Multi-Level Integrated Classifications Based on the 2001 Census

Daniel William Vickers

The University of Leeds

School of Geography

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

January 2006

The Candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

First and foremost I must thank my supervisors Phil Rees and Mark Birkin. Phil especially has provided support, advice and guidance far beyond that required of a PhD supervisor. I will be forever grateful and indebted to him for his constant encouragement and the unwavering belief that he has shown in me. I would also like to thank Graham Clarke and John Stillwell, for their valuable nuggets of advice in their role as members of my research support group.

I would like to acknowledge the funding that was provided for the project by the Economic and Social Research Council (ESRC) and CASE partner, the Office for National Statistics (CASE Award PTA-033-2002-00067). Without such generous financial support the project would not have been possible.

Thanks must go to all three census agencies ONS, GROS and NISRA for providing 2001 Census data free of charge through the ‘click-use’ licensing agreement. Also to the dissemination units of ESRC/JISC Census Programme, whose websites I made extensive use.

Staff at the ONS have been invaluable in the completion of the classification and were instrumental in its publication as a ‘National Statistic’. John Charlton my initial CASE supervisor, is worthy of a special mention. John was the driving force of the project within the ONS and provided much needed support in the early stages of the project. Sadly John passed away before the completion of the classification and never got to see its publication. Other ONS staff who helped in the development of the classification include Andrew Botterill, Simon Compton, Chris Denham, Greg Phillpotts, Samantha Poole, Nargis Rahman, Nick Ormiston-Smith, Dev Virdee, Alison Whitworth and Susan Williams. I would also like to thank anyone else within ONS who was involved in the publication of the classification, but of whom I am unaware.

Many thanks must go to Keith Dugmore (Demographic Decisions Ltd.), and Martin Callingham (Birkbeck College, University of London). At the start of the project I could not have hoped to find two such enthusiastic and vocal supporters of my work.

Thanks to Pete Shepherd for all the umpteen discussions about classification and the sharing of ideas and references. Talking to someone who was equally obsessed with cluster analysis as me was a great help. Thanks to Ian Shuttleworth (Queens University Belfast) for casting his experienced eye over my great simplification of demographic issues in Northern Ireland.

Thank you to everybody who took part in the consultation exercise, the small amount of time and effort that you gave me was of immeasurable value.

I would like to thank all my friends, family and colleagues with whom I have shared the last three years of my life and to anyone else who offered, help, support, advice or just showed an interest.

Abstract

The purpose of this thesis is to describe and explain the processes and decisions that were involved in the creation of the National Area Classification of 2001 Census Output Areas (OAs). The thesis describes the creation of the classification: selection of the variables, assembly of the classification database, the methods of standardisation and the clustering procedures, together with some discussion of alternative methodologies that were considered for use. The processes used for creating the clusters, their naming and description are outlined. The classification is mapped and visualised in a number of different ways.

In order to enable a classification of OAs to be possible the document starts with a review of the history of area classification and issues surrounding its future development. The methodological and theoretical issues in the creation of a classification system are also discussed. In order to test out the practicalities of creating a classification system, a classification of UK local authorities was created prior to the construction of the OA classification.

The thesis describes the quality assurance procedures that the OA classification was put through. This included an innovative consultation exercise. This ensured that the classification was of enough quality and without error, enabling it to be published as a 'National Statistic'. Examples of use of the classification are presented, outlining the value and relevance of the classification to social research. The OA classification is connected to other scales of classification to form a multi-scale classification system enabling the socio-demographic pattern of the UK to be examined at multiple scales.

The project had to overcome numerous methodological issues due to the size of dataset that was used. The project used a new methodology to create the first free-to-use small scale classification of the UK. The classification was published as a 'National Statistic' on the 29th July 2005 and is freely available. The classification can be downloaded from the ONS website at http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/oa/default.asp or via the School of Geography, University of Leeds website at: http://www.geog.leeds.ac.uk/people/d.vickers/OAclassinfo.html. Alternatively it can be ordered on CD by contacting info@statistics.gov.uk. The publication of the classification as a 'National Statistic' has created a resource that can be used by private, public and academic researchers.

Table of Contents

List of Figures

List of Tables

Abbreviations

ACORN	A Classification of Residential Neighbourhoods
BBC	British Broadcasting Corporation
BMRB	British Market Research Bureau
BCS	British Crime Survey
CACI	Consolidated Analysis Centers Inc.
CAMS	Controlled Access Microdata Sample
CASE	Collaborative Awards in Science and Engineering
CASWEB	Census Area Statistics Website
CCG	Centre for Computational Geography
CHCC	Collection of Historical and Contemporary Census Data and Related Materials
CIA	Central Intelligence Agency
Con	Conservative Party
DEFRA	Department of the Environment, Food and Rural Affairs
DfES	Department for Education and Skills
DUG	Demographic User Group
DVLA	Driver and Vehicle Licensing Agency
E&W	England and Wales
ED	Enumeration District
EDINA	Edinburgh Data and Information Access
ESRC	Economic and Social Research Council
ESRI	Environmental Systems Research Institute
EuroStat	European Office for Statistics
EU	European Union
GB	Great Britain
GIGO	garage in, garbage out
GIS	Geographical Information System
GOR	Government Office Region
GROS	General Register Office Scotland
HBoS	Halifax Bank of Scotland
HE	Higher Education
IMD	Indices of Multiple Deprivation
KS	Key Statistics
LA	Local Authority
Lab	Labour Party
LD	Liberal Democrats
LLTI	Limiting Long Term Illness

log	logarithm
LSE	London School of Economics & Political Science
MAUP	Modifiable Areal Unit Problem
MIMAS	Manchester Information and Associated Services (Manchester University)
NHS	National Health Service
NI	Northern Ireland
NISRA	Northern Ireland Statistics Research Agency
NUTS	Nomenclature of Units for Territorial Statistics
OA	Output Area
ODPM	Office of the Deputy Prime Minister
ONS	Office for National Statistics
OPCS	Office of Population Censuses and Surveys
OS	Ordnance Survey
PAM	Partitioning Around Medoids
PC	Personal Computer
PCA	Principal Components Analysis
PDF	Postdoctoral Fellowship
RAC	Royal Automobile Club
RSG	Research Support Group
S4C	Sianel Pedwar Cymru
SARs	Sample of Annonomised Records
SDF	Social Democratic Federation
SDRC	Social Disadvantage Research Centre
SIR	Standardised Illness Ratio
SOA	Super Output Area
SPSS	Statistical Package for the Social Sciences
TGI	Target Group Index
UK	United Kingdom of Great Britain and Northern Ireland
UN	United Nations
US	United States of America
v	Variable

Chapter One – Introduction: Project Outline, Aims and Objectives

1.1 The 2001 Census, the importance of place and area classification

The Office for National Statistics (ONS) is responsible for the collection and publication of statistics on a wide variety of topics, including information on the demographic, social and economic attributes of the UK population. The ONS use periodic and continuous surveys, registers of births, deaths, marriage, divorce, decennial censuses of population and a variety of other official registers (electoral, national health service). Together with their equivalents in Scotland and Northern Ireland, the ONS has collected, processed and published data from the 2001 Census of Population.

The census of the UK population is a means of counting people and recording their attributes and characteristics (Rees and Martin *et al.* 2002). The primary purpose of the census is to provide government with data about the population of the country on which to base funding decisions. However, the value of the census goes way beyond its primary purpose providing an essential source of data for public, private and academic research (Cook 2004). The 2001 Census has produced a very large and rich dataset for the 58,789,194 people, 24,479,439 households, 223,060 output areas, 10,553 Statistical Wards and 434 local authorities of the UK. The Census Area Statistics, for example, have delivered 190 tables containing about 6 thousand unique counts for output areas and all higher geographies (ONS/GROS/NISRA 2001).

The ONS together with the General Register Office Scotland (GROS) and the Northern Ireland Statistics Research Agency (NISRA) has produced Standard Area Statistics for the whole of the UK. Standard Area Statistics is the collective term for four data products produced from the 2001 Census for the whole United Kingdom: counts of persons and households (available for postcodes and all higher areas), Key Statistics (output areas and all higher areas), Census Area Statistics (output areas and all higher areas), Standard Tables (wards and all higher areas) (Rees and Martin *et al*. 2002).

Faced with this cornucopia of statistics the average census user will feel overwhelmed. In response a wide variety of simplified census outputs have been designed. One of the ways of dealing with this volume and complexity of information is to reduce it to much simpler terms, through the development of composite indicators (e.g. of deprivation) or through the creation of area classifications (Rees and Denham *et al.* 2002). An area classification is a bringing together of multiple pieces of data about areas to provide single, easy to understand generalised identifiers and descriptions of each area. An area classification would be the ideal way to simplify and describe the copious statistics produced from the 2001 Census (Rees and Denham *et al.* 2002). Multivariate based area classifications are a long established method of presenting the characteristics of residential areas in a simple and easy to understand way.

The basic concept that underpins area classifications is that people who live close to each other have a tendency to display similar characteristics and behaviours (Harris *et al.* 2005). This is what is known as 'spatial autocorrelation' (Cliff and Ord 1973), the basic premise being that similar phenomena tend to be found close to each other, or to put it simply *Geography Matters* (Ballas *et al.* 2005). The premise of 'spatial autocorrelation' enables the grouping of statistics about places or neighbourhoods to provide descriptions of the character of localities and people living within them.

The neighbourhood has re-emerged as a setting to examine social processes that influence social identity, cohesion and life-chances (Forrest and Kearns 2001). The focus on the neighbourhood has reinvigorated the notion that people are inherently linked to the locality in which they reside and consequently people shape the places in which they live (Champion *et al.* 1987; Harris 2003). However, people's perception of place is not uniform. Rather, a view of a particular place is an individual interpretation of the location's significance to them, influenced by each person's culture and experience (Altman and Low 1992; Canter 1977). Longley and Batty (1996) contend that *"The behaviour of individuals in space together contribute to the development of places over time and these place effects in turn condition subsequent spatial behaviour"* (p76). Place is not just fundamental to how people live their lives, but indicative of the way they live them (Weiss 1988).

People's lives are embedded in particular places, in which they were either, born, live or have lived in the past. These places could be as large as towns or cities, or as small as neighbourhoods or individual streets and houses. People identify with these places because they represent part of themselves; it is the people who live within a place that gives it an identity (Rose 1995). Places are human creations; any study of the social geography must appreciate the sense of identity that each place bestows upon the people who live there, and that certain characteristics of people can be established from the place in which they reside (Massey 1995; Eisenhauer *et al.* 2000). Each place is unique; studying data about a particular place provides a collective snapshot of the nature of that place and the people who live within it (Dorling and Thomas 2004).

Longley (2003) argues that *"We need to be better able to differentiate between locations, not just on account of their physical attributes but also by virtue of their identification with specific identities"* (p116). Data about neighbourhoods is essential to sustain social policies that are based around the concept of the neighbourhood (Martin 2004). The census provides much of the information that is necessary to make sense of the geography of the complex social patterns of the UK, revealing an unbiased picture of the social make-up of every place in the country. However, the cornucopia of information that is presented in the census needs to be greatly simplified to enable the patterns they display to be understood.

The start of the new millennium represents an ideal time to investigate the patterns of social groupings across the country. Changes in the social structure of modern society have been recognised by the Registrar General, who has redesigned the Registrar General's Index of Social Class for the 2001 Census for the first time since its introduction in 1911. These changes would undoubtedly be reflected in area classifications created from census data. Longley (2003) goes on to propose that *"The challenge to today's urban geography is to provide a nexus for interdisciplinary social science and create truly generalized representations of social structure"*(p116). This project will aim to meet Longley's challenge with the creation of a multivariate small area classification providing a description and visualisation of the social structure of the UK.

Openshaw and Wymer (1995) argue that a multivariate classification of small area data provides a simple and useful descriptive summary of the characteristics of the zones in a spatial system. Blake and Openshaw (1995) suggest that the classification of small area census data has, in the commercial world become a valued and trusted resource. However, such methodologies are still under utilized in modern geographic study (Longley 2005). By creating an equivalent free at the point of access classification, the benefits that have been discovered and enjoyed by the commercial sector to public and academic researchers.

It is impossible to understand the complexity of information that the census tells us about each area of the country without an attempt to summarise the information in the dataset. This thesis aims to provide an answer for those who do not have the time or skills to wade through endless census outputs. The goal of the thesis is to provide free of charge to the wider academic and research community a multivariate classification of areas at a fine level of aggregation. This unprecedented study will for the first time make available a classification for the whole UK. The new small area classifications that will be developed in this project will make innovative use of new census geographies, and will be the first such academic investigation carried out using the new census output areas. This thesis expands the notion of small area classifications from a black box, expensive business tool to a transparently created, free and easy to use, quality assured statistical product. In order to do this, the project will be developed with an open and published methodology, enabling replication and investigation into the quality of the outcomes.

This project is sponsored by an ESRC Collaborative Awards in Science and Engineering (CASE). These are awards for research students to carry out projects in the social sciences in collaboration with companies/business. They provide PhD students with the opportunity to gain experience of work outside an academic environment. The Office for National Statistics is the CASE partner for this study. This studentship can assist ONS in developing small area classifications in which the academic community can have confidence: the methods and assumptions used will be placed in the public domain through publication of results.

1.2 Aims and Objectives

The principal aim of this thesis is to *create a general purpose classification of UK Output Areas from the 2001 Census of population*. This will be complemented by a series of further aims:

2nd aim: *Compare existing classification methods and choose the most suitable.*

3rd aim: *Assure the quality of the classification produced.*

4th aim: *Show the value of the classification with examples of its use.*

5th aim: *Link the Output Area Classification to Ward and Local Authority Level Classifications created by the ONS, creating a Multi-Level Integrated Classification System of the UK.*

In order to achieve these aims; the following research objectives have been identified:

1) *To investigate the concept of clustering that underpins the main premise behind area classification, that objects and people that are in close proximity to each other are likely to share similar characteristics.*

2) *To investigate the development of area classification and geodemographics over time, the people who have been behind its development and the products that it has produced.*

3) *To investigate and review the methods and procedures involved in creating an area classification.*

4) *To use the knowledge that has been gained in objectives 1-3 to create a classification at a broad geographic scale (local authority districts) to gain an understanding of the difficulties and practicalities of creating an area classification.*

5) *To use the knowledge and experience from objectives 1-4 to create a classification of UK output areas (OAs).*

6) *To provide a detailed description of the methodology and the creation of the OA classification.*

7) *To describe, map and name the clusters produced by the OA Classification.*

8) *To quality assure the OA classification with a mixture of statistical techniques and surveys.*

9) *To show evidence of the value of the OA classification by using it to predict and account for trends and patterns seen in a number of current socio-demographic issues.*

10) *To link the OA classification to the higher level geography classifications created by the ONS, to investigate diversity within the classifications and to illustrate the importance of the choice of scale in geographic analysis.*

11) *To evaluate the success of the project and examine potential uses for the OA Classification and discuss potential for further work.*

The project to create a classification of UK output areas will form part of a larger ONS project to create a suite of area classifications covering the UK at different geographies and geographical scales. The ONS project will develop general purpose classifications of output areas, wards, local authorities and health authority areas. The Methodology Group of ONS will be responsible for the local authority, ward and health area classifications and this project will work in close liaison with staff within the ONS who are tasked to produce the area classifications. The output area level classification is the finest level of geography for which an area classification is being produced. The ONS developed classifications for local authorities, wards and health areas from the 1991 Census, but did not produce a classification for the finest geography for which census data was released, then being enumeration districts (EDs). Therefore the creation of an area classification at output area level represents a step forward for the ONS.

