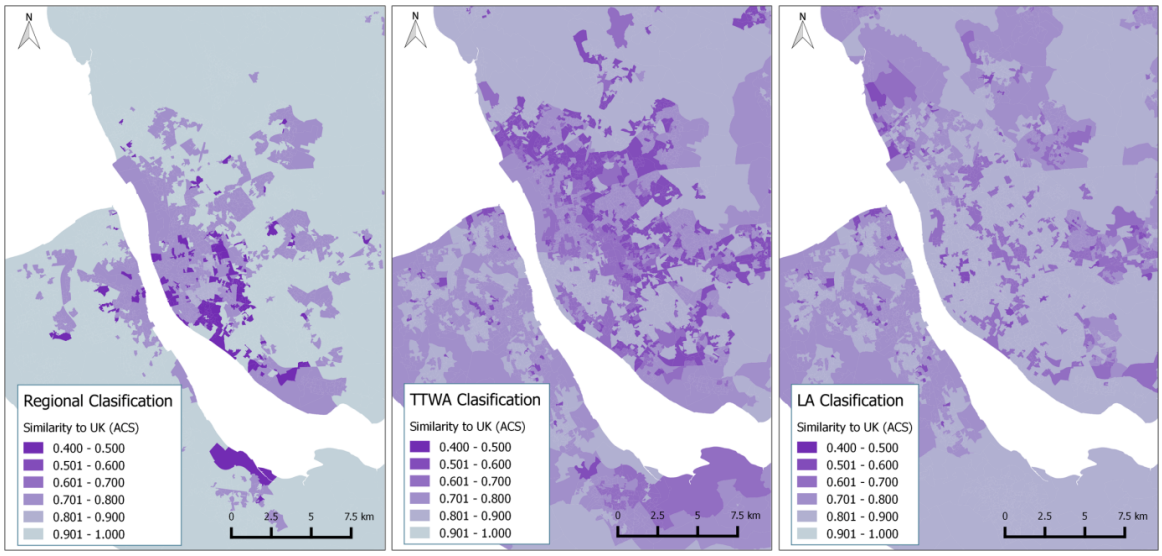

Figure 6.10 Maps demonstrating the disparities between local and national classifications based on their cluster attribute means per OA level (Liverpool area).

The regional classification is the most interesting as the spatial pattern follows to some extent the spatial pattern of the least affluent areas around the Liverpool region. Further analysis

revealed that similarity levels are also associated with the spatial performance of the national 2011 OAC (Fig. 6.11). This provides some more evidence towards the hypothesis that OAC SED scores do follow specific spatial patterns, which implies that some underlying spatial attributes have not been accounted for in the classification.

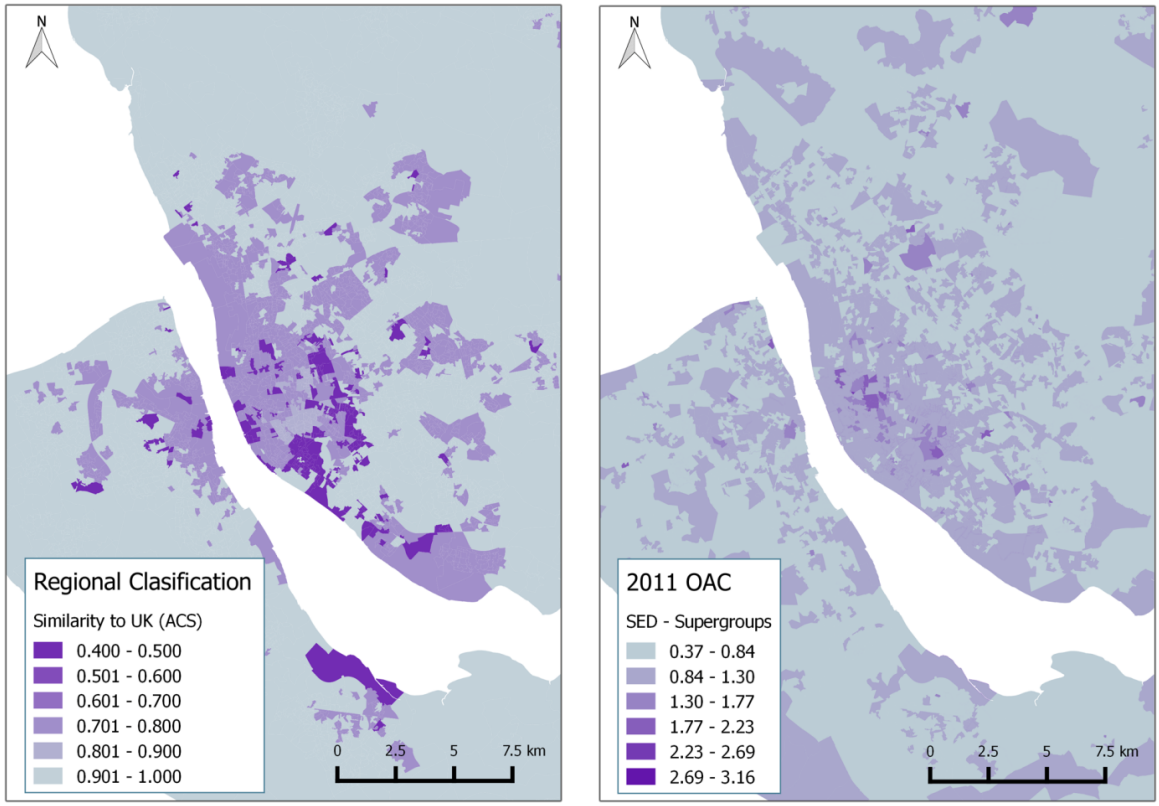

Figure 6.11 A comparison between Regional - UK classification disparities and SED levels of the 2011 OAC at the OA level, Liverpool area.

In general, changes in the Regional Classification appear to take place the most in areas that national Classifications cannot classify accurately. To what extent these changes reflect actual socio-spatial patterns is still uncertain without external evaluation. Similar results were observed for the Greater London area, although the SED variation in the region is more uniform than expected (Fig. 6.12).

One way to interpret the nature of changes is by looking directly at the transitions of cluster typology. For this purpose, the areas of Liverpool, Brighton and Greater London have been selected. A series of maps were created in order to demonstrate the cluster analysis results between different geographic contexts.

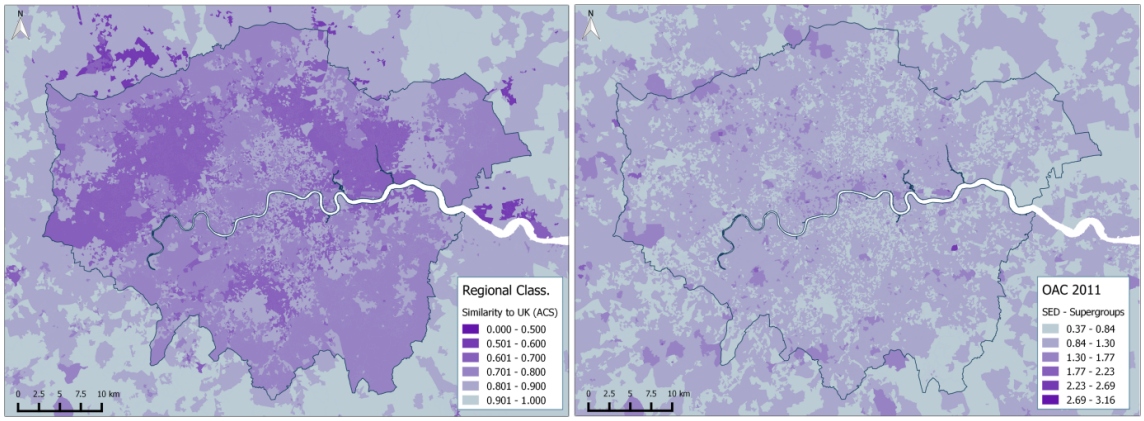

Figure 6.12 A comparison between Regional - UK classification disparities and SED levels of the 2011 OAC at the OA level, London Region.

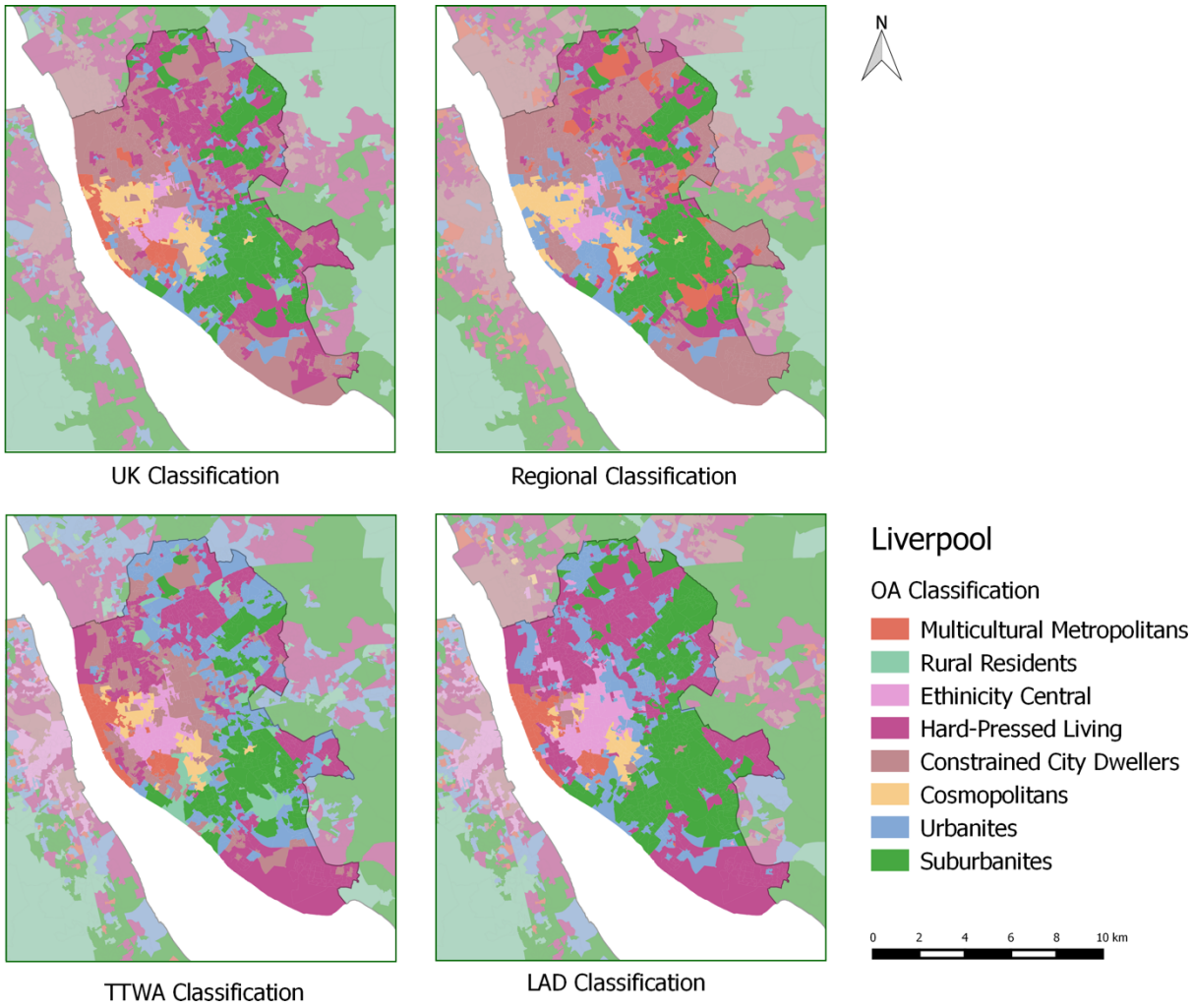

Figure 6.13 Classification comparisons through mapping typologies for each contextual geography used in the adjustment of attributes, Liverpool LAD.

Figure 6.13 illustrates the city of Liverpool. At first glance, a distinct differentiation is the suburban typology. At higher scales, there is a visible tendency of rural typologies to be considered suburban along the city fringes. The majority of changes however are within the clusters *Hard-Pressed Households* and *Constrained City Dwellers*. In general, as the analysis progresses to higher scales, the less the latter class appears in the spatial pattern. The exception seems to be the Regional level, which has a higher percentage compared to the UK level. Neighbourhoods with a *Multicultural Metropolitans* typology in the city centre are also replaced by *Urbanites*. Overall, the majority of the differentiation can be found within the transition of *Hard-Pressed Households* to *Constrained City Dwellers and Urbanites*. TTWAs also offer a very interesting pattern, where *Urbanites* seem to act like a "buffer" zone between hard-pressed neighbourhoods and other classes, indicating transitional relationships between these classes.

A likely explanation for this phenomenon is the heterogeneity of the North West Region as a spatial unit of analysis. Taking a look at the LAD rankings, areas such as Preston and Chester appear to score very high, Liverpool has a moderate score, while Warrington and most Greater Manchester LADs appear to score very low, such as Wigan and Oldham. In this framework, some classes such as *Constrained City Dwellers* are a lot more distinctive at the Regional level and a lot more common at more local levels, hence the differentiation.

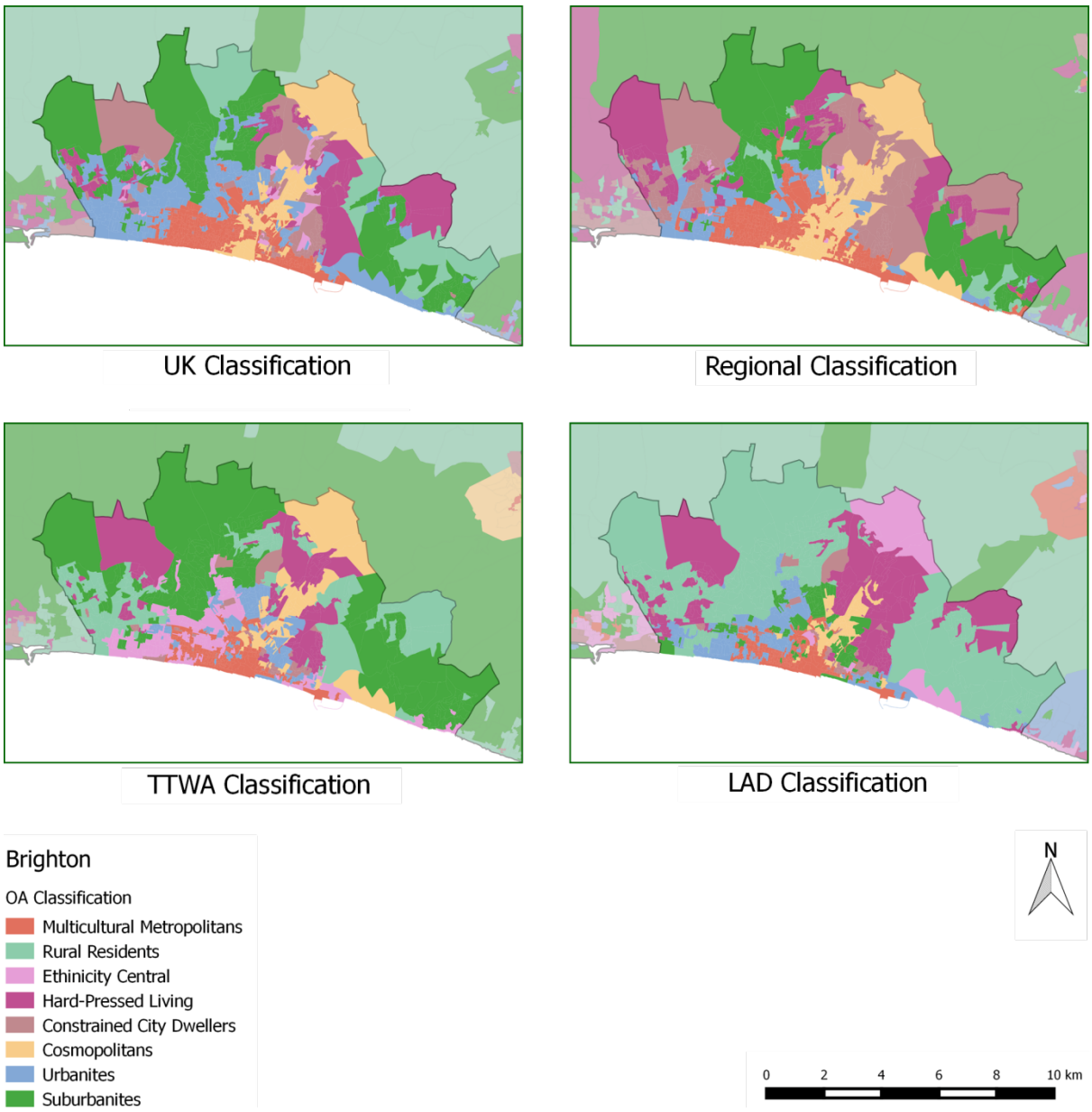

Figure 6.14 Classification comparisons through mapping typologies for each contextual geography used in the adjustment of attributes, Brighton LAD.

In Brighton, the general trend of class transitions is very similar to Liverpool with some notable exceptions (Fig. 6.14). At the regional level, there are visible spatial patterns of Suburban and Rural neighbourhood types in the UK context that are classified as *Hard-Pressed Households* and *Constrained City Dwellers*. A way to interpret this is that the South East Region is generally very affluent, so attribute standards for what is considered affluent in the national sense might be a lot higher in the regional context. On the other hand, areas around the city centres remain similar between the UK and Regional classifications. This could be due to the nature of settlements in the South East; most residents living near city centres are likely to live and work in that city and so

the general pattern remains relatively constant, while settlements outside or at the fringes of urban areas could be potential London commuters, which suggests higher incomes and different household characteristics.

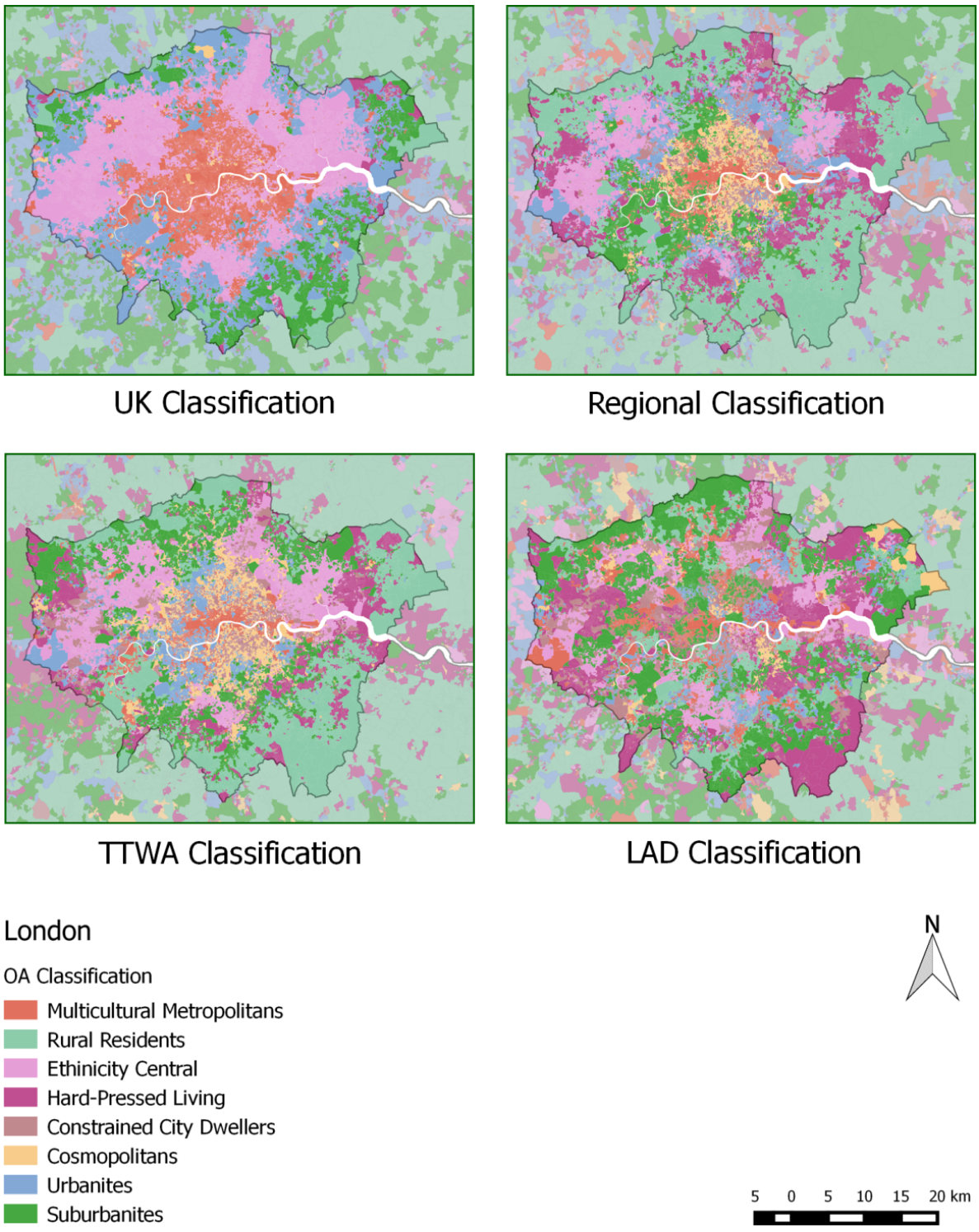

Figure 6.15 Classification comparisons through mapping typologies for each contextual geography used in the adjustment of attributes, London Region.

Another interesting remark is that, unlike Liverpool, suburban areas start to appear closer to the city centre as geographic context becomes smaller. A probable explanation is that in the regional or national context suburban communities, i.e. characterized by middle-aged families in detached houses, higher education levels and higher car ratios are not likely to be assigned near city centres. Attributes in these OAs are not distinctive enough to be considered suburban compared to the average values of areas at city fringes, suburban towns or countryside settlements, so despite these areas having a more suburban nature, they are still classified as *Urbanites* or *Cosmopolitans*. In the local level however these distinctions become more obvious and some central areas can be classified as suburban.

Finally, a similar analysis was carried out for the Greater London Area, where the mixture of populations is much more complex. Figure 6.15 shows the differentiation of cluster membership between UK, Regional, TTWA and LAD clustering geographies. The UK classification shows first-hand how little variation national classification systems ascribe to large metropolitan areas.

At the Regional and TTWA level cluster outcomes are significantly more diverse, showing some level of concentric zone topology within neighbourhood typology; a metropolitan core surrounded by cosmopolitan neighbourhoods, followed by urbanites and ethnic neighbourhoods which expand at the West and North-East. Hard-pressed households are located at the fringes of the city (possibly due to the high cost of housing) while most of the suburban population is located at the outer South-West of the city. Since the London Region only has two TTWAs (north and south of river Thames), the two maps are very similar. At the LAD level however outcomes are much more chaotic. The extents of LADs in the London Region are simply too small for them to behave as contextual areas, and as such the classification produces an amalgam of cluster types.

In conclusion, there is a prevailing trend regarding the transition of areal classes as the analysis progresses from smaller to higher geographic scales. One transition seems to take place between *Rural Residents* and *Suburbanites* (and in some cases further down to *Urbanites*), another transition between *Hard-Pressed Households* and *Constrained City Dwellers*, while more complex exchanges take place among the *Multicultural Metropolitans*, *Cosmopolitans* and to some extent, *Ethnicity Central*.

6.6. Evaluation outcomes

From the exploration results it is clear that there is considerable divergence from national socio-spatial patterns for many regions across the UK. This affirms the hypothesis that national

classification performance is not constant across the UK, and while some areas could be adequately represented by geodemographic systems such as the OAC, there are areas that perform very poorly. Initial exploration of attribute means at the various levels of geography generally shows the same underlying pattern, although average similarity scores drop significantly when classification scale becomes larger. This is not surprising, given that the larger the scale, the more "localized" and unique the socio-spatial pattern will be.

More specifically, a number of conclusions can be made regarding the regional / local performance of national geodemographic systems:

- There are significant disparities between local and national classifications at the Super-Group level and when the number of clusters remains constant. The higher the scale of what is considered local, the more the disparities increase.
- The disparities are evident both in relation to cluster attribute means and to cluster assignment. Although the majority of clusters retain their nature (e.g. rural, suburban, etc.), there are significant changes in the individual attributes of the clusters (e.g. rural socio-spatial patterns in the North West Region differ compared to those in Wales). Furthermore, small changes in attribute means induce considerable changes in neighbourhood classification, producing diverging local socio-spatial patterns.
- The intensity of disparities appears to exhibit distinctive spatial patterns. In general, urban areas and city cores demonstrate more disparities, but overall intensity at a local level is dependent on a case by case basis. Moreover, some neighbourhood typologies such as deprived or multicultural neighbourhoods seem to be impacted more by local conditions than others.

Results raise a series of important policy issues regarding the use of geodemographic classifications as a guidance tool in policy making. Arguably, one of the principal purposes of national geodemographic classifications is to be used by local authorities since most of them lack the necessary resources to create them. Exploration results showed that, excluding several large conurbations, middle-sized urban areas perform better, while smaller Local Authorities and rural towns score consistently low. In this framework, the areas that national classification seems to perform worse are the same areas that would benefit the most out of an open geodemographic system, considering they are more likely to lack resources and expertise to carry out classification at their local level. Furthermore, economically lacking and remote areas are prospective targets of national socio-economic policies, e.g. subsidizing economic performance or social upscaling. Such discrepancies are seriously undermining the usefulness of national classification systems;

spatial identification might actually be misleading in regions where it is needed the most.

The disparities between national and regional classifications can be traced on how individual attribute distributions impact cluster formation and has not been addressed adequately within geodemographic research. A critical assumption that can be made is that the amount of information that can be retrieved from an attribute value at a particular area is dependent on the area's locality. This relates to how absolute attribute values represent the nature of the neighbourhood, compared relative values within a contextual geography.

It is reasonable to believe that a geographically sensitive geodemographic model can indeed be useful in cases where geographic context plays an important role in the analysis, for instance it can be used by Local Authorities to identify neighbourhoods in need of upscaling or evaluate the impact of Area Initiatives within national policy frameworks. A national classification cannot be based on an amalgam of local classifications since technically every local classification produces unique clusters. A national classification however for which observations reflect local socio-economic conditions may provide a good basis for a geographically sensitive model. The theoretical framework suggests that standardizing attributes regionally will allow geodemographic classifications to operate to some extent on the regional means and standard deviations of attributes, thus reflecting a form of spatial dependence of values that can be further carried on into the cluster analysis. The next Chapter presents an initial approach to such a model where a level of spatial dependency is introduced to the traditional geodemographic methodology.

Chapter 7. A Geographically Sensitive Geodemographic Model

7.1. Theoretical Framework

The second part of the analysis will be to present a geodemographic model which will take into account such contextual measures, demonstrate classification results and measure its effectiveness. As such, the main research question that this part of the analysis answers is:

1. *How can information about geographic context be most appropriately incorporated within geodemographic classifications?*

In order to incorporate geographic sensitivity to a geodemographic classification, the aim of the model would be the creation of a typology of neighbourhoods that reflects the nature of socio-economic and built environment conditions from a regional/local rather than national perspective. Contrary to previous attempts of incorporating spatial interaction within the clustering algorithm (Theiler and Gisler, 1997) or on an ex post facto basis (Feng and Flowerdew, 1998; See and Openshaw, 2001), the method proposed here is built upon the data preparation step of geodemographic analysis, similar to Adnan et al. (2012).

The rationale is that adjusting values to reflect the distribution of attributes within a locality directly affects the attribute distances between areas, which will therefore make proximal zones appear more similar. This way, a form of spatial dependency is incorporated in the clustering process, as the attribute values of a neighbourhood are now dependent on those attributes around it. For instance, two OAs within the same geographic context will tend to be more similar that those further apart.

The approach follows the methodological response to the local classification critique provided by Webber (1980). If local classifications are incomparable to national classifications because they operate on different means and standard deviations, then there must be some way to mitigate these effects by adjusting data points. Arguably, it would be safe to suggest that any adjustment on the attribute values based on near-geography would involve the mean value of the locality, the dispersion (standard deviation), the range of values (min – max) or a combination of those. Since value standardization does exactly that (although for a different purpose), values can be simply adjusted using one of the several standardization techniques available, and achieve both spatial dependency between observations and commensurability between attributes.

The method proposed here is relatively simple and straightforward. The main difference however is that although values at a conventional national level classification are standardized based on the national scale, by standardisation at a regional scale attribute values are adjusted based on regional attributes. If, for instance, a locality has attribute values that compared to the national indexes seem very consistent, by standardizing at a regional scale values could potentially "spread out" and interesting local patterns could emerge.

This would insert a certain level of spatial dependency between areas within the same geographic context. Areas with low attribute values surrounded by areas with similarly low values would be transformed to average, without however applying a "smoothing" effect like the classification provided by Adnan et al. (2012). It would also affect the entire classification distribution and not the cluster membership of a particular area relative to its neighbours, as the case with fuzzy clustering or contiguity enhanced *K*-means (see Chapter 3, section 4). The trade-off would be that classification outcomes would be closely dependent on the delineation of the geographic context within which attributes are adjusted, as illustrated in the evaluation presented in Chapter 6.

