Geodemographics: Creating a Classification at the Level of the Individual

Luke Peter Burns

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

The University of Leeds
School of Geography

February 2014

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Acknowledgements

It is important to recognise the support of several people without whom this research project would have been much more difficult and possibly unachievable. Firstly, I must thank my supervisors for their ongoing help and support throughout this project. Special thanks therefore go to Mark Birkin, Alison Heppenstall and Linda See for their valuable assistance, time and support throughout. In particular, thanks must go to Linda for continuing to act as a remote supervisor following her departure from Leeds during the early stages of my studies and Alison for maintaining communication during two periods of maternity leave.

Secondly, I would very much like to place on record my thanks to Research Support Group members Graham Clarke and Phil Rees in addition to Paul Norman (independent assessor) for the valuable nuggets of information they provided at different junctures of this research. Having such a strong, helpful and supportive group willing to provide impartial advice when needed was of enormous value. Mention must also go to the Centre for Spatial Analysis and Policy (CSAP) in the School of Geography for providing an excellent working environment in which to complete this research.

Thirdly, I would also like to pay thanks to my fellow PhD candidates, especially within CSAP, with whom I have worked closely over the past three to four years both within the department and on-the-road through conference attendance. Through research and teaching duties we have supported one another enabling each to succeed and move forward onto new challenges. This research would have been much harder without such a supportive and social group around me.

Lastly, this research would not have been possible without the generous support of the Economic and Social Research Council (ESRC). The ESRC kindly provided nominated 1+2 funding enabling me to fulfil my Masters and PhD commitments. Thanks also go to my supervisors and the School of Geography for providing the final year of financial support and tuition fees.

Abstract

This research challenges the existing geodemographics ethos by investigating the benefit to be gained from a move away from conventional areal unit categorisation to systems capable of classifying at the individual level. This research will present a unique framework through which classifications can be developed at this level of resolution. Inherently methodological, a local classification for Leeds (UK) will be presented plus further examples of this applied framework. Issues such as ecological fallacy, Modifiable Areal Unit Problem and generalisation are aspects to be considered when interpreting spatially aggregated data. A move away from such problems is one of the central objectives of this research. Data variables from the UK's 2001 Small Area Microdata file underpin this research. These variables undergo transformation from categorical states into scale variables based on gross monthly income data present in the British Household Panel Survey therefore enabling effective clustering. Micro-simulation is then employed to create an individual-level population.

The framework presented comprises entirely census variables but also demonstrates a linkage capability to other non-census datasets, such as the British Household Panel Survey (now Understanding Society), for deeper profiling, classification validation and enrichment.

Contents

List of Figures

List of Tables

List of Acronyms

ACORN	A Classification Of Residential Neighbourhoods
BCS	British Crime Survey
BHPS	British Household Panel Survey
CACI	Consolidated Analysis Center Incorporated (company)
CAS	Census Area Statistics
CASWEB	Census Area Statistics Website
CBD	Central Business District
CCSR	Centre for Census and Survey Research
CEO	Chief Executive Officer
CSAP	Centre for Spatial Analysis and Policy
CSISS	Center for Spatially Integrated Social Science
DEFRA	Department for Food and Rural Affairs
ED	Enumeration District
EFS	Expenditure and Food Survey
EPSG	European Petroleum Survey Group
ESRC	Economic and Social Research Council
ESRI	Environmental Systems Research Institute
EWN	England, Wales and Northern Ireland
GB	Great Britain
GENESIS	Generative e-Social Science for Socio-Spatial Simulation
GHS	General Household Survey
GIS	Geographical Information System
GLS	General Lifestyle Survey
GOR	Government Office Region
HRP	Household Reference Person
HSE	Health Survey for England
IBM	International Business Machines Ltd. (company)
ID	Identification (number)
IMD	Index of Multiple Deprivation
ISER	Institute for Social and Economic Research
ITV	Independent Television (UK)
LA	Local Authority
LFS	Labour Force Survey
LLTI	Limiting Long-Term Illness

LSOA	Lower Layer Super Output Area
MAUP	Modifiable Areal Unit Problem
MIMAS	Manchester Information and Associated Services
MMIS	Marketing Management Information Systems
MORI	Market & Opinion Research International Ltd. (company)
MSOA	Middle Layer Super Output Area
NP	Nondeterministic Polynomial Time
NS-SEC	National Statistics Socio-Economic Classification
NTS	National Travel Survey
OA	Output Area
OAC	Output Area Classification
ONS	Office for National Statistics
PAF	Postal Address Format
PCA	Principle Component Analysis
PLASC	Pupil Level Annual School Census
PRIZM	Potential Rating Index for Zip Marketers
SAM	Small Area Microdata
SAR	Sample of Anonomised Records
SDF	Social Democratic Federation
SEC	Socio-Economic Classification
SOA	Super Output Area
SPSS	Statistical Package for the Social Sciences (software)
SRMSE	Standardised Residual Mean Squared Error
TGI	Target Group Index
TV	Television
UK	United Kingdom of Great Britain and Northern Ireland
UKDS	UK Data Service
US	United States
USA	United States of America
WSS	Within (Cluster) Sums of Squares

Chapter 1: Introduction: Research Outline, Justification, Aims and Objectives

1.1. Brief Introduction to the Research Project

This research project builds on work undertaken in the past, from early area classifications such as Charles Booth's poverty mapping (1889), identified by Rothman (1989) as the first form of urban classification, to more recent work in the sphere of geodemographics. Such recent academic work, for example the 2001 Output Area Classification (Vickers, 2006), and certain private sector influences, including the efforts of Acxiom and Experían to classify at the household level, have acted as inspiration for this project.

Geodemographics, relative to the notion of area classification, is relatively new having surfaced as a technique within the last 40 years (Harris *et al.*, 2005). Brown (1990) states how the availability of enumeration district (ED) level data in the 1960's coupled with an interest in quantitative methods through the quantitative revolution arguably led to the sudden rise of geodemographics as a discriminatory tool, albeit initially only in the private sector. Brown (1990) points to a pragmatic shift in marketing strategy and a movement from mass marketing in the 1950's and 1960's to niche marketing in the 1970's as another reason for its rise to prominence and hence a need to identify the right type of consumer for a given good/service.

With an estimated worth (in 2003) of circa £200 million per annum, the benefits to be gained from adopting geodemography are clear and not only to business and retail industries. With geodemographics now adopted across industries as diverse as motor insurance and policing, the benefit of knowing what people in certain areas 'look like' as far as their characteristics are concerned is of tangible benefit.

With geodemographics having been firmly rooted in area-based methodologies since its inception, this research aims to challenge existing methodologies and propose alternative means of classifying populations. Geodemography is now used real-time for 'pay as you drive' motor insurance

through insurers such as Norwich Union (Osborne, 2006) and with novel datasets now being incorporated into research, such as Twitter and cell phone usage (Day, 2009), there is no reason for geodemographics to move away from its dated areal unit foundations. With the increase in computational power and availability of fine-level datasets, this work aims to be one of the first academic research projects designed to classify at the level of the person.

It is widely accepted that problems exist in any area-level classification scheme, whether geodemographic or otherwise, and these are well documented (Fotheringham and Rogerson, 2009; Openshaw 1984; Wong 1995; Greenland and Robins 1994). Issues such as ecological fallacy, the Modifiable Areal Unit problem (MAUP), and generalisation are three aspects to be considered when interpreting any spatially aggregated data. One fundamental purpose of this research is to add value to the notion of modern-day geodemographics through the creation of a system capable of discriminating at the level of the person. This research aims to prove that any system with the ability to classify populations at this finest level of detail will by far surpass existing schemes which are primarily confined to areal units.

1.2. Aim

The research adopts one over-arching aim. This is as follows:

To investigate the benefit of adopting individual-level data in geodemographic classification schemes through the creation of a framework designed to enhance existing area-based methodologies through person-level classification.

1.3. Objectives

In order to fulfil this broad aim, a series of independent objectives has been formulated. These objectives range from positioning this work amongst the existing literature to devising a sophisticated framework to enable individual-level classification. Rationale for each objective is also included below.

- ***#1 Conduct a review of the literature pertaining to (1) geodemographic classifications and (2) population generation techniques.***

The purpose of this is to investigate not only the evolution and application of geodemographics but also the tension between the aggregate and the individual. Alternative (and potentially complementary) means of population generation techniques will also be explored.

- ***#2 Present an assessment of common methodologies adopted when formulating geodemographic classification schemes.***
  This objective is crucial given the research remit of producing a framework through which individual-level classification can take place. In order to achieve this, existing approaches must be explored and means of modifying/adapting such methods considered.

- ***#3 Formulate a framework through which general-purpose individual-level geodemographic classification schemes can be generated.***
  Without doubt, a leading output from this research should be a framework through which individual-level classifications can be readily generated. This framework must be transferable to different regions and provide effective clustering methods based on individual-level datasets.

- ***#4 Apply the classification framework to two case study locations and investigate the performance relative to these.***
  In order to determine the robustness and transferability of the framework, this will be applied to two areas of differing demographic composition to ensure that the framework is able to differentiate between individual people-types.

- ***#5 Facilitate a link from the classification to other social scientific datasets for the purpose of validation and enrichment.***
  Over and above individual-level geodemographics, which alone is highly novel, the ability to append the classification codes to external datasets to allow for wider profiling represents innovation and will be presented as part of this research.

1.4. Thesis Structure

This thesis is sub-divided into nine chapters. Each chapter links to the objectives stated in section 1.3 as follows:

Chapter 1 (this chapter): **Introduction: Project Outline, Research Justification, Aims and Objectives.**

Chapter 2: Classifications and Geodemographics – From the Beginning…

This chapter presents a comprehensive review of general object classifications, area classifications and geodemographics. It includes information on evolution, formulation and applications of the above and hence addresses objective #1.

Chapter 3: Methods for Creating Realistic Synthetic Populations.

This chapter reviews several methodologies through which synthetic populations have been created and validated. This addresses objective #1.

Chapter 4: Conventional Geodemographics: A Dated Approach?

This chapter provides a detailed look at how common area-based geodemographic systems are formulated and considers factors such as variable inclusion, cluster methodologies and interpretation. This chapter (in addition to chapter 5) will consider if such approaches are applicable today and suitable for finer-level classifications. This addresses objective #2.

Chapter 5: Proposing A Framework: From Raw Data to Individual-Level Classification.

This chapter, with reference to general observations made in chapters 2 and 4, presents a detailed framework through which an individual-level geodemographic classification can be created. A novel approach to classification is put forward and one that can be regarded as the first of its kind in geodemographic academic research. This framework is one of the key outputs expected from this research. This chapter addresses objective #3.

Chapter 6: Presenting the SAM Individual-Level Classification

This chapter applies the framework set out in chapter 5 and demonstrates the extent to which it is able to differentiate between different people-types and cluster individual into homogeneous groups. This chapter addresses objective #4

Chapter 7: Linking to Microsimulated and External Datasets

This chapter extends the application of the framework through facilitating the link to both a microsimulated dataset (for complete population modelling and to aid visualisation) and other external non-census datasets (for general validation and enrichment). This chapter addresses objectives #4 and #5.

Chapter 8: Summary, Conclusions and Way Forward

This chapter completes the thesis by revisiting each of the above listed objectives and stating how each have been addressed. By re-visiting the objectives, this chapter completes the research circle and reviews the framework/outcomes generated. The relative merits of individual-level geodemographics as a discriminative tool are also discussed in addition to research limitations, extension opportunities and adopting the framework more widely.

Chapter 2: Classifications and Geodemographics – From the Beginning…

2.1. Introduction and Chapter Preface

The concept of geodemographics is relatively new, having risen to prominence from simple coarse scale census-based classifications to sophisticated address-based systems over the past 40 years (Harris *et al.*, 2005) and developed into an endemic resource in both the public and private sectors; however, the notion of area classification is more dated and has been around since the late nineteenth century. It is the work of Charles Booth (1889), as identified by Rothman (1989), which is recognised as being the first form of urban area classification. Since then, area classification has developed at a rapid rate progressing from work by Park and Burgess (1925) on Urban Ecology in Chicago and climaxing with today's multimillion pound geodemographic systems with an estimated worth of £200 million per annum - an increase of almost 90% on 1992 (Sleight, 2003). Sleight (2003) points to the growing number of end-users spending vast amounts of funds on securing these data, techniques and software for this swift rise. Mid-decade estimates suggest that the market growth of geodemographics is increasing at a rate of circa 10% per annum (*ibid)*.

This chapter will present a comprehensive review of work in the field of area classification schemes and, more specifically, geodemographic systems. The review covers a series of aspects relating to development and evolution and is structured as follows; definitions (2.2), history and origins (2.3), construction (2.4), applications (2.5), known problems associated with area-based classifications (2.6), present-day systems and their purveyors (2.7), system ethics (2.8), and recent and future directions, including a move towards individual- or household-level systems as the ultimate means of classification (2.9).

2.2. Classifications: Simple, Area and Geodemographic

The starting point of this review is the concept of general classification. The human brain embraces classification on a daily basis. The following quotation emphasises the importance of grouping analogous entities in a bid to improve comprehension;

"An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object at hand" (Pinkner, 1997, p.1, cited in Everitt *et al*., 2001).

Furthermore, Bowker and Star (1999, p.1) state how "*to classify is human*" because "*human physical abilities are limited so the amount of information provided to us is constrained by our ability to see*" (Weinberger, 2007, p.4, cited in Singleton, 2007).

For example, the human brain can, on average, distinguish between up to ten million colours through the eye, however, if the number of identifiable colours is exceeded, then the brain immediately assigns the colour in view into a comparable category; red, blue, green, etc thus making it far easier to process (Pointer and Attridge, 1998). Area classification and geodemographic schemes adopt a similar principle – that of simplifying reality. In geodemographics, this is largely achieved through capturing the socio-economic status of underlying populations through a relatively small number of groupings. In turn, groupings then describe an area's socioeconomic standing or conditions, often on a hierarchical geodemographic scale. Vickers (2006) discusses how this principle is far from new and likens area classifications in geography to physical groupings in other fields, for example elements on the periodic table being grouped by properties, or animals in a biological sense being categorised by physical features; mammals, birds, reptiles etc. A possible difference in geography is the supplementing of such methods with areal units in a bid to capture, measure and classify phenomena over space. Without geographical classification and this ability to simplify multivariate datasets, how could one distinguish between 1.8 million postcodes or 28 million individual addresses (Withnall, 2014; PAF, 2014)? In short, this would be near impossible and would most likely utilise more brain and computer power than deemed efficient.

The Dictionary of Human Geography, Fifth Edition (2009), expresses area classification as "*procedures for combining individual observations into categories*" and "*splitting a population into mutually exclusive categories on predetermined criteria*" (Gregory *et al.*, 2009, p.89). More specifically, a definition of geodemographics appears for the first time in the Fourth Edition (2000), thus highlighting the extent of recent growth and interest in the topic in the mid-to late 1980's, as emphasised by Birkin (1995). In this instance, geodemographics is depicted as *"the analysis of social and economic data in a geographical context for commercial purposes related to marketing, site selection, advertising and sales forecasting*" (Johnston *et al.*, 2000, p.297).

Definitions of geodemographics are quite varied and often open to criticism. For example, Birkin and Clarke (1998, p.88) state that "*Demography is the study of population types and their dynamics therefore geodemographics may be labelled as the study of population types and their dynamics as they vary by geographical area"*. This description emphasises the importance of population structure and that fundamentally geodemographics is concerned with the analysis of this overlying structure in relation to areal units. Debenham (2003) argues, however, that this definition is slightly misleading as geodemographics is in no way dynamic given that it is based purely upon the characteristics of an area in a single census year (or even day) and is therefore completely static in nature. In many ways, the recent provision of lifestyle data acts to lessen this over-reliance on 'static' data with such 'soft' datasets being more readily available and accessible on a more frequent basis, thus opening up the opportunity for a more dynamic approach.

Despite the small disagreements and generally wide-ranging descriptions which surround geodemographics, Rothman (1989) provides two indisputable premises. The first premise, which is strongly allied to Tobler's first law of geography (1970), states that the concept is based fundamentally on the fact that *"two people who live in the same neighbourhood are more likely to have similar characteristics than are two people chosen at random*" (p.1). Secondly, Rothman (1989) discusses how *"neighbourhoods can be categorised in terms of the characteristics of the population which they contain, and two neighbourhoods can be placed in the same category, i.e. can contain similar types of people, even though they are widely separated"* (Rothman, 1989, p.1, cited in Debenham, 2003, p.10).

If a simple and logical definition is sought, Birkin (1995, p.105) best provides this; he refers to homeland geodemographics as "*a classification of the entire UK population according to the type of area in which they live."* One could argue that this reference to geodemographics fuelled Sleight's (1997, p.16) conjecture referring to the concept as; "*the analysis of people by where they live*." The latter statement is now regarded as synonymous with geodemographics and in many ways has been adopted as the marketing techniques' tagline.

Birkin (1995) and Harris *et al.* (2005) both emphasise the broad range of indicator variables used in such systems; housing, socioeconomic and demographic characteristics, which, when supplemented by geography, reaffirm the term and its composition, hence geography plus demography and its name of '*geo-demo*-graphics'.

2.3. Evolution

As stated in section 2.2, area classification has a long history and can be dated back to as early as 1889 and the work of Charles Booth, a philanthropist and social researcher. Revolutionary at the time, Booth produced a series of coded maps reflecting social class in London with data acquired largely from visiting each street in the city (Feardon, 2007). Spatial units were defined by individual streets and these were categorised based on the number of rooms occupied per family, estimated family income and level of overcrowding (Charles Booth On-line Archive, 2002). Booth's primary goal with this assessment was to prove that past analyses of London's poverty position were exaggerated. The Social Democratic Federation (SDF) concluded that of London's total population, circa one quarter were living below the poverty line (Hyndman, 1911). What followed was far from expected. Assisted by his wife, brother and a team of researchers, Booth concluded that in excess of 30% of London's residents were living below the poverty line (CSISS, 2009), a result which exceeded the SDF's study by approximately 5%.

The final classification for Booth's assessment of London was completed by combining the notes made by both himself and his team of researchers with information from school board visitors who, over a period of time, visited all addresses which housed children of school attending age. Figure 2.1 illustrates this final classification for the St Pancras region of Greater London together with the seven groupings used to code (or classify) each individual

street. Each grouping is based upon observations made by Booth and his researchers and range from the lowest class (black shading) to the most affluent (yellow shading). The classification is therefore hierarchical - a structure many of today's leading purveyors also adopt.

BLACK: Lowest class. Vicious, semi-criminal.

DARK BLUE: Very poor, casual. Chronic want.

LIGHT BLUE: Poor. 18s. to 21s. a week for a moderate family

PURPLE: Mixed. Some comfortable others poor

PINK: Fairly comfortable. Good ordinary earnings.

RED: Middle class. Well-to-do.

YELLOW: Upper-middle and Upper classes. Wealthy.

Figure 2.1. Charles Booth Online Archive [St. Pancras, London] (2002).

In spite of Booth's seemingly early development of an area classification scheme, further notable strides in the field failed to materialise until 1925, this

time through Park and Burgess and the 'Chicago School' and work on urban ecology and the development of the concentric zone theory (first published in *The City* (Park and Burgess, 1925)). This work instigated paradigmatic interest among sociologists and geographers alike in determining the principles underpinning the spatial and social structure of cities (Harris *et al.*, 2005). Ultimately, the work culminated with features of an urban structure / system being summarised through a multidimensional classification. The Burgess model, comprising five urban 'zones', and the Hoyt model, split by a series of wedges extending from the central business district (CBD), were two such depictions (*ibid*). Figure 2.2 provides pictorial representation of these structures - both of which are inherently different and examples of both exist in the developed world.

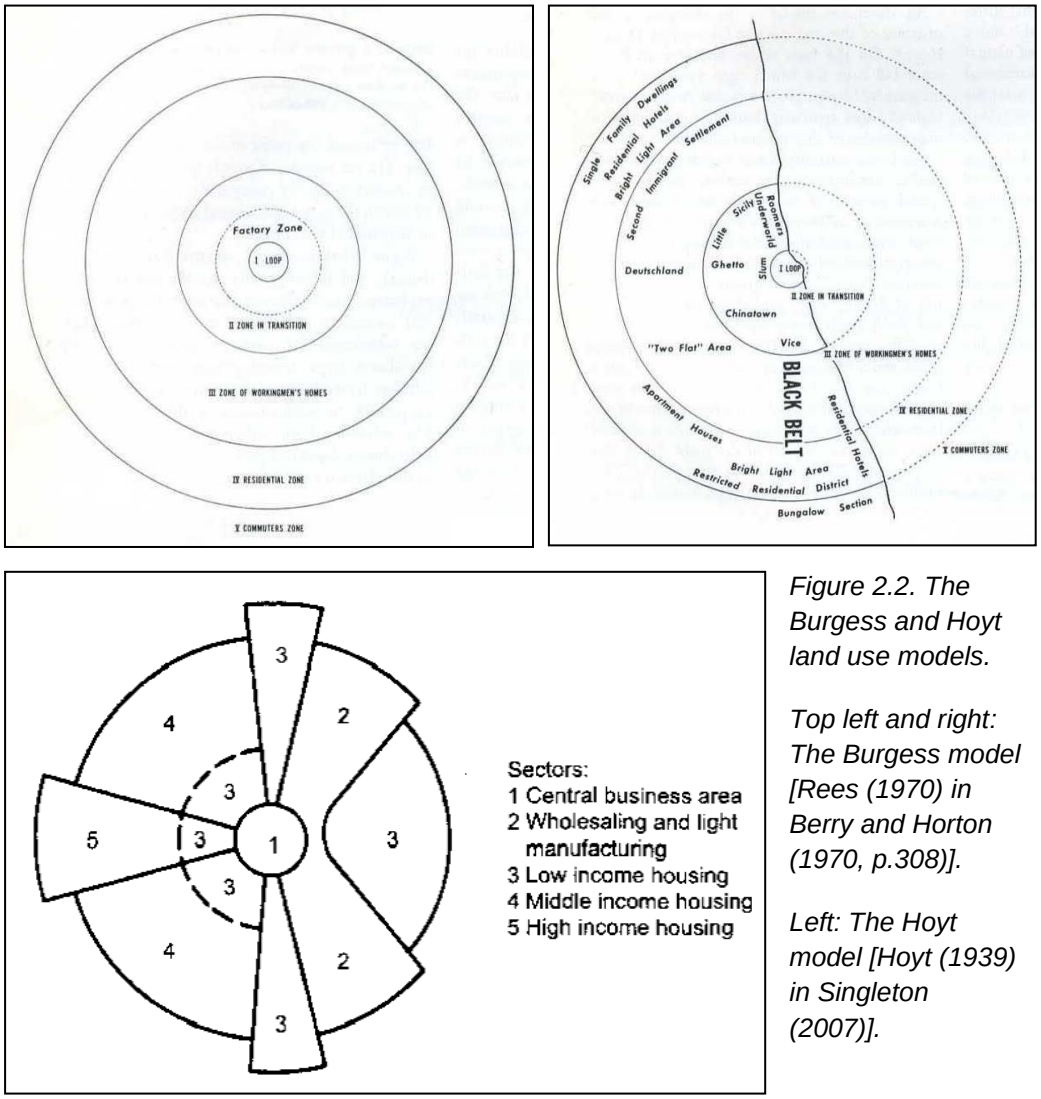


Figure 2.2. The Burgess and Hoyt land use models.

Top left and right: The Burgess model [Rees (1970) in Berry and Horton (1970, p.308)].

Left: The Hoyt model [Hoyt (1939) in Singleton (2007)].

There was then a further lull of activity until the 1960's, which was a key period in area classification history, most likely fuelled by the increasing availability of census data at Enumeration District (ED) level (*ibid*). Brown (1990) in particular argues that the surfacing of geodemographics as a discriminatory tool in the private sector corresponded with the availability of ED-level data together with a sudden interest in quantitative methods, as part of the quantitative revolution. Brown (1990) specifically highlights the move from mass marketing in the 1950's and 1960's to *'niche marketing'* in the 1970's and 1980's as a key motivator behind the rise of geodemographics. This increased the need to target the *right* type of consumer due to the more sharply segmented makeup of consumer markets (Beaumont and Inglis, 1989, cited in Brown, 1990). Rees (1970) also presents a study of city structures, this time in relation to residential and social geographical patterns in Chicago, USA. This work illustrates how both of the city depictions shown in Figure 2.2 co-exist within Chicago, the city in which the both these models were first developed.

Although Richard Webber is recognised as being the so-called father of geodemographics following his development of both the ACORN (1979) and Mosaic (1985) systems respectively (Ronson, 2005), a study by Baker *et al.* in 1979 proved to be momentous in its rapid rise to prominence. This work found that respondents in different neighbourhood types showed considerably different propensities to purchase specific products and services. The research deemed the newly created classification system (in essence, the forerunner to ACORN) to be a far more effective discriminator than methods previously adopted, for example social class (Baker *et al.*, 1979). What Baker *et al.* (1979) showed was how a simple neighbourhood classification scheme could add very useful segmentation to the target group index. Sleight (1995) attributes this work by Baker and his colleagues as being the first commercially available geodemographic classification system.

2.4. Creating a Geodemographic System

The end product of a standard geodemographic system is a simplified depiction of reality. Harris *et al.* (2005) and Gibson and See (2006) both present detailed descriptions of how geodemographic systems are formulated. Given that this chapter is fundamentally a review of past work within the field of geodemographics, knowledge of the composition and construction of a

system is unimportant. For this reason, this section will not be drawn into presenting a stepwise discussion of system formulation and instead acts as a signpost to Chapter 4 where a full explanation of how systems are built is presented. Given the cross-cutting nature of this review, mention of how systems are built is important hence this section's inclusion.

2.5. The Applications of Geodemographics

Geodemographic systems offer highly useful information capable of supplementing additional intelligence and pinpointing primary population areas. Classic uses include targeting a specific market, identifying people at risk (e.g. from health or crime cases) or identifying areas of affluence and deprivation. This does not mean, however, that a classification is of automatic benefit. The key to making effective use of geodemographics is based upon using the classification system for the most relevant applications.

In terms of previously adopted and broad application areas, CACI (1993) (cited in Birkin *et al.*, 2002, p.206) identified nine application areas for its flagship ACORN product. Modern day applications listed by CACI (2003) are almost in direct parallel to those given below.

- Site analysis
- Sales planning
- Planning for public services
- Media buying
- Database analysis
- Market research sample frames
- Direct mail
- Coding
- Door-to-door leaflet campaigns

Wallace *et al.* (1995) discuss how the call for area classification schemes, whether geodemographic or otherwise, transpired from a need for an uncomplicated and robust indicator of socio-economic information capable of contrasting the similarities and differences between small areas. In many ways, the list of application areas given above clearly demonstrates the importance and uses of socio-economic area differentiation within the private sector.

Brown (1990) discusses how the technique can also be used in retail site analysis and store location, credit scoring and target marketing, while more recently, Sleight (2004) outlines its uses more generally in retail and catchment analysis, marketing and media applications and market share studies. With regards to marketing, and in particular direct mailing, as put forward as a key application area by CACI, Openshaw (1989) argues the case for geodemographics and emphasises how the technique can add intellect to direct mailing campaigns through targeting the right kind of consumer. Even a one percent response rate is deemed a success with respect to returned surveys and questionnaires so distributing such material to the correct segment of the population in the first instance is both time and cost effective (*ibid*).

Debenham (2003, p.25) discusses the work of Birkin (1995), Clarke (1999) and Birkin *et al.* (2002) and emphasises how links must be made between geodemographics and GIS for effective retail analysis and thus effective Marketing Management Information Systems (or MMIS). Clarke (1999) best exemplifies this through the use of a grocery retailing example to show how the information in a geodemographic classification scheme can be used in combination with GIS and retail modelling to find the most profitable site for a new supermarket.

Johnson (1989), cited in Brown (1990), discusses how geodemographics can be applied when assessing the most appropriate range of products to be offered at different branches of a store and even the most appropriate arrangement of products in terms of display and shelving. Brown (1990) emphasises the usefulness of cross-referencing the Target Group Index (TGI) with one or more geodemographic systems in a bid to successfully determine the propensity of individual household types to consume a particular product or service.

Sleight (2004) presents an array of general retail, media and marketing applications of geodemographic systems, including innovative uses by ITV and Cable TV in the UK. Sleight (2004) points to the ability of advertisers to broadcast a number of advertisements simultaneously and thus select the most appropriate advert for screening at a given household as a pioneering use of geodemographics. Although such systems undoubtedly operate at the small-area level, this method provides an efficient technique for differentiating

between households most suited to viewing different types of advertisement as a means of increasing product sales through intelligence. Categorising households by small-area presents a far more effective discriminator than blanketing the entire country under one generalised or averaged perception. Despite this so-far effective use of separation through small areal units, some authors still believe it to be ineffective (Harris and Longley, 2004; Harris and Johnston, 2003). Harris and Longley (2004) criticise the use of small areas for census related applications or any other area-led analysis. The authors rightly argue that such areal units assume uniformity over space and fail to display any form of diversity, hence a forced assumption of homogeneity. Harris and Longley (2004) condemn this unavoidable supposition when making use of aggregate area-level data. However, they also introduce the possible worthwhile sacrifice of within-area diversity in a bid to enhance generalisation. Clearly, any simple depiction of reality through classification must find the optimal balance between simplicity of separation of space and diversity. One may argue that the resolution of the system needs to hold some form of meaning within the context of its overriding purpose if an area-based system is to be deemed fit-for-purpose. For example, a classification of small areas in one sense may refer to neighbourhoods whereas in a totally separate study any given small-area may be deemed representative of so-called crime zones. It is possible that an assumption of homogeneity within such zones is acceptable if the areas are indeed fit-for-purpose with respect to the final classification. Harris and Johnston (2004) discuss how electoral wards are often used in the UK as a means of describing the loose definition of neighbourhood. This is an example of a set geography being taken on board and 'made to fit' a problem, as opposed to vice versa. This is an example where making assumptions about individuals who fall into arbitrary (or easy-fit) groupings can lead to problems, thus, an individual-level classification as proposed in this research would be one step towards overcoming these problems. Further discussions on problems surrounding areal units and data dissemination are presented in Chapter 3 (Methods for Creating Realistic Synthetic Populations).

Vickers (2006) presents a brief synopsis of geodemographic usage and, although he reiterates its importance commercially by presenting some of the usages detailed above, he also mentions how such classifications are now widespread across the public sector and within the academic community and

how such systems range from general purpose classifiers to highly specific and bespoke systems. Vickers (2006) points to Rees *et al.* (1996) as a pioneering use of geodemographics within academia. In this case, Rees and colleagues made use of ONS' classification of districts to contrast rates of internal migration in the UK. With regards to public sector usage, Vickers (2006) lists a series of examples ranging from public health and crime to education. Abbas *et al.* (2009) emphasise how geodemographics is becoming increasingly recognised within the field of health, largely due to its capabilities of differentiation;

"Evidence has shown that different neighbourhood types are characterized by varying types and levels of deprivation. Accordingly, the 'one size fits all' approach will not work in public policy making. Policy makers are increasingly concerned with effective delivery of workable results at the local level. This can only be realized by appropriating differentiation strategies such that resources can be allocated effectively" (Abbas *et al.,* 2009, p.35).

Furthermore, Ashby and Longley (2005) also discuss how geodemographics can be effectively employed within the public sector, and specifically within policing, to assist attempts to ensure an equitable deployment of resources across police units. In this instance, a general purpose classification, albeit tailored to suit, by far surpasses alternative methods adopted by other authorities. Crime is just one area for the deployment of geodemographics on a public sector scale and Shepherd (2006) illustrates this through the development of a neighbourhood profiler and classification for use in community safety. Longley (2005) also provides further examples

In spite of the apparent intellect that geodemographic area classification schemes undoubtedly offer, this tool also has some fundamental flaws.

2.6. Problems Surrounding Area-Based Classifications

Making use of any area classification scheme, whether this be geodemographics, the index of multiple deprivation or any other categorisation which requires continuous space to be subdivided into a series of arbitrary features, presents problems which can often go unnoticed. Such problems include differences in observed patterns when mapping at different spatial scales or using different units (Modifiable Areal Unit Problem (MAUP)), basing assumptions about individuals on aggregate data (ecological fallacy) and

common generalisation. Elaborations on such problems common in the analysis of any spatial aggregate data are presented in sections 2.6.1 to 2.6.3.

Further problems identified by fuzzy geodemographics (both geographical and in attribute space) are discussed in Section 2.9.

2.6.1. The Modifiable Areal Unit Problem

The Modifiable Areal Unit Problem (or MAUP) was coined by Openshaw and Taylor (1979) through the assessment of correlation coefficients and, in particular, how these values can change when smaller areal units are aggregated to form larger areal units, either hierarchically or otherwise. The conclusion reached was that the coefficient can carry a range of values over different levels of spatial aggregation. Thus fundamentally, the MAUP lies with the division of artificial or ad hoc boundaries which are used to divide continuous space. Ultimately, when boundaries are drawn to demarcate space, analyses of data tabulated according to different boundaries may provide very different results (Fotheringham and Rogerson, 2009). For the purpose of this review, the concept of MAUP need only be recognised given the use of varying areal units within modern day geodemographic 'area' classifications (output areas, postcodes etc). Comprehensive reviews of the problem, including thorough discussions on the zoning and scale effects, have been written by Openshaw (1984), Wong (1995), and, more recently, Fotheringham and Rogerson (2009).

2.6.2. Ecological Fallacy

The ecological fallacy is concerned with the false assumption that relationships observed for groups also hold for individuals. Thus, any analysis conducted at area level using aggregate statistics cannot be assumed to apply to the individual level without due consideration. For example, if an area in Belfast has a majority Protestant population, this does not mean that all individuals are Protestants (Freedman, 2001). Such an inference may prove correct but is only weakly supported by aggregate-level data (*ibid*). Greenland and Robins (1994) present a broad review of this aggregate versus individual inference problem and Openshaw (1984) presents a very detailed review of how this concept applies to the collection and dissemination of census data in the UK.

More recently, Singleton (2004) states how the ecological fallacy is a general caveat to the use of geodemographic systems. He discusses how geodemographic systems naturally fall foul of ecological fallacy through the prediction of individual behaviour from variables pertaining to areal aggregations. Tranmer and Steel (1998) argue that these aggregation effects take place because individuals who live in close proximity to one another tend to exhibit similar characteristics (hence area homogeneity and a notion in geodemographics that "*birds of a feather flock together*" (Nelson, 2003, p.1)). Arguably, however, the strength of this association depends on the precise area of aggregation being examined (Martin, 1991).

2.6.3. Misrepresentation in a Classification

It is inevitable when clustering data into homogeneous groups that some areas will fit the cluster description far better than others. This is something unavoidable in geodemographics, and is a clear example of where cluster labels may not portray an area's correct image. Furthermore, Birkin (1995) provides two examples where cluster labels are not always indicative of continuous populations. For the 'SuperProfiles Lifestyle' classification, Birkin points to clusters labelled "*Young Married Suburbia*" and "*Metro Singles*" and emphasises how these names are more than slightly misleading. For the former cluster, this grouping in fact accounts for over one quarter of the population whose age is 45 plus. Meanwhile, for the latter named cluster, this category encompasses only 21% of single workers – rather misleading when you consider the cluster label contains the words "metro" and "single". The potential error with regards to targeting the wrong type of consumer is substantial and is in many ways fuelled by a desire to pursue traditional classification methods in standard geodemographic systems. Chapter 4 provides further evidence of this inaccuracy within conventional area-based geodemographics.

All of the above issues could be improved upon by implementing a geodemographic classification that classifies at the individual / household level as opposed to an areal unit. A classification of this nature would (1) alleviate problems surrounding MAUP as the spatial scale in use would most likely be as fine as is achievable, (2) reduce the need to generalise and infer characteristics about persons based on their neighbours (hence, ecological fallacy) and, (3) enable cluster labels to be far more accurate in describing

what would be individuals as opposed to collections of people with vaguely similar traits.

One point to bear in mind, however, is the importance of neighbourhood. Drilling down to the lowest geography (or person / household) is not always the optimal solution as emphasised by Harrop and Heath (1991). In this article, the authors discovered how voting behaviour was impacted upon by locality – and how this had a far greater bearing on behaviour than household. This implies that that the purpose to which such a proposed individual-based classification is applied must be carefully considered.

2.7. Today's Systems and Their Purveyors

The ACORN and Mosaic systems, as developed by Webber during his time at now-named CACI and Experían, are just two of the geodemographic classifications available commercially today. According to Birkin (1995), these two systems remain the most important tools for planning and business in the UK.

As with most current systems, CACI's ACORN and Experían's Mosaic systems follow a hierarchical structure. ACORN operates with six categories, seventeen groups and fifty-six types (CACI, 2003) whereas Experían's Mosaic functions with 11 groups, 61 types and 243 segments (Experían, 2007). The similarities, however, lie in the vast array of consumer variables used for classification; ACORN alone uses over 125 demographic statistics and 257 lifestyle variables (CACI, 2003). It is the more recent adoption of lifestyle variables that has revolutionised geodemographics and lessened its over-reliance on census, and hence more static, data.

The following four subsections provide more information on individual schemes and in particular detail three leading commercial and one academic geodemographic classification system. It is important to note that each of the systems adopt *'crisp technology'* thus allowing for a one-to-one mapping of areas to cluster types with no overlap or exception (Birkin *et al.*, 2002).

An additional point should also be raised prior to presenting the following product descriptions. Such is the nature of private sector geodemographics (thus, three of the four systems which follow) and the lack of methodological transparency which surrounds these systems, any second-hand explanations can only be as detailed as the information available in the public domain.

Unfortunately, an inability to penetrate this secrecy prohibits detailed descriptions of the data, methodologies and processes which surround ACORN, Cameo and Mosaic. The OAC, on the other hand, is of complete contrast with fully accessible documentation available given its total transparency. Arguably, only the Output Area Classification (OAC) passes the test of scientific rigour for this reason.

2.7.1. ACORN (CACI)

The name ACORN is an acronym for the description "**A C**lassification **O**f **R**esidential **N**eighbourhoods." CACI Online (2003) defines the ACORN system as *"…the leading geodemographic tool used to identify and understand the UK population and the demand for products and services.*"

ACORN was first developed in 1979 as CACI moved from its 'SITE' census analysis system on a bureau basis to the more up-to-date ACORN classification system. ACORN is still regarded as the company's flagship product over thirty years hence and is now in its fourth version (Sleight, 2003).

The 2001 system operates at postcode level, having previously functioned at enumeration district then output area level during its embryonic stages. The system currently combines six categories, seventeen groups and fifty-six types to form a postcode level area classification. The system classifies all 2.1 million UK postcodes into categories, groups and types using in excess of 125 demographic variables and 287 lifestyle indicators across the UK (CACI, 2003). This classification thus encompasses each of the UK's 28 million addresses, albeit based on an area-level classification. Until 2000, ACORN was purely a census-based classification tool and it was not until after this period that lifestyle and market research data were added to enrich the classification. At present, census, income, house price, shareholdings, lifestyle surveys, electoral roll, PAF and neighbourhood statistics are all employed in the classification process (Sleight, 2003). CACI's classification procedure is two-fold; firstly census output areas are clustered before postcode level data are added at stage two with data not matching the OA classification being reassigned using a complex best-fit algorithm (*ibid*).

The three-tier hierarchical nature of the ACORN system means that the most affluent population groups are generally categorised in the upper reaches of the classification, for example '*Wealthy Achievers'* being the top category, with

this then scaling down to those less prosperous groups, such as '*Moderate Means'* and '*Hard Pressed'*. This makes for an easily understandable and interpretable system for businesses, planning practitioners, and any further persons looking to embrace this marketing product.