To achieve the principal aim, the project's second aim will be to compare existing classification methods and choose the most suitable. Among methods to be implemented and compared will be cluster analysis using hierarchical or k-means methods such as that used in the ONS 1991 Census local authority and ward classifications (Wallace, Charlton and Denham 1995; Wallace and Denham 1996) and the revised district classification (Bailey *et al.* 1999a, 1999b).

A third aim of the project will be to provide data and evidence on the quality of the classification created, to justify the publication of the OA classification as a 'National Statistic'. This quality assurance will be comprehensive and take many different forms, ensuring that the methodology is sound and that the classification that is produced is a satisfactory representation of the geo-social distribution of people throughout the UK. The classification must also be presented in a format which is appropriate for use and can be easily understood.

A fourth aim of the work will be to test the power of the general multivariate classifications to predict other "behaviours" compared with alternative determinants such as composite indicators or single variable classifications. The other "behaviours" might be census indicators not used in the classification such as: religion, migration or non-census indicators such as house prices, election results or crime statistics. Patterns displayed by the classification will also be tested to see how the classification is distributed by different geographies, such as: distribution by region, a comparison of cities, or how the north of the country differs from the south.

A fifth aim is to create a multiple scale classification system by linking the classifications in the ONS suite of area classifications together. This will exemplify the importance of scale in the analysis of areal data and enable the classifications to be used together as a combined product. The linking of the classifications will enable the examination of the diversity within areal units and classifications. By examining diversity within the different levels of classification it can be established if diversity occurs at different geographic scales in different types of area.

1.2.1 Methods

A transparent and reliable methodology must be developed in order to produce a classification that can be published as a 'National Statistic'. The steps in the classification exercise are as follows:

1. Review carefully the purpose of the classification and the demographic-social-economic-behaviour domains that should be covered.
2. Develop a suitable set of variables that cover those domains, exploring the degree of collinearity and selecting variables that are independent.
3. Decide on a method of indicator construction that treats chosen variables in a comparable way.
4. Assemble a database of indicators for the units at each spatial level.
5. Choose a general classification method (after review of the literature and assembly of suitable software).
6. Decide (in advance) on the characteristics desired in the classification (number of classes, degree of homogeneity within classes etc.).
7. Experiment with the classification methods, selecting a variety of options.
8. Prepare statistical and visual (graphs, maps) summaries of the classifications.
9. Label the classifications with descriptions of varying lengths.

Once the set of classifications has been chosen, methods will be developed to link postcodes to the classifications so that users can place their own observations from surveys or medical records into the classification or attach the classifications to a set of cases as in the GBProfiler system developed by Openshaw and Blake, described in Rees and Denham *et al.* (2002).

1.2.2 Outputs

The outputs from this project will be more numerous and varied than is expected from a PhD thesis. Not only will a series of papers outlining the classification methods and a comprehensive description of the classes be developed, the classification will also be published as an official 'National Statistic' by the Office for National Statistics.

A 'National Statistic' is a quality marker applied to certain of the United Kingdom's official statistics (ONS 2004a). The label 'National Statistics' ensures the quality of a statistical product, which is required to meet certain criteria. 'National Statistics' are obliged to be: fit for purpose, methodologically sound, politically independent and transparently produced. Data and information released under the 'National Statistics' banner supply an up-to-date, comprehensive and meaningful portrait of the UK's economy and society (ONS 2004a). To ensure that all 'National Statistics' meet the necessary criteria they are produced in accordance with the 'Framework for National Statistics' to ensure that they comply with the principles outlined in the 'National Statistics Code of Practice' (ONS 2000; ONS 2004a). Only products issued by the Office for National Statistics are designated 'National Statistics', though many are produced by other parts of the Government Statistical Service in collaboration with ONS.

The outputs from this project are a set of classifications which have been made available to users via both the ONS website www.statistics.gov.uk and the academic website www.census.ac.uk (maintained by the Census Dissemination Unit of the MIMAS service of Manchester Computing). Arrangements have also been made to host the classification with additional information and outputs on the University of Leeds website. A detailed outline of the publication of the classification can be found in § 9.3.

1.3 Thesis Structure

In order to achieve the research objectives set out in Section 1.2, the thesis is organised into nine chapters as outlined in Table 1.1. Each chapter relates to one or more of the stated research objectives.

Table 1.1: Thesis Outline

	Chapter	Objective
Chapter 2	Introducing Clustering, Area Classification and Geodemographics.	1&2
Chapter 3	Making a Classification System: a Guide to Methods and Procedures.	3
Chapter 4	A Classification of the UK's Local Authorities.	4
Chapter 5	A Classification of the 2001 Census Output Areas.	5,6&7
Chapter 6	Quality Assuring and Adding Value to the OA Classification.	8
Chapter 7	Testing the OA Classification: Accounting for Behaviours and Patterns.	9
Chapter 8	A Multi-scale Integrated Classification System: Investigating Diversity within Area Classifications.	10
Chapter 9	Conclusions: the Way Forward for a Newly Classified Nation.	11

Chapter 2 (Introducing Clustering, Area Classification and Geodemographics) introduces the concept of clustering and links this to area classification and geodemographics. The history and development of area classification from a philanthropic Victorian study, through the 'Chicago School' and factorial ecology into the modern geodemographics industry is summarised. Definitions of geodemographics within relevant literature are discussed, as are some of the many uses to which area classifications have been applied. There is a discussion of some of the limitations and criticisms that have been laid at the door of area classifications. The chapter concludes with a discussion of some of the current issues within geodemographics and area classification.

Chapter 3 (Making a Classification System: a Guide to Methods and Procedures) reviews the processes, procedures and methods involved in creating an area classification system, which can be summarised as the 'seven steps of cluster analysis' (Milligan 1996). The chapter reviews all of the procedures and decisions that are required to produce a classification in sequence; this is done in three sections. Firstly, there is a discussion of inputs covering, data sources, variable selection and data reduction. Secondly the processes section provides an overview of standardisation, weighting of variables and a variety of clustering procedures. Finally, the outputs section reviews the production of a classification structure, variable and cluster portraits, attaching photos, maps and postcodes to the classification (Harris 1999). A short conclusion then reflects on the methods and processes outlined in the chapter.

Chapter 4 (A Classification of the UK's Local Authorities) provides a detailed outline of the creation of the Local Authority Classification, which was developed to enable the author to experiment and gain experiences with clustering procedures. The chapter provides a sequential run through of the creation of the Local Authority Classification covering all the decisions and methods that were used in the creation of the classification, based on the three stages of area classification as outlined in Chapter 3. In the conclusion to the chapter reflections on the success of the classification are given. Lessons that have been learnt from the creation of the Local Authority Classification are commented upon to be taken forward into the creation of the Output Area Classification.

Chapter 5 (A Classification of the 2001 Census Output Areas) describes in detail the creation of the Output Area Classification, which is the major output and main aim of the study. With the implementation of lessons learnt from the Local Authority Classification and information gained investigations of output area level data. The chapter runs through the development and creation of the National Classification of Output Areas. Each element of the creation of the classification is described in detail, from the variable selection to the clustering methodology used, including the changes made to the methodology after the original techniques failed to cope with some of the extremes within the dataset. The chapter goes on to name, map and describe the clusters that have been produced.

Chapter 6 (Quality Assuring and Adding Value to the OA Classification) tests the quality and reliability of the Output Area Classification with a number of statistical tests and qualitative investigations of the classification. Statistical analysis of the classification includes analysis of the reduction of variability provided by the classification as well as sensitivity analysis and the examination of the change in within cluster and between cluster variability. An analysis of atypical areas (for their clusters) and the reasons behind them ensures that these are due to real world features rather than methodological problems. A groundtruthing exercise conducted around the country provides typical photographs of areas, to check that what areas look like usually matches how they are described and explained statistically. Further evidence of the reliability is given by an undergraduate field class exercise carried out in Bangor, North Wales. The chapter also contains the results of an innovative and unprecedented consultation exercise, where selected experts were asked to identify cluster groups for a selected area known to them and comment on the quality of the classification. The consultation exercise not only showed the quality of the classification, but suggestions from participants provided excellent ideas for this and future projects, as well as affirming there is a large number of people who will make great use of the classification.

Chapter 7 (Testing the OA Classification: Accounting for Behaviours and Patterns) makes use of the Output Area Classification for the first time by using a selection of case study examples. The chapter shows how the classification can be used to explain and account for patterns and processes within the selected examples. Examples explained include: an explanation behind the swing seen in the British General Election, accounting for the distinct geographic pattern displayed by the Welsh language, an examination into religious segregation in Northern Ireland, an analysis of the north south divide and examinations of deprivation and rurality.

Chapter 8 (A Multi-scale Integrated Classification System: Investigating Diversity within Area Classifications) outlines the creation of a multi-scale classification system by integrating the Output Area Classification with the ward and local authority classifications created by the ONS. The different geographic levels of classifications are used to investigate diversity within each other, showing how different types of clusters show different amounts of diversity within them at different scales. The lowest geographical level the Output Area Classification is investigated using household variables from the census. The difference in diversity between different types of area exemplifies the value of multi-scale system for socio-demographic analysis.

Chapter 9 (Conclusions: the Way Forward for a Newly Classified Nation) provides a conclusion to the study by summarising the findings of the research. The chapter reviews how well each of the aims of this study have been fulfilled and discusses the successful publication of the classification. It then moves on to discuss the many uses of the classification and the limitations of the research. The chapter then looks to the future with the discussion of several ideas for future research, including an ESRC postdoctoral fellowship, which has received funding to examine the changing residential patterns of the UK 1991-2001. Further proposals for the future of area classifications within geographical and social research are also outlined.

Chapter Two - Introducing Clustering, Area Classification and Geodemographics

2.1 Introduction

This chapter introduces the theories, principles and practices behind clustering, area classification and geodemographics. Section 2.2 introduces the principle of clustering and classifying using, theories concepts and real world examples. Section 2.3 introduces and defines the terms area classification and geodemographics. Section 2.4 charts the history and development of area classification over the last 100 years. Section 2.5 gives examples of some recent applications of area classifications. Section 2.6 introduces the administrative geography of the UK and the importance of its different scales. Section 2.7 investigates some of the criticisms that have been made against geodemographics and area classification. Section 2.8 introduces the concept of fuzzy classification, which is probably the biggest current and future issue within geodemographic discourse. Section 2.9 concludes and links the discussions of the previous sections to the aims of the project, extracting a research agenda.

2.2 Why Classify?

Area classifications provide a unique way of bringing together areal patterns from a range of variables, and identify similarities and dissimilarities between areas (Webber & Craig 1978). However, the idea of sorting things into categories based on similarities is not a new one; the basic premise of classification is fairly primitive. The nouns of the English language are little more than labels to describe classes into which objects can be placed (Everitt *et al.* 2001). In its widest sense, a scheme of classification represents a convenient technique for the organisation of a large dataset to enhance the efficiency of information recovery. Class labels describing arrangements of differences and similarities between objects provide a convenient summary of data (Everitt *et al.* 2001).

The human mind classifies objects into groups without conscious thought, simplifying information in this way helps us understand the world around us. No two things are ever identical, but by grouping similar things together our mind's understanding of objects is increased, the next time we see a similar object we know what it is and what to do with it having learnt from our experiences of a similar object (Pinker 2004). This can be seen as learning by similarity, a very useful tool and one which ensures we always eat soup with a spoon rather than a fork.

Classification is fundamental to most branches of science. For example, the periodic table in chemistry groups elements with similar properties (for example, magnesium, potassium and lithium) together to aid understanding. In biology species of animals are put into taxonomic classes based on their physical features e.g. mammals, birds, fish, reptiles etc. These groups are then broken down in a number of levels ending in each individual species being named. This process of animal taxonomy began with Aristotle around 350BC dividing animals into two groups vertebrates and invertebrates (Everitt *et al*. 2001).

As humans we can't help classifying to make things simple. It is an intrinsic way in which our brains work to make sense of the world by grouping similar things together. This was recognised as far back as the eighteenth century by Carolus Linnæus (1707 – 1778) who stated in his book Genera Plantarum:

> *"All the real knowledge which we possess depends on methods by which we distinguish the similar from the dissimilar. The greater number of natural distinctions this method comprehends the clearer becomes our idea of things. The more numerous the objects which employ our attention the more difficult it becomes to form such a method and the more necessary."* (Linnæus 1737)

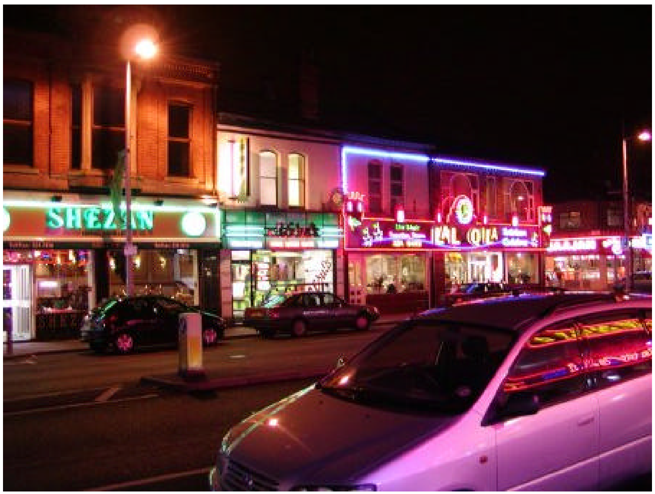
Classifying residential areas works in the same way. By grouping areas together into similar types our understanding of them can be greatly enhanced. The complexities of 223,060 individual and different census output areas is too much information for the human mind to process. However, by clustering these areas into a handful of groups which share similar properties, our understanding of the areas is greatly increased. The reduction in the amount of data makes it much easier for our brains to process the information; we can begin to see patterns in the distribution of the different types of areas, and infer what processes are taking place.

2.2.1 Clusters All Around You

Geographic clustering in many different forms can be seen all around in everyday life; you may or may not have noticed them, as some are easier to spot than others. Some examples of clustering make logical sense, while others need a little more thought to understand the reasoning behind them. In this section examples of such clusters are explained.