The premise of this approach is simple; many socio-economic conditions are not included in classification, so by using relative attribute values there is more control for unobserved variables that would otherwise be assumed to be homogenous. For instance, cost of living or housing prices certainly differ quite radically from place to place, even within the same country. By not including such attributes there is, for example, an erroneous assumption that type of housing reflects socio-economic conditions. Certainly, owning a detached house in Oxford would tell the classification creator a lot about the socio-economic status of the residents, but would detached housing reveal the same information about residents in Thurso in Northern Scotland? Since different areas have different underlying conditions, and assuming that cluster formation and labelling is carried out relative to what is considered "low" or "high" within attribute scores, relative attribute values can give a much better representation of neighbourhood types within a locality.

As such, geographic standardization must take into account the differences in mean and standard deviations of the region. To illustrate this differentiation, Figure 7.1 shows the variation in average values between contexts for variable *K30: Percentage of households who live in a flat*, as reported for the region of Cambridge. This density plot, superimposed as a percentage of OAs that belong to each kernel, also shows the mean values between geographical contexts (dashed lines). For instance, the Cambridge LAD has a much higher average than the rest of the UK, although the Region of East of England as well as the TTWA it belongs to have a much lower value than the national average. As seen by the distributions, the UK lies in between these geographies

mainly because of a large concentration of areas that are comprised by over 90% flat dwellers in some other parts of the UK. It is clear by this figure that the selection of the contextual geography will impact results significantly.

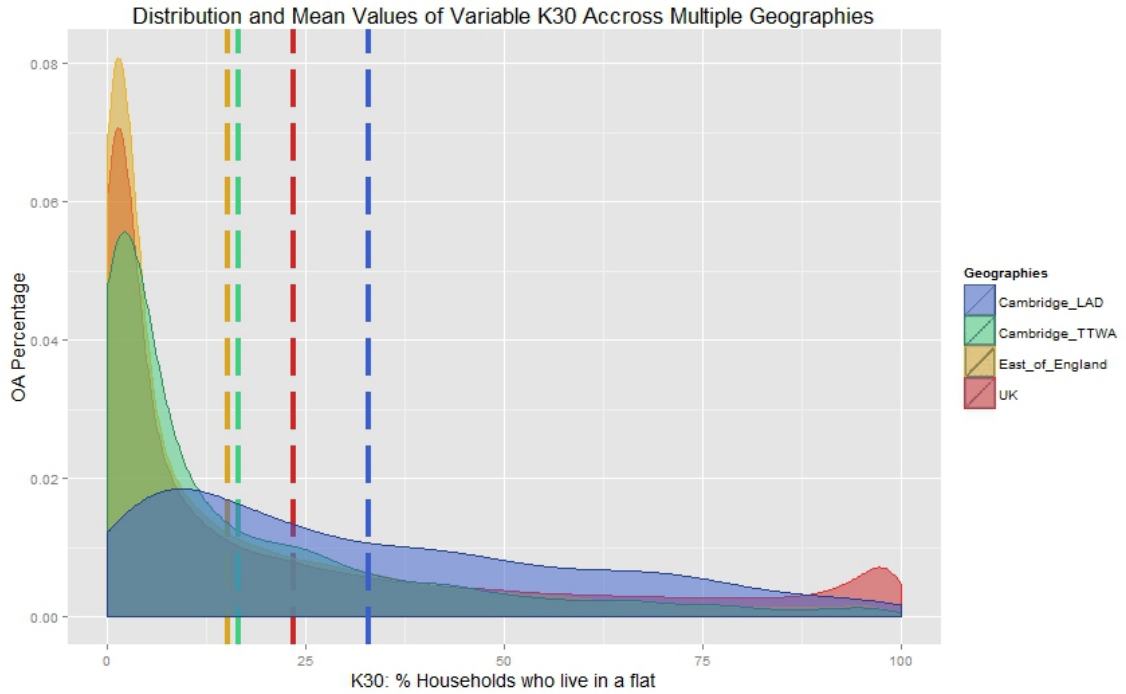

Figure 7.1 Standardized density plots and variation in average values between geographical contexts for variable K30: Percentage of households who live in a flat (Data source: Census 2011).

In this instance, *z*-score standardization has been used, which takes into account the mean and standard deviation of the population. The selection of *z*-score standardization is based on the same reasons outlined in the classification methodology in Chapter 6, Section 4. With regards to local geographic contexts, range standardization only affects the width of the base of the distribution, as illustrated by Figure 7.2. Values of variable *K04: Percentage of Persons aged 45-64* are extracted at the UK, Regional, TTWA and LAD level and range standardized individually creating 4 independent datasets. Then the OAs for the LADs of Cambridge and Glasgow are extracted for every dataset respectively and their standardized densities are plotted.

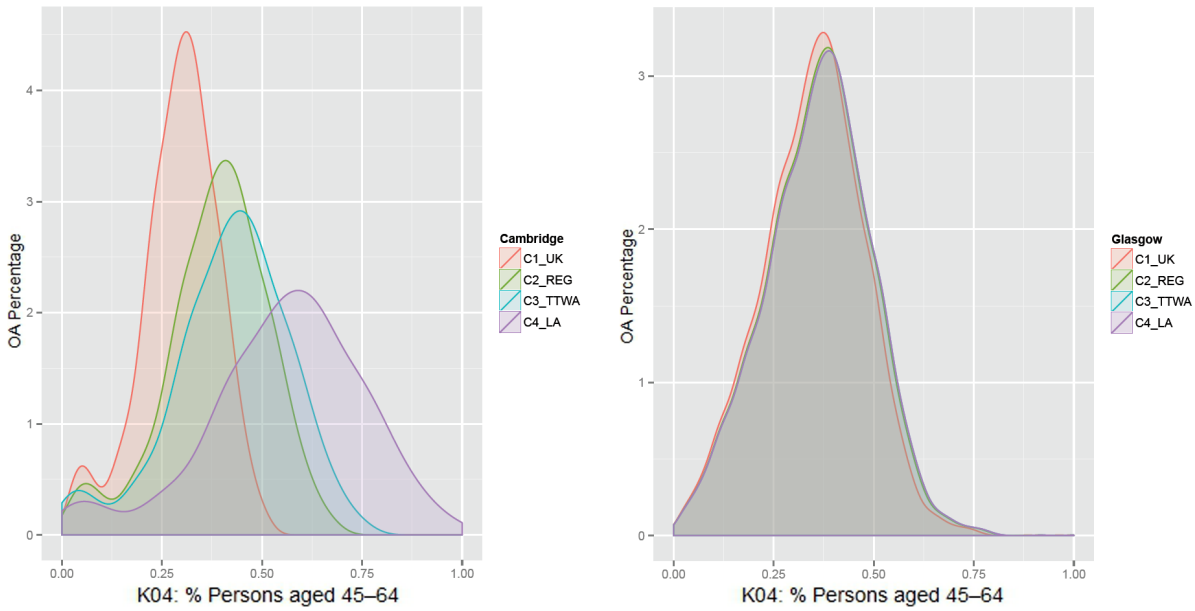

Figure 7.2 Range standardization of the LAD of Cambridge and Glasgow respectively (Data source: Census 2011).

In general, LADs with very few OAs such as Cambridge could perform well with range standardization (which is one of the best case scenarios), while larger urban areas and conurbations such as the City of Glasgow typically have the same attribute extents with other higher geographies (i.e. there is a high possibility that the maximum value is found within them). In some cases, if the local maximum lies within the contextual area under consideration, range standardization will not produce any level of attribute value differentiation between geographies. This property renders range standardization ineffective for our model.

A standardization technique such as the *z*-scores on the other hand is based on subtracting every attribute value from the mean and dividing by the standard deviation and it is much better suited for our analytical framework. The following example illustrates the results of a contextual *z*-score standardization using the LAD of Liverpool.

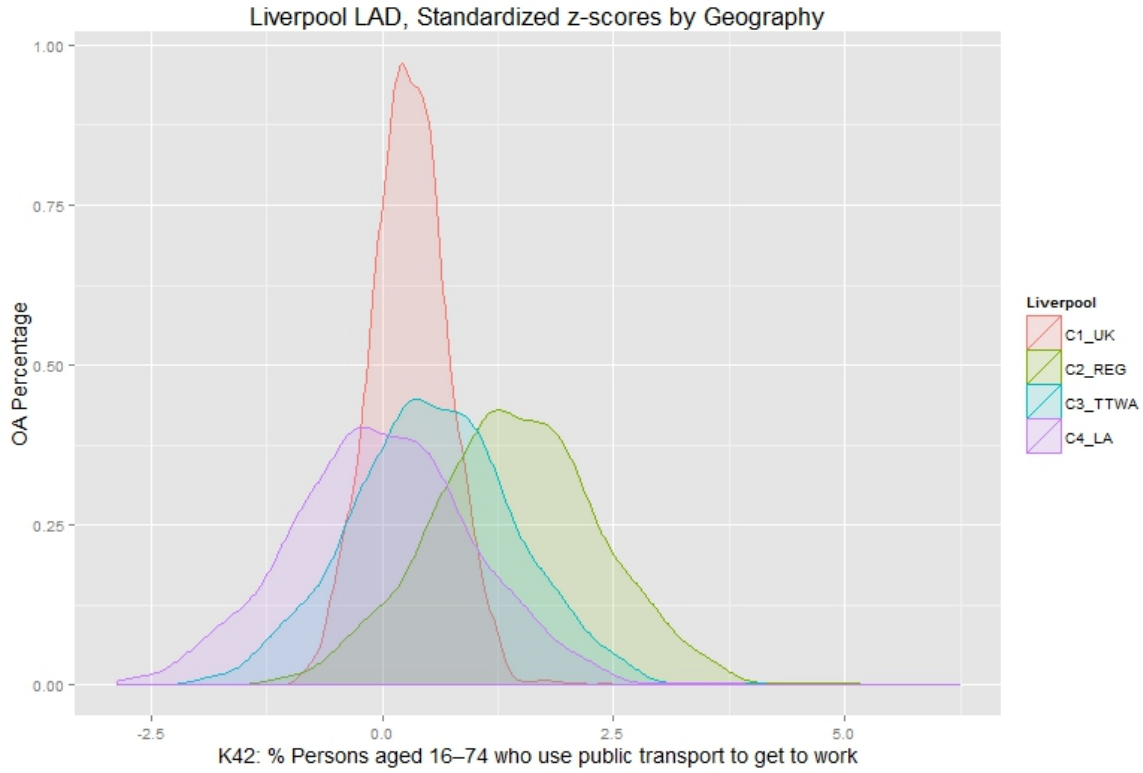

Figure 7.3 The distribution of attribute values of variable K42 for the 1584 OAs of Liverpool LAD, as standardized using *z*-scores at different geographic contexts (Data source: Census 2011).

Figure 7.3 shows the distribution of attribute values of variable *K42: Percentage of persons aged 16–74 who use public transport to get to work* for the LAD of Liverpool, as standardised using *z*-scores for different geographies. The variable is standardized per geography and then the 1584 OAs for the area of Liverpool are extracted and plotted, similarly to Figure 7.2. The red density plot shows attribute values of Liverpool relative to the UK context, the green relative to the North West Region and the blue and violet relative to the TTWA and LAD respectively. The plots show that compared to the national attribute values, Liverpool is slightly more prone to use public transport to go to work. However, in the larger TTWA area surrounding Liverpool and within the Region this propensity is even higher, meaning that in a regional context Liverpool OAs score much higher than the rest of the region.

This approach will affect the similarity between OAs based on their localities. For instance, Output Area *E00032990*, has a K42 value of 0.07 in the national context, and thus is considered average. In the Regional and TTWA context, this value is 0.95 and 0.89, much higher than the national average. At the Local Authority level however, the value is -0.53, significantly lower than

the Liverpool average. If we consider another OA area in Chester, OA *E00092350*, this also appears to have the same average value in the UK context (0.03), although in the regional, TTWA and Local context this value raises significantly to 0.84, 2.60 and 2.86 respectively. Although both areas appear very similar under a national lens, these areas are far from similar in their local contexts (-0.53 and 2.86). If we were to adjust values based on regional contexts, both areas would look similar (0.95 and 0.84) since they belong to the same region.

The above example shows how much contextual geographies could affect attribute values. The adjusted values can be described as *relative attribute values,* meaning that values reflect conditions relatively to a contextual geography. In a model where relative attribute values are used, an area exhibiting low or high values is subjective to the geographic framework that values are contrasted against.

In this framework, different contextual geographical can provide different relative attribute scores when standardized accordingly. An OA that can be considered an outlier in the national context could be very much closer to the average values in a local context and vice-versa. If values are standardized on a contextual geography basis, they will reflect the local conditions of the regions selected to serve as geographic contexts. This will enable the classification to group an area with regard to their *relative* attribute values. In doing so, there is a notion of control over conditions that are not necessarily taken into account in the classification process. For instance, a family living in London has a much lower propensity to have a car than the rest of the country, even if their other socio-economic conditions would suggest it, because of restrictions and limitations that are based on location.

The method of choice regarding variable adjustment is therefore geographic standardization. This approach offers the advantage of simplicity and model integrity. This modification takes place in the data preparation step of cluster analysis, so any further parameterization of the model such as variable weighting, clustering algorithm, etc. can be applied without adjustments to already established models, and any further changes in the classification will only be limited to the clustering process. A critical assumption made regarding the assessment of the resulting classification system is the premise that a "good" local performance is a necessary and sufficient condition for a "good" national classification. In so doing no claim is made that this type of methodology is better than the conventional, only that one the proposed type of classification would be more sensitive to local data patterns, and as such would perform sufficiently well on local policy applications.

7.2. Model Description

For simplicity, the proposed model will be applied to the same dataset and using the same classification methodology outlined in Chapter 6. Since the proposed geodemographic model only adjusts attribute values based on the local distribution of a contextual geography, the final input dataset is identical to the combination of all local input datasets. The adjustment takes place within the data preparation step of the geodemographic analysis; methodologically, the complete UK dataset will be standardized based on the number of geographies outlined in the previous Chapter, i.e. at the Regional, TTWA, and LAD level. In doing so, three different national classifications will be produced.

A critical aspect of the model will be the ability to adjust the level of spatial dependency of data points depending on purpose, similarly to Theiler and Gisler (1997). The method adopts this approach with a generalised model that can incorporate any degree of spatial dependency within the geographic context that can be pre-specified by the user. In this sense, the classification creators can adjust the level of impact of the contextual geography to an area's attribute values as they see fit. Specifically, the proposed method replaces data scaling with a new function that can adjust attribute values based the attribute distribution of the area's contextual geography and a geographic impact factor g:

$$z_{i,\alpha} = (1-g)\frac{x_{i,a}-\mu_N}{\sigma_{\mathrm{N}}} + g\frac{x_{i,a}-\mu_c}{\sigma_c}\,, \qquad g=[0,1] \qquad (7.1)$$

where $x_{i,a}$ is the attribute value *i* of area a, μ_N and σ_N is the average value and standard deviation of the attribute for the whole dataset, μ_c and σ_c is the average value and standard deviation of the attribute for the contextual area and g is a parameter that takes values between 0 and 1 (see *Appendix I: E. g-Factor Attribute Adjustment*).

The geographic parameter, denoted here as the g factor, effectively adjusts the level of impact of contextual geography to attribute values. For $g=1$, the first element of equation (7.1) becomes zero and the attribute values are scaled based on the geography of the locality, similarly to the exploration results. For $g=0$, the second element of the equation becomes 0 so local geography is not incorporated at all in the model, and the attribute values are effectively calculated as z-scores. Any value between 0 and 1 will adjust how much impact local geography has on the attribute scores and by extent on the final classification.

The method provided here is very simple and straightforward, yet it can provide a good basis for geodemographic models that need to incorporate some level of spatial dependence in a

national classification. It does not rely on arbitrary buffer zones or complex weighting schemes within the cluster analysis, although it is dependent on the contextual geography selected. It is also worth noting that the basic principle of the geographic parameter can be applied in parallel with any standardization method, by applying some modifications to equation (7.1).

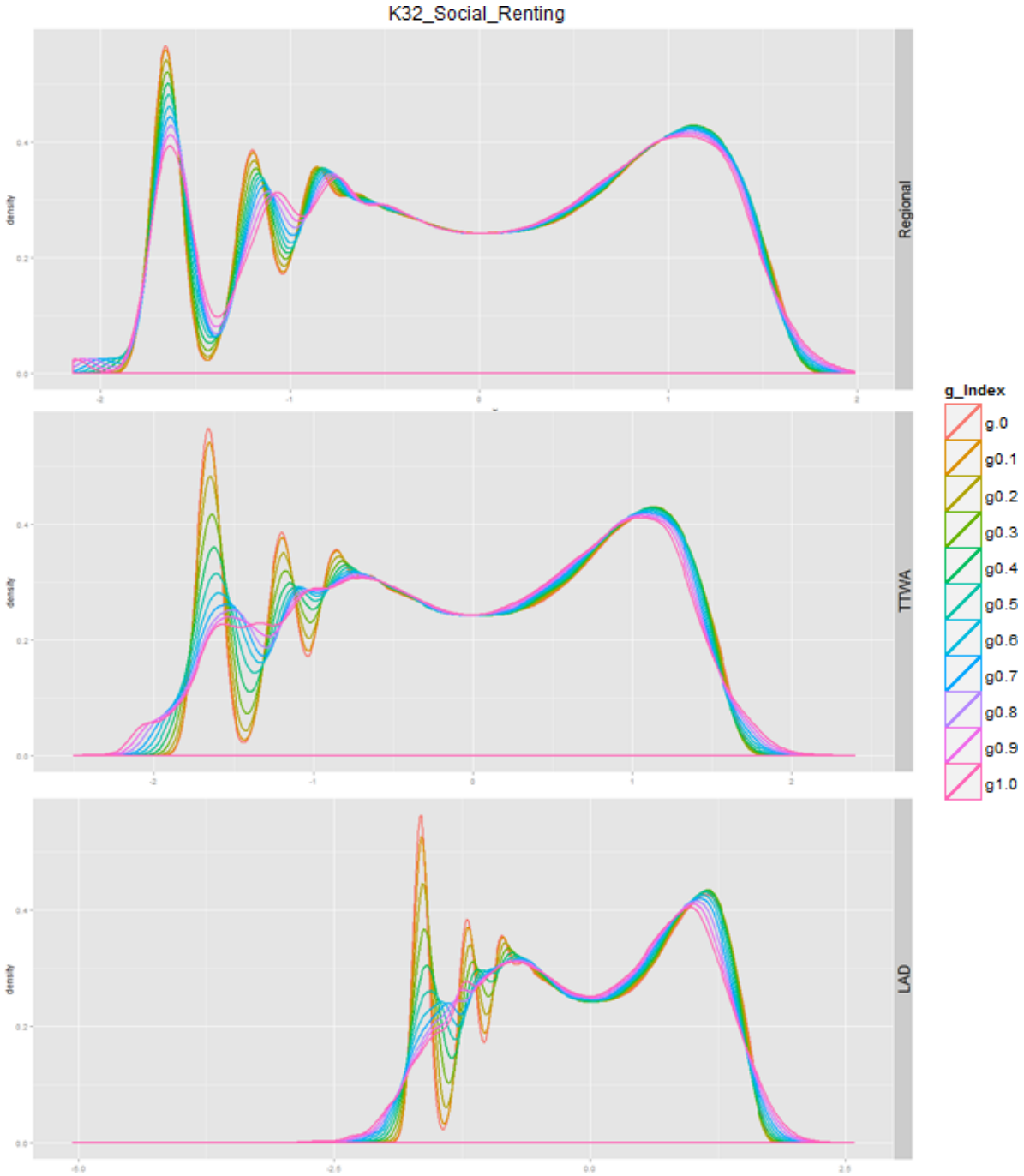

Figure 7.4 Impact of the g factor on value distribution of attribute *K32: Percentage of households who are social renting* for each spatial context.

Figure 7.4 demonstrates the transformation of attribute values for g = 0 to g = 1 at 0.1 increments, as seen for variable *K32: Percentage of households who are social renting*. Between all three geographic contexts, there is a general observation that higher levels of g tend to normalise attribute values, particularly local minima and maxima in distributions. This is reasonable, in the sense that at higher scales neighbourhoods become more similar, hence attributes values approach normality. Nevertheless, this effect seems to apply to only some of the variance; in this instance the impact of local geography seems to affect only the left side of the distribution. The higher values on the right side remain relatively constant.

This behaviour is related to the nature of the attribute and to what spatial variation patterns exist at the local level (obviously, social housing is to some extent related to areal deprivation which is very spatially distinguishable in the UK). In this framework, the data points incorporate some level of spatial dependence that can be carried on to the cluster analysis. Another observation that can be made is the spread of the distribution. Higher geographies seem to affect the standard deviation of values. Particularly in the LAD context, values demonstrate considerably higher variance, which could potentially indicate that LAD contextual geographies could produce significant outliers that may skew classification results.

Without a single optimization function in mind however, it is difficult to pinpoint the "best" contextual geography. How geographically sensitive attribute values must be is heavily dependent on the classification purpose and what theory dictates in the analysis of a particular social phenomenon. Ultimately, the scope of this research is to provide a methodological framework for geographically sensitive geodemographic models, and not provide a new or "better" geodemographic system per se. For the remainder of this section and for simplicity, analysis will focus on results demonstrated for the Regional contextual geography, which seems to provide the best basis for general-purpose classifications.

7.3. Model Evaluation

A first step to evaluating model outcomes is visualizing the classification and comparing it to known models. Figure 7.5 shows the transition of socio-spatial patterns for g values between 0.1, 0.25, 0.5, 0.75, 1.0 and the baseline model, the 2011 OAC in the Liverpool area. The maps were plotted taking into account Regional contextual standardization, so typology variation is more subtle. Upon closer inspection, the processes that take place in regards to class transition are very similar to those observed between regional and national classifications in the exploration section

of the analysis. The more visible transition is taking place between the *Hard-pressed Living* and *Constrained City Dwellers*, while the *Urbanites* Super-Group seems to move from a buffer zone between suburban population and more deprived neighbourhoods to a buffer zone around the city centre.

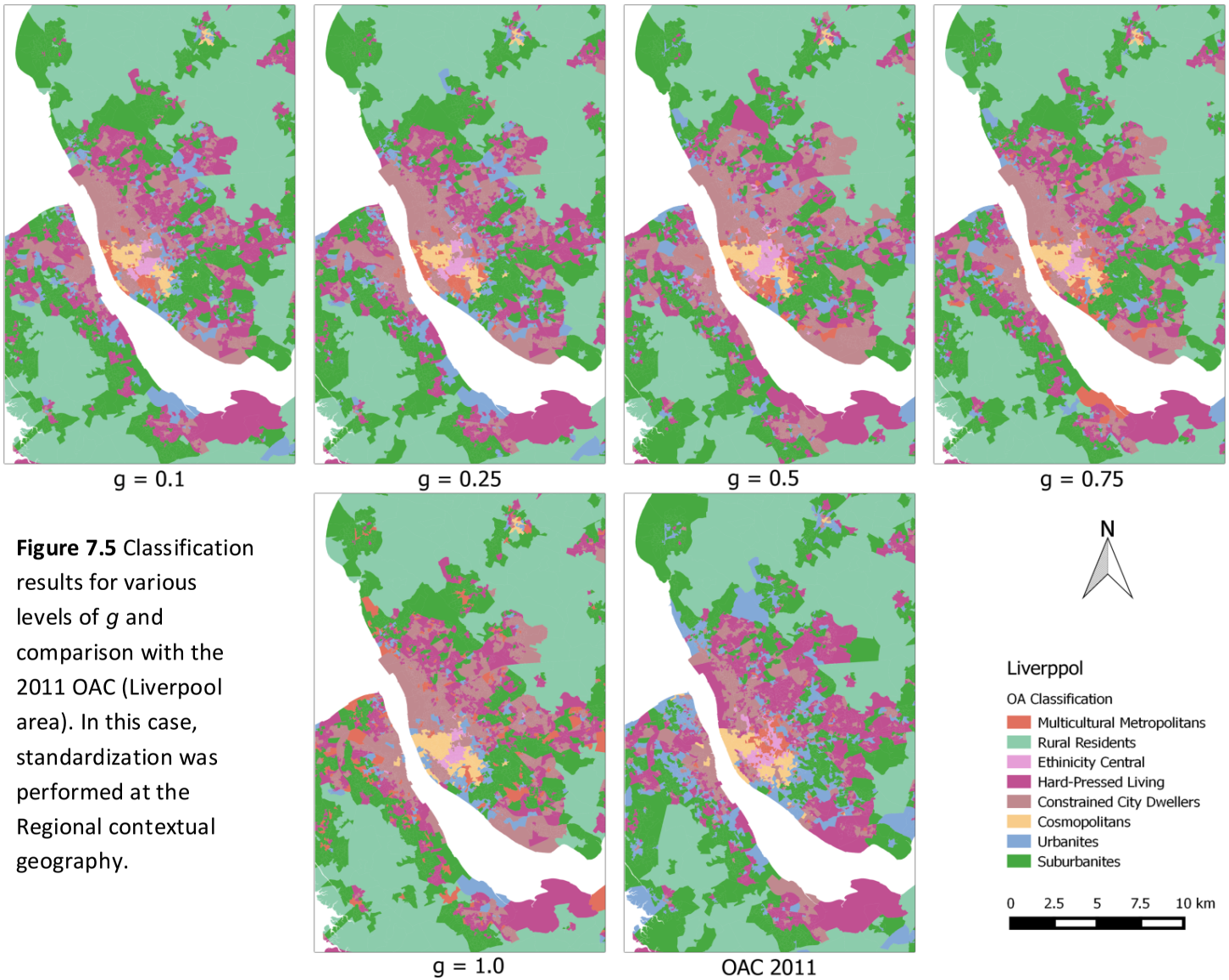

Figure 7.5 Classification results for various levels of *g* and comparison with the 2011 OAC (Liverpool area). In this case, standardization was performed at the Regional contextual geography.

This measure can be calculated by labelling clusters and measuring the degree of class consistency for different values of g, described as the ratio of OAs that did not change class between classifications. For simplicity, only the differences between $g = 0$ and $g = 1$ will be analysed. Figure 7.6 shows the OAs that have changed cluster membership between the two classifications in the area of Greater Manchester, superimposed over the baseline classification.

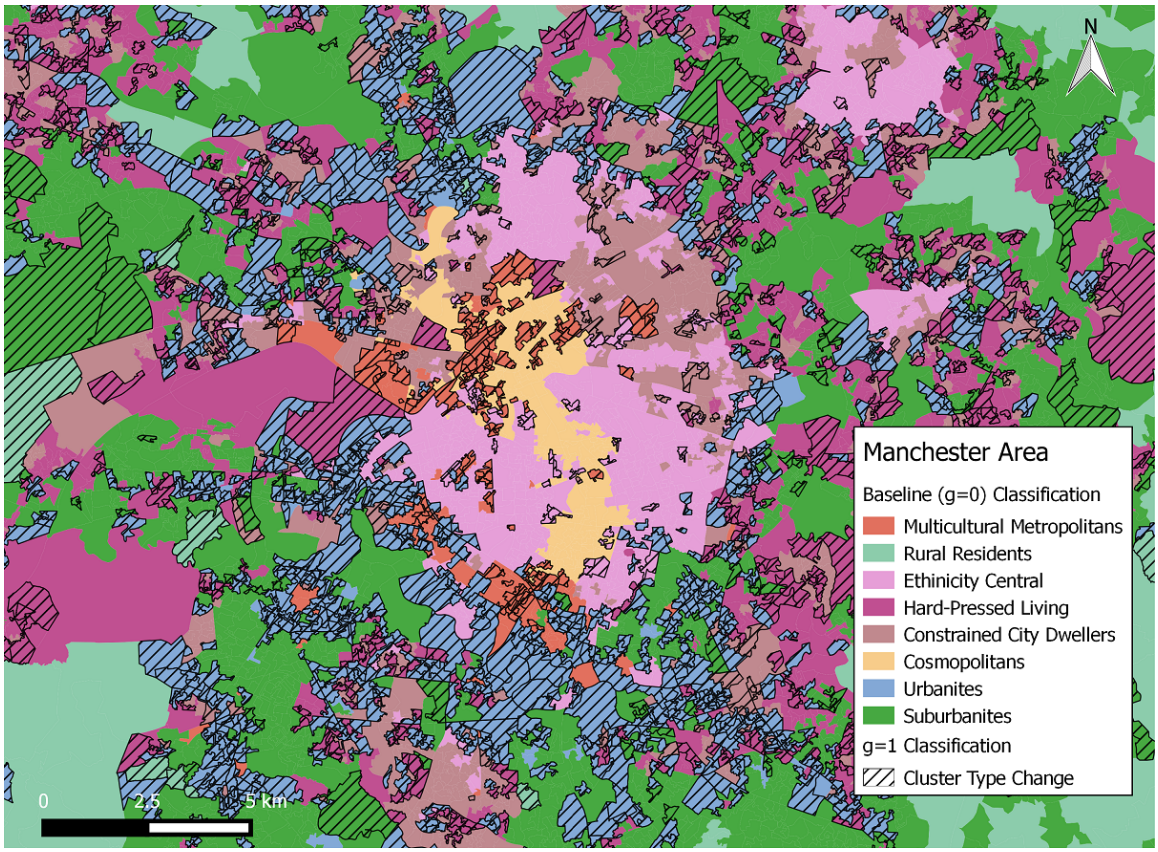

Figure 7.6 Changes in cluster membership at the Output Area level between a baseline classification and a fully geographically sensitive classification for Greater Manchester.