The success of the general-purpose ACORN system in particular and the continued uptake of geodemographics commercially led CACI to develop more specialised products; Financial ACORN and Health ACORN. These new products now sit firmly alongside the company's flagship classification and highlight the ever-growing demand for products capable of differentiating consumers and, as a result, enabling effective consumer targeting.

2.7.2. CAMEO (EuroDirect)

The CAMEO system (previously named Neighbours & Prospects) is described by EuroDirect (2006) as *"A hugely powerful and well-established consumer geodemographic classification developed for the analysis and targeting of UK consumers."* Previous versions of the CAMEO system adopted purely census-based variables but more recent versions have spanned out to incorporate a wide range of consumer and lifestyle variables. Current datasets inputted into the classification include consumer credit data (in conjunction with sister organisation, Callcredit), Household Council Tax Band and Property Valuation Data and individual shareholder data. Such datasets, EuroDirect (2006) argue, introduces the *wealth* aspect not covered by the Census. EuroDirect (2006) also claim that the fusion of individual and household-level data enables a methodical differentiation of UK postcodes (*ibid*).

The system classifies small areas based what it calls *"ten key marketing groups" (ibid)* and fifty-seven neighbourhood types and is an example of a two-tier non-hierarchical scheme.

2.7.3. Mosaic (Experían)

Experían Online (2007) advertise the Mosaic system as a product that *"provides decision makers with the tools and services they need to successfully implement micromarketing strategies within their business."* The latest system, Mosaic UK, is Experían's most advanced Mosaic area classification tool that covers the whole of the United Kingdom. It is

Experían's third British version and builds on the first Mosaic as launched in 1986 under the influence of Webber (Sleight, 2003).

The system cuts the population into 11 groups and 61 types and is thus another two-tier classification system. A key principle regarding the Mosaic system is that it operates at both household and postcode level (*ibid)*.

The system comprises 400 data variables, 54% from the 2001 census and the remaining 46% derived directly from Experían's consumer segmentation database providing data on all of the UK's 46 million adult residents and 23 million households (Experían, 2007). Additional datasets incorporated into the classification include house price and tax information, ONS local area statistics and the edited electoral roll. With the exception of the 2001 census data, Experían endeavours to update all its data annually to maintain an advanced classification scheme *(ibid)*. The company's classification development methods are given added value through links to the TGI, the British Crime Survey, MORI Financial Research and Forrester Technographics & Internet User Monitor (Sleight, 2003).

The system categorises postcodes into groupings from "*Symbols of Success*" to "*Rural isolation*" and unlike ACORN the categories do not appear to fall into a hierarchical pyramid of affluence but are juxtaposed together. Experían adopted a 'bottom up' approach to clustering, beginning with collating residents and household-level data before combining those datasets with higher levels of geography, namely postcode and output areas to form the classification (*ibid).*

Mosaic is the only member of the leading group of systems thus far to attempt household-level classifications and therefore the functioning at this level is a key differentiator between Mosaic and its rivals. In a recent brochure, Experían (2009) discuss how 62% of the information used to build the Mosaic UK system was sourced from privacy-compliant datasets. The leading dataset referred to is Experían's own Consumer Dynamics Database which is formulated based on a range of datasets including personally completed surveys and externally purchased information such as house sale prices. This move towards household-level classification is very much in its infancy in the commercial sector.

Mosaic's classification at the household-level appears to have advanced further in the Netherlands with a 'Mosaic Household' brochure now available (Experían Netherlands, 2013). This classification segments households into one of forty-four types (across ten groups) and was built under the guidance of Professor Richard Webber from University College London. The classification comprises household types such as 'Conservative Students', 'Homes for the Elderly' and 'Young in Apartments'. The UK-version comprises ten leading groups and is evolving with Warwickshire local authority currently making use of the system (Warwickshire Observatory, 2011).

2.7.4. Output Area Classification (ONS)

The Output Area Classification (OAC) is a non-commercial geodemographic classification scheme which categorises at output area level. Initially ninety-one census variables were selected to comprise the scheme but this was later reduced to forty-one through rejection and merging of indicator variables (Rees *et al.*, 2005). The input variables used span demographic, household, socio-economic and employment data and create 9,145,460 individual data points (*ibid*). The final classification encompasses seven super groups, twenty-one standard groups and fifty-two subgroups all based purely on census data. The difference with this system is that the third-level of classification is not named (or given a pen-portrait description). It was claimed that the time and effort needed for this process was not justified (Vickers & Rees, 2006). The scheme employs census data only and differs from more commercially dominated systems both in terms of its output level and a lack of 'soft' variables which strongly relate to lifestyle / wealth. GB Profiles is another example of a similar public sector geodemographic classification scheme.

2.7.5. Other Systems

Smaller companies, such as Beacon Dodsworth and Claritas, have also developed areal unit segmentation systems such as the P2 People & Places Scheme and SuperProfiles respectively, the latter now discontinued. The P2 system adopts 14 Trees, subdivided into 41 Branches, with a lower level made up of 157 Leaves and again follows the standard structure for hierarchical geodemographic segmentation (Beacon Dodsworth, 2007).

Other companies with post-2001 geodemographic systems include; AFD Software Ltd., Allegram Ltd., The Clockworks, Claritas (with PRIZM), MapInfo

Predictive Analytics (previously GeoBusiness Solutions Ltd, now rebranded as Pitney Bowes), Intermediary Systems Ltd. and Streetwise Analytics Ltd. (Sleight, 2003). Acxiom have also developed the PersonicX system and this will be further discussed in section 2.9.

Table 2.1 provides a summary of the four leading geodemographic systems currently available.

System	Purveyor	Divisions			Variables		Finest Level of Operation
		Tier 1	Tier 2	Tier 3	Census	Other	
ACORN	CACI	5	17	56	120	280	Unit Postcode
CAMEO UK	EuroDirect	10	57	-	116 (Census to Other variable ratio unknown)		Unit Postcode
Mosaic UK	Experían	11	61	-	216	184	Unit Postcode
OAC	ONS	7	21	52	41	-	Output Area

Table 2.1. Overview of four leading geodemographic systems (CACI, 2003; EuroDirect, 2006; Experían, 2007; Rees et al., 2005).

Table 2.1 clearly illustrates the differences between each of the four systems. Although inherently business tools, ACORN, Cameo UK and Mosaic UK are widely embraced in other sectors, such as local government and public sector planning (Birmingham City Council, 2010). Although very little research has been undertaken contrasting the effectiveness of each system, Brown (1990), Leventhal (1995) and Voas and Williamson (2000) do provide some assessment of systems available at the times of writing but fail to conclude on any obvious superiority. Leventhal (1995), in particular, states that *"…no single classification outperformed all others and that the differences between them were generally small"* (p.8). It therefore fails to be proven if a higher number of clusters, greater number of variables or mix between census and behavioural inputs lead to more discriminative systems.

As can be seen in Table 2.1, each of the leading systems widely embraced today segment the population based on aggregate data and hence into distinct geographical areas (such as output areas). The exception is in the case of postcode classifications albeit with uses almost exclusively for business

applications. This doctoral research will extend such methodologies such that individual-level classifications are possible and, given the transparent framework that will emerge, application-areas beyond the business sector can benefit.

2.8. Issues Pertaining to Privacy and Ethics

The purveyors of geodemographic systems claim that *"if you're trying to find a person with particular attributes, we can point you to his doorbell"* (Hill [CEO Experían], 1990, cited in Roberts, 1992, p.26), and that, if you *"tell me someone's zip code ... I can predict what they eat, drink, drive or even think"* (Robin, 1980, cited in Weiss, 1988, p.1). Understandably, such suggestions have fuelled the imagination of various companies with the need to target consumers and although this has propelled the geodemographic industry to its standpoint today, one may point to matters of privacy and ethics.

It is often the case that when devising any form of classification, whether this be of areas, households or individuals, there is the assumption that reality can be precisely portrayed by the typologies which describe such areas or individuals. Singleton (2007) points to the dangers that such assumptions can bring, particularly when adopting geodemographics in the public sector where the application of such a technique may directly impact upon the life chances of those classified – either accurately or otherwise. Although Singleton fails to mention any areas within the public sector where geodemographics has resulted in disadvantage, education, health and crime are potential areas of worry. It is also foreseeable that life chances may also be affected by the use of area-level geodemographics within the private sector, perhaps most notably through house prices and the classification of so-called ‘desirable’ and ‘less desirable’ residential neighbourhoods. Other private sector dominated life chances include insurance (house or car), life assurance and eligibility to bank accounts/loans, etc (Vickers, 2006). With respect to the latter, Levene (1999) writes how an unnamed bank stipulated that customers from ‘less desirable’ postcodes needed to source and deposit larger initial sums of money when first opening a bank account when compared to customers from a preferable geodemographic group. Sui (1998) discusses how such classifications can cause harm, particularly when the social position of a researcher or system purveyor is totally independent from the research he/she generates. Singleton (2007) embraces this statement and implies that data-led empirical

investigations, such as geodemographics, may not always be adequate when attempting to represent complex and highly dynamic real world social processes.

When you consider statements such as the following two, it is no surprise that geodemographic classification schemes are criticised for their ethical functioning:

[1] *"We and a couple of hundred other companies are going to appropriate your name, match it, store it, rent it, swap it; we'll evaluate your geodemographic profile, determine your ethnic heritage, calculate your propensity to consume. We'll track you the rest of your consuming life, pitch you baby toys when you're pregnant, condos when you're fifty. In return for the use of your name, we won't pay you a penny."* (Larson, 1992, cited in Goss 1995, p.178).

[2] *"Using caller ID to call up a postcode, company call centres can move you to the front of a queue if they think you are more likely to buy their goods, divert your call to a call centre in India, or let you hang on if you are likely to sap their profits."* (Highfield and Fleming, 2007).

It is surprising that no regulatory body exists within this area capable of monitoring geodemographics, both in terms of composition and application, to ensure that it operates within the bounds of ethical principles.

Goss (1995) is one strong critic of geodemographics, largely on two fronts. Firstly, he states that it is often the case that simple misspecification in a database can unconsciously discriminate even if the use of the data are legitimate in lawful terms. Goss' (2005) second privacy concern, and perhaps the one which holds most weight when data collection is considered, concerns data being fit for purpose. He states how it is often the case in geodemographics that data collected for one purpose are then re-used for another without the permission of the data subject. Singleton (2007) elaborates on this and discusses how "off the shelf" systems are formulated using only legally available data. However, Goss' (2005) second concern only transpires when such systems append to external data sources. This issue of data matching is also a problem flagged by Curry (1997).

2.9. Innovations in Geodemographics

Before exploring the possible future of geodemographics, it would be sensible to assess any key innovations to have taken place since the inception of the technique in 1979. Arguably, the primary shift in emphasis arrived courtesy of Openshaw (1989) and his suggested movement from Boolean to fuzzy geodemographic clustering. In short, fuzziness is concerned with uncertainly which transpires from imprecision and ambiguity and, within geodemographics, there are two different types (Feng and Flowerdew, 1998). The first kind of fuzziness is in attribute space and this refers to small areas being classified within one grouping but residing very closely within the taxonomic space to one or more others (*ibid*). An example is one area, say an output area, having only marginally different characteristics to one another but being categorised into a very different cluster based on the simplicity of the algorithm and a Boolean approach to clustering.

The second form of fuzziness relates to geographical fuzziness, and this can be further split into two types. The first is concerned with linking postcodes to census geographies. This linkage problem arises from different causes and results in the associated error that incorrect postcodes are included and correct postcodes excluded from a cluster (Openshaw, 1989 cited in Feng and Flowerdew, 1998). Perhaps more importantly in fuzzy terms, the second form of geographical fuzziness is that of the neighbourhood effect. Given that census or postal geography boundaries are by no means an accurate disaggregation of populations based on socioeconomic conditions, as emphasised by Morphet (1993), one may argue that residents from a neighbouring area to that formally classified are rather likely to possess the same (or similar) lifestyle traits to residents classified in the target area (Feng and Flowerdew, 1998). Openshaw (1989) stresses how, based on Boolean geodemographics, such neighbourhood effects are simply overlooked.

Considering the above notion of fuzziness, it can be argued that geodemographics has seen two phases in its development to date, that of conventional Boolean area-level geodemographics as we know it and that of fuzzy geodemographics. This research seeks to take geodemographics into its third development phase and ensures that many of the issues discussed in previous sections are considered.

There has been much written on the projected path of geodemographics into the future, including work by Debenham (2003) on integrating more supply-side variables and Singleton and Longley's (2009) research into more innovative geodemographic visualisation techniques and real-time or on-the-fly type systems – the latter of which has already been trialled (and recently discontinued) through 'pay as you drive' car insurance by Norwich Union (Osborne, 2006). However, in many ways the definitive level classification will be one capable of operating at the finest resolution, that of the household or, ultimately, the individual.

The idea of classifying households is not a new one. Both Claritas and Acxiom have developed and utilised systems capable of classifying individual households, and, at present, this remains the most extreme level of geodemographic disaggregation. The two systems, PRIZM Household and PersonicX respectively, are described in more details in the sections that follow. Again, however, all descriptions are based only on information the vendors are happy for the customer to be made aware of – often via enticing brochures or advertising literature.

2.9.1. PersonicX Household (v2.1)

PersonicX Household version 2.1 is a consumer segmentation system that classifies each UK household into one of one hundred and fifty micro-clusters and then fifty-two PersonicX clusters (Acxiom, 2009). The system, initially proposed in 2004, comprises in excess of five hundred variables collected from twenty-five million households which in turn feed into the clustering algorithm and develop each of the cluster types, for example: Just Retired, High Flying Solos, Rich Returner's etc (*ibid*). Variables are largely behaviour orientated and include; hobbies, car ownership, internet usage, credit card usage, TV, education, financial products, newspaper readership, grocery shopping, residence type, mobile phone usage, charity donations etc (*ibid*).

What Acxiom omits from their marketing material is that a large proportion of the data which are input into the system is in fact modelled or simulated. The exact proportion is unknown. However, estimates suggest that as much as 30% of the variables are modelled so as to ensure 100% coverage (Bradbrook, 2009). The remaining 70% comes as a result of the many

questionnaires issued by the marketing company at various intervals per year. This idea of simulation is discussed by Farr and Webber (2001) with respect to the division of an individual-level classification.

A brief methodology of how the system is constructed, together with detailed cluster descriptions and illustrations, can be found in Acxiom (2009).

2.9.2. PRIZM Household

Information on Claritas' PRIZM Household system is far less available than Acxiom's equivalent, perhaps largely due to its dated nature. However, standard non-methodological information is available publicly. Similar to all other types of geodemographic classification, the PRIZM system divided the population (in this case, households) in to a series of segments. The PRIZM classification operates with sixty-six "*demographically and behaviourally distinct types*" in a bid to aid marketers discern those consumers' likes, dislikes, lifestyles, purchase behaviours and media partiality (Neilson, 2005). The clusters, arranged hierarchically by estimated household income, range from type 1; Blue Blood Estates (income: ~ US$113,000) through to the final type; Southside City (income: ~ US$15,800) (Weiss, 2000).

2.10. Why the Individual?

Based on the review of existing geodemographic systems (as summarised in Table 2.1), it is clear that the majority of existing systems are based on areal units. Moreover, those that are based on postcodes have generally been classified at the area level initially and the postcode level classification has been modelled based on further individual-based data. The innovations of the Axciom and PersonicX systems represent a step forward towards smaller scale classifications, namely at the level of the household. Nevertheless, present-day systems typically fall foul of the following issues: ecological fallacy, the Modifiable Areal Unit Problem (MAUP) and generalisation as discussed in section 2.6. In this research, the focus is on constructing a classification at the individual level making use of microsimulation with sample-based data from the census to start at the level of the individual. This research represents the first attempt at such a classification in the academic literature and hence the thesis will propose a framework for further exploration and extension.

Given that a classification at this finest of levels has never been attempted before, robust methodologies do not exist in the same way as for area-level schemes. However, Farr and Webber (2001) discuss three possible alternatives for generating an individual-level population capable of classification. Full details of this and other means of synthesising populations can be found in Chapter 3.

One reason for producing an individual-level classification over a system output at a larger spatial unit is that it has far less scope for error or misinterpretation, as noted previously with reference to ecological fallacy, MAUP and generalisation. Consider the two following fundamental points as described previously:

- **Modifiable Areal Unit Problem:** A classification output at the level of the person is disaggregated down to its finest level thus negating the possibility of changing spatial patterns when viewed at differing scales. The presentation of a classification at this scale is predicted to nullify the needs to view the classification through any alternative means of aggregation and reduce the generalisation of more conventional systems operating at higher spatial resolutions.

- **Ecological Fallacy:** In conventional geodemographics there is the erroneous assumption that patterns observed for a body of people collected together in a specified spatial unit are also directly applicable to the individuals, for example everyone in a *"Young Married Suburbia"* area is deemed to be young and married. Due to generalisation which inevitably transpires when classifying through areal units, such assumptions are unavoidable. One may argue that ecological fallacy even exists, albeit to a lesser degree, in Axiom and Claritas' products if a household contains more than one resident. A classification at the level of the person would supersede previous classifications and improve the longstanding notion of ecological fallacy within geodemographics by providing a classification to the most detailed of levels.

Another interesting issue that can be addressed is the static nature of geodemographic systems and their incompatibility. This incompatibility is caused primarily by changing zonal systems over time. By developing

synthetic populations for the specified time periods (or obtaining suitable SAM samples), individual-level classifications can be built and assessed to compare changes in the populations, neighbourhoods and key demographics (such as affluence and deprivation) which have taken place across census periods.

2.11. Summary and Conclusions

This chapter has presented a detailed overview of both area classification and latterly geodemography. Information on history and early beginnings, evolution, a taster on system formulation, specific problems linked to area-based classifications, many of which this research aims to overcome, applications, current leading systems, ethical considerations and possible future directions for the discipline. The latter point of crucial importance given the focus of this research and hence a move towards person-based classification. Furthermore, given the deep-rooted foundations of geodemographics in area-based classification theory and little evolution in the recent past, an increase in computational power means such methods can be challenged. Examples of this computational power are shown in chapter 3 where methods for generating synthetic populations are discussed. Such population data being of paramount importance if an individual-level classification is to be developed.

Chapter 3: Methods for Creating Realistic Synthetic Populations

3.1. Introduction and Chapter Preface

As discussed in Chapter 1, the purpose of this research is to overcome the spatial and generalisation issues surrounding geodemographic classifications constructed using aggregate or area-level data. Ecological fallacy, MAUP, misrepresentation, and erroneous or misleading cluster labelling are fundamental problems which are largely unavoidable when producing area portraits based on collective data. Given the general lack of data available at the person-level of nationwide coverage, previous analyses have been restricted and have often succumbed to the aforementioned problems.

The main focus of this chapter is to explore a series of methods for population synthesis, in particular; deterministic reweighting (Smith *et al.*, 2009), conditional probability (or Monte Carlo simulation) (Birkin and Clarke, 1988; 1989) and simulated annealing (or combinatorial optimisation) (Openshaw and Rao, 1995; Williamson *et al.*, 1998; Voas and Williamson 2000; 2001). Consequently, this chapter is structured as follows: introduction (3.2), synthetic versus aggregate data critique (3.3), discussion of the various population synthesis algorithms (3.4), an overview of synthetic data applications (3.5), a more specific illustrative discussion on making use of individual data within geodemographics (3.6), and how this relates to the research undertaken here (3.7).

3.2. Overview of Population Generation Techniques

The use of population generation techniques have seen a rapid rise in recent years with many applications now requiring realistic individual-level data or complete synthetic populations. This trend can be attributed to a number of factors including: an increase in computational power and storage, a wealth of individual-level data of acceptable geographical coverage (for example, the British Household Panel Survey or Understanding Society as it is now known) and the appearance of new computational paradigms, such as cellular automata and agent-based modelling (Harland *et al.,* 2009a).

Static (as opposed to dynamic) spatial microsimulation produces a synthetic population (i.e. a population built from true but anonymous data at the person-level) which accurately portrays the observed ground population in a certain geographical zone for a given collection of criteria, for example: sex, age, social economic position, etc. Uses of such synthetic data are very wide-ranging and span research/policy areas from health (Smith *et al.,* 2009; Tomintz and Clarke, 2008) to water demand (Williamson *et al.*, 1996). Smith *et al.* (2009) also list taxation, child benefit policy, crime and education as past areas of research with respect to population synthesis. A fuller review of diverse application areas can be found in Ballas and Clarke (2008) and Ballas *et al*. (2005) and in section 3.5 in this Chapter.

3.3. Synthetic versus Aggregate Data

Synthesised data by far surpasses the aggregate-level equivalent available primarily from the Census. Sources such as CASWEB (2001), InFuse (2011) and Neighbourhood Statistics [courtesy of ONS] (2001) enable the free downloading of area-level datasets. CASWEB (2001) has recently opened its services to all whereas InFuse (2011) remains restricted to academic users only. Such datasets provide statistics similar to those given in Figure 3.1 (note: such statistics are for illustrative purposes only and do not reflect the ground situation in any given output area).

Output Area: Example001

Total Population: 325

Males: 200 **Females:** 125

Age 0-4: 22 **Age 5-9:** 35 **Age 10-15:** 21 **Age 16-19:** 43 **Age 20-24:** 32 …

Single: 158 **Married:** 100 **Widowed/Divorced:** 67

White British: 290 **White Irish:** 4 **Other White:** 1 **Indian:** 10 **Pakistani:** 6 …

Figure 3.1. Example of aggregate-level statistics available from online sources.

As can be seen from Figure 3.1, such statistics can provide informative results, for example in this area there is a 200:125 male-female ratio and a population which is single dominated. Furthermore, and despite authorities' efforts to manage output area populations for reasons of confidentiality,

percentages can be calculated given a total population statistic, i.e. 61.5% males, 38.5% females, or a total single population of 48.6%. Such figures can assist marketers and business planners to target the correct type of consumer with respect to a given product/service if presented at the small-area level. A plethora of high quality research has also been undertaken making use of data disseminated at this level (see; Longley and Batty, 2002; Stillwell and Clarke, 2004 as examples). However, one might argue that that is where the usefulness ends. Although the census clearly collects data at the level of the individual (hence completing a household form with individual/household-level data), confidentiality prohibits its distribution at this level and instead geographical zones are used as a means of release. Consequently, as the resolution of a census area become coarser (e.g. output area to lower super output area) and the risk of data disclosure diminishes, the availability of attributes increases (Harland *et al.,* 2009a).

Aggregate area-level data such as that shown in Figure 3.1 are very abstract and provide only an overview of the ground situation in any given census area. There is also an assumption of uniformity over space. Based on the data available from the census or that provided in Figure 3.1, it is not possible to determine the total number of single males within a given area or the number of married persons aged 20-24 without requesting commissioned cross-tabulations of variables. Observations such as these can only be generated through synthetic (estimated) populations whereby each individual is assigned a series of personal traits through simulation. Table 3.1 presents an example and makes use of the same illustrative output area as in Figure 3.1.

Person# / Characteristic	Sex	Age	Marital Status	Ethnicity	More Traits…
1 / 325	Male	0-4	Single	White British	…
2 / 325	Male	25-29	Married	White British	…
3 / 325	Female	20-24	Single	Pakistani	…
4 / 325	Male	65-74	Widowed/Divorced	White Irish	…
… / 325	…	…	…	…	…
325 / 325	Female	25-29	Married	White British	…

Table 3.1. Example of a synthesised population dataset.

In this instance, assuming a complete population, one could determine the total number of married persons aged 20-24, or, for example, the number of single females aged 20-24 of Pakistani origin. It is possible to deduce the total number of persons who share any series of traits when making use of synthetic data.

Harland *et al.* (2009a) describe how making use of a synthetic population is equivalent to filling in the blanks with respect to the added dimension the data provides when compared to that of aggregate-level census outputs. Furthermore, Wilson (2000, p.98 cited in Ballas and Clarke, p.278) describes microsimulation as "*a critical concept in the future development of modelling because it provides a way of handling complexity that cannot be handled analytically.*"

3.4. Microsimulation Algorithms

There are several established methodologies for generating synthetic populations which enable the creation of populations with traits in line with the above. The following sections will discuss deterministic reweighting (Smith *et al.*, 2009), conditional probability (Monte Carlo simulation) (Birkin and Clarke, 1988, 1989) and simulated annealing (combinatorial optimisation) (Openshaw, 1995; Williamson, Birkin and Rees, 1998; Voas and Williamson, 2000, 2001). These methods were selected due to their common application in geography. Many recent spatial microsimulation studies including Anderson (2007), Ballas *et al.* (2005), Voas and Williamson (2000, 2001), Tomintz *et al.* (2008) Smith *et al.* (2009) and Morrissey *et al.* (2008) have adopted a variation of at least one of these three approaches. Harland *et al.* (2009a) discuss similar findings.

Despite the varying microsimulation algorithms open for use, including the three mentioned in the preceding paragraph, all share one condition. In order to formulate an accurate synthetic population there is a requirement for a sample population from which to construct the synthetic populace. This sample population tends to be a dataset such as the British Household Panel Survey [BHPS] (Ballas *et al.,* 2005; 2007), Health Survey of England (Smith *et al.,* 2009) or other survey-based dataset. To further supplement the sample dataset, aggregate-level data such as that in Figure 3.1 is incorporated into the modelling process in the form of univariate (i.e. sex) or cross-tabulated (i.e. sex by marital status) constraint tables (Harland *et al.,* 2009b). Such

area-level data usually span both the attributes and areas in use (age, sex, marital status.....zone1, zone2, zone3....etc).

3.4.1. Deterministic Re-Weighting

The deterministic reweighting method produces a synthetic population by reweighting a survey population (such as the BHPS) or household dataset to fit to individual or household characteristics known at the small-area level through standard census variables (Smith *et al.,* 2009). This algorithm follows a two-stage process when synthesising data for small areas and is a large iterative proportional fitting routine (Harland *et al*., 2009b). Firstly, a weight is calculated denoting the likelihood of each individual record from within the sample population residing within the zone in question using information looked-up from the constraint tables (ibid). Secondly, the weights are then proportionally fitted to the known population for the zone and then repeated thereafter for each zone until a population of full coverage is generated (Smith *et al.*, 2009 cited in Harland *et al.*, 2009b). This form of population synthesis is very time-efficient with populations often generated within minutes on machines of modest computing power. Furthermore, Harland *et al.* (2009b) discuss how deterministic re-weighting is affected by the order of constraint incorporation and thus, as a result, the approach can integrate various model configurations for zones which may share similar characteristics. Such a possibility enables the incorporation of changing relationships between attributes over space. The authors provide the example of two economically active persons, one living on the city fringes and the other in the city. Although both these individuals share the same economic status, the person residing on the city fringe may be more likely to own a car than his/her counterpart in the heart of the city.

Harland *et al.* (2009b), in addition to Smith *et al.* (2009), list various shortcomings associated with using this method. In particular, the model is sensitive to the order in which the constraints are entered and the results are equally sensitive to constraint misconfiguration which can ultimately lead to overall data inaccuracy. With this considered, Harland *et al.* (2009b) note how the result from each model must be vigorously examined in a bid to ensure robustness and identify any error at an early stage.

3.4.2. Conditional Probability

This model, as developed by Birkin and Clarke (1988), is a further example of a reweighting algorithm. The algorithm is capable of operating with or without survey data or a sample population.

As the name suggests, the model operates chiefly on probabilities and in particular the likelihood of individuals residing in geographical zones and possessing a given series of traits. To take a hypothetical example, a geographical zone contains 150 males and 150 females where three quarters of the males and 80% of the females are economically active. The probability of being a male or female in this area is 0.5. To determine the number of economically active males one would multiply both the probability of being male with that of the probability of being economically active, in this case 0.5 multiplied by 0.75 giving a probability of 0.375 (or 56[.25] persons). This example is inherently simplistic and, inevitably, as the number of constraints increase, the calculation of conditional probability distribution becomes far more sophisticated. Birkin and Clarke (1989) provide a full review of this means of population synthesis.

In a bid to further ensure accurate populations are constructed, once the probability calculation has been established, its outcome is compared to counts in the observed population and the conditional probabilities are adjusted iteratively until a near match develops (Harland *et al*, 2009b). Harland *et al.* (2009a) present a more detailed description on how this comparison process is undertaken. Once a match is developed, the sample population is searched (assuming one is used) to find individuals that best represent the traits needed using a Monte-Carlo search algorithm (ibid).

Harland *et al.* (2009b) present a concise synopsis of the above technique, paying attention to its strengths and weaknesses. In summary, the specified population synthesis technique operates rather time-efficiently with respect to population allocation – albeit not quite as fast as deterministic reweighting. The authors discuss how conditional probability can synthesise a population of circa one million residents in less than three hours if running on an adequately resourced platform. The authors do mention, however, that this form of simulation creates entities (in this case individuals or households) which are 'likely' to appear in a given geographical area. It cannot be deemed to replicate the population with respect to matching records. By this, one means

that conditional probability creates a *predicted* population but it is by no means assured that a record in the simulated population will appear in the sample.

The overarching advantage of adopting conditional probability over other techniques is its non-reliance on a sample population. As mentioned previously, if a sample population is unavailable, this synthesis technique can construct a population using solely aggregate-level data. At the time when this method was first developed, sample populations were scarce, thus resulting in the method being required to synthesise without such datasets. In instances where a sample does not exist, Iterative Proportional Fitting can be used to create an initial probability distribution for individual/households likely to be contained within the population using the separate aggregate categories in the constraint tables (Harland *et al*., 2009b). Birkin and Clarke (1989) provide further details.

3.4.3. Simulated Annealing

Simulated annealing is a technique coined in statistical mechanics in 1983 as a method for optimizing functions of many variables (Buckham, 1999). Buckham (1999) discusses how simulated annealing is a heuristic approach that provides a means for optimisation of *non-deterministic polynomial time* (*NP*) *complete* problems, for example those for which an exponentially increasing number of steps are required to generate an exact solution. Buckham (1999) lists various uses for simulated annealing, principally within engineering. Examples include: the travelling salesman problem (see also; Press *et al.,* 1992; Harland *et al.,* 2009b), image reconstruction, integrated circuit design, path generation, and Planar Mechanism Synthesis. The process, however, can also be successfully applied within social science. Harland *et al.* (2009b) state how, in the social sciences, combinatorial optimisation problems tend to demand the minimisation of a given fitness statistic or measure representing the extent to which a certain configuration matches observed information. The named authors present a flow diagram illustrating the simulated annealing process when applied to synthetic population generation as shown in Figure 3.2.

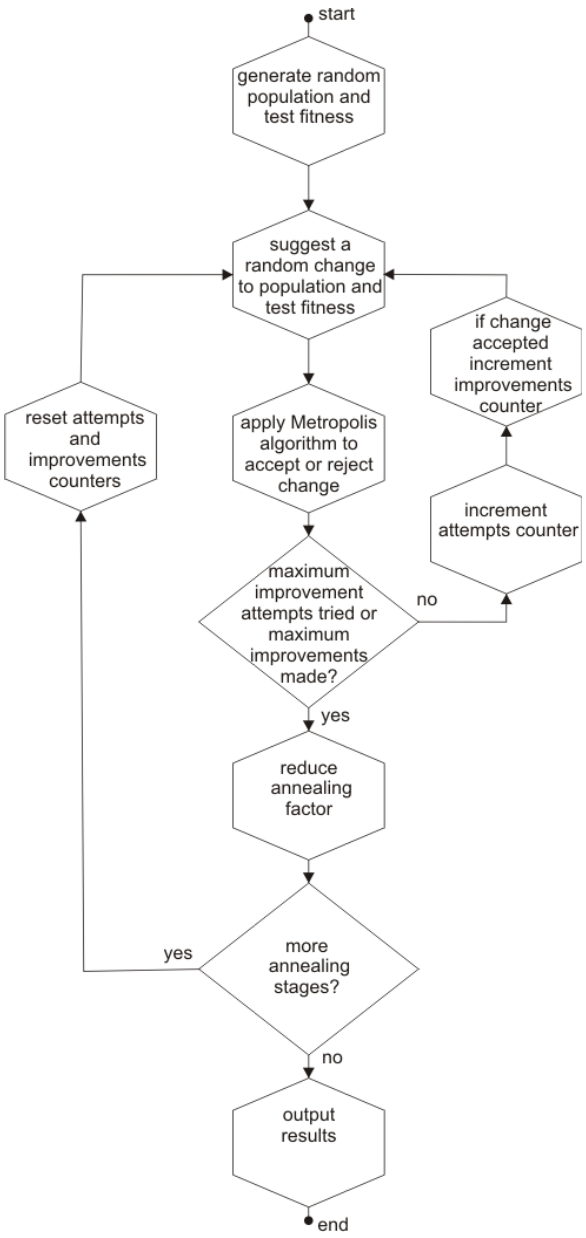


Figure 3.2. Simulated annealing algorithm used to construct synthetic population (Harland et al., 2009b)

Harland *et al.* (2009a; 2009b) present a detailed discussion of the algorithm's functionality. In short, the simulated annealing process begins with a synthetic population as arbitrarily generated from the sample survey population for a specified geographical zone. The fit of the synthetic population can be assessed to determine how well or otherwise it reproduces the traits of the zone, i.e. the individual characteristics, using a conventional goodness of fit statistic such as Chi-Squared or, as suggested by Harland *et al.* (2009b), Standardised Residual Mean Squared Error [SRMSE]. Various authors provide a review of such statistics, including; Lemeshow and Hosmer (1982), McKinley and Mills (1985), Legates and McCabe (1999) and more recently,

Genest *et al.* (2009). Voas and Williamson (2001) also present research specifically on the assessment of goodness-of-fit measures for synthetic data and concur with Legates and McCabe (1999) who criticise the ongoing use of such statistics often regarded as misleading.

The main testing process in simulated annealing involves the swapping of members from the synthetic and sample populations on a one-to-one basis with the synthetic population re-tested following each exchange. This test enables the fit of the known characteristics of the area to be contrasted with those in the synthetic population, if the new population sees an improvement of fit then the exchange is accepted by default. If the fit has not seen an improvement (or has worsened), a decision on whether to accept/reject the change is made through the use of the Metropolis algorithm (see Metropolis *et al.,* 1953). It is the inclusion of this Metropolis algorithm which differentiates the simulated annealing process from that incorporating the combinatorial algorithm (see Voas and Williamson, 2000). The enabling of a 'backward step' at a time when the match has worsened allows for the true best-fit solution to be determined by also rejecting as opposed to accepting any change.

Harland *et al.* (2009b) again critique this process by assessing its positive and negative aspects, in the same way as for the preceding methods. One rather obvious observation is the high number of exchanges necessary to facilitate the division of the optimum synthetic population. Although simulated annealing would appear very successful at synthesising populations from sample datasets, these exchanges impact heavily on both time and computational power, something also discussed by Goffe *et al.* (1994). However, large strides in computational capabilities in recent years mean that array indexing and further developments enable simulated annealing to function in a similar time frame to the other methods discussed, assuming comparable populations and conditions. Furthermore, this method requires by far the least pre-processing of input data when compared to both conditional probability and deterministic re-weighting.

3.5. Application areas within Social Science

So far this chapter has explored the functionality of three techniques capable of population synthesis. Although application areas have been discussed in

passing, particularly with reference to the added usefulness a synthetic population can provide over more conventional aggregate data, no more than brief mentions of wide-ranging applications have been provided. This section will provide a review of synthetic population usage.

Although, according to Ballas and Clarke (2008), the first geographical application of microsimulation came about in 1967 by Hagerstrand with respect to the spatial diffusion of innovation (see Hagerstrand, 1967), arguably the basis for microsimulation of households/individuals was introduced by Wilson and Pownall (1976) when assessing traditional models of urban systems. More recently, however, microsimulation has spanned a variety of research and policy areas. Ballas and Clarke (2008) provide an illustrative overview of its wide diffusion. Figure 3.3 displays the results from an academic journal search when filtering for 'microsimulation' in the title and/or abstract of published research.

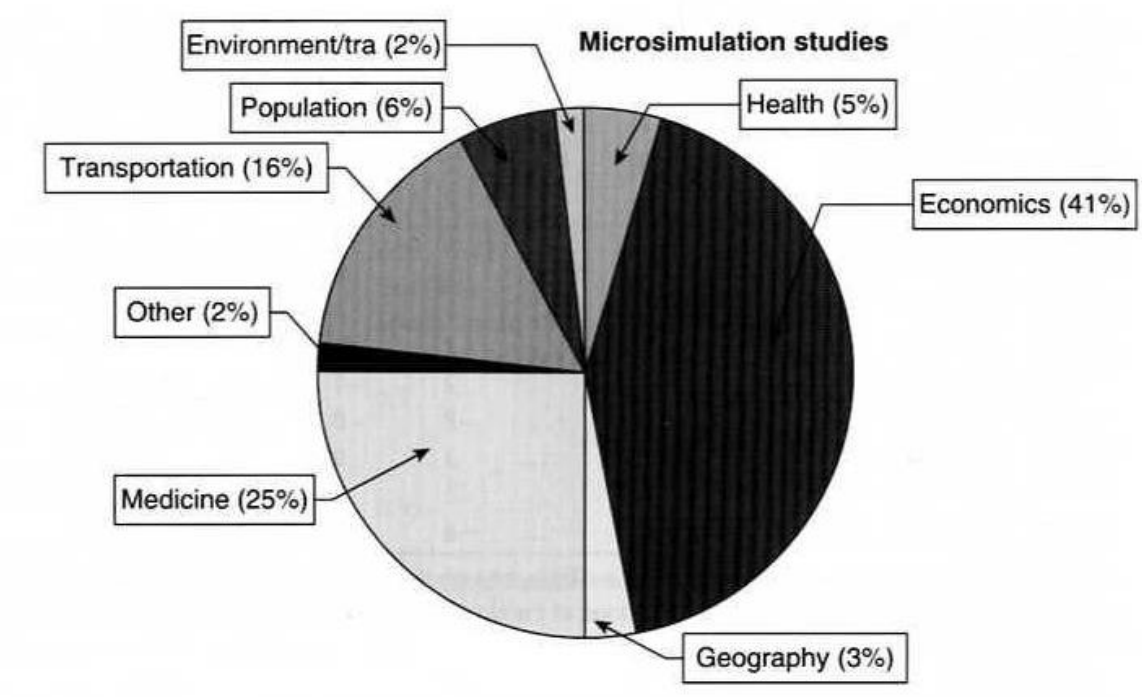


Figure 3.3. The distribution of 'microsimulation' in ScienceDirect academic studies in the period 1967-2003. Source: Ballas and Clarke (2008) in Fotheringham and Rogerson (2009).

As can be seen in Figure 3.3, the predominant segment of microsimulation applications during the specified period was undertaken in applied economics with very little in core geography (3%). One may argue, however, that research in health (5%), population (6%) and transportation (16%) hold strong

links with modern day social geography and its disaggregation into such categories does not represent geography's true uptake. As referred to by Smith *et al.* (2009), taxation and income modelling is one application area to have seen a healthy uptake in microsimulation methods, and Ballas and Clarke (2008) point to work by Neilson (1993), Propper (1995), and Birkin and Clarke (1989) as examples. Research by the latter was probably the first attempt to forecast income by small-area available in the literature (Ballas and Clarke, 2008). For a detailed critique of applications in tax, labour and housing markets, transport and land-use models, and retail, Ballas and Clarke (2008) and Ballas *et al.* (2005) provide concise discussions.