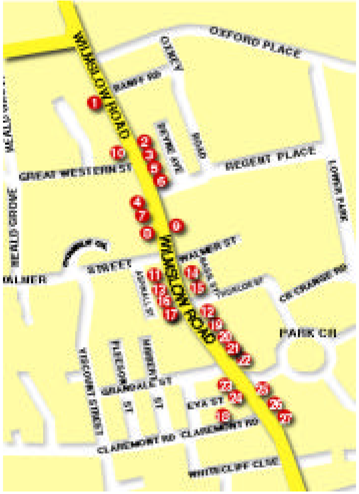
For the first example we take a trip to Manchester, Rusholme, Wilmslow Road to an area that has become known as *'The Curry Mile'*. Since the early 1970s thousands of immigrant families (mainly from Pakistan) settled in the area three miles south of the city centre. Twenty years ago the area consisted of just a few Asian businesses, but in the intervening period the area has seen extensive growth, 27 curry houses now occupy a one mile stretch of Wilmslow Road alone (Greenlees 2004), as shown in Figure 2.2. In all, over 150 Asian businesses including restaurants, takeaways, sweet houses, Asian grocers, kebab houses, sari shops, Asian music, video and book shops and 'golden jewellers' are squeezed into this one mile strip (Greenlees 2004). Many of the shops are adorned with large amounts of florescent lighting and the whole road is lit up at night creating a spectacular sight as shown in Figure 2.1.

Figure 2.1: The 'Curry Mile' by night



Source: http://alligevel.blogspot.com/2005/02/we-have-visual.html

Figure 2.2: Map of the Curry houses on Wilmslow Road



Source:http://www.restaurantsofmanchester.com/rusholme-map.htm

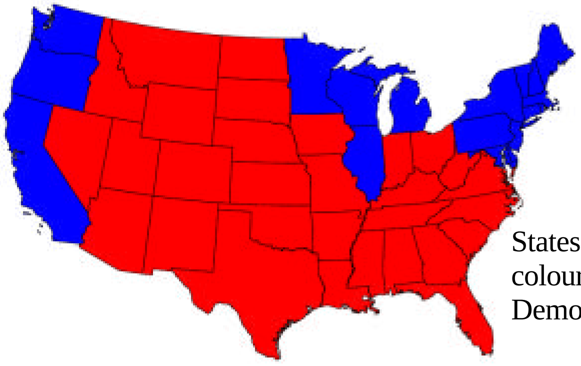
Why has this happened? These businesses started up because of the large immigrant population in the area, so initially there was a market for the first businesses. Immigrants often setup their own businesses when arriving in a new country as they find it hard to enter the job market and are often offered only the most menial of jobs; this problem is exacerbated for those with poor language skills (Robinson 1981). Setting up their own business not only provides work for the head of the household, but often for the extended family (Engstrom 1997). As the immigrant community established itself in the area new immigrants and extended families moved in, more businesses set up until there were so many it was dubbed *'The Curry Mile'*.

How do all these businesses survive with so much competition around them? There is serious competition between the businesses and prices are kept low because of the availability of similar product locally. However, the reason they all survive is that because the area has become so well known for Asian cuisine people are automatically drawn to the area. There is a wide variety of choice and the number of restaurants ensures that you are sure to get a table somewhere. The number of restaurants has made the area famous so the cluster works as a form of combined advertising for all of the businesses. Bus trips are run from neighbouring cities to visit the *Curry Mile* even though there are many Asian restaurants nearer their origin. It is a common challenge for students at the nearby Manchester University to try and visit every curry house and kebab shop during their period of study (Greenlees 2004).

By clustering together, the curry houses of Rusholme have managed to build up a reputation to the benefit of all. This works to the advantage of all the businesses despite the increased level of competition that it creates between them. The curry house industry benefits from the 'economics of agglomeration' (Fujita and Thisse 2002). This example shows how that similar things gain an advantage from their proximity to each other, socially, culturally and economically.

The second example is a very different form of clustering. The 2004 US Presidential election results were controversial and split the country down the middle. The Republican president George W. Bush held on to power by a small margin from the Democrat John Kerry. Not only was the result close, but it also shows a fascinating clustering of opinion across the United States. Each of the 50 states vote for their candidate of choice and the winning candidate in each state is given a number of Electoral College votes loosely based on the size of population of that state. Therefore we are able to investigate not only who was elected president, but who won each individual state. Figure 2.3 shows the distinct geography of the election result. The map gives the impression that the Republican vote dominates the country, since it covers a larger area than the Democratic vote. However, this is misleading as most of the Republican states have small populations.

Figure 2.3: The geography of the 2004 US presidential election results (48 contiguous states)



States voting Republican are coloured red; states voting Democrat are coloured blue

Source: http://www-personal.umich.edu/~mejn/election/

Figure 2.3 shows three distinct clusters: firstly, there is a cluster of Democrats on the west coast, next to this is the large central Republican cluster and finally there is a second cluster of Democrats in the north east. Figure 2.3 not only shows a distinct pattern to the election results but a distinct clustering of opinion across the country, so much so it can almost be seen as a schism in the union of the states. These patterns show how in general people with similar opinions live in close proximity to each other, suggesting that people prefer to be live amongst people with similar ideas and values to themselves.

The two examples shown are simple and clear examples of real world clustering. What they tell us is that clustering does occur naturally in the human world. This is an important point as we can only be confident that clusters created through statistical analysis and examination are representative of the world and not just a characteristic of a statistical manipulation, if there is evidence of clusters occurring naturally without statistical manipulation.

2.3 What are Area Classifications and Geodemographics?

Area classification is the classifying of areas into groups of similarity based on the characteristics of selected features within them; this could include anything from fish stocks to the risk of explosions from natural gas (Everitt *et al.* 2001). However, in the context that it is being used in this project, area classification will refer to the classification of areas into groups based on the socio-economic characteristics of their residents.

Geodemographics is a term that has grown in prevalence over the last 20 years, with the development an industry that produces small scale area classifications (usually at postcode level) for commercial purposes such as target marketing and business or service site selection. Geodemographics is understood to be the analysis of information about population location.

However, one can see geodemographics as a much broader geographic paradigm which has many applications. Geodemographics is not just a set of off the shelf consumer targeting products it is *"the analysis of people by where they live"* (Sleight 2004, p18). Sleight's definition of geodemographics needs little elaboration as it is simple and to the point. Another definition that is worth noting is *"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"* (Birkin and Clarke 1998 p.88). This definition identifies the blend of geography and demography that underpins geodemographics.

These definitions suggest that geodemographics is a wider discipline than the commercial classification systems with which the term has become associated. With the growth of the industry creating area classifications, geodemographics has become misrepresented to mean almost solely small scale area classifications. The reason this has happened is that it is in the interest of the commercial firms producing area classifications to have a succinct term that describes their product. The term is now so associated with commercial area classifications that its broader meaning is rarely used. It is unusual to see the term 'geodemographic classification' used accurately; classification is often dropped in favour of the more concise, but vaguer term geodemographics.

Geodemographics works on the principle that the place and population are inextricably linked. Knowing where somebody lives, can reveal a certain amount of information about that person. Geodemographics can be said to work to the proverb that 'Birds of a feather flock together', this is what gives geodemographics its strength. Information about the different characteristics such as age, ethnicity, education, employment and housing type is used to determine a picture of the type of people who live in an area. If similar people live in similar places then by knowing information about one person enables information about others in that locality to be inferred (Sleight 2004; Weiss 2000).

The notion that distance between locations has an important role in determining their similarity is not one that is limited to population geography. It is a concept that is fundamental to all geographical study. Tobler's first law of geography states that *"Everything is related to everything else, but near things are more related than those far apart"* (Tobler 1970). Geodemographics takes Tobler's law and gives it a twist, using his principle that two houses next to each other are likely to be fairly similar and contain people with comparable characteristics. It is straightforward to visualise how zones of similarity can be created within an individual town or city (indeed geodemographics has its roots in single city studies). Building on this geodemographics not only can group areas in the same locality, but can also group similar areas together which are not connected. There is no reason why an area of Bournemouth cannot share similar characteristics with an area of Inverness even though they are at opposite ends of the British Isles.

By adding geodemographics to Tobler's law we can define as the first law of geodemographics that *people who live in the same neighbourhood are more similar than those who live in a different neighbourhood, but they may be just as similar to people in another neighbourhood in a different place*. The term neighbourhood has been used to illustrate the point as enables conceptualisation of the area in which you live. The term could be replaced with any scale of geographic entity dependent on the study. It has been established that similar people gravitate towards each other; this creates neighbourhoods containing comparative homogeneity within them. It is important to note, however, that people within groupings are not identical by any means they simply share similar characteristics.

2.4 The Development of Area Classification

This section describes the development of social area classifications through a 100 year history. An attempt is made to cover briefly each of the major pieces of work in the development of social area classifications and later the geodemographics industry. Some researchers are interested in the summary profiling of areas, other researchers area interested in finding clusters of similar areas. These are two sides of the same coin.

2.4.1 Modest Beginnings

Charles Booth is widely seen as the father of Area Classification (Rothman 1989). Booth saw the need to use data about more than one thing to get a true impression of what an area is like. First published in 1889 Booth's work on *the Life and Labour of the People of London* contains detailed maps of every street in London, placing every house into one of seven classes (as shown in the key to Figure 2.4). The maps he created can be found in the Charles Booth Online Archive housed at the London School of Economics (LSE 2005). Reprints of Booth's work have been issued over the years such as *Charles Booth's London* in 1969.

Booth set out to prove that poverty in London was not as widespread as had been reported. This was in response to the Social Democratic Federation (SDF) who carried out a sample survey of the people of London and concluded that approximately 25% of its residents were living below the poverty line (Hyndman 1911). Booth, who was essentially conservative in his views was riled by the publication of such a high figure, and set out on his survey with the desire to prove the Socialists wrong (Norman-Butler 1971). Booth actually found that the extent of London's poverty was greater than anyone had thought. His study concluded that 30.7% of London's population lived below the poverty line (Simey and Simey 1960; Pfautz 1967).

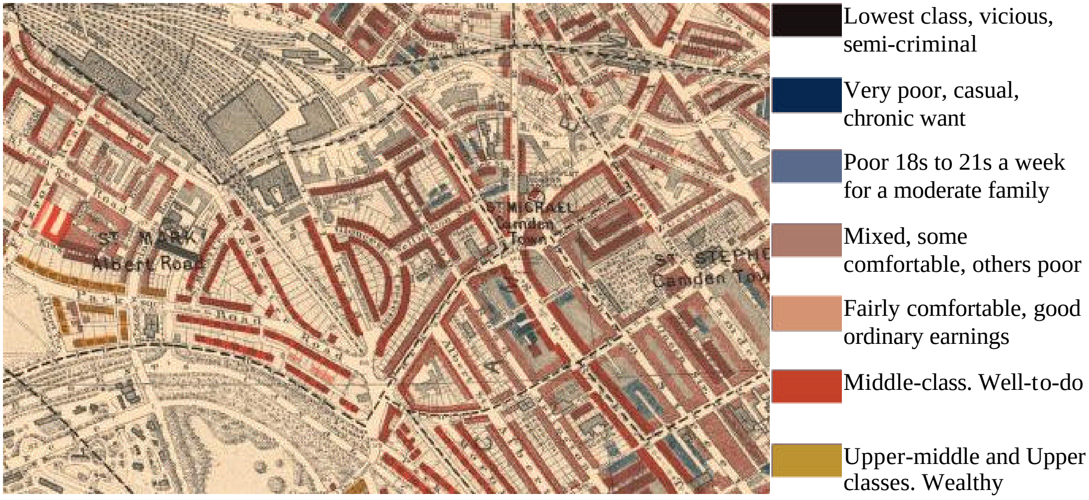
Booth did not undertake his study alone. He employed a team of trusted researchers including his cousin and his wife to help him with the mammoth task of surveying the whole of London. Along with what his team found Booth made extensive use of data gathered by the School Board visitors who visited every house containing children of school age. Every house in the capital was visited and notes were made on the conditions within every household. Below is an extract from one of Booth's notebooks describing the conditions in 34 Carver Street which falls into the 'mixed, some comfortable, others poor' class.

> *"No. 34 is occupied by the widow of a boatman. He committed suicide and left her with eleven children. Some have died, and she has five here now, two of whom go to work, and three to school. She makes sailor jackets, but is nearly blind. Struggles hard for her children. There are also living in this house, in one room, Coleman and his wife, and two children. Coleman was a porter but does nothing, preferring to smoke his pipe. His wife takes in washing and keeps him. In another room there lives Brough, a maker of dolls,*

> *working for his father who keeps a shop in Drumlow Road. He has a wife and two children. A third room is occupied by Owen, a labourer, often out of work, with wife and three children. They are nearly starving. The children are always ill."* (Extract from 'the Life and Labour of the People of London' reproduced in Charles Booth's London 1969)

The information in Booth's note books along with information from school board visitors were used to establish the general socio-economic conditions in which the residents lived. This information was then used to make a judgement as to which group each street should be assigned. A colour to represent each group was then shaded on to a base map of London to give a graphical indication of the general socio-economic status of the people living in each street. Figure 2.4 shows a section of Booth's 'Descriptive Map of London Poverty'.

Figure 2.4: An extract of Camden from Charles Booth's poverty map of London


ST MARK
Albert Road
ST MICHAEL
Camden Town
ST STEPHE
Camden Tow
Lowest class, vicious, semi-criminal
Very poor, casual, chronic want
Poor 18s to 21s a week for a moderate family
Mixed, some comfortable, others poor
Fairly comfortable, good ordinary earnings
Middle-class. Well-to-do
Upper-middle and Upper classes. Wealthy

© London School of Economics & Political Science. Source: LSE (2005)

Creating the world's first social area classification, Booth recognised that, although there are differences between houses in close proximity it was better to generalise his maps, ignoring minor differences to better illustrate the location of social classes within the city (Harris *et al.* 2005). Booth's recognition of the importance of place in the poverty of London was an early recognition as to the importance of geography in understanding how society functions.

The first form of official socio-economic classification (not based on area) was the Registrar General's social class groups based on occupation and employment variables of the male head of household, which were introduced in reports from the 1911 census. The social classes were used to investigate socio-economic differences in mortality (Jackson 1998). The classes are summarised below:

(I) Professional etc occupations
(II) Managerial and Technical occupations
(III) Skilled occupations
    (N) non-manual
    (M) manual
(IV) Partly-skilled occupations
(V) Unskilled occupations

These classes were used virtually unchanged in every census up to and including 1991. The 2001 Census saw the introduction of a new classification in response to changing household structures. However, this was not area classification and despite Booth's survey it was to be a long time until anyone attempted anything similar.

The next stage in the development of social area classification is dominated by one city, not London this time but Chicago. A group of urban sociologists at the University of Chicago, who became known as the 'Chicago School', developed a number of representations of the social structure of cities based initially on Chicago, and then applied to other American cities (Robson 1971). The school was founded by Robert Ezra Park who did much of the school's early work (Park and Burgess 1925). Burgess developed a concentric ring model of functional areas of the city, which consisted of five rings moving from zone one representing the centre of the city, through zone two 'Transition', zone three 'Workingmen's Homes', zone four 'Residential Zone' to zone five 'Commuters Zone'. The school went on to develop various models, notably Hoyt's Sector model, which combines concentric rings with cross cutting sectors defining areas of differing land use (Carter 1995). The Harris and Ullman Multiple Nuclei Model incorporates multiple centres within urban environments (Robson 1971).