From the above map there is a clear indication that the classes *Urbanites* and *Multicultural Metropolitans* seem to be the most volatile in terms of cluster consistency. This suggests that some particular typologies produced by the baseline model (which is almost identical to the 2011 OAC) are very sensitive to local variation of attribute values, while other classes (e.g. *Cosmopolitans* and *Suburbanites*) are generally adequately represented at the national level. Indeed, a closer analysis of the within-cluster ratio of class transitions shows varying results (Fig. 7.7). In general, the *Urbanites* class seems to be the most sensitive to regional standardization by a huge margin, followed by *Ethnicity Central* and *Multicultural Metropolitans.* Results are

consistent with the exploration outcomes as some typologies being more spatial dependant than others.

Sequence analysis could be used to identify the transition course of classes at small intervals of g and obtain some more insight on why some specific socio-economic patterns are more spatially dependant than others or whether the original clusters were too "dispersed" originally. It is important to note that although the label of the cluster remains the same (based on the dissimilarity measure), the nature of the cluster, i.e. what it socio-economic conditions it represents in terms of attribute means, could be very much different. This issue makes any robust generalisation complicated.

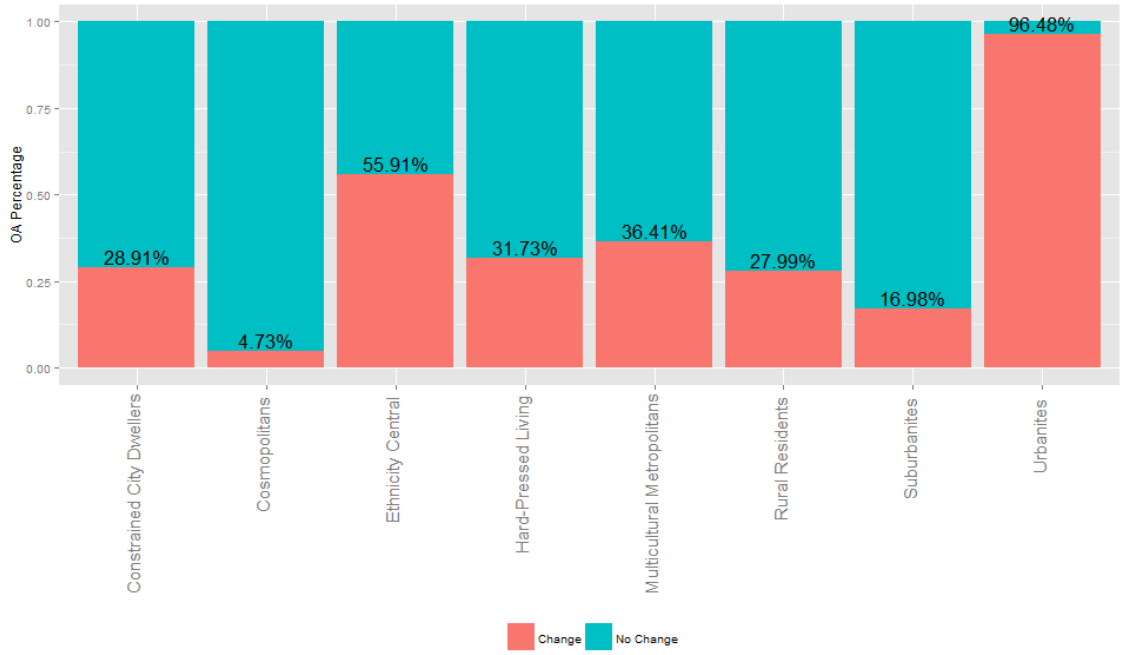

Figure 7.7 Class transitions between the baseline classification (g = 0, no geographic sensitivity) and a fully geographically sensitive classification (g = 1). The plot shows the percentage of OAs that have changed class from the baseline classification by cluster type.

It would also be useful to assess the impact of the g factor in clustering performance, specifically in terms of cluster compactness (as presented in Chapter 3, Section 3). For this purpose, an evaluation of classification results will be performed for various levels of g, and for the selected set of geographic contexts, Regional, TTWAs and LADs. Unfortunately, external evaluation criteria cannot be used to test the validity of the model, as there is no information on a known cluster structure of the data set. The other set of criteria is the internal criteria, which

measure the validity of the cluster structure. A variety of methods have been proposed to evaluate internal cluster quality using graphical, bootstrapping or data point distance methods (Everitt et al., 2011).

One of the most common internal criterions used in cluster analysis is evaluating cluster structure in terms of within-cluster cohesion and between-cluster isolation. The clustering error in *K*-means (SSE) is calculated as the squared distance between all data points and the global sample mean, SSD. A good clustering approach should minimize the total within-cluster sum of squares compared to the total. This internal criterion, IC, can be written as a ratio of:

$$IC = \frac{\sum_{\mathrm{i}}^{\mathrm{N}} SS - \sum_{\mathrm{k}}^{\mathrm{K}} WCSS}{\sum_{\mathrm{i}}^{\mathrm{N}} SS} = \frac{BSS}{TSS} \qquad (7.2)$$

where the sum of SS is the total sum of squared distance, TSS, from the global mean, and WCSS is the within-cluster sum of squared distances from the k cluster mean. Since the between-cluster sum of squares, BSS, is equal to the total sum of squares minus the within-cluster sum of squares, the IC ratio can be defined as BSS/TSS. The IC index takes values between 0 and 1 and is usually presented as a percentage. Values closer to 1 indicate a good fit, i.e. clustered data points are very close to their assigned mean, which decreases the WCSS and offers a good clustering solution.

In order to provide a closer inspection of the effects of the g value in classification performance, a number of *K*-means clustering application were carried out for varying levels of g at 0.1 increments. Evaluation results are based on the IC index, which was plotted for every g value for each of the 3 geographic contexts (Fig. 7.8).

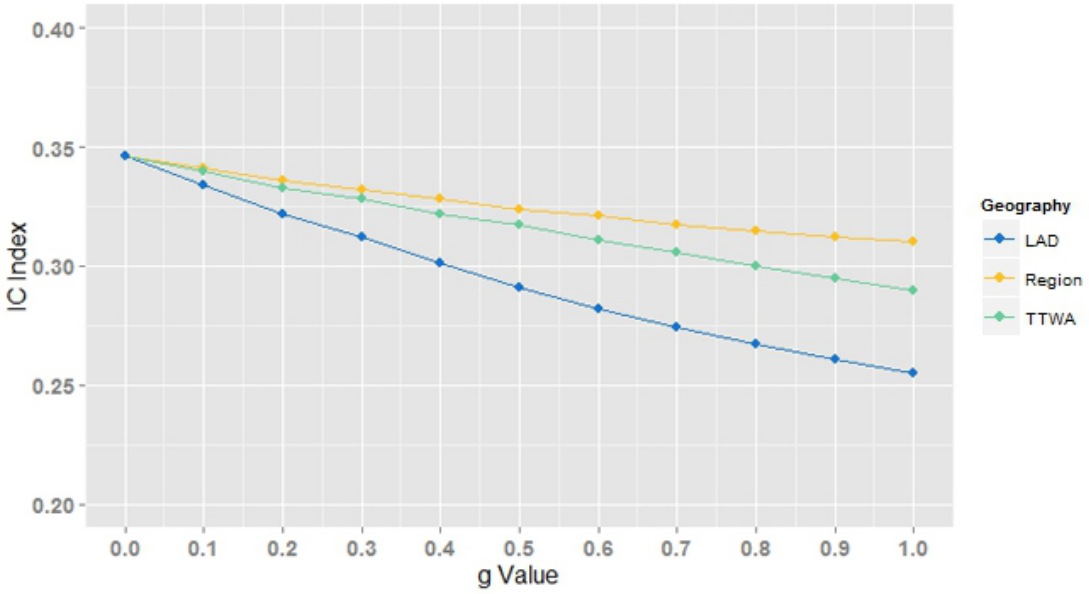

Figure 7.8 IC index (BCSS/TSS) score by g factor per contextual geography.

Results show a linear relationship between IC and geographic factor g. In a strictly clustering evaluation scheme, the Regional context appears to explain more variance than the TTWA and LAD counterparts. The total differences between none and complete geographic dependency is 3.6%, 5.6% and 8.9% for the Regional, TTWA and LAD contexts, which is directly comparable to the 4.3% difference between the Tyne and Wear and National Classification reported by Openshaw et al. (1980). Compared to the baseline model, this classification scores significantly lower in terms of variance explained. The 2011 OAC with range standardization of values scores 39.1%, while the UK classification without any spatial dependency and *z*-score standardization scores 34.6%. However, results from the IC index are not directly comparable, as explained below.

Measures of internal cohesion give a lot of insight on how well the clusters centres represent the data, and how much variability is "lost" as a result of clustering (e.g. in this case between 65% and 85%). The internal cohesion criterion also gives a lot of insight on the amount of clusters that should be produced, as a few more clusters reduce the variability loss significantly. Secondly, these methods are highly dependent on the nature of data. Cluster evaluation is dependent on the Euclidean distances of data points from the cluster means. The evaluation is thus related to the attribute values of the input dataset. In this case, since attributes are standardized in a different way within every geographic model, comparisons are nonsensical.

In general, the more "spread out" the distributions of attributes, the less cohesive clusters will be, so when certain standardization techniques and transformations are used, the classifications are bound to score higher in such indexes. Clusters tend to appear to be more homogenous in the

case of range standardization when few extreme outliers exist in the dataset. If range standardization confines values in a small attribute space, some attributes may not have the desired impact in the classification. In certain cases, this could give the impression of a better clustering solution. In this framework, it is best to be used to evaluate "good" or "bad" clustering solutions in the strict sense only when assessing the performance of clustering algorithms (Milligan, 1996).

It is important to note as well that the model presented here is based on a theoretical framework and built for a specific purpose, and as such undermines comparison with other geodemographic models. The key idea is that neighbourhoods are clustered based on their relative attribute values and not their absolute ones. The internal evaluation criteria used only provides a good basis as to the extent of the variation between different levels of g and between different levels of geographical contexts. Regarding the latter, the slope of the IC plot indicates that as g becomes larger, attribute values are adjusted more radically within higher geographies.

From the plot, it appears that classification at the Local Authority or Travel to Work Areas simply lose too much information too quickly, which would provide in turn very poor clustering solutions. This relates to the level of diversity each area has when examined individually; their small geographical extents escalate the presence of outliers. Obviously, the inconsistency of their extents plays an important role in these shortcomings (for instance, most Local Authorities within the London Region). For the above reasons it seems that the Regional classification provides currently the most cohesive results. It is important to reiterate however that the IC index can hardly provide an accurate evaluation of the classification accuracy without any external evaluation.

7.4. Conclusions

This Chapter provides the theoretical rationale and methodology on how national classification systems can operate on local means and standard deviations, thus introducing some level of spatial dependency across regions. A key aspect of the extended model is the adjustment of attribute values to reflect relative (to the geographical context) values rather than absolute. Although a conventional standardization of *z*-scores is selected to adjust values, other adjustments can be used as well, such as Rank Standardization (see Table 4.4), depending on purpose. In this case, besides *z*-score standardization three levels of contextual geography were used to define localities: Regions, Travel-to-Work Areas and Local Authority Districts. Results

demonstrate the impact of spatial dependence on classification outcomes through the comparison of a baseline model (a conventional geodemographic model with no spatial dependencies) with the extended model. With the introduction of the g parameter, the level of dependency can be adjusted by the creator, providing flexibility on how strong the geodemographic classification will be in terms of geographic sensitivity.

In conjunction with Chapter 6, it seems that the contextual geography used to adjust attributes greatly affects the outcomes of the classification. While all three contexts were considered, some contexts may perform better in capturing the socio-economic conditions within a region. Ideally, these areas should represent some level of organization of actual communities. In this case, there are some issues regarding LADs and TTWA that skew results. In practice, the model performance is dependent on the size of contextual geographies used. Identified issues are not only a result of average area size but also area consistency. Some areas are too small to capture the range of socioeconomic conditions across neighbourhoods, so input geographies are not independent enough to be considered individually.

Specifically, certain LADs face certain shortcomings due to their very small size, particularly within large metropolitan conurbations such as London and Greater Manchester. TTWAs on the other hand are the only non-administrative geographic unit present in the analysis. Although they may look promising, they are very diverse in terms of geographic extents and population they capture. Some TTWAs are even smaller than LADs (e.g. rural or semi-rural areas), while areas such as Greater London only have two TTWAs (North and South of river Thames). This may produce some inconsistencies in neighbourhood typologies, particularly between suburban and rural classes.

Results from the Regional classifications seem to outperform the other two in terms of neighbourhood representation and cluster cohesion. However, the selection of geographic contexts primarily regards the scope and theoretical framework of the classification. Without any objective function to maximize and without any means of external evaluation relative to the classification targets, there is no one universal context that can maximise accuracy and no single geography can be considered best.

Secondly, one of the most interesting outcomes of the model evaluation is that classification differences are not uniformed spatially. The analysis showed that certain typologies such as *Urbanites, Ethnicity Central* and *Multicultural Metropolitans* are much more dependent on the contextual geography used, while others, such as *Cosmopolitans* and *Suburbanites* are more robust. This finding has certain implications in the use of national geodemographic systems.

Within the public sector, targeted policies may face substantial difficulties in identifying areas of a specific profile, as this will reflect global and not local socio-economic conditions. For instance, the definition of neighbourhoods with strong presence of ethnic diversity may be considerably different between London and Lancaster. The implications can be severe in the private sector as well, since inconsistent neighbourhood typologies may lose value as the primary indicators of income levels and/or product preference.

A model where attribute values are linked spatially can help mitigate such effects. By introducing a model that can have any level of near-geography incorporated, classifications can be freely customised to fit the needs of the creator. The limitation of the approach is mainly the selection of the extents of near-geography, i.e. the contextual geography used to standardise values, and the value of the g factor, which are both biased parameters and as such should reflect the theoretical rationale and purpose of the classification creator. Finally, it is important to note that not all geodemographic classification need to geographically sensitive. Sometimes by looking at the relative distribution of values regionally might also obscure phenomena that are important in the national scale, for instance when identifying characteristics of families below poverty levels or when addressing issues regarding social justice.

Chapter 8. Discussion

8.1. Thesis Outputs

This Thesis tries to elucidate some of the inner workings of Geodemographics. While geodemographic analysis can be viewed as an established methodology, the nature of the theoretical framework along with the lack of a single global optimization function produces a lot of uncertainty regarding whether geo-classifications provide a good representations of socio-spatial patterns. The analytical weaknesses, often coupled with a lack of any statistical clothing, often make it difficult to assess either the significance of apparent trends found in data or the importance of predictor variables that might explain those (Harris et al., 2007).

The Thesis tried to immerse the reader into Geodemoraphics by firstly providing a comprehensive review of the relevant literature, from the early geodemographic precursors to factorial ecologies and the emergence of geodemographic methods in the late 1970's. In order to understand the methodological framework, a review of current clustering methods and techniques serves to prepare the reader on the analytical steps carried out throughout this research. The Thesis proceeds to describe the analytical steps needed in creating a conventional geodemographic classification, from data gathering and preparation to the clustering methodology and cluster evaluation.

A detailed, practical example of building a bespoke geodemographic classification is presented in Chapter 5. This chapter not only provides detailed information on building a geodemographic classification but also tries to address some issues regarding the availability of data in the future and the decoupling of geodemographics from Census dependencies, as discussed in Chapter 4, Section 5. The bespoke classification for the built environment can be used in a variety of ways, such as increasing methodological robustness and classification performance, evaluating other classification, or analyse the correlations of the built environment with other social phenomena.

Chapter 6 and 7 have focused on the main topic of this research, the evaluation of national classification performance within regional contexts and the creation of an extension to the conventional methodology that account for spatial dependencies within the classification process. This research is not developed as a critique to Geodemographics, but rather tries to systematically review certain aspects of classification methodology, particularly the spatial dimension of geodemographic analysis; geodemographics are bound to be dependent on the

complex relations between spatial scale, spatial measurement and spatial variation, as in any geographic analysis (Atkinson and Tate, 2000). Thus far, there have been very few attempts to build a unified framework where the relative benefits of both spatial interaction and geodemographic approaches can be maximised (Debenham et al., 2003; Singleton et al., 2012). Moreover, a review of the relevant literature has shown that little has been done within geodemographic research as a response to issues of classification uncertainty and system-wide accuracy. Systematic evaluation of classification performance is limited, at least within the academia (Openshaw et al., 1980; Twigg et al., 2000; Voas and Williamson, 2001; Petersen et al., 2011; Reibel and Regelson, 2011). Evaluation constrains are further enhanced by the lack of classification transparency, that would otherwise enable replication and modification which are necessary in order to advance the field (Fisher and Tate, 2015).

The aim of this research was to determine how information about an area's spatial context can be captured and incorporated within geodemographic analysis. By including geographic context into neighbourhood typologies, spatial patterns that do not occur frequently enough in a nationwide analysis could be more distinctive. This enables a classification to be more sensitive to local variation of attributes despite it being carried out at national scales, while also addressing concerns about the inclination among scholars to perceive empirical studies that are larger in scale as more complete, compared to classification studies of individual cities (Reibel and Regelson, 2011). This quality is particularly useful in the context of national policy and public sector delivery, where national geodemographic classifications can be utilised by local authorities that lack resources to carry out their own classifications. Geographic sensitivity will allow policy applications to be more accurate, particular local policies and initiatives, and reduce the social and fiscal implications of mistargeting.

The scale-dependency of geodemographics is crucial in defining the representational and discriminatory power of neighbourhood classes, and has received some attention within academic research (Feng and Flowerdew, 1998; See and Openshaw, 2001; Reibel and Regelson, 2011; Harris and Feng, 2016). Hitherto, previous attempts to address the problem have been limited to accounting for spatial dependencies in the immediate area of a neighbourhood (Feng and Flowerdew, 1998), or use attribute transformation that incorporates such dependencies (Adnan et al. 2012).

This research adds and expands current attempts by providing a new analytical framework, accounting for the impact of proximal effects over wider regions. Such spatial contexts are defined as near- or contextual geographies. Their impacts are explored through a series of comparisons between regional and national classifications at the UK level, with an emphasis on the intensity

as well as the geography of the spatial variation.

Comparisons illustrate considerable divergence from national socio-spatial patterns. This affirms the hypothesis that national classification performance is not constant across the UK; some areas could be adequately represented by geodemographic systems such as the OAC, but there are areas that perform very poorly. While private sector users could experience fiscal implications, such as a reduced uptake of a product or service delivery, in public sector uses the consequences may be more severe, with mistargeting having potential implications on life chances, health and wellbeing.

The theoretical and methodological limitations presented raise a series of important policy issues regarding the usefulness of geodemographics as a guidance tool. As aforementioned, the value of national classification systems is still compelling, and one of the principal purposes of public national geodemographic systems is to be used by Local Authorities. The level of disagreement between national and regional classifications seems to be considerably higher for smaller Local Authorities and rural towns. Areas where the national classification performs worse are the same areas that would benefit the most out of an open geodemographic system, considering they are more likely to lack resources and expertise to carry out classification at their own local level. Furthermore, economically lacking and remote areas are prospective targets of national socio-economic policies, e.g. subsidizing economic performance or social upscaling. Such discrepancies are seriously undermining the usefulness of national classification systems; spatial identification might actually be misleading in regions where it is needed the most, such as policies regarding ethnic diversity or deprivation.

Results show variation on the level of disparity on a case by case basis. In particular, the higher the geographic scale, the more the disparities increase. The disparities are evident both in cluster attribute means and cluster membership of neighbourhoods. The intensity of disparities appears to follow specific spatial patterns at the UK level, with some areas scoring considerably lower regardless of the geographical context selected. Similarity levels were also found to vary depending on cluster type, as neighbourhood typologies such as deprived or multicultural neighbourhoods seem to be impacted more by local conditions than others. Overall, the exploration analysis showed that small changes in attribute means induce considerable changes in neighbourhood assignment, indicating that localized patterns are impacted more by attribute variance rather than average values.

The main contribution of this exploration is the introduction of a methodological extension to conventional geodemographic research that accounts for spatial proximity and context, designed

to assume better correlation between places and social identity. The central hypothesis of the model is that the amount of information that can be retrieved from an attribute value at a particular area is dependent on the area's geographic context. This relates to how absolute and relative attribute values are treated within a geodemographic analysis, and what kind of impact they have in the representation of an area. This model operates on local means and standard deviations of attributes, thus reflecting a form of spatial dependence of values that can be further carried on into the cluster analysis.

Evaluation of model results did not offer concrete information on one universal geographic context that can maximize accuracy, although there are certain shortcomings regarding very high geographic scales, such the Travel-to-Work Areas and Local Authority Districts. Results from the Regional classifications seem to outperform the other in terms of neighbourhood representation and cluster cohesion. However, without any means of external evaluation relative to the classification targets, no single geography can be considered best. The selection of geographic contexts primarily regards the scope of the classification. In this sense, the model should be treated more as a framework than an analysis tool.

The model evaluation also illustrated that the disparity within classes is not uniform either. The analysis showed that certain typologies such as *Urbanites, Ethnicity Central* and *Multicultural Metropolitans* are much more associated with spatial dependence, while others, such as *Cosmopolitans* and *Suburbanites* are more robust nationally. This finding has certain implications in the use of national geodemographic systems. Within the public sector, targeted policies may not or may not be effective enough depending on the target populations. It also highlights that the nature (and definition) of some typologies may not be spatially constant, as cluster centres "evolve" through classifications. The implications can be severe in the private sector as well, since inconsistent neighbourhood typologies may lose value as the primary indicators of income levels and/or product preference.

The introduction of the *g* factor was deemed necessary for that purpose, as the classification creator can adjust the level of impact of spatial context in accordance to the classification purpose. Another limitation of the approach is the selection of the extents of near-geography, i.e. the contextual geography used to standardise values, which is also a biased parameter. The evaluation showed to some extent that very small regions (such as LADs) do not respond well to the model extension.

8.2. Challenges and Limitations

As aforementioned, a key issue in any attempt to systematically document classification performance is the high susceptibility of Geodemographics to the operational decisions during the creation process (Openshaw and Gillard, 1978). This was one of the major challenges of this research. Without a single operational objective to optimize, identifying the best classification model through all variations is exponentially complex. It is next to impossible to consider all the combinations of all the methodological options in every step of the analysis, so the selection of appropriate decisions was based on theoretical grounds. Despite the efforts to provide as much evidence towards rationale, some of the operational decisions are carried out on the subjective level.

Furthermore, it also suggests that small modifications to the models presented may produce significantly diverse results. This presents an immense obstacle towards establishing relationships and generalizing results. Much of the classification parameters described in the MODUM classification, for instance, are very specific to the underlying data and methodology. The basis of the methodology is the introduction of proximal measures in relation to built environment features. Without systematic evaluation, it is difficult to assume what kind of differentiation the final model would present if, for example, adjacency effects were calculated not on a 100m distance measure but on 200m or 50m.

The same holds true when addressing the issues of similarity and classification scale. Results are biased to the similarity measures used in the exploration, as well as the selection of geographies that can be used as spatial contexts. The latter proved to be particularly difficult; the theoretical framework suggests that these areas must be based on the organization of actual communities. However, the nature of data availability combined with lack of any prior research on the extents of these areas posed significant limitations on the selection of appropriate contextual geographies. Considerations were made to include one more level of subregional geography, such as the NUTSII level. Still, their historically high volatility and the absence of consistent methodology disqualified them from any further analysis.

Another limitation is the assumption that the optimal number of clusters between classifications remains unchanged. Although actions were taken to mitigate these effects by including only the number of clusters present in the 2011 OAC classification, there is still no concrete evidence that the optimal amount of clusters should be carried on between scales. Deciding the number of *K* in cluster analysis is highly subjective, and while there are many

approaches to evaluate it, comparing hundreds of classifications without any baseline hypothesis on cluster amount would be nonsensical.

Since the adopted clustering methodology is top-down, analysis on subsequent clustering hierarchies beyond Super-Groups was not deemed necessary. If substantive differences occur between classifications at the Super-Group-level, then by default differences will occur to at least one sub-cluster, in this case at Group-level. However, measuring similarity at lower hierarchies would provide more accurate results on the variation of similarity level. For instance, different conclusions would be drawn from a local classification comparison if all sub-clusters were equally different, or if one sub-cluster was exceptionally different and the rest appeared the same. Of course, this would make the analysis unnecessary complex, as the amount of Group-level clusters would also vary significantly per locality. Furthermore, measuring similarity between Group-level clusters would be problematic; locally produced group-level clusters may look very similar to national group-level clusters of another Super-Group. This would produce erroneous results, taking into account that the methodology used is top-down. If however, this analysis were to explore comparisons within a bottom-up classification methodology, such as an agglomerative hierarchical clustering, then the analysis of lower hierarchies would be necessary.

Finally, another limitation that this research faced regards the extended geodemographic model and particularly the evaluation of the clustering performance itself. Similarly to cluster amount, evaluating the clustering performance based on internal criteria is dependent on the underlying parameters that have been used in cluster analysis, such as data preparation, transformations and standardizations. Typically, such techniques are reserved to evaluate clustering performance between clustering algorithms. In this case, external evaluation criteria would be much more meaningful but that would assume the existence of information on a known cluster structure of the data set.

8.3. Conclusions and Further Research

This research raises a series of important issues regarding the spatial behaviour of geodemographic classifications. Although results are inherently of tentative nature, they can provide a better understanding of the effects and extents of the problems. A key future research would combine built environment summary characteristics and spatial dependence into one consolidated geodemographic model. The model can be then equipped with a unified framework in order to produce a typology of neighbourhoods. Under a unified framework, the relative

benefits of both spatial dependencies and geodemographic approaches on new forms of data can be maximised. One way to accomplish the task of merging both concepts into one geodemographic system would be to include some of the built environment variables used as identifiers of neighbourhood morphology into the data inputs after socio-economic attributes have been adjusted regionally.

Before however building the unified model, there is still the challenge of the selection of the two biased parameters, geographic factor and geographic context, which are needed in order to produce relative attribute values. In general, any future development to the model would benefit greatly from any further work regarding classification parameters other than the ones used in this analysis. The parameters under investigation would not only include geographic contexts but also different standardization techniques, weighting and clustering algorithms.

The issue of geographic context can also be better addressed through a regionalization approach; as mentioned in Chapter 3 of this Thesis, regionalization methods are a special case of clustering methods, the aim of these methods is to produce analytical or functional regions. These regions can be formed from smaller areas (such as OAs) into homogenous yet continuous regions using some geographical criteria. The introduction of zone design (Openshaw, 1977) into the model may provide a way to produce consistent and non-overlapping geographic contexts within which the homogeneity of socio-spatial or built-environment patterns is maximized.

To conclude, the aim of this research is to provide a comprehensive account on the spatial behaviour of geodemographic classifications, while also trying to fill in some of the theoretical gaps in geodemographic research. As aforementioned, this research is not a critique to Geodemographics, but rather aims to contribute to the growing need for classification systems that are accurate and versatile enough to handle the abundance of big data that are currently available. This research is based on an exploratory analysis and on certain classification parameters and conditions; as such, it is important to stress that this research provides a new perspective of the analytical framework rather than a ready-to-use model.

The central premise of this methodology is the assumption that a "good" local performance is a necessary and sufficient condition for a "good" national classification. In so doing, no claim is made that this type of methodology is “better” than the conventional, only that the proposed type of classification would be more sensitive to local data patterns, and as such would perform sufficiently well under certain circumstances where sensitivity to local patterns is important.

References

Abler, R., Adams, J.S. and Gould, P. (1971). *Spatial organisation: the geographer's view of the world.* New Jersey: Prentice Hall.

Abu-Lughod, J.L. (1969). Testing the theory of social area analysis: the ecology of Cairo, Egypt. *American Sociological Review* 134: 198-212.

Adnan, M. and Singleton, A.D. and Longley, P. (2012). Spatially Weighted Geodemographics. *Proceedings of the 20th GIS Research UK*, April 11-13, Lancaster.

Aggarwal, C., Hinneburg, A. and Keim, D.A. (2001). On the surprising behaviour of distance metrics in high dimensional space. *Proceedings if the 8th International Conference on Database Theory*, London, UK, January 4-6, pp. 420-434.

Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD* 27: 94–105.

Aldenderfer, M.S. and Blashfield, R.K. (1984). *Cluster Analysis*. Beverly Hills, CA: Sage Press.

Alexiou, A, Singleton A.D. and Longley P.A. (2016). A Classification of Multidimensional Open Data of Urban Morphology. *Built Environment* 42(3): 463-476.

Alexiou, A. and Singleton, A.D. (2015a). Geodemographic Analysis. In Singleton A.D. and Brunsdon C. (Eds.), *Geo computation: a practical primer*. London: Sage.

Alexiou, A. and Singleton, A.D. (2015b). The Role of Geographical Context in Building Geodemographic Classifications. *Proceedings of the 23rd GIS Research UK*, April 15–17, Leeds, pp. 40-46.

Alvanides, S., Openshaw, S. and Rees, P. (2002). Designing your own geographies. In P. Rees, D. Martin and P. Williamson (Eds), *The Census Data System* (pp. 47–65). Chichester, UK: Wiley.