A further example of a microsimulation application, this time more central to the conventional perception of geography, is that of water demand. Williamson *et al.* (1996) discuss how water authorities bill households based on arbitrary estimates of water consumption. The lack of water metres in most UK households has meant that households were often charged a standard flat rate across a given area rather than a fee more in line with resource consumption; often rates were set based on the rateable value of properties in an area. Williamson *et al.* (1996) made use of microsimulation in a bid to define an individual household's propensity to consume and hence set up an infrastructure for a more equitable water billing procedure. This is a prime example of making use of microsimulation to advance the understanding of a social phenomenon and generate individual information from sample sources – something this research will explore.

3.6. Microsimulation and Individual-Level Geodemographics

Synthetic populations and, more specifically, individual-level classifications, have been the topic of discussion within geodemographics for a number of years. Evidence for this is provided by EuroDirect, amongst other sources. EuroDirect is continually investigating the possibility of incorporating person-level data into their 'Cameo UK' system with the view to creating an individual-level system (Bradbury, personal communication, 16/12/2009). As discussed at various junctures previously, a classification at this resolution possesses a series of advantages and overcomes a multitude of problems which surround both Boolean and fuzzy area-based schemes. Farr and Webber (2001) describe the benefits to be gained from moving from aggregate systems to systems capable of individual-level classification as being "*intuitively obvious*"

(p.58), particularly with reference to the added discrimination such systems provide. The same authors also suggest that previous analyses have proven this observation but fail to list any such examples. Farr and Webber (2001) exemplify their concerns regarding area-based systems by describing a hypothetical example. The authors suggest that two types of neighbourhood (or small-area) deemed to encompass similar concentrations of, for example, unemployment, lone parenthood and overcrowding, may in fact be far more diverse than any aggregate classification can suggest. It is not beyond reason that one neighbourhood may contain high numbers of persons suffering from all three disadvantages or even none at all whilst the other may only house residents suffering from one or two of the disadvantages. Farr and Webber (2001) concur that blanketing populations under one label is both wasteful and often ill-informing, particularly if the purpose is to identify individual people at risk. Such classifications may also be regarded as quite dangerous if the interpreter does not have some prior knowledge as to their formulation.

Farr and Webber (2001) suggest that attempts to construct geodemographic person-level classifications fall into three categories; the first is those making use of lifestyle survey datasets (e.g. those collected by Experían), while the second, largely influenced by the failing results of the previous attempt, are those integrating solely publicly available datasets (e.g. from the electoral roll). The final method is those adopting data from the census' Sample of Anonymised Records (SARs). This research falls into the final category and, according to Farr and Webber (2001), presents a highly practical method for classifying at the person-level. An overview of each method is presented below.

According to the authors, previous attempts at making use of commercial lifestyle survey data (including work undertaken by the authors themselves) to generate a classification fail as, inevitably, such datasets are highly behavioural and fail to correlate with one another. For example, basing a classification on responses to questions such as 'do you play golf?' or 'do you drive a company car?' result in clusters which discriminate very poorly on the input variables. Webber and Farr (2001) argue that this is primarily caused by low correlations between data characteristics and a general lack of 'natural' clusters within the data. Furthermore, one may argue that respondents to such surveys are by no means representative of the population as a whole

and the authors do make reference to this by means of considering strong geographical disparities in response rates.

Farr and Webber (2001) discuss how the failings associated with this method of person-level classification influenced a second method, that of using only publicly available data – for example, that held in the electoral roll. This method has seen a healthy uptake, albeit to varying extents, by commercial geodemographic vendors including Experían, Claritas and CACI (*ibid*). Again, however, one must understand that not everybody completes the electoral roll and those who do not are not likely to be a random representative sample. The authors argue how this then has implications for 'social exclusion' and represents a negative point for means of classification which on the whole has been well embraced.

The final method for individual-level classification construction involves adopting data from the census' Sample of Anonymised Records (SARs). The authors also state that such is the geographically referenced nature of SARs data, it is possible to analyse the frequencies of personal classifications on an area-basis. As this forms the root of this research, adopting this approach will be consider in greater depth in chapter 5.

3.7. Summary and Conclusions

This chapter has reviewed different methodologies through which individual population datasets can be constructed, specifically the chapter has discussed deterministic reweighting, conditional probability and simulated annealing (or combinatorial optimisation) and through an assessment of the pros and cons of each, a method for use in my research has been determined.

This research will incorporate synthetic data formulated through the combinatorial optimisation using the simulated annealing method. As discussed previously and emphasised by Harland *et al.* (2009a; 2009b), this method presents the most accurate means of synthesising populations despite its rather intensive computational requirements. Furthermore, with support from the work of Farr and Webber (2001) suggesting that a census-based approach (from the SAR) is the most effective way forward, this research will also pursue this route in bid to create an effective classification of individuals.

The following chapter will explore common methodologies adopted in geodemographic classification systems through the presentation of a stepwise

approach to system formulation. It will be argued that such methodologies remain firmly rooted in the disciple (hence the continual persistence with area-based /aggregate data classifications).

Chapter 4: Conventional Geodemographics: A Dated Approach?

4.1. Introduction and Chapter Preface

The purpose of this chapter is to build on the foundations set out in section 2.4 regarding the building blocks of geodemographic systems. Presented herewith is a detailed review and explanation of common methods adopted in geodemographic system formulation. This chapter takes a sequential approach to reviewing the key phases commonly regarded as necessary in the construction of any such system.

4.2. Geodemographic System Formulation

Although the end product of a geodemographic system is a simplified depiction of reality (as per any geographical model) typically through the creation of a series of clusters each linked to a geographical area, the underpinning processes required to reach this are far from straightforward. Harris *et al.* (2005) and Gibson and See (2006) both present detailed descriptions of how geodemographic systems are constructed, the former describing the processes adopted by Experían when creating the Mosaic system (see section 2.7.3), the latter presenting a more generic approach albeit later linked to uses within sustainable development. Milligan (1996) also presents a seven phase approach, albeit more dated, similar to that proposed by Gibson and See (2006) for common cluster analysis. Figure 4.1 illustrates the stages involved in creating a geodemographic system.

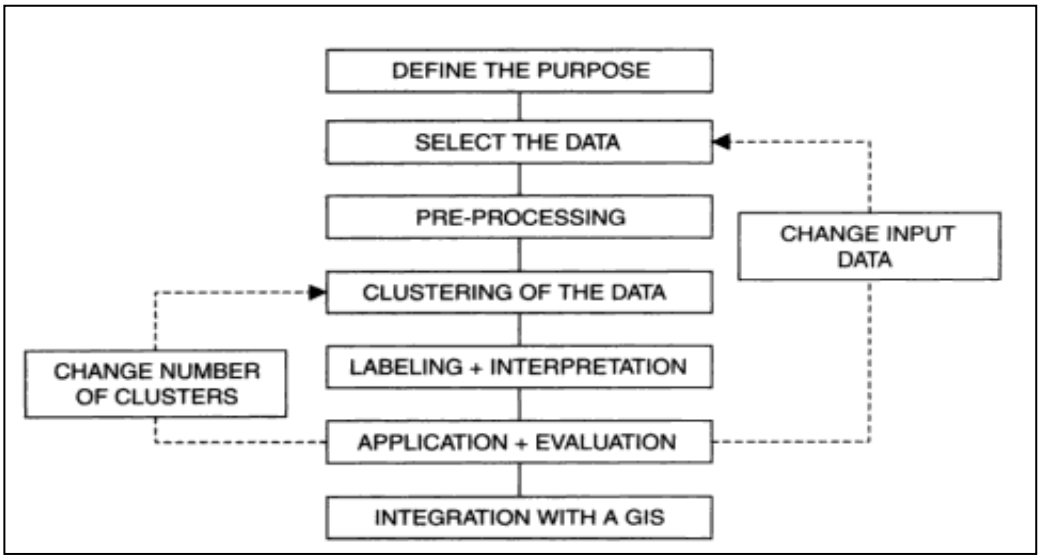


Figure 4.1. Flow diagram showing the processes required to devise a geodemographic area classification scheme (Gibson and See in Campagna, 2006, p.214).

Figure 4.1 illustrates a minimum of seven phases required when constructing any such system. However, it is not uncommon to re-visit or loop back to previously accomplished steps when striving for the optimal classification. The first two phases are relatively uncomplicated. Ensuring a successful output from a geodemographic system is determined primarily by the following three phases; Pre-Processing, Clustering, and Labelling. This report will take a stepwise approach to discussing each phase - with particular emphasis placed on these three fundamental phases. Given that the purpose of this chapter is to discuss system creation, the final two phases of 'Application/Evaluation' and 'Integration with a GIS' will not be discussed here and will instead become apparent in later chapters as part of visualisation and evaluative measures.

4.3. Defining the Purpose

The commencement point in system development is to define the purpose of the classification - that is, is a general-purpose all-encompassing system desired, one that is designed to paint a picture of ground conditions in an area independent of any specific application. Many of today's leading commercial systems are general-purpose and they are built in this way to ensure market penetration. Any system designed in this way can, theoretically, be sold to a range of organisations and markets and potentially achieve a goal. General purpose systems may incorporate a range of variables, for example, UK census variables such as age, sex, car ownership, ethnicity, etc plus, if

available, a plethora of more behavioural variables that define propensities to buy high-end goods, eat foreign food or watch live sport amongst other things.

Gibson and See (2006) argue that although together, variables within multi-purpose systems can be used to develop a range of diverse area typologies, more bespoke systems are often more fit for purpose.

Systems designed with a specific purpose in mind are now more widespread than ever before. Conventional geodemographic vendors are releasing more explicit classifications, for example Experían (2013) now has a portfolio of systems ranging from CAMEO UK (the flagship product) to CAMEO Property, CAMEO Choices and CAMEO Welfare. The former of these is designed to understand levels of affluence through house prices, the second to understand purchasing behaviour and the latter for evaluating economic hardship. These are just three in a portfolio of seventeen classifications offered by EuroDirect (2013). In academia, systems built for specific purposes are also apparent. Abbas *et al.* (2009) presents a system within the domain of health intelligence and Burns (2009) also emphasises how geodemographics can be tailored to the health/deprivation sector.

4.4. Selecting the Data

The second phase in system formulation is to determine the data or input variables. In the UK, given that the census is the most comprehensive dataset available, both in terms of depth and coverage, it naturally forms the bulk of variables utilised in geodemographic systems. This may well change moving forward with the uncertainty surrounding future national censuses but for now it is seen as the leading repository for geographically-referenced demographic data. Gibson and See (2006) do, however, put forward cases for using alternative data sources such as share ownership, unemployment, county court judgments (CCJs), and registers of company directors as these tend to provide more financial input to the classification. Webber (2007) also emphasises how, in many countries, non-census data can prove useful and lists the electoral registers (UK, Australia, Spain), the files of mail order companies (Netherlands), car registration files (Italy), Property Registers (Germany, New Zealand, UK), registers of shareholders and of directors (UK), statistics on house prices and on council tax bands (UK) and registers of addresses (Australia, UK) as prime examples. Webber (2007) also states that in the Netherlands, where census statistics are not published at the small-area

level, market research respondent files are used to add value to classifications.

Webber (2007) is a strong advocate of using non-census data in geodemographics and states that such sources of information can be useful for three key reasons. Firstly, questions in national censuses justifiably tend to centre more strongly on measures of disadvantage than on measures of affluence or prosperity, asking their populations about their literacy (Brazil, China), long term illness (UK) or unemployment (Australia). Information from non-census sources, such as those mentioned above, is often helpful in redressing this prejudice and in providing greater detail about the location of more privileged population groups.

A second advantage of using non-census sources is that, in many instances, the data are available at a finer level of geographical detail than that at which census statistics are published. This latter point is very apparent in the UK with the 2001 census data released at its finest resolution to output area level and, on average, each output area being comprised of up to five postcodes. Inevitably, in instances such as this the scope for finer-level analysis is clear but this is only the case if aggregate census data forms no part of the data input process.

The third advantage put forward by Webber (2007) is that in many markets the use of non-census sources makes it relatively easy and highly advisable to update the classification codes given to existing areas (or clients) as their population character changes over the interval between censuses (decennial in the UK). Furthermore, Webber (2007) states that by using alternative data sources it is possible to assign classification codes to neighbourhoods built since the date of the previous census and hence ensure systems are kept current.

This phase is arguably one of two largely subjective junctures in devising a classification - the second being in determining the total number of clusters or sub-clusters (see section 4.6). Deciding on which variables to input and how many should underpin the final classification is a topic that has seen much research. Openshaw and Wymer (1995) citied in Gibson and See (2006) argue that one should attempt to make use of the smallest number of

variables for the purpose required and that one should also try to avoid variables that are highly correlated (hence a need to assess all variables for multi-collinearity prior to classification). The data should also be examined for outliers or rogue data, which should be duly removed or verified prior to classification. Openshaw and Wymer (1995) emphasise that there is no unique or correct single set of variables to use, hence the subjectivity at this phase. The authors do, however, stress the need to treat this phase as being *"iterative"* (p.215) in a bid to eliminate subjectivity and find the optimal classification output.

A slightly different viewpoint is offered by Webber (2007) who states that generally, the more variables that are used in the clustering process and the greater the breath of sources they come from, the more meaningful the resulting clusters are likely to be. Webber (2007) does, however, concur with previous authors that only variables that claim to be of value to the specific nature of the classification should be included.

4.5. Pre-Processing the Data

In pre-processing, various techniques can be employed; however, the simplest procedure is that of normalisation. Normalisation is carried out to ensure that all variables operate on equal grounding, thus variables with large ranges are equally weighted against variables which have much smaller ranges. This method typically linearly re-scales data on to a scale of zero to one (Gibson and See, 2006). Alternative procedures for data preparation can also be employed, for example Principal Component Analysis (PCA), or the related technique of factor analysis. Voas and Williamson (2001) present a detailed discussion of the former. In summary, these are a series of methods used to isolate the key differentiating factors (or 'components') of a collection of correlated variables (Robinson, 1998). Gibson and See (2006) argue that PCA is not the ideal method for data pre-processing due to its sensitivity toward skewed variables. This apparent lack of support for the PCA process is also endorsed by Harris *et al.* (2005) and, ultimately, Experían, who according to Harris *et al.* (2005, p.157), failed to make use of this technique in their Mosaic system due to its tendency to *"blur rather than clarify fine distinctions between cluster types"*. It should be noted though that PCA has been successfully employed in pre-processing techniques involved in the

construction of other geodemographic systems, notably the post-1981 UK census version of the SuperProfiles classification (Brown and Batey, 1994).

4.6. Clustering Overview

The next phase is the clustering phase and this is arguably what separates a geodemographic classification from a simple hybrid index. In the simplest of terms, this process reduces multivariate datasets into a descriptive and manageable number of area typologies (Gibson and See, 2006). Harris *et al.* (2005) discuss how clustering algorithms can take one of two forms; stepwise, top-down methods or iterative allocation-reallocation methods. Alternatives, however, do exist and include simulated annealing and neural network classifiers – see Openshaw and Wymer (1995). There are several methodologies in place which allow for crisp or fuzzy approaches to clustering and Gibson and See (2006) point to a comprehensive review of such multivariate clustering techniques/algorithms by Krzanowski and Marriott (1995) for a fuller description.

Gibson and See (2006) emphasise how there are many different clustering algorithms available, ranging from commonly adopted K-Means approaches to algorithms developed in the field of artificial intelligence, examples include the self-organizing map and sophisticated machine learning techniques. The authors also refer to fuzzy clustering algorithms that can assist to characterize areas more accurately in cases where they could easily belong to more than one defined cluster-type.

Deciding on the correct number of clusters is a rather subjective process with many systems, such as Mosaic (from Experían) and ACORN (from CACI), choosing to segment areas into an odd number of clusters – 5, 7, 9 groupings seen as popular amongst smaller classifications. In the latest manifestations, Experían opted for fifteen key groups and CACI seven. Milligan (1996, p.343) describes the process of determining a set number of clusters as being very difficult "*if no a priori information exists as to the expected number of clusters in the data*." Gibson and See (2006) also emphasise the subjectivity linked to cluster number selection and add that, in the main, the best method for deducing this is to create multiple classifications and assess the changes in the slope of the scree or information loss.

4.7. Labelling and Interpretation

Upon completion of the clustering processes, a phase synonymous with devising a geodemographic scheme is that of cluster labelling. Harris *et al.* (2005) discuss how geodemographic system purveyors do not purposely set out with the intention of creating X number of clusters with names of rural isolation, mortgaged families and so on, but in fact such groupings naturally emerge from the classification and, as a result, require the assigning of a descriptive label. This label is usually a short-hand description of the characteristics of the population which the given cluster encompasses and is often supplemented by a more detailed area pen portrait (*ibid*).

Although these more detailed 'portraits' are used for more thorough analyses, the short-hand labels are necessary for more everyday and generalised interpretation. Gibson and See (2006) discuss the importance of comparing the individual cluster centres (or average behavioural characteristics) with the global average before assigning any descriptive labels. The global average in this instance refers to the mean characteristics of the whole area spanned by the analysis. One common approach, according to Gibson and See (2006) and Harris *et al.* (2005), is to determine z-scores. Z-scores describe the deviations from the global average for each individual variable which then enables clusters to be labelled based on their standing with respect to the global (national, regional or more local) average. Harris *et al.* (2005) refer to the importance of assigning true labels to clusters in a bid to portray an area in the most accurate fashion, however, the authors discuss how short-hand labels can often detract from the longer area portraits and even paint a less than accurate area description. With the continuing growth of geodemographics in the private sector, and in marketing especially, one can understand the over-reliance placed on short-hand area labels and, in particular, the 'favourable massaging' (or bias) of such labels in a bid to describe all areas in a positive light.

Experían's Mosaic system, itself an example of a private sector scheme, adopts both short-hand labels and descriptive portraits and four partial examples of these are shown in Figures 4.2 and 4.3. Figure 4.2 illustrates two of the highly ranked cluster-types whereas Figure 4.3 provides examples of the more deprived segments of society.

Group A: Alpha Territory

Piers and Imogen

3.54% of UK households (Types 1-4)

Alpha Territory people are the most wealthy and influential individuals in the UK. They have risen to positions of power in the private and public sectors, whether as owners of their own businesses, as bankers in the city, as senior managers in industry or as top lawyers, surgeons or civil servants, as well as a small but influential cadre of celebrities in sport, the arts and entertainment.

An increasingly large proportion of this group are wealthy foreign nationals. Some are top executives employed by international companies, others wealthy entrepreneurs, increasingly from Russia and the Middle East, who find Britain a convivial location in which to manage their business interests.

Alpha Territory neighbourhoods are concentrated in London, in particular the fashionable inner suburbs of South Kensington, Islington and Notting Hill. They are also common in prime residential areas such as Surrey and the Chilterns, and in the semi rural commuter belt of Britain's largest regional centres. They reside in quite different types of housing, in smart private flats in inner London, town houses in Georgian squares in Bath or Edinburgh's new town, or in large properties common to the outskirts of the UK's big cities.

Group B: Professional Rewards

Graham and Carolyn

8.23% of UK households (Types 5-10)

Professional Rewards are the UK's executive and managerial classes. Often in their 40s, 50s or 60s, some may be owners of small or medium sized businesses whilst others will have risen to senior positions in large multi-national organisations.

These people have significant equity in their own homes. With incomes from their investments as well as from employment, their affluence means they are often paying a higher rate of tax.

Professional Rewards are usually married and are living in spacious family homes with children, many of whom will be at university or finding their feet in successful careers. Professional Rewards neighbourhoods are found in residential areas in the outer suburbs of large cities or in semi-rural dormitory villages from where their occupants often travel to work by car.

The house is likely to be detached, have four or more bedrooms, is well maintain and equipped, and surrounded by property of a similar style. Tastes in clothing and furniture are generally conservative, with preferences for good taste over ostentation. Professional Rewards people will shop at retail stores where service is a key component of the overall proposition. They seek value for money rather

Figure 4.2. Two examples of Mosaic's cluster types. Group A and Group B represent more affluent members of society. Experían (2009).

Group K: Upper Floor Living

Jamal and Chantel

5.18% of UK households (Types 45-49)

Upper Floor Living people are on limited incomes and rent small flats from local councils or housing associations. Typically these people are young single people or young adults sharing a flat. They may also be single people of older working age or even pensioners.

Upper Floor Living neighbourhoods are most common in inner London, Scotland and, to a lesser extent, large provincial cities in the north of England.

Quite a few of the larger blocks that these people live in have turned out to be less attractive places to live than their architects and planners had originally envisaged. As a result many of them are hard to let.

Not necessarily living in housing of their own choice, many residents are disadvantaged by living among neighbours who suffer high levels of vulnerability. On the other hand, in those parts of London where money has been invested in the renovation of these blocks, some of the flats have been sold to private owners and can prove attractive to young professionals as an alternative to buying a flat in a more modern block a long commute away from their jobs in central London.

Group L: Elderly Needs

Henry and Violet

5.96% of UK households (Types 50-53)

Elderly Needs are pensioners whose faculties are now fading and who can no longer easily manage the responsibility of looking after a house and garden. This group contains a large number of older pensioners, typically in their 70s, 80s and even 90s, who are no longer as physically active as they once were.

Some members of this group may at one time have had well paid jobs but the majority are people who, on account of their low incomes, could never realistically look forward other than to a state pension. Those members of the group who years ago benefited from the sale of a family home and who still enjoy some form of private pension have also found their retirement incomes reduced by inflation.

Elderly Needs are found in a mix of different types of accommodation. These range from nursing homes and sheltered accommodation serviced by a resident warden, to homes designed for semi-independent older people, such as modest bungalows on a council estate. These neighbourhoods are scattered throughout the country but with notable concentrations in the UK's principal coastal retirement destinations, in Northern industrial cities and in Scotland.

Figure 4.3. Two examples of Mosaic's cluster types. Group K and Group L represent less affluent members of society. Experían (2009).

As can be seen from Figure 4.2 and 4.3, the inclusion of typical imagery indicative of the ground situation in these clusters together with photographs of stereotypical members are designed to immerse the reader into painting their own mental image of life in these areas. The commercial element is further enhanced by the naming conventions, ranging from 'Alpha Territory' to 'Elderly Needs'. Furthermore, adopting person names such as 'Henry and

Violet' (Group L) and 'Piers and Imogen' (Group A) and other names that one may instinctively relate as falling into a certain demographic grouping helps the lay reader to convey a sense of understanding of the people-types who inhabit such areas. Although the pen portraits are purely descriptive in nature, they do offer the reader some insight into the conditions, people, facilities, behaviour and affluence levels present in areas categorised into each grouping.

4.8. A Dated Approach?

On top of a review of methodologies and the stages surrounding the creation of geodemographic systems, a second objective of this chapter is to assess how fit for purpose such methods are today. With all commonly available and trusted systems operating at the small-area level, the conventional methods discussed above do tend to provide suitable enough discrimination to render them effective. However, ambiguity in geodemographics is a word that comes to the fore quite often, hence discussions on a move towards a more fuzzy-based approach to cater for areas that do not fit a single stereotype. As this research proposes a classification at the individual-level, the level of ambiguity brought about through crisp (or best-fit) cluster matching should diminish.

As a fore-runner to creating the individual-level classification, several tests were undertaken on the Output Area Classification (OAC) as devised by Vickers (2006) and now widely used by the Office for National Statistics (ONS). Although the OAC is an area-based classification, the transparent methodology means that tests such as those discussed below can be freely undertaken without a fear of prejudice. This testing process involves contrasting the OAC with some 2001 simulated data for Leeds and hence data similar to that which will be utilised for the creation of an individual-level classification.

The 2001 pre-simulated data used in this analysis comprises six variable constraints; Age, Marital Status, Sex, Highest Qualification, Socio-Economic Classification, and Ethnicity and was formulated through combinatorial optimisation (simulated annealing) by Heppenstall (date unknown). The data are complete for Leeds (715,402 persons) and synthesised at output area level (2,439 areas) with constraint data acquired from the Census of Population via CASWEB (2001) and survey data courtesy of the British Household Panel Survey.

A simple summation of the total number of people who possess each characteristic per output area can, when contrasted with the Output Area Classification (OAC), generate interesting findings and highlight some of the problems associated with conventional systems which classify populations at the area (aggregate) level. Given the simplicity of the data required for this analysis, data used in the following examples could easily have been extracted from CASWEB (2001) given that only aggregate data are required. However, for the purpose of validation, the 2001 synthetic data was used as a substitute.

As discussed by Birkin (1995), area cluster labels can be far from indicative of the populations to which they are expected to encompass. Recall from Section 2.6.3 and the SuperProfiles example. Areas in clusters labelled "*Young Married Suburbia*" and "*Metro Singles*" contained rather different populations to those implied by the cluster labels. One area classified under the former label encompassed in excess of one quarter of residents over the age of 45 whilst a separate area within the latter contained only circa 20% of single workers (*ibid*). Such labels are very weak short-hand descriptors of the populations to which they are expected to describe rather accurately. Furthermore, such labels are wrongfully misleading and could easily result in failure or inaccurate population targeting when applied to a given scenario, in either the public or private sector, if interpreted by somebody with no understanding of how geodemographic systems operate.

The use of Leeds' 2001 simulated data together with the open source nature of ONS' Output Area Classification (by Vickers (2006)) enables similar contrasts to be made. The OAC comprises seven super-groups, twenty-one standard groups and fifty-two subgroups and is a system constructed from purely census data. The following pages will present an analysis of several clusters and standard groups within this wider OAC when contrasted with individual-level data simulated to encompass the six personal characteristics stated above. The observations made here highlight the key reasons for moving away from area-based classification towards finer-level systems.

The OAC cluster selected for discussion here is the "*Multicultural*" cluster (Super-group 7). This cluster is disaggregated into two standard groups; "*Asian Communities*" (Group 7a) and "*Afro-Caribbean Communities*" (Group 7b) with both of these groups further split into three and two subgroups

respectively. However, given that these final-tier groups possess no explicit label and for the purpose of this testing process, this assessment will operate down to tier-two of the OAC only.

The analysis that follows is made possible through the availability of an ethnicity characteristic in the 2001 synthesised population.

It is reasonable to make the assumption that any output area selected from the 2,438 which comprise the district of Leeds and which falls into either of the "*Asian Communities*" or "*Afro-Caribbean Communities*" categories should, one would expect, contain a high concentration of the given ethnicity; Asian or Afro-Caribbean. Furthermore, one would also expect any "*Multicultural*" area classified within the "*Asian Communities*" group to contain a higher percentage of persons of Asian ethnicity than those of Afro-Caribbean, and vice versa. However, an assessment of the OAC proves this to be correct for the majority of cases but this observation is by no means consistent across the 273 output areas classified as "*Multicultural*" across Leeds. In fact, of the 217 areas deemed to be in the *"Asian Communities"* subgroup, 30 (13.82%) actually contain higher concentrations of Afro-Caribbean residents. A similar observation is also evident when the assessment is reversed. Of the 55 areas categorised within the "*Afro-Caribbean Communities*" subgroup, 22 (40%) include a higher percentage of Asian inhabitants. In one extreme case, the number of Asian residents in an *"Afro-Caribbean Communities"* area exceeded the Afro-Caribbean population by over one-fifth (22.9%).

Based on the pre-simulated data used in this comparison and the availability of the six population characteristics, further contrasts could also be made to assess the composition of areas classified as "*Older Blue Collar*", "*Older Workers*" and *"Young Blue Collar"* (using the Age and Socio Economic Classification (SEC) variables) and *"Senior Communities"* (using the Age variable). Results show that the patterns observed with regards to age and economic status replicate those as seen with ethnicity. Thus, rather large disparities exist. Contrasts such as these, however, are more open to the interpretation of the researcher, unlike ethnicity. In these cases, definitions of 'Senior', 'Young' and 'Older' with regards to age and 'Blue Collar' in terms of socio-economic classification are required.

To take the *"Senior Communities"* subgroup as an example. This subgroup accounts for fifty-eight of Leeds' total output areas. If 'Senior' in this case is

defined as 65 and over (hence what is commonly regarded as retirement age for females), then of the ten output areas containing the highest concentration of 'senior' residents, seven are captured by the *"Senior Communities"* cluster with the remaining three described as *"Settled in the City"* – the latter not influenced by age. One may argue that this is a reasonable returns ratio. However, one must also consider the concentrations of 'Senior' residents within an area compared to both other areas and alternative age groups. In this instance, of the 58 small areas classified as being within the *"Senior Communities"* subgroup, the range of concentrations of persons aged 65 plus is quite extreme. Concentrations vary from 12.18% to 75.29% and, when it is considered that output areas are designed to enclose circa 300 inhabitants, such a variation is rather large (63.11%). Furthermore, areas categorised within the lower reaches of *"Senior Communities"* (e.g. with concentrations closer to 12.18%) enclose higher percentages of people in other age ranges; however, this again depends on how the age structure is disaggregated.

The above analysis is enough to suggest that a classification constructed at the level of the person will discriminate to a far greater extent than current systems operating at higher resolutions, such as the OAC. Furthermore, the issues surrounding generalisation, cluster concentrations and, to some extent, erroneous cluster labelling, can be overcome.

4.9. Summary and Conclusions

Geodemographics is widely embraced for consumer or person targeting. With the issues discussed above, the success of such operations is largely restricted and often succumbs to the error imposed when making decisions based on collective data. In the public sector, it is often necessary to identify specific people 'at risk' from a given phenomenon, something very difficult when adopting a classification based on aggregate and hence 'averaged' data (as per the examples above). As a result, people 'at risk' often go unnoticed due to the resolution at which systems are constructed. A classification at the level of the person would eliminate such problems and enable a far more efficient means of population targeting.

This chapter has presented a stepwise approach to formulating an area-based geodemographic system, following the flow diagram structure proposed by Gibson and See (2006). The chapter has presented a detailed synopsis of each stage in the process ranging from the importance of identifying a

purpose (for variable selection, scale etc) to more ambiguous stages including how to select variables, how many variables and the number of clusters. A key point to be taken from this chapter is the analysis undertaken on the OAC and the issues that transpire when attempting to attribute a cluster label to aggregate data – hence, the notion and problem of ecological fallacy and something an individual-level classification may not eradicate but could certainly improve upon.

The methods presented in this chapter are commonplace when constructing area-based systems. However, given the need in this research to formulate a classification at the level of the person, new and adapted methods are required. One key objective of this research is to propose a universal framework through which individual-level classification can take place and these methods are discussed in the following chapter (chapter 5).

Chapter 5: Devising a Framework: From Raw Data to Individual-Level Classification

5.1. Introduction and Chapter Preface

As mentioned at various junctures in this research, the predominant aim of this project is to develop a framework through which an individual-level classification can be created. The purpose of this chapter is, therefore, to present the methods through which a classification at this level will be developed. The development phases as put forward by Gibson and See (2006) in chapter 4 are loosely followed for the purpose of structure but adapted to suit a classification being constructed at the individual level.

5.2. Defining a Purpose & Proving Rationale

As discussed in chapter 4, classifications can be developed such that they are designed for a specific purpose such as health (see Abbas, 2009) or deprivation (see Burns, 2009) or they can adopt a wider appeal through being general purpose. Given that this research will make use of entirely census variables (see section 5.5), a wide selection of variables has been used in order to create a general purpose classification to demonstrate the system's usefulness across a broader set of applications. In many ways, a classification of this nature can be likened to the Index of Multiple Deprivation (IMD) which, although a simple index by definition, is often applied to a wide range of problem areas given its composition of: income variables (22.5% of total), employment (22.5%), health (13.5%), education (13.5%), housing (9.3%), crime (9.3%) and environment (9.3%) (Data.Gov.UK, 2010).

A key motivator behind the construction of a classification at this level is discrimination and the added benefit to be gained by moving away from areal unit categorisation, regardless of whether the system is general purpose or application-specific.

Aggregate systems, by definition, have to contend with a plethora of different variables which are expected to describe the population in these areas. Although such area populations are largely homogeneous based on the

premise that "*birds of a feather flock together"* (Harris *et al.*, 2005, p.16*)* and hence people with similar traits tend to gravitate to similar locations, with an increase in variables comes an increase in prospective error as demonstrated in Figure 5.1.

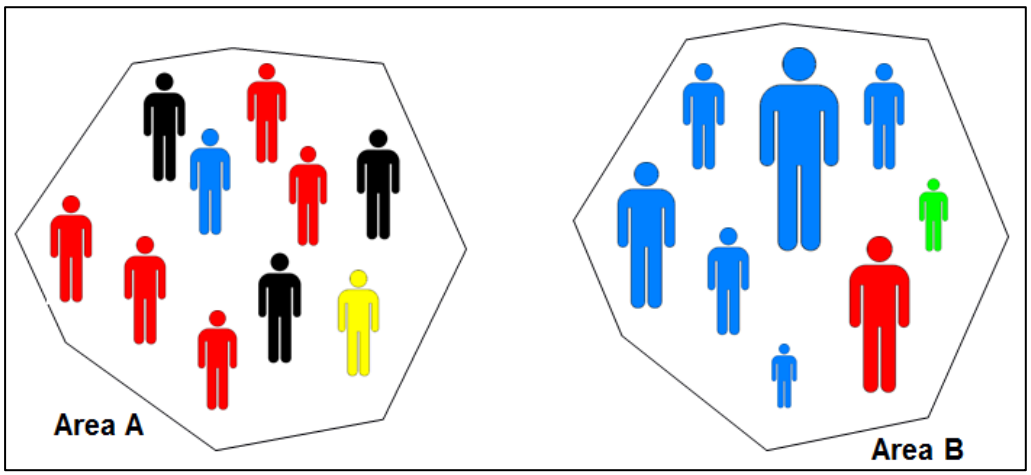


Figure 5.1. Problems in aggregate level classification caused by increasing variables and people traits.

In Figure 5.1, two hypothetical areas exist; Area A and Area B. If Area A was taken independently and clustered into a crisp 'best fit' grouping, one may expect that based solely on one variable, that of person shading, this area would be assigned to a ‘red shaded’ dominated cluster. By doing this, even with one solitary variable, collective error is apparent given that all members of this area do not fit this typology. There are in fact four people types resident in this area; nevertheless the level of error is minimal when focusing on one variable. When a second variable is introduced, that of person height, the collective error increases yet further. In Area B, two variables are apparent - person shading and person height. If this area were to be assigned to a single ‘best fit’ cluster, it may fall into a ‘tall blue’ grouping (or similar). In this instance, the ambiguity is greater and the ability to classify with minimum error becomes more difficult. Therefore, it is clear that as variable numbers are increased, the level of collective error that transpires as a result also increases, making it near impossible to accurately position areas into groups based on large numbers of variables. This demonstrates the need to keep variable numbers to a minimum when adopting a cluster-led approach to data analysis, which is also supported by Openshaw and Wymer (1995) cited in Gibson and See (2006).

The above observations may seem rather obvious and somewhat critical, even when based upon such a simplistic and hypothetical example. However,

given that geodemographics is seen by many as a type of modelling and hence a simplification of reality, the intention is to reduce complexity and aid understanding, and, by grouping people with 'similar' traits into best-fit groupings, it effectively achieves this purpose. Nonetheless, being able to classify persons independently of their geography (hence individually) should reduce the collective error discussed above and allow for more meaningful clusters to be formulated. As discussed in section 4.8, both through the SuperProfiles example as put forward by Birkin (1995) and the independent analysis undertaken on the OAC, cluster labels that do not accurately describe the populations that they are attributed to can be seen as poorly representing the population. Although this may merely be down to poor interpretation of the data, if such collective error is apparent at the area level then accurate cluster descriptors or pen portraits are difficult to assemble. By classifying at the person level, one expects this process to be far simpler and hence clusters should be more homogeneous than under present area-based schemes.

5.3. Defining the Geographical Scope

Section 5.2 has clearly put forward reasons for constructing a classification at the individual level, plus explicitly defined the nature of the system, hence a general purpose system comprising census variables. In order to formulate a strong framework, a case study region (or regions) must be selected and carried forward to demonstrate how this process works. For reasons of familiarity, Leeds (UK) will be the primary case study region on which the methodology will be applied. The selection of Leeds as a case study region is sensible for two reasons; firstly, given the level of personal familiarity with the region and its social geography, and secondly given that only London (7.17 million) and Birmingham (977,099) had greater populations in 2001 (the data year chosen for this research) (CASWEB, 2001). This therefore ensures that the Leeds classification is conducted on enough records to deem it meaningful and with an ability to test the general robustness of the framework. Furthermore, with Leeds residing extremely closely to the UK national average for a myriad of fundamental census statistics (Rees, 2013, personal communication), it makes sense to utilise Leeds to see how well or otherwise the classification methods segment the city's population.

So to further test the methodology, a second area will also be subjected to the same process. The second area selected is Richmondshire (UK). Both Leeds

and Richmondshire were selected from within the Yorkshire region due to familiarity and to enable ease of validation. Both Leeds and Richmondshire represent sensible districts through which to apply this methodology due to their differing population sizes and different classifications on the Office for National Statistics' (ONS) Local Authority classification. Leeds, with a population (in 2001) of 715,402 is classified in the Cities and Services supergroup and Regional Centres subgroup whereas Richmondshire, in 2001, had an official population of 47,010 and fell into the ONS classification supergroup of Prospering UK and subgroup of Prospering Smaller Towns. With such different population sizes and structure and contrasting classifications on ONS' Local Authority divisions, both areas will test the level to which this research can accurately segment people-types. Table 5.1 presents each of the districts / unitary authorities in the Yorkshire region (sorted by population size, smallest to largest) and clearly shows how the two selected areas differ - both in terms of basic demographics and regional classification. For ease of presentation, the final tier of the ONS classification has been removed from Table 5.1 given that it offers no new cluster naming over and above tier 2. Tier 3 naming conventions include 'Prospering Smaller Towns - A', 'Prospering Smaller Towns - B', 'Prospering Smaller Towns - C' etc.

Zone Name	Area Type Metropolitan District (NM Dist), Non-Metropolitan District (NM Dist) or Unitary Authority (UA)	AC Supergroup Label (Tier 1)	AC Group Label (Tier 2)	Population (Smallest to Largest)
Richmondshire	**NM Dist**	**Prospering UK**	**Prospering Smaller Towns**	**47010**
Ryedale	NM Dist	Coastal and Countryside	Coastal and Countryside	50872
Craven	NM Dist	Coastal and Countryside	Coastal and Countryside	53620
Selby	NM Dist	Prospering UK	Prospering Smaller Towns	76468
Hambleton	NM Dist	Prospering UK	Prospering Smaller Towns	84111
Scarborough	NM Dist	Coastal and Countryside	Coastal and Countryside	106243
Harrogate	NM Dist	Prospering UK	Prospering Smaller Towns	151336
North Lincolnshire	UA	Mining and Manufacturing	Manufacturing Towns	152849
North East Lincolnshire	UA	Mining and Manufacturing	Manufacturing Towns	157979
York	UA	Prospering UK	Prospering Smaller Towns	181094
Calderdale	M Dist	Cities and Services	Centres with Industry	192405
Barnsley	M Dist	Mining and Manufacturing	Manufacturing Towns	218063
Kingston upon Hull, City of	UA	Mining and Manufacturing	Industrial Hinterlands	243589
Rotherham	M Dist	Mining and Manufacturing	Manufacturing Towns	248175
Doncaster	M Dist	Mining and Manufacturing	Manufacturing Towns	286866
East Riding of Yorkshire	UA	Prospering UK	Prospering Smaller Towns	314113
Wakefield	M Dist	Mining and Manufacturing	Manufacturing Towns	315172
Kirklees	M Dist	Cities and Services	Centres with Industry	388567
Bradford	M Dist	Cities and Services	Centres with Industry	467665
Sheffield	M Dist	Cities and Services	Regional Centres	513234
Leeds	**M Dist**	**Cities and Services**	**Regional Centres**	**715402**

Table 5.1. Districts and Unitary Authorities (Local Authorities) within the Yorkshire region sorted by population size. Red shading denotes two selected areas.