Further work in the US became possible with the publication of data for census tracts (small scale US census areas containing between 2,500 – 8,000 people). This work focused on the social areas of two cities on the west coast Los Angeles and San Francisco. Shevky and Williams (1949) produced a detailed volume of work on 'the Social Areas of Los Angeles' with excellent maps and statistics based on data from the 1940 US census. Shevky and Bell (1955) did similar work on San Francisco. These were two important pieces of work in the development of social area classifications as for the first time they used solely statistical methods to classify areas in terms of their social composition. They are based on the theory of social stratification, the life cycle concept and ethnic segregation (Timms 1971).

By the 1960 US Census, census tract data were available for 180 cities/areas across the US. This enabled classifications to be developed for an increasing number of cities based on a larger selection of variables (Batey and Brown 1995). This led on to work such as that by Rees (1979), who examined residential patterns across twelve American cities chosen from groups in a classification of metropolitan areas. Along with work such as Rees in Berry and Horton (1970), this kind of work can be seen as empirical analysis, which extracts underlying patterns of residential structure.

Despite the lead set by Booth no work of note was done on social area classifications in Britain until the 1960s, when the increasing availability of small scale census data and a paradigm shift within geography saw an increase in the use of quantitative methods (Batey and Brown 1995). In *British Towns: A Statistical Study of their Social and Economic Differences* Moser and Scott (1961) conducted one of the first comparative studies of the socio-economic variations across Great Britain. They grouped 157 British towns and cities into 14 groups, themselves arranged in three types, with London left unclustered, being unlike other cities in Britain. They used factor analysis to measure common dimensions of variables across a wide set of socio-economic variables. They produced 4 factors: Social class, Population change 1931–51, Population change 1951–8, and Overcrowding. This enabled them to make a judgement as to which towns shared similarities, based on just 4 components rather than their original 57 variables. By graphing the correlation values for each town against each other for each of the four components they were able to estimate which towns should be grouped together (Moser and Scott 1961). This and work by Gittus (1964) in Merseyside and south-east Lancashire saw the return of area classification to Britain.

The 1966 sample census saw the first release of Small Area Statistics (SAS) in machine readable from in Britain. It was the public sector who picked up the classification batten and started to run with it. Several local authorities became interested in social area classification as a way of looking for social divisions and areas in need of investment. Liverpool City Council (1969) carried out what they termed a 'Social Malaise' study, which they used for the allocation of social services. As described in Kelly (1969 and 1971) the Greater London Council created a classification of the (then) 32 London boroughs.

The origins of the modern geodemographics industry can be traced back to the work of Webber and Craig (1976 and 1978). The Office of Population Censuses and Surveys (OPCS), where both Webber and Craig worked, commissioned three national classifications. This was a major step forward as it enabled comparisons between places at small scales. It had previously been difficult to compare across the country as all the local studies that had been carried out used different methodologies. At the smallest scale every ward and parish in Great Britain with a population of 50 or more was classified into one of 36 clusters based on 40 variables from the 1971 Census (Webber 1977). The key variables used included unemployment, students, two car households, industry sector, social class, age, migration, tenure, overseas immigration, overcrowding and household amenities.

At this point the commercial sector started to see the potential benefits of area classification and the British Market Research Bureau (BMRB) took a keen interest in the OPCS classifications, which they used to examine variations in consumer patterns (Baker 1997). BMRB re-structured and re-named Webber's classification *A Classification of Residential Neighbourhoods* (ACORN) and launched it at the Market Research Society's 1979 conference, as a marketer's dream (Batey and Brown 1995). In the same year Webber left OPCS to work at CACI (a company which provides marketing solutions and information systems for private clients). Webber continued to develop versions of the re-named ACORN and the modern geodemographics industry was born (Baker 1997).

2.4.2 Contemporary Geodemographics and the Rise of the Commercial Segmentation System (Post 1981 Census)

CACI through ACORN dominated the geodemographics market at its conception, but as interest in the system grew, competitors emerged and many CACI staff were lured to other companies to create new systems (Sleight 2004). The release of data from the 1981 Census was the ignition for the development of geodemographics into a major new industry and by the middle of the decade four main systems were competing for dominance, ACORN, PiN, Mosaic and Super Profiles.

The use of area classifications based on census data has developed rapidly especially within the private sector. Geo-demographic companies who are licensed users of census data create value-added classifications for a range of marketing applications, targeting either specific areas or specific types of people (Wallace *et al.* 1995). Information about types of neighbourhood can then be used to classify areas into types, joining areas with similar characteristic across the country. For instance the residents of Hyde Park (inner city Leeds) and Rusholme (inner city Manchester) live in similar neighbourhoods made up of terraced houses, having a large student population in two of Britain's biggest cities. By grouping areas such as these in a common class it is possible to build up a picture of similar areas across the country.

Commercial companies added new sources of data to geodemographic classifications. Webber (1977) solely used census data in the creation of a classification. However, the development of commercial segmentation systems has taken geodemographics away from being a totally census based exercise. Administrative data are often added to the census data to provide greater context. Datasets used include: the electoral register (age classification of names), vehicle registration data (quality/expense of cars as well as quantity), county court judgements and credit references (debtors' data) many of which are held by some of the firms that create the classifications (Harris *et al.* 2005). Another form of data that has become integrated into geodemographic classifications is "lifestyle" data. This is based on surveys carried out by commercial firms, analysis of commercial data, sales and warranty records (Sleight 2004). Millions of records have been accumulated over time, which can be integrated into classification systems. The main benefit of adding these forms of data to classification systems is that they contain some information that is not covered in the census, especially in terms of affluence and income. (Harris *et al.* 2005). Adding new forms of data to classification systems seems to have benefits in terms of increased amounts of and different data. However, the data are not comprehensive (lacking the 100% coverage of census data) and are biased towards non-poor consumers (Vickers *et al.* 2005). The data are also not publicly verified or available, so no kind of quality assurance of these data can be made.

This information can be used for a variety of purposes such as targeting a specific market for mail shot, getting an accurate representation of society in a stratified sample for opinion polls or the location of services and facilities. Throughout the 1980s and 1990s there were a handful of competing systems in the UK, with development of further commercial systems hindered by the licence fees payable for the use of census data. However, the market was to expand after the 2001 Census when the licence fee for the use of census data was removed, through the Census Access Project, a partnership between the census agencies and their main customers, including the ESRC (Sleight 2004).

The field of area classification was now dominated by the private sector and little work was visible beyond the commercial classifications. Social area classification had almost come full circle, from its origins as a survey done by a Victorian philanthropist, to Chicago academics, through the OPCS and on to the private sector. Although academics had seemed to have forgotten about area classifications one had not. Stan Openshaw developed GB Profiles using 1991 Census data based on the geography of the 1991 Census enumeration districts (EDs) the smallest British census geography with an average size of 200 households and 450 people. The system developed by Openshaw is available online via the Centre for Computational Geography, School of Geography, University of Leeds (Blake and Openshaw 1995; Rees and Denham *et al.* 2002). See CCG (2005) for more details. Geodemographic classifications are becoming ever more sophisticated and smaller in scale, with a movement towards household classifications (Webber and Farr 2001).

Following the removal of the large licence fee (£250,000 approx) from commercial firms to use census data following the 2001 Census, opportunities for smaller firms to create census based geodemographic systems were opened up. Coupled with the increased power and reduced cost of computing, the investment needed to create a geodemographic classification system had reduced significantly between the release of data from the 1991 and 2001 censuses. To create a commercial classification from 1991 census data would have require a large well backed company and state of the art computing. In contrast, to make a classification system from 2001 Census data requires as little as one person with the relevant skills and little more computing power than the average home PC. This was an opportunity that several people saw to get into the lucrative geodemographics market with little investment, unsurprisingly over the last few years the number of British/UK geodemographic systems available has more than trebled. Table 2.1 lists the Britain/UK based geodemographic classifications currently available. Table 2.1 is taken from Sleight's 2004 book *Targeting Customers*, which provides an excellent account of all the details of the commercial geodemographics sector in the UK. The market has grown from four systems before the release of 2001 Census to thirteen, with the likelihood being that further systems will be developed.

Table 2.1: List of British/UK geodemographic systems currently available.

Company/Organisation	Classification
The Old Boys	
CACI	ACORN
Claritas	PRIZM
EuroDirect	CAMEO
Experían	MOSAIC
The New Guys	
Acxiom	Personicx Geo
AFD Software	Censation
Allegran	Gnuggets
Beacon Dodsworth	P² People and Places
Business Geographics	Locale
The Clockworks/TRAC	SONAR
GeoBusiness	Locale
ISL	RESIDATA Lifetypes
Streetwise Analytics	Likewise

Adapted from Sleight 2004 p49

Not only did the release of the 2001 Census prove to be a catalyst for new commercial systems, but interest in area classifications has also seen a resurgence in public and academic research. The Office of the Deputy Prime Minister (ODPM) produced an index system called 'Indices of Multiple Deprivation 2004' (IMD) (replacing earlier versions published in 1998 and 2000) to assess deprivation in the England. It works to the theory that deprivation, in fact, different components of deprivation, vary over space. It combines a large number of indicators into 7 domains (Income, Employment, Health Deprivation and Disability, Education, Skills and Training, Housing and Geographical Access to Services) that are combined to produce a single measure of deprivation at both ward and local authority level (Noble *et al.* 2004). The IMD uses government administrative data such as various measures of benefits, health etc. This information can then be used to distribute funds and resources for a variety of regeneration projects (Noble *et al.* 2004).

All the classification systems described so far are based on "people in their homes". However, there are examples of other forms of classifications, Debenham (2003) classified "people in their workplaces". 'Supply side' and change variables were used in the classification to add characteristics of the labour market and their inclination to change over time to more commonly used social indicators (Debenham 2002; Debenham *et al.* 2003).

2.5 Some Current Uses of Area Classifications

The call for area classifications whether general or more specific evolved from a need for a straightforward and robust indicator of socio-economic information, contrasting the similarities and differences between areas (Wallace *et al.* 1995). Geodemographic classifications are being used for an increasingly wide variety of applications. Classifications are being made by commercial, governmental and academic institutions for general all-purpose use and more specific applications.

Geodemographic classifications are used heavily in the marketing industry. They are used in the planning stages of a project usually at a relatively broad geographic level. They are also used in data profiling to code geographically sparse data to an area typology to enable further analysis. Consumer profiling by geographical area can help to establish how a product may sell in a certain area. Identification of product use across geographical areas can help to establish who is buying a certain product. Classifications are often used as a method of stratified sampling for opinion polls that are used to gauge the views of the nation most notably before a general election. Many academics regularly use classifications in their research; Rees *et al.* (1996) employed the ONS classification of districts to compare rates of migration across the UK.

The Department of Health saw a classification of local health authorities as a useful tool to be available to the National Health Service (NHS) in terms of planning and service coverage. In 1996 the Office for Population Censuses and Surveys (OPCS) produced a classification system based on the geography of local health authorities. This was updated in 1999 when the geography of local health authorities was changed (Wallace *et al.* 1995). The uses include location and provision of services, such as schools, hospitals, emergency services or refuse collection. At a smaller level classifications are used to manage, allocate resources and monitor performance and also to enable precision within targeting resources, allowing reorganisation and reallocation to take place. Classifications are used to support bids for funding or resources to a higher level of government (Wallace *et al.* 1995).

Geodemographic classifications were used with the British Crime Survey to assess if different types of area suffer from different rates of crime. It was shown that conclusions could be drawn from classification systems as to the likely extent of crime in the area (Home Office 2005). Home Office (1997) showed that the areas of highest risk of fire in the home were ACORN group 17, multi-ethnic, low-income area, followed by ACORN group 16, areas of predominantly council estates associated with the worst hardship (ODPM 2003).

Classification systems have been used to monitor the participation of groups of different backgrounds in higher education. This has become increasingly significant in recent years as it has been the policy of the current Labour government under Prime Minister Tony Blair to increase higher education participation for people from working class backgrounds. At one point a target of 50% participation in higher education was set. Universities in England are paid extra for taking students from disadvantaged backgrounds. The Super Profiles system was used as a means of assessing which areas were most deprived; the universities gain extra revenue for taking students from those areas. Although the system worked it was modified after some universities were accused of 'postcode chasing' by favouring students from areas for which they received extra money in preference to equally gifted students from other areas (Eason 2002).

2.6 The Geographical Building Blocks of the Census

The classifications that are to be created in this project will be based on census geography. It is therefore imperative that before using the data and creating any classifications a good knowledge of how the census geography works and what each level represents is needed (Rees and Martin 2002). The geography of the census is far from simple. The census is simultaneously administered by three different agencies: the Office for National Statistics (ONS) is responsible for the undertaking of the census in England and Wales; the General Register office for Scotland (GROS) is responsible for the undertaking of the census in Scotland and the Northern Ireland Statistics and Research Agency (NISRA) is responsible for the undertaking of the census in Northern Ireland (Rees and Martin *et al.* 2002).

The data from the 2001 Census could have been aggregated to any level of spatial unit. However, the two base geographies are Enumeration Districts (EDs) and Output Areas (OAs). EDs were used for the purposes of data collection (Martin 2002a). OAs were employed for data publication they are being used across the UK for the first time for the 2001 Census. In 1991 EDs were used for both data collection and output (Rees and Martin 2002). However, in 2001 they were used primarily in the data collection process: the shape and size of the EDs were designed to form a suitable workload for each census enumerator (Martin 2002b). A new, smaller unit known as Output Area was introduced in the 1991 census in Scotland, and the 2001 Census saw the introduction of OAs to the whole of the UK (Exeter *et al.* 2005). OAs are smaller than EDs and so allow for a finer resolution of data analysis. They are built up of clusters of unit postcodes, but they also fit within administrative boundaries down to ward and parish level. OAs are designed to have approximately similar populations and be as socially homogenous as possible (Martin 2002b). Where possible, urban/rural mixes have been avoided. Ideal OAs consist entirely of urban postcodes or entirely of rural postcodes. OAs are discussed further in Chapter 5.

2.6.1 Geographical Hierarchies

Regional hierarchies have for a long time been a central theme within geography (Haggett *et al.* 1977). The surface of the Earth can be viewed as a hierarchy of areal units: Continents contain countries, Countries contain regions, Regions contain towns and cities, Towns and cities contain Streets and Streets contain individual buildings. A hierarchical provision of information enables a researcher to focus in on a small area, whilst enabling a comparison between the area of study and similar or dissimilar regions at the same scale. Comparisons with data at higher scales are also possible which enables the attributes of a sub-region to be compared with regional and national averages/trends (Haggett *et al.* 1977).

The regional hierarchies within the UK enable a researcher to focus in on different parts of the country at different scales, this can be extremely useful when making a comparison of information about non-connected regions. In conjunction with the simplification of data through classification, a hierarchical system of geography can significantly aid statistical investigation and comparison. Classifications can be built for one or many of these hierarchies.