Ankerst, M., Breunig, M.M., Kriegel, H. and Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. *ACM SIGMOD International conference on management of data*. ACM Press. pp. 49–60.

Farr, M. and Evans, A. (2005). Identifying "Unknown Diabetics" using geodemographics and social marketing. *Interactive Marketing* 7(1) : 47-58.

Arribas-Bel, D., Nijkamp, P. and Schoelten, H. (2011). Multidimensional urban sprawl in Europe: a self-organizing map approach. *Computers, Environment and Urban Systems* 35(4): 265 – 275.

Arribas-Bel, D. and Schmidt, C.R. (2013). Self-organizing maps and the US urban spatial structure. *Environment and Planning B* 2: 362-371.

Arsdol, M.D. van Jr., Camilleri, S.F. and Schmid, C.F. (1958). The generality of the Shevky social area indexes. *American Sociological Review* 23, 277-284.

Ashby, D.I. and Longley P.A. (2005) Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS* 9(1): 53–72.

Assuncao, R.M., Neves, M.C., Camara G. and Freitas, C.D.C. (2006). Efficient Regionalization Techniques for Socio-Economic Geographical Units Using Minimum Spanning Trees. *Geographical Information Science* 20(7): 797-811.

Atkinson, R., and Flint, J. (2004). Fortress UK?: Gated communities, the spatial revolt of the elites and time-space trajectories of segregation. *Housing Studies* 19(6): 875-892.

Atkinson, R., and Tate, N.J. (2000). Spatial scale problems and geostatistical solutions: a review. *Professional Geographer* 52: 607-623,

Atlas, M., (1989). Gambling with Elections: The Problem with Geodemographics. In Sabato, L. (Ed.), *Campaigns and Elections: A Reader in Modern American Politics*. New York: Longman.

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):308–314.

Ball, G.H. and Hall, D.J. (1965). *Isodata: a method of data analysis and pattern classification. Office of Naval Research*. Information Sciences Branch. Menlo Park, US: Stanford Research Institute.

Banfield, J.D. and Raftery, A.E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* 49(3): 803-821.

Barbosa, O., Tratalosa, J.A., Armsworth, P.R., Davies, R.G, Fuller, R.A., Johnson, P. and Gaston, K.J. (2007). Who benefits from access to green space? A case study from Sheffield, UK. *Landscape and Urban Planning*, 83:187–195.

Batey, P. and Brown, P. (1995) From human ecology to customer targeting: the evolution of geodemographics. In Longley, P.A. and Clarke, G. (Eds.), *GIS for business and service*

planning. Cambridge: GeoInformation International.

Batey, P., Brown, P. J. B. and Corver, M. (1999). Participation in higher education: a geodemographic perspective on the potential for further expansion in student numbers. Journal of Geographical Systems 1:277 – 303.

Batey, P.W.J. and Brown, P.J.B. (2007). The spatial targeting of urban policy initiatives: a geodemographic assessment tool. *Environment and Planning A* 39: 2774-2793.

Bearman, N. and Singleton, A. (2014). Modelling the potential impact on CO2 emissions of an increased uptake of active travel for the home to school commute using individual level data. *Journal of Transport and Health* 1(4): 295-304.

Beaumont, J. R., and Inglis, K. (1989). Geodemographics in practice: developments in Britain and Europe. *Environment and Planning A* 21 (5): 587–604.

Bell, W. and Greer, S. (1962). Social area analysis and its critics. *The Pacific Sociological Review* 5(1): 3-9.

Berry, B. J. L. (1972). The goals of city classification. In B. J. L. Berry and K. B. Smith (Eds.), *City classification handbook: Methods and applications* (pp. 1–26). New York: Wiley-Interscience.

Berry, B.J.L. and Kasarda, J.D. (1977). *Contemporary urban ecology*. New York: Macmillan Publishers.

Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press.

Bezdek, J.C., Ehrlich, R. and Full, W.(1984). FCM: the fuzzy c-means clustering algorithm. *Computers and GeoSciences* 10: 191- 203.

Birkin, M. and Clarke, G. (1998). GIS, Geodemographics, and Spatial Modeling in the U.K. Financial Service Industry. *Journal of Housing Research* 9(1): 87-111.

Birkin, M.H., Clarke G.P. and Clarke M.C. (2002). *Retail geography and intelligent network planning.* London: John Wiley and Sons Ltd.

Birkin, M. and Clarke G. (2012). The enhancement of spatial microsimulation models using geodemographics. *Annals of Regional Science* 49:515-532.

Birkin, M. (1995). Customer targeting, geodemographics and lifestyle approaches. In Longley, P.A. and Clarke, G. (Eds.), *GIS for business and service planning*. Cambridge: GeoInformation

International.

Blake, M. and Openshaw, S. (1994). *Selecting census variables for use in classification research.* Working Paper, School of Geography, Leeds University.

Blashfield, R. and Aldenderfer, M.S. (1978). Literature Review on cluster analysis. *Multivariate Behavioral Research* 13(3): 271-295.

Bock, H.H. (1985). On some significance tests in cluster analysis. *Journal of Classification* 2(1): 77-108.

Boots, B. (2003). Developing local measures of spatial association for categorical data. Journal of Geographical Systems 5, 139–160.

Bowers, R.V. (1939). Ecological Patterning of Rochester, New York. *American Sociological Review* 4(2): 180-189.

Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society* B 26: 211-234.

Brindley, T. S. and Raine, J. W. (1979). Social Area Analysis and Planning Research. *Urban Studies* 16: 273–289.

Brown, P.J.B., Hirschfield, A.F.G. and Batey, P.W.J. (1991). Applications of geodemographic methods in the analysis of health condition incidence data. *Papers in Regional Science* 70(3): 329-344.

Brown, P. J. B., Hirschfield, A. F. G. and Batey, P. W. J. (2000). Adding Value to Census Data: Public Sector Applications of the Super Profiles Geodemographic Typology. *Journal of Cities and Regions* 10: 19–32.

Brunsdon, C., (2015). Reproducible research and quantitative geography. *Progress in Human Geography*, Quantitative methods I.

Brunsdon, C., Longley, P., Singleton, A., 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): 17–30.

Bulmer, M. (1984). *The Chicago School of Sociology: Institutionalization, Diversity, and the Rise of Sociological Research*. Chicago: University of Chicago Press.

Burgess, E.W. (1925). The Growth of City: An introduction to a research project. In Park, R.E., Burgess, W. and McKenzie, R.D. (Eds.). *The City*. Chicago: University of Chicago Press.

Burgess, E.W. (1964). Introduction. In Burgess E. and Bogue, D.J. (Eds.) *Contributions to Urban Sociology* (pp. 7-13). Chicago: University of Chicago Press.

CACI, (2013). *The acorn user guide: The consumer classification*. London: CACI Limited [electronic source, Accessed April 2014 <URL: http://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf>].

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

Catney, G., (2014). Religious concentration and health outcomes in Northern Ireland. In Social-Spatial Segregation. In C. D. Lloyd, I. G. Shuttleworth, & W. S. Wong (Eds.), *Social-Spatial Segregation: Concepts, Processes and Outcomes* (pp. 335-362). Bristol: Policy Press.

Charlton, M., Openshaw S., and Wymer C., (1985). Some new classifications of census Enumeration Districts in Britain. A poor man's ACORN. *Journal of Economic and Social Measurement* 13:69-96.

Clark, P.J., and Evans, F.C. (1954). Distance to nearest neighbor as a measure of spatial relationaships in populations. *Ecology* 35(4): 445-453.

Clarke, G., (1998). Changing methods of location planning for retail companies. *GeoJournal* 45: 289.

Colwell, P., Dehring, C. and Turnbull, G. (2002). Recreation demand and residential location, the influence of sensitivity for road traffic noise on residential location: Does it trigger a process of spatial selection? *Journal of Urban Economics* 51:418 – 428.

Corcoran, J., Higgs, G., Brunsdon, C., Ware, A. and Norman, P. (2007). The use of spatial analytical techniques to explore patterns of fire incidence: A South Wales case study. Computers, *Environment and Urban Systems* 31 (6), 623-647.

Cox, P.R. (1976). *Demography*. Cambridge: Cambridge University Press.

Davie, M. (1937). The pattern of urban growth. In G. Murdock, (Ed.) *Studies in the Science of Society.* New Haven: Yale University Press.

De Soete, G., Desarbo, W.S., and Carroll, J.D. (1985). Optimal Variable Weighting for Hierarchical Clustering: An Alternating Least-Squares Algorithm. *Journal of Classification* 2:

173–192.

Dear, M. and Flusty, S., (1998). Postmodern Urbanism. *Annals of the Association of American Geographers* 88(1): 50–72.

Dear, M. (2002). Los Angeles and Chicago School: Invitation to Debate. *City and Community* 1(1): 5–32.

Dever, A.G.E., Smith, L.T. and Stamps, B.V. (2005). Marketing And Quality Of Life: A Model For Improving Perinatal Health Status. *Quality-of-Life Research in Chinese, Western and Global Contexts* 25:145-181.

Domingos, P. (2012). A few useful things to know about machine learning. *ACM Communications* 55(10): 78-87.

Dorling, D., (2014). Class Segregation. In Social-Spatial Segregation. In Lloyd, C., Shuttleworth, I. and Wong, D.W. (Eds), *Social-spatial segregation: Concepts, processes and outcomes* (pp. 363-388). Bristol: Policy Press.

Duque, J.C., Anselin L. and Rey. S. J. (2011). The max-p-regions problem. *Journal of Regional Science* 52(3):397:419.

Duque, J.C., Ramos, R. and Suriñach, J. (2007). Supervised Regionalization Methods: A Survey. *International Regional Science Review* 30:195–220.

Easley and Kleinberg, (2010). Ne*tworks, Crowds, and Markets: Reasoning About a Highly Connected World* (Chap. 4). Cambridge: Cambridge University Press.

Ester, M., Kriegel, H., Sander, J. and Xu,tried to provide a comprehensive review of the relevant literature, from the early geodemographic precursors to factoFeng, Z. and R. Flowerdew (1998), Fuzzy Geodemographics: a contribution from fuzzy clustering methods. In S. Carver (Ed.), *Innovations in GIS 5*. London: Taylor and Francis.

Fisher, P. and Tate, N.J. (2015). Modelling Class Uncertainty in the Geodemographic Output Area Classification. *Environment and Planning B*: *Planning and Design* 42, 541–563.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2): 179–188.

Flowerdew, R. and Goldstein, W. (1989). Geodemographics in practice: developments in North America. *Environment and Planning A* 21: 605–616.

Fotheringham, A. S., and Wong, D.W.S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23: 1025–44.

Frank R.E. and Green P.E (1968). Numerical Taxonomy in Market Analysis. *Journal of Market Research* 5: 83-98.

Frawley, W.J., Piatetsky-Shapiro G, and Matheus, C.J. (1991). Knowledge Discovery in Databases: An Overview. In Piatetsky-Shapiro G. and Frawley, W.J. (Eds.), *Knowledge Discovery in Databases*. MA: AAAI/MIT Press.

Fried, A., and Elman, R. M., (1969). *Charles Booth's London: a portrait of the poor at the turn of the century, drawn from his Life and labour of the people in London*. London: Hutchinson.

Gidlöf-Gunnarsson, A. and Öhrström, E. (2007). Noise and well-being in urban residential environments: The potential role of perceived availability to nearby green areas. *Landscape and Urban Planning* 83: 115-126.

Gierl, H., and Schwanenberg, S. (1998). A comparison of traditional segmentation methods with segmentation based upon artificial neural networks by means of conjoint data from a Monte Carlo simulation. In Balderjahn, I., Mathar, R., and Schader M., (Eds.), *Classification, data analysis, and data highways* (pp. 386–392). Berlin: Springer.

Gittus, E. (1964). The structure of urban areas. *Town Planning Review* 35: 5-20.

Good, I.J. (1965). Categorization of Classification. In *Mathematics and Computer Science in Biology and Medicine* (pp. 115-125). London: HMSO.

Goodchild, M.F. (2001). Geographic information systems. In Smelser, N.J. and Baltes, P.B. (Eds.), *International Encyclopedia of the Social and Behavioral Sciences*, (pp. 6175–6182). Oxford: Pergamon.

Goss, J. (1995). Marketing the new marketing: The strategic discourse of geodemographic information systems. In: Pickles J., (Ed.) *Ground truth* (pp. 130–70). New York: Guilford.

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27: 857—874.

Graham, S.D.N. (2005). Software-sorted geographies. *Progress in Human Geography* 29(5): 562–580.

Graham, M. and Shelton, T. (2013). Geography and the Future of Big Data; Big Data and the Future of Geography. *Dialogues in Human Geography* 3(3) 255-261.

Green, P.E., Frank, R.E. and Robinson, P.J. (1967). Cluster analysis in test market selection. *Management Science* 13: 387–400.

Grekousis G. and Hatzichristos, T. (2012). Fuzzy clustering analysis in geomarketing research. *Environment and Planning B: Planning and Design* 40 (1), 95-116.

Gustafson E. and Kessel, W. (1979). Fuzzy clustering with a fuzzy covariance matrix. *Proceedings of IEEE CDC*, 1979.

Haining , R., Wise , S. and Ma, J., (2000). Designing and implementing software for spatial statistical analysis in a GIS environment. *Journal of Geographical Systems* 2: 257–286.

Harris, R.J. (2003). Population mapping by geodemographics and digital imagery. In Mesev, T. (Ed.). *Remotely-Sensed Cities* (pp. 223 – 241). Taylor and Francis Group.

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, 553-79.

Harris, R., Sleight, P. and Webber, R. (2005). *Geodemographics, GIS, and Neighbourhood Targeting.* Chichester: John Wiley and Sons.

Harris, R.J. and Feng, Y. (2016). Putting the geography into geodemographics: Using multilevel modelling to improve neighbourhood targeting - a case study of Asian pupils in London. *Journal of Marketing Analytics* (in press).

Hartigan, J.A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics* 28(1): 100-108.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The *elements of statistical learning* (Chap.14). New York: Springer.

Hedges, B., di Salvo, P. and Purdon, S. (1997). Health variations by 'ACORN' area classifications. In Prescott-Clarke, P. and Primatesta, P (Eds.), *The Health Survey for England 1996*. London: The Stationery Office.

Herbert, D.T. (1967). Social Area Analysis: a British Study. *Urban Studies* 4:41-60.

Hohenegger, J. (1986). Weighted Standardization — A General Data Transformation Method. Preceeding Classification Procedures. *Biometrical Journal* 28: 295–303.

Hoyt, H. (1939*). The Structure and Growth of Residential Neighbourhoods in American Cities*. Washington: Federal Housing Administration.

Hui, E., Chau, C., Pun, L. and Law, M. (2007). Measuring the neighboring and environmental effects on residential property value: using spatial weighting matrix. *Building and Environment* 42 (6):2333–2343.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data*. Prentice-Hall advanced reference series, Upper Saddle River, NJ: Prentice-Hall Inc.

Jackson, C., Best, N. and Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in Medicine* 25: 2136–2159.

Jain, A. Murty M. and Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys* 31(3): 264-323.

Janson, C.G. (1980). Factorial social ecology - An attempt at summary and evaluation. *Annual Review of Sociology* 6: 433–456.

Jardine N. and Sibson, R. (1971). *Mathematical Taxonomy*. London and New York: Jwiley and Sons Ltd.

Kaufmann, L. and Rousseeuw , P.J. (1990). *Finding Groups in Data.* New York: John Wiley and Sons, Inc.

Kelly, L.F., (1969). Classification of Urban Areas. *Quarterly Bulletin of the Research and Intelligence unit* 9: 13-19.

Kemper, R. (2006). Park, Robert Ezra (1864–1944). In Birx, H. (Ed.), *Encyclopedia of anthropology* (pp. 1829-1830). Thousand Oaks, CA: SAGE Publications.

Kim T., Horner, M.W. and Marans, R.W. (2005). Life Cycle and Environmental Factors in Selecting Residential and Job Locations. *Housing Studies* 20 (3): 457–473.

Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine Learning 30*: 271–274.

Kohonen, T. (2001). *Self-organizing maps*. Berlin: Springer.

Krzanowski W. J. and Lai, Y.T. (1985). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 44: 23–34.

Lage, J. P, Assunção , R. M. and Reis , E.A. (2001). A Minimal Spanning Tree Algorithm Applied to Spatial Cluster Analysis. *Electronic Notes in Discrete Mathematics* 7.

Liu, B. (2006). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. New York, Inc., Secaucus, NJ: Springer-Verlag.

Lloyd, C.D. (2014). *Exploring Spatial Scale in Geography*. UK: Wiley-Blackwell.

Lloyd, C.D., Catney, G., and Shuttleworth, I.G. (2014). Measuring neighbourhood segregation using spatial interaction data. In Lloyd, C. D., Shuttleworth, I. G., & Wong, D. W. (Eds.), *Social-Spatial Segregation: Concepts, Processes and Outcomes* (pp. 65-90). Bristol: Policy Press.

Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2): 129–137.

Longley, P. A. (2007). Some challenges to geodemographic analysis and their wider implications for the practice of GIScience. *Computers, Environment and Urban Systems* 31(6): 617–622.

Longley, P.A. (2005). Geographical information systems: a renaissance of geodemographics for public service delivery. *Progress in Human Geography* 29 (1): 57-63.

Longley, P.A. and Goodchild, M.F. (2008). The use of geodemographics to improve public service delivery. In J. Hartley, C. Donaldson, C. Skelcher, and M. Wallace, (Eds.), *Managing to Improve Public Services*, (pp. 176–194). Cambridge, UK: Cambridge University Press.

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): 362-382.

Longmoor, E. S. and Young, E. F., (1936). Ecological Interrelationships of Juvenile Delinquency, Dependency, and Population Mobility: A Cartographic Analysis of Data from Long Beach, California, *American Journal of Sociology* 41(5): 598-610.

MacQueen, J., (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability Statistics* 1: 281–297.

Marcuse, P. (2001). The Academic Formulations: Explanations of the Partitioned City. In Marcuse, P and van Kempen, R. (Eds.), *Of states and cities: the partitioning of urban space*. Oxford: Oxford University Press.

Martin, D. (2000). *Census 2001: making the best of zonal geographies*. Paper presented at The Census of Population: 2000 and Beyond, University of Manchester, 22-23 June 2000.

Martin, D., (1998). Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Science* 12: 673–685.

Martin, D., Nolan, A. and Tranmer, M. (2001). The application of zone-design methodology in the 2001 UK Census. Environment and Planning A 33: 1949-1962.

Mason, G.A. and Jacobson, R.D. (2007). Fuzzy Geographically Weighted Clustering. *Proceedings of the 9th International Conference on Geocomputation*, pp. 1–7.

Massey, D.S. and Denton, N.A. (1988). The Dimensions of Residential Segregation. *Social Forces* 67: 281–315.

McElrath, D.C. and Dennis C. (1962). Social areas of Rome: a comparative analysis. *American Sociological Review* 27(3): 376-391.

McKenzie, R.D., (1926). The Scope of Human Ecology. *Publications of the American Sociological Society* 20: 141–154.

Milligan, G.W. (1980). An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika* 45: 325–342.

Milligan, G. W. (1996). Clustering validation: results and implications for applied analyses. In Arabie, P., Hubert L.G.and De Soete, G. (Eds.), *Clustering and Classification*. Singapore: World Scientific Press.

Milligan, G. W. and Cooper, M.C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification* 5: 181-204.

Milligan, G. W. and Cooper, M.C. (1987). Methodological Review: Clustering Methods. *Applied Psychological Measurement* 11: 329–354.

Milligan, G. W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50: 159–179.

Mooi E. and Sarstedt, M. (2011). A *Concise Guide to Market Research*. Berlin Heidelberg: Springer-Verlag.

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: 553–579.

Nightingale, C. (2012). *Segregation: A Global History of Divided Cities. Historical Studies of Urban America*. Chicago: University of Chicago Press.

Nijman, J. (2000). The Paradigmatic City. *Annals of the Association of American Geographers* 90(1): 135-145.

Norman, P. (1969). Third Survey of London Life and Labour: a new typology of London districts. In Dogan, M. and Rokken, S. (Eds.), *Quantitative Ecological Analysis in the Social Sciences* (pp. 371-96). Cambridge, MA: MIT Press.

Novak, T.P., Leeuw J.D and MacEvoy, B. (1992). Richness Curves for Evaluating Market Segmentation. *Journal of Marketing Research* 29(2): 254-267.

O'Day, R. and Englander, D. (1993). *Mr Charles Booth's Inquiry: Life and Labour of the People in London Reconsidered*. London: Hambledon Press.

ONS (2012). *Changes to Output Areas and Super Output Areas in England and Wales, 2001 to 2011.* Office of National Statistics, Crown Copyright. [electronic source, accessed April 2014. <URL:https://data.gov.uk/data/resource_cache/e3/e3d20aab-d953-4705-8e58-79ce123f1ae5/report--changes-to-output-areas-and-super-output-areas-in-england-and-wales--2001-to-2011.pdf>].

ONS (2015a). *Methodology note on 2011 Travel to Work Areas*. Office for National Statistics, August 2015. [electronic source, accessed Feb. 2016 <URL:https://data.gov.uk/dataset/travel-to-worK-areas-uK-2011-guidance-and-information-v4>].

ONS (2015b). *Methodology Note for the 2011 Area Classification for Output Areas*. Office of National Statistics. [electronic source, accessed May 2015 <URL: www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/methodology-and-variables/methodology-oa.pdf>].

Openshaw, S., Cullingford D. and Gillard A. (1980). A critique of the national classifications of OPCS/PRAG. *Town Planning Review* 51 (4): 421.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning A* 16(1):17-31.

Openshaw, S. and Blake, M. (1995). Geodemographic segmentation systems for screening health data. *Journal of Epidemiology and Community Health* 49(2):34-38.

Openshaw, S., Rao, L. (1995). Algorithms for re-engineering 1991 Census geography. *Environment and Planning A* 27(3): 425-446.

Openshaw, S. and Taylor, P.J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In Wriggley N. (Ed.) *Statistical Methods in the Spatial Sciences,* (pp. 127-144). London: Pion.

Openshaw, S. (1998). *Towards the Marketing System that Thinks*. Institute of Direct Marketing, Lecture. [electronic source, accessed May 2016 <URL: http://www.geog.leeds.ac.uk/ presentations/98-8/ tsld104.htm>].

Openshaw, S. (1973). A regionalisation program for large data sets. *Computer Applications* 3(4): 136–47.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling. *Transactions of the Institute of British Geographers* 2: 459–72.

Openshaw, S. and Gillard, A.A., (1978). On the stability of a spatial classification of census enumeration district data. In P.W.S. Batey (Ed.) *Theory and Methods in Urban and Regional Analysis* (pp. 101-119). London: Pion.

Openshaw, S., and Wymer, C. (1995). Classifying and regionalizing census data. In Openshaw S. (Ed.) *Census users handbook* (pp. 239–70). Cambridge, UK: GeoInformation International.

Ordnance Survey (2015). *Open Map – User guide and technical specification* v1.4. Crown copyright. [electronic source, accessed January 2016 <URL: http://digimap.edina.ac.uk/webhelp/os/data_files/ os_manuals/os-vector-map-local-user-guide.pdf>].

Orr, J.M., Sackett, P.R. and DuBois, C.L.Z. (1991). Outlier detection and treatment in I/O Psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology* 44: 473-486.

Osborne, J.W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research and Evaluation* 15(12).

Paelinck, J.H.P. (2000). On aggregation in spatial econometric modeling. *Journal of Geographical Systems* 2: 157–65.

Tan, P.N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Park, R. (1936). Succession: An Ecological Concept. *American Journal of Sociology* X(2): 171-179.

Parkes, A., Kearns, A. and Atkinson, R. (2002). What makes people dissatisfied with their neighborhoods? *Urban Studies* 39:2413–2438.

Peeples, Matthew A. (2011) *R Script for K-Means Cluster Analysis*. [electronic source, accessed

June 2, 2016 <URL:http://www.mattpeeples.net/kmeans.html.>].

Petersen, J., Longley, P.A., Gibin, M., Mateos, P. and Atkinson, P. (2011). Names-based classification of accident and emergency department users. *Health and Place* 17(5): 1162-1169.

Petersen, J., Ashby, D. and Atkinson, P. (2007). Health Applications for Open Geodemographics. *Proceedings of the 21st GIS Research UK*, April 3–5, Liverpool.

Pickles, J. (2006). Ground Truth 1995–2005. *Transactions in GIS* 10: 763–772.

Quinn, J. (1940). Human ecology and interactional ecology. *American Sociological Review* 5(5): 713–722.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66 (336): 846–850.

Reardon, S.F. and O'Sullivan, D. (2004). Measures of Spatial Segregation. *Sociological Methodology* 34: 121–162.

Rees, P. (1972). Problems of classifying subareas within cities. In Berry B. J. L. and Smith, K. B. (Eds.), *City classification handbook: Methods and applications* (pp. 265–330). New York: Wiley-Interscience.

Reibel, M. (2011). Classification Approaches in Neighborhood Research: Introduction and Review. *Urban Geography* 32(3): 305-316.

Reibel, M. and Regelson, M. (2011). Neighborhood racial and ethnic change: the time dimension in segregation. *Urban Geography* 32(3): 360–382.

Renkow, M. and Hoover, D. (2000). Commuting, Migration, and Rural-Urban Population Dynamics. *Journal of Regional Science* 40(2):261-287.

Rhind, D.W. (1991). Counting the People In: Maguire, D. J., Goodchild, M. F. and Rhind, D. W. (Eds) *Geographical Information Systems: Principles and Applications* (pp. 127-137). Harlow: Longman.

Riddlesden, D. and Singleton, A.D. (2014). Broadband speed equity: A new digital divide?. *Applied Geography* 52:25-33.

Richards, S.J., (2008). Applying Survival Models to Pensioner Mortality Data. *British Actuarial Journal* 14(02): 257-303.

Rijnders, E., Janssen, N.A., van Vliet, P.H. and Brunekreef, B. (2001). Personal and outdoor

nitrogen dioxide concentrations in relation to degree of urbanization and traffic density. *Environmental Health Perspectives* 109(3):411–41.

Robson, B.T. (1969). *Urban Analysis: A study of city structure with spatial reference to Sunderland*. Cambridge: Cambridsge University Press.

Rogerson, P. (2014). *Statistical Methods for Geography*. London: Sage.

Rousseeuw, P. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* 20: 53–65.

Schelling, T.C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology* 1: 143–186.

Schnore, L.F. and Winsborough, H.H. (1972). Functional classification and the residential location of Social Classes. In Berry B. J. L. and Smith, K. B. (Eds.), *City classification handbook: Methods and applications* (pp. 123–141). New York: Wiley-Interscience.

See, L. and Openshaw, S. (2001). Fuzzy geodemographic targeting. In Clarke, G. and Madden, M. (Eds), Regional *Science in Business* (pp.269-282). Berlin: Springer.

Shevky E. and W. Bell, (1955). *Social Area Analysis*. Stanford, CA: Stanford University Press.

Shevky, E. and Williams M. (1949). *The Social Areas of Los Angeles: Analysis and Typology*. Berkeley and Los Angeles: University of California Press.

Singleton, A.D. (2010). The geodemographics of educational progression and their implications for widening participation in higher education. *Environment and Planning A* 42(11): 2560–2580.

Singleton, A. D. and Longley, P.A. (2009). Geodemographics, visualisation, and social networks in applied geography. *Applied Geography 29* (3): 289–298.

Singleton, A.D., Spielman, S. and Brunsdon, C. (2016b). Establishing a framework for Open Geographic Information science. *International Journal of Geographical Information Science* 30(8): 1507-1521.

Singleton, A. D., Wilson, A. G., and O'Brien, O. (2012). Geodemographics and spatial interaction: an integrated model for higher education. *Journal of Geographical Systems* 14(2), 223-241.

Singleton, A., Pavlis, M., and Longley, P. A. (2016a). The stability of geodemographic cluster assignments over an intercensal period. *Journal of Geographical Systems* 18(2), 97-123.

Singleton, A.D. and Spielman, S.E. (2013). The Past, Present and Future of Geodemographic Research in the United States and United Kingdom. *The Professional Geographer* 66 (4): 558-567.

Sivadas, E. (1997). A preliminary examination of the continuing significance of social class to marketing: a geodemographic replication. *Journal of Consumer Marketing* 14(6): 463 – 479.

Sleight, P. (1997). *Targeting customers: How to use geodemographic and lifestyle data in your business.* Henley-on-Thames: NTC Publications.

Sleight, P. (2004). An introductory review of geodemographic information systems. *Journal of Targeting, Measurement and Analysis for Marketing* 12: 379–388.

Sokal, R.R., and Sneath, P.H.A. (1963). *Principles of numerical taxonomy*. San Francisco: W.H. Freeman.

Spielman, S. and Singleton, A. (2015). Studying Neighborhoods Using Uncertain Data from the American Community Survey: A Contextual Approach. *Annals of the Association of American Geographers* 105(5):1003-1025.

Spielman, S.E. and Folch, D.C., (2015). Social Area Analysis with Self-Organizing Maps. In A. Singleton and C. Brundson (Eds.), *Geocomputation: A Practical Primer.* London: Sage.