Furthermore, with both Leeds (Metropolitan District, West Yorkshire) and Richmondshire (Non-Metropolitan District, North Yorkshire) being situated in different parts of the region, they are also geographically dissimilar. The Euclidean distance between the centroids of the two districts is close to 40 miles. See Figure 5.2 for a visual representation of the location of both case study districts relative to the rest of the region.

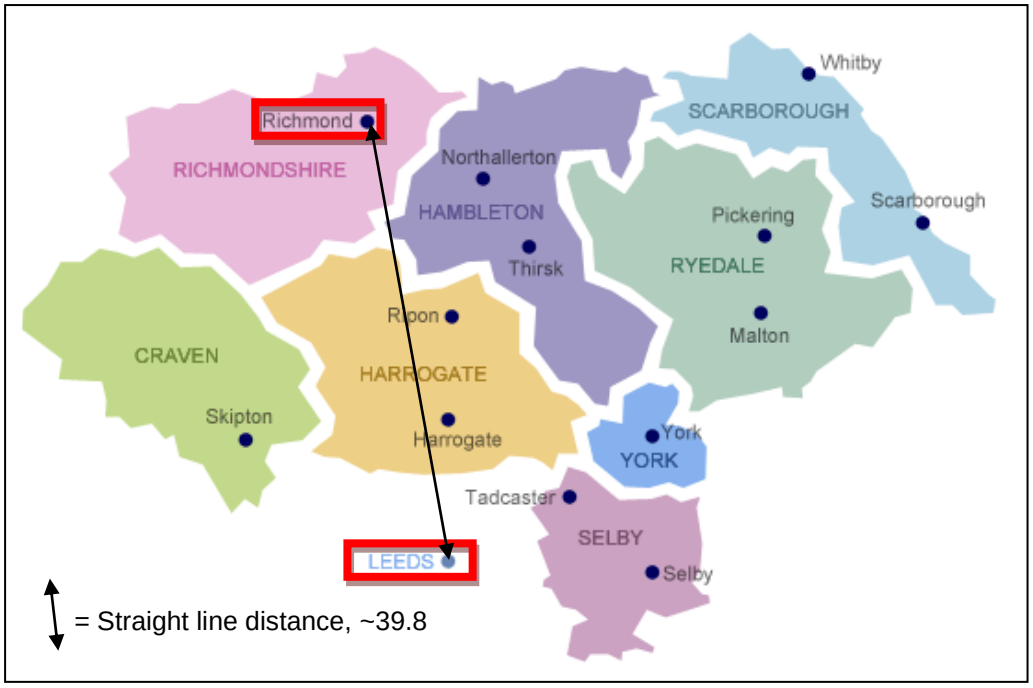


Figure 5.2. The North Yorkshire Region and Leeds: Showing locations of Leeds and Richmondshire (in red) (Adapted from Invest North Yorkshire, 2013).

5.4. Selecting the Raw Data Source

Once a purpose and rationale for the classification have been established and a geographical region determined, the variables for use in the classification must be chosen. In this research all variables will be census characteristics and will come from the Small Area Microdata (SAM) file. SAM is the term used for an individual-level sample of anonymised records from the 2001 Census (ONS, 2008). The SAM is similar to the SAR (Sample of Anonymised Records) with regards to variable inclusion, however, broader banding is adopted to preserve individuals' confidentiality given the personal nature of the multivariate dataset. Furthermore, the SAM provides a finer geography-level to aid analysis releasing data at the local authority level (as opposed to government office region in the standard SAR). The SAM sample accounts for 5% of the population and contains circa 2.9 million records from people in the UK (and ~35,000 for the Leeds metropolitan district (SAM code 67) and ~2,350 from Richmondshire (SAM code 279)) (CCSR, 2001). Across the SAM, each record is identifiable down to local authority level across a broad range of census topics, including; employment, personal demographics (such as age, sex, ethnicity etc) and residential arrangements (*ibid*). Northern

Ireland is an exception, however, and is only available at Parliamentary constituency level.

The data held in this file varies by type and each variable categorised into one distinct category of (1) individual, (2) household, or (3) family. The file contains a total of seventy-four real variables, one unique identifier, plus thirteen ONS / DEFRA variables and additional imputed variables - the latter not considered for the purpose of this classification (see Appendix A for comprehensive list). Some of variables include; number of cars / vans owned or available for use (household category), presence / number of dependent children in family (family category) and fundamental demographic variables such as age / sex / ethnicity / social-economic classification of respondents (individual category).

Such is the nature of individual-level data when compared to small-area aggregations, which are also controlled for data disclosure using similar methods, great care is taken to ensure that no information is included which would allow an individual to be identified (*ibid*). The SAM does, however, seek to maintain a balance whereby an optimum point is reached at which information is maximised whilst ensuring the risk of disclosure remains negligible (CCSR, 2009). For this reason, SAM data are ideal for such fine-level classifications and will be incorporated in this research.

The SAM and SAR files are currently held by the UK Data Service in Essex and are available free of charge to researchers in academia and for a small fee to those in a business or local authority environment (UKDS, 2014). With 2011 individual-level data unavailable at the time of this project, this research adopts 2001 data under the premise that future census iterations (if applicable) can be used to update the classification. However, with future censuses still under consideration, proposing a framework through which such segmentation can take place is of primary concern.

Alternative datasets were considered for use, including the British Household Panel Survey (BHPS) (now Understanding Society). However, given than such a dataset contains only 10,300 individuals (at wave 1) from 250 areas of Great Britain, the SAM represents a far more comprehensive selection. Other datasets also failed to compare for reasons of completeness (hence surveys) and validity.

5.5. Methods for Variable Selection

As with any geodemographic classification, one of the principle decisions which governs success or failure is the number and choice of input variables (see Section 4.4). However, this decision is made all the more difficult when constructing a general-purpose system, largely due to the fact that any such system lacks an overriding function through which input variables are expected to predict or describe.

Abbas (2009) presents work in the growing field of health geodemographics and any system designed to identify small areas by health risk is likely to make use of census variables such as self-reported health, limiting long-term illness and even the number of care hours provided. Clearly, a classification of this nature has a very clear purpose or question through which the classification is expected to inform and hence the selection of variables follows a logical procedure. Furthermore, any variable weighting decisions will also follow a similar process and be instigated as a result of the perceived level of importance.

For the purpose of this research, the final classification is expected to inform the Generative e-Social Science for Socio-Spatial Simulation project (GENESIS) and thus variables have been selected to closely align with the project's key themes (UCL, 2009). These themes, chosen to span key policy areas in social science, include planning and problem solving with respect to healthcare, housing, transportation and retail. Further themes including education, crime and employment will also supplement GENESIS' primary themes in a bid to ensure that variables are selected across the breadth of social science and therefore provide reason for inclusion within a broad general-purpose classification scheme. Table 5.2 illustrates the seven themes which arguably underpin social science in addition to British survey datasets which may be expected to best capture the associated phenomena.

Theme (A-Z)	GENESIS Project Theme	Generic Social Science Theme	Parallel-most Survey/s
Crime		✓	British Crime Survey (BCS)
Education		✓	Pupil Level Annual Schools Census (PLASC)
Employment		✓	Labour Force Survey (LFS)
Health(care)	✓	✓	Health Survey for England (HSE)
Housing	✓	✓	General Household Survey (GLS) / General Lifestyle Survey (GLS) British Household Panel Survey (BHPS)
Retail (and consumption)	✓	✓	Expenditure and Food Survey (EFS)
Transportation	✓	✓	National Travel Survey (NTS)

Table 5.2. Seven broad themes through which classification variables will be selected.

As can be seen in Table 5.2, it is possible to source a comparable survey dataset relative to each of the broad social science themes listed. One would expect each of these surveys to contain variables of a similar nature (for example, the British Crime Survey should provide details on the Crime theme). Each of these datasets are made available to academic communities free of charge subject to agreeing to the respective terms and conditions. One would expect the datasets listed to provide extensive details on the social scientific phenomena. Common datasets include the BHPS and Health Survey for England, amongst others. These survey datasets will each be used to determine the SAM variables put forward for inclusion in the individual-level classification. This process of selection will be conducted as described in section 5.6.

5.6. Selecting the Input Variables

In order to select the SAM variables for inclusion in the classification, each of the sixty-eight real variables contained in the SAM file was contrasted with the eight surveys listed in Table 5.2 and a matrix was formulated denoting the level of inclusion. For example, the variable ‘Accommodation Type’ (acctypa), as present in the SAM file, was searched for in each of the eight surveys. Depending on the number of times this variable was found (up to a maximum of eight), this value was recorded as that variable’s 'inclusion value'. This process assigned each SAM variable with a rank value to then assist with variable selection (ranging from 0 (no reference) to 8 (reference in all eight surveys)).

Table 5.3 shows this matrix and the resultant output from the rating process. Table 5.3 was sorted by SAM variable type; Individual (I), Household (H) or Family (F) as defined by the SAM data dictionary. A green tick in Table 5.3 denotes inclusion and a red cross denotes an absence in the survey datasets. The 'inclusion value' for each variable can then be seen in the right-hand most column.

A lookup table for SAM variable abbreviations/definitions can be found in Appendix A.1.

SAM Variable	Household (H), Individual (I), or Family (F) Variable Type	British H/hold Panel Survey	Labour Force Survey	Health Survey for England	British Crime Survey	Expenditure & Food Survey	General H/hold Survey	National Travel Survey	Pupil Level Annual Schools Census	Inclusion Value
Accommodation Type	H	✓	✓	✓	✓	✓	✓	✗	✗	6
Use of Bath/Shower/Toilet	H	✓	✗	✗	✗	✗	✓	✗	✗	2
Cars/Vans owned or available for use	H	✓	✗	✓	✗	✓	✓	✓	✗	4
Type of communal establishment	H	✗	✗	✗	✗	✓	✓	✗	✗	2
Central heating	H	✓	✗	✗	✗	✓	✓	✗	✗	3
Status in communal establishment	H	✗	✗	✗	✗	✗	✗	✗	✗	0
DEFRA: Urban/rural	H	✗	✗	✓	✓	✗	✗	✓	✗	3
No. Residents per room	H	✗	✗	✗	✗	✗	✓	✗	✗	1
Accommodation furnished (Scotland)	H	✓	✓	✓	✗	✓	✗	✗	✗	4
Household education indicator	H	✗	✗	✗	✗	✗	✗	✗	✗	0
Household employment indicator	H	✗	✗	✗	✗	✗	✗	✗	✗	0
Household housing indicator	H	✗	✗	✗	✗	✗	✗	✗	✗	0
Household health & disability indicator	H	✗	✗	✗	✗	✗	✗	✗	✗	0
Household headship	H	✓	✗	✗	✗	✗	✓	✗	✗	2
No. Carers in the household	H	✗	✗	✗	✗	✗	✓	✗	✗	1
No. Employed adults in household	H	✗	✗	✗	✗	✓	✓	✓	✗	3
No. of household members with LLTI	H	✗	✗	✓	✓	✗	✓	✗	✗	3
No. of household members with poor health	H	✗	✗	✓	✗	✗	✓	✗	✗	2
No. usual residents in household	H	✓	✗	✓	✗	✓	✓	✓	✗	5
Social grade of household reference person	H	✗	✗	✓	✓	✓	✓	✓	✗	5
ID within country	H	✗	✗	✗	✗	✗	✗	✗	✗	0
Lowest floor level of household living accommodation	H	✗	✗	✗	✗	✓	✗	✗	✗	1
Occupancy rating of household	H	✗	✗	✗	✗	✗	✗	✗	✗	0
ONS LA indicator	H	✗	✗	✗	✓	✗	✗	✗	✗	1
Relationship to HRP	H	✓	✓	✓	✗	✓	✗	✗	✗	4
Number of floor levels (N.Ireland)	H	✗	✗	✗	✗	✓	✗	✗	✗	1
Number of rooms occupied in household space	H	✗	✗	✗	✗	✓	✗	✗	✗	1
Accommodation self-contained	H	✗	✗	✗	✗	✗	✗	✗	✗	
Household with students away during term time	H	✗	✗	✗	✗	✗	✓	✗	✗	1
Tenure of accom. (country specific)	H	✓	✗	✓	✓	✓	✓	✗	✗	5
Family Type	F	✗	✗	✗	✗	✗	✗	✗	✗	0
Dependent children in family	F	✓	✓	✓	✗	✓	✓	✗	✗	5
Economic position of FRP	F	✗	✗	✓	✓	✓	✓	✗	✗	4
NS-SEC of FRP	F	✗	✗	✓	✓	✓	✓	✗	✗	4
Sex of FRP	F	✗	✗	✗	✓	✓	✓	✗	✗	3

Table 5.3. (part 1): Sixty-eight SAM variables (2001) and their presence or non-presence in other survey datasets (sorted in order of appearance in the SAM file).

SAM Variable	Household (H), Individual (I), or Family (F) Variable Type	British H/hold Panel Survey	Labour Force Survey	Health Survey for England	British Crime Survey	Expenditure & Food Survey	General H/hold Survey	National Travel Survey	Pupil Level Annual Schools Census	Inclusion Value
Age of respondents	I	✓	✓	✓	✓	✓	✓	✓	✓	8
County of birth	I	✓	✓	✗	✓	✗	✓	✗	✗	4
Community background - religion or religion brought up in (N.Ireland)	I	✓	✗	✗	✗	✗	✗	✗	✗	1
Country	I	✓	✓	✗	✓	✓	✓	✗	✗	5
Distance of move for migrants	I	✗	✗	✗	✗	✗	✗	✗	✗	0
Distance to work (inc. study in Scotland)	I	✗	✗	✗	✗	✗	✗	✗	✓	1
Economic activity (last week)	I	✓	✗	✓	✓	✓	✓	✗	✗	5
Ethnic group (country specific)	I	✓	✓	✗	✓	✓	✓	✓	✓	7
Ever worked	I	✓	✓	✓	✗	✓	✓	✗	✗	5
Generation indicator	I	✗	✗	✗	✗	✗	✗	✗	✗	0
General health over the last 12 months	I	✓	✗	✗	✗	✗	✓	✗	✗	2
Hours worked weekly	I	✓	✓	✓	✓	✓	✓	✗	✗	6
Local authority (GB) or Parliamentary Constituency (N.Ireland) [or other geography]	I	✓	✓	✗	✗	✓	✓	✓	✓	6
Year last worked	I	✓	✓	✗	✗	✓	✓	✗	✗	4
LLTI	I	✗	✗	✓	✗	✗	✓	✗	✗	2
Marital status	I	✓	✓	✓	✓	✓	✓	✓	✗	7
Migration indicator	I	✗	✗	✗	✗	✗	✗	✗	✗	0
Region of origin	I	✗	✗	✗	✗	✗	✗	✗	✗	0
NS-SEC (8 classes)	I	✓	✓	✓	✓	✓	✓	✓	✗	7
Record identified within country	I	✗	✗	✗	✗	✗	✗	✗	✗	0
Population base qualifier	I	✗	✗	✗	✗	✗	✗	✗	✗	0
Professional qualification (England and Wales)	I	✓	✓	✓	✓	✗	✓	✗	✗	5
Number of hours care provided per week	I	✓	✗	✗	✗	✗	✓	✗	✗	2
Level of highest qualification (16-74) (country specific)	I	✓	✓	✓	✓	✗	✓	✗	✗	5
Region of usual residence	I	✓	✓	✗	✗	✓	✓	✗	✗	4
Religion (country specific)	I	✓	✗	✗	✗	✗	✗	✗	✗	1
Sex	I	✓	✓	✓	✓	✓	✓	✓	✓	8
Schoolchild / student in full-time education	I	✓	✗	✗	✓	✗	✓	✗	✓	4
Supervisor/Foreman	I	✗	✗	✗	✗	✗	✗	✗	✗	0
Term-time address of students / schoolchild	I	✗	✗	✗	✗	✗	✗	✗	✗	0
Transport to work, UK (inc to study in Scotland)	I	✓	✗	✗	✗	✗	✗	✓	✓	3
Size of workforce	I	✓	✓	✗	✗	✗	✓	✗	✗	3
Workplace	I	✗	✗	✗	✗	✗	✗	✓	✓	2

Table 5.3. (Part 2): Sixty-eight SAM variables (2001) and their presence or non-presence in other survey datasets (sorted in order of appearance in the SAM file).

An inspection of Table 5.3 suggests that very few variables possess particularly high inclusion values. Only two SAM variables were contained in each of the eight survey datasets and unsurprisingly, these are fundamental person-level characteristics; the age and sex of the respondent. To the contrary, it is surprising that 25% of the SAM variables were not included across any of the surveys, however, one must be aware that such data do include more unconventional variables such as 'Population Base Quantifier' and 'Supervisor/Foreman' and may be defined subtly differently by the respective survey data dictionaries. The latter mentioned variables are only included in the list for the purpose of completeness. Such variables will not be considered for inclusion in the final classification.

Table 5.4 summarises the counts given in Table 5.3 though cumulative summation where X = total number of survey datasets.

Threshold (X = Number of Survey Datasets)	**Number of SAM Variables**	**Percentage (%) Inclusion (to two decimal places)**
x=8	2	2.94
x≥ 7	5	7.35
x≥ 6	8	11.76
x≥ 5	17	25.00
x≥ 4	26	38.24
x≥ 3	33	48.53
x≥ 2	41	60.29
x≥ 1	51	75.0
x≥ 0	**68**	**100.00**

Table 5.4. SAM variable inclusion values across all eight survey datasets.

As stated by Openshaw and Wymer (1995), variable numbers should be kept to a minimum; one should test for multi-collinearity and not be afraid to make selections based on intuition or, according to Gibson and See (2006), trial and

error (iteration). For the purpose of demonstrating this framework, any SAM variable with an inclusion value equal to or greater than five is used (thus, a variable present in at least half of the survey datasets). The classification therefore comprises seventeen independent person-level variables (~23% of the SAM file) as listed in Table 5.5. In addition, variables defined as Scotland, Wales or Northern Island are included here should the framework need to be applied to alternative regions of the UK outside of England. The list also includes the unique record identifier. Further rationale for the selection process is provided in section 5.5.

Variable	Type
Country	Nominal
Record identifier within country	Nominal
Local authority (GB) or parliamentary constituency (NI)	Nominal
Age of Respondents	Interval
Cars/Vans Owned or Available for Use	Interval
Central Heating	Dichotomous
Country of Birth	Nominal
Ethnic Group for England and Wales	Nominal
Ethnic Group for Northern Ireland	Nominal
Ethnic Group for Scotland	Nominal
Family Type	Nominal
General Health Over the Last Twelve Months	Nominal
Number of Usual Residents in Household	Interval
Hours Worked Weekly	Interval
Marital Status	Nominal
Relationship to HRP	Nominal
Number of Hours Care Provided per Week	Interval
Level of Highest Qualifications (Aged 16-74, EWN)	Ordinal
Level of highest qualifications (16-74)	Ordinal
Sex	Dichotomous
NS-Social Economic Classification - 8 Classes	Nominal

Table 5.5. Selection of variables for SAM individual-level classification.

It is important to note that each of the variables put forward for adoption in the classification (Table 5.5.) are present in the BHPS. As discussed later in section 5.12, this is fundamental to the proposed framework. Also fundamental are the data types, again shown in Table 5.5, and discussed in detail in section 5.10.

The variables in Table 5.5 were also compared to those included in the Output Area Classification (see Vickers, 2006) plus other more purpose-specific classifications which focus on subject-areas closely aligned to GENESIS' key themes (Crime, Education, Employment, Health(care),

Housing, Retail (and consumption), Transportation). Examples include Abbas *et al.* (2009), Burns (2009) and Picket and Pearl (2001). It was determined that only (1) Relationship to Family Reference Person, (2) Sex and (3) Number of Hours Worked per Week were not present in the OAC (2001) but they were present in related classifications (e.g. Abbas *et al.*, 2009, Burns (2009) and Picket and Pearl (2001)). Thus the choice of variables in the individual-based classification presented in this research aligns closely to previous area-based classifications that use census data.

5.7. Proposed Technique/s for Classification

The classification for this research will be undertaken through implementing a cluster algorithm, as opposed to any simple index calculation (like the IMD). Clustering can be defined as the process of organising objects in a database into clusters (or groups) such that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity (Kaufman and Rousseeuw, 1990). See section 2.2 for a greater discussion of clustering techniques.

Given the dependence on individual-level data in this research, it is difficult to make use of standard classification techniques which partition data to form homogeneous clusters, for example K-Means, in their standard form. Furthermore, the variables in the SAM are of three kinds: dichotomous, categorical – nominal and categorical – ordinal, which do not lend themselves to simple K-Means operating on continuous variables only. The definitions of these data types are presented in Sections 5.7.1 to 5.7.3.

5.7.1. Nominal (Categorical)

A nominal variable is represented by categories with no intrinsic ranking (from the point of view of intensity); for example, nominal variables contained in the SAM include religious affiliation and ethnic group. Further common examples of nominal variables include region, postcode, or any other geographical reference (SPSS Log, 2006).

5.7.2. Ordinal (Categorical)

A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (SPSS Log, 2006); for example, ordinal variables in the SAM include age and highest level of qualification. It should be noted that

although such variables represent some degree of polarity, it is not usually possibly to conduct arithmetic operations with them as it depends on the relations among categories (Rezanková, 2009).

5.7.3. Dichotomous

Dichotomous variables are often coded by the values zero and one and contain two opposite categories. Although dichotomous variables are often likened with nominal data, for similarity measuring it is necessary to take into account whether the variables are symmetric or asymmetric. To use SAM examples, in the first case, both categories have the same importance, for example Sex: male and female. In the second case, one category is regarded as more important (depending on the classification purpose), for example, ever worked: yes or no (Rezanková, 2009).

5.8. Why Not Conventional K-Means?

As mentioned in section 5.7, conventional classification methods are inappropriate in their common form due to the data types contained within the SAM file. K-Means clustering is widely used for partitioning large datasets into homogeneous clusters and is often the first technique considered when pursuing a cluster-led approach (Jain and Dubes, 1988). K-Means is defined as an iterative relocation algorithm based upon an error sum of squares measure (*ibid*). The fundamental operation of the algorithm is to move a case from one cluster to another to see if the move would enhance the sum of squared deviations within each cluster (Aldenderfer and Blashfield, 1984 cited in Vickers, 2008). The case will then be allocated (or re-allocated) to the cluster to which it brings the maximum improvement. The next iteration takes place when all the cases have been processed. A stable classification is therefore achieved when no moves occur during a full iteration of the data. After clustering is complete, it is then possible to inspect the means of each cluster (or cluster centres) for each case in order to gauge the distinctiveness of the clusters (Everitt *et al.*, 2001 cited in Vickers, 2008).

Despite K-Means' widespread usage within geodemographics and clustering in general, it is not always applicable. Batcher (2000) lists three disadvantages with the technique. Firstly, variables must be commensurable (Fox, 1982 cited in Batcher, 2000). This means that interval, ratio or any non-continuous variable must be reconfigured in scale format. Secondly, each

pattern (or case) is assigned deterministically to one cluster and one cluster only (hence 'crisp' geodemographics) and thirdly, no accepted statistical basis exists for the technique, although it seems that many approaches are currently available, for example Bryant (1991).

It is the first of these three points, as raised by Batcher (2000), which is of greatest relevance in this research. As mentioned previously, conventional clustering algorithms are unable to handle data of differing types. K-Means classifications (and related methods) are capable of effectively handling continuous numerical data once standardisation and polarity have been ensured. Polarity, in this case, refers to incorporating only data that run in the same direction – hence where high values in all variables are positive and low values negative (excluding any variables than may be regarded as neutral). Such algorithms can also successfully work with ordinal data (hence a degree of ranking). These methods then proceed through a clustering algorithm to formulate X number of distinct clusters as defined by the user of the relevant software package (such as SPSS). However, such methods will not usually successfully classify data based on a combination of, for example, categorical ordinal and nominal variables. Batcher (2000) does propose a solution to this issue which involves the replacement of distances (as used in K-Means) with probabilities to form a probabilistic clustering model, however, the assessments were less than favourable.

San *et al.* (2004) also present an alternative extension of the K-Means algorithm to ensure the successful clustering of mixed data; categorical (nominal and ordinal) and scale / numerical. This method introduces a new notion of cluster centres called *representatives* for categorical objects. This works on the basis that arithmetic operators are totally redundant with categorical objects and thus fuzziness is applied to define representatives instead of arithmetic cluster means (*ibid*). With this notion, the authors state how it is also possible to formulate a clustering procedure of categorical objects as a partitioning problem in a fashion similar to that of K-Means clustering.

As K-Means in its conventional form is not appropriate, two other classification approaches were considered here; Decision Tree Classification and Adapted K-Means Classification. Sections 5.9 and 5.10 will discuss both of these approaches in turn.

5.9. Decision Tree Classification

With the variables contained in the SAM spanning both nominal (e.g. family type) and ordinal (e.g. age) data-types, and even variables such as NS-SEC which, although nominal by design, are perceived to follow an ordinal structure, it is important to make use of a technique capable of handling such categorical data and decision trees are one such approach.

There has been considerable work in the field of 'decision tree learning' with regards to variable categorisation and cluster analysis when working with data of differing types, including some work in a geographical context. Rokach and Maimon (2005) present an assessment of modern-day methods for using decision trees as a means of classification.

Rokach and Maimon (2005) discuss how decision trees are structured in a data mining context; "*A decision tree is a classifier expressed as a recursive partition of the instance space*" (p.2). This means that decision trees conduct a separation of all available examples until a solution is reached. A tree's structure consists of nodes which form a rooted tree, meaning it is a directed tree with a node called a root that has no incoming edges. All remaining nodes have just one incoming edge. A node with an outgoing edge is termed an internal node (hence, within the tree structure – below the root itself). And any remaining nodes are classed as leaves (*ibid*).

In the decision tree, each internal node splits the instance space (or available examples) into two or more subspaces according to a certain function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range (*ibid*).

Each leaf is allocated to one class representing the most suitable target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target value having a specific value.

Instances are classified by navigating from the root of the tree down to a leaf, according to the outcome of the tests along the path (*ibid*).

Such an approach would seem a sound choice given its ability to (1) handle data of differing types (including data not in continuous / scale format) and (2) successfully partition data until effective clusters are reached. Hence, it has an ability to work with data applicable to this research project and determine best-fit groupings through a process of cluster analysis.

5.10. Adapted K-Means Classification

The second proposed approach concerns the use of conventional K-Means classification methods albeit using modified data. For example, Huang (1998) proposes two algorithms which extend the K-Means algorithm to categorical domains and domains with mixed numeric and categorical values but concludes that the complexities associated with such approaches are not necessarily worth the time expense when contrasted with alternative approaches.

The point made by Huang (1998) is that any data of a categorical nature is very difficult to incorporate into a K-Means algorithm in its original state. In this research, all SAM variables are coded in this way, for example, Marital Status is nominal; 1 = Single, 2 = Married, 3 = Separated / Widowed / Divorced. Running such values through a hierarchical K-Means algorithm will not produce effective results, largely due to the fact that the data are categorical and clustering will take place based on the value alone which is unlikely to be indicative of the individual as it does not have any direct meaning. Furthermore, the mathematical difference (magnitude) in value between 1 and 2 (Single and Married) and 2 and 3 (Married and Separated / Widowed / Divorced) is the same and thus the clustering algorithm will partition the data with equal separation between categories even if, in reality, the difference is likely to be more pronounced. For example, a person attributed with a value of 2 (Married), based on clustering nominal data, is equally like somebody who is 1: Single and 3: Separated / Widowed / Divorced. In reality, one may expect persons Married or Separated / Widowed / Divorced to be more like each other than somebody defined as Single.

If three clusters were formed based on the above example (one variable) and adopting the nominal data structure, one may except results similar to those given in Figure 5.3 (a). However, in reality the data should identify how close or otherwise people in the groups are relative to one another and look more

like Figure 5.3 (b). In order to achieve this, an independent variable must be employed to convert certain data types (nominal and sometimes ordinal) into a continuous format to enable the effective clustering of individuals.

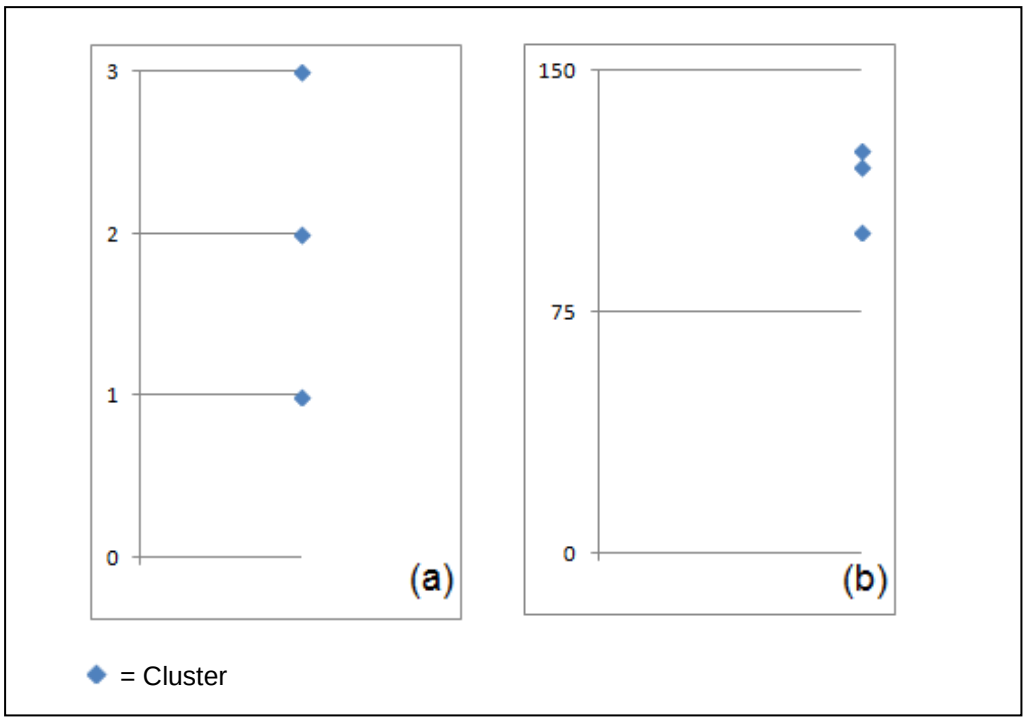


Figure 5.3. (a) Nominal marital status data clustering and (b) Continuous marital status data clustering.

The above discussions and illustration given in Figure 5.3 consequently provides reason to convert the SAM data into a continuous format based on an independent variable and adopt an adapted K-Means approach.

5.11. Completing the Categorical to Continuous Conversion Process

In order to make the transition from categorical to continuous data, various openly available independent variables were considered. Variables reflecting income and/or wealth were favoured given the need for a continuous scale. With no such variables present in the census (excluding proxy measures), the British Household Panel Survey (BHPS) was selected. The 'Monthly Gross Income' variable is the fundamental variable from within this dataset that reflects income (BHPS ref: RPAYG) and hence this was selected over other less complete options such as 'Last Payment Received', 'Income in X Month', 'Annual Labour Income', 'Household Income', etc. The variable selected for this process is most complete with regards to individual responses. Gross Monthly Income also offers the possibility of re-scaling to annual (estimated) income if necessary.

Each of the variables selected to form this classification were transformed into monetary values based on this BHPS variable. This was achieved by sourcing the equivalent SAM variable in the BHPS and recording the average gross monthly income for persons falling into each category. Inevitably, in order for this process to work efficiently, each of the variables selected for use in the classification had to be present in the BHPS and this was also taken into account at the variable selection phase.

Table 5.6 illustrates the original categorical data and Table 5.7 displays the results of this transition process. This process was largely automated through a program developed in Fortran designed to look-up and replace values in large datasets.

It should be noted that the data displayed in Table 5.6 and Table 5.7 is only a subset of the data to be used in the classifications (in this instance for Leeds), in terms of the number of individual records and the number of variables. It does, however, give a clear indication of the results of the conversion process.

ID	Country	LA Code	Age	Sex	Car Ownership	Central Heating	Health	Marital Status
11283325	1	67	30	2	2	1	1	3
11283381	1	67	30	1	1	2	2	1
11283448	1	67	30	1	1	2	1	1
11283449	1	67	30	1	0	2	1	2
11283450	1	67	30	1	0	1	1	2
11284353	1	67	30	2	2	1	2	3
11284354	1	67	30	2	1	1	1	2
11284355	1	67	30	2	2	2	1	1

Table 5.6. Original SAM data prior to conversion to gross monthly income.

ID	Country	LA Code	Age	Sex	Car Ownership	Central Heating	Health	Marital Status
11283325	1,131.58	67	1,702.20	1,392.84	1,077.38	1,793.19	1,826.79	1,468.27
11283381	1,131.58	67	1,702.20	2,225.09	923.26	1,478.51	1,626,93	1,516.29
11283448	1,131.58	67	1,702.20	2,225.09	923.26	1,478.51	1,826.79	1,516.29
11283449	1,131.58	67	1,702.20	2,225.09	549.68	1,478.51	1,826.79	2,006.11
11283450	1,131.58	67	1,702.20	2,225.09	549.68	1,793.19	1,826.79	2,006.11
11283453	1,131.58	67	1,702.20	1,392.84	1,077.38	1,793.19	1,626,93	1,468.27
11283454	1,131.58	67	1,702.20	1,392.84	923.26	1,793.19	1,826.79	2,006.11
11283455	1,131.58	67	1,702.20	1,392.84	1,077.38	1,478.51	1,826.79	1,516.29

Table 5.7. Newly created gross monthly income values for each SAM category based on BHPS.

Continuing with the marital status example discussed earlier in this section, recall that the value '1' under Marital Status (Table 5.6) denotes that the individual listed in this record is Single (never married). The value '2' denotes Marriage, and the value '3' denotes that this person is Separated / Widowed / Divorced. Once converted to grossly monthly income, the results are as

follows; Single individuals have an average gross monthly income of £1,516.29 (Table 5.7), those married earn on average £2,006.11, whereas those who are now separated / widowed / divorced tend to take home around £1,468.27 gross income per month. As can be seen from these statistics, people who are legally defined as single are far more similar to those who are separated / widowed / divorced than they are to those involved in a marriage as far as income is concerned. The movement of data from a categorical (nominal) format into a continuous format is therefore hugely beneficial for clustering purposes. If clustering were to take place using the original categorical variables, naturally somebody defined as single (value 1) and somebody separated (value 3) would be regarded as being highly dissimilar on a 1 to 3 scale with individuals involved in a marriage (value 2) residing somewhere in-between.

One should also note that the gross income figures listed above are averaged across the entire population (including those out of work, e.g. under 16's, retired, unemployed) and are therefore lower than any average official salary estimates noted elsewhere. Without incorporating the entire population, the process would be misleading. Nevertheless, for the purpose of effective clustering, it is the magnitude and level of difference of values between groupings which is of most importance – more so than producing accurate salary estimates. This is a key point to raise as simply attributing individuals with estimates of income is not the principal purpose of this process.

In order to fully facilitate the process shown above, each of the SAM variables selected for inclusion in the classification were sourced within the BHPS. Once located, the variables were extracted together with the Gross Rate of Pay Per Month variable (ref: RPAYG variable in BHPS table RINDRESP). This then enabled the gross monthly income to be calculated by taking the average of the income variable once broken-down by SAM grouping.

A certain re-coding of the data was necessary due to the structure / aggregation of the data between the SAM and BHPS. For example, some variables in the BHPS are continuous (e.g. age, hours worked per week, number of care hours provided) and therefore needed to be aggregated up to match the categories as put forward by the SAM. This was a simplistic summation process. However, other BHPS variables are also categorical and

if the categorical variables in the BHPS fail to match the categorical variables in the SAM then some data matching is necessary – and this is not always straightforward. For example, the BHPS contains a far greater number of groupings for Marital Status than the SAM thus requiring some interpretation and matching. For example, taking Marital Status as an example once more, the SAM groups all individuals into one of three [legal] categories; Single (never married), Married / re-married and Separated (but still legally married) / divorced / widowed. The BHPS separates individuals into nine groupings – including several extra categories not covered by the SAM, for example: Living as a couple, Have a dissolved civil partnership, Separated from a civil partnership and Surviving partner of a civil partnership. Clearly, matching these individuals into the categories as put forward by the SAM required some decisions to be made. Nevertheless, by following these principles and making some informed decisions, each variable was transformed into a monetary value and therefore of a format more suitable for a K-Means algorithm. Figure 5.4 illustrates the process through which Marital Status within the BHPS was matched to that in the SAM. A similar matching process was also necessary for several other variables, including: Relationship to HRP (30 BHPS vs. 6 SAM categories), NS-SEC (33 BHPS vs. 8 SAM categories).

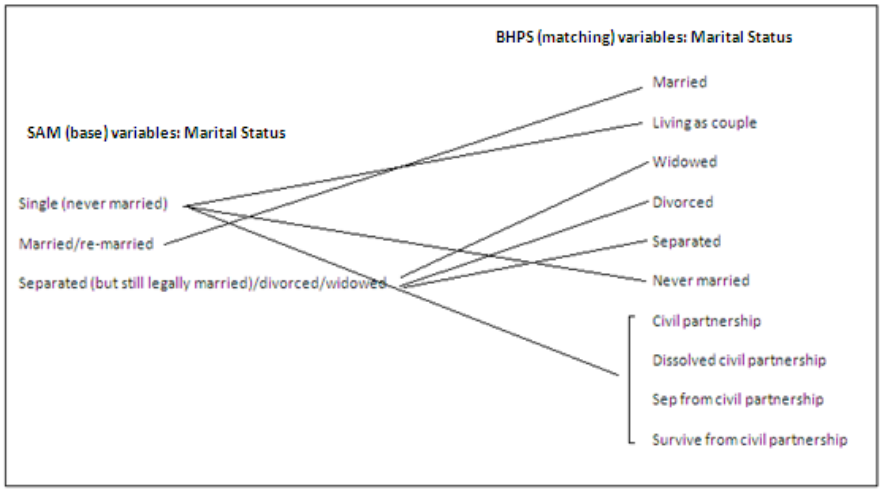


Figure 5.4. Example of problems faced when re(aggregating) data from BHPS to match SAM categories.

The completed table of results for each of the variables selected for the classification and their respective gross monthly income values can be seen in Table 5.8. For the purpose of ease of interpretation / reading, alternative rows have been made bold.

Table 5.8 clearly indicates a widespread distribution of gross monthly earnings across each of the variables and their separate aggregation categories. As mentioned previously, the monetary values are marginally less than one would expect individuals to earn per month in many cases, however, such is the need to encompass the full population within the classification, all earnings were averaged across the full spectrum of society and thus take into consideration those out of or ineligible to work (hence an income of zero – with any additional payments, e.g. benefits, excluded). The age variable is a perfect example of this and demonstrates that the chosen method does work and that the results reflect some degree of expectancy – across a normal distribution in this case.

Each of the variables listed in Table 5.8 display results that one would expect – at least with respect to the differences between variable categories. Socio-Economic Classification does show some unexpected variations as does Ethnic Group and Number of Hours Worked per Week, however, the latter can probably be explained by more menial employment requiring longer working hours and hence less overall pay. There is also some unexpected variation in Country (of residence), however, for the purpose of the England classifications, this does will not affect the results.