The geography of the UK is not as simple as a single hierarchical system of geography for which all statistics are reported. There are several different geographic hierarchies within the United Kingdom including (administrative, health, electoral, postal). Data from the census may be used at all the different levels of each of the different hierarchies, so it is important to understand the geography of each of the hierarchies and what they represent. The structure of the hierarchies is also different in the constituent countries of the UK. The boundaries of the different systems are not aligned with each other and are subject to frequent change. In 2002 the boundaries of 1,549 electoral ward and divisions were changed in the UK, as many as the rest of the Europe combined (ONS 2003a). Other geographies within the UK are also subject to constant revision especially postcodes. The inconsistent geography of the UK is a major problem when trying to produce and compare meaningful statistics (ONS 2003a).

2.6.2 Administrative Geography of the United Kingdom

Administrative geography is the hierarchy of areas relating to national and local government in the UK. This hierarchy is far from simple. There several layers and the structure is different in each constituent country of the UK. Figure 2.5 shows how the UK consists of four constituent countries, three of which make up Great Britain. Unlike the other countries of the UK, England does not have its own devolved parliament and is thus entirely subject to the administration of the UK Government in Westminster. Despite appearing on maps and areas for reporting statistics metropolitan counties and Government Office Regions (GORs) have little administrative power. Electoral wards/divisions are 'building blocks' from which higher units are constituted they are used for reporting statistics but they are controlled at a higher level (ONS 2003a).

Electoral wards/divisions are the spatial units for which local government councillors are elected. Wards cover the whole of the UK which makes them ideal base unit for collating and presenting statistical data. They are also used to construct other geographies e.g. parliamentary constituencies, health authorities and Nomenclature of Units for Territorial Statistics (NUTS), a European Union hierarchical classification of administrative areas used for statistical purposes (ONS 2003a).

Parishes are a historical anomaly of geography and relate back to when the country was more influenced by the church, originally representing areas of both civil and ecclesiastical administration. Parishes have their own council and used to be significant local government areas, but now have a limited or no function. However, they are isolated from England's geographic structure as, unlike electoral wards, they are not found across the whole of the country (ONS 2003a). They can be affected by boundary changes to the local authority in which they fall, but not by those to wards. At the start of the year 2003 there were 10,373 parishes in England.

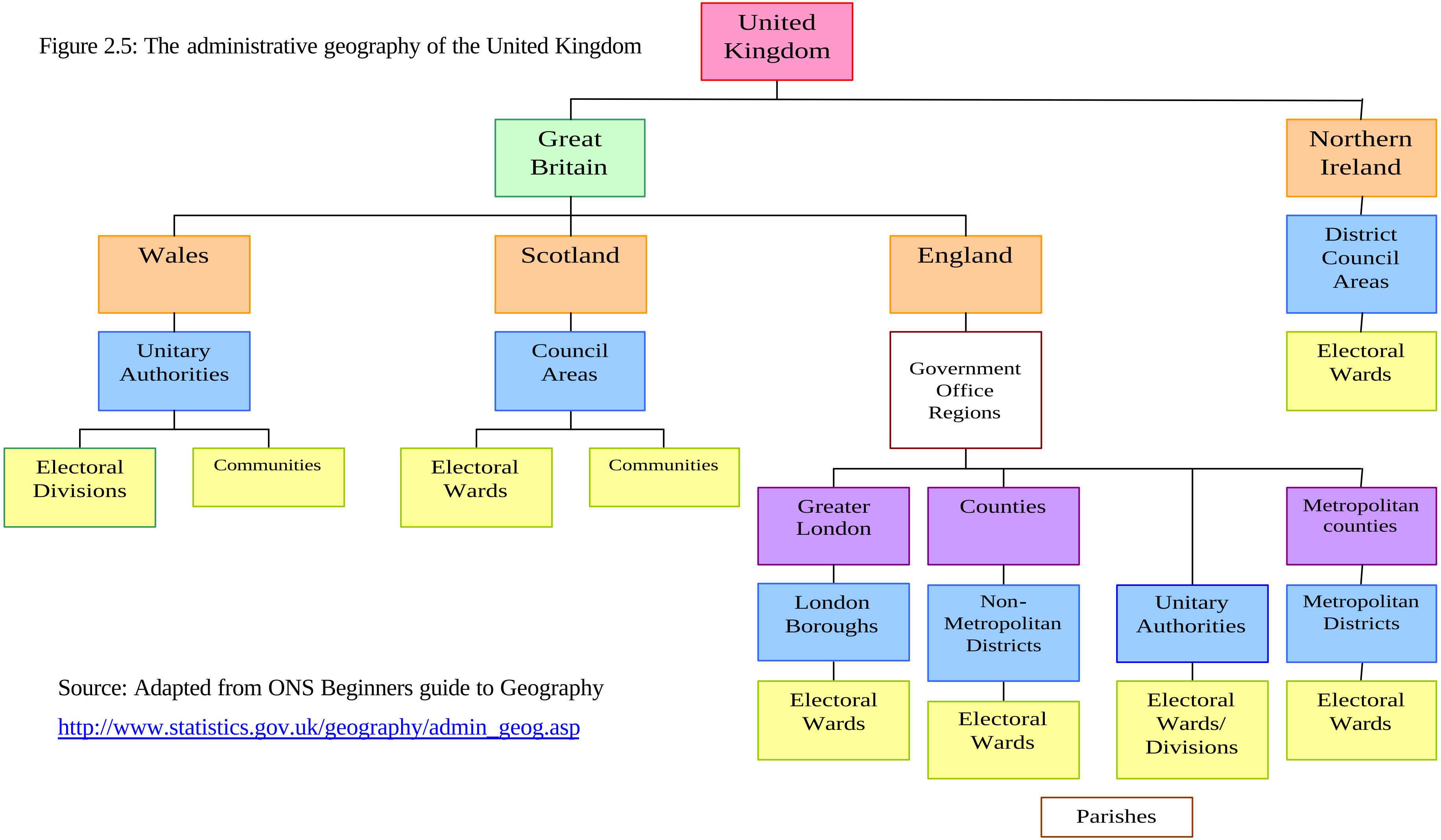
Scotland is subject to the administration of two parliaments: the UK Parliament in Westminster and the Scottish Parliament in Edinburgh. In 1996 the existing structure of 9 regions and 53 districts was abolished it was replaced by dividing the country into 32 units known as council areas (ONS 2003a). Council areas are built from electoral wards and are also divided into communities. Communities replaced parishes in Scotland, but did not stick to their geography they are a disaggregation of Council Areas created to represent the view of the local community, but are largely without any real power. Work has been done to harmonise the geography of statistical areas in Scotland with the creation of 'Consistent Areas Through Time' CATTs (Exeter *et al.* 2005)

The hierarchy of administrative geography in Wales is similar to that of Scotland. Communities cover the whole of Wales and fit into unitary authorities. As in Scotland they are the equivalent of parishes in England, they have similar powers although some act as town councils. As they nest into council areas they have greater potential as a statistical unit than the English parish. There are 867 communities in Wales; communities are not aligned to the old parish boundaries. However, parishes were abolished when communities were created in 1974 (ONS 2003a).

Northern Ireland is subdivided into 26 district council areas, which in turn are divided into electoral wards. The six counties (Antrim, Armagh, Down, Fermanagh, Londonderry and Tyrone) are still referred to but do not constitute a level of administration (ONS 2003a).

By understanding the hierarchy of geographical units of the UK, a researcher can appreciate what the complex array of statistics they are faced with represents.

Figure 2.5: The administrative geography of the United Kingdom



Source: Adapted from ONS Beginners guide to Geography

http://www.statistics.gov.uk/geography/admin_geog.asp

2.7 Criticisms and Limitations: Lies, Damn Lies and Geodemographics

Geodemographics has its advocates, but it also has its critics. The best proof that geodemographics works is its continued use and development; users like its simplicity and its applicability. However, many criticisms have been thrown at geodemographics. Some criticisms are general data or scale issues that could be levelled at any form of geo-statistical analysis and others are more targeted at the specifics of area classification.

2.7.1 Ecological Fallacy

The scale and areal extent at which a classification system is produced is very important to allow it to be used appropriately and accurately. Changes in, or the inappropriate use of the wrong scale or areal extent could produce significantly different results. The ecological fallacy arises when there is an error in the interpretation of statistical data. The fallacy occurs when information about individuals are based entirely on statistics for a group to which the individual belongs, an assumption is made that all members of a group display the characteristics of the whole group (Robinson 1950). The fallacy can also be seen when analysing different scales of aggregate data, if statistics relating to an aggregated areal unit are incorrectly assumed to represent an individual or a smaller unit within the original area (Tranmer and Steel 1998).

All forms of geographical grouping suffers from the effect of the ecological fallacy. Table 2.2 shows values from the 2001 Census for limiting long term illness (LLTI) for the same place at different scales. The first three levels (GOR, County and LA) the values of the areas are similar ranging from 19.5% to 18.0%. However, at the ward level the value is 10.5%, just over half of the value at the higher geographies. The value for the OA (4.8%) is half of that of the ward and a quarter of the higher geographies.

Table 2.2: An example of ecological fallacy using 2001 Census data

Geographical level	Name/code	percentage of people with LLTI
Government Office Region	Yorkshire and the Humber	19.5%
County	West Yorkshire	18.8%
Local Authority	Leeds	18.0%
Ward	Headingley	10.5%
Output Area	00DAFN0007	4.8%

If the percentage of people with LLTI were needed for Headingley and the value for the GOR, County and LA level were used; the value given would be unrepresentative of the area of interest. An ecological fallacy would have taken place as data at a large scale would have been assumed to be representative of every smaller area within it.

2.7.2 The Modifiable Areal Unit Problem

A dataset can appear to have significantly different values depending upon how and where the data are aggregated; this is called the Modifiable Areal Unit Problem (MAUP) (Openshaw and Taylor 1991; Monmonier 1996). The MAUP is a fundamental geographic problem that is endemic to all studies of spatially aggregated data (Wrigley 1995). Sensitivity to MAUP is unpredictable and further conclusions cannot be made, as the severity of the problem appears to be specific to each dataset (Openshaw 1984a).

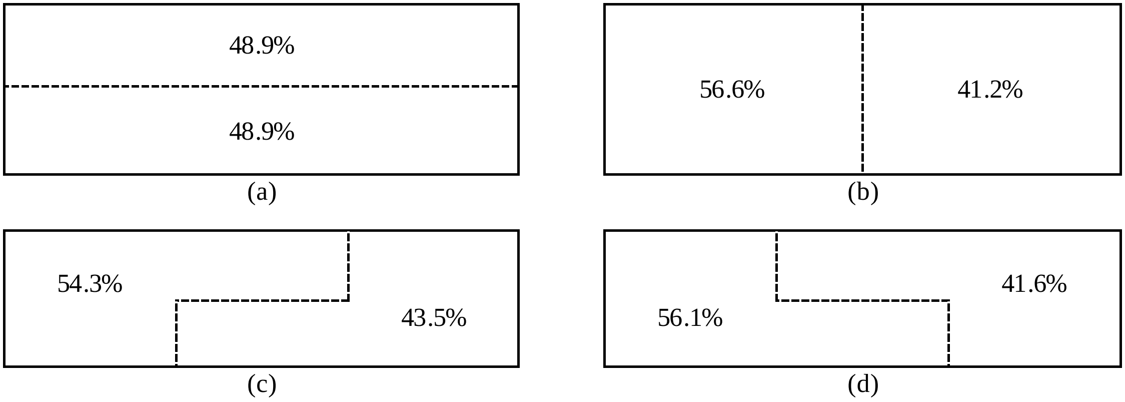
It is possible to produce significantly different values by choosing a variety of shapes and sizes of unit area on which to base a study. Therefore the results of studies based on modifiable units will depend on the units used (Openshaw 1984b). Figures 2.6 and 2.7 show an example of the MAUP. Figure 2.6 represents a grid of 24 fictitious census areas. The numbers inside the boxes represent the percentage of males who live in each OA. For purposes of illustration we can assume the population of each area is the same.

Figure 2.6: Sample dataset to illustrate the effect of aggregation on areal units

62.4%	58.7%	54.3%	23.7%	32.1%	50.1%
47.5%	50.9%	65.2%	72.0%	52.6%	17.5%
44.2%	60.5%	73.1%	55.1%	59.9%	11.6%
54.9%	57.4%	50.4%	58.5%	23.5%	37.6%

Figure 2.7 (a – l) illustrate the values that are created if the values in Figure 2.6 are aggregated in different ways, the numbers can be made to look significantly different by splitting the grid in different places. Split into two equal areas horizontally (Figure 2.7 a) and both areas have the same value. When split into two equal areas vertically (Figure 2.7 b) there is a 27% difference in value between the two areas. If the region is split diagonally in two different ways, producing four areas of the same size and shape as in (Figure 2.7 c & d), all values produced are different from each other and the previous examples.

Figure 2.7 a – d: An illustration of the effect of data aggregation of socio-economic data



It is clear that by splitting the grid in different places it is possible to make the value of the two areas look both uniform and irregular. By splitting the grid into different numbers of areas, of varying shapes and sizes further manipulations of the population of the area can be made. The percentage of males in whole area (Figure 2.7e) is 48.9%. Many more different values can be produced by aggregating the data into different sized and shaped areas illustrated in Figures 2.7f-l.

Figure 2.7 e - l: An illustration of the effect of data aggregation of socio-economic data

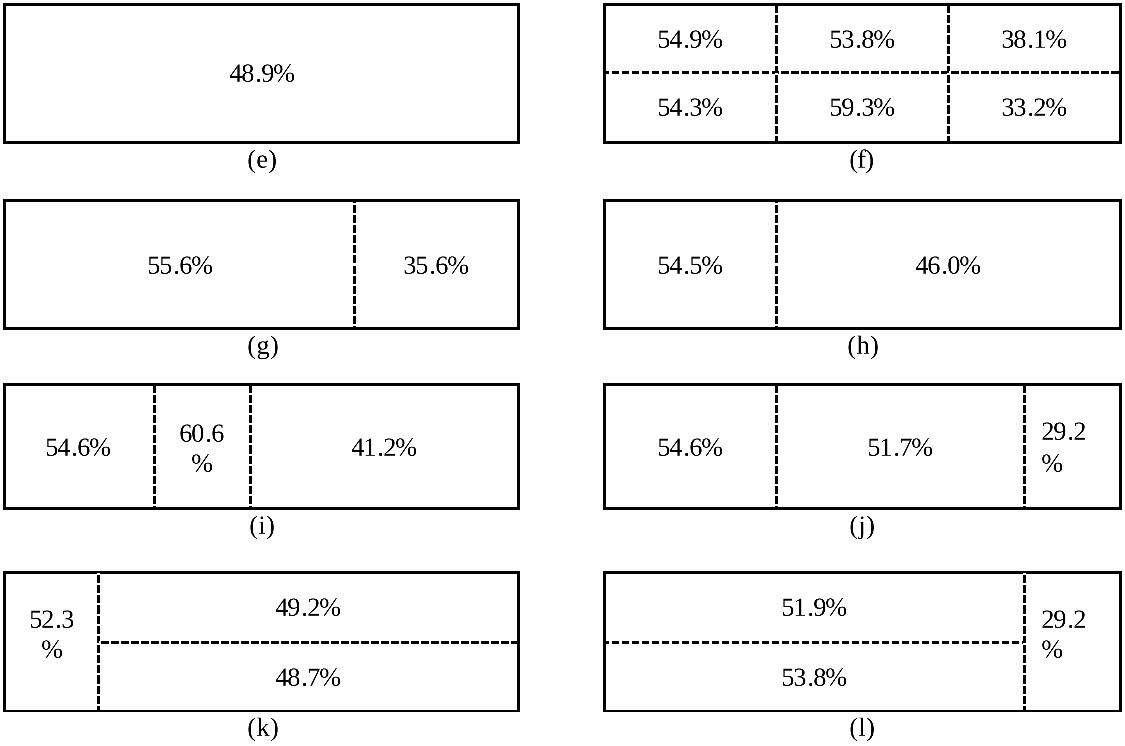


Table 2.3 demonstrates that by aggregating the data in many different ways it is possible to make the red square in Figure 2.6 have many different values ranging from 41.2% (-13.9%) to 59.3% (+4.2%) a difference of 18.1%. None of the aggregations of the data kept the original value of the square.