Spielman, S.E., and Thill, J.C., (2008). Social area analysis, data mining and GIS. *Computers, Environment and Urban Systems* 32(2):110-122.

Steinley, D. (2006) K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology* 59: 1–34.

Struyf, A., Hubert, M. and Rousseeuw, P. (1997). Clustering in an Object-Oriented Environment. *Journal of Statistical Software* 1, 30.

Sugar, C.A. and Gareth M.J. (2003). Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *Journal of the American Statistical Association* 98(463): 750-763.

Sweetser, F.L. (1965). Factor structure as ecological structure in Helsinki and Boston. *Acta Sociologica* 8:205-2.

Theiler, J., and Gisler, G. (1997). A contiguity-enhanced K-means clustering algorithm for unsupervised multispectral image segmentation. Proceedings of SPIE (International Society for Optical Engineering), pp. 108–118.

Thorndike, R.L. (1953). Who belongs in the family? *Psychometrika* 18: 267–276.

Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 411–423.

Tickle, M., Blinkhorn, A., Brow n, P. (2000). A geodemographic analysis of the Dental patient population in the North West Region. *British Dental Journa*l 189(9): 494-499.

Timms, D.W.G. (1971). The *urban mosaic: Towards a theory of residential differentiation*. Cambridge: Cambridge University Press.

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46: 234–240.

Tryon, R.C. (1955). *Identification of social areas by cluster analysis: a general method with an application to the San Francisco bay area*. Berkley: University of California Press.

Twigg, L., Moon, G. and Jones, K. (2000). Predicting Small-Area Health-Related Behaviour: a Comparison of Smoking and Drinking Indicators. *Social Science and Medicine* 50(7-8): 1109–1120.

Vale, D.S. (2015). Transit-oriented development, integration of land use and transport, and pedestrian accessibility: combining node-place model with pedestrian shed ratio to evaluate and classify station areas in Lisbon. *Journal of Transport Geography*, 45:70–80.

Van Arsdol, M.D, Cammilleri, S.F. and Schmid, C.F. (1958). The generality urban social area indexes. *American Sociological Review* 23:227-284.

Van Ommeren, J., Rietveld, P., and Nijkamp, P. (1999). Job moving, residential moving, and commuting: a search perspective. *Journal of Urban Economics* 46: 230-253.

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): 379-403.

Villeneuve, P.J., Jerrett, M., Su, J.G., Burnett, R.T., Chen, A.J. Wheeler, et al. (2012). A cohort study relating urban green space with mortality in Ontario, Canada. *Environmental Research*, 115: 51–58.

Voas, D. and Williamson, P. (2001). The diversity of diversity: a critique of geodemographic classification. *Area* 33(1): 63–76.

Wallace, M. and Denham, C. (1996). *The ONS Classification of Local and Health Authorities of Great Britain.* London: Stationery Office.

Walter B. and Wirt, F.M. (1972). Social and political dimensions of American Suburbs. In B. J. L. Berry and K. B. Smith (Eds.), *City classification handbook: Methods and applications* (pp. 92–111). New York: Wiley-Interscience.

Ward, J.H.Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58:236–244.

Webber, R.J. and Craig J. (1976). Which local authorities are alike? *Population Trends* 5:13–19.

Webber, R.J. (1977). *An introduction to the national classification of wards and parishes*. London: Centre for Environmental Studies.

Webber, R.J. (1980). A response to the critique of the national classifications of OPCS/PRAG. The *Town Planning Review* 51 (4): 440–450.

Webber, R.J. and Farr, M. (2001). MOSAIC: From an area classification system to individual classification. *Journal of Targeting, Measurement and Analysis for Marketing* 10:55.

Webber, R.J. (2004). *The relative power of geodemographics vis a vis person and household level demographic variables as discriminators of consumer behaviour*, CASA Working Paper. London: CASA, UCL.

Webber, R.J. (2007). The metropolitan habitus: its manifestations, locations, and consumption profiles *Environment and Planning A* 39: 182–207.

Webber, R.J. (1978). Making the most of the census for strategic analysis. *The Town Planning Review* 49(3): 274–284.

Webber, R.J. (1975). *Liverpool Social Area Study, 1971 data: PRAG Technical Paper 14*. London: Centre for Environmental Studies.

Wehrens, R. and Buydens, L.M.C., (2007). Self- and Super-Organising Maps in R: the kohonen package. *Journal of Statistical Software* 21(5): 2007.

White, M. (1987). *American Neighborhoods and Residential Differentiation*. New York: Russell Sage Foundation.

Wijayanto A.W., Purwarianti, A., and Son L.H. (2015). Fuzzy geographically weighted clustering using artificial bee colony: An efficient geo-demographic analysis algorithm and applications to the analysis of crime behaviour in population. *Applied Intelligence* 44:377–398.

Wilkinson, R. and Pickett, K. (2010). *The Spirit Level: Why Equality is Better for Everyone*. London: Penguin.

Williamson, T., Ashby, D. and Webber, R. (2006). Classifying Neighbourhoods for Reassurance Policing. *Policing and Society* 16: 189-218.

Willis, I., Gibin, M., Barros, J., and Webber, R. (2010). Applying neighbourhood classification systems to natural hazards: a case study of Mt Vesuvius. N*atural Hazards and Earth System Science* 70:1–22.

Wilson, H. and Womersley, L. (1976). *Social area analysis — Liverpool 1971, Inner Areas Study*. Liverpool: Department of the Environment, City of Liverpool.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edn, Morgan Kaufmann Series in Data Management Systems. San Francisco, CA: Elsevier.

Zimek, A., Schubert, E. and Kriegel H.P. (2012). A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. *Statistical Analysis and Data Mining* 5(5): 363–387.

Appendix I

R Code Scripts

A. Data Input

This is the R script used to create the input dataset. The 2011 OAC raw datasets can be downloaded at *http://geogale.github.io/2011OAC/.* Lookup tables can be found at https://data.gov.uk/.

Preparing Lookups

```
# Census Variables mangled
OA_Input <- read.csv("2011_EW_60Var_Data_Percentages_and_OAC.csv")

# Travel to Work Areas
ttwa <- read.csv("LSOA11_TTWA11_UK_LU_V2.csv")

# England and Wales
OA_LSOA_lookup_EW <- read.csv("OA11_LSOA11_MSOA11_LAD11_EW_LUv2.csv")

# Scotland (Datazones = LSOA)
OA_LSOA_lookup_SC <- read.csv("OA_LSOA_Lookup/OA_DZ_IZ_2011.csv")

# N. Ireland (SOA = LSOA)
OA_LSOA_lookup_NI <- read.csv("OA_LSOA_Lookup/NI_SA_to_SOA.csv")
```

Join Data

```
OA_LSOA_lookup_EW <- merge(OA_LSOA_lookup_EW, ttwa, by = "LSOA11CD", all.x
= T)

OA_LSOA_lookup_SC <- merge(OA_LSOA_lookup_SC, ttwa, by.x =
"DataZone2011Code", by.y = "LSOA11CD", all.x = T)

OA_LSOA_lookup_NI <- merge(OA_LSOA_lookup_NI, ttwa, by.x = "SOA", by.y =
"LSOA11CD", all.x = T)

# Join all three into one lookup table, OA, TTWA code and TTWA name

# first make column names the same
colnames(OA_LSOA_lookup_NI)[2] <- "OA_CODE"
colnames(OA_LSOA_lookup_SC)[2] <- "OA_CODE"
colnames(OA_LSOA_lookup_EW)[2] <- "OA_CODE"

# Join
OA_LSOA_lookup <- rbind(OA_LSOA_lookup_EW[, c(2, 10, 11)],
OA_LSOA_lookup_SC[, c(2, 5, 6)], OA_LSOA_lookup_NI[, c(2, 14, 15)])
```

Append TTWAs to OA census data

```
# OAC2011 output table with cluster types
oac11 <- read.csv("2011OAC.csv")

# OAC input dataset, preprocessed into ratios
OAC_Input_PCT_RATIO <- read.csv("01_OAC_Percentages.csv")



# Merge Admin + OAC with census data
```

```
OA_Input <- merge(oac11, OAC_Input_PCT_RATIO, by.x = "Output.Area.Code",
by.y = "OA", all.x = T)

# Merge Merged data with ttwas
OA_Input <- merge(OA_Input, OA_LSOA_lookup, by.x = "Output.Area.Code",
by.y = "OA_CODE", all.x = T)

# Clean data frame
OA_Input <- OA_Input[, c(1:5, 72:73, 6:71)]
colnames(OA_Input)[1:5] <- c("OA_CODE", "LA_CODE", "LA_NAME", "RE_CODE",
"RE_NAME")
```

Transformations

```
# According to Gale's code, transformation comes first
OA_Input[, 14:73] <- asinh(OA_Input[, 14:73])

# Test for NAs
table(is.na(OA_Input))

# Range standardization function
Range_stand <- function(x){(x-min(x))/(max(x)-min(x))}

# Test extents
apply(OA_Input[, 14:73], 2, summary)

# Write final csv file, OA_Input #
write.csv(OA_Input, "OA_Input.csv")
```

Note that range standardization was not applied at the end of the script, as the original OAC 2011 prompted.

B. Find k per Geography

This code finds and stores in a list how many super-group clusters are present in OAC2011 per geography. This example uses the TTWAs, but can be changes to Regions and LADs by changing the TTWA11NM column accordingly.

```
# The position of the supergroup code (or similar) column in the input #
dataframe
pos_groupcode <- 8

# The stores the number of k per i geography code
cl_per_TTWA <- list()

for(i in 1:nlevels(OA_Input$TTWA11NM)) {

  cl_per_TTWA [[i]] <- nrow(as.data.frame(
                       table(OA_Input[OA_Input$TTWA11NM ==
                       levels(OA_Input$TTWA11NM)[i], pos_groupcode])))
  }
```

C. K-means per Spatial Context

The script to run k-means per geography, using z-scores (`scale` function). Note that columns 14 to 73 are the ones that the 60 variables are stored.

```
### Declarations ####

REG_km <- list()
TTWA_km <- list()
LA_km <- list()

# Assuming 8 clusters at supergroup
kcen <- 8

# rounding digits
v_digits <- 4

### UK K-means ###

# Data prep: Zscore standardization, rounding
km_df <- scale(OA_Input[, 14:73])
km_df <- round(km_df, digits = v_digits) # for algorithm to converge

# K-means clustering algorithm for UK
UK_km <- kmeans(km_df, centers=kcen, nstart = 500, iter.max = 10000000)

### Regional K-means ###

for(i in 1:nlevels(OA_Input$RE_NAME)) {

  # Subset, Zscore standardization, rounding
```

```
  km_df <- OA_Input[OA_Input$RE_NAME == levels(OA_Input$RE_NAME)[i],
           14:73]

  km_df <- scale(km_df)
  km_df <- round(km_df, digits = v_digits)

  # K-means
  print(paste(i, levels(OA_Input$RE_NAME)[i] , sep = " - "))
  print(table(is.na(km_df)))
  REG_km[[i]] <- kmeans(km_df, centers=kcen, nstart = 500, iter.max =
                        10000000)
}

### TTWA K-means ###

for(i in 1:nlevels(OA_Input$TTWA11NM)) {

  # Subset, Zscore standardization, rounding
  km_df <- OA_Input[OA_Input$TTWA11NM == levels(OA_Input$TTWA11NM)[i],
           14:73]
  km_df <- scale(km_df)
  km_df <- round(km_df, digits = v_digits)

  # K-means
  print(paste(i, levels(OA_Input$TTWA11NM)[i] , "centres:",
        cl_per_TTWA[[i]], sep = " - "))
  print(table(is.na(km_df)))
  km_df <- replace(km_df, is.na(km_df), 0)

  TTWA_km[[i]] <- kmeans(km_df, centers=cl_per_TTWA[[i]], nstart = 500,
                         iter.max = 10000000)
}

### LAD K-means ###

for(i in 1:nlevels(OA_Input$LA_NAME)) {
  # Subset, Zscore standardization, rounding
  km_df <- LA_Input[OA_Input$LA_NAME == levels(OA_Input$LA_NAME)[i],
           14:73]
  km_df <- scale(km_df)
  km_df <- round(km_df, digits = v_digits)

  # K-means
  print(paste(i, levels(OA_Input$LA_NAME)[i] , "centres:",
        cl_per_LA[[i]], sep = " - "))
  print(table(is.na(km_df)))
  km_df <- replace(km_df, is.na(km_df), 0)

  LA_km[[i]] <- kmeans(km_df, centers=cl_per_LA[[i]], nstart = 500,
                       iter.max = 10000000)
}
```

D. Cluster Comparison

The code to make cluster comparisons and output a ranked table based on the average Angular Cosine Similarity. This example uses TTWAs, change when prompted for other geographies.

```
# Garbage collector
gc()

# For COSINE tables
library(gmodels)
library(lsa)
library(gtools)

# Number of clusters that have been selected
kcen <- 8

# Change this for other geographies
SUB_km <- TTWA_km
cl_per_sub  <- cl_per_TTWA
geo_levels <- nlevels(OA_Input$TTWA11NM)

# Global Declarations
acos_matrix <- list()
sim_table <- data.frame(matrix(0, ncol = 5, nrow = 0))
colnames(sim_table)<- c("TTWA", "CLUSTERS","AVG_SIM", "PERMUTATION",
                        "CL_SIM")

for(sub_geo in 1:geo_levels) { # start main loop

  print(paste("Generating matrix ",
              sub_geo, " of ", geo_levels, sep = ""))

  # Angular Cosine Similarity
  cross_class <- data.frame(1,2,3)
  colnames(cross_class) <- c("Clust_SUB", "Clust_NAT", "Similarity")
  cross_class <- cross_class[-1,]

  for(i in 1:(cl_per_sub[[sub_geo]])) {
    for(j in 1:kcen)                    {

      Similarity <- 1 - (acos(cosine(SUB_km[[sub_geo]]$centers[i, ],
                    UK_km$centers[j, ])))/pi
      cross_class <- rbind(cross_class, c(i, j, Similarity))

    }
  }

  # The crossclass table is in long Format:
  colnames(cross_class) <- c("Clust_SUB", "Clust_NAT", "Similarity")
  cross_class <- cross_class[order(cross_class$Clust_NAT,
                             decreasing = F), ]
```

```
  # Make into wide format for easier manipulation
  acos_matrix <- matrix(data = cross_class$Similarity, nrow =
                        cl_per_sub[[sub_geo]], ncol = kcen,
                        dimnames = list(paste("SUB",
                        as.character(1:cl_per_sub[[sub_geo]]), sep =
                        ""), paste("NAT", as.character(1:kcen), sep =
                        "")))

  # Find best average similarity based on the (kcen) clusters
  ksub <- cl_per_sub[[sub_geo]]
  acos_max <- 0
  attr_fit <- list()

  # Find all permutations based on n taking r at a time, n!/(n-r)!
  k_perm <- permutations(n = kcen, r = ksub, repeats.allowed = F)

  for (p in 1:nrow(k_perm)) {
    temp_comb <- k_perm[p, ]

      for(h in 1:ksub) {
      attr_fit[h] <- acos_matrix[h, temp_comb[[h]]]
    }

  # Find best average similarity and keep it
    acos_mean <- mean(unlist(attr_fit))
    if(acos_mean >= acos_max) {acos_max <- acos_mean
                               acos_perm <-temp_comb
                               acos_sims <- unlist(attr_fit)
    }
  }

  # Output Table, subregion results per row
  sim_table[sub_geo, ] <-c(levels(OA_Input$TTWA11NM)[sub_geo], ksub,
                           round(acos_max, 4),
                           paste(acos_perm[acos_perm[1:ksub]],
                           collapse = "/"),
                           paste(round(acos_sims, 2),
                           collapse = "/"))

} # ends loop for all geography units

# Order output results (and possibly save into .csv if needed).
sim_table_ordered <- sim_table[order(sim_table$AVG_SIM,
                     decreasing = F), ]
sim_table_ordered$TTWA <- factor(sim_table_ordered$TTWA,
                          levels = sim_table_ordered$TTWA)
sim_table_ordered$AVG_SIM <- as.numeric(sim_table_ordered$AVG_SIM)
```

E. g-Factor Attribute Adjustment

This function can be used to adjust values based on the contextual geography and the g-Factor to values [0,1]. The function checks standard deviation values in case they are 0, which would produce `NaN` otherwise.

```
# df_input: the dataframe containing the input dataset, e.g. OA_Input.
# geography: the column of the dataframe that stores the spatial
# context codes, e.g. OA_Input$TTWA11NM.
# nvar: the variable range, e.g. 14:73
# g: the value of factor g, e.g. 0.75.

gs_adjust <- function(df_input, geography, nvar, g)  {

  for(v in nvar) {
    nat_mean <- mean(df_input[, v])
    nat_sd <- sd(df_input[, v])

    for(subregion in 1:nlevels(geography)) {

      sub_values <- df_input[geography ==
                             levels(geography)[subregion], v]
      sub_mean <- mean(sub_values)
      sub_sd <- sd(sub_values)

        if(sub_sd != 0) {

          df_input[geography == levels(geography)[subregion], v] <-
                   (1-g)*((sub_values - nat_mean)/nat_sd) +
                   g*((sub_values - sub_mean)/sub_sd)
        } else {
          df_input[geography == levels(geography)[subregion], v] <- 0}
    }

  }

  return(df_input[, c(1, nvar)])
}

# example of data input adjustment:
adjusted_OA_Input <- gs_adjust(OA_Input, OA_Input$RE_NAME, 14:73, 0.25)
```

Appendix II

Selected Publications

1. Alexiou, A, Singleton A.D. and Longley P.A. (2016). A Classification of Multidimensional Open Data of Urban Morphology. Built Environment 42(3): 463-476.
2. Alexiou, A. and Singleton, A.D. (2015a). Geodemographic Analysis. In Singleton A.D. and Brunsdon C. (Eds.), *Geo computation: a practical primer* (pp. 137-152). London: Sage.
3. Alexiou, A. and Singleton, A.D. (2015b). The Role of Geographical Context in Building Geodemographic Classifications. *Proceedings of the 23rd GIS Research UK*, April 15–17, Leeds, pp. 40-46.

A Classification of Multidimensional Open Data for Urban Morphology

ALEXANDROS ALEXIOU, ALEX SINGLETON and PAUL A. LONGLEY

Identifying socio-spatial patterns through geodemographic classification has proven utility over a range of disciplines. While most of these spatial classification systems include a plethora of socioeconomic attributes, there is arguably little to no input regarding attributes of the built environment or physical space, and their relationship to socioeconomic profiles within this context has not been evaluated in any systematic way. This research explores the generation of neighbourhood characteristics and other attributes using a geographic data science approach, taking advantage of the increasing availability of such spatial data from open data sources. We adopt a SOM (Self-Organizing Maps) methodology to create a classification of Multidimensional Open Data Urban Morphology (MODUM) and test the extent to which this output systematically follows conventional socioeconomic profiles. Such an analysis can also provide a simplified structure of the physical properties of geographic space that can be further used as input to more complex socioeconomic models.

Geodemographics is a field of quantitative geography that engages in the analysis and classification of populations into discrete classes based on socioeconomic and built environment characteristics of small-area geography. Simply put, geodemographics is the 'analysis of people by where they live' (Sleight, 1997, p. 16). Such classifications have demonstrated utility over a range of public and private sector applications (Longley, 2005; Longley and Goodchild, 2008; Reibel, 2011; Singleton and Spielman, 2013). A geodemographic analysis is essentially a data reduction methodology that aggregates populations, so that correlations between sub-populations can be drawn on with ease. It involves the process of producing key statistics of a particular area, on the basis of the characteristics of its residents and their contexts.

Geodemographic applications were initially developed as a strategy to analyse and systematically document socio-spatial segregation. The associated data reduction methods were established in the 1970s (Webber, 1978), although a wider review and interpretation would extend right back to the 'human ecology' studies from the Chicago School of Sociology in the 1920s (Burgess, 1925), social area analysis in the 1950s (Shevky and Bell, 1955) and the factorial ecologies of the 1970s (Janson, 1980). Although that geodemographics has evolved considerably over the years (Singleton and Spielman, 2013), its conceptual background is still wedded to the principle that people tend to align themselves with the behaviour and aspirations of the local communities in which they live. The inferential nature of the aggregations rely on the notion of societal homophily, or in other words, that 'birds of a feather flock together' (Harris *et al*., 2005). As such, people who live close by (e.g. in the same neighbourhood) are more likely to have commonalities in attributes and behaviours than a randomly selected group of people.

Although geodemographic frameworks can

capture a wide set of input attributes, current classification systems typically include little to no input of explicitly spatial attributes regarding the built and physical attributes of neighbourhoods. There is, however, an abundance of variables that might be collected on the built forms and relative locations that underpin neighbourhood differentiation. For instance, proximity to certain amenities is important to residential decisions such as transport nodes, parks, retail and healthcare-facilities. There has, for example, been extensive research into the topic of analysing relationships between accessibility and urban development patterns, (e.g. land use-transportation interaction (LUTI) models); and connectivity has been advanced as a key feature in shaping urban residential dynamics and socio-spatial segregation (Dear, 2002). Research on residential decisions has also attracted a lot of attention over the years, particularly through hedonic modelling. While most of the relevant research focuses on the importance of work location (Van Ommeren *et al*., 1999; Renkow and Hoover, 2000), there is strong evidence that certain demographic groups favour some relative locations over others, and that the nature and configuration of the local built environment and land-use characteristics are also relevant (Hui *et al*., 2007). For instance, individuals with children often favour green space and recreational opportunities nearby, while those without children prefer smaller residences that offer closer proximity to central services (Colwell *et al*., 2002). Other characteristics may impact the area as unfavourable due to negative externalities, such as high-speed roads or railway tracks within the vicinity of the neighbourhood (Parkes *et al*., 2002). It is unclear exactly how such characteristics impact upon residential decisions as there are many synergies involved across lifecycles (Kim *et al*., 2005). For instance, moderate proximity (200 m to 300 m) to a green space may mitigate negative effects of noise pollution (Gidlof-Gunnarsson and Ohrstrom, 2007).

Some census variables reflect limited built environment characteristics, for instance housing type and population densities. For classification systems that have been developed entirely from census variables, such as the publicly open ONS (Office of National Statistics) Output Area Classification (OAC) for 2011, attributes such as density can, however, be misleading; the arbitrary nature of the geographic extents of the administrative areas for which population measurements are offered renders comparisons between the physical features ineffective. Other proprietary geodemographic classifications, such as Mosaic by Experian (Nottingham, UK) and Acorn by CACI (London, UK) include some measures of relative location (CACI, 2013; Experian, 2014). However, to what precisely these attributes pertain, how they are used in the clustering process and the weight they are assigned in the final classification remains obscure, because of the commercial sensitivities that are inherent in 'black box' commercial solutions (Singleton and Longley, 2009).

In this paper, we test whether specific and multidimensional urban morphologies systematically correspond with socioeconomic characteristics at the neighbourhood level. In order to identify and analyse such attribute patterns, we adopt a geodemographic approach, which involves the creation of a classification for a national extent, based on clustering at the small area level. In essence, we try to identify the physical and built environment characteristics that might be used to supplement neighbourhood typologies.

Open Data Inputs

This research captures a variety of physical attributes collected for a small-area geography, and in order to enhance reproducibility, replication and extension these inputs are assembled from Open Data sources (Singleton *et al*., 2016). We produce a classification at the 2011 UK Census Output Area level for the 181,408 Output Areas (OAs) that make up England and Wales. One of the main providers of geographical data for England and Wales is the

national mapping agency Ordnance Survey (OS), and there are many datasets available within their repository, with varying degrees of granularity, depending on whether they are publicly accessible or available for purchase. As this paper focuses on Open Data sources, we use OS Open Map – Local, the most recent and detailed open OS vector data product currently available (Ordnance Survey, 2015). However, within different contexts, such data might also be supplemented by other national mapping agency data, or alternative sources such as OpenStreetMap (www.openstreetmap.org). The OS vector data product provides a variety of information including outlines of buildings, street network with hierarchy, railways, woodland areas, surface water and important functional sites.

While the OS Open Map – Local provides the main source of this data, there were a few other sources within England and Wales deemed of utility. These included data about listed buildings and historic parks and gardens supplied by the *Historic England Archive* (https://services.historicengland.org.uk/NMRDataDownload/) which is regularly updated (November 2015 update used here) and also under Open Data License. For Wales, the corresponding provider is the Cadw heritage organization (available through the UK data Service, https://data.gov.uk/dataset/listed-buildings-in-wales-gis-point-dataset), although the data are slightly outdated (September 2011). Commercial buildings for local retail centres were identified using data from the Local Data Company, an Open version of which is available through the ESRC Consumer Data Retail Centre. Finally, we included aggregated data on housing type from the 2011 Census supplied by the Office for National Statistics (ONS). Unfortunately, there are currently no Open Data available on building age or height.

Table 1 summarizes the range of inputs

Table 1. Description of the spatial dataset compiled for England and Wales.

Variable Name	*Variable Description*
D1: OA Boundaries	181,408 Output Area boundaries, as defined by the 2011 Census. All other data were spatially joined with the respective OAs that they fall into (data features were split when falling into more than one OA).
D1: Buildings	12,878,666 Building objects represented as polygons. Note that these areas do not represent individual households.
D2: Road Network	Road network is represented as line segments, approximate to the road centre. The categories include 'Motorway', 'Primary Road', 'A Road', 'B Road', 'Minor Road', 'Pedestrianized Street', 'Local Street' and 'Private Road Publicly Accessible', as well as their 'Collapsed Dual Carriageway' counterparts.
D3: Woodland	Areas of trees represented as polygons, described as coniferous and non-coniferous.
D4: Functional Sites/ Important Buildings	120,677 Building polygons that can be found within functional sites. They are categorized into themes such as Air Transport, Education, Medical Care, Road Transport and Water Transport, which are further classified into numerous more discrete classes.
D5: Railway Stations and Tracks	Railway tracks and tunnels represented as lines (in this instance we used tracks only in the analysis) and Railway Stations defined as points.
D6: Surface water	Polygons of surface water. Small rivers and streams are represented as lines and were not included in the dataset. The dataset was also supplemented with 'seawater', derived from the country's coastline.
D7: Registered Historic Buildings	406,496 listed historic buildings defined as points, which were geolocated.
D8: Registered Parks and Gardens	2,007 Polygon features with extents of the parks / gardens, classified as I, II*, or II, from most to least important. For Wales, the 372 sites were identified from points from a 'Named Places' dataset and given an approximate 200 m radius.
D9: Retail Centres	1,312 Retail Centres across England and Wales. There is no recent update for this dataset which dates back to 2004. The centres are only depicted as points and have no typology attached. We assumed an average radius of 200 m to convert them to areas.
D10: Housing Type	Percentage of households that are classified by the Census as Detached, Semi-detached, Terraced or Flat.
D11: Population	Population of total persons per OA.

used to derive measures featured in this analysis.

The classification presented later was created for Output Areas (LSOAs), and as such the input measures were assembled for this geography. These zones offer advantage over other administrative units in England and Wales since many other socioeconomic classifications are offered at the OA level, such as the 2011 ONS Output Area Classification, thus making comparisons possible. Additionally, such geography also allows the incorporation of Census data which is distributed for these units. However, for the range of the derived measures that are described in the remainder of this section, there are problems with this approach. OA borders were designed to maximize within zone homogeneity in population characteristics (population normalization), without regard to the geographical features of the area (Martin *et al*., 2001; see figure 1). As such, for proximity based inputs there were challenges about how such measures might be calculated, and to which area they should be attributed.

A similar attempt to create such a dataset was made by the Department for Communities and Local Government in 2005, within the framework of the ONS Neighbourhood Statistics, described as Land Use Statistics. The dataset was described as a generalized land-use database aggregated into OAs. The dataset contained estimates of built environment attributes, such as roads, paths, domestic and non-domestic buildings, domestic gardens, water, rail etc. Despite the fact that the proprietary OS Enhanced Basemap was used to create this resource, ONS classified it as experimental, as there were issues of accuracy, mainly arising because only the centroids of features were taken into account in class assignments of aggregations.

To facilitate these methodological shortcomings, we adopted three different types of attribute measures for each OA that related to either two types of proximity measures including *adjacency effects* or *intermediate effects*; and additionally *direct measures*. The last of these are simply attributes captured at the OA level, while the first two assume buildings as the initial unit of analysis which are then later assigned to OAs. Building polygon

Figure 1. Maps looking at the un-generalized Output Area borders (black lines) around Sefton Park, Liverpool. *Left*: Notice how the area of the park is divided arbitrarily between proximal OAs (crosshatched pattern). *Right*: Output Area borders usually coincide with the street network, making simple street network-to-area assignments impracticable.

features serve as observations in this input dataset, and represent homogenous built-up areas which can include one or more households. A graphical representation of the model is described in figure 2. All the attributes collated as input across all domains are summarized in table 2.

For both types of proximity measure, we used a series of spatial queries that identified buildings that fulfil certain criteria, for instance, which buildings are within a set distance of a major street? The buildings that met each criterion were then assigned to OA aggregations with weights determined by their attributed area. Thus, within each OA, a ratio of the area of buildings meeting the criteria relative to the total built areas was calculated for each of the attributes considered in the analysis. The necessity to differentiate between adjacency and intermediate proximity effects follows the logic that not all built environment characteristics have the same effect, and these effects may vary in scale. For example, when considering the location of a residential property, being adjacent to a very major road might be perceived as having a negative impact, given the noise/pollution associated with increased traffic volumes, whereas being near, but not adjacent to a busy road might be perceived as advantageous, given the enhanced connectivity this might facilitate.