SAM Variable Ref#	SAM Variable Name	Average Monthly Gross Earnings (from BHPS, (re)aggregated based on SAM categories) [UK£]
2	Country	England: £1,131.58 **Scotland: £972.54** Wales: £1,209.62 **Ireland: £1,462.50** Other: £1,197.32
7	Age of Respondents	0-4: £0.00 **5-9: £0.00** 10-15: £0.59 **16-19: 411.34** 20-24: £839.86 **25-29: £1,467.78** 30-39: £1,702.20 **40-49: £1,837.32** 50-59: £1,582.41 **60-64: £1,107.10** 65-74: £547.64 **75-84: £21.31** 85+: £0.00
9	Cars/Vans Owned or Available for Use	3+: £1,021.09 **2: £1,077.38** 1: £923.26 **0: £549.68**
11	Central Heating	Yes: £1,793.19 **No: £1,478.51**
13	Country of Birth	England: £1,131.58 **Scotland: £972.54** Wales: £1,209.62 **Ireland: £1,462.50** Other: £1,197.32
20	Ethnic Group for England and Wales	White British: £1,138.49 **White Irish: N/A** Other White: £1,027.00 **Mixed – White & Black Carib/Black African/Black Other: £2,500** Mixed - White and Asian/Other Mixed/Other: £2,500 **Indian (Asian/Asian British): N/A** Pakistani (Asian/Asian British): N/A **Bangladeshi (Asian/Asian British): N/A** Other Asian (Asian/Asian British): £1,946.50 **Caribbean (Black/Black British): N/A** African (Black/Black British): £1,208.00 **Chinese: £944** Other N/A

Table 5.8. (Part 1): Results of categorical to continuous conversion process. Income values are gross per month and are extracted from latest wave (18) of BHPS. Bold vs. standard text on alternate is lines are simply to aid readability.

SAM Variable Ref#	SAM Variable Name	Average Monthly Gross Earnings (from BHPS, (re)aggregated based on SAM categories) [UK£]
24	Family Type	**Lone parent: £1,298.31** Married /cohabiting couple - no children: £1,952.67 **Married/ cohabiting couple – children: £1,761.74** Ungrouped individual (not in a family): £2,021.30
31	General Health Over the Last Twelve Months	Good: £1,826.79 **Fairly Good: £1,626,93** Not Good: £1,550.65
41	Number of Usual Residents in Household	0-1: £2,002.65 **2-4: 1,020.31** 5+: £496.26
42	Hours Worked Weekly	1-15: £393.24 **16-30: £1,021.02** 31-37: £2,197.04 **38-48: £2,111.21** 49+: £2,759.27
47	Marital Status	Single (nvr married): £1,516.29 **Married: £2,006.11** Sep/Div/Wid: £1,468.27
53	Number of Hours Care Provided per Week	0: £1,799.41 **1-19: £1,698.63** 20-49: £0.00 **50+: £0.00**
54	Highest Qualification	**No Quals: £1,172.70** Level 1: £1,244.60 **Level 2: £1,269.28** Level 3: £1,656.19 **Level 4/5: £2,602.94** Other: 1,654.20
60	Relationship to HRP	**Household Reference Person: £1,813.24** Husband or wife: £1,425.60 **Partner: £1,279.45** Son or daughter/ Step-child: £190.90 **Other related: £361.83** Unrelated: £466.47 **Unknown: £1,382.20**
64	Sex	Male: £2,225.09 **Female: £1,392.84**
74	NS-SEC 8 Classes	**Large employers & higher managerial occupations: £3,906.41** Higher professional occupations: £2,770.11 **Lower managerial and professional occupations: £2,263.30** Intermediate occupations: £1,411.17 **Small employers and own account workers: £0.00** Lower supervisory and technical occupations: £1,820.76 **Semi-routine: £989.87** Routine occupations: £1,105.30 **Never worked and long-term unemployed: £0.00**

Table 5.8. (Part 2): Results of categorical to continuous conversion process. Income values are gross per month and are extracted from latest wave (18) of BHPS.

5.12. Refinements to Data Structure

At this stage, data for both Leeds and Richmondshire were in a form ready for clustering following the data conversion process described in section 5.11. However, in line with previously stated observations made about the SAM age variable (see Table 5.8) and the gross income BHPS variable used to instigate this categorical to continuous conversion, some slight alterations were made to the data. Figure 5.5 illustrates how, under the current structure, differentiating between the young and elderly is difficult given that both are attributed very similar levels of gross (earned) income.

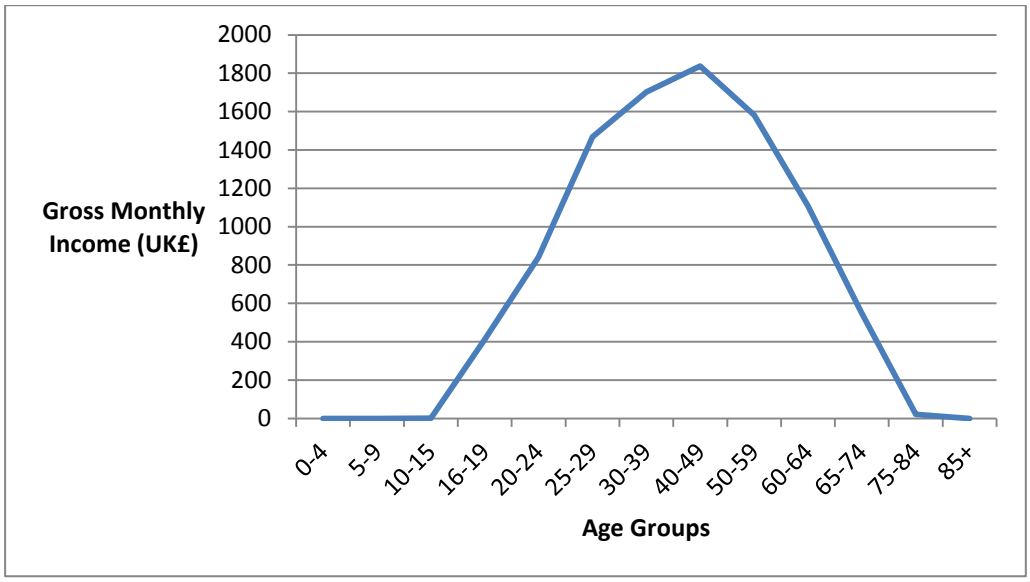


Figure 5.5. Graph illustrating results from age variable transformation into gross monthly income.

As can be seen in Figure 5.5, persons in age groups 0-4, 5-9 and 10-15 unsurprisingly do not earn form any income. The graph then proceeds in an upward direction whereby an individuals' gross monthly income increases by age, peaking at circa fifty years of age. Monthly earned income then gradually starts to decrease (excluding pensions, benefits etc) before falling in line with those in the early age groups when it reaches persons 85+. Due to the curved nature of the graph and the mirroring of average income across age groups, variables deemed to be ordinal in their structure and of a format suitable for clustering in their original forms will remain as such. The variables in question are: (1) Age, (2) Number of Hours Worked Per Week, (3) Number of Care Hours Provided, (4) Number of Residents in Household and (5) Number of Cars/Vans Available for Use.

With further irregularities identified in the income transition process for certain of the above named variables (for example, persons with access to 2 cars/van earning more per month than those with access to 3+ cars), this would seem like a rational decision.

For the purpose of completeness, it should be noted that in the SAM file all age groups are recorded by the first age to form that category. For example, somebody residing on the 0-4 age group is recorded as 0, somebody in the 5-9 age group is recorded as 5. Such values are deemed suitable for the purpose of illustrating this framework. With regards to 'Hours Worked per Week' and 'Number of Care Hours Provided', the values provided are linked to groupings, for example 1 refers to 1-15 hours. In cases such as this, the higher number of the category is utilised. Therefore, in the example 1-15 hours, fifteen is put forward for clustering. A similar process is employed with the variable 'Number of Usual Residents in Household'.

The National Socio-Economic Classification (NS-SEC 8 Classes) variable also underwent some refinements. This variable was considered suitable for classifying under its original data structure (despite being nominal by definition) in the same way as the five variables stated above (e.g. 1 - Large employers & higher managerial occupations, 2 - Higher professional occupations, 3 - Lower managerial and professional occupations etc). Leaving this variable in its original form was considered mainly due to the fact that two of the categories are incomplete. However, regardless of the fact that this variable is not designed to be hierarchical (ONS, 2010) it was decided to keep this in continuous monetary form and estimate the two missing categories based on a combined average of the two nearest categories, or, in the case of the final category, an overall positional average. Similar to the importance of gauging how (dis)similar individuals in the marital status category are, the same thought process applies here.

Table 5.9 displays the full collection of variables and the data structures adopted.

SAM Variable Ref#	SAM Variable Name	Data format	Format Recorded for Classification Illustration
2	Country	Gross Monthly Income	England: £1,131.58 **Scotland: £972.54** Wales: £1,209.62 **Ireland: £1,462.50** Other: £1,197.32
7	Age of Respondents	Original Data	0-4: 0 **5-9: 5** 10-15: 10 **16-19: 16** 20-24: 20 **25-29: 25** 30-39: 30 **40-49: 40** 50-59: 50 **60-64: 60** 65-74: 65 **75-84: 75** 85+: 85
9	Cars/Vans Owned or Available for Use	Original Data	3+: 3 **2: 2** 1: 1 **0: 0**
11	Central Heating	Gross Monthly Income	Yes: £1,793.19 **No: £1,478.51**
13	Country of Birth	Gross Monthly Income	England: £1,131.58 **Scotland: £972.54** Wales: £1,209.62 **Ireland: £1,462.50** Other: £1,197.32
20	Ethnic Group for England and Wales	Gross Monthly Income	White British: £1,138.49 **White Irish: N/A** Other White: £1,027.00 **Mixed – White & Black Carib/Black African/Black Other: £2,500** Mixed - White and Asian/Other Mixed/Other: £2,500 **Indian (Asian/Asian British): N/A** Pakistani (Asian/Asian British): N/A **Bangladeshi (Asian/Asian British): N/A** Other Asian (Asian/Asian British): £1,946.50 **Caribbean (Black/Black British): N/A** African (Black/Black British): £1,208.00 **Chinese: £944** Other N/A

Table 5.9. (Part 1): Final variables and their composition ready for classification. Bold vs. standard text on alternate are lines is simply to aid readability.

SAM Variable Ref#	SAM Variable Name	Data format	Format Recorded for Classification Illustration
24	Family Type	Gross Monthly Income	**Lone parent: £1,298.31** Married /cohabiting couple - no children: £1,952.67 **Married/ cohabiting couple – children: £1,761.74** Ungrouped individual (not in a family): £2,021.30
31	General Health Over the Last Twelve Months	Gross Monthly Income	Good: £1,826.79 **Fairly Good: £1,626,93** Not Good: £1,550.65
41	Number of Usual Residents in Household	Original Data (average)	0-1: 1 **2-4: 4** 5+: 5
42	Hours Worked Weekly	Original Data (average)	1-15: 15 **16-30: 30** 31-37: 37 **38-48: 48** 49+: 49
47	Marital Status	Gross Monthly Income	Single (nvr married): £1,516.29 **Married: £2,006.11** Sep/Div/Wid: £1,468.27
53	Number of Hours Care Provided per Week	Original Data (average)	0: 0 **1-19: 19** 20-49: 49 **50+: 50**
54	Highest Qualification	Gross Monthly Income	**No Quals: £1,172.70** Level 1: £1,244.60 **Level 2: £1,269.28** Level 3: £1,656.19 **Level 4/5: £2,602.94** Other: 1,654.20
60	Relationship to HRP	Gross Monthly Income	**Household Reference Person: £1,813.24** Husband or wife: £1,425.60 **Partner: £1,279.45** Son or daughter/ Step-child: £190.90 **Other related: £361.83** Unrelated: £466.47 **Unknown: £1,382.20**
64	Sex	Gross Monthly Income	Male: £2,225.09 **Female: £1,392.84**
74	NS-SEC 8 Classes	Gross Monthly Income (+ one averaged category)	**Large employers & higher managerial occupations: £3,906.41** Higher professional occupations: £2,770.11 **Lower managerial and professional occupations: £2,263.30** Intermediate occupations: £1,411.17 **Small employers and own account workers: £1,615.57** Lower supervisory and technical occupations: £1,820.76 **Semi-routine: £989.87** Routine occupations: £1,105.30 **Never worked & long-term unemp: £0**

Table 5.9. (Part 2): Final variables and their composition ready for classification.

5.13. Cluster Analysis

The clustering K-Means method was then applied in the same manner as outlined in section 4.6. The number of clusters was experimented with based on suggestions made by Milligan (1996) and Gibson and See (2006) until an optimal solution was reached. The results are presented and critiqued in chapters 6 and 7.

5.14. Supplementing with Small-Area Geography

As emphasised by Farr and Webber (2001), the benefits of moving from areal to individual-level classification are *"intuitively obvious" (p.58).* However, despite the distinct advantages, it is also important to emphasise the value of geography to the classification and steer clear of a sociological classification of individuals.

Despite the fact that a system classifying individuals is predicted to bypass many of the problems observed in area-based systems, it is important to consider the notion of space and this is best emphasised by Harrow *et al.* (1991). Harrow *et al.* (1991) discuss the social characteristics of neighbourhood on voting behaviour and conclude that the immediate area in which people live is strongly correlated with how they vote, and this relationship persists despite controlling for individual characteristics. Likewise, Rice and Sumberg (1997) emphasise how newspaper readership may be as much a function of education as it is community and thus, the decision on which newspaper to read is not merely a decision made solely by the individual but one impacted upon by various external pressures.

The above information is important in the context of this classification as not only is it necessary to maintain the geographical element given the research area but also a firm understanding is required of how an individual's characteristics arise. SAM variables such as 'general health over the last twelve months' may be influenced as much by environment as the individual themselves and this must not be overlooked. Classifications constructed at the postcode or output area level can easily identify such patterns, however, when working with SAM data which is referenced only to local authority level in England, Scotland and Wales and parliamentary constituency level in Northern Ireland, such observations are far easier to overlook. As a result, this classification must make use of a novel method to incorporate

neighbourhood into its findings. This will be achieved by linking the classification to a microsimulated dataset.

The datasets in question have been produced through combinatorial optimisation (simulated annealing) by Heppenstall (date unknown). In the case of Leeds, the data are complete (715,402 persons (2001)) and synthesised at output area level (2,438 areas) with constraint data acquired from the Census of Population via CASWEB (2001) and survey data courtesy of the British Household Panel Survey.

The sole purpose of this link is to attribute each member of the complete population a cluster code based on the classification generated on the modified SAM data. This will then ensure all members of the population have a cluster code (indicative of their behaviour, traits etc) and also an output area reference enabling the aforementioned notion of neighboured to be assessed. Should this be based purely on the SAM data, any analysis would be restricted to the local authority level and hence the influence of neighbourhood would be lost.

This link is instigated through converting all variables common between the microsimulated dataset and the SAM file into SAM-identical format (hence monetary income values or equivalent). Then, through a process akin to that of Sum of Squares, the cluster codes are matched across. This is instigated by calculating the Euclidean distance between the variables in the microsimulation and the cluster centres from the classification. Then the cluster with the minimum distance is assigned to that microsimulated individual.

5.15. Validation and Enrichment

The final phase of this research will undertake a combined validation and enrichment exercise. Given the predominant census make-up of the classification (supplemented by BHPS), the ability to link the final output to external non-census datasets will provide a means of profiling far deeper against more behavioural (or non-census) datasets. No only will this add value to the classification but it can also be used for validation. For example, one may expect any cluster categorised as being predominantly young, city-living types to being technologically advanced. Given that a system built

entirely on census variables cannot benchmark against such a variable, the ability to link the classification to survey datasets like the BHPS which contains variables of this nature adds real value. Furthermore, although the presentation of a framework to enable the construction of an individual-level general-purpose classification is the primary objective of this work, the ability to link (or 'bolt on') the classification to external datasets, such as those presented in Table 5.2, is a means of adopting the classification for a specific purpose should this be necessary. For example, attaching the classification to the British Crime Survey or National Travel Survey enables the classification to adopt a firmer focus. It also gives users of these external datasets an alternative method through which to view their data. Chapters 6-7 will demonstrate this statistical link and the added benefit this can generate - both in terms of validation and enrichment.

As with linking to the microsimulated dataset, this process will be achieved through statistical matching and the Sum of Squares technique and will be illustrated in forthcoming chapters.

5.16. Summary and Conclusions

This chapter has presented a detailed review of the framework through which an individual-level geodemographic system will be generated. Although loosely linked to the structure proposed by Gibson and See (2006) and discussed in detail in chapter 4, the framework comprises a series of phases ranging from data selection (in this case linked to ‘inclusion values’) and data preparation (with regards to SAM-adapted data).

The proposed framework is highly transferable and, given the sole reliance on census (SAM) data, can be applied readily for different regions of different data years (based on availability). The framework proposed handles data of differing types and makes use of income data from the BHPS to re-code categorical data where necessary in a bid to emphasise the differences that exist between sub-variables (for instance, the variation in marital status between those recorded as Single, Married and Separated/Widowed/Divorced). This data adaptation means that more conventional clustering methodologies can be employed to partition the data given its cluster-ready format. The transferability of the framework for both further research and industrial applications is further emphasised in chapter 8 where a flow diagram and process discussion is presented for wider adoption.

The following chapter will begin employing this framework on the case study regions of Leeds and Richmondshire. As discussed in section 5.3, Leeds represents a sensible district on which to apply the methodology given personal social geography familiarity and due to a meaningful population size. Richmondshire, due to its dissimilarity to Leeds, also represents a useful district on which to test the framework.

Chapter 6: Presenting the SAM Individual-Level Classification

6.1. Introduction and Chapter Preface

The purpose of this chapter is to present the results generated from following the framework set out in chapter 5. Specifically, this chapter will focus on classifying the populations of both Leeds and Richmondshire using the SAM-adapted data. This will be followed by as assessment of both areas' populations relative to the clusters produced.

Chapter 7 will then extend this analysis through the linkage to a microsimulated dataset (for complete population modelling) and external non-census datasets (for validation and enrichment).

6.2. Framework Validation

In line with the framework put forward in chapter 5, two classifications were developed, for Leeds and Richmondshire, using the cluster-ready SAM data. This cluster-ready (adapted) data refers to the newly coded data once transformed from categorical data into gross monthly income data (where applicable). Recall that certain variables, such as Number of Cars per Household, remained in their initial formats given the ordinal structure and data direction / polarity (hence 0, 1, 2, 3+ etc).

For the purpose of validating the framework, five clusters were agreed upon for both classification schemes. This decision was largely down to the data loss that is generally experienced when extending beyond a higher number of groupings, something illustrated by the percentage of Within cluster Sums of Squared (%WSS) and to ensure ease of comparability between districts. A test-run with five clusters also produced, on the most part, a sensible distribution of areas per cluster. Furthermore, when classifications with greater/less than five clusters were trialled, the results were far from satisfactory and often resulted in individual clusters being highly dominant and, in many cases, one cluster describing in excess of sixty percent of the individuals. For the purpose of completeness, it should be noted that the initial seed starting points were randomised during the K-Means classification process in a bid to ensure the classifications were not constructed based on

SPSS' default starting points. This process involved providing an additional file containing the starting seeds which SPSS could read prior to segmenting the data. Figure 6.1 illustrates how this process of seed randomisation is instigated in IBM SPSS (v.21). Multiple classification runs with different initial seeds were generated so as to enable the classification with the optimal cluster differentiation to be selected in both cases. In both cases, up to twenty runs were instigated and the most favourable outputs selected based on even cluster membership distribution and results.

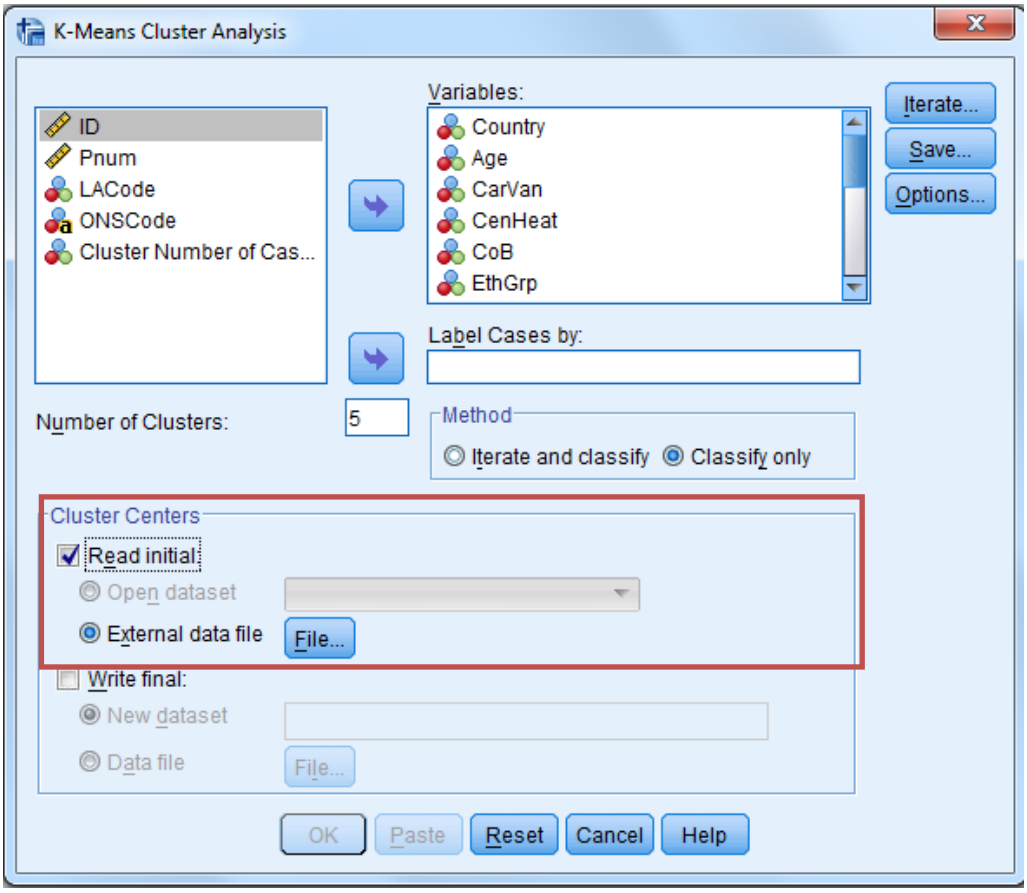


Figure 6.1. IBM SPSS (v.21) 'K-Means Cluster Analysis' window showing the facility to read-in a file specifying cluster centres.

An overview of the results of both the Leeds and Richmondshire classifications are shown in the following section.

It should also be noted that although the selection of five clusters is likely to affect the outcome of the classification, given that the overarching purpose of this research is to propose a framework through which individual-level classifications can be undertaken, the number of proposed clusters is far less important than a demonstration of methods operating as planned. Naturally, the level of differentiation and the extent to which the population is segmented

is dependent on the number of clusters in a geodemographic system, however, such a selection does not determine the success or failure of the underlying methods. Differentiation can be evident at various levels and this framework aims to demonstrate this using a five cluster approach.

6.3. Overview of Classification Results

In the results that follow, sixteen variables are eligible for classification having been adapted in line with the individual-level classification framework (chapter 5). The only variable not deemed to add value to either of the proposed classifications to be illustrated here is the Country variable. Given that Leeds and Richmondshire both reside in the country of England, the value attributed to this variable will not affect the classification outcome and has hence been removed. Should this framework be used to develop a UK-wide classification or to contrast two areas in different countries (England, Wales, Scotland or Northern Island) then such a variable is worthy of use - hence its inclusion as part of the wider framework. In this case, however, and in line with data redundancy rules, it has been removed. The two classifications that follow therefore segment populations based on the remaining fifteen individual-level variables as shown in Figure 6.2.

Age of Respondents
Cars/Vans Owned or Available for Use
Central Heating
Country of Birth
Ethnic Group for England and Wales
Family Type
General Health Over the Last Twelve Months
Number of Usual Residents in Household
Hours Worked Weekly
Marital Status
Relationship to HRP
Number of Hours Care Provided per Week
Level of Highest Qualifications (Aged 16-74, EWN)
Sex
NS-Social Economic Classification - 8 Classes

Figure 6.2. Fifteen variables used in SAM-adapted classifications of Leeds and Richmondshire, England.

Although all variables have been adapted and converted to monetary values in cases where an inherently categorical (nominal) structure was previously adopted (e.g. Marital Status), this conversion does not make the data entirely continuous. The conversion was designed to move categories away from using nominal data structures (e.g. 1= Single, 2= married, 3= Widowed/Separated/Divorced) and instead emphasise the differences

between categories (hence, replacing the above 1-3 codes with gross monthly income values to emphasise the differences / magnitude between sub-categories). For this reason, testing for multi-collinearity is not appropriate as the majority of variables are not in true scale format. Standard classification evaluation techniques are also more difficult (for example, benchmarking against a global average) given the format of the classification and the use of predominantly categorical variables (e.g. those with sub-categories as opposed to simple scale values on a linear trajectory). However, an approach similar to this is presented in a bid to understand cluster composition.

True evaluation of these classifications will come as a result of linking to a microsimulated dataset (to add finer-level geography, down to output area level) and external non-census datasets as shown in chapter 7. For example, do individuals from area X, described by this classification as being married, of high socio-economic classification and with 2+ cars, have a tendency to travel abroad (using a variable from, for example, the British Household Panel Survey) more frequently than those classified into a less favourable cluster, such as one described with high unemployment, low qualifications and poor health status?

Given that the above will be assessed in chapter 7, this chapter will assess each cluster based predominantly on an analysis of the final cluster centres. Such an assessment will be conducted on both the Leeds and Richmondshire classifications.

With regards to cluster membership, the output for both classifications is shown in Tables 6.1 and 6.2.

Cluster Number		Total Number of Cases	% of Cases per Cluster
Cluster	1	10772.000	29.9
	2	4738.000	13.2
	3	1488.000	4.1
	4	324.000	0.9
	5	18664.000	51.9
Valid		35986.000	100.0
Missing		0.000	

Table 6.1. Cluster membership, Leeds.

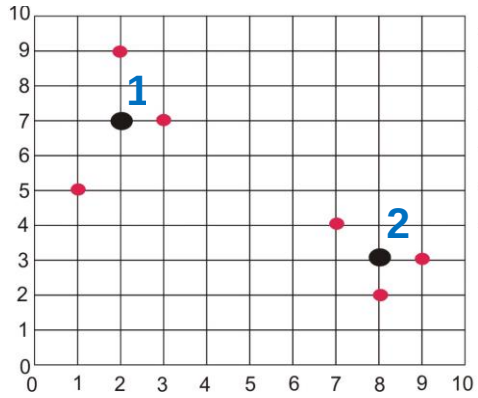
Cluster Number		Total Number of Cases	% of Cases per Cluster
Cluster	1	157.000	14.9
	2	92.000	8.7
	3	154.000	14.6
	4	178.000	16.9
	5	473.000	44.9
Valid		1054.000	100.00
Missing		0.000	.000

Table 6.2. Cluster membership, Richmondshire.

As can be clearly seen in Tables 6.1 and 6.2, Leeds has a greater number of cases (individuals) making up its population even when incomplete records are removed (as evidenced as part of Leeds' case study selection process). Incomplete records (of individuals) are defined as records where two or more of the variables are missing. In cases where one variable is omitted, a process of averaging across the region is undertaken to fill in the blanks. In cases of two or more omissions, the individual is removed to prevent over-averaging and hence the risk of creating a false predominantly homogeneous population – something one may argue is over-representative in area-based systems.

Cluster centres are an important way of analysing cluster composition. A cluster centre refers to the average of a set of observations in a given cluster

(hence 'means' in the term K-Means) where K refers to the number of clusters. Figure 6.3 illustrates this through diagrammatic representation.



Observation	% Unemployment	%2+ Cars
1	2	9
2	1	5
3	3	7
Cluster Center = Average	**2**	**7**

Figure 6.3. An illustration of cluster centres. Adapted from See (2009).

As can be seen in Figure 6.3, two clusters exist. The first, as detailed in the accompanying table, with average observations for percentage employment and percentage 2+ cars resulting in a centre of 2,7. This can easily be represented in two dimensional space given the inclusion of only two variables in the classification. When more variables are clustered, the process becomes more complex but functions in exactly the same way making use of average calculations of each observation in order to gauge the cluster centre in often multi-dimensional space.

With regards to final cluster centres for the two case study regions, these are displayed in Tables 6.3 and 6.4. A lookup table for the variable names is available in Appendix A.2 if required.

	Cluster				
	1	2	3	4	5
Age	37	12	27	34	62
CarVan	2	1	1	1	1
CenHeat	1763.04	1734.71	1710.29	1688.30	1727.87
CoB	1501.15	1501.15	1501.15	1501.15	1501.15
EthGrp	1240.7823	1321.7341	1332.7169	1292.8687	1324.4129
HholdFamTyp	1728.85	1732.57	1726.25	1777.04	1856.35
Health	1779.41	1798.71	1763.25	1753.40	1735.88
No.UsualRes	4	4	2	3	2
HrsWrkdWkly	37	42	40	37	39
MarStat	1789.32	1535.78	1750.31	1978.24	1773.27
HrsCareWkly	3	1	3	3	2
ReltoHRP	1597.48	241.13	446.55	1586.08	1664.01
Sex	1785.56	1784.42	1788.36	1781.15	1780.24
NS-SEC	2170.17	1708.64	867.35	732.26	1587.72
Qual	2491.07	1566.84	1304.71	1323.25	1344.75

Table 6.3. Final cluster centres for Leeds classification.

	Cluster				
	1	2	3	4	5
Age	38	56	43	26	36
CarVan	2	2	1	1	1
CenHeat	1785.17	1786.35	1776.84	1782.58	1783.21
CoB	1133.03	1131.83	1131.39	1144.84	1135.29
EthGrp	1137.78	1147.27	1137.77	1150.66	1141.37
HholdFamTyp	1808.52	1857.88	1860.43	1764.40	1820.60
Health	1784.71	1672.20	1760.53	1800.11	1780.27
No.UsualRes	4	1.99	2.03	2.19	1.98
HrsWrkdWkly	40	37	40	37	39
MarStat	1810.09	1816.82	1824.02	1468.15	1811.52
HrsCareWkly	1.16	1.10	1.10	1.12	1.13
ReltoHRP	1467.29	1686.80	1650.69	686.30	1705.85
Sex	1726.80	2107.49	1981.90	1748.18	1876.71
NS-SEC	2205.88	3449.42	.00	1226.53	1798.26
Qual	2596.91	2200.70	1268.58	1363.47	1337.65

Table 6.4. Final cluster centres for Richmondshire classification.

The results from the final cluster centres provide indications as to the population groups within each cluster. However, in order to fully understand the composition it is necessary to translate the data in Tables 6.3 and 6.4 from current monetary format (or original format in five cases) to something easier

to interpret. In cases where variables are scale (five cases - age, cars/van ownership, number of residents per household, hours worked per week and number of car hours provided), such variables can be readily interpreted, however, the re-categorised variables require translation.

6.3.1. Interpretation of the Results (Leeds)

The results presented in Table 6.5 define the translated final cluster centres (for Leeds) as discussed in section 6.3. These results are somewhat best-fit based on how each variable is best categorised and do not show how near or far a variable is from falling into a separate category (fuzziness). Although this process was automated in some cases, it does give us a qualitative indication as to how the clusters look in terms of predominant individual-level demographics.

	Cluster				
	1	2	3	4	5
Age	37	12	27	34	62
CarVan	2	1	1	1	1
CenHeat	Yes	Yes	Yes	Yes	Yes
CoB	England	England	England	England	England
EthGrp	White British	White British	White British	White British	White British
HholdFamTyp	Married/Child	Married/Child	Married/Child	Married/Child	Married
Health	Good	Good	Good/Fair	Fair	Fair
No.UsualRes	4	4	2	3	2
HrsWrkdWkly	37	42	40	37	39
MarStat	Married	Single	Single	Married	Married
HrsCareWkly	3	1	3	3	2
ReltoHRP	Husband/wife	Son/Daughter	Unrelated	Husband/wife	Husband/wife
Sex	M/F	M/F	M/F	M/F	M/F
NS-SEC	Large manag.	Small emps	Semi-Routine	Semi-Routine	Lower manag.
Qual	Level 4/5	Level 3	Level 2	Level 2	Level 2

Table 6.5. Final cluster centres for Leeds SAM classification translated to predominant variable sub-categories to add meaning.

From both Tables 6.4 and 6.5, it is possible to infer five quite distinct clusters. These are described overleaf through simple and descriptive pen portraits and cluster names designed to describe the typical residents in each grouping.

Cluster 1: Affluent Managers

This cluster is a middle-aged cluster with an average age of thirty-seven years. Typically households are quite affluent as reflected by access to two cars and being employed in large managerial capacities. Members of this cluster provide some weekly care for relatives and work typical hours. Members tend to be married and live in households with circa four people, likely to include children. Individuals in this cluster are typically of White British ethnicity and well educated.

Cluster 2: Young People living with Family

This cluster contains a youthful and healthy demographic with an average age of twelve years. These individuals live with their parents who are married, have good general health and are of White British ethnicity. The household has access to one car, is heated and on average houses around four people. They are the son/daughter of the head of household.

Cluster 3: Co-habiting Couples

This cluster is categorised by young individuals with start-up families. Members of this cluster tend to be single by legal definition but may be cohabiting. Individuals are in their mid/late twenties, have access to a car and work predominantly in semi-routine occupations with employment taking up to circa forty hours per week. Members have some education and are typically in good to fair health.

Cluster 4: Average Resident

This cluster is categorised by individuals in their mid thirties who are married with children. Health is recorded as fair and members have some education and working typical length weeks. Households typically contain three individuals with care provided for family on a weekly basis. Education levels are fair and access to a car is common.

Cluster 5: Nearing Retirement

This cluster contains an elderly demographic with a typical age of sixty-two. Members tend to be married without children at home and in fair health. Of those still working, most work in lower managerial occupations and have some

education. The average sized household is two persons with most married and of a White British Ethnicity.

6.3.2. Interpretation of the Results (Richmondshire)

In line with interpreting the Leeds results in section 6.3.1, the same will now be undertaken for Richmondshire. Table 6.6 presents the qualitative interpretation based on best-fit assignment to variable sub-categories. A descriptive insight into each cluster composition then follows.

	Cluster				
	1	2	3	4	5
Age	38	56	43	26	36
CarVan	2	2	1	1	1
CenHeat	Yes	Yes	Yes	Yes	Yes
CoB	England	England	England	England	England
EthGrp	White British	White British	White British	White British	White British
HholdFamType	Married/Child	Married	Married	Married/Child	Married/Child
Health	Good	Good/Fair	Good	Good	Good
No.UsualRes	4	2	2	2	2
HrsWrkdWkly	40	37	40	37	39
MarStat	Married	Married	Married	Single	Married
HrsCareWkly	1	1	1	1	1
ReltoHRP	Husband/Wife	HRP	Husband/Wife	Unrelated	HRP
Sex	M/F	M	M/F	M/F	M/F
NS-SEC	Lower manag.	Higher prof.	Unemp	Intermediate	Lower super.
Qual	Level 4/5	Level 4/5	Level 1	Level 2	Level 2

Table 6.6. Final cluster centres for Richmondshire SAM classification translated to predominant variable sub-categories to add meaning.

From this, as previously undertaken for Leeds, it is possible to infer five different clusters as summarised below by pen portrait description and a cluster name designed to describe the typical resident.

Cluster 1: Long hours, Middle-Management

This cluster is categorised by individuals in their late thirties who fall into the lower managerial socio-economic classification. Individuals tend to have access to two cars (per household), be of White British ethnicity and married with children who still live at home. Typically, members work in lower

managerial occupations working the average number of hours per week. Health status is generally good.

Cluster 2: Educated Professionals

Members of this cluster are in their mid/late fifties and have access to two cars. Health status is good to fair with most members being White British persons. The cluster has a heavy male presence although the majority of individuals are married. Level of education is high and members tend to work in higher professional occupations and work typical length weeks.

Cluster 3: Middle-Aged Unemployed

This cluster contains individuals in the early forties with good health. Members tend to live in two-person households, usually with spouse. Unemployment is widespread in this grouping although members do have access to one car. Education levels are low.

Cluster 4: Start-up Families

This is the youngest of the five clusters with members typically in their mid/late twenties with access to one car and starting families, although single by legal definition. Households tend to be two to three person residences and members have good general health. Employment is within the intermediate occupations domain and members have some education.

Cluster 5: Average Resident

This cluster contains members in their mid/late thirties with, on average, access to a single car. Members are of White British ethnicity with some education and work marginally above average length weeks in lower supervisory occupations. Health status is good and households are typically two-person in size, usually with spouse.

6.4. Discussion of Clustering Outcomes

As presented in sections 6.3.1 and 6.3.1, two five-cluster schemes have been overviewed for both case study districts of Leeds and Richmondshire. On both occasions, the final cluster centres (Tables 6.3 and 6.4) were interpreted and transformed into word-based descriptors (Tables 6.5 and 6.6) in cases where the numeric values were hard to interpret. This, in turn, led to a pen

portrait being developed for each cluster and, in most cases, an accurate cluster name tag designed to give a snapshot view of the cluster.

The tabular results and associated pen portraits describe distinctly different clusters in most cases. For Leeds, one can identify a youthful cluster with children living with parents ranging to a more elderly (but not retired) cluster containing individuals living with their spouse. It is clear that some variables differentiate the population particularly well and, in general, correlate with variables one would expect. For example, single co-habiting couples living in households with a size of two-persons and higher education leading to higher socio-economic classification (and employment) status. Such is the format of the data and efforts to make it somewhat continuous, these patterns of correlation do corroborate the framework adopted and in particular the SAM data adaptation that has taken place. However, with fully continuous datasets, such observations of multi-collinearity could have been identified at earlier stages through correlation or regression analysis. The data adopted in this research is not purely continuous/scale nor is it entirely categorical hence the selected novel approach.

Although some variables do appear to segment the population particularly well, for example age given the broad clusters produced, it is clear from the results that certain variables do not differentiate between the population groups to quite the same level. Such variables include sex, ethnic group, country of birth and central heating in particular. The inclusion of sex as a fundamental demographic characteristic was included given its very high inclusion value (when assessed based on its presence in eight survey datasets) and inclusion in other geodemographic systems. Ethnic group and country of birth were included for the same reason, however, suffered as a result of incomplete records and hence either a process of global averaging or omission. These variables may well differentiate populations more effectively in different areas assuming completeness or when employed on a wider basis, for example a nationwide classification. Central heating, again incorporated due to its inclusion score and as a proxy for deprivation, failed to differentiate in any cases and, similar to sex, may simply be due to its binary format - yes or no - with no variation or further sub-categories.

Of the remaining variables, most segment the population rather well. Age, as discussed, given its breadth of values, arguably forms the best segmentation

and this very much supports the decision to leave age in its original format and not convert to monetary values. Other variables in pure continuous format (e.g. those not transformed to monetary values given their suitable original format) also add value to the classification, in particular; number of residents per household, hours worked per week and cars/van available for use. The 'number of care hours provided per week' variable also suffers from incompleteness and small numbers and hence averaging/omissions. Of those variables transformed, marital status, socio-economic classification and highest-level of qualification are three that effectively partition the data and formulate distinct groupings.

As evidenced from Tables 6.5 and 6.6, some variables do contradict themselves. For example, within the Leeds classification in cluster 3, certain individuals were described as single by legal marital status definition but within a married with children family-type environment. Such contractions do not impact greatly on the ability of the classification to segment the population, however, for the purpose of any further classifications adopting this framework, it is a consideration to be taken forward and assessed as part of the variable selection process.