Table 2.3: The effect of data aggregation on the sample dataset (as shown in figure 2.7)

Example Number	Original Value	(a)	(b)	(c)	(d	(e)	(f)
Value	55.1	48.9	41.2	43.5	56.1	48.9	59.3
Difference		-6.2	-13.9	-11.6	+1.0	-6.2	+4.2
Example Number	Original Value	(g)	(h)	(i)	(j)	(k)	(l)
Value	55.1	55.6	46.0	41.2	51.7	48.7	53.8
Difference		+0.5	-9.1	-13.9	-3.4	-6.4	-1.3

There is no real solution to ensuring that multi-level and scale aggregations of data that display the similar geographic patterns whatever the aggregation. The only real way of getting round the problem is to store all data in the least aggregated form possible (i.e. individual level where possible). When stored at this level the data can then be aggregated to the required level or scale, whether this is postal sectors, census OAs, or electoral wards.

2.7.3 The Labelling of Clusters

The labelling of groups produced within a classification can be a contentious issue. It is important for the groupings to be accurately labelled so that the classes can be clearly identified in terms of their makeup. However, the naming of the classification should not seek to order the groups using words such as 'best' or 'worst' or use language that is derogatory to the less well off areas. Often names can be too specific and give a stereotype of a cluster that although may be an accurate representation of the mean values or the cluster centre it does not represent any of the diversity within the cluster. This is often a result of commercial firms wanting to give their clusters catchy and memorable names. These are often attractive to what can be termed superficial users of geodemographics. People such as advertising executives may want catchy names, but do they take the time to look through any data to see if the labels they are using are in anyway meaningful?

Even more dangerous territory is occupied by those labels that make inferences about areas and are based on the opinion of who created the names rather than the data used to create the classification. An example of this is 'Rural Isolation', a name used to describe a cluster in the top level of the Mosaic classification system. The suggestion that this puts across is that everyone who lives in rural areas feels isolated, a suggestion that may well be challenged by those who enjoy a country life. It paints a negative image of these areas. Contrast this with another Mosaic label 'Urban Intelligence'. This gives out very positive connotations about the people in these areas, the use of the word 'intelligence' suggesting that people live in these areas have some form of intellectual prowess, and consequently some form of superiority over other groups who do not have such a name.

The labels given to the clusters do not in any way affect the integrity of the input data, methods used or the clusters produced. The clusters would still be exactly the same whatever they are called. However, the names matter, because they are a first impression that the classification gives to a user. It will matter little how representative the clusters are if the first thing the user sees is an unrepresentative name. No matter how much descriptive, statistical and illustrative material is provided for each cluster, many users will not look past the names to provide them with an impression of what an area is like. Unless someone is standing over the shoulder of everyone who uses a classification there is no way that we can ensure that they use every bit of available information to make a judgement about a cluster. It is therefore imperative that great care is taken when labelling clusters.

2.7.4 Out of Date as Soon as it is Made?

A criticism of geodemographics has been the inability to update data as it ages. The majority of variables used in all geodemographic classifications are derived from the census, which is two years out of date when it is first published. With 12% of people moving home each year only 77% live in the place that they were enumerated by the time the data are released, before even considering any births and deaths that may have occurred in that period. Census data cannot reliably be updated between censuses other than by predicting changes in the total population.

Several commercial systems also use additional information that can be more frequently updated such as credit listings, county court judgements, and share ownership. This kind of information needs to be updated more frequently than census data because it is much less stable. As much of the data are based on individual records whereas census data are based on areal units, the effect of the movement of one person is much greater on these other forms of data in comparison to census data.

Another consideration is the size of the areas being clustered. The smaller the areas are, the less stable they are over time and the more likely it is that significant changes will have occurred. The majority of commercial classifications are based on postcodes with an average of 14 addresses per postcode; these will be less stable over time than a classification based on OAs which have a minimum of 40 homes and 100 people, and an average of 250 people/110 homes, being roughly three times larger. This means that creating a classification using OAs, rather than postcodes and using wholly census data rather than data from other sources as well, will make a classification less susceptible to change over time. Classifications based on much larger geographic areas such as wards or local authorities, are even less likely to suffer from change over time as the numbers needed to create a significant change in the social make up of the area would be very large.

Changing social patterns have always been a source of great interest for social scientists. There is no doubt that over time social patterns evolve, but how, why and to what extent is a much harder call to make. Orford *et al.* (2002) looked at changing social patterns in reference to the Charles Booth's work on *'the Life and Labour of the People of London'*. They wanted to see how well Booth's mapping of the social areas of London reflected the modern city, posing the question "is Booth's study still valid after over 100 years?"

To answer this question they digitised and geo-referenced Booth's poverty maps for entry into a GIS. A poverty index was constructed using Booth's original data and occupation data from the 1991 Census, using 1991 Census wards as the units for comparison (Harris *et al*. 2005). By comparing the 1991 Census data with Booth's findings they made several significant discoveries. It was clear that in absolute terms poverty had decreased to a level way below that at the end of the nineteenth century. However, in relative terms the picture of affluence and poverty in London has remained fairly stable, the inequalities seen by Booth persist today (Orford *et al.* 2002). Their statistical analysis of the areas suggested that there was a close relationship between those areas of relative poverty in the time of Charles Booth and the areas of relative poverty in 1991. They concluded that social polarisation had reduced due to the growth of the middle classes (Orford *et al.* 2002).

Orford *et al.* (2002) found startling results when comparing the two classifications against health data. Booth's poverty index was actually more reliable in predicting deaths from strokes and stomach cancer than the index that they derived from the 1991 Census data (Orford *et al.* 2002). They concluded that, even over a 100 year period, wholesale neighbourhood shifts in the geo-social hierarchy are a very rare occurrence. Only areas which had seen significant redevelopment (such as parts of London's docklands) had seen a significant change in their place in the social hierarchy. The poorest people of London still live in the same areas as they did when Charles Booth undertook his survey over 100 years ago. This relative stability adds power to geodemographics and is perhaps one of the reasons why area classifications work because in relative terms the geography of social patterns appears to change very little over time (Harris *et al.* 2005).

So area classifications do age over time, but the extent to which this will affect the classification depends on the size of the areal units used and the stability of the input data. Classifications are helped in combating this problem of ageing by only small changes in relative terms in the geography of the social hierarchy over time. However, it is often areas of change that are of interest especially when large scale investments have been made.

2.7.5 Arbitrariness, Transparency and Lack of Validation

Criticism has been levelled at geodemographics for the lack of transparency about the classifications, especially in terms of methodology and the exact details of the variables that have been used in their creation. The reason why so little is known about how geodemographic classifications are created is mainly because it is such a competitive market. Virtually all the classifications available are commercial products that are licensed out to customers for a fee. Revealing the methodology used to make the product and the variables within it would firstly give information to their competitors and may even enable customers to create their own classifications rather than having to buy them in. Development of area classification within academia and the public sector would bring openness and transparency to geodemographics, as any academic publications require a full explanation of the methodology used. The choice of variables can always be criticised as being subjective, although it is hard to think of any form of analysis where subjective decisions don't have to be taken. Subjectivity is not necessarily a bad thing as long as there are good and considered reasons for the choices made. By having to make choices it is possible to discover new things about a variable or a group of variables that otherwise have not been found out.

There is a perceived lack of validation of commercial geodemographic classification systems. At least they do not publicise that they do any form of validation of their products so it is assumed it does not happen, although some systems such as Mosaic, do publish photos of typical areas. Therefore the user may be left wondering, how do I know that the areas are really like the analysis suggests? There are several ways in which a classification could have some validation added to it. A ground truthing exercise could be run, researchers leaving the office and going to different areas of the country to see if how the classification describes an area is similar to what can be seen there in reality. Consultation with experts on local areas could be asked if maps of the classifications accurately reflect the make-up of their area.

2.7.6 Lacking in Theory and Statistical Grounding

Tobler's first and only law of geography states that two things close to each other are more likely to be similar that two things which are further away (Tobler 1970). This law would suggest that geodemographics has a sound theoretical framework. Geodemographics works on the principle that people who live close to each other share broadly similar characteristics. However, others have argued differently as the quote below from Flowerdew and Leventhal (1998) shows:

> "*there is no formal proof and no 'theory of geodemographics' either, only the concept that 'birds of a feather flock together'. All the evidence is empirical and … tend[s] to stay within the companies who have tried the technology. The systems are used simply because they work and have become established …*" Flowerdew and Leventhal 1998 p36

The reason geodemographics has continued to grow is not because there is an increasing amount of evidence that the theories and concepts behind it are sound. Quite simply the success of geodemographics is because the proof of the pudding is in the eating and geodemographics tastes very good. There is a general dichotomy of positions between on the one hand academia and the public sector who, want to know how the systems are built so they can have some form of confidence and justification for using them. On the other hand the business community are happy enough to use geodemographics without any hard and fast evidence, because they do not need to justify what they do to anyone, but themselves. For geodemographics to move forward systems need to be developed by the academic community and provided for use in a transparent way.

2.7.7 Ethics, Privacy and Persecution

Geodemographic classifications along with forms of consumer information such as loyalty card schemes, lifestyle databases and credit reference information have enabled companies to gather information on the population for the purposes of target marketing (Curry 1998). This has caused growing concern amongst some academics that see the use of geodemographics as a threat to privacy and individual freedoms (Curry 1997). This view is not helped by the amount of mis-information that is relayed by people who don't have a full understanding of geodemographic classifications or the data within them. A perfect example is Evans (2004) who suggests that by linking census data to postcodes and the electoral roll, characteristics about individual households can be ascertained. This is untrue, census data cannot be disaggregated further down than its lowest level of geography and no one can be identified. Although what Evans writes is not true, if the reader is unaware of the facts, they may well believe what they read. This has led to some unjustified criticism of the ONS who administer the census and caused already rigorous disclosure control procedures to be increased. This is to the detriment of researchers who want to carry out sound and ethical research using census data. Goss (1995) argues that geodemographics is a self fulfilling prophecy, stating that its use for target marketing ostracises the next generation of consumers who were not targeted. As long as geodemographics is used once it will always work as the same consumers will always be targeted.

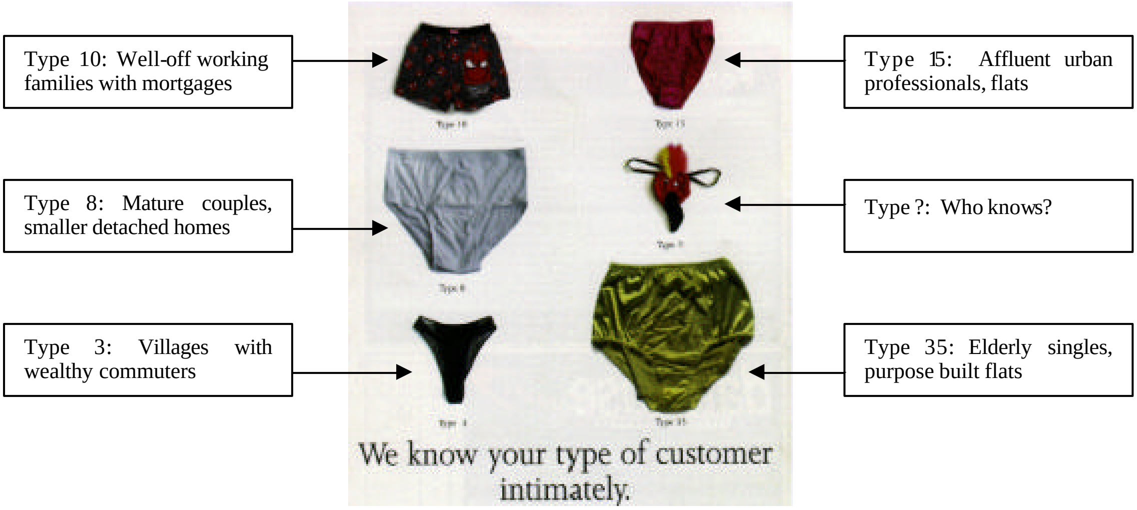
Where you live affects access to and the cost of such things as health care, house insurance, car insurance, life assurance, eligibility to schools, ease of opening a bank account. The media often label this as 'postcode lottery' (BBC Online 2002). This leads to postcode persecution, where the labelling of areas may lead to all people within being treated unfairly. Examples of postcode persecution include a bank that was asking customers from 'poorer areas' to provide higher opening balances than those from 'better-off areas' based on their postcode (Levene 1999). Although geodemographics can't be totally blamed for postcode persecution the use of postcodes as geography in geodemographics has certainly contributed to this by adding labels, descriptions and images of areas to postcodes.

Geodemographics helps companies target their mail shots at a more suitable audience. Despite this every home still gets a lot of mail which they don't want. People are targeted on the basis of their postcode, which is linked to a geodemographic cluster. The total spend on direct marketing in the UK in 2002 was £11.85 billion and has been experiencing a growth of about 6% year on year (Sleight 2004). Much of this growth has been enabled by the use of geodemographic classification systems which has made possible more specific targeting of customers for individual products. Despite the vast majority of targeted mail that comes through British letter boxes going straight into the recycling bin, this form of direct marketing is said to make those who use it an average of £11 for every £1 spent (Winterman 2004).

2.7.8 Misrepresentation: Over-precision and Over-stating of Capabilities

Another problem that is caused by the competitive commercial geodemographic classification market is that the competing companies are forced to out do each other in terms of claims about the quality of their products. It would be unfair to say that they are lying about how good their systems are, as again the black box nature of commercial geodemographic classification systems makes this impossible to check. What can be said is that in their advertising and within the information sent with their classifications, many of the systems make claims about things that can be no more than inferred from the data that has been used to make the system. They pile information about each cluster into the literature that accompanies their classifications to make their product seem to have more value. By providing so much inferred information and it in very precise terms they are in danger of mis-representing the classification and ignoring the diversity within the clusters. By overstating the abilities of their classifications and using unnecessarily precise descriptions, there is a danger that users will misunderstand and misuse the classification, or be dissatisfied when some of the precise descriptions do not match reality.