We defined *adjacency effects* to features measured within 100 m linear distance, as commonly used in the literature on negative externality effects of built environment features, such as noise or pollution from roads (Rijnders *et al*., 2001). For *intermediate effects* a distance of 600 m was used, on the basis of various Western international definitions of 'within walking distance'. The distance figure generally varies depending on the context of analysis, but distances between 300 m and 900 m are considered appropriate for urban

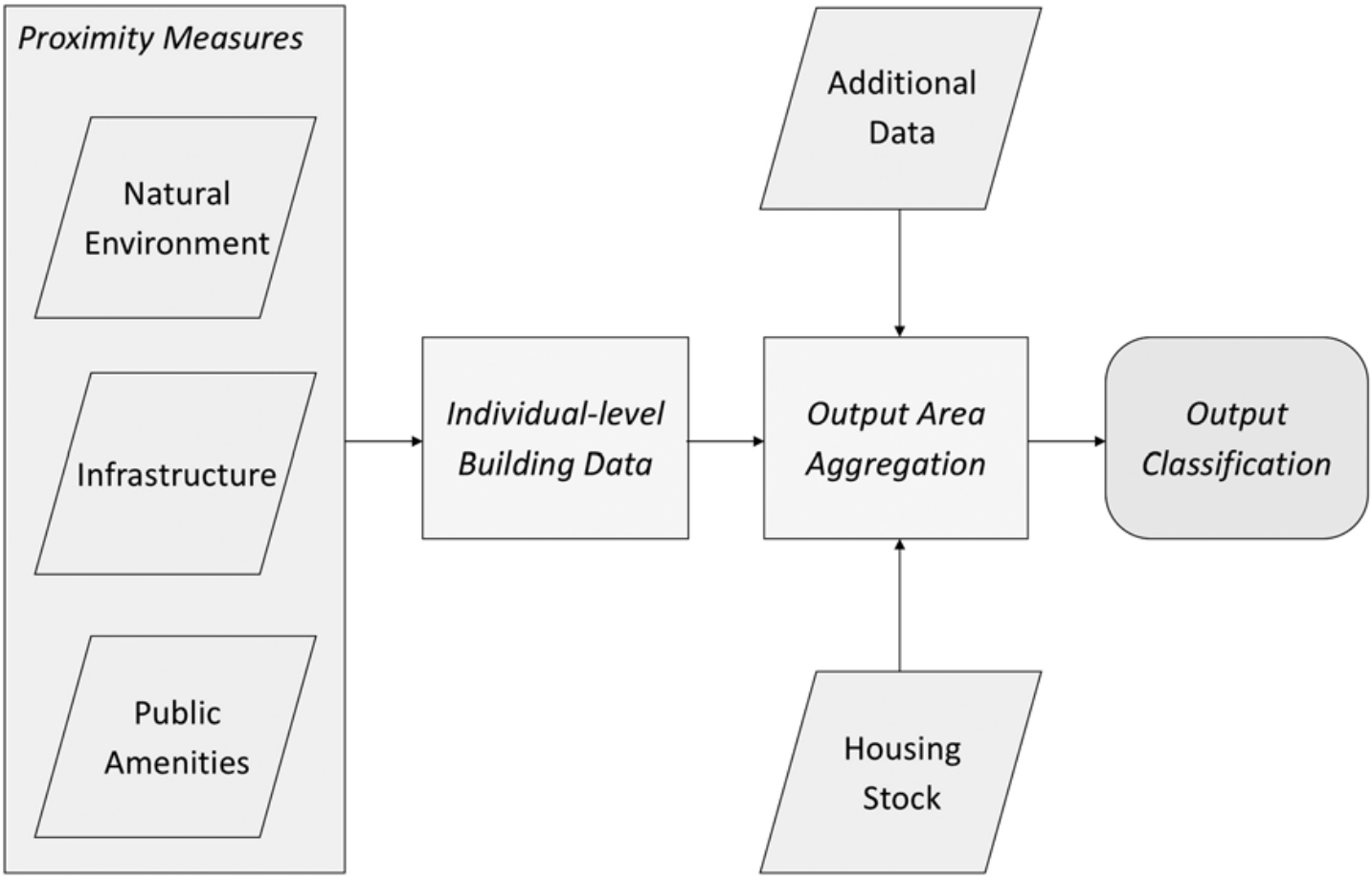

Figure 2. The spatial data model used to process data and produce Output Area inputs to the classification.

features (Hui *et al*., 2007; Barbosa *et al*., 2007; Villeneuve *et al*., 2012; Vale, 2015).

Beyond these distances we assume there are no adjacency or intermediate effects. The delineation of *adjacency effects* or *intermediate effects* brings additional practical considerations which relate to the overall density of the built environment features being considered. In common with practice when creating inputs to multidimensional classifications, preference should be for those attributes which, in addition to theoretical rationale, also provide useful differentiation between areas (Spielman and Singleton, 2015). For example, in this application, when 600 m buffers were used for major roads, this resulted in more than 50 per cent of buildings meeting this criterion, thus providing a weak differentiation. These tasks were computationally expensive, as the complete dataset contains more than 12.8 million observations (building polygons). Thus the database was pre-processed into regional datasets which were then computed separately using the R programming language.

Finally, there were two further types of *direct* measures: those which were derived from geographic features, and those which were simple inputs from secondary data. The derived *direct* measures included listed buildings and culs-de-sac (dangling segments in the road network). The latter of these was defined geocomputationally as the end of a line segment that did not intersect with any other such segment. A sensitivity of 10 m was applied to this criterion in order to avoid topological errors and intermittent street segments. The results show that such measures can capture specific urban morphologies even at the small-area level as we show in figure 3.

For the other non-derived *direct* measures, the variables were simply aggregated directly at the OA level, such as the housing type. Population density was calculated using a

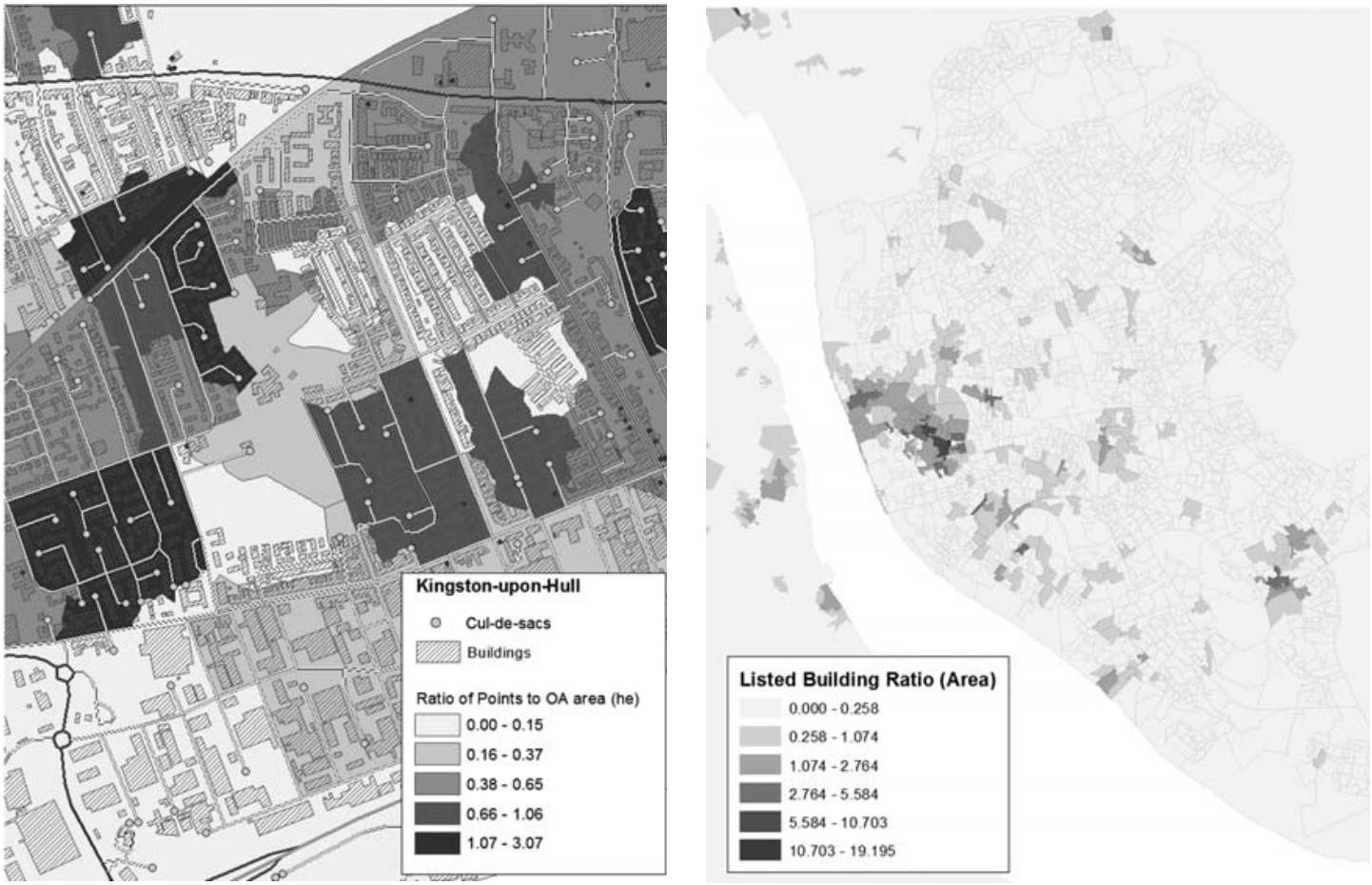

Figure 3. *Left*: Attribute of cul-de-sac ratio per OA at Kingston-upon-Hull, Yorkshire. *Right*: The ratio of listed (registered) buildings per OA area in Liverpool.

ratio of persons per total building area, which potentially would give more accurate results regarding housing conditions. The final OA attributes along with their descriptions are provided in table 2.

A Multidimensional Classification of the Built Environment

Methodologically, our cluster analysis follows a conventional approach as detailed in Harris *et al*. (2005); however, here we use only built environment data to create the typology. A common clustering technique used in geodemographic analyses is the iterative allocation – reallocation algorithm, known as k-means. Although this algorithm has been used in a variety of geodemographic applications, our dataset is sparsely populated, and k-means is known not to respond well to the non-Gaussian distributions that characterize such datasets (Everitt *et al*., 2011).

Table 2. Built environment attributes used in the classification.

Variables	*Variable Description, Aggregated per OA Code*
Adjacent effects	
1. Major Roads	Percentage of the area of buildings that the centroid is within 100 m of a major road to the total building area. We defined major as those of type 'Motorway', 'A Road' and 'Primary Road'.
2. Arterial Roads	Percentage of the area of buildings that their centroid is within 100 m of an arterial road to the total building area. We defined Arterial roads as those with type 'B Road'.
3. Pedestrian Roads	Percentage of the area of buildings that their centroid is within 100 m of a pedestrian road or footway to the total building area.
4. Railway Tracks	Percentage of the area of building units that their centroid is within 100 m of railway tracks, excluding tunnels, to the total building area.
5. Woodland Areas	Percentage of the area of building units that their centroid is within 100 m of woodland features to the total building area.
6. Surface Water	Percentage of the area of building units that their centroid is within 100 m of surface water (inland) and seafront (calculated by the distance from the coastal line), but excluding small rivers and streams, to the total building area.
Intermediate effects	
7. Railway Stations	Percentage of the area of building units that their centroid is within 600 m from the centroid of a railway station to the total building area.
8. Parks & Gardens	Percentage of the area of building units that their centroid is within 600 m from the registered site extents to the total building area.
9. Retail Centres	Percentage of the area of building units that their centroid is within 600 m from the retail centre centroid plus 200 m to the total building area.
10. Schools	Percentage of the area of building units that their centroid is within 600 m from the sites that are identified as primary through secondary education to the total building area.
11. Higher Education	Percentage of the area of building units that their centroid is within 600 m from the sites that are identified as further and higher education to the total building area.
Direct measures	
12. Detached Ratio	Percentage of unshared households that are classified by the 2011 Census as detached housing to the total building area.
13. Semi-Detached Ratio	Percentage of unshared households that are classified by the 2011 Census as semi-detached housing to the total building area.
14. Terraced Ratio	Percentage of unshared households that are classified by the 2011 Census as terraced housing to the total building area.
15. Flat Ratio	Percentage of unshared households that are classified by the 2011 Census as Flats to the total building area.
16. Density	Ratio of persons to total building area (people/he).
17. Cul-de-sac	Ratio of culs-de-sac or dead-end road points to the total OA area (points/he).
18. Registered Buildings	Ratio of listed buildings to the total OA area (points/he)

As such, in this framework we adopt the alternative technique of a Self-Organizing Map (SOM). A SOM is an unsupervised classifier that uses artificial neural networks to classify multidimensional observations in two-dimensional space based on their similarities (Kohonen, 2001). A SOM typically organizes observations by projecting them onto a plane, and through consecutive iterations finds the best configuration of observations so that every observation is most similar to the others closest to them. Typically, the SOM mapping process employs a lattice of squares or hexagons as the output layer, and the results are therefore easily mapped as they retain their topology. SOMs have many applications in a broad range of fields, from medicine and biology to image analysis and computer science. SOMs have also been tested as an alternative classifier of census data (Spielman and Thill, 2008; Arribas-Bel and Schmidt, 2013) where they seem to perform well for socioeconomic data at the US Census tract scale. Arribas-Bel *et al*. (2011) have also demonstrated the algorithm capabilities to measure urban sprawl in Europe using a similar attribute set, specifically six variables: connectivity; decentralization; density; scattering; availability of open space; and land-use mix. The technique also has the advantage of not assuming any hypotheses regarding the nature or distribution of the data, and responds well to geographic sensitivity. A further advantage of using a SOM is the capacity to visualize the structure of data values aiding initial data exploration. This feature can be very useful when analyzing datasets such as our built environment measures, where there are little to no *a-priori* hypotheses on their underlying distribution.

As input to this analysis the dataset comprising the eighteen variables described in table 2 was transformed into z-scores in order to standardize the measures. The majority of the analysis and output production was performed in the R programming language using the 'Kohonen' library (Wehrens and Buydens, 2007). More specifically, we adopted a SOM approach to cluster our input dataset using the methodology described by Spielman and Folch (2015). A relatively unexplored built environment classification with too many clusters would be difficult to interpret, so we selected a 4-by-2 hexagonal grid, which produces eight distinct clusters. We implemented a hexagonal geodesic grid to project results. A geodesic plane forces the cells' relations to 'loop' around the edges, while the hexagonal representation is typically favoured over grids, as this configuration benefits from every cell having six immediate neighbours. The other main parameters of the SOM algorithm are the learning rate alpha, which we defined to progress linearly from 0.05 to 0.01 over fifty reconfigurations (updates), and the initial size of the neighbourhood, in this instance a distance chosen in such a way that two-thirds of all distances of the map units fall within the topological extents. The neighbourhood decreases linearly during training until the algorithm reaches equilibrium. The algorithm has achieved equilibrium at ~25 iterations, meaning that no more changes to the observations' configuration were required, with the mean distance to the closest unit in the map at 11.34. Once areas were assigned to clusters, we then implemented a radar plot to map their characteristics on the basis of the input variables as we show in figure 4. This enables classes to be labelled and the following short descriptions to be created:

High Street and Promenades. These clearly depicted areas represent the main retail centres of urban regions located along the main commercial streets. This cluster also includes areas with significant pedestrianized street networks, especially along seafronts, where a lot of recreational and leisure venues can be found.

Central Business District. The area often called city centre. Typically high-rise buildings with a lot of commercial and office spaces, hence the relatively low net population density. These areas have proximity to the majority

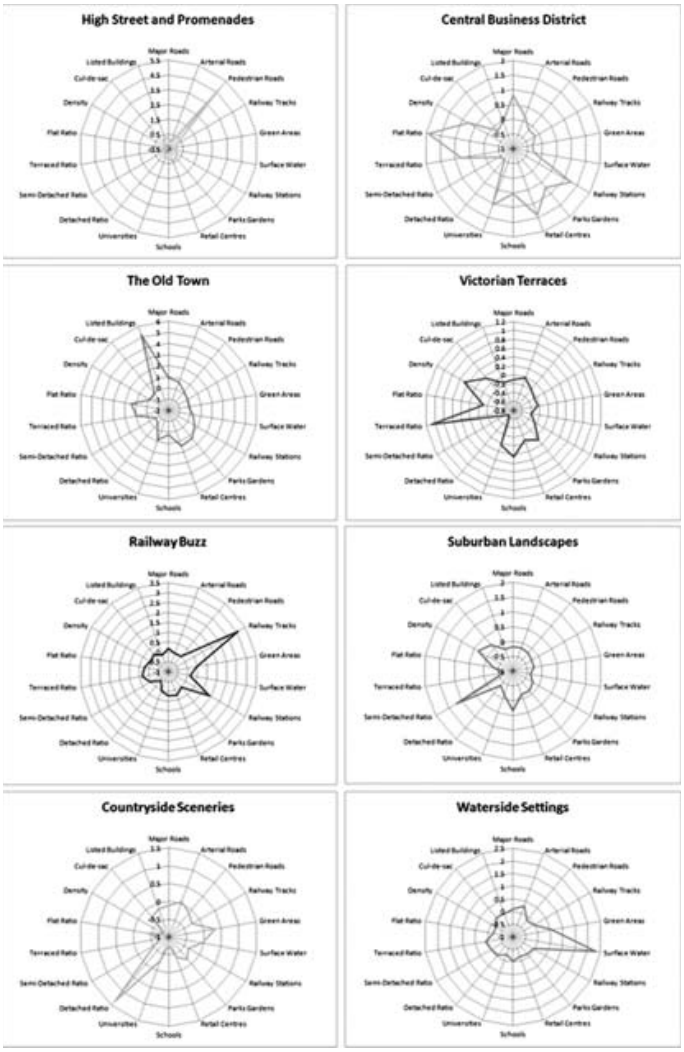

Figure 4. Final cluster results produced by the SOM, with mean attribute centres per cluster.

of public amenities, and have plenty of access via major roads and railways. For moderate-size cities the title holds true, but in areas such as London they tend to be too expansive to be labelled as central (figure 6).

The Old Town. The traditional town centre, usually close by the main high street. It is strongly defined by the amount of registered buildings. Typically a lot of recreational facilities can be found there, like pubs and restaurants, along with many administrative buildings and some historical major roads. Although it does have a considerable amount of flats, densities remain low, potentially due to refurbishments and change of usage.

Railway Buzz. These areas are dominated by railway tracks and railway stations. They have no other major distinguishing attributes which may suggest that they are actually rather heterogeneous in physical structure.

Suburban Landscapes. These areas are typically of semi-detached houses, with good access to parks. They tend to be quite distant from town centres. They are primarily residential areas, and close to schools. Culs-de-sac are relatively common, probably because of organized developments and gated communities.

Countryside Sceneries. These areas are dotted with detached houses, and are located either near or within open countryside. Most rural villages fall into this category, along with some city fringe developments that lie beyond the classic suburbs.

Waterside Settings. The principal defining attribute of these neighbourhoods is their proximity to surface water such as rivers, canals or sea. Some of these areas are ports, industrial or post-industrial sites. Distinctive infrastructure is arterial roads, i.e. roads wide enough to be used by lorries for the distribution of goods.

A Comparison of MODUM and OAC

In order to test whether the Multidimensional Open Data Urban Morphology (MODUM) classification systematically follows the conventional OAC geodemographic classification, we correlate the two sets of output classes via a contingency table. Table 3 shows the frequency distribution of MODUM within OAC 2011. *Supergroup 6. Rural residents* seems to be identified fairly well by the morphological features, with a correlation of more than 82 per cent, followed by a small percentage of *Waterside Settings* and *Suburban Landscapes*. About half the areas categorized as suburban also fall into this category, which is to be expected taking into account that typologies tend to blend out at the urban

Table 3. Contingency tables showing frequencies of OAC 2011 classes within MODUM.

	Output Area Classification 2011 – Supergroup Level								
MODUM Cluster Description	1. Rural residents	2. Cosmo-politans	3. Ethnicity central	4. Multi-cultural metro-politans	5. Urbanites	6. Suburban-ites	7. Constrained city dwellers	8. Hard-pressed living	OA Amounts
	%	%	%	%	%	%	%	%	
1. Suburban Landscapes	5.53	2.83	3.38	24.82	23.77	38.97	22.12	43.33	46,788
2. Railway Buzz	0.99	10.61	13.50	10.09	8.31	3.08	7.31	5.33	12,186
3. The Old Town	0.25	17.87	5.35	0.58	4.05	0.05	4.76	0.30	2,812
4. Victorian Terraces	1.20	14.43	16.56	43.93	24.59	1.79	39.38	34.98	49,860
5. Waterside Settings	8.43	5.03	3.56	6.98	12.08	6.73	8.04	8.82	12,468
6. Countryside Sceneries	82.45	2.05	0.43	2.91	18.89	47.79	2.14	3.90	3,172
7. High Street and Promenades	1.07	6.20	4.28	3.00	4.03	1.50	4.98	2.47	1,299
8. Central Business District	0.08	40.99	52.94	7.68	4.26	0.09	11.27	0.88	52,823
Sum (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	181,408

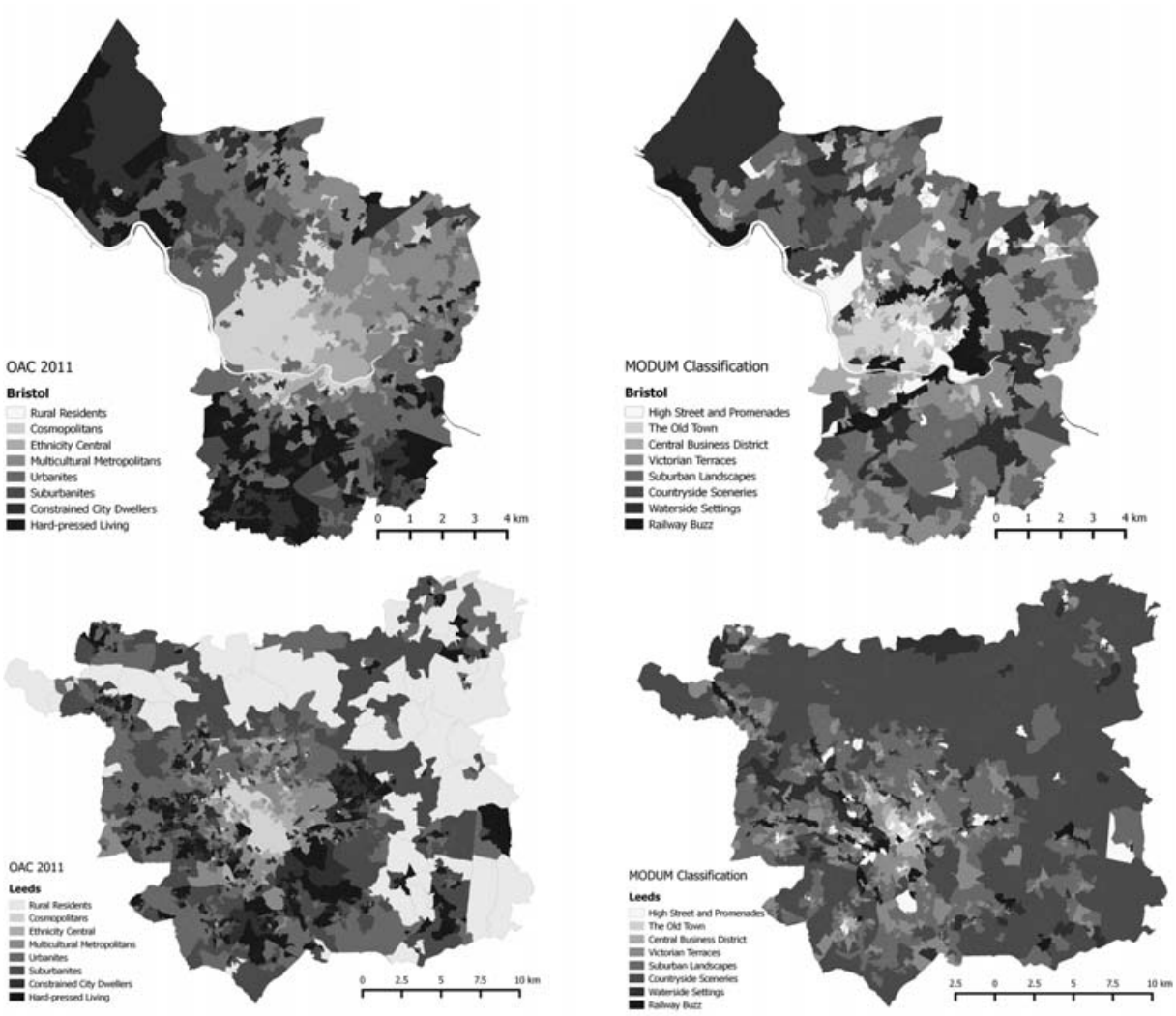

Figure 5. Built environment and socio-spatial patterns for the cities of Bristol (*top*) and Leeds (*below*). The two classifications, MODUM and OAC 2011, share many common locations, especially towards the city centre. In general, axial zones exhibit much more strongly in the morphological classification, while OAC seems to have a more 'regionalized' patterning, at least within local extents.

edges. The expansive central areas seem to be mainly populated by *Supergroup 2. Cosmopolitans* and *Supergroup 3. Ethnicity Central*. Moving out of the centre, Victorian Terraces seem to be scattered across three classes, *Supergroup 4. Multicultural Metropolitans*, *Supergroup 7. Constrained City Dwellers* and *Supergroup 8. Hard-Pressed Living*. The suburban class is most interesting, as 43 per cent of the areas classified as suburban is populated by areas identified as hard-pressed living. Generally speaking, unique classes in the MODUM classification such as the old city centre and railway-heavy areas seem to be equally dispersed among classes. Some further analysis could provide better insight as to why, and even reveal interesting patterns. Figure 5 provides two different sets of maps of the area of Bristol and Leeds, in order to demon-strate the overall pattern relationships between MODUM and OAC.

A chi-square test of the two categorical values shows that the two classifications have a significant relationship between them. We can measure the strength of the association by calculating the Cramer's V value $\varphi_c = 0.328$, which indicates an important level of association, given that φ_c can take values between 0 (no association) and 1 (complete association).

Discussion and Further Research

The development of MODUM illustrates that the production and analysis of a classification of the built environment using Big and Open Data can offer unique insights into some aspects of geodemographic structure of urban areas. The results capture, through the multi-

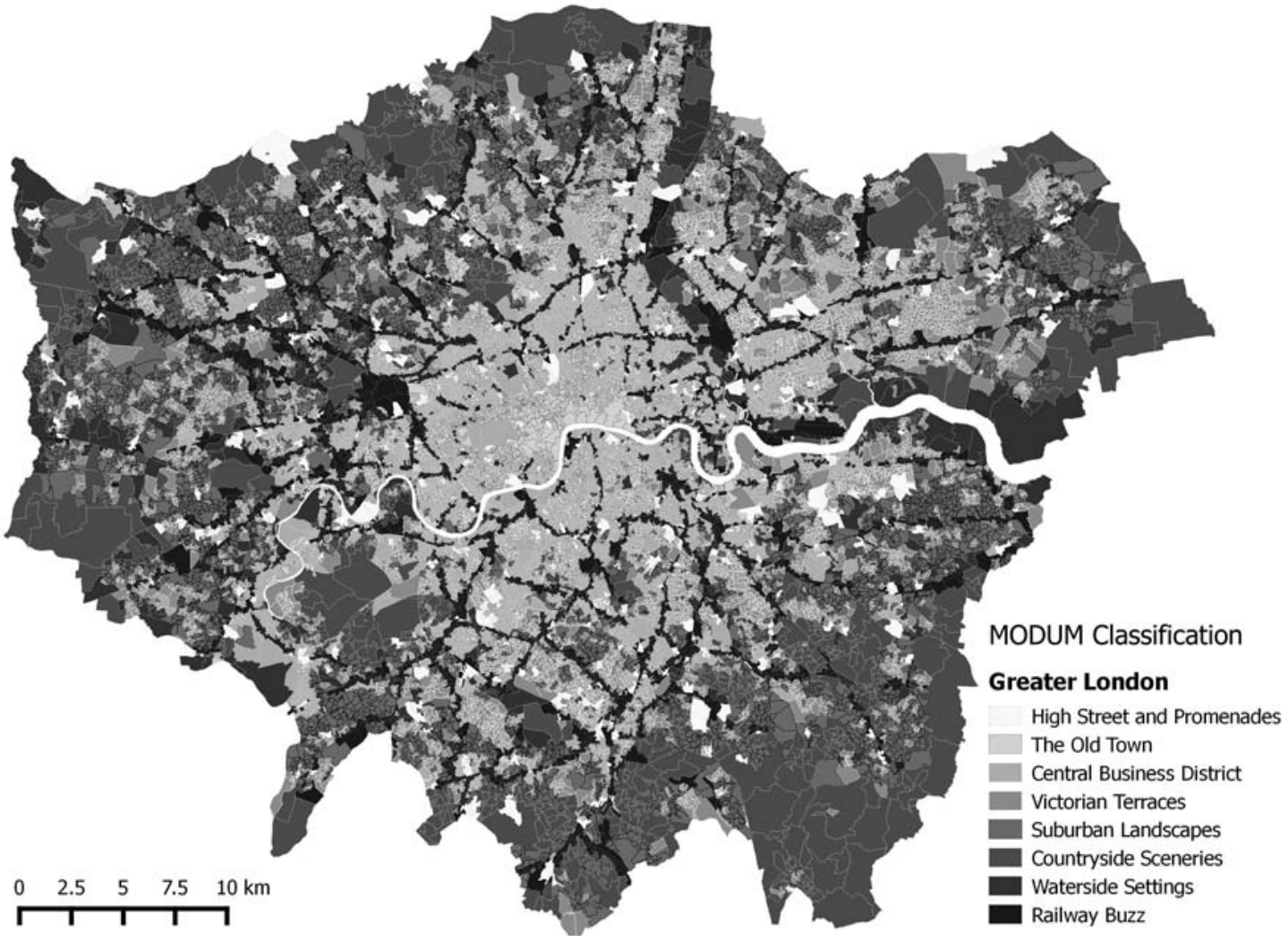

Figure 6. Mapping the MODUM classification for the Greater London Area.

dimensionality of the data, both microscopic and macroscopic identifiers of urban morphology. The classification can be used as input to more complex socioeconomic models, increasing robustness. There is strong evidence that residential preference is in significant part related to form of the built environment, suggesting that there is an important dimension to residential decisions beyond homophily. This raises some logical discrepancies in current socioeconomic geodemographic classifications; the conceptual 'control by aggregation' does not account for these unobserved variables. For instance, one would expect house prices to drop significantly very close to railway tracks. However, these localized phenomena are aggregated in the general context of the area, and thus patterns get 'smoothed away', raising some issues about the success of geo-classifications (Voas and Williamson, 2001). While gathering this type of behavioural data would be next to impossible, their outcomes can be observed through peoples' residential decisions on local morphology.