6.5. Summary and Conclusions

This chapter has shown how the framework put forward in chapter 5 leads to meaningful and largely effective clustering outcomes. The ability to handle both continuous variables (such as age) in addition to categorical variables (both ordinal and nominal) demonstrates the success of the framework in the data adaptation process. However, as evidenced in section 6.4, this route is by no means error free. These problems will be discussed in chapter 7 in addition to demonstrating the remaining phases of the framework, including linking to a microsimulated population and external datasets for reasons of modelling and validating/enrichment. The latter is a highly innovative focus of this research and hence demonstrates an ability to profile individuals against alternative data.

Chapter 7: Linking to Microsimulated and External Datasets

7.1. Introduction and Chapter Preface

The purpose of this chapter is to evidence the remaining phases of the framework not discussed in chapter 6. Specifically, this refers to linking the classification results to both a microsimulated dataset and other external datasets; the former to aid visualisation though supplementing the classification with finer-level geography and the latter to add value to the classification through enrichment but also to validate the results.

Following the above, a visual representation of the classification for Leeds will be presented in addition to examples of the link facilitated to external non-census data.

One fundamental aim of this chapter is to illustrate that the framework presented is robust enough to be carried forward and used in future individual-level classifications and modelling.

7.2. Linking to Microsimulated Dataset

As described in chapter 5, the individual-level classifications developed as part of this research are constructed based solely on the use of census data through the SAM file which is a 5% sample of records (CCSR, 2001). The SAM file, for reasons pertaining to risk of data disclosure, only geographically references individuals to the Government Office Region (GOR) to which they belong. England is divided into nine GORs (moving up from ten in 1998 when Merseyside was combined with the North West) (ONS, 2011). More recently, GORs have inherited the name 'regions' (as of April 2011) but still maintain the original GOR names in the 2001 SAM. Both classifications presented in chapter 6 reside in the Yorkshire and the Humber GOR (or Region). Reasons for this rather coarse identification are fully understandable but far from ideal when it comes to classification visualisation. The cluster membership and composition analysis presented in chapter 6 provides an indication as to the individual-level demographics and people types that inhabit Leeds and

Richmondshire based on this sample, however, given the geographical focus of this research, simple sociological identification of people types is only the beginning. By linking the classification to a microsimulated dataset containing small-area references, one will be able to provide a visual/spatial analysis through cartography and model the complete populations. This chapter will demonstrate this approach using the Leeds SAM.

In order to add this fine-level geography, the SAM classification will be linked to a microsimulated dataset. As initially referred to in chapter 3, the 2001 microsimulated data used in this analysis comprises six variable constraints; Age, Marital Status, Sex, Highest Qualification, Socio-Economic Classification, and Ethnicity and was formulated through combinatorial optimisation (simulated annealing) by Heppenstall (date unknown). The data are complete for Leeds (715,402 persons) and synthesised at output area level (2,439 areas) with constraint data acquired from the Census of Population via CASWEB [now UK Data Service, UKDS] (2001) and survey data courtesy of the British Household Panel Survey (BHPS) (more recently known as part of Understanding Society).

Inevitably, a dataset with such a fine geographical resolution supercedes one at the GOR (or Region) level for the purpose of visualisation and detailed analysis. Although this work has set out a framework to enable individual-level classification, given the inability to visualise at a level beyond the GOR unit, linking the classification codes (1-5) to the population data generated through the microsimulated dataset will enable each output area to be assigned a predominant (modal) cluster code. Furthermore, the relative popularity of each cluster per output area can also be calculated and spatially visualised as necessary (fuzziness).

As previously referred to in section 5.14, in order to complete this linking process, all variables common between the microsimulated dataset and the SAM file require translation into SAM-identical format (hence monetary income values or equivalent, as opposed to categorical data). Then, through a process of statistical matching using the Sum of Squares technique, the cluster codes can be transferred. This statistical link was achieved by calculating the Euclidean distance between the variables in the microsimulated dataset and the final cluster centres from the classification. Once complete, the cluster with the minimum distance was automatically

assigned to that microsimulated individual. Figure 7.1 clearly illustrates this process in addition to the formulae adopted as undertaken in Microsoft Excel.

As shown in Figure 7.1, the eight variables common between both datasets were extracted from the microsimulated dataset and the Euclidean distance calculated to the final cluster centres of the SAM classification. The final distance was divided by 10,000 in all cases to reduce the magnitude of the values and allow for ease of interpretation. Each individual was then assigned to its best-fit cluster based on the shortest distance. This allocation process was automated as evidenced in column O in Figure 7.1. This matching process was conducted on all 715,402 of Leeds' individuals hence classifying the entire Leeds population in to one of five distinct groups (see Figure 6.5 for a review of Leeds' cluster descriptors). For the purpose of visualisation, each output area (column A in Figure 7.1) can be used to aggregate individuals and produce cartographical representation showing the principal (most popular) cluster per output area.

	A	B	C	D	E
1	Area_code (OA)	PersonalID	Household_Size	Age	Sex
2	00DAFA0001	1	1	71	1393
3	00DAFA0001	2	2	26	2225
4	00DAFA0001	3	2	23	2225
5	00DAFA0001	4	1	76	1393
6	00DAFA0001	5	1	69	1393
7	00DAFA0001	6	1	36	2225
8	00DAFA0001	7	3	52	2225
9	00DAFA0001	8	3	71	2225
10	00DAFA0001	9	3	48	2225
11	00DAFA0001	10	2	67	1393
12	00DAFA0001	11	3	44	1393
13	00DAFA0001	12	3	36	1393
14	00DAFA0001	13	3	33	2225
15	00DAFA0001	14	4	35	1393

Illustrative example only - not all common variables shown here

K	L	M	N	O	P
Dist-1	Dist-2	Dist-3	Dist-4	Dist-5	Cluster
525.438	297.6633	16.05524	257.2207	17.50072	3
526.464	297.9999	21.5028	261.517	19.73483	5
526.471	297.9933	21.50891	261.5311	19.73656	5
524.723	298.3036	16.14557	256.8468	18.10877	3
524.673	298.2209	16.09264	256.8124	18.04561	3
53.4451	19.46609	300.7199	22.45764	298.9667	2
25.3755	86.30477	639.5446	108.4729	637.8145	1
529.129	301.61	21.87214	262.1434	20.76321	5
528.84	302.6068	22.17784	262.0912	21.59636	5
21.5658	83.7045	634.2216	103.5854	635.6287	1
21.6656	82.30182	633.7058	103.5495	634.5181	1
522.76	294.0871	15.6777	256.5596	16.4783	3
528.449	300.6883	21.37224	261.7121	20.15718	5
524.755	296.7488	15.55444	256.7833	16.89626	3

Variables common between microsimulated dataset and SAM Classification. Only a subset shown for purpose of illustration.

Individual person ID in Microsimulated dataset.

Output area code of each individual in microsimulated dataset.

Euclidean distances between microsimulated variables and final cluster centres from SAM classification.

In this analysis, the final cluster centre values (see Tables 6.3 and 6.4) were held in a separate Excel tab called 'FinalCC'.

Hence, the classification to calculate distances for cluster one (Dist-1), record one (cell K2) in this example was:

=(($C2-FinalCC!C$2)^2+($D2-FinalCC!C$3)^2+($E2-FinalCC!C$4)^2+($F2-FinalCC!C$5)^2+($G2-FinalCC!C$6)^2+($H2-FinalCC!C$7)^2+($I2-FinalCC!C$8)^2)+ ($J2-FinalCC!C$9)^2)/10000

Automated process adopted to determine nearest match.

Formula:

=MATCH(MIN(K2:O2),K2:O2,0)

Figure 7.1. Visual illustration of SAM classification to microsimulated dataset linking process.

Subset of spreadsheet shown.

7.3. Spatial Visualisation the Results

By following the route described in section 7.2 and illustrated in Figure 7.1, it was possible to generate the number of individuals per output area who were classified into each of the five clusters. A lengthy sample list of these results are shown in Appendix B.2 (full list available on request - see Appendix C for details) and the first ten are presented in Table 7.1 for illustrative purposes. The 2001 CAS ward name is also included to assist with interpretation. It should be noted that in the twelve cases where two clusters shared the highest count of individuals, individuals were automatically assigned to the first cluster numerically. This was deemed a suitable process given such few occurrences and as the principle focus is the highlight the functionality of the framework. Such cases are noted in the appendix (B.1).

OA_Code	Ward_Name	Total Pop.	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster Member-ship
00DAFA0001	Aireborough	304	71	182	8	6	37	2
00DAFA0002	Aireborough	286	84	93	47	16	46	2
00DAFA0003	Aireborough	436	255	5	75	62	39	1
00DAFA0004	Aireborough	437	25	62	56	237	57	4
00DAFA0005	Aireborough	294	103	37	46	61	47	1
00DAFA0006	Aireborough	363	141	50	51	48	73	1
00DAFA0007	Aireborough	370	32	21	211	95	11	3
00DAFA0008	Aireborough	318	19	122	70	45	62	2
00DAFA0009	Aireborough	384	110	96	100	37	41	1
00DAFA0010	Aireborough	287	65	58	49	64	51	1

Table 7.1. First ten Leeds output areas (sorted A-Z) and associated cluster codes.

As can be seen in Table 7.1 (and Appendix B.2), the linking process does differentiate between individuals pairing them with different clusters. As this matching process makes use of circa half (eight of the fifteen) SAM classification variables given their presence in the microsimulated datasets, there is clearly scope for improvement, however, upon studying the spatial distribution of clusters by Leeds output area (presented in Figure 7.2), one can see patterns that may be expected given a knowledge of cluster composition. For example, Cluster 1, categorised by individuals in higher managerial occupations and in 2+ car households tends to be distributed in the more affluent areas of the city, in particular to the north and with some presence in the east. To the contrary, Clusters 3 and 4, which may be regarded as the less affluent cluster-types given the semi-routine occupations (probably

leading to longer working weeks as also identified in the classification), fair health and, in the case of cluster 3, persons sharing houses who are unrelated, show different patterns. These clusters are focused more around inner Leeds (in the case of cluster 4) and to a lesser extent cluster 3, the latter also being more sporadic it its spatial patterns.

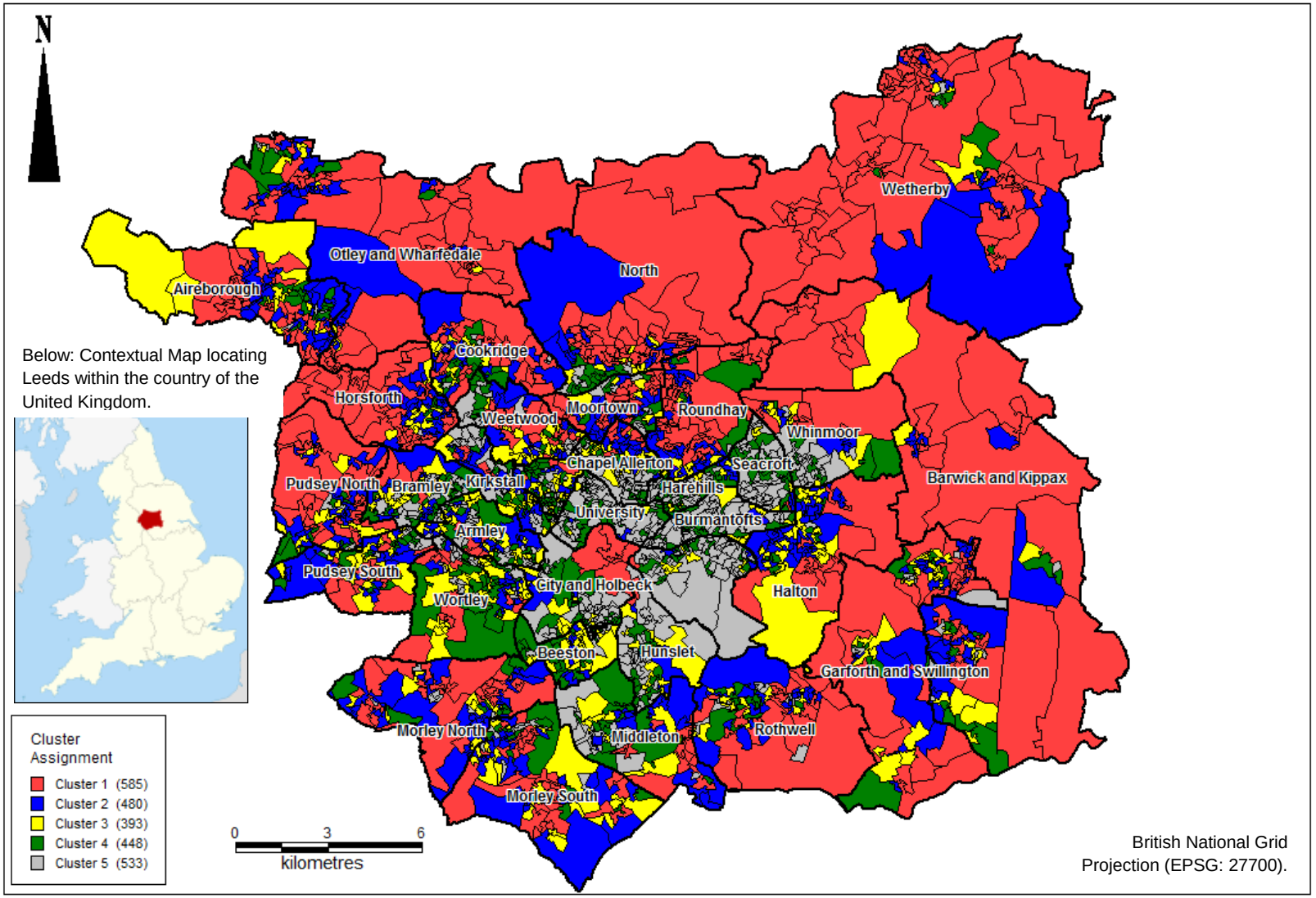


Figure 7.2. Leeds-wide illustrative map of five cluster types by output area. CAS Wards overlaid to add context.

See Table 6.5 for full details of cluster types.

***Cluster 1:** Affluent Managers*

***Cluster 2:** Young People living with Family*

***Cluster 3:** Co-habiting Couples*

***Cluster 4:** Average Resident*

***Cluster 5:** Nearing Retirement*

To take a more statistical stance and to contrast the results with an alternative system, of the two clusters from the 2001 Output Area classification (OAC) that appear to match those clusters in this research (by definition), 'Typical Traits' from the OAC and 'Average Resident' from this work and 'Prospering Suburbs' from the OAC and 'Affluent Managers' from this research, both do correlate in terms of classification of areas. The former categorises 572 output areas (as opposed to 447 in this work) and the latter 530 output areas (versus 591 in this research). When a spatial assessment is undertaken looking at cases where these classifications classify the same area in the said clusters, the results are mixed. Of the 572 output areas categorised into the 'Typical Traits' cluster by the OAC, 41 fall into the 'Average Resident' grouping in this work (7.6%). Whereas, of the 592 small areas deemed to reside in the 'Prospering Suburbs' cluster in this research, 325 from the OAC fall into the 'Prospering Suburbs' cluster (54.9%). Such results support differentiation towards the top of the hierarchy but less so for the more disadvantaged segments of society. As this is by no means a like-for-like comparison, true evaluation of the classification is presented in later sections.

Regardless of how well or otherwise one interprets the spatial patterns presented in Figure 7.2 (based on any local knowledge of the social geography of Leeds or other statistical analysis similar to that given above), the fact of the matter is an individual-level classification has been generated based on SAM data and visualised through a linkage to a microsimulated data with fine-level geographical units. This represents new ground within the domain of geodemographics. Although arguably the process of visualisation and hence aggregating individual-level cluster assignments to pre-determined geographical units and taking the modal occurrence to then map at the output area level may be regarded as returning to the conventional format of area-based geodemographics, one must be aware of the benefits. At a time when Experían is pushing its household-level classification (see section 2.7.3), systems capable of classifying to this level of detail are no doubt on the rise. Visualisation is particularly difficult with individual-level data, however, the linkage to the microsimulated dataset is not entirely for visualisation purposes alone. This link has generated a complete population classification for the city of Leeds (715,402 individuals) through the sample classification conducted on the 5% SAM file (~35,000 individuals). It is this tabular data that is arguably of greater benefit given a knowledge of how many people per output area fit into

each grouping - as opposed to a simplified visualisation process highlighting the predominant cluster per output area, the results of which (by output area) can be found in Appendix B.2. By this, an element of fuzziness is introduced given an awareness of how close or otherwise an output area is to residing in a different cluster. Table 7.2 illustrates this for the same group of records.

Ref ID	OA_Code	Ward_Name	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)	Cluster 4 (%)	Cluster 5 (%)	Cluster Member-ship
1	00DAFA0001	Aireborough	23.36	59.87	2.63	1.97	12.17	2
2	00DAFA0002	Aireborough	29.37	32.52	16.43	5.59	16.08	2
3	00DAFA0003	Aireborough	58.49	1.15	17.20	14.22	8.94	1
4	00DAFA0004	Aireborough	5.72	14.19	12.81	54.23	13.04	4
5	00DAFA0005	Aireborough	35.03	12.59	15.65	20.75	15.99	1
6	00DAFA0006	Aireborough	38.84	13.77	14.05	13.22	20.11	1
7	00DAFA0007	Aireborough	8.65	5.68	57.03	25.68	2.97	3
8	00DAFA0008	Aireborough	5.97	38.36	22.01	14.15	19.50	2
9	00DAFA0009	Aireborough	28.65	25.00	26.04	9.64	10.68	1
10	00DAFA0010	Aireborough	22.65	20.21	17.07	22.30	17.77	1

Table 7.2. Percentage cluster composition of ten selected Leeds output areas, 2001.

As evidenced from assessing Tables 7.1 and 7.2, certain output areas are very close to being re-categorised into different best-fit (or crisp) clusters, for example output areas 00DAFA0009 and 00DAFA0010 are both within fourteen and seven re-classified individuals of having the predominantly-allocated cluster changed. This relative fuzziness is evident from the data when available in tabular format but is very hard to visualise cartographically. One could assign symbology to each zone indicating the cluster spread, as visualised in Figure 7.3, using the data from Table 7.2 and labelled by the reference ID for the ten named output areas. Adopting such visualisation techniques for a wider area would lead to problems such as symbology overlap and, consequently, poor means of interpretation. This very much supports the statement made earlier that the tabular output when working at the level of the individual, even when aggregated to the zone level, is the most suitable means of outputting such research.

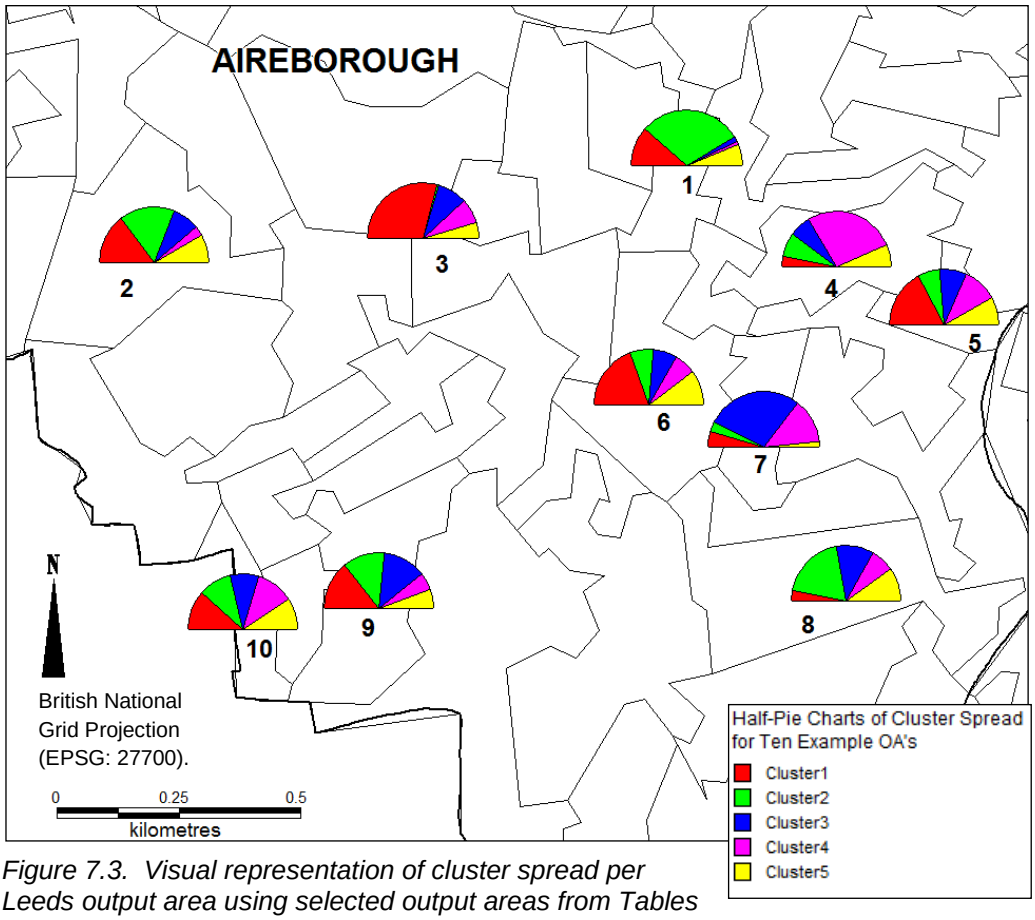


Figure 7.3. Visual representation of cluster spread per Leeds output area using selected output areas from Tables 7.1 and 7.2 (Aireborough, North-West Leeds). Numerical labelling refers to output area ID as shown in Table 7.2.

One distinct advantage of being able to visualise the classification at the output area level is that it enables the results of this individual-level classification to be critiqued, to some extent, against previously created systems at the areal unit level. This has already been demonstrated to some extent with statistical comparisons with the OAC (see section 4.8). Although such critiques are inherently visual by nature, Figure 7.4 shows the OAC, as developed by Vickers (2006), also using 2001 data - at the aggregate level. Even though the system devised by Vickers (2006) segments output areas into one of seven pre-determined groupings (as opposed to five in this research), some common visual patterns are apparent. Firstly, output areas categorised by Vickers as falling into the 'Countryside' grouping do mirror the patterns observed in this research with regards to Cluster 1: Affluent Managers. There is also some correlation between Vickers' 'Typical Traits' grouping and Cluster 4: Average Resident in this work (as evidenced previously). Such assessments are, however, incredibly visual and alternative

means are necessary to determine the true validity of the framework/classification. Section 7.4 presents such an approach.

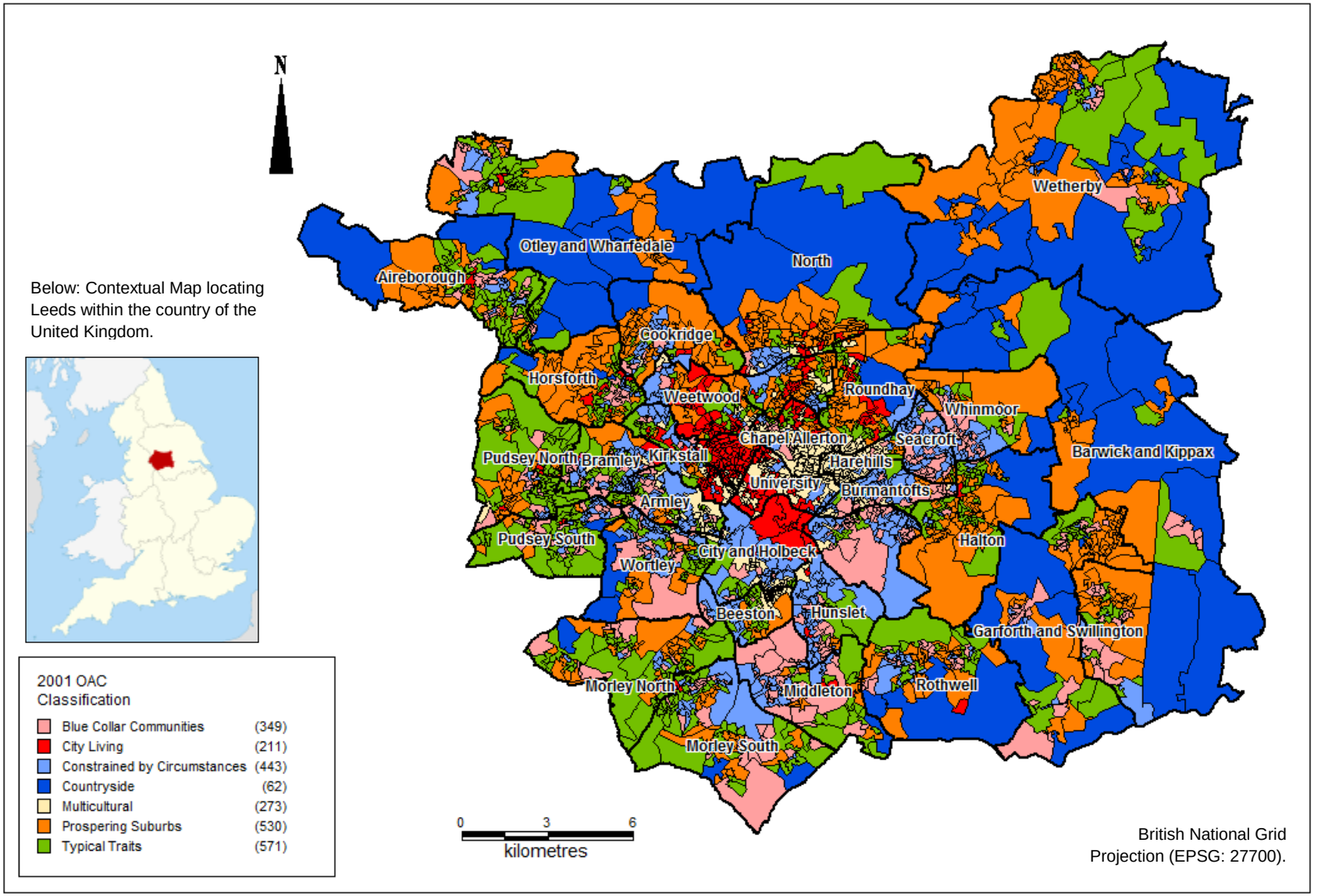


Figure 7.4. Leeds-wide illustrative map 2001 OAC by output area. CAS Wards overlaid to add context.

Page 118.

7.4. Adding Value and Validation: Linking to Non-Census Datasets

Through adopting a process akin to the statistical matching method as discussed in section 7.2 (and illustrated in Figure 7.1), that of Sum of Squares, it was possible to link the SAM classification to external datasets based on common variables. This is particularly useful for validation but it also allows the classification to be 'bolted-on' to more behavioural datasets to profile alongside non-census characteristics - providing of course there are some common variables between the two datasets to enable the cluster codes to be matched. Therefore, for reasons pertaining to both validation and enrichment of the classification, a link to the BHPS (wave 18) was established. This link enabled each of the five clusters to be assigned directly to one individual in the BHPS – and hence, each record in the BHPS was assigned a cluster code 1-5.

The variables present in the BHPS are designed to describe socio-economic conditions at both individual and household level (ISER, 2011). Variable categories include; household organisation, employment, accommodation, tenancy, income and wealth, housing, health, socio-economic values, residential mobility, marital and relationship history, social support, and individual and household demographics (ISER, 2011) and hence add value over and above the variables present in the original SAM file.

Given that this research is set to inform the GENESIS project, one or two variables per GENESIS theme were selected for validation purposes (recall the themes in Table 5.2). A deliberate attempt was made to ensure that the selected variables were of a lifestyle / behavioural type and thus in no way similar to the census variables which comprise the classification. The selected BHPS variables are shown in Table 7.3.

Theme	GENESIS Project Theme	Generic Social Science Theme	BHPS Variable
Crime		**X**	• Not available
Education		**X**	• Has a degree? • Has a higher degree? • Political allegiance
Employment		**X**	• Job satisfaction
Health(care)	**X**	**X**	• Any recent health problems? • 'Meals on wheels' uptake • Play/participate in sport?
Housing	**X**	**X**	• Value of property • Type of property • Internet in property
Retail (and consumption)	**X**	**X**	• Regularity of eating out in restaurants
Transportation	**X**	**X**	• Number of flights taken in last twelve months

Table 7.3. Selecting lifestyle / behavioural BHPS variables in line with GENESIS' key themes.

7.5. Linking SAM Classification to BHPS

Achieving this linkage process again required some rather novel methods and these are summarised in the sections which follow. Further details are available in earlier chapters given the repeated process.

At this point in the overall framework, a completed SAM classification (for Leeds) has been devised with each of the 35,986 individuals assigned to a single cluster (1-5) on a one-to-one best-fit basis - often referred to as crisp geodemographics. Fuzziness values were also available following the Euclidean distance to cluster centres calculation. A city-wide classification has also been formed following the microsimulation linkage process with fuzziness values again available, although, for the purpose of analysis and visualisation, crisp membership to single output areas was again adopted.

In order to facilitate the linkage of the SAM classification to the BHPS file, each of the fifteen SAM variables were searched for in the BHPS, fourteen were located (with the exception of Family Type), and extracted into a new file. Table 5.3 shows the relationship between variables in the SAM and BHPS. The person ID variable was also extracted so as to enable a join process at a later stage should this be required. These variables were then re-coded under

the SAM structure (e.g. sub-variable definitions – similar to that discussed in section 5.11) and assigned the same gross monthly income values.

With a newly generated BHPS file containing variables quantified based on income by SAM definition, it was then possible to deduce an individual's proximity to each of the five SAM clusters. This was again achieved by calculating the Euclidean distance between each of the fifteen variables (per individual) and the final cluster centres – hence, a process of statistical matching was adopted as described previously. This then produced a value indicating the relative distance between the BHPS individual and each of the five SAM clusters. As before, the smaller the value, the closer that individual to the said cluster. A automated process of identifying the smallest value from a list was then employed in Excel and each of the 14,419 BHPS individuals were assigned a best-fit single cluster. Naturally, and to re-emphasise, the scope to incorporate a fuzzy classification is once more apparent at this stage given the proximity of individuals to cluster centres, however, for the purpose of this evaluative work and to validate the framework's functionality, a crisp classification was deemed sufficient.

7.6. Reviewing the SAM-BHPS Link Results

The presence of more behavioural variables, such as the frequency of dining out and an individual's political alignment, makes the critique all the more interesting. Furthermore, the vast array of variables in the BHPS across each of the GENESIS themes (with the exception of crime) and the widespread coverage such variables provide with respect to social science make the critique incredibly encompassing of extra datasets, all of which can be used to supplement the classification output and add a lifestyle / behavioural angle. A selection of the results from this data linkage process can be seen in Table 7.4 with four BHPS variables contrasted with the five clusters.

Recall the five Leeds cluster definitions and their make up (see Tables 6.3 and 6.5):

Cluster 1: *Affluent Managers*

Cluster 2: *Young People living with Family*

Cluster 3: *Co-habiting Couples*

Cluster 4: *Average Resident*

Cluster 5: *Nearing Retirement*

Cluster	Dine Out?	Play Sport?	Political Allegiance (if voting tomorrow)?	Watch Live Sport (at sporting venue)?
1	Once Per Month – 60% Several Times per Year – 22%	Once per week – 3%	No vote – 50% Conservative – 25% Other Party – 8.3% Labour – 8% LibDem – 8%	Several Times per Year – 20%
2	Once Per Month – 35% Several Times per Year – 40%	Once per week – 7%	No vote – 40% Can't Vote – 18.6% Labour – 12% Conservative – 9.3% LibDem – 5.8%	Several Times per Year – 8%
3	Once Per Month – 47% Several Times per Year – 30%	Once per week – 6%	No vote – 50.4% Conservative – 12.2% Labour – 8.5% LibDem – 6%	Several Times per Year – 14%
4	Once Per Month – 44% Several Times per Year – 31%	Once per week – 6%	No vote – 48.2% Conservative – 11.7% Labour – 10% Can't Vote – 8% LibDem – 6.2%	Several Times per Year – 14%
5	Once Per Month – 36% Several Times per Year – 31%	Once per week – 8%	No vote – 47.5% Conservative – 12.5% Labour – 12.5% LibDem – 7%	Several Times per Year – 9%

Table 7.4. Contrasting SAM Classification with BHPS Individuals and extracting new Information

As can be seen in Table 7.4, if taking the full cluster descriptors into consideration (section 6.3.1), the results appear to corroborate the cluster characteristics to some degree, however, there are a series of anomalies which may be explained by the methods adopted.

As reflected in Table 7.4 and the accompanying pen portraits, Cluster 2 is categorised by predominantly young individuals (circa aged 12) living with family. One should therefore not be surprised that this cluster is one of the least likely to dine out on a monthly basis and is one of the more active clusters when it comes to sport and physical activity but one of the least active when it comes to attending costly sporting events. Furthermore, the categorisation of individuals in this cluster being of non-voting age is supported by BHPS statistic which denotes that 18.6% of individuals in this cluster are ineligible to vote in elections. The results referred to here suggest

some degree of success with regards to this matching process as far as validation goes.

An second example can be seen from assessing Cluster 1 (Affluent Managers). As many of the variables presented in Table 7.4 can be linked to availability of disposable income, it is unsurprising that members of this cluster have a high tendency to dine out once per month (greater than other clusters) and attend sporting venues. The high proportion of members willing to vote (the only cluster where 'No Vote' is not highly ranked) in addition to an alignment towards the Conservative Party are also statistics one can contend with. The low percentage partaking in sport is rather surprising but not one which should mask what appear to be statistics that corroborate a health data match.

A key observation to be highlighted is the use of Leeds' final cluster centres when classifying the complete BHPS (wave 18). Naturally, different parts of the UK look rather different in terms of their demographic profiles and the use of Leeds' cluster centres may have impacted on the results of this BHPS linkage process – particularly given that the BHPS is UK-wide. Furthermore, adopting the complete BHPS file as opposed to a more regionalised subset may also have had some bearing on the results presented in Table 7.4.

It should be noted that the results presented in Table 7.4 represent a selection of the variables assessed. As overviewed in Table 7.3, alternatives were also explored, however, those put forward in Table 7.4 display the greatest level of perceived corroboration to the classification. Variables where the match brought about results one may regard as unexpected, in particular with reference to number of flights taken and recent health problems, suggest that although in its infancy, the framework and/or linking process may require further testing before being deemed robust.

The linkage demonstrated in this research was undertaken on the BHPS for two reasons. Firstly, due to the array of behavioural variables matching those of GENESIS' themes and secondly due to the high number of common variables between the SAM classification and BHPS thus enabling an effective match. Other datasets, such as the British Crime Survey (eight common variables), Health Survey for England (nine) and National Travel Survey (six) would make for interesting linkages albeit based on fewer variable matches.

7.7. Summary and Conclusions

This chapter has built upon the SAM classification presented in chapter 6 by evidencing the linkage process between the base classification and both a microsimulated dataset (for complete population modelling and visualisation) and external survey dataset (BHPS) for validation and classification enrichment.

The linkage to microsimulation is of particular importance given that the SAM classification is constructed based on a 5% sample of Leeds' individuals (35,986 in 2001). The linkage enables each of Leeds' 715,402 individuals to be attributed with a cluster code (1-5 in this case) denoting their positioning across the give groupings. The link is facilitated through a statistical matching process using the Sum of Squares Technique whereby the Euclidean distances between variables and cluster centres are calculated. Although each individual is assigned one single cluster code, the scope to incorporate fuzziness is clear given the availability of the Euclidean distances. Classification at this level goes some distance to reducing the problems of MAUP and ecological fallacy often regarded as synonymous with geodemographics. Furthermore, the availability of individual fuzziness scores and hence a knowledge of how near/far an individual is from re-categorisation arguably reduces such problems further.

The ability to link to the microsimulated dataset also offers a means to visualise. By making use of the small-area geography in the simulated data, cartographical outputs could be generated denoting crisp clustering (Figure 7.2) and fuzziness (Figure 7.3) when aggregating by output area. Although such aggregation arguably returns geodemographics to the point of area-based functionality, such methods are useful not only for visualisation but also to contrast with alternative data/systems in operation at the area level.

The final sections of this chapter discussed further linkage abilities, this time to external and non-census datasets for validation and in-depth profiling. The BHPS was selected as the dataset to demonstrate this on given the number of common variables (fourteen) between this and the SAM base classification. The linkage enabled non-census variables such as the propensity to dine (eat) out and sport participation to be profiled alongside the five clusters. The results enabled a corroboration of the clusters and framework in addition to

generating value-added information over and above variables used within the classification process.

Chapter 8 will summarise the research presented and emphasise the usefulness of the proposed framework for future classifications of this nature whilst also providing discussion on some of its limitations. In particular, the processes required to classify at the level of the individual will be illustrated and explained such that the framework can be replicated for further research and non-academic usage.

Chapter 8: Summary, Conclusions and Way Forward

8.1. Introduction and Chapter Preface

This final chapter of the thesis is divided into four main sections. Firstly, the objectives as initially presented in chapter 1 are re-visited and a summary of the research findings relating to each is presented (8.2). Secondly, a final word is given on the framework put forward in chapter 5 with particular reference given to its utilisation in a wider context (8.3) and its strengths, limitations and suggestions for improvement (8.4). Finally, further research opportunities are briefly discussed (8.5).

8.2. Research Outcomes

This section will provide a summary of the research outcomes relating to each of the five research objectives as initially set out in chapter 1.

Objective #1: Conduct a review of the literature pertaining to (1) geodemographic classifications and (2) population generation techniques.

The literature pertaining to geodemographics very much positions the discipline/technique as an area-based procedure designed to partition aggregate data into what it perceives to be homogeneous groupings. The operation at the aggregate/areal level is not without its problems though, as discussed at various junctures in this thesis. The Modifiable Areal Unit Problem (MAUP) (predominantly the scale effect), particularly within commercial systems with a tendency to operate at different spatial scales (e.g. postcode and output area), is highly apparent. Changing geographical units or boundaries impacts noticeably on the categorisation of areas given how the data are partitioned. For example, analysis undertaken at Lower Layer Super Output Area (LSOA) level may result in different patterns to the same analysis conducted at Middle Layer Super Output Area (MSOA) level given the different spatial resolutions of the individual zones. This leads onto the second problem, and one arguably more rife within area-based geodemographics, that of ecological fallacy. When areas are best-fitted to crisp clusters, the general assumption, particularly from those not aware of the

methodologies that underpin such systems, is that clusters with names such as 'Young Married Suburbia' or 'Metro Singles' encompass exclusively people categorised by those tag lines. Section 8.3 explores these problems in more depth when discussing the need for a system capable of classifying to the person level and, more importantly, how the framework proposed in this research lessens such issues.

The review of literature established that although geodemographics has remained firmly rooted in area-based classification since its inception, more recently the emergence of Experían's household Mosaic system in addition to Acxiom's PersonicX system suggests that a pragmatic shift is commencing such that future geodemography can classify beyond the areal unit level. When problems such as those discussed above are considered, the benefits are apparent (and will be illustrated in section 8.4).

With regards to synthetic populations, the key observation to be drawn from the literature is the rise of computational power since the early days of geodemography. With such powerful machines and sophisticated algorithms, an ability to formulate accurate synthetic populations from aggregate constraint data means that classifications beyond those of areal units can be developed. Methods such as simulated annealing, deterministic re-weighting and conditional probability have all been employed with success in other research and, with the benefit to be gained from person level classification, such methods represent a sensible and largely accurate approach.

Objective #2: Present an assessment of common methodologies adopted when formulating geodemographic classification schemes.