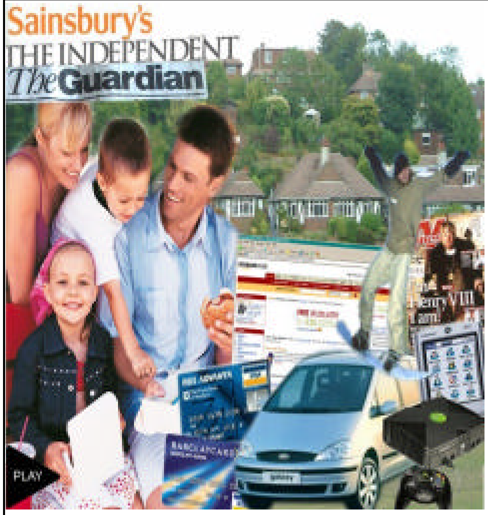
Figure 2.8: Advert for CACI's ACORN in the People Newspaper


Type 10: Well-off working families with mortgages
Type 15: Affluent urban professionals, flats
Type 8: Mature couples, smaller detached homes
Type ?: Who knows?
Type 3: Villages with wealthy commuters
Type 35: Elderly singles, purpose built flats
We know your type of customer intimately.

© CACI (UK) Source: Harris 1999

Figure 2.8 shows such a claim in a light hearted way, but valid for illustrative purposes. The picture (an advert for CACI's ACORN system) displays six undergarments and gives each an ACORN type number and suggests CACI know what type of undergarment people in each type wear. If we actually consider what they are saying it will become clear how ridiculous their claim is. ACORN contains 56 types. On average each type should have a population of about one million people. The advert has given each type shown just one type of undergarment suggesting that everyone of each type wears that type of undergarment. If someone knew their type and looked at the advert, they might think 'I don't wear anything like that', and start to doubt the integrity of the product. Although the advert is clearly meant to be light-hearted, the same principle would hold true for any other product, cars, breakfast cereal, newspapers for example. Suggesting that the clusters as so homogeneous that one item can represent each type is a very misleading statement to make.

Figure 2.9: Visual and verbal portrait of MOSAIC: Group C19 Original Suburbs



"Appearances are not that important. These aren't the sort to make judgements about people based on how they look or what car they drive, and they don't choose clothes or goods themselves to stand out or make a statement. Fashion and looking well dressed is a low priority, and they have little interest in designer brands, probably dressing in a more relaxed, individual way. Among the higher spenders on groceries, they do buy frozen ready meals probably for convenience when feeding a family. Quite a high proportion may also be vegetarian. They enjoy eating out in good restaurants and also like foreign food but they probably also like entertaining at home."

© Experían 2004

Figure 2.9 shows how this perpetuates into material that is provided with classification systems. The visual and verbal portrait of the Mosaic Original Suburbs cluster makes some precise and unapologetic statements about the people within this cluster. Remember that the statements made need to represent the entire population of the cluster. Therefore any all encompassing statements that are made need to be suitably broad. However, it is clear to see that Mosaic does not adhere to this philosophy. Take this statement for example, *"These aren't the sort to make judgements about people based on how they look or what car they drive"* Over a million people live in this cluster so how can this statement possibly be true? As it is a statement of the opinion of the residents it is hard to see how they could know this kind of information about anything but a small sample of people. They provide no proviso that this statement applies to the majority or just some people in this cluster. There are other statements too *'they do buy frozen ready meals"*. What, all of them? They do not qualify the statement with any kind of scale. It is easy for the critical and analytical mind to add provisos to these statements, but quite simply this type of description should not have been written. However, this problem of over precision can affect someone's view of the quality of the product. If a judgemental, fashion conscious, ready meal hater lives within this cluster (and it is more than likely there is a least one), the false statements that are made about them will cloud their judgement of the classification regardless of whether they think everything else about i is brilliant. By overstating the case of what the classification can tell us about an area or a household, it endangers the whole classification because it will not be able to live up to its own hype and will become an easy target for criticism.

2.8 A Fuzzy Future?

The concept of fuzzy logic is one that is becoming increasingly prevalent within both the physical and social sciences. It has entered into the field of area classification with the development of fuzzy geodemographic systems.

The axioms of science and mathematics work in a very black and white way. Something is true or something is not true, this is hard logic. To think in a fuzzy way introduces several new possibilities; something maybe true or it probably is true or it probably is not true. This can be seen as a grey or fuzzy scale. The easiest way to exemplify this is to use that most British of obsessions the weather. We can regard a temperature of 30°C as being hot and 0°C as being cold, but when does hot become cold, and cold become hot. There is no correct answer to this. The answer is hard to quantify and different people may have different thresholds between hot and cold. Someone in Reykjavík will have a different idea as to what is hot than someone in Singapore. This is an idea that is easy for the human brain to grasp, but much harder for a computer to understand. A computer works in a binary code of 1 and 0 some thing is either on/true and therefore 1 or off/false and therefore 0. The challenge is to make the computer think as a human does. Fuzzy thinking can be seen as an East-meets-West amalgamation while still at its early stages in western culture it is an accepted way of working in Japanese industry (Macrone 1998). One of the foremost thinkers in fuzzy logic is Bart Kosko and an excellent overview of the concept is given in his book; *Fuzzy Thinking: The New Science of Fuzzy Logic* (Kosko 1994).

So how does fuzzy thinking link to geodemographics? The idea of fuzzy geodemographics is that areas are not seen as a member of one type but as partial members of all types dependent on values. There are arguments for and against using fuzzy logic in geodemographics. The argument for is that by using fuzzy logic in geodemographics the membership information that is created is better representative of the real world. This sounds great, so why not go full steam ahead with fuzzy geodemographics? The answer is simple; it is the simplicity of geodemographics that has generated its success. "One area in one group" is a simple concept and easy to use. The current users of geodemographics are happy with this. The membership of each group is either a 1 (yes it is a member of this group) or 0 (no it isn't a member of this group). With fuzzy geodemographics things become much more complicated. An answer could now be 0.7 of cluster A, 0.0 of cluster B, 0.2 of cluster C and 0.1 of cluster D. This provides answer that is mathematically more correct than a 0 or 1 solution, but something of the simplicity of geodemographics is lost. Take fuzzy thinking too far and you have lost your classification. The choice is there to be made, but it can be seen as catch 22 situation. Fuzzyfying geodemographics improves the result, but it reduces the simplicity of the method which has been one of its main attractions and contributors to its success.

So should this project develop a fuzzy classification rather than a conventional one? The answer is no. The idea of fuzzy geodemographics is not being dismissed. Fuzzy geodemographics looks set to increase in usability and popularity as the method becomes more refined. The reason for not using a fuzzy method for the main output of the project is that, the project aims to produce a set of free to use general purpose classifications of the UK. Because nothing exists which meets this remit at present already a simple 1 or 0 classification will be the foremost objective of the project. Experiments of developing a fuzzy classification from the existing classification could be an area of further work to which the project leads.

2.9 Conclusions: Extracting a Research Agenda

The development of area classifications and geodemographics has been stop/start since Charles Booth's study of London at the end of the nineteenth century. Research has tended to come in patches rather than a steady development, with the Chicago School representing the next step in the 1930s then people such as Shevky, Williams and Bell, and Moser and Scott further developing the study in the 1950s and 1960s. Then Webber and Craig sowed the seeds of what was to become the modern geodemographics industry in the late 1970s. From then on, with a few exceptions, area classifications have developed as commercial marketing products, which is the situation that we find today. Away from the commercial products whose acronyms are widespread (discussed in § 2.4) national level classifications are seldom used, and barely exist at small scales.

There is an opportunity not only to create a set of classifications that will have widespread use but also to bring classification theory and methodology to the fore in an academic setting. Basic classifications can be made relatively simply so that they can be included in the mainstream of both academic research and teaching. Undergraduate students would gain significant knowledge and information from the power of classifications.

There needs to be more honesty and transparency within geodemographics to take area classifications forward in the eyes of the academic community. The processes used to create area classifications are sound and straightforward (as outlined in Chapter 3). However, for area classifications to be used in academic research the researcher must be confident about the product they are using. It is not enough to just accept that a commercial system seems to produce sensible and useful results. Researchers must know what data are input to a classification and how it was made to have confidence in using it. There is little hope of commercial geodemographic firms releasing such information as it would endanger their product by passing on details to their competitors. It would therefore be judicious to see a watershed in geodemographics, where the academic community cease to use commercial geodemographic classifications and began to develop their own classification systems. Commercial systems should no longer be seen as a general purpose geodemographic representation, but as a marketing tool. This is the purpose for which they were created. To use such a tool with an unknown methodology in academic research is no longer a valid option.

This projects represents the right opportunity at the right time to provide to the wider academic community with the kind of small scale geography area classification that can be easily understood, that is fully documented and that can be reliably used in academic research. This project can provide geodemographics to meet research needs for useful areal geographies, providing free access and ease of use.

Chapter Three - Making a Classification System: a Guide to Methods and Procedures

3.1 Introduction

The biggest question in many areas of investigation is how to organise observed data into meaningful structures? Clustering of data enables this to take place (Tyron 1939). Cluster analysis is not a typical statistical test, but a process in which a cluster algorithm is used to assign each object into a group of similar objects. Each object is represented by a point in multi-dimensional space; each dimension representing a different variable, the values of which fix the location of each object (Anderberg 1973). Unlike many statistical procedures cluster analysis does not require a prior hypothesis. It is very much a technique in the data exploration phase of research.

Although objects are clustered according to similarity, it must be noted that the variance between values within a cluster can be as large, or larger than the difference of values the between two classes. The view that the areas can be classified into mutually exclusive groups must therefore be challenged. Classifications must be viewed as object sets of fuzzy groupings where the outer points of each classification can overlap resembling clouds in a summer sky (Voas and Williamson 2001a). There are many different methods of classifying objects into groups of similarity. Geographic areas are the objects to be clustered in this instance.

Area Classifications are created by the clustering of geographical entities with the use of cluster analysis. The process of cluster analysis, although based on a fairly simple clustering algorithm, is much wider than the clustering of the objects themselves. To run a cluster analysis and therefore create an area classification requires a series of steps, with multiple decisions to be made at each stage (Milligan and Cooper 1987). Each decision has an incalculable, but real effect on the result of the analysis. This makes classification as much of an art form as a science. There is a great deal of skill and knowledge required to make these decisions with confidence. There are no right or wrong answers to any of the decisions that have to be made, they merely produce different results. Consequently different decisions could be more or less suitable dependent on the purpose of the classification that is to be created (Lorr 1983).

The steps involved in cluster analysis are excellently summarised by Milligan (1996), who outlines the 'seven steps of cluster analysis'. Milligan's seven steps were further summarised by Everitt *et al.* (2001) who add their own comments and ideas to Milligan's framework. The steps are described as *"fairly predictable"* (Milligan 1996 p341). Each step represents a major or critical decision that has to be taken to successfully run a cluster analysis. While recognising that, dependent on application, some steps may be more or less important than others. Milligan suggests it is vital that the user recognises the critical decisions that need to be made, and the importance that they may have on the final results. A clear distinction needs to be made between cluster analysis and clustering method. The clustering method is simply the method by which the clusters are formed, while cluster analysis refers to the much wider sequence of steps that have to be followed to complete the whole analysis. Cluster analysis is much more than simply running a dataset through a clustering algorithm (Milligan 1996).

It is essential for users of cluster analysis, especially those hoping their classifications will be used by others, that they record and report decisions taken at each step of the cluster analysis and the reasoning behind each decision. This enables others to not only critically evaluate what the researcher has done, but also gives them the possibility of adding to, or extending the results of the analysis (Milligan 1996). There are many examples of authors who have failed to provide significant information about the decisions taken. Milligan sites Harrigan (1985) who failed to even name the clustering method that was used in the study. Although examples such as Harrigan (1985) can be found within academic literature, no one is as guilty of failing to provide information about the creation of classifications and the steps used in cluster analysis as the firms who create and license out commercial geodemographic classifications.

Harris *et al.* (2005) recognise that little is known about how geodemographic classifications are built or what information goes into them. While appreciating that the problem exists due to the constraints of commercial confidentiality, they fail to point out the implications for anyone who wishes to use these potentially rich data sources in an academic study. A link has been made between a commercial geodemographic firm and a geography department at a high ranking British university. Researchers at the institution have been given free access to a commercial geodemographic classification to aid their research, although there is no way of knowing exactly how much information the commercial firm has passed to the institution about the creation of the classification system. None of the information about the system is available outside the institution. The following question has to be asked. Can any of the research that they have conducted be considered valid if no external party is privy to any of the information within the classification?

Milligan's seven steps are outlined below with comments comprising an amalgamation of the original description by Milligan (1996), additional comments by Everitt *et al.* (2001) and the addition of some further points that relate more directly to area classification.

Step 1. Clustering elements (Objects to cluster, also known as "operational taxonomic units")

a. Should where possible be defined to give a 100% geographical coverage.
b. Should be representative of the cluster structure believed to be present.
c. Should be sampled properly if generalisation to a larger population is required.

Step 2. Clustering variables (Attributes of objects to be used)

a. The variables represent the measurements taken on each entity/area that is to be clustered.
b. Variables should only be included if there is a good reason for their presence such as adding definition to the clusters.
c. Irrelevant or masking variables should not be included as they can hide more significant patterns within the clusters.

Step 3. Variable standardisation

a. There is no requirement that standardisation must be performed on any set of data. It is up to the researcher to decide if standardisation is necessary and if so which method should be used.
b. Standardisation over the range of each variable shows a good recovery of clusters (Milligan and Cooper 1988).

Step 4. Measure of association (Proximity measure)

a. A measure of similarity or dissimilarity must be selected. This reflects the degree of closeness or separation between objects to be clustered. These can work in different ways. For example, Euclidean distance as a dissimilarity measure reports larger values as two entities become less similar, so that the distance between them in Euclidean space is greater. In contrast a similarity measure such as a Pearson correlation, assumes the opposite reporting larger values as two objects become more similar.
b. Either linear or non-linear measures can be used.
c. Few general guidelines. However, knowledge and context of the data may suggest an appropriate measure.

Step 5. Clustering method

a. Methods used should be those designed to recover the type of clusters suspected to be present. This is important as different types of clustering method are better at finding different types of cluster structures.

b. Robustness of method. Some are able to handle different amounts of data, and show different amounts of sensitivity to certain types of data.

Step 6. Number of clusters

a. This is the most difficult decision to be made in cluster analysis. It is especially troublesome if there is no prior information as to the number of clusters expected to be in the dataset.

b. There are several different rules that can be followed for the selection of the most suitable number of clusters. However, these can often be contradictory for the same application.

c. If you can’t choose between two solutions, then the larger number of clusters should be selected.

d. You also need to consider if there are actually any clusters present with the data. When there is no obvious difference between the different solutions produced.

e. There is no right answer to the selection of the number of clusters. The choice is not based on scientific theory and the solution selected should be judged on its usefulness rather than being a correct representation of the patterns within the dataset.