Furthermore, the MODUM classification can not only enhance socioeconomic classifications, and take into account microscopic variation, but also prove useful in itself; it can provide a simplified structure of the physical properties of geographic space that can be used to explore correlations with other spatial phenomena, potentially in a variety of applications, from real estate and house prices to health and wellbeing. In a dynamic sense, it can be used by urban planners and investors in the built environment to identify the areas in which the physical preconditions exist for neighbourhood renewal or upscaling.

On the other hand, the classification process described here is very specific to the underlying data and methodology. An inherent disadvantage of all geodemographic classifications is that lack of a single global optimization function during the classification procedure, making them highly susceptible to the operational decisions during the creation process (Openshaw and Gillard, 1978). However, geodemographics are nevertheless still valuable in many circumstances, mainly because they are practicable. Our own classification is easy to use, and offers the ability to append and update data as it becomes available, while keeping the same model infrastructure intact. In general, it meets the growing need for geodemographic systems that are open and versatile enough to handle the abundance of big data that is currently available.

REFERENCES

Arribas-Bel, D., Nijkamp, P. and Schoelten, H. (2011) Multidimensional urban sprawl in Europe: a self-organizing map approach. *Computers, Environment and Urban Systems*, **35**(4), pp. 265–275.

Arribas-Bel, D. and Schmidt, C.R. (2013) Self-organizing maps and the US urban spatial structure. *Environment and Planning B*, **40**(2), pp. 362–371.

Barbosa, O., Tratalosa, J.A., Armsworth, P.R., Davies, R.G, Fuller, R.A., Johnson, P. and Gaston, K.J. (2007) Who benefits from access to green space? A case study from Sheffield, UK. *Landscape and Urban Planning*, **83**, pp. 187–195.

Burgess, E.W. (1925) The growth of city: an introduction to a research project, in Park, R.E., Burgess, W. and McKenzie, R.D. (eds.) *The City*. Chicago, IL: University of Chicago Press.

CACI (2013) *The ACORN User Guide: The Consumer Classification.* London: CACI. Available at: http://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf.

Colwell, P., Dehring, C. and Turnbull, G. (2002) Recreation demand and residential location: the influence of sensitivity for road traffic noise on residential location: does it trigger a process of spatial selection? *Journal of Urban Economics*, **51**, pp. 418–428.

Dear, M. (2002) Los Angeles and Chicago School: invitation to debate. *City and Community*, **1**(1), pp. 5–32.

Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*, 5th ed. Chichester: Wiley.

Experian (2014) *Mosaic: The Consumer Classification Solution for Consistent Cross-Channel Marketing.* Nottingham: Experian Ltd. Available at: http://www.experian.co.uk/assets/marketing-services/brochures/mosaic_uk_brochure.pdf.

Gidlöf-Gunnarsson, A. and Öhrström, E. (2007) Noise and well-being in urban residential environments: the potential role of perceived availability to nearby green areas. *Landscape and Urban Planning*, **83**, pp. 115–126.

Harris, R., Sleight, P. and Webber, R. (2005) *Geodemographics, GIS, and Neighbourhood Targeting*. Chichester: Wiley.

Hui, E., Chau, C., Pun, L. and Law, M. (2007) Measuring the neighboring and environmental effects on residential property value: using spatial weighting matrix. *Building and Environment*, **42**(6), pp. 2333–2343.

Janson, C.G. (1980) Factorial social ecology – an attempt at summary and evaluation. *Annual Review of Sociology*, **6**, pp. 433–456.

Kim T., Horner, M.W. and Marans, R.W. (2005) Life cycle and environmental factors in selecting residential and job locations. *Housing Studies*, **20**(3), pp. 457–473.

Kohonen, T. (2001) *Self-organizing Maps*. Berlin: Springer.

Longley, P.A. (2005) Geographical information systems: a renaissance of geodemographics for public service delivery. *Progress in Human Geography*, **29**(1), pp. 57–63.

Longley, P.A. and Goodchild, M.F. (2008) The use of geodemographics to improve public service delivery, in Hartley, J., Donaldson, C., Skelcher, C. and Wallace, M. (eds.) *Managing to Improve Public Services*. Cambridge: Cambridge University Press, pp. 176–194.

Martin, D., Nolan, A. and Tranmer, M. (2001) The application of zone-design methodology in the 2001 UK Census. *Environment and Planning A*, **33**, pp. 1949–1962.

Openshaw, S. and Gillard, A.A. (1978) On the stability of a spatial classification of census enumeration district data, in Batey, P.W.S. (ed.) *Theory and Methods in Urban and Regional Analysis*. London: Pion, pp. 101–119.

Ordnance Survey (2015) *Open Map – User Guide and Technical Specification* v1.4. Crown Copyright, London: HMSO.

Parkes, A., Kearns, A. and Atkinson, R. (2002) What makes people dissatisfied with their neighborhoods? *Urban Studies*, **39**, pp. 2413–2438.

Reibel, M. (2011) Classification approaches in neighborhood research: introduction and review. *Urban Geography*, **2**(3), pp. 305–316.

Renkow, M. and Hoover, D. (2000) Commuting, migration, and rural-urban population dynamics. *Journal of Regional Science*, *40* (2), pp. 261–287.

Rijnders, E., Janssen, N.A., van Vliet, P.H. and Brunekreef, B. (2001) Personal and outdoor nitrogen dioxide concentrations in relation to degree of urbanization and traffic density. *Environmental Health Perspectives*, **109**(3), pp. 411–441.

Shevky, E. and Bell, W. (1955) *Social Area Analysis*. Stanford, CA: Stanford University Press.

Singleton, A.D. and Longley, P.A. (2009) Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, **29**(3), pp. 289–298.

Singleton, A.D. and Spielman, S.E. (2013) 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, Spielman, S.E. and Brunsdon, C. (2016) Establishing a framework for open geographic information science. *International Journal of Geographical Information Science*, **30**(8), pp. 1507–1521.

Sleight, P. (1997) *Targeting Customers: How to use Geodemographic and Lifestyle Data in your Business*. Henley-on-Thames: NTC Publications.

Spielman, S.E. and Folch, D.C. (2015) Social area analysis with self-organizing maps, in Singleton, A. and Brundson, C. (eds.) *Geocomputation: A Practical Primer*. London: Sage.

Spielman, S.E. and Thill, J.C. (2008) Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, **32**(2), pp. 110–122.

Spielman, S.E. and Singleton, A.D. (2015) Studying neighborhoods using uncertain data from the American Community Survey: a contextual approach. *Annals of the Association of American Geographers*, **105**(5), pp. 1003–1025.

Vale, D.S. (2015) Transit-oriented development, integration of land use and transport, and pedestrian accessibility: combining node-place model with pedestrian shed ratio to evaluate and classify station areas in Lisbon. *Journal of Transport Geography*, **45**, pp. 70–80.

Van Ommeren, J., Rietveld, P. and Nijkamp, P. (1999) Job moving, residential moving, and commuting: a search perspective. *Journal of Urban Economics*, **46**, pp. 230–253.

Villeneuve, P.J., Jerrett, M., Su, J.G., Burnett, R.T., Chen, H., Wheeler, A.J. and Goldberg M.S. (2012) A cohort study relating urban green space with mortality in Ontario, Canada. *Environmental Research*, **115**, pp. 51–58.

Voas, D. and Williamson, P. (2001) The diversity of diversity: a critique of geodemographic classification. *Area*, **33**(1), pp. 63–76.

Webber, R. J. (1978) Making the most of the census for strategic analysis. *Town Planning Review*, **49**(3), pp. 274–284.

Wehrens, R. and Buydens, L.M.C. (2007) Self- and super-organising maps in R: the kohonen package. *Journal of Statistical Software*, **21**(5), 23–29.

ACKNOWLEDGEMENTS

This research was funded by Economic and Social Research Council grant ES/L011840/1 (Retail Business Datasafe).

8

GEODEMOGRAPHIC ANALYSIS

Alexandros Alexiou and Alexander Singleton

Introduction

Geodemographic classification has been defined as 'the analysis of people by where they live' (Sleight, 1997: 16); it involves categorical summary measures that aim to capture the multidimensional characteristics of both built and socio-economic characteristics of small geographical areas. This chapter outlines the origins of geodemographic classifications, how they are typically constructed, and their application through an illustrative case study of Liverpool, UK.

Within sociology and geography there is a legacy of identifying aggregate socio-spatial patterns within urban areas through a variety of empirical methods. From the early 1900s onwards, researchers tried to systematically document spatial segregation and establish a series of general principles about the internal spatial and social structure of cities, commonly motivated by the ill effects of residential segregation of the poor and ethnic minorities (van Kempen, 2002). Within the UK, Charles Booth's poverty maps were one of the first attempts to map the socio-spatial structure of London in the early 1900s, although it was not until the late 1920s that the Chicago School formulated a comprehensive model of urban ecology, such as the concentric zone model of Burgess and Park (Burgess, 1925). Their research was largely based on the then recently introduced census data, alongside extensive fieldwork and map-making (Burgess, 1964: 11–13).

The analysis of detailed demographic, social and economic census data was further developed through the work of Shevky and Bell (1955). Their work introduced 'social area analysis', a methodology focused on a three-factor hypothesis that aimed to assert a typology of urban places measured in terms of urbanisation, segregation and 'social rank' (Brindley and Raine, 1979). This analytic framework inspired the adoption of a set of tools and techniques encapsulating a broader range of socio-economic census variables (Tryon, 1955; Rees, 1972), and such theoretical

approaches were later collectively known as 'factorial ecologies', due to a widening of those aspects used to explain urban structure (Janson, 1980). Factor analysis (and similarly, principal component analysis) dominated such quantitative geography in the 1970s, and was largely used to identify major underlying attributes of spatial structure, albeit with debatable results. Factorial studies were criticised not only because of their lack of theoretical context (Berry and Kasarda, 1977), but also because of their methodological weaknesses, for example their lack of extendability that contained them to being city-specific (Batey and Brown, 1995).

During this period, much scholarly concern was also focused on the interpretation and categorisation of the fundamental processes by which cities operate. In spite of the numerous attempts to classify cities per se, studies failed to find a unified theory of city typology – if such a functional typology ever existed. Classifications started to focus alternatively on smaller-area geography, and on the 'methods flowing from identification of variations of cities and following from the selection of dimensions relevant to a specific purpose' (Berry, 1972: 2). There was a common belief that typologies aid in generalisation and prediction, and urban classification was much more comprehensive when applied with a narrow scope, in terms of both area and purpose.

Within such context, geodemographics emerged in both the United States and United Kingdom during the late 1970s as an extension of those earlier empirically driven models of urban socio-spatial structure. Geodemographic classifications organise areas, typically referred to as neighbourhoods, into categories or clusters that share similarities across multiple socio-economic attributes (Singleton and Longley, 2009).

Despite a lineage of use, geodemographic classifications lack a solid theory. In nomothetic terms, many view geodemographics as methodologically unsatisfactory since the underlying theory can be considered as 'simplistic' and 'ambiguous' (Harris et al., 2005). The conceptual framework is based on a fundamental notion in social structures, homophily – the principle that people tend to be similar to their friends. This manifests spatially as a general tendency for people live in places with similar people, much like the 'birds of a feather flock together' adage suggests; and it is consistent with Tobler's first law of geography, that 'everything is related to everything else, but near things are more related than distant things' (Tobler, 1970: 236). However, one paradox is that despite geodemographic representations showing spatial autocorrelation between taxonomic groups, the methods for building geodemographics as currently construed can be considered contradictory to Tobler's statement. The central concept of geodemographics has only found limited application to the clustering processes, and not to the geographical context of each area. The aggregations of zones into categorical measures based on attributes sweeps away contextual differences between proximal zones; and as such, the final classifications assume that areas within the same cluster have the

same underlying characteristics. Standard geodemographic techniques have failed to incorporate near geography in a sophisticated way, and despite the term, geodemographics are in fact aspatial. Thus far, there have been very few attempts to build a unified framework, at least within which the relative benefits of both spatial interaction and geodemographic approaches can be maximised (see, for example, Singleton et al., 2010). For many applications, the issue of geographic sensitivity is usually experienced when normalising input variables globally and without taking into account local variation extents, thus obscuring potentially interesting local patterns. For instance, some argue that the relationship between areal typology and behaviour might not be spatially constant (Twigg et al., 2000). This type of ecological fallacy raises a series of methodological questions regarding the success of geoclassifications, given the high within-cluster variation that is already smoothed away (Voas and Williamson, 2001).

Geodemographic Classification Systems

Geodemographic analysis was initially developed as a 'strategy' that can be used to identify patterns from multidimensional census data (Webber, 1978). However, current geodemographics may use a variety of public and private data to generate profiles (Birkin, 1995). Some of the pioneering studies were applied in the UK to identify neighbourhoods suffering from deprivation (Webber, 1975). However, in the USA, geodemographics were first utilised in the private sector, as the macroeconomic conditions, alongside the freedom-of-information tradition, created an environment that quickly enabled the exploitation of census data commercially (Flowerdew and Goldstein, 1989), and the first commercial applications started appearing during the early 1980s. In the following years, geodemographic classifications gained large popularity as their utility was demonstrated across a variety of applications – from strategic marketing and retail analysis to public sector planning (Birkin, 1995; Brown et al., 2000).

Despite a common starting point, there are arguably critical differences between the UK and the USA, as geodemographics evolved through different paths. While the US classifications have typically been commercial, in the UK context there is a long history of free and more recently open classifications, and they have seen greater application in public policy and academia (for a detailed review, see Singleton and Spielman, 2014). More generally, in the UK there has been a recent renaissance of interest in geodemographics from the public sector, mainly driven by government pressure to demonstrate value for money and the advent of new application areas (Longley, 2005).

For instance, Batey and Brown (2007) developed a method of evaluating the success of area-based initiatives by using a geodemographic classification to produce spatially targeted socio-economic profiles. In this way, they assessed the efficiency

of urban policies by examining how many of the people they contain are in fact not those for whom the initiative is intended, in which case it is defined as inefficient or incomplete. Singleton (2010) and Singleton et al. (2010) explored patterns of access to higher education by linking summary measures of local neighbourhood characteristics with individual-level educational data; and through a spatial interaction framework, demonstrated the size of spatial flows between socio-economically stratified areas and institutions, with the aim that such a tool could be used by key stakeholders to examine potential policy scenarios.

Geodemographics have also been recently used in health screening, and specifically geographic epidemiology, where detailed geographical information is often unavailable. In these studies, finer geographic granularity is essential in order to produce accurate ecological estimates and infer correlations or interaction effects between health and demographics (Aveyard et al., 2002). Small-area aggregates can also be used to increase statistical power, as small-area ecological data can alleviate bias due to measurement errors in individual-level data (Jackson et al., 2006). Other notable examples include the application of geodemographics in policing (Ashby and Longley, 2005). Geodemographic analyses of local policing environments, crime profiles and police performance can provide a neighbourhood classification that is produced explicitly to reflect differing policing environments and help allocate policing resources accordingly.

The composition of geodemographic classification differs quite radically depending on the scope and probable usage by the intended stakeholders; as a result, available geodemographic products include a variety of classification systems. Among the conventional general purpose classification systems are some privately developed classifications such as the Mosaic (Experian), Acorn (CACI), P2 People and Places (Beacon Dodsworth), MyBestSegments (Nielsen) and CAMEO (EuroDirect). Commercial geodemographic systems produce discrete classes primarily designed to describe consumption patterns. Their respective databases are not only populated with census data but compiled from large consumer dynamics databases such as credit checking histories, product registrations and private surveys (Singleton and Spielman, 2014). Open classifications, on the other hand, are those that have been produced and can be accessed by the public without cost, have transparent published methodologies, and comprise freely available input data. One of the most popular open classifications available in the UK is the Output Area Classification (OAC) provided by the Office of National Statistics (see Vickers and Rees, 2007).

Building a geodemographic classification

Building a successful classification may seem fairly straightforward but it can be a difficult and very time-consuming process. It is important that a classification addresses end-user needs, but is also impacted by data availability, coverage and

potential weighting (Webber, 1977). Harris et al. (2005) provide a good basis for the methodologies typically used to build geodemographic classifications, and also provide some examples in the UK context. Vickers and Rees (2007) also provide a detailed step-by-step analysis of the process of creating the OAC geodemographic classification, which was built upon previous work on clustering methodologies by Milligan (1996) and Everitt et al. (2001). Less is known about how geodemographic classifications are built within the private sector, beyond those details usefully presented in Harris et al. (2005). Commercial geodemographic classifications have an inherent commercial confidentiality, and as suc, most of their methodologies remain a 'black box', which some have argued impairs not only reproduction, but also scientific questioning of the ways in which the clusters emerged from the underlying data (Longley, 2007; Singleton and Longley, 2009).

Scale, variable selection and evaluation

The first stage in building a geodemographic classification is to assemble a database of inputs that are deemed important for differentiating areas. The geographical unit of reference used to collate such data will depend on the purposes of the classification, and also pragmatically on those data available to the classification builder at different scales (including licencing constraints). For example, most open (and some commercial) geodemographic systems in the UK are based on data aggregated at the output area level, which represents an average population of approximately 300 people, and is the smallest scale at which census data are provided. However, different sets of variables can have different scales and there are various ways in which these are managed, ranging from simple apportionment from aggregate to disaggregate scales, small-area estimation or microsimulation (Birkin and Clarke, 2012).

From the outset (Webber, 1977), geodemographic methods have typically employed a pragmatic variable selection strategy, combining the experience of the classification builder (what is deemed to work) with the overarching purpose of a classification (what is required), alongside some degree of empirical evaluation. Attributes can be collected and compiled with a variety of measurement types including percentages, index scores, ratios or composite measures (e.g. principal components, weighting). When standardising values it is important to remember that sometimes variables have varying propensities among different groups of people, typically by age or sex (Table 8.1). For instance, long-term illness indices frequently have higher values between groups of older people. An area that has a higher ratio of older to younger people will, ceteris paribus, tend to have higher rates of illnesses as well. In these cases, age standardisation is recommended since it can scale values in accordance with age structure; scaled ratios are calculated as the sum of the age-specific rates multiplied by the area population per age group. If area specific rates are not provided, they could be obtained from the national or regional average.

TABLE 8.1 Data formatting per aerial unit

Obtaining ratios per areal unit		
Percentages	$\left(x'_{a,i}\right) = \frac{x_{a,i}}{P_a}$	where $x_{a,i}$ is the attribute value i of area a and P_a is the population of reference (denominator) of area a, i.e. total population, number of households, etc.
Standardised by group	$x'_{a,i} = \frac{x_{a,i}}{\sum_g r_{N,g} P_{a,g}}$	where $x_{a,i}$ is the attribute value i of area a, $r_{N,g}$ is the observed national ratio N for group g and $P_{a,i}$ is the population of group g in area a.

When managing quantitative data, in many cases variables will not seem appropriate to use in their raw format. Available data can have skewed distributions, contain a high rate of missing values or originate from sample sizes smaller than desired, thus generating uncertainty. In general, a detailed assessment of each variable is typical prior to the clustering process in order to identify 'unfit' data. Evaluation typically includes mapping, distribution plots (such as histograms) and correlation analysis.

A particular issue for effective cluster formation is non-normality of attributes or skew. Common techniques used to address this issue include normalisation of the variables when applicable, or weighting to adjust their influence on the final classification when normalisation is deemed by the classification builder not to be appropriate. Normalisation is the process of transforming the variable values to approximate normal distributions, usually through various power transformations. Other treatments include weighting or using principal component analysis to identify common vectors of variables that help reduce data complexity and noise (Harris et al., 2005). Table 8.2 summarises those common transformations used in geodemographics to deal with problematic data observations that are associated with the census.

TABLE 8.2 Variable transformations used for normalisation

Normalisation transformations		
Box – Cox	$x'_i = \begin{cases} \frac{x_i^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log x_i, & \text{if } \lambda = 0 \end{cases}$	The power λ achieves the best normalisation and can be estimated algorithmically
Square root transformation	$x'_i = \sqrt{x_i}$	
Log transformation* *(holds the place of zero)	$x'_i = \log x_i$	
Inverse hyperbolic sine	$x'_i = \sinh^{-1} x_i$	
Square transformation	$x'_i = x_i^2$	

Finally, a universal scale of measurement should be applied to every observation prior to clustering, such as range standardisation or standardised *z*-scores (Table 8.3), given that disproportionate measurements will frequently affect the dissimilarity function of the clustering technique towards variables with higher values. Techniques such as interquartile and interdecile range standardisation are useful when our data contain outliers (e.g. densities).

TABLE 8.3 Variable transformations used for scaling

Variable scaling		
z-scores	$z_i = \frac{x_i - \mu}{\sigma}$	
Range standardisation	$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$	
Interquartile and interdecile range standardisation	$x'_i = \frac{x_i - x_{Q2}}{x_{Q3} - x_{Q1}}$	$x'_i = \frac{x_i - x_{Q2}}{x_{90} - x_{10}}$

Clustering approaches and techniques

Clustering approaches and techniques can differ quite radically, depending not only on the purpose, but also on the nature of the data to be clustered (for a more in-depth analysis of clustering techniques, see Everitt et al., 2001; Hastie et al., 2009). A geodemographic typology is usually presented as a hierarchy; with different clusters produced for varying tiers of aggregated areas (Table 8.4). Such a hierarchy can be created from the top or the bottom. A top-down approach includes the creation of larger groups of cases that are subsequently divided into smaller subgroups. This method is typically implemented with the

TABLE 8.4 An example of a nested hierarchy for the 'blue collar communities' supergroup cluster from the 2001 Output Area Classification

Supergroup	Group	Subgroup
1: Blue collar communities	1a: Terraced blue collar	1a1: Terraced blue collar (1)
		1a2: Terraced blue collar (2)
		1a3: Terraced blue collar (3)
	1b: Younger blue collar	1b1: Younger blue collar (1)
		1b2: Younger blue collar (2)
	1c: Older blue collar	1c1: Older blue collar (1)
		1c2: Older blue collar (2)
		1c3: Older blue collar (3)

K-means clustering algorithm, and was used to produce the 2001 OAC, which included seven supergroups, which were respectively split into 21 groups and further into 52 subgroups.

A bottom-up approach is, however, more prevalent within the commercial sector, and includes the creation of numerous smaller groups (using *K*-means), which are then aggregated based on their similarities into larger groups (typically with hierarchical algorithms such as Ward's clustering).

K-means clustering uses squared Euclidean distance as a dissimilarity function, and so can be used only when variables are of a continuous measurement type. Essentially, *K*-means clustering assigns *N* observations into *K* clusters in such a way that, within each cluster, the average distance of the variable values from the cluster mean is minimised. Taking into account that for any set of observations *S* there is an argument that describes the minimum squared distance defined as

$$\bar{x}_s = \arg\min_m \sum_{i \in S} \|x_i - m\|^2$$

then for the aggregate of the total clusters there is a set of arguments that minimise the total within cluster variation of the multidimensional data points:

$$\mathrm{WCSS} = \min_c \sum_{k=1}^{K} N_k \sum_{C(i)=k} \left\|x_i - \bar{x}_k\right\|^2$$

where WCSS is the within-cluster sum of squares for a cluster distribution *C* with *K* seeds, $x_i \in \mathbb{N}$ is the data observations and $\bar{x}_k$ is the *k*-cluster mean.

K-means is typically initiated with a random set of initial seeds, and then the algorithm assigns every observation to a seed based on the least squared distance. New means based on the assignments and then calculated, and observations reassigned to their new nearest cluster mean, again based on the least squared distances. The algorithm 'converges' when the within-cluster sum of squares is minimised, i.e. when the cluster assignments no longer change. This technique is straightforward to implement and perhaps explains the popularity in the classification of multidimensional inputs; however, the *K*-means algorithm needs a specific predetermined number of clusters (*K*), and furthermore, results can differ based on the initial *k* centres that are selected. As such, it is typical to run *K*-means multiple times for an analysis, extracting the results for each converged cluster set, and evaluating them on the basis of some metric – most commonly, an effort to minimise the within sum of squares (i.e. more compact, and therefore homogeneous clusters).

In hierarchical cluster analysis, Ward's algorithm can be applied to merge clusters with the least amount of between-cluster variance, thus producing the minimum

increase in total within-cluster variance after merging (Everitt et al., 2001). Ward's clustering criterion is typically used in those geodemographics created from the bottom up to produce the more aggregate hierarchy of clusters.

Although more prevalent in research rather than commercial applications, there are multiple other clustering techniques that have been implemented within the context of area classification. A self-organising map (SOM) is an unsupervised classifier that uses a type of artificial neural network to classify space, based on the configuration of attributes that 'fit' each neuron (Skupin and Hagelman, 2005). Typically the SOM mapping process employs a lattice of squares or hexagons as the output layer, and the results are therefore easily mapped. SOMs have been tested as an alternative classifier of census data in the UK (Openshaw and Wymer, 1995) and the USA (Spielman and Thil, 2008) where they seem to perform well for socio-economic data at the census tract scale. They also have the advantage of not assuming any hypotheses regarding the nature or distribution of the data, and respond well to geographic sensitivity.

Another methodology to classify areal units is based on fuzzy logic algorithms or 'soft' classifiers. Fuzzy classifications have the inherent ability to assign spatial units to more than one cluster with varying membership values (i.e. probabilities). The degree of membership reflects the similarities or dissimilarities between groups and therefore is often addressed as a soft classifier (in contrast to hard classifiers such K-means). Most studies regarding geodemographic analysis that use fuzzy classification employ the fuzzy *C*-means algorithm or the Gustafson–Kessel algorithm (Feng and Flowerdew, 1998; Grekousis and Hatzichristos, 2012).

Other probabilistic classifiers that have been used less prevalently are multinomial logistic regression models, also known as m-logit models. A logit model has the advantages of using continuous, binary or categorical data to generate clusters, and these can also be considered as a soft classifier as they output the probability of areas belonging to each cluster category. Such models have been used in health geodemographics and epidemiology, where detailed geographical information is often unavailable so small-area aggregate data can be utilised to increase power (Jackson et al., 2006).

Cluster analysis and interpretation

The final step in building a geodemographic classification includes the review and testing of the cluster results, alongside description of the typology. For example, checking the size of clusters is one of the basic steps in the optimisation procedure. Clusters with relatively low representation of cases should generally be avoided, by either adjusting the number of clusters or by the re-evaluation of the data input. Furthermore, if, measured in terms of variance, two or more of the output clusters look very similar, merging might be considered, and inversely split if the clusters are too large. Harris et al. (2005) provides a 'rule of thumb' for merging similar

clusters, if the loss of variance within the dataset is less that 0.22%. Other ways to test an output classification is to correlate it with existing classification systems, or via sampling, such as cross-tabulation with geocoded survey data.

If the classification appears successful, a final step in interpretation is naming and describing the resulting clusters with written 'pen portraits' that best fit the profile of areas represented by the clusters. The process of creating such descriptions can be quite difficult, especially in lower hierarchies, where the cluster dissimilarities are more subtle (Vickers and Rees, 2007). Here is an extract of the profile for the 'affluent achievers' cluster from the Acorn commercial classification by CACI:

> These are some of the most financially successful people in the UK. They live in wealthy, high status rural, semi-rural and suburban areas of the country. Middle aged or older people, the 'baby-boomer' generation, predominate with many empty nesters and wealthy retired. Some neighbourhoods contain large numbers of well-off families with school age children, particularly the more suburban locations. These people live in large houses, which are usually detached with four or more bedrooms. (CACI, 2013).