As discussed earlier in this section under the research outcomes of objective #1, standard geodemographic clustering methodologies have not changed greatly since the early days of use and nor have the units used to release such classifications (hence, remain area-based). Clustering methodologies such as hierarchical or stepwise K-Means in software packages such as IBM's SPSS lead the way for classifications to be undertaken with certain leading commercial vendors adopting more sophisticated and specialist approaches.

The methodology (flow-diagram) put forward by Gibson and See (2006) in chapter 4 is commonplace in area-based system formulation and begins with identifying a purpose, whether this be general or application-specific, and

includes data selection, pre-processing, clustering approaches, labelling/interpretation, application/evaluation and, in most cases, integration with a GIS such as ESRI's ArcGIS or Pitney Bowes' MapInfo product.

Such methods are somewhat suitable for geodemographic clustering but, given the reliance on aggregate data (for which they are designed), the problems as previously referred to in this thesis transpire (ecological fallacy, MAUP, generalisation etc). Furthermore, given that this research has a clear remit of devising a framework through which individual-level classifications can be achieved, such methods in their original form are not entirely suitable. Knowledge of their composition is, however, important and links to objective #3.

Objective #3: Formulate a framework through which general-purpose individual-level geodemographic classification schemes can be generated.

This objective represents the root of the research, not least to overcome the aforementioned problems, but also to challenge the current geodemographic ethos and progress towards more efficient means of population targeting.

The framework put forward makes use of microdata from the 2001 SAM file and outlines the phases to develop a general-purpose system at the level of the person. The framework adopts data-handling methods such that both categorical (ordinal and nominal) data are equally incorporated into the system and no bias ensues when handling differing data types. This is achieved through the re-coding of, in particular, nominal variables with gross monthly income data extracted from an external dataset (BHPS). This re-coding emphasises the magnitudinal variances between variable sub-categories (e.g. marital status: single vs. married vs. separated/widowed/divorced) and hence ensures effective data partitioning.

The framework adopted represents the first of its kind available in the public domain and certainly within academia. The framework has been devised with updatability and transferability in mind. As such, the proposed route to system formulation is based on census products, in this research from 2001. The updating of the system to classify 2011 data is also achievable once data availability is confirmed given likely comparable formats. The system may require revision beyond 2011 to cater for other census-replacement datasets

(see section 8.4). Furthermore, the ability to append the classification to external datasets adds value, not only for validation (as illustrated), but more importantly for future users to profile against external datasets. In this work, the BHPS was demonstrated given the plethora of non-census variables present in the dataset and parallelism of such variables to the broad field of social science and GENESIS.

Objective *#4: Apply the classification framework to two case study locations and investigate the performance relative to these.*

This objective highlighted the functionality of the framework in terms of its ability to successfully partition and differentiate between the individual data, whether this be continuous or categorical. The primary case study region of Leeds and also that of Richmondshire were presented, both selected due to their differing locations, population sizes and general demographics.

The results identified five largely distinct clusters in both cases, ranging from clusters of individuals with high levels of education and employment attainment to those with very low education/qualifications and unemployed statuses. Such clusters suggest that, on the most part, the framework and methods are able to handle and partition data of this nature into homogeneous clusters.

Objective #5: Facilitate a link from the classification to other social scientific datasets for the purpose of validation and enrichment.

Although the framework proposal represents the key innovation behind this research, an ability to link the classification to other datasets is of great benefit. This was demonstrated in this research for validation purposes first and foremost and latterly for enrichment and added value.

Through a statistical matching routine, cluster codes can be appended to datasets sharing at least some common variables (the greater the number of common variables, the more accurate the match). This process was illustrated on the BHPS where all but one of the SAM variables were present hence enabling a strong statistical match. The linkage corroborated the cluster centres and pen portrait definitions for the most part although anomalies were evident. However, given the novel approach taken in this research, limitations and modifications moving forward were always likely.

8.3. Adopting the Individual-Level Geodemographic Framework in a Wider Context

The framework presented and discussed in this research is designed to be transferable such that it can be applied by other researchers and practitioners in a variety of contexts. Although applied more generally in this research through the construction of a general-purpose classification designed to highlight the framework's functionality, other applied opportunities include health, crime and retail profiling (see section 2.5). Local authorities therefore may be one such beneficiary of this research.

The flow diagram presented in Figure 8.1 acts as an illustrative guide through which interested parties can employ the framework with alternative data (outside academia). The diagram comprises eight phases, some comparable to those listed by Gibson and See (2006) when formulating area-based classifications [see section 4.2], and others that involve new steps designed to enable the handling of individual data. The following sections discuss each phase in turn and provide supplementary information to that presented in Figure 8.1.

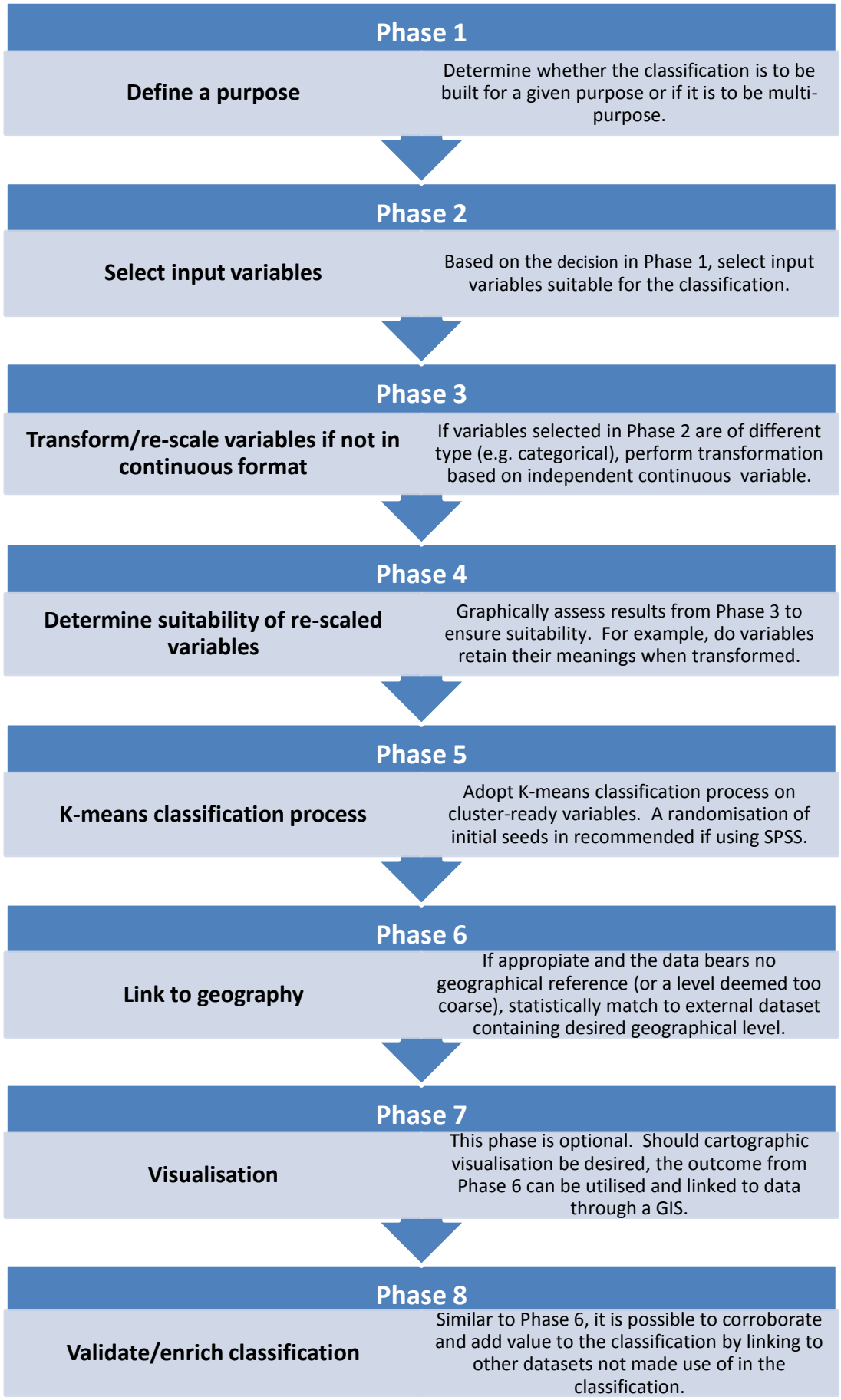


Figure 8.1. Demonstrating the transferability of the individual-level framework.

8.3.1 Phase 1: Define a Purpose

As with area-based classifications, knowledge of a given purpose to which the classification is expected to support is of particular benefit, in particular with reference to input variable selection. However, general-purpose classifications are commonplace in the market and have been discussed in this thesis (ACORN, Mosaic and CAMEO being three such examples). This research made use of the framework to construct a multi-purpose classification. This development phase is common to the construction of area-based systems.

8.3.2 Phase 2: Select Input Variables

As discussed in section 4.4, input variable selections are often aided by an overriding purpose which the classification is designed to support, for example health profiling (see Abbas, 2009). The presence of a specific purpose makes selection easier as it narrows down the scope of the classification. This work has presented a multi-purpose classification and therefore variables were selected based on other means. One such approach is to consider variables included in alternative multi-purpose geodemographic systems or those found in datasets spanning key domains within areas of interest. This work took a high-level approach and considered variables from the breadth of social science and selected those variables common across datasets such as the British Crime Survey and National Travel Survey. Naturally, this resulted in fundamental census variables being put forward to demonstrate the framework's functionality. This phase is common to the construction of area-based systems and variable selection methods such as those discussed can be adopted.

8.3.3 Phase 3: Transform/Re-Scale Variables

This phase represents one of the novel aspects of this research and distinguishes it from area-based classification. Individual-level data are often stored in a categorical format, for example ordinal, nominal or dichotomous. Such data types, particularly if conflicting or mixed, are not suited to conventional clustering methodologies such as K-Means. It is often therefore necessary to transform variables into a (quasi)continuous state suitable for a clustering process. The framework achieves this by making use of an independent continuous variable (in this case, monthly earned income from the BHPS) and converting all categorical data into this format. The conversion

process is facilitated by sourcing an external dataset with an independent continuous variable and also each of the variables selected for input to the classification algorithm. The average value from the independent variable is then attributed to the categorical data, for example: male and female would receive two separate valuations, in this research linked to average income. If a single external dataset containing all input variables is not sourceable, alternatives can be used to ensure each variable is re-coded onto a quasi-continuous scale. If more than one external dataset is used, a process of standardisation will be required to linearly re-scale all data and ensure parity.

In some cases, it may be acceptable to retain the format of ordinal variables (such as highest level of qualification where values range from 0 for no qualifications to 5 for highest possible level of qualification) or interval variables (such as age, e.g. 0-4, 5-10 etc). In the case of interval data, one of the lower, middle or highest values per interval can be used providing this remains consistent throughout. Should certain variables remain in their original form, normalisation/re-scaling may be required as discussed previously.

Variable polarity should also be considered at this stage.

8.3.4 Phase 4: Determine Suitability of Re-Scaled Variables

Regardless of whether the previous phase is applied to all or a selection of the input variables, it is necessary to ensure the meaning of any variable is not lost by the conversion from one data type to another. This research has demonstrated this through the age variable but other cases may exist when adopting this framework with new data. Age is stored in interval format such that individuals may be 0-4, 5-10 years of age etc. Retaining the interval format (or lowest value per interval, for example 0, 5...) for each individual preserves the meaning of the variable. If this is instead converted into a continuous format based on an average earned income value as per other variables, a great deal of meaning is lost. For example, those individuals below the working age (<18) and those in retirement (>~70) inherit highly comparable earned incomes making them inseparable when it comes to classification. This is the case in this example as the continuous variable used represents earned income only and fails to take into account benefits or any other forms of financial support.

8.3.5 Phase 5: K-Means Classification Process

This phase is comparable to any classification created using aggregate data given that the variables have been converted to continuous formats where appropriate (phase 3).

A recommendation at this stage is to randomise the initial seeds of the cluster centres prior to commencing the clustering algorithm (if the software chosen does not do this by default). This research was conducted using IBM's SPSS package and therefore seed randomisation was required manually.

8.3.6 Phase 6: Link to Geography

In many instances, individual-level data are not available with a fine-level of geography given sensitivity and confidentiality. This is certainly the case with census data but may be different with company-specific data (e.g. data collected by retailers on customers). If a finer-level of detail is desired, it is possible to link the classification codes generated in phase 5 to a dataset offering this level of detail. This process is termed 'statistical matching' in this research and is detailed in depth in section 7.2.

If the framework is being pursued with company or self-collected data, the resolution to which the data are captured may be sufficient.

The purpose of this phase is to enable a greater understanding of the spatial patterns of the clusters. It is also required to fulfil phase 8.

This research also made use of the statistical matching process to obtain a fuller dataset, thus moving from the 5% SAM sample to that of a complete population dataset for the case study regions. As with completing this phase to obtain finer-level geographies, if the data collected by an organisation are of a suitable sample then statistical matching is not necessarily a requirement.

8.3.7 Phase 7: Visualisation

This is an optional phase and feeds on from phase 6 where finer-level geography was introduced (if required). Linking the attribute data to geographical boundaries through a GIS enables spatial patterns/relationships to be easily identified and decisions made on population targeting with respect to retail/sales marketing, at-risk population identification, resource prioritisation etc.

8.3.8 Phase 8: Validation and Enrichment

Although phase 7 denotes the completion of building and visualising the classification and making use of its output, this phase represents an opportunity to both validate and enrich the results with supplementary information. Through a process of statistical matching (identical to phase 6), as discussed in section 7.2, the cluster codes from the classification can be appended onto other datasets enabling deeper profiling. In the case of this research, it was possible to match the cluster codes onto the BHPS dataset and profile the results against other variables (principally behavioural) such as one's propensity to dine out of an evening or take flights abroad during the course of a twelve month period. Such outcomes not only add value and enrich the classification but also offer an opportunity to corroborate the clustering process. For example, to take a crude case, one may opt to validate the affluent-most cluster of a classification against the BHPS with particular reference to the aforementioned flights abroad variable. One would expect the affluent-most clusters to demonstrate a greater propensity to take air travel than other clusters in the classification.

This phase completes the framework and is very much supplementary to the remainder of the development phases but is a useful tool for validation, enrichment and/or profiling against other datasets if required.

8.4. Strengths, Weaknesses and Considerations of the Individual-Level Geodemographic Classification Framework

As discussed at previous junctures in this research, an approach to individual-level classification has not been attempted before in academic literature, therefore the framework put forward in this work is designed to offer a first route to classification. However, inevitably the framework is not the finished product nor is it without its problems. It does, however, function and produce homogeneous clustering outcomes. Pros and cons of the framework are discussed below.

8.4.1. Strengths

The key strengths to be drawn from this research can be linked to the two problems identified at the outset of this chapter; those of MAUP and ecological fallacy.

As brought to attention earlier, area-based geodemographic systems rarely classify small areas into best-fit groups without some degree of ambiguity and the example in section 4.8 under the heading of 'misrepresentation in classifications' is a prime example. Recall the 'Metro Singles' example discussed by Birkin (1995) when assessing the SuperProfiles Lifestyle classification; the true composition of this cluster when compared to what one may expect is very different. In fact, rather than exclusively (or even principally) encompassing single workers, the cluster in fact only accounts for 21% of such a demographic meaning that circa 80% of the cluster is categorised by people not recorded as single and working. This is misrepresentation. The 'Young Married Suburbia' example also returns the same findings with over 25% of the cluster recorded as aged over forty-five years.

Birkin's (1995) assessment of ecological fallacy within geodemographics was built upon by new research in this thesis which critiqued a more recent and fully open-source system, the Output Area Classification (OAC) (2001). Despite the OAC being made available circa twenty years after SuperProfiles was first released in the early/mid 1980's (Harris *et al.,* 2005), the same problems were highlighted. This research identified, in particular, the 'Multicultural' cluster (amongst others) in Leeds and how its two sub-groups, 'Asian Communities' and 'Afro-Caribbean Communities' did not truly describe the population demographics of areas categorised in these clusters. The former grouping in fact contained over 13% more Afro-Caribbean persons than Asians and in the latter 40% more Asians than Afro-Caribbeans, observations that one may have expected in reverse given the cluster definitions.

The above discussion suggests that a movement away from areal-unit categorisation would make for far more effective and focused methods for population targeting. However, one must also bear in mind that similar to geographical modelling and the notion of simplifying reality (words discussed in chapter 2 as part of the dictionary definition of classification), geodemographics cannot be expected to describe the real world down to infinite detail. The ability to segment areas very much aids understanding even if errors transpire. Nevertheless, reducing the frequency and magnitude of such errors is something this research looked to address.

Farr and Webber (2001) describe the benefits to be gained from moving from aggregate systems to systems capable of individual-level classification as being "*intuitively obvious*" (p.58), particularly with reference to the added discrimination such systems provide. Arguably, it is a combination of added discrimination and improved levels of ecological fallacy that this research has overcome through operating at the level of the individual. If a system is deemed to discriminate better than alternatives then it will, as a consequence, reduce the level of ecological fallacy as the clusters are likely to be more homogeneous. As discussed in section 5.2, as the quantity of variables increases in an aggregate-data classification, the scope for misrepresentation also increases as fewer people are likely to fit the described cluster demographic. At the level of the person, although this is also the case, it is easier to maintain a greater level of homogeneity as one individual can easily be re-classified should he/she not fit a given cluster definition. At the area level, for a small-area to be re-classified, a bigger shift is required and even then, the degree of homogeneity is likely to be less than that of a system operating at the level of the person.

This research has demonstrated the above by aggregating from the individual-level to output area. As can be seen from the data subset in Table 7.2 and the list in Appendix B.2, although area-based systems profess to cluster areas based on homogeneity, as evidenced in the two said tables, a variety of individuals reside in these clusters and by simply allocating an area to a crisp cluster, this variety (or heterogeneity) is lost. Take output area '00DAFA0015' in Aireborough CAS ward (2001), in north-west Leeds as an example. Any area-based classification would attribute a cluster code to this area (for example, in the OAC this is 'Prospering Suburbs' (super group) and 'Thriving Suburbs' (group)). Whereas, in reality, the variation that exists in this area is captured to a greater extent by classifying at the level of the individual as, although the area may have some degree of homogeneity, finer-level classifications pick up any heterogeneity that exists. In this research, this particular output area clearly has a predominant cluster ('Affluent Managers') but it also has variations within, as evidenced in Table 8.1, something an area-based system very much overlooks.

OA Zone_Code	Ward_Name	TotalPop	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster M'Ship
00DAFA0015	Aireborough	385	132	109	125	4	15	1

Table 8.1. Example output area and the benefits of individual-level classification.

A further strength evidenced in this research is the ability to aggregate the classification to chosen geographies with relative ease. This has been shown through a linkage to output area geography. At times when changing geographical boundaries making spatial-temporal analysis difficult, a classification at the individual level enables aggregations to chosen geographies for the purpose of benchmarking and analysis.

8.4.2. Weaknesses and Considerations

As referred to in chapters 6 and 7, certain variables do not appear to differentiate between individuals particularly well. Such variables include ethnic group and sex amongst others. Even though these variables may be termed fundamental census characteristics (and poses high inclusion values when looking for occurrences in survey data sets), later versions of this framework may be required to make more detailed decisions on the variables included. Given the format of the data and the relative inability to test for multicollinearity and other common statistical measures, such patterns are best evidenced through classification iteration and hence may be re-considered before a final variable set is decided upon. The same can also be said for deciding on a number of clusters, the key here was to propose a framework and any fine-tunings (and decision on number of clusters) are likely to be as much a function of the purpose of the classification as the framework itself. This work has proposed a multi-purpose classification in a bid to test the framework's functionality. Having a clear purpose to the classification will aid decision making on such aforementioned matters.

More generally, and having discussed ecological fallacy at length, it is important to emphasise that the interpretation of results generated from this framework should not fall foul of exception fallacy. Exception fallacy, being the reverse of ecological fallacy, is the process whereby inferences are made about the characteristics of groups based on individual traits. Should results be aggregated to a selected geography, as in this work, for the purpose of visualisation, this renders the problem insignificant, however, if results are interpreted in tabular format (e.g. all 715,402 Leeds individuals) with some form of geography attached then it is a consideration. Visualisation of the classification results remains difficult and exploration of innovative means to achieve this would go some way to improving the usability of the outputs and negating both ecological and exception fallacies.

Finally, given this classification's total reliance on census data for its input variables, continued use of the framework remains achievable for the foreseeable future given the recent announcement of a further census in 2021 (albeit predominantly online census replacing the traditional paper-based approach) (ONS, 2014). With no confirmation of a national census beyond 2021, the framework may require adaptation with possible alternative datasets including Post Office records, local government data and credit checking agency data.

8.5. Further Research Opportunities

The inter-disciplinary opportunities that profiling at this level generates, in particular with an ability to profile against external datasets, offers a broad appeal to further research using this framework. This work has demonstrated an ability to link to the BHPS and explore behavioural datasets over and above pure census characteristics held directly within the classification. Opportunities therefore exist to profile against datasets such as the Health Survey for England, British Crime Survey and National Travel Survey to name three. When one considers the refinement of the framework in addition to such diverse profiling opportunities, scope for research extension is clear and policy implications brought about from more accurate and finer-level classifications offer incentive to pursue this research avenue. Given the new approaches put forward in this research and the lack of any directly parallel methods, the framework can no doubt be enhanced based on further research and testing, however, once fine-tuned, the benefits of such fine-level classification are highly apparent.

References

Abbas, J., Ojo, A. & Orange, S. (2009), Geodemographics – A Tool for Health Intelligence?, *Public Health*, 123 (1), p.35-39.

Acxiom (2009), *Giving You a Fresh Perspective on Your Customers: PersonicX Household v2.1,* Available online at: http://www.personicx.co.uk/2012planner/UserDocumentation/PersonicX%20Household%20Brochure%20%28short%29.pdf, [Accessed: 09/12/2009].

Ahmadi, P. & Samsami, F. (2010), Pharmaceutical Market Segmentation using GA K-Means, *European Journal of Economics, Finance and Administrative Sciences, 22, p.72-82.*

Aldenderfer, M. & Blashfield, R. (1984), *Cluster Analysis*, Beverly Hills, CA, Sage Press.

Ashby, D. and Longley, P. (2005), Geocomputation, Geodemographics and Resource Allocation for Local Policing, *Transactions in GIS,* 9, p.53-72.

Baker, K., Bermingham, J. & McDonald, C. (1979), *Proceedings of the MRS Conference*.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B. & Rossiter, D. (2005), SimBritain: A Spatial Microsimulation Approach to Population Dynamics, *Population, Place and Space*, 11, p.13-34.

Ballas, D. & Clarke, G. (2008), Spatial Microsimulation, in Fotheringham, A. & Rogerson, P. (2009), The SAGE Handbook of Spatial Analysis, Sage, London.

Beacon Dodsworth, (2007), *P2 People and Places Classification,* http://www.beacon-dodsworth.co.uk/products/people-classification/, [Accessed: 08/12/2009].

Beaumont, J. & Inglis, K. (1989), Geodemographics in Practice: Developments in Britain and Europe, *Environment and Planning A,* 21(5), p.587-604.

BFSC (2009), *Urban Land Use Patterns*, http://geographyfieldwork.com/UrbanModelsMEDCs.htm, [Accessed: 04/12/2009].

Birkin, M. & Clarke, M. (1988), SYNTHESIS - A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples, *Environment and Planning A,* 20, p.1645-1671.

Birkin, M. & Clarke, M. (1989), The Generation of Individual and Household Incomes at the Small Area Level Using Synthesis, *Regional Studies,* 23, p.535-548.

Birkin, M. (1995), Customer Targeting, Geodemographics and Lifestyle Approaches. In Longley, P. and Clarke, G. (Eds.), *GIS for Business and Service Planning,* Cambridge, GeoInformation.

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

Birkin, M., Clarke, G. & Clarke, M. (2002), *Retail Geography & Intelligent Network Planning,* Wiley, Chichester.

Birmingham City Council (2010), *An Analysis and Commercial Review of Geodemographic Classifications within the West Midlands,* www.westmidlandsiep.gov.uk/download.php?did=2583*,* [Accessed: 17/05/2013].

Booth, C. (1889) *Life and Labour of the People of London*. London, Macmillan.

Bowker, G. & Star, S. (1999), *Sorting Things Out*, MIT Press Cambridge, Mass.

Bradbrook, C. (2009), unpublished presentation / personal communication.

Bradbury, M. (2009), Personal Communication, face-to-face. On: 16/12/2009.

Brown, P. (1990), *Geodemographics: A Review of Recent Developments and Emerging Issues – Towards an R.R.L. Research Agenda*, Regional Research Laboratory Initiative Discussion Paper No. 5. Department of Town and Regional Planning, University of Sheffield.

Brown, P. & Batey, P. (1994), *The Design and Construction of a Geodemographic Targeting System: Super Profiles 1994*, URPERRL Working Paper 40, Department of Civic Design, University of Liverpool.

Buckham, B. (1999), *Simulated Annealing Applications,* Mechanical Engineering Department, University of Victoria. Available online at: http://www.me.uvic.ca/~zdong/courses/mech620/SA_App.PDF, [Accessed: 01/02/2010].

Burns, L. (2009), *Devising a Health/Deprivation Geodemographic Area Classification System*, http://cdu.mimas.ac.uk/case-studies/burns/index.htm, [Accessed: 12/02/2013].

CACI (2003), *Welcome to the New Acorn,* http://www.caci.co.uk/acorn/, [Accessed: 29/11/2009].

CACI, (1993), CACI's Insite System in Action, *Marketing Systems Today,* 8(1), p.10-13.

CASWEB (2001), (now: UK Data Service), http://casweb.mimas.ac.uk/ [Accessed: 03/01/2014].

CCSR (2001), *The Sample of Anonymised Records: Small Area Microdata*, Available online at: http://www.ccsr.ac.uk/sars/2001/sam/, [Accessed: 07/10/2011].

Charles Booth Online Archive (2001), *Charles Booth Online Archive*, http://booth.lse.ac.uk/, [Accessed: 09/12/2009].

Clarke, G. (1999) Geodemographics, Marketing and Retail Location. In Pacione, M. (Ed.) *Applied Geography: Principles and Practice,* London, Routledge.

CSISS (2009), *Charles Booth: Mapping London's Poverty, 1885-1903*, http://www.csiss.org/classics/content/45, [Accessed: 06/12/2009]

Curry, M. (1997), The Digital Individual and the Private Realm, *Annals of the Association of American Geographers,* 87, p.681-699.

Data.Gov.UK (2010), English Indices of Deprivation 2010, http://data.gov.uk/dataset/index-of-multiple-deprivation, [Accessed: 01/12/2013].

Day, P. (2009), *BBC Radio 4: Location Location*, http://www.bbc.co.uk/programmes/b00k4g5b, [Accessed: 09/11/2013].

Debenham, J. (2003), *Extending Geodemographics with Supply-side Variables*, Unpublished Document.

ESDS (2011), *An ESDS Guide: Guide to British Household Panel Survey*, Available online at: http://www.esds.ac.uk/longitudinal/access/bhps/L33196.asp, [Accessed: 04/10/2011].

EuroDirect (2006), *Cameo Online*, http://www.cameo-online.co.uk/cameo/aboutcameo.aspx, [Accessed: 03/12/2009].

Everitt, B., Landau, S. & Leese, M. (2001), *Cluster Analysis* (4th ed.), Arnold, London.

Experían (2007), *Analysis and Consultancy Services,* http://www.Experían.co.uk/business/products/data/232.html, [Accessed: 01/12/2009].

Experían (2009), *Mosaic United Kingdom*, http://www.Experían.co.uk/assets/business-strategies/brochures/Mosaic_UK_2009_brochure.pdf, [Accessed: 26/10/2013].

Experían (2013), *CAMEO Customer Segmentation & Analysis for the UK & Overseas*, http://www.callcredit.co.uk/products-and-services/consumer-marketing-data/segmentation-analysis, [Accessed: 12/12/2013].

Experían Netherlands (2013), *Mosaic Household*, available online at: http://documents.esd-toolkit.eu/Download.ashx?ID=169279 [Accessed: 05/01/2014].

Farr, M. & Webber, R. (2001), MOSAIC: From an Area Classification System to Individual Classification, *Journal of Targeting, Measurement and Analysis for Marketing*, 10(1), p.55-65.

Feardon, D. (2007), *Charles Booth: Mapping London's Poverty, 1885-1903,* Centre for Spatially Integrated Social Science, http://www.csiss.org/classics/content/45, [Accessed: 20/11/2009].

Feng, Z. & Flowerdew, R. (1998), Fuzzy Geodemographics: A Contribution from Fuzzy Clustering Method. In Carver S (Ed.), *Innovations in GIS 5*. London, Taylor and Francis, p.119-127.

Fotheringham, S. & Rogerson, P. (2009), *The SAGE handbook of Spatial Analysis*, SAGE, London.

Freedman, D. (2001) *Ecological Inference and the Ecological Fallacy*, in Smelser and Baltes (eds.), p. 4027-30.

Genest, C., Rémillard, B. & Beaudoin, D. (2009), Goodness-of-Fit Tests for Copulas: A Review and a Power Study, *Insurance: Mathematics and Economics*, 44(2), p.199-213.

Gibson, P. & See, L. (2006), Using Geodemographics and GIS for Sustainable Development. In Campagna, M. *GIS for Sustainable Development*, Taylor and Francis, London.

Goffe, W.L., Ferrier, G. & Rogers, J. (1994), Global Optimization of Statistical functions with Simulated Annealing, *Journal of Econometrics*, 60, p.65–99.

Goss, J. (1995), 'We Know Who You Are And We Know Where You Live': The Instrumental Rationality of Geodemographic Systems, *Economic Geography*, 71(2), p.171-204.

Greenland. S. & Robins J. (1994), Invited commentary: Ecologic Studies-Biases, Misconceptions, and Counter Examples, *American Journal of Epidemiology*, 139, p.747-760.

Gregory, D. (ed), Johnston, R., Pratt, P., Watts, M. & Whatmore, S. (2009), *Dictionary of Human Geography: Fifth Edition,* Blackwell Publishing, Oxford.

Hagerstrand, T. (1967), *Innovation Diffusion as a Spatial Process*, University of Chicago Press, Chicago.

Harland, K., Birkin, M., Smith, D. & Heppenstall, A. (2009a), *Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques*. In review.

Harland, K., Birkin, M., Smith, D. & Heppenstall, A. (2009b), *Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques*. Non-Published, pre-submitted draft version of 2009a.

Harris, R. & Longley, P. (2004), Targeting Clusters of Deprivation in Cities. Chapter 6 in Clarke ,G. & Stillwell, J. (eds.) *Applied GIS and Spatial Analysis*. Chichester, John Wiley and Sons, p.89-110.

Harris, R., Sleight, R., & Webber, R. (2005), *Geodemographics, GIS and Neighbourhood Targeting,* Wiley Publishing, Chichester.

Harris, R. & Johnston, R. (2003), Spatial Scale and Neighbourhood Regeneration in England: A Case Study of Avon, *Environment and Planning C: Government and Policy,* 21(5), p.651-662.

Highfield, R., & Fleming, N. (2007), *Postcodes Help Big Brother Keep an Eye on us,* http://www.telegraph.co.uk/scienceandtechnology/science/sciencenews/3306417/Post-codes-help-Big-Brother-keep-an-eye-on-us.html, [Accessed: 27/11/2009].

Hill (1990), No Reference Available (cited in Roberts, 1992).

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

Huang, Z. (1998), Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and Knowledge Discovery*, 2 (3), p.283-304.

Hyndman, H. (1911), *The Record of an Adventurous Life*, Macmillan, New York.

Invest North Yorkshire (2013), *The Best Life,* http://www.investnorthyorkshire.co.uk/life/, [Accessed: 01/12/2013].

Institute for Social and Economic Research (ISER), 2011, *British Household Panel Survey,* https://www.iser.essex.ac.uk/bhps, [Accessed: 02/12/2013].

Jain, A. & Dubes, R. (1988) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ

Johnson, M. (1989), The Application of Geodemographics to Retailing – Meeting the Needs of the Catchment, *Journal of Research Society,* 31, pp.7-36.

Johnston, R. (ed), Gregory, D., Pratt, & P., Watts, M. (2000) *Dictionary of Human Geography: Fourth Edition,* Blackwell Publishing, Oxford.

Kaufman L. & Rousseeuw P.J. (1990), *Finding Groups in Data*, New York: Wiley.

Krzanowski, W. & Marriott, F. (1995), *Multivariate Analysis Part 2: Classification, Covariance Structures and Repeated Measurements,* Edward Arnold, London.

Larson, E. (1992), *The Naked Consumer: How our Private Lives Become Public Commodities,* Henry Holt, New York.

Legates, D. & McCabe (1999), Evaluating the Goodness-of-Fit Measures in Hydrologic and Hydroclimatic Model Validation, *Water Resources Research*, 31(1), p.233-241.

Lemeshow, S. & Hosmer, D (1982), A Review of Goodness of Fit Statistics for use in the Development of Logistic Regression Models*, American Journal of Epidemiology*, 115(1), p. 92-106.

Leung, Yee. (2010), *Knowledge Discovery in Spatial Data*, Springer-Verlag, Berlin.

Levene, T. (1999), *Postcode Persecution*, published in The Guardian, 06/02/1999.

Leventhal, B. (1995), Evaluation of Geodemographic Classifications*, Journal of Targeting, Measurement and Analysis for Marketing*, 4, p.173-183.

Longley, P. & Batty, M. (2003) (eds), *Advanced Spatial Analysis,* The CASA Book of GIS, Redlands, CA, ESRI Press.

Longley, P. A. (2005) A Renaissance of Geodemographics for Public Service Delivery, *Progress in Human Geography*, 29, p. 57-63.

Martin, D. (1991), Geographic Information Systems and their Socioeconomic Applications, London, Routledge.

McKinley, R. & Mills, C. (1985), A Comparison of Several Goodness-of-Fit Statistics, *Applied Psychological Measurement*, 9(1), p.49-57.

Metropolis, N., Rosenbluth, A., Rosenbluth N., Teller A., & Teller, E. (1953), Equation of State Calculation by Fast Computing Machines, *Journal of Chemical Physics*, 21(6), p.1087–1092.

Milligan, G. (1996), Clustering Validation: Results and Implications for Applied Analyses, in Arabie, P., Hubert, L. J. and De Soete, G. Eds., *Clustering and Classification*, Singapore, World Scientific.

Morphet, C. (1993), The Mapping of Small-Area Census Data – A Consideration of the Effects of Enumeration District Boundaries, *Environment and Planning A*, 25, p.1267-1277.

Morrissey, K., & Clarke, G., Ballas, D., Hynes, S. & O'Donoghue, C. (2008), Examining Access to GP Services in Rural Ireland Using Microsimulation Analysis, *Area*, 40(3) p.354-364.

Neilson, J. (1993), Labour Market, Income Formation and Social Security in the Microsimulation Model NEDYMAS, *Economic Modelling*, 10, p.225-272.

Neilson (2005), *PRIZM*, http://en-us.nielsen.com/tab/product_families/nielsen_claritas/prizm, [Accessed: 04/12/2009].

Nelson, C (2003), *Do Birds of a Feather Flock Together?,* http://www.foreseechange.com/Geodemographic%20Segmentation.pdf [Accessed: 04/12/2009].

Office for National Statistics [ONS] (2011), *Regions (Formers GORs)*, http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/government-office-regions/index.html, [Accessed: 12/02/2014].

Office for National Statistics [ONS] (2014), *The Census and Future Provision of Population Statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority, and the Government's Response*, Available online at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-report-on-autumn-2013-consultation--and-recommendations/index.html, [Accessed: 06/08/2014] .

Openhaw, S. (1984), Ecological Fallacies and the Analysis of Areal Census Data, *Environment and Planning A*, 16, p.17-31.

Openshaw, S. & Rao, L. (1995), Algorithms for Reengineering 1991 Census Geography, *Environment and Planning A,* 27, p.425-446.

Openshaw, S & Taylor, P. (1979), A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem, *Statistical Methods in the Spatial Sciences*, p.127-144.

Openshaw, S. & Wymer, C. (1995), Classification and Regionalization*, Census Users*' *Handbook,* S, Openshaw (Ed.) GeoInformation International, Cambridge p.239 – 270.

Openshaw*,* S. (*1989*), Making Geodemographics More Sophisticated, *Journal of the Market Research Society*, 31(1), p.111-32.

Osborne, H. (2006), *Norwich Union offers 'pay as you drive' insurance,* http://www.guardian.co.uk/money/2006/oct/05/business.motorinsurance, [Accessed: 04/12/2009].

Postcode Address File (PAF) (2014), *UK Postcodes, 28 Million Addresses, Counties, Wards, Grid References, Boundary files and so much more...*, http://www.postcodeaddressfile.co.uk, [Accessed: 01/08/2014].

Park, R., & Burgess, E. (1925), *The City*, Chicago University Press, Chicago.

Picket, K. & Pearl, M. (2001), Critical Review Socioeconomic Context and Health Outcomes: A Multilevel Analyses of Neighbourhood, *Journal of Epidemiol Community Health,* 55, p.111-122.

Pinker, S. (1997), *How the Mind Works*, Penguin, London.

Pointer, M. & *Attridge*, G. (*1998*), The Number of Discernible Colours, *Colour Research and Application*, 23, p.52-54.

Propper, C. (1995), *For Richer, For Poorer, In Sickness and in Health: The Lifetime Distribution of NHS Health Care*. In: Falkingham, J & Hills (eds), The Dynamic if Welfare: The Welfare State and The Liftcycle, p.184-203. New York, Prentice Hall.

Rees, P. H. (1970), Concepts of Social Space: Towards an Urban Social Geography, in *Geographic Perspectives on Urban Systems,* Eds. Berry, B. & Horton, F. Prentice-Hall, Englewood Cliffs, NJ.

Rees, P. (1979), *Residential Patterns in American Cities: 1960*, Research Paper no. 189, Department of Geography, University of Chicago.

Rees, P., Durham, H. and Kupiszewski, M. (1996), *Internal Migration and Regional Population Dynamics in Europe: United Kingdom Case Study*, Working Paper 96/20, School of Geography, University of Leeds.

Rees, P., Vickers, D. & Birkin, M. (2005), *The ONS 2001 OA Classification*, Paper presented at the Conference on Census: Present and Future, ESRC/JISC 2001 Census of Population Programme, Gilbert Murray Conference Centre, University of Leicester, 16-17 November 2005.

Rees, P. (2013), personal communication on 21/11/2013.

Rezanková (2009), *Cluster Analysis and Categorical Data*, available online at: http://panda.hyperlink.cz/cestapdf/pdf09c3/rezankova.pdf, [Accessed: 18/02/2011].

Roberts, K. (1992), *Aiming for a Direct Hit*, Information Week, 20 April, p.26-30.

Robin, J. (1980), *Geodemographics: The New Magic* Campaigns and Elections, Spring edition, p.25-46.

Robinson, G. M. (1998), *Methods and Techniques in Human Geography*, Chichester, UK, Wiley.

Ronson, J. (2005), *Who Killed Richard Cullen?*, The Guardian, http://www.guardian.co.uk/money/2005/jul/16/creditcards.debt, [Accessed: 08/12/2009].

Rothman, J. (1989) Editorial. *Journal of the Market Research Society*, 31 (1), p.17.

See , L. (2009), *Building a Geodemographic System*, GEOG5191M Geodemographics and Database Marketing Unit 2 Notes, unpublished document, University of Leeds, UK.

Shepherd, P. (2006), *Neighbourhood profiling and classification for community safety*, Unpublished PhD Thesis, School of Geography, University of Leeds, Leeds.

Singleton, A. (2004), *A State of the Art Review of Geodemographics and their Applicability to the Higher Education Market*, Working paper. Available online at: http://www.bartlett.ucl.ac.uk/casa/publications/working-paper-74 [Accessed: 21/12/2013].

Singleton, A. & Longley, P. (2009), *Geodemographics, Visualisation, and Social Networks*, Applied Geography, 23(3), p.289-298.

Singleton, A. (2007), *A Spatio-Temporal Analysis of Access to Higher Education,* PhD Thesis, Department of Geography, University College London.

Sleight, P. (1995), *Explaining Geodemographics: What It Is, What Do You Use It For, and Where Is It Going?*, Admap Magazine, January Edition.

Sleight, P. (1997), *Targeting Customers: How to use Geodemographic and Lifestyle Data in your Business*, NTC Publications, Henley-on-Thames, UK.

Sleight, P. (2003) *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*, 3rd edition. Henley-on-Thames, UK. World Advertising Research Centre Publications.

Smith, D., Clarke, G., & Harland, K. (2009), Improving the Synthetic Data Generation Process in Spatial Microsimulation Models, *Environment and Planning A,* 41, p.1251-1268.

SPSS Log (2006), *Nominal, Ordinal and Scale,* http://www.spsslog.com/2006/05/03/nominal-ordinal-and-scale/, [Accessed: 16/02/2011].

Stillwell, J. & Clarke, G. (2004) (eds), *Applied GIS and Spatial Modelling*, London, Wiley.

Sui, D. (1998), Deconstructing Virtual Cities: From Reality to Hyper Reality. *Urban Geography.* 19(7), 657-676.

Tobler, W. (1970), A Computer Movie Simulating Urban Growth in the Detroit Region, *Economic Geography*, 46, p.234-40.

Tomintz, M. & Clarke, G. (2008), The Geography of Smoking in Leeds: Estimating Individual Smoking Rates and the Implications for the Location of Stop Smoking Services, *Area*, 40(3), p.341-353.

Tranmer, M., & Steel, D. (1998), Using Census Data to Investigate the Cause of the Ecological Fallacy, *Environment and Planning A*, 30, p.817-831.

UK Data Service (UKDS) (2014), *Census Microdata*, available online at: http://census.ukdataservice.ac.uk/get-data/microdata.aspx, [Accessed: 05/08/2014].

Vickers, D. & Rees, P. (2006), Creating the UK National Statistics 2001 Output Area Classification, *Journal of the Royal Statistical Society: Series A*, 170(2), p.379-403.

Vickers, D. (2006), *Multi-level Integrated Classifications Based on the 2001 Census*, Unpublished PhD Thesis, School of Geography, University of Leeds, Leeds.

Vickers, D. (2008), *Creating a Geodemographic Classification*, http://www.mrs.org.uk/pdf/03_11_08_dan_vickers.pdf, [Accessed: 11/11/2013].

Voas, D. & Williamson, P. (2000), An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata, *International Journal of Population Geography,* 6, p.349-366.

Voas, D. & Williamson, P. (2001), Evaluating Goodness-of-Fit Measures for Synthetic Microdata, *Geographical and Environmental Modelling,* 5, p.177-200.

Voas, D. & Williamson, P. (2001), The Diversity of Diversity: A Critique of Geodemographic Classification, *Area*, 33 (1), p.63-76.

Wallace, M., Charlton, J. & Denham, C. (1995), The new OPCS area classifications, *Population Trends*, 79, p.15-30.

Warwickshire Observatory (2011), *An Introduction to Mosaic*, Available online at: http://www.warwickshireobservatory.org/observatory/observatorywcc.nsf/0/1DDF5456EFC1D9448025779D0039DDEC/$file/Mosaic%202010%20Briefing%20Note.pdf, [Accessed: 22/01/2014].

Webber (2007), *How Geodemographic Classifications Are Built,* http://www.mosaic.geo-strategies.com/wp-content/uploads/2007/11/article-building-mosaic.pdf, [Accessed: 10/11/2013].

Weinberger, D. (2007), *Prologue: Information in Space*, http://www.everythingismiscellaneous.com/wp-content/samples/eim-sample-prologue.html, [Accessed: 09/12/2009].

Weiss, M. (2000), *The Clustered World*, Little Brown and Company, New York.

Williamson, P., Birkin, M. & Rees, P. (1998), The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymised Records, *Environment and Planning A*, 30, p.785-816.

Williamson, P., Clarke, G. & McDonald, A. (1996), *Estimating Small-Area Demands for Water With The Use Of Microsimulation.* In: Clarke, G. (1996), Microsimulation for Urban and Regional Policy Analysis. London, Pion Ltd.

Wilson, A. & Pownall, C. (1976), A New Representation of the Urban System for Modelling and for the Study of Micro-Level Interdependence, *Area*, 8, p.246-254.

Wilson, A. (2000), *Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis*, London, Prentice Hall.

Withnall, A. (2014), *Royal Mail postcode 40th anniversary: Study reveals what your postcode says about you*, http://www.independent.co.uk/news/uk/home-news/royal-mail-postcode-40th-anniversary-study-reveals-what-your-postcode-says-about-you-9274139.html, [Accessed: 01/08/2014].

Wong, D. (1995), *Aggregation Effects in Geo-Referenced Data*, in Arlinghaus, S. (ed.) Practical Handbook of Spatial Statistics. CRC Press, p.83-106.

Appendix A

A.1. Full SAM Variable List (in format: SAM code – SAM description)

Below is a list of SAM variables as referred to in the SAM data dictionary.

acctypa - Accommodation Type
agea - Age of Respondents
bathwc - Use of Bath/Shower/Toilet
carsh - Cars/Vans Owned or Available for Use
cemtyp - Type of communal establishment
cenheat0 - Central Heating
ceststat - Status in Communal Establishment
cobirta - Country of Birth
combgn - Community Background - Religion or Religion Brought Up In (N.Ireland)
country - Country
defra - DEFRA: urban/rural type (numerical)
densitya - No. of Residents per Room
dfdisp - DEFRA: Dispersed Pop
dflgmtwn - DEFRA: Large Market Twn Pop
dflgurb - DEFRA: Large Urb Pop
dfmjurb - DEFRA: Major Urb Pop
dfothurb - DEFRA: Other Urb Pop
dfrutnp1 - DEFRA: Rural Twn Pop
dfrutnp2 - DEFRA: Rural Twn Pop (Includ Lrge Market Twn Pop)
dftotal - DEFRA: Total Pop
dftotrupa - DEFRA: Total Rural Pop (Includ Lrge Market Twn Pop)
dftotrupb - DEFRA: Total rural % (Includ Lrge Market Twn Pop)
dftoturb - DEFRA: Total Urb Pop(Exclud. Lrge Market Twn Pop)
dfvilp - DEFRA: Village Pop
distmova - Distance of Move for Migrants (km)
distwrka - Distance to Work (Including Study in Scotland)
econach - Economic Activity (last week)
edisdono - Number of times information donated
ethewa - Ethnic Group for England and Wales
ethn - Ethnic Group for Northern Ireland
ethsa - Ethnic Group for Scotland
everwork - Ever Worked
famtypa - Family Type
fndepcha - Dependent Children in Family
freconac - Economic Position of Family Reference Person
frnssec8 - NS-SEC Social-Economic Classifications of Family Reference Person
frsex - Sex of Family Reference Person
furn - Accommodation Furnished (Scotland)
genind - Generation Indicator
health - General Health Over the Last Twelve Months
hedind - Household Education indicator
hempind - Household Employment indicator
hhsgind - Household housing indicator
hhtlhind - Household health & disability indicator

hmptpuk - Hhd headship (ODPM)
hncarers - Number of Carers in the Household
hnearnra - Number of Employed Adults in Household
hnllti - Number in Household with Limiting Long-term Illness
hnprhlth - Number of Household Members with Poor Health
hnresida - Number of Usual Residents in Household
hourspwg - Hours Worked Weekly
hrsocgrd - Social Grade of Household Reference Person
id - ID within country
lacode - local authority (GB) or parliamentary constituency (NI)
lastwrka - Year Last Worked
llti - Limiting Long Term Illness
lowflora - Lowest floor level of household living accommodation
marstata - Marital Status
miginda - Migration Indicator
migorgn - region of origin
nssec8 - NS-SEC 8 classes
occupncy - Occupancy Rating of Household
oncperim - One Number Census status
onscode - ONS LA indicator
pnum - record identifier within country
popbasea - Population Base qualifier
profqual - Professional Qualification (England and Wales) .
provcare - Number of Hours Care Provided per Week
qualvewn - Level of Highest Qualifications (Aged 16-74, EWN)
qualvs - Level of highest qualifications (16-74) (Scotland)
regiona - Region of usual residence
relgew - Religion (England and Wales)
relgn - Religion (Northern Ireland)
relgs1 - Religion Belongs to (Scotland)
reltohra - Relationship to HRP
roomsflr - Number of Floor Levels (Northern Ireland)
roomsnum - Number of Rooms Occupied in Household Space
selfcont - Accommodation Self-Contained
sex - Sex
stahuka - Household with Students Away During Term Time
student - Schoolchild or Student in Full-Time Education . . .
supervsr - Supervisor/Foreman
tenurewa - Tenure of Accommodation, England and Wales
tenursna - Tenure of Accommodation, NI, Scotland
termtima - Term time Address of Students or Schoolchildren .
tranwrka - Transport to Work, UK (Including to Study in Scotland)
workforc - Size of Work Force
wrkplcea - Workplace
zacctypa - acctypa imputation flag
zagea - agea imputation flag
zbathwc - bathwc imputation flag
zcarsh - carsh imputation flag
zcemtyp - cemtyp imputation flag
zcenheat - cenheat imputation flag
zceststa - ceststat imputation flag
zcobirta - cobirta imputation flag
zcombgn - combgn imputation flag
zdensity - density imputation flag
zdistmov - distmov imputation flag

zdistwrk - distwrka imputation flag
zeconach - econach imputation flag
zethewa - ethewa imputation flag
zethn - ethn imputation flag
zethsa - ethsa imputation flag
zeverwor - everwork imputation flag
zfamtypa - famtypa imputation flag
zfndepch - fndepch imputation flag
zfrecona - frecona imputation flag
zfrnssec - frnssec imputation flag
zfrsex - frsex imputation flag
zfurn - furn imputation flag
zgenind - genind imputation flag
zhealth - health imputation flag
zhedind - hedinid imputation flag
zhempind - hempind imputation flag
zhhlthin - hhtlhind imputation flag
zhhsgind - hhsgind imputation flag
zhmptpuk - hmptpuk imputation flag
zhncarer - hncarer imputation flag
zhnearnr - hnearnr imputation flag
zhnllti - hnllti imputation flag
zhnprhlt - hnprhlt imputation flag
zhnresid - hnresid imputation flag
zhourspw - hourspw imputation flag
zhrsocgr - hrsocgr imputation flag
zlastwrk - lastwrk imputation flag
zllti - llti imputation flag
zlowflor - lowflor imputation flag
zmarstat - marstat imputation flag
zmiginda - miginda imputation flag
zmigorgn - migorgn imputation flag
znssec8 - nssec8 imputation flag
zoccupnc - occupncy imputation flag
zprofqua - profqual imputation flag
zprovcar - provcare imputation flag
zqualvew - qualvew imputation flag
zqualvs - qualvs imputation flag
zregiona - regiona imputation flag
zrelgew - relgew imputation flag
zrelgn - relgn imputation flag
zrelgs1 - relgs1 imputation flag
zreltohr - reltohr imputation flag
zroomsfl - roomsflr imputation flag
zroomsnu - roomsnum imputation flag
zselfcon - selfcont imputation flag
zsex - sex imputation flag
zstahuka - stahuka imputation flag
zstudent - student imputation flag
zsupervs - supervsr imputation flag
ztenure - tenure imputation flag
ztenursn - tenursni imputation flag

Full 2001 SAM Codebook available online at:
http://www.ccsr.ac.uk//sars/2001/sam/variables/samcodebook20070604.pdf

A.2. SAM Variable Look-Up

Below is a look-up table for SAM variables as referred to at various junctures in this thesis, particularly chapter 6 (Tables 6.3 to 6.6) when referring to the fifteen selected variables.

SAM-listed Variable	SAM Variable Description
Age	Age (in Years)
CarVan	Car/Vans available for use (per household)
CenHeat	Central Heating provision
CoB	Country of Birth
EthGrp	Ethnic Group
HholdFamTyp	Family Type
Health	Self-Reported health status
No.UsualRes	Number of residents in household (typically)
HrsWrkdWkly	Number of hours in paid employment per week (typically)
MarStat	Legal marital status
HrsCareWkly	Number of hours spent caring for relative(s) per week (typically)
RelToHRP	Relationship to Household Reference Person
Sex	Legal sex definition
NSSEC	National-Statistics Socio-Economic Classification
Qual	Highest Level of Qualification

Appendix B

B.1. Individual to 2001 OA Special Case Assignments

In cases where two or more clusters share the same highest count of individuals per output area, the first cluster (numerically) is assigned as its membership. For the purpose of completeness, those output areas in question are listed below together with the final cluster membership. Green highlighting denotes instances where output areas share equal cluster count. The final column in the table, Cluster M'Ship, specifies the final cluster assignment based on the aforementioned ruling.

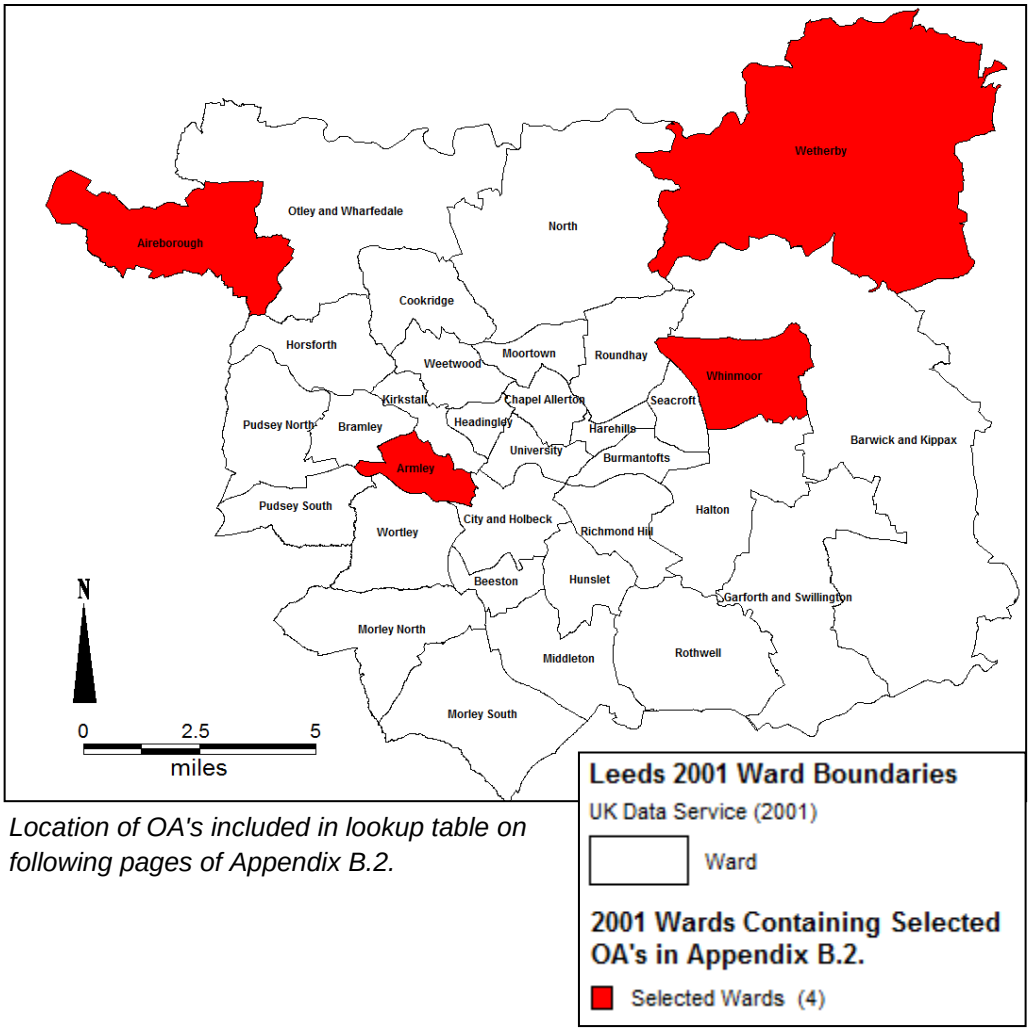
Given that this research is primarily to highlight a framework through which individual-level classifications can take place, taking this approach for output areas with more than one highest cluster value seems appropriate as a means of testing the model.

OA Zone Code	Ward Name	Total Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster M'Ship
00DAFE0022	Bramley	309	79	79	18	71	62	1
00DAFF0001	Burmantofts	312	62	73	52	52	73	2
00DAFF0034	Burmantofts	272	58	56	52	58	48	1
00DAFL0076	Halton	258	65	65	43	28	57	1
00DAFM0033	Harehills	283	78	34	64	29	78	1
00DAFQ0043	Hunslet	235	64	3	8	80	80	4
00DAFR0035	Kirkstall	307	62	58	66	55	66	3
00DAGA0017	Pudsey South	282	72	4	69	72	65	1
00DAGA0061	Pudsey South	330	82	69	29	68	82	1
00DAGB0017	Richmond Hill	322	92	67	61	10	92	1
00DAGC0057	Rothwell	294	26	78	41	78	71	2
00DAGF0066	University	234	38	12	56	64	64	4

B.2. Individual to 2001 OA Lookup Table

This section lists a sample of 2001 Leeds census output areas with context added through the inclusion of census ward names. The total number of individuals classified into each of the five clusters per output area are shown. The final column indicates the cluster membership based on the highest presence of individuals in each of the five clusters.

Included for the purpose of illustration are Leeds output areas in: Aireborough (north-west), Armley (central/west), Whinmoor (central/east) and Wetherby (north-east). See map below.



Location of OA's included in lookup table on following pages of Appendix B.2.

The full lookup table is available on the CD-ROM at the back of this thesis (Appendix C) should this copy include one (file name: OALookupTable.pdf). Otherwise, it can be obtained by contacting Luke Burns at: L.P.Burns@leeds.ac.uk.

OA Zone Code	Ward Name	Total Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster M'Ship
00DAFA0001	Aireborough	304	71	182	8	6	37	2
00DAFA0002	Aireborough	286	84	93	47	16	46	2
00DAFA0003	Aireborough	436	255	5	75	62	39	1
00DAFA0004	Aireborough	237	25	62	56	237	57	4
00DAFA0005	Aireborough	294	103	37	46	61	47	1
00DAFA0006	Aireborough	363	141	50	51	48	73	1
00DAFA0007	Aireborough	370	32	21	211	95	11	3
00DAFA0008	Aireborough	318	19	122	70	45	62	2
00DAFA0009	Aireborough	384	110	96	100	37	41	1
00DAFA0010	Aireborough	287	65	58	49	64	51	1
00DAFA0011	Aireborough	298	66	74	27	59	72	2
00DAFA0012	Aireborough	271	88	40	31	56	56	1
00DAFA0013	Aireborough	291	101	32	44	87	27	1
00DAFA0014	Aireborough	359	97	75	35	61	91	1
00DAFA0015	Aireborough	385	132	109	125	4	15	1
00DAFA0016	Aireborough	333	90	88	51	24	80	1
00DAFA0017	Aireborough	204	45	50	18	44	47	2
00DAFA0018	Aireborough	312	30	111	70	29	72	2
00DAFA0019	Aireborough	287	2	51	92	93	49	4
00DAFA0020	Aireborough	329	71	72	30	84	72	4
00DAFA0021	Aireborough	338	87	91	68	90	2	2
00DAFA0022	Aireborough	310	45	98	54	40	73	2
00DAFA0023	Aireborough	253	15	115	21	11	91	2
00DAFA0024	Aireborough	343	108	67	63	103	2	1
00DAFA0025	Aireborough	157	41	31	44	19	22	3
00DAFA0026	Aireborough	279	48	64	50	63	54	2
00DAFA0027	Aireborough	307	65	85	65	77	15	2
00DAFA0028	Aireborough	303	68	81	46	46	62	2
00DAFA0029	Aireborough	108	18	11	28	26	25	3
00DAFA0030	Aireborough	342	126	91	6	102	17	1
00DAFA0031	Aireborough	143	19	38	42	34	10	3
00DAFA0032	Aireborough	314	30	107	49	77	51	2
00DAFA0033	Aireborough	300	94	76	37	40	53	1
00DAFA0034	Aireborough	263	44	70	52	39	58	2
00DAFA0035	Aireborough	245	77	43	71	31	23	1
00DAFA0036	Aireborough	414	106	53	49	105	101	1
00DAFA0037	Aireborough	225	2	81	38	48	56	2
00DAFA0038	Aireborough	242	58	42	17	109	16	4
00DAFA0039	Aireborough	309	9	114	9	122	55	4
00DAFA0040	Aireborough	288	1	16	155	16	100	3
00DAFA0041	Aireborough	311	44	76	54	66	71	2
00DAFA0042	Aireborough	254	57	72	11	58	56	2
00DAFA0043	Aireborough	362	63	50	64	97	88	4
00DAFA0044	Aireborough	325	98	46	41	86	54	1

OA Zone Code	Ward Name	Total Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster M'Ship
00DAFA0045	Aireborough	303	112	46	92	18	35	1
00DAFA0046	Aireborough	285	67	69	26	84	39	4
00DAFA0047	Aireborough	253	8	70	65	77	33	4
00DAFA0048	Aireborough	269	32	72	71	61	33	2
00DAFA0049	Aireborough	258	59	117	23	40	19	2
00DAFA0050	Aireborough	299	22	56	111	84	26	3
00DAFA0051	Aireborough	285	16	38	100	53	78	3
00DAFA0052	Aireborough	240	24	12	79	86	39	4
00DAFA0053	Aireborough	347	118	69	91	64	5	1
00DAFA0054	Aireborough	288	19	112	3	89	65	2
00DAFA0055	Aireborough	314	69	65	65	57	54	1
00DAFA0056	Aireborough	309	29	59	77	84	60	4
00DAFA0057	Aireborough	258	4	88	45	118	3	4
00DAFA0058	Aireborough	322	80	18	77	69	78	1
00DAFA0059	Aireborough	312	126	0	36	59	91	1
00DAFA0060	Aireborough	291	59	62	113	56	1	3
00DAFA0061	Aireborough	272	44	78	75	28	47	2
00DAFA0062	Aireborough	334	92	61	29	87	65	1
00DAFA0063	Aireborough	268	13	97	50	81	27	2
00DAFA0064	Aireborough	319	92	62	63	34	68	1
00DAFA0065	Aireborough	257	25	58	59	74	41	4
00DAFA0066	Aireborough	321	114	50	101	28	28	1
00DAFA0067	Aireborough	236	14	48	25	70	79	5
00DAFA0068	Aireborough	307	6	98	69	87	47	2
00DAFA0069	Aireborough	209	54	74	69	0	12	2
00DAFA0070	Aireborough	210	12	66	9	113	10	4
00DAFA0071	Aireborough	258	75	67	53	23	40	1
00DAFA0072	Aireborough	248	26	124	41	7	50	2
00DAFA0073	Aireborough	351	169	22	79	62	19	1
00DAFA0074	Aireborough	305	101	62	93	38	11	1
00DAFA0075	Aireborough	225	0	110	3	50	62	2
00DAFA0076	Aireborough	265	73	86	72	8	26	2
00DAFA0077	Aireborough	257	40	57	32	78	50	4
00DAFA0078	Aireborough	271	30	23	71	62	85	5
00DAFA0079	Aireborough	208	47	42	52	47	20	3
00DAFA0080	Aireborough	304	128	53	59	28	36	1
00DAFA0081	Aireborough	328	51	73	110	86	8	3
00DAFA0082	Aireborough	271	53	41	75	70	32	3
00DAFA0083	Aireborough	230	57	59	51	58	5	2
00DAFA0084	Aireborough	278	31	142	19	49	37	2
00DAFA0085	Aireborough	232	62	38	45	38	49	1
00DAFA0086	Aireborough	343	53	111	46	60	73	2
00DAFA0087	Aireborough	250	35	80	42	28	65	2
00DAFA0088	Aireborough	282	98	74	45	21	44	1

OA Zone Code	Ward Name	Total Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster M'Ship
00DAFA0089	Aireborough	262	68	55	51	35	53	1
00DAFB0001	Armley	428	50	65	68	80	165	5
00DAFB0002	Armley	427	84	43	107	62	131	5
00DAFB0003	Armley	301	67	1	112	58	63	3
00DAFB0004	Armley	336	64	93	145	7	27	3
00DAFB0005	Armley	374	109	60	25	73	107	1
00DAFB0006	Armley	537	19	187	25	181	125	2
00DAFB0007	Armley	310	71	21	71	64	83	5
00DAFB0008	Armley	352	63	140	70	41	38	2
00DAFB0009	Armley	274	12	68	81	26	87	5
00DAFB0010	Armley	285	44	25	34	38	144	5
00DAFB0011	Armley	283	78	36	40	43	86	5
00DAFB0012	Armley	315	75	66	60	95	19	4
00DAFB0013	Armley	288	72	8	103	55	50	3
00DAFB0014	Armley	277	11	11	91	107	57	4
00DAFB0015	Armley	308	38	89	83	52	46	2
00DAFB0016	Armley	262	41	1	105	35	80	3
00DAFB0017	Armley	251	63	56	33	84	15	4
00DAFB0018	Armley	297	56	88	77	12	64	2
00DAFB0019	Armley	353	1	122	78	47	105	2
00DAFB0020	Armley	250	13	61	42	94	40	4
00DAFB0021	Armley	281	55	33	29	72	92	5
00DAFB0022	Armley	325	50	51	129	84	11	3
00DAFB0023	Armley	286	12	78	87	45	64	3
00DAFB0024	Armley	333	11	44	88	111	79	4
00DAFB0025	Armley	293	83	43	152	10	5	3
00DAFB0026	Armley	333	77	42	113	74	27	3
00DAFB0027	Armley	299	46	99	40	30	84	2
00DAFB0028	Armley	250	35	102	7	42	64	2
00DAFB0029	Armley	337	12	1	175	113	36	3
00DAFB0030	Armley	315	51	152	3	63	46	2
00DAFB0031	Armley	373	220	21	23	25	84	1
00DAFB0032	Armley	297	7	29	123	68	70	3
00DAFB0033	Armley	224	38	44	9	63	70	5
00DAFB0034	Armley	375	114	27	113	112	9	1
00DAFB0035	Armley	353	74	118	103	6	52	2
00DAFB0036	Armley	386	19	128	96	26	117	2
00DAFB0037	Armley	243	44	25	120	31	23	3
00DAFB0038	Armley	379	30	95	105	99	50	3
00DAFB0039	Armley	264	25	126	19	16	78	2
00DAFB0040	Armley	319	42	120	33	111	13	2
00DAFB0041	Armley	247	29	65	16	87	50	4
00DAFB0042	Armley	253	16	86	90	21	40	3
00DAFB0043	Armley	158	17	32	13	44	52	5

OA Zone Code	Ward Name	Total Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster M'Ship
00DAFB0044	Armley	209	61	14	7	65	62	4
00DAFB0045	Armley	288	21	59	81	45	82	5
00DAFB0046	Armley	369	25	139	70	64	71	2
00DAFB0047	Armley	280	50	40	46	88	56	4
00DAFB0048	Armley	196	54	0	51	25	66	5
00DAFB0049	Armley	319	93	89	57	21	59	1
00DAFB0050	Armley	183	40	50	35	54	4	4
00DAFB0051	Armley	309	66	9	120	27	87	3
00DAFB0052	Armley	263	100	10	23	71	59	1
00DAFB0053	Armley	220	62	12	11	62	73	5
00DAFB0054	Armley	289	82	13	72	3	119	5
00DAFB0055	Armley	294	67	19	89	36	83	3
00DAFB0056	Armley	290	59	20	105	58	48	3
00DAFB0057	Armley	193	38	41	46	48	20	4
00DAFB0058	Armley	665	182	173	207	6	97	3
00DAFB0059	Armley	297	72	76	15	78	56	4
00DAFB0060	Armley	278	86	19	89	41	43	3
00DAFB0061	Armley	323	22	96	102	76	27	3
00DAFB0062	Armley	229	31	34	47	69	48	4
00DAFB0063	Armley	199	30	36	43	40	50	5
00DAFB0064	Armley	267	56	66	29	74	42	4
00DAFB0065	Armley	337	26	71	140	53	47	3
00DAFB0066	Armley	233	62	6	148	14	3	3
00DAFB0067	Armley	270	70	4	133	24	39	3
00DAFB0068	Armley	310	18	121	17	28	126	5
00DAFB0069	Armley	223	37	47	64	2	73	5
00DAFB0070	Armley	182	14	23	25	54	66	5
00DAFB0071	Armley	219	21	13	61	56	68	5
00DAFB0072	Armley	260	4	98	31	114	13	4
00DAFB0073	Armley	281	29	50	79	61	62	3
00DAFB0074	Armley	328	61	105	66	94	2	2
00DAGH0001	Wetherby	359	108	5	93	54	99	1
00DAGH0002	Wetherby	353	94	84	30	59	86	1
00DAGH0003	Wetherby	313	98	52	23	92	48	1
00DAGH0004	Wetherby	395	200	45	43	100	7	1
00DAGH0005	Wetherby	266	74	56	39	46	51	1
00DAGH0006	Wetherby	339	99	98	59	53	30	1
00DAGH0007	Wetherby	360	121	84	76	4	75	1
00DAGH0008	Wetherby	302	42	113	50	84	13	2
00DAGH0009	Wetherby	267	108	53	7	85	14	1
00DAGH0010	Wetherby	280	32	84	63	17	32	2
00DAGH0011	Wetherby	294	92	91	33	75	3	1
00DAGH0012	Wetherby	325	101	24	7	96	97	1
00DAGH0013	Wetherby	296	65	56	51	61	63	1

OA Zone Code	Ward Name	Total Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster M'Ship
00DAGH0014	Wetherby	249	89	25	75	29	31	1
00DAGH0015	Wetherby	246	16	60	7	156	7	4
00DAGH0016	Wetherby	305	145	28	45	61	26	1
00DAGH0017	Wetherby	288	28	89	105	66	0	3
00DAGH0018	Wetherby	312	64	91	22	75	60	2
00DAGH0019	Wetherby	278	38	65	16	125	34	4
00DAGH0020	Wetherby	242	10	84	50	62	36	2
00DAGH0021	Wetherby	322	98	30	78	55	61	1
00DAGH0022	Wetherby	294	104	46	46	61	37	1
00DAGH0023	Wetherby	294	163	34	1	11	85	1
00DAGH0024	Wetherby	352	102	69	53	64	64	1
00DAGH0025	Wetherby	184	64	68	36	4	12	2
00DAGH0026	Wetherby	305	105	40	47	25	88	1
00DAGH0027	Wetherby	326	88	31	71	89	47	4
00DAGH0028	Wetherby	234	15	91	42	85	1	2
00DAGH0029	Wetherby	281	81	39	73	73	15	1
00DAGH0030	Wetherby	317	72	62	50	70	63	1
00DAGH0031	Wetherby	295	76	87	47	82	3	2
00DAGH0032	Wetherby	289	95	57	22	62	53	1
00DAGH0033	Wetherby	225	74	74	47	21	9	1
00DAGH0034	Wetherby	343	149	35	62	15	82	1
00DAGH0035	Wetherby	308	95	82	27	48	56	1
00DAGH0036	Wetherby	331	90	82	78	45	36	1
00DAGH0037	Wetherby	299	76	44	74	76	29	1
00DAGH0038	Wetherby	352	157	35	115	36	9	1
00DAGH0039	Wetherby	252	82	74	1	17	78	1
00DAGH0040	Wetherby	260	99	3	9	67	82	1
00DAGH0041	Wetherby	256	101	72	6	52	25	1
00DAGH0042	Wetherby	301	123	65	36	67	10	1
00DAGH0043	Wetherby	265	23	78	59	83	22	4
00DAGH0044	Wetherby	273	85	66	73	40	9	1
00DAGH0045	Wetherby	367	162	3	59	30	113	1
00DAGH0046	Wetherby	324	152	9	8	39	116	1
00DAGH0047	Wetherby	260	93	89	17	24	37	1
00DAGH0048	Wetherby	305	89	85	4	74	53	1
00DAGH0049	Wetherby	818	239	161	121	216	81	1
00DAGH0050	Wetherby	217	65	28	51	21	52	1
00DAGH0051	Wetherby	369	29	132	116	66	26	2
00DAGH0052	Wetherby	305	87	72	80	44	22	1
00DAGH0053	Wetherby	332	130	65	106	26	5	1
00DAGH0054	Wetherby	290	86	19	31	69	85	1
00DAGH0055	Wetherby	488	163	54	59	105	107	1
00DAGH0056	Wetherby	302	135	52	7	39	69	1
00DAGH0057	Wetherby	246	45	86	5	33	77	2

OA Zone Code	Ward Name	Total Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster M'Ship
00DAGH0058	Wetherby	251	99	15	13	48	76	1
00DAGH0059	Wetherby	287	66	88	19	108	6	4
00DAGH0060	Wetherby	351	103	102	57	68	21	1
00DAGH0061	Wetherby	300	91	24	37	77	71	1
00DAGH0062	Wetherby	317	84	47	73	56	57	1
00DAGH0063	Wetherby	329	125	10	33	92	69	1
00DAGH0064	Wetherby	361	131	88	127	10	5	1
00DAGH0065	Wetherby	593	237	4	149	140	63	1
00DAGH0066	Wetherby	286	88	13	87	63	35	1
00DAGH0067	Wetherby	187	19	62	25	41	40	2
00DAGH0068	Wetherby	271	73	69	63	22	44	1
00DAGH0069	Wetherby	292	147	41	38	61	5	1
00DAGH0070	Wetherby	311	87	13	57	75	79	1
00DAGH0071	Wetherby	311	113	79	32	59	28	1
00DAGH0072	Wetherby	326	93	35	78	90	30	1
00DAGH0073	Wetherby	228	79	68	71	4	6	1
00DAGH0074	Wetherby	287	93	58	38	67	31	1
00DAGH0075	Wetherby	169	17	27	46	8	71	5
00DAGH0076	Wetherby	236	88	5	40	77	26	1
00DAGH0077	Wetherby	297	74	57	46	48	72	1
00DAGH0078	Wetherby	301	101	22	84	63	31	1
00DAGH0079	Wetherby	267	108	51	7	97	4	1
00DAGH0080	Wetherby	322	4	126	78	1	113	2
00DAGH0081	Wetherby	263	12	1	123	113	14	3
00DAGH0082	Wetherby	291	79	70	85	56	1	3
00DAGH0083	Wetherby	195	34	16	37	47	61	5
00DAGH0084	Wetherby	367	45	60	60	157	45	4
00DAGH0085	Wetherby	278	155	8	49	20	46	1
00DAGH0086	Wetherby	379	136	53	5	126	59	1
00DAGH0087	Wetherby	170	40	22	24	23	61	5
00DAGJ0001	Whinmoor	287	26	84	41	35	101	5
00DAGJ0002	Whinmoor	301	64	2	11	95	129	5
00DAGJ0003	Whinmoor	324	58	8	61	93	104	5
00DAGJ0004	Whinmoor	365	3	90	84	87	101	5
00DAGJ0005	Whinmoor	309	99	44	3	142	21	4
00DAGJ0006	Whinmoor	250	25	9	20	87	109	5
00DAGJ0007	Whinmoor	339	113	17	92	40	77	1
00DAGJ0008	Whinmoor	308	68	123	10	51	56	2
00DAGJ0009	Whinmoor	264	5	3	181	33	42	3
00DAGJ0010	Whinmoor	367	111	1	78	97	80	1
00DAGJ0011	Whinmoor	271	73	46	100	37	15	3
00DAGJ0012	Whinmoor	330	33	162	23	39	73	2
00DAGJ0013	Whinmoor	325	118	10	90	5	102	1
00DAGJ0014	Whinmoor	219	47	10	5	60	97	5

OA Zone Code	Ward Name	Total Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster M'Ship
00DAGJ0015	Whinmoor	244	59	8	67	49	61	3
00DAGJ0016	Whinmoor	342	76	75	142	10	39	3
00DAGJ0017	Whinmoor	350	87	41	77	79	66	1
00DAGJ0018	Whinmoor	374	87	111	19	50	107	2
00DAGJ0019	Whinmoor	279	7	40	93	72	67	3
00DAGJ0020	Whinmoor	237	24	33	71	36	73	5
00DAGJ0021	Whinmoor	329	138	33	23	34	101	1
00DAGJ0022	Whinmoor	299	77	26	74	81	41	4
00DAGJ0023	Whinmoor	289	79	53	32	57	68	1
00DAGJ0024	Whinmoor	326	57	13	103	96	57	3
00DAGJ0025	Whinmoor	339	50	58	41	95	50	4
00DAGJ0026	Whinmoor	300	77	63	50	28	82	5
00DAGJ0027	Whinmoor	271	51	7	66	66	81	5
00DAGJ0028	Whinmoor	275	51	46	17	5	156	5
00DAGJ0029	Whinmoor	314	92	12	96	2	112	5
00DAGJ0030	Whinmoor	193	44	28	39	34	48	5
00DAGJ0031	Whinmoor	236	22	34	3	73	104	5
00DAGJ0032	Whinmoor	206	31	33	52	33	57	5
00DAGJ0033	Whinmoor	334	23	87	76	92	56	4
00DAGJ0034	Whinmoor	352	81	71	79	121	0	4
00DAGJ0035	Whinmoor	287	25	66	39	35	122	5
00DAGJ0036	Whinmoor	310	42	54	2	90	122	5
00DAGJ0037	Whinmoor	348	68	27	97	109	47	4
00DAGJ0038	Whinmoor	266	4	63	58	65	76	5
00DAGJ0039	Whinmoor	275	49	49	14	71	92	5
00DAGJ0040	Whinmoor	252	41	27	55	26	103	5
00DAGJ0041	Whinmoor	338	56	39	91	30	122	5
00DAGJ0042	Whinmoor	291	49	52	76	85	29	4
00DAGJ0043	Whinmoor	293	19	20	88	113	53	4
00DAGJ0044	Whinmoor	300	38	30	107	88	37	3
00DAGJ0045	Whinmoor	321	3	94	47	80	97	5
00DAGJ0046	Whinmoor	261	33	75	19	54	80	5
00DAGJ0047	Whinmoor	374	165	66	4	41	98	1
00DAGJ0048	Whinmoor	293	78	36	98	75	6	3
00DAGJ0049	Whinmoor	345	33	75	103	113	21	4
00DAGJ0050	Whinmoor	300	61	112	11	23	93	2
00DAGJ0051	Whinmoor	144	7	27	38	44	28	4
00DAGJ0052	Whinmoor	227	8	68	46	32	73	5
00DAGJ0053	Whinmoor	253	77	7	50	93	26	4
00DAGJ0054	Whinmoor	271	62	25	73	47	64	3
00DAGJ0055	Whinmoor	424	128	89	52	58	97	1
00DAGJ0056	Whinmoor	342	85	83	59	63	52	1
00DAGJ0057	Whinmoor	298	50	121	5	120	2	2
00DAGJ0058	Whinmoor	288	53	69	67	37	62	2

Appendix C

C.1. CD-ROM of Thesis and Supporting Data

Electronic copy of thesis (L.Burns_Thesis.pdf) and full OA Lookup Table (OALookupTable.pdf) available on CD-ROM located on inside of back cover.