Classification systems also commonly augment such descriptions with other visual materials such as photographs, maps and bar graphs or radar charts. Depending on the intended end-users, labelling and description must be selected appropriately in order to expand the user's understanding of the group, while taking into account that the end user might not be accustomed to geodemographic classifications.

Liverpool Case Study

In this final section, a practical example of creating a geodemographic classification will be presented. For this purpose, the Local Authority of Liverpool will define the extent of the classification, which includes 1584 output areas. The analysis uses the R statistical programming language, and the dataset is assembled in its entirety with 2011 census variables, provided by the Office for National Statistics and aggregated at the output area level.

Methodologically, the cluster analysis follows a similar approach to that of the 2001 OAC, although it only aims to capture broad socio-economic categories for illustrative purposes. This analysis utilises the *K*-means clustering algorithm and produces a single aggregate typological level. As a first step, consideration was required to identify those variables that would form useful inputs to the classification. Although the census includes a very wide variety of potential candidate variables, a large number of them are homogeneous across space or highly correlated. For variables to be effective in a classification they should ideally show variation over space. For instance, any variation by sex is considered to be of lower importance, since the majority of output areas have the same overall ratio of males to females.

Furthermore, given the urban location of this case study area, variables that captured dichotomies between urban and rural space might also be considered as less useful for any resulting classification.

Three elements were initially selected to guide the classification process and included demographic, housing and economic activity indicators. In total, 29 preliminary attributes were selected over the three taxonomical elements, attempting to describe the broad socio-economic profile of each output area (Table 8.5).

TABLE 8.5 Initial dataset used for the Liverpool classification

Variables	Variable Definition
Demographic	
V1: Age 0–4	Percentage of resident population aged 0–4 years
V2: Age 5–14	Percentage of resident population aged 5–14 years
V3: Age 15-24	Percentage of resident population aged 15-24 years
V4: Age 25–44	Percentage of resident population aged 25–44 years
V5: Age 45–64	Percentage of resident population aged 45–64 years
V6: Age 65+	Percentage of resident population aged 65 or more years
V7: Ethnic Group, White	Percentage of people identifying as white
V8: Ethnic Group, Black	Percentage of people identifying as black African, black Caribbean or other black
V9: Ethnic Group, Asian	Percentage of people identifying as Indian, Pakistani, Bangladeshi, Chinese or Other Asian
V10: Population Density	Number of people per hectare
Housing	
V11: Privately Owned	Percentages of households that are privately owned
V12: Rent (Private):	Percentage of households that are private sector rented accommodation
V13: Rent (Public):	Percentage of households that are public sector rented accommodation
V14: Detached	Percentage of all household spaces that are detached
V15: Semi-Detached	Percentage of all household spaces that are semi-detached
V16: Terraced	Percentage of all household spaces that are terraced
V17: Flats	Percentage of households which are flats
V18: Central heating	Percentage of occupied household spaces with central heating
V19: No central heating	Percentage of occupied household spaces without central heating

(Continued)

(Continued)

Economic Activity	
V20: Working full-time	Percentage of household representatives who are working full-time
V21: Working part-time	Percentage of household representatives who are working part-time
V22: Unemployed	Percentage of household representatives who are unemployed
V23: Retired	Percentage of household representatives who are retired
V24: Student	Percentage of household representatives who are full-time students
V25: No Qualifications	Percentage of people over 16 years without further education qualifications
V26: Higher Education	Percentage of people over 16 years for which the highest level of qualification is level 4 qualifications and above
V27: No car household	Percentage of households with no cars
V28: 1 Car household	Percentage of households with 1 car
V29: 2+ Car household	Percentage of households with 2 or more cars

The variables were each transformed into percentages, taking into account their respective denominator, with the exception of density, which was the only non-percentage variable. The next stage was to check how the variables were distributed and correlated, and assess for any that might negatively affect the clustering process. On the basis of variables with problematic distributions, these were removed form the initial dataset. Following the 2001 OAC methodology (Vickers and Rees, 2007), a log transformation was fitted to the variables to create more normal distributions. A cross-correlation table was then generated to show those variable pairs with high correlation, and a number of further attributes were selected for removal on this basis, thus aiming to

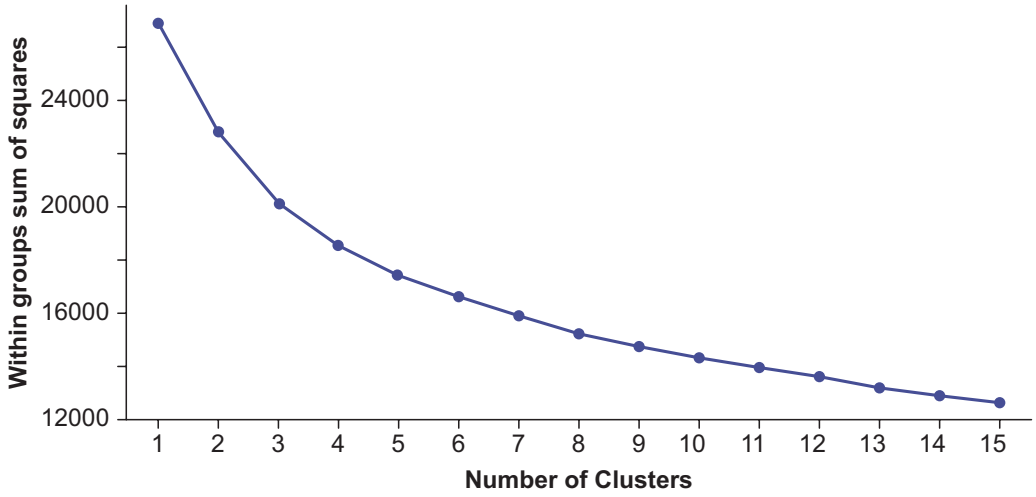

FIGURE 8.1 *K*-means: distance from mean by cluster frequency

reduce redundancy within the input data, and also limit bias towards any particular dimension being measured.

The variable selection process returned 17 variables that would form the input data to the *K*-means clustering, and were then scaled uniformly with *z*-scores. In order to address the question of how many clusters might be suitable, a within-cluster sum of squares distance graph (scree plot) was used to help identify a point at which the total distance only marginally improves the cluster homogeneity (also known as the elbow or knee criterion). However, in this case (Figure 8.1) there is no significant elbow visible, and as such, for these illustrative purposes we select $K = 5$ as a number of clusters that would be useful when mapping urban areas – increasing the classes would create a more detailed, but potentially less easily interpretable representation.

The *K*-means algorithm was subsequently run 10,000 times, and the result returning the least within-cluster total distance through these multiple iterations was extracted as the optimal result. The cluster sizes were then checked, and these varied between 72 and 522 output areas. This size variation is within acceptable limits, taking into account the limited extent of the analysis area. A useful way of obtaining information about how variables load onto each cluster is through a radar plot. Figure 8.2 shows a summary of the distribution of values within Cluster 2 (note that the Liverpool mean is 0). Cluster 2 consists mainly of neighbourhoods of middle-aged families, the majority of which are full-time workers with higher education degrees. Families are more prevalently living in low-density, detached houses, while the high

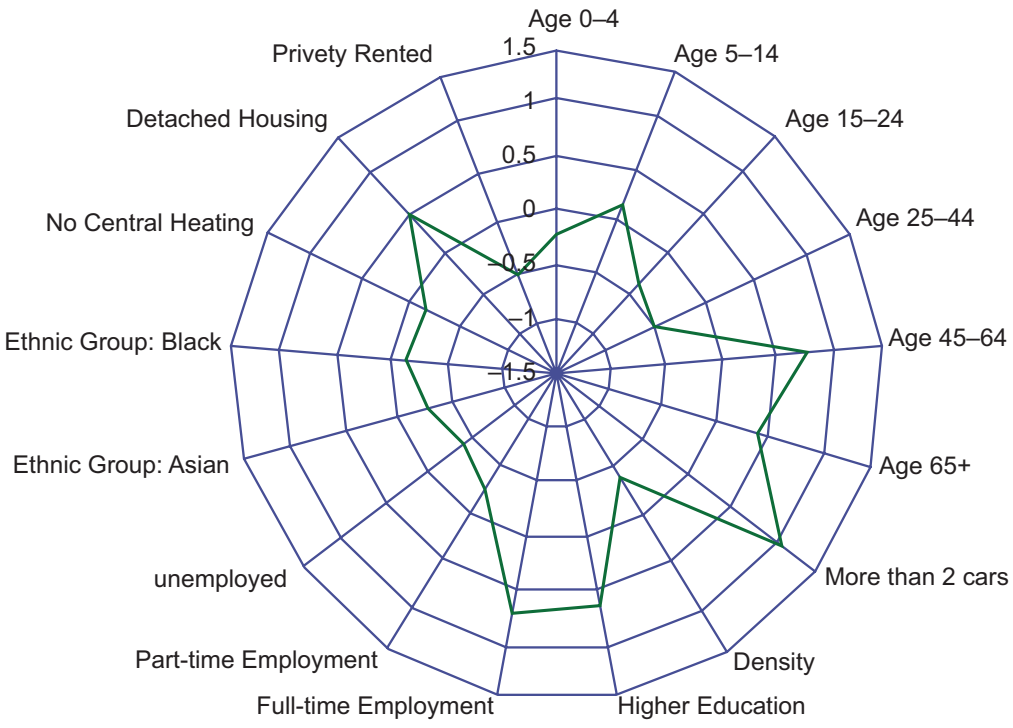

FIGURE 8.2 Within-cluster variable analysis of Cluster 2

(Continued)

(Continued)

ratio of car ownership indicates these areas may be more affluent. This cluster was named 'white collar families'. A map of the other clusters and their attributed names can be seen in Figure 8.3. As discussed earlier, patterns exhibit a degree of spatial autocorrelation, despite locational proximity being absent from the classification.

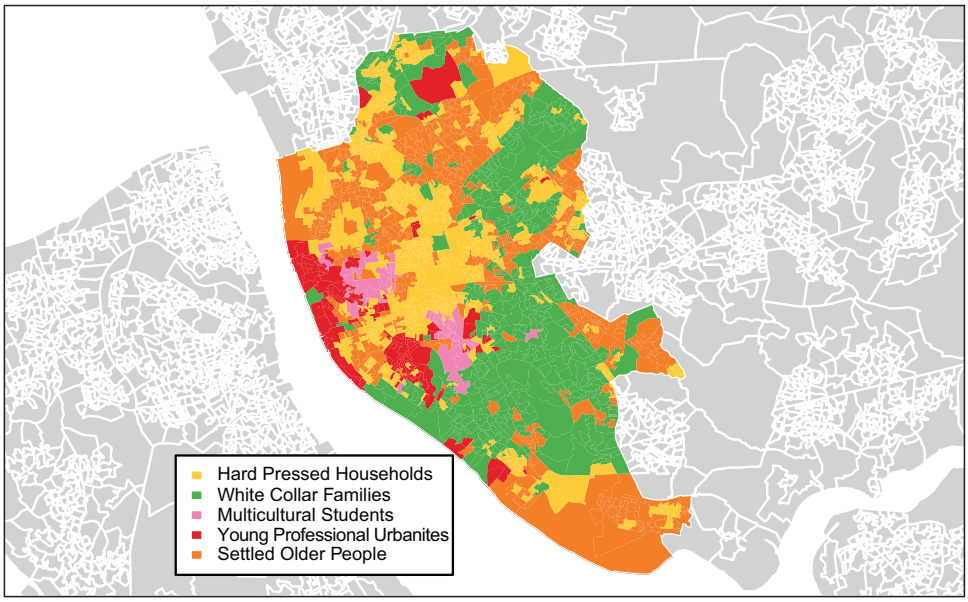

FIGURE 8.3 The final classification results, grouping the output areas of Liverpool into five clusters

Conclusions

In the previous sections we have briefly outlined the history and application of geodemographic classifications, concluding the chapter with an overview of the basic process of building a geodemographic using the case study of Liverpool. While it is true that such applications can produce reliable results, geodemographic research may face substantial challenges in the near future. Many geodemographics have historically relied on the analysis of the decennial census of the population, but institutional shifts in both the USA and UK are already changing the nature and availability of such data, given the growing costs associated with their collection (Singleton and Spielman, 2014). As such, the granularity currently offered by census data might not be readily available in the future; and as such, more research is needed into how the linkage of non-census attributes (both commercial and non-commercial) can be both validated and made more accessible.

Secondly, geographic classifications, as currently construed, do not account for spatial relations between proximal zones. This traditional 'aspatial' approach has a number of implications when generating profiles. For marketing-related applications of geodemographics, a lack of local sensitivity may have fiscal implications, such as a reduced uptake of a product or service. However, in public sector uses, the consequences may be more severe, with mistargeting having potential implications on life chances, health and wellbeing. Hitherto, methods used to take into account near geography are typically geographically crude, accounting for spatial context through either an arbitrary zonal distance, or by division of areas into administrative units that may not correspond with the organisation of actual communities. Future research is needed to produce measures of near geography that can capture such associations and evaluate these vis-à-vis traditional geodemographic models.

FURTHER READING

For an excellent introduction to geodemographics, we would highly recommend Harris et al. (2005). More generally, the majority of research articles utilising geodemographic models can be found online at https://www.zotero.org/groups/geodemographics.

The Role of Geographical Context in Building Geodemographic Classifications

Alexandros Alexiou[*1], Alexander Singleton[♦1]

[1] Department of Geography and Planning, University of Liverpool

November 7, 2014

Summary

Geodemographic analysis is a methodology that simplifies differentiated patterns of socio-economic and built environment structure for sets of small area geography. A particular issue with many current geodemographic classifications is that these lack any explicit specification of geographic context within the clustering process. Within the broad range of geodemographic applications, current techniques arguably smooth away geographic differences between proximal zones, thus limiting classification sensitivity within local contexts. This research begins to address the issue of geographic context by analyzing and evaluating various local, regional and national extents that can be used as attribute contextual weights.

KEYWORDS: Geodemographics, Geographic sensitivity, K-means

1. Introduction

Geodemographic analysis is an established methodology that can provide a simplified measure of socio-spatial structure of small area geography. Such classifications have demonstrated utility over a range of public and private sector applications (Longley, 2005; Singleton and Spielman, 2013). Geodemographic analysis typically uses the K-means clustering algorithm of multidimensional socio-economic variables. This methodological framework can capture a wide set of input attributes, taking advantage of the plethora of census variables and other geographically referenced data to generate aggregate multidimensional profiles (Harris et al., 2005).

A particular issue when constructing such classification is the way attributes are used in the clustering process. Due to the aspatial nature of the K-means clustering algorithm, geodemographic classifications account only for similarities in the clustering process and not the geographical context of each area; areas are essentially treated as independent from one another. Arguably, national aggregations could sweep away contextual differences between proximal zones, reducing the local sensitivity of classifications and thus obscuring potentially important patterns. This type of ecological fallacy raises methodological questions regarding the accuracy of geo-classifications, given the inherent loss of within-cluster variation (Voas and Williamson, 2001).

Proposed methodologies use a number of techniques to address these limitations, typically through the implementation of radial buffers for zones, and selecting attribute locational contextual measures. Although there are many national and proprietary classifications available (i.e. the OAC National Classification by the ONS, MOSAIC by Experian and ACORN by CACI), these classifications may not be suitable when assessing local patterns for policy applications. There are indicators that private classifications incorporate locational attribute sensitivity, however, underlying techniques are typically

[*] a.alexiou@liverpool.ac.uk

[♦] alex.singleton@liverpool.ac.uk

obscured and impeding thus impede reproduction, and as such, there are no established tests to their validity (Harris et al., 2005; Longley, 2007). Counter to this argument is that classifications constructed at the national, regional and local extent are effectively built for different purposes, and as such undermines comparison. This is a longstanding debate originating in the earliest of UK classifications (see Openshaw, Cullingford and Gillard, 1980 and Webber, 1980).

2. Methodology

This research uses a set of fixed input attributes for Output Area zonal geography to build classifications with different geographic extents. For this purpose, a number of scales are considered (local, regional, national) to demonstrate the impact on final classification outcome when input variables are kept constant.

Following the methodology of Harris, Sleight and Webber (2005) and Vickers and Rees (2007), a data set was assembled (Table 1) that includes demographic, economic and housing attributes of England and Wales. The dataset is assembled in its entirety with 2011 census variables, provided by the Office for National Statistics and aggregated at the Output Area (OA) level. Values where converted into percentages in accordance to their respective denominator (with the exception of *V10: Population Density*). In order to minimize the influence of certain attributes in the clustering process, highly correlated variables were later discarded at a cut-off point of 70% and above. The final remaining dataset was then normalized using a Box-Cox transformation and converted into z-scores for standardization:

$$z_{i,\alpha} = \frac{x_{i,a} - \mu_S}{\sigma_S} \tag{1}$$

where $x_{a,i}$ is the attribute value i of area a and μ_S is the mean and σ_S is the standard deviation of the observations in the dataset S. In order to measure the contextual differences between the three geographical levels, the mean and standard deviation of the OA observations for the Local, Regional and National datasets S_L, S_R, S_N where calculated, and z-scores where adjusted accordingly in equation (1). Each of the three final datasets produced where used for the clustering process in order to measure differences in classification performance.

Table 1 Initial attribute dataset used. Attributes are aggregated per OA code.

Variables	*Variable Definition*
Demographic	
V1: Age 0–4	Percentage of resident population aged 0–4 years
V2: Age 5–14	Percentage of resident population aged 5–14 years
V3: Age 15-24	Percentage of resident population aged 15-24 years
V4: Age 25–44	Percentage of resident population aged 25–44 years
V5: Age 45–64	Percentage of resident population aged 45–64 years
V6: Age 65+	Percentage of resident population aged 65 or more years
V7: Ethnic Group, White	Percentage of people identifying as white
V8: Ethnic Group, Black	Percentage of people identifying as black African, black Caribbean or other black
V9: Ethnic Group, Asian	Percentage of people identifying as Indian, Pakistani, Bangladeshi, Chinese or Other Asian
V10: Population Density	Number of people per hectare
Housing	
V11: Privately Owned	Percentages of households that are privately owned
V12: Rent (Private):	Percentage of households that are private sector rented accommodation
V13: Rent (Public):	Percentage of households that are public sector rented accommodation
V14: Detached	Percentage of all household spaces that are detached
V15: Semi-Detached	Percentage of all household spaces that are semi-detached
V16: Terraced	Percentage of all household spaces that are terraced
V17: Flats	Percentage of households which are flats
V18: Central heating	Percentage of occupied household spaces with central heating
V19: No central heating	Percentage of occupied household spaces without central heating

Economic Activity	
V20: Working full-time	Percentage of household representatives who are working full-time
V21: Working part-time	Percentage of household representatives who are working part-time
V22: Unemployed	Percentage of household representatives who are unemployed
V23: Retired	Percentage of household representatives who are retired
V24: Student	Percentage of household representatives who are full-time students
V25: No Qualifications	Percentage of people over 16 years without further education qualifications
V26: Low Qualifications	Percentage of people over 16 years with some qualifications but not a HE qualification
V27: Higher Education	Percentage of people over 16 years for which the highest level of qualification is level 4 qualifications and above
V28: No car household	Percentage of households with no cars
V29: 1 Car household	Percentage of households with 1 car
V30: 2 Car household	Percentage of households with 2 cars
V30: 3 Car household	Percentage of households with 3 cars
V31: 4+ Car household	Percentage of households with 4 or more cars
V32: NSeC - managerial	Percentage of households with an HRP with a managerial position
V33: NSeC - intermediate	Percentage of households with an HRP with an intermediate occupation
V34: NSeC - semi	Percentage of households with an HRP with a semi-routine occupation
V35: NSeC - none	Percentage of households with an HRP with no occupation

The classification methodology to produce clusters is the iterative allocation–reallocation algorithm, known as the *K-means clustering* detailed in Milligan (1996) and Everitt, Landau and Leese (2001). K-means clustering uses squared Euclidean distance as a dissimilarity function. Essentially, K-means clustering assigns n observations into K clusters in such a way that within each cluster, the average distance of the variable values from the cluster mean is minimized. For the aggregate of the total clusters there is a set of arguments that minimize the total within cluster variation of the multidimensional data points:

$$\mathrm{WCSS} = \min_{c} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2 \tag{2}$$

where WCSS is the within-cluster sum of squares for a cluster distribution C with K seeds, $x_i \in N$ is the data observations and $\bar{x}_k$ is the k cluster mean. Since the algorithm is dependent on the initial seeds, it must run multiple times in order to obtain optimal results (typically minimizing the WCSS).

Once the optimised sets of K cluster assignments are calculated for each scale of input, clusters within each set are matched in order to determine which cluster ID from one classification fits best to another. Besides the typical qualitative way, i.e. cross-tabulation of the within-cluster distribution, an algorithm was also developed that for a set of different classifications calculates the minimum absolute distance between cluster attribute means to test if this process could be used for a wider set of comparisons in the future.

If $k_i = \begin{pmatrix} \mu_1 \\ \cdots \\ \mu_n \end{pmatrix} \in K_i$ represents a vector with the average attribute values μ cluster k_i of the set K_i, then that cluster is more similar to another cluster $k_j \in K_j$, given they come from the same set of observations S, when:

$$k_i - k_j = \underset{\mu}{\operatorname{argmin}} \sum_{n} \left\| \mu_{k_i}^n - \mu_{k_j}^n \right\| \tag{3}$$

Finally, this research uses the R programming language in order to perform the analysis and map the output classifications.

3. Preliminary results and future directions

In this particular example, the Local Authority of Liverpool is considered, which contains 1584 Output Areas, and is used as a basis to compare different classification outcomes. Figure 1 demonstrates how the local, national and regional classifications are mapped within this context. Between the classifications, there are differences in the emergent cluster patterns, with the local classification appearing to offer the greatest differentiation between areas. The cluster mapped with a red colour represents the most affluent residents (e.g. "*White Collar Families*").

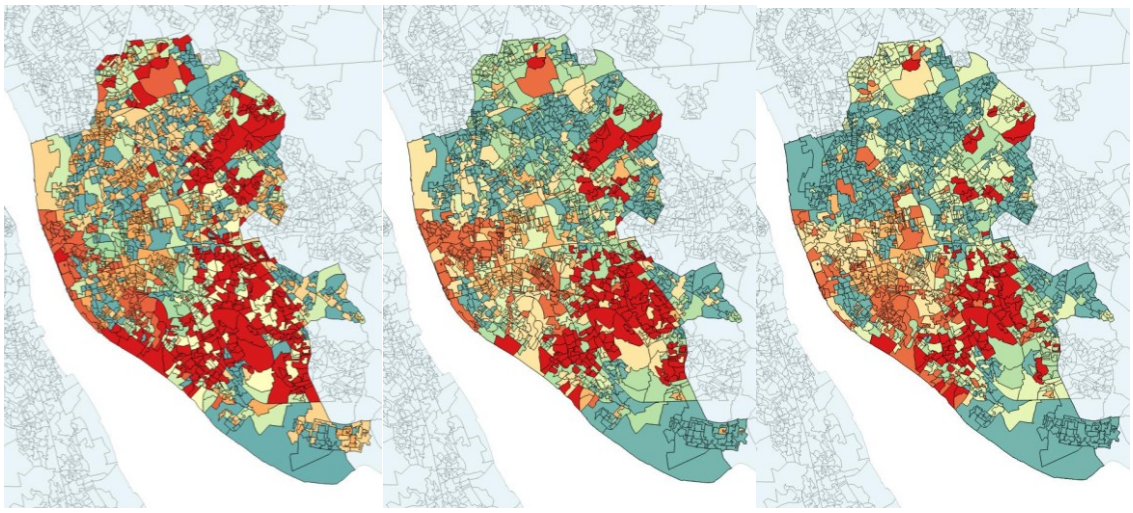

Figure 1 Differences in cluster patterns in Liverpool, UK. From left to right: Local, Regional and National geographical contexts used to calculate attribute extents.

Figure 2 shows a summary of the distribution of the values within this cluster portrayed as "*White Collar Families",* and gives an example how the developed cluster-fit algorithm works (in this case it fitted best the clusters 4, 5 and 7). It is also evident that the number of OAs in the cluster decreases as the attribute extents are scaled more globally in the case of Liverpool. For instance, an affluent family by local standards may not be as affluent by national ones. Since the Liverpool area is considered generally deprived, this number decreases from 234 OAs to 172 in the regional context and 118 in the national one.

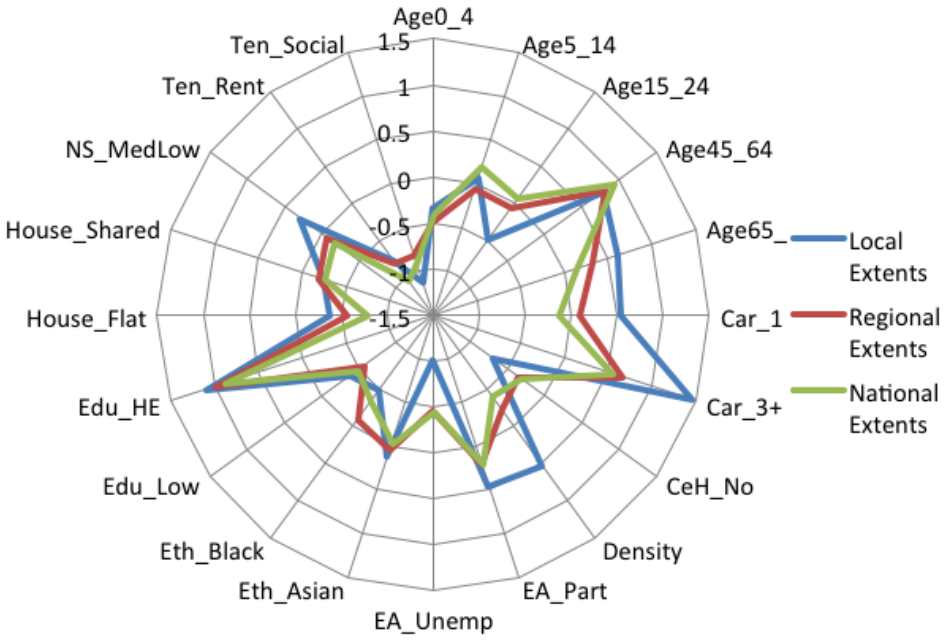

Figure 2 Distribution of average attribute values of Clusters 4, 5 and 7 (mapped red in Figure 1) for local, regional, and national extents respectively.

Although preliminary results show some degree of differentiation, a more extensive analysis is required to explore how these patterns may map between different geographic contexts, for example,

how might such patterns differ between Leeds, Liverpool, Manchester or Lancaster. Furthermore, research is needed to explore how classifications created at local or regional extents can be assembled in a way that national comparisons become possible. A challenge for future research is how these differences can be measured, and how between classifications created for different scales impacts upon the performance of the classifications when used for real world applications.

Finally, for simplicity, administrative definitions of context have been used for this study, however, we recognise that these may not represent true functional regionals or localities, and as such, further work is required about how local or regional extents might be defined, and what impact these geography will have on the final classification. In particular, at a local level, further work is also required to examine how built environment / transport infrastructure can be used to measure geographic extents and how this may impact up emergent patterns.

4. Acknowledgements

This research is part of a PhD project funded by the ESRC with an Advanced Quantitative Methods (AQM) award at the North West Doctoral Training Centre.

5. Biography

Alexandros Alexiou holds a diploma in Engineering with an MSc (Transport Planning) from the Aristotle University of Thessaloniki and an MPhil (Land Economy) from the University of Cambridge. Alexandros is currently a PhD candidate at the University of Liverpool, with research interests in the creation of new models of urban socio-spatial structure that better account for both geographic context and the dynamics of population.

Alex Singleton is a Reader in Geographic Information Science at the University of Liverpool. His research interests extend a geographic tradition of area classification and have developed a broad critique of the ways in which geodemographic methods can be refined through modern scientific approaches to data mining, geographic information science and quantitative human geography.

References

Everitt B S, Landau S and Leese M. (2001). Cluster Analysis. 4th edn. London: Arnold.

Harris R, Sleight P and Webber R (2005). Geodemographics, GIS, and Neighbourhood Targeting. Chichester: John Wiley & Sons.

Longley P A (2005). Geographical information systems: a renaissance of geodemographics for public service delivery. Progress in Human Geography 29(1): 57-63.

Longley, P. A. (2007). Some challenges to geodemographic analysis and their wider implications for the practice of GIScience. Computers, Environment and Urban Systems 31(6): 617–622.

Milligan G W (1996). Clustering validation: results and implications for applied analyses. In P. Arabie, L. J. Hubert & G. De Soete (Eds.), Clustering and Classification (Vol. 2, pp. 120-125). Singapore: World Scientific Press.

Openshaw S, Cullingford D and Gillard A (1980). A critique of the national classifications of OPCS/PRAG. Town Planning Review 51 (4): 421.

Singleton A D and Spielman S E (2013). The Past, Present and Future of Geodemographic Research in the United States and United Kingdom. The Professional Geographer.

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): 379-403.

Voas D and Williamson P (2001). The diversity of diversity: a critique of geodemographic classification. Area 33(1): 63–76.

Webber R J (1980). A response to the critique of the national classifications of OPCS/PRAG. The Town Planning Review 51 (4): 440–450.