Step 7. Interpretation, testing and replication

a. Interpretation of the results in the context of the applied problem and an assessment of whether the solution adequately meets the needs of the investigation should be undertaken. This requires knowledge and expertise in the discipline in which the investigation has been carried out.

b. Re-run the analysis to make sure the same solution is found on all occasions.

c. Test to determine whether there is a significant cluster structure within the data. Follow by cross-validation to instigate if the clusters are representative of data not originally included in the analysis.

d. Perturbation: examine of the difference to the result by the removal of each of the variables included in the analysis.

Milligan’s seven steps provide a good outline of what is involved in the creation of a classification. However, Milligan provides only general guidelines for all datasets; adaptations will need to be made depending on the specifities of the dataset being clustered, in this case, spatial data and the creation of a general purpose area classification. The creation of an area classification can be seen as a combination of three general steps, inputs, processes and outputs (Harris 1999).

Section 3.2 outlines the inputs into a classification system, including data issues and perspectives about selecting variables. Section 3.3 introduces the processes involved in cluster analysis, including: standardisation, weighting and clustering. Section 3.4 discusses issues relating to the output of area classification systems. Section 3.5 concludes with a summary of the chapter.

3.2 Inputs

The inputs to an area classification are spatial data, predominantly areal data at whatever scale of geography the classification is to represent, but any form of geo-referenced individual or point data can also be used. There are several view points on the data that should be included in an area classification. Areal data are generally more geographically comprehensive and show greater stability over time than point data, due to the aggregating effect of the areal units. Individual or point data can often provide additional information that is not available for areal units. However, individual data rarely have 100% geographic coverage and are much more susceptible to change over time.

3.2.1 Data, the More the Better?

Commercial companies advertise their geodemographic classifications as having hundreds of variables, suggesting that their classification is better than their competitors as it is built from more data. However, there is little evidence to suggest this is true. Milligan (1996) contends that the opposite to be true, suggesting that variables should only be included if there is a very good reason. The inclusion of less relevant variables can mask and reduce the effectiveness of more relevant variables within the clustering process. Multidimensional datasets are very difficult to understand and difficult to represent graphically. Adding further redundant information to the classification process serves no purpose other than to make the results of the analysis harder to interpret and unnecessarily complicated (Milligan 1996).

3.2.2 Census Data

The UK Census provides much of the data needed to create a geodemographic classification; on the 29th April 2001 the numbers and composition of the present day UK population were surveyed with the undertaking of the 20th Census of UK population. The Census of the UK has taken place once every ten years since 1801 (with the exception of 1941 due to World War II). As well as the decennial census there was also a 'sample' census which took place once in 1966, but was not repeated. The territorial extent of the UK has not stayed constant during the history of the census: from 1801-1911 the United Kingdom represented Great Britain and Ireland, but following the Irish war of independence (1919-21) a north-south partition of Ireland was established and the South of Ireland gained independence from the UK. Therefore, from 1921 onwards the census of the UK represents Great Britain and Northern Ireland, but not the South of Ireland.

The Census is the most complete source of information on the number, characteristics and location of the UK population. The census collection and dissemination process requires three main components: firstly, the people who complete census returns; secondly, the census offices who are responsible for the collection, editing and production of data; and thirdly, the licensed census partners who disseminate census data and produce value added data products (Rees and Martin *et al.* 2002). The importance of the census should not be underestimated, as results from the census provide an input into a large number of the reports, findings and policies of both national and local government, research into decisions to open or close facilities such as schools, hospitals, and clinics use information revealed by the census (Boyle and Dorling 2004). Census information is also a valuable tool for marketing companies, business planners and academic researchers (Raper *et al.* 1992).

Superficially a census looks relatively simple as it is based on a form containing only 20-40 questions (Dale 1993). However, each of those questions has a number of categories, each of which can lead to an indicator (e.g. the percentage of the population aged under 5, the percentage of the population aged 5 – 14, the percentage of the population aged 15 -19 etc.). The categories of one question can be crosstabulated against several others. For example, the percentage of females in a particular age band who are married, and who work full time. This piece of information involves the answers to four separate questions. The number of possible crosstabulations and associated indicators is a number of gargantuan proportions; the number of sub-populations for which crosstabulations and indicator variables can be generated is also very large. There were 223,060 output areas generated in the 2001 Census of the United Kingdom. Even if the results are confined to a simple rectangular matrix of 223,060 rows by say 10^6 indicators, this will produce 2.23E11 or 2.23×10^{11} cells of information.

Data available from the UK Census include such things as Population present, Population resident, Age, Living arrangements, Marital status, Country of Birth, Ethnic Group, Religion, Health and provision of unpaid care, Economic activity, Hours worked, Industry of employment, Occupation groups, Qualifications and students, National Statistics Socio-economic Classification, Travel to work, Household spaces and accommodation type, Cars or vans, Tenure, Rooms and amenities, Household composition, Communal establishments, Migration and for Wales only, Knowledge of Welsh.

3.2.3 Issues of Census Data Quality

The census is a high quality and comprehensive dataset, but there are still several data quality issues, of which all users of census data need to be aware. It is all too easy to jump into using a dataset without examining issues of quality that would affect how the data should be best used. Knowledge of these issues can prevent even an experienced researcher from falling into some potentially serious traps.

The 2001 Census is the most comprehensive survey of the UK population ever undertaken and attempts to count everyone present in the UK on census day. However, this does not mean that everybody in the UK was counted in the enumeration process. A large number of people failed to fill in and return their census forms; even after follow up attempts to problem residences there were a large number of missing returns. Any people who failed to respond had to be imputed into the census using knowledge of who they expected to find at each non-responding residence (ONS 2003b). Not only was whole record imputation necessary, but answers to individual questions had to be imputed or changed where contradictory responses had been given. This would be enough of a problem, but the response rate differs greatly by geography and demography (ONS 2005a). People were much less likely to not respond to the census in the centre of large cities than in less urban areas. Irregular housing patterns in these areas did not help in this matter. Younger people, especially young males, were less likely to return their census form than other sections of society (Simpson 2002). Some people such as those who live in the UK illegally were unlikely to fill in their forms for fear the information given could be used against them (ONS 2005a).

Another issue relating to census data is its time reference or currency. The census is only carried out once every ten years and some of the data are not released until up to four years after the census enumeration. This means some 2001 Census data will still be in use in 2015, fourteen years after its capture. While population estimates are made for the intervening periods, this only helps with the actual number of people not their social make-up. Simply by examining the average migration rate of 12% a year, would mean that only 17% ($100\times (1-0.12)^{14}$) of people will live in the same place in 2015 as they did in 2001. When births and deaths are also taken in to account, can we have any confidence in the long term value of census data? Well to a certain extent yes we can. When some body moves out of an area they are likely to be replaced by someone broadly similar in terms of socio-economic status. Residential social patterns change only very slowly over time (Orford *et al.* 2002). Although census data will of course date over time, it should still be broadly representative of the population present at the time of the next census. However, with the passing of time users of census data must be aware that the data are constantly ageing. Precautions can be taken to limit the effect of this ageing. For example by using the largest feasible geographic scale of data.

The final major issue of census data quality is that of disclosure control. The census agencies have an obligation to anonymise the data that they produce to ensure that the characteristics of any one person are not disclosed. For aggregate statistics of large areas this is not a problem as the counts of people in each cell are likely to be numerous. However, for much smaller geographies such as Output Areas (OAs) the chances of finding a cell with a single count are much higher, therefore the data are altered to prevent the identification of people and the disclosure of information about them (Rees and Martin 2002). The most obvious evidence of disclosure control in the 2001 Census is the absence of any ones or twos in any of the census data. This presents itself as an abundance of zeroes and threes in the dataset, especially at the output area scale. Much comment has been made about the methods of disclosure control used in the 2001 Census as they have had a much greater effect on the data then methods used in previous censuses (Williamson 2005). The problem becomes especially apparent when multiple variables are used when small numbers acting together can produce values of over 100%. There is little that can be done to overcome these problems, although the problem is not a serious one for the creation of an area classification as it is the large numbers displaying distinctive patterns that are indicative, not the small numbers. The problem will come when a variable is considered for inclusion that only has a very small membership nationally. The difference between zero and three for these kinds of variables could have a significant impact on the classification especially at OA scale. Careful consideration will have to be given before including such variables in the classification.

3.2.4 Other Data Sources

As well as using data derived from the 2001 Census, other data sources can be used to supplement the census information and hopefully add new dimensions to the classification. However, the principal role of adding such data to a census based classification should be to provide data that is not provided in the census. The most obvious topic not covered by the census is information on income and wealth. Commercial geodemographic firms add non-census data to their classifications to principally provide information on wealth and affluence that is the major weakness of the census, which adequately provides information on the very poorest in society but struggles to identify the best off quite so well. Commonly used non-census datasets that are used in commercial classifications include the electoral roll, county court judgements, Land Use Surveys, Financial data such as share ownership, Monthly updated unemployment figures, Indices of Deprivation 2004, Land Registry data, DVLA - Car Registration, DfES - School results, Lifestyle data from consumer loyalty schemes, Credit referencing data and Companies House data (Sleight 2004). Additional data sources that can be used will depend heavily upon availability, confidentiality, and the spatial scale at which the data are produced. Some of the data are freely available or available on request and under licensed conditions. However, some data such as credit reference information and lifestyle data are only available for inclusion in commercial classifications as they are collected by the companies who make the classification or companies with whom they have data sharing agreements.

3.2.5 Data Quality Issues from Other Data Sources

There are obvious benefits to adding non-census data to a classification to enrich the pool of information from which the classification is built; additional data can provide information that is not available from the census. Another benefit of non-census data is that it is likely to be updated at much more regular intervals than the census data, often annually and in some cases monthly.

There are several dangers that should taken into account when using different sources along with data from the census. Firstly, the accuracy of the data has to be assured even from reputable sources such as other government departments. It is important to know how and when all the data were constructed. Few datasets are as well documented as the census in terms of the enumeration and processing methods and it is unusual to find significant data support for any of these other data sources. The coverage of the data will not be the 100% that is available in the census. Most of these datasets will be based on a sample of the population unlikely to represent more than 10% of the country. Additionally these samples are unlikely to be representative either geographically or demographically; certain sections of society are likely to be over or under represented in the data. There will also be many other sources of uncertainty, it is impossible to know if the data contain undocumented and unidentifiable errors.

A classic example of errors that can be put into a classification system by adding data from other sources is the set of DVLA statistics on car registration, which are used by several commercial geodemographic systems. Car firms register their cars at the factory to facilitate a quick sale especially when a new model is released. This can be seen clearly in Experían postal sector data which shows the Swindon postal sector SN5 6 to have a population of 5,034 and 106,644 registered cars which works out at 21.2 cars per person including children and non-drivers. Postal sector SN5 6 is not home to several thousand car collectors; it is the location of a

Honda car factory (Vickers 2003). The dangers of using such data are obvious to see. Even with the removal of such anomalies from the dataset it would be very difficult to have any confidence in the rest of the data.

Data from other sources should only be used if they add further dimensions to the classification. By adding more variables to a classification that just reinforce trends already provided by the census data only hinders the formation of the classification through the added complexity that it brings to the variable list. It is important that any non-census data which is brought into the classification are at the same spatial scale and refers to the same geographical system. Although several methods of transferring data between systems of spatial registration have been formulated and indeed some are used widely, no system has yet been formulated that can transfer data between overlapping areal units to a satisfactory level of accuracy (Vickers 2003).

3.2.6 The Theory of Selecting Input Variables

The goal of the variable choice for the creation of an area classification is to select the minimum possible number of variables that satisfactorily represent the main dimensions of the 2001 Census and therefore, to get the most information possible in the fewest variables possible into the classification (Bailey *et al.* 1999a, 1999b and 2000). Although in the previous section the use of non-census data was discussed, for simplicity here only choices and comparisons between census variables will be discussed.

There are two main reasons why the minimum possible number of variables should be used to avoid co-linearity and to reduce computational demands. To prevent co-linearity, each variable that is included should add something that the other variables do not give to the classification. As the data are all from the same sources and about the same geographic areas, it is likely that any selection of variables will contain a host of interrelated variables. The problem that co-linearity gives is that it makes it difficult to assess the effect that a variable is having on the classification. It will not only have its own characteristics, but will be working with any correlated variables making it difficult to assess the strength of effect each is having on the classification. The more variables that are added to the classification the less likely they are to add any new information, and the more likely they are to be just repeating information which is covered by one or more variables already selected.

Voas and Williamson (2001a) go beyond suggesting that fewer variables should be used, calling for the increase in 'problem-specific' classifications. In response to comment by Harris (2001) on their paper the same authors suggest that *"By conflating a range of marginally correlated measures, such as income and newspaper readership, there is an inherent tendency to obscure* *the actual between-area differences on matters of specific interest"* (Voas and Williamson 2001b p335). Voas and Williamson (2001b) make the important point that, by adding in every variable which the architect of a classification can get his/her hands on, patterns of interest can be clouded by other irrelevant variables. They see this as a reason to change tack and produce a series of 'problem-specific' classifications. However, one could take a different stance on this point to say that the process of variable selection is a very important one and great consideration should be given to the inclusion/exclusion of any variable. Variables should only be included if they inherently contain information that you wish to be displayed in the geographic patterns represented by the classification.

There is another, less intellectually based reason for selecting fewer variables, which is fundamental to the successful creation of any classification. The greater the number of variables used the more computer processing power is required to generate the clusters and therefore the more time it takes for the procedure to run. If the number of variables to be clustered goes beyond a certain level, it could go past the current capabilities of the computer.

Selecting the fewest possible variables is not an argument that is put across by everyone. Many people reason the more variables used the better. Harris *et al.* (2005) state that: *"As a general rule, but with limits, the more variables that are used in the clustering algorithm and the more different sources they come from the more meaningful (nuanced not idiosyncratic) the resulting set of clusters is likely to be"* (Harris *et al.* 2005 p151). This view is not supported by anyone except commercial geodemographics companies. Academic literature especially within mathematics suggests that the minimum possible number of variables should be used (Everitt *et al.* 2001).

It is not easy to gauge the opinion of the creators of commercial classification systems on this issue as they have traditionally been reluctant to disclose how their classifications are made. However, it is generally regarded that they have a "the more the better attitude" in terms of variables. The latest Mosaic UK brochure states that *"A total of 400 data variables have been used to build Mosaic"* (Experían 2005). A EuroDirect promotional leaflet for their Cameo classification claims it includes *"over 9,000 pieces of information for over 150,000 Census units"* (EuroDirect Unknown p5) while their current website states that the latest version contains *"over 2 billion items of data"* (EuroDirect 2005). So what evidence is there of the effect that too much data can have on a classification? Harris *et al.* (2005 p160) displays a table of the age data used in the creation of Mosaic UK, which contains 22 age groups as follows: Aged 0-4, Aged 5-9, Aged 10-14, Aged 15-19, Aged 18-24, Aged 20-24, Aged 25-29, Aged 25-44, Aged 30-34, Aged 35-39, Aged 40-44, aged 45-49, Aged 45-64, Aged 50-54, Aged 55-59, Aged 60-64, Aged 65+, Aged 65-74, *Aged 75-84*, **Aged 85+, Aged 85-89** and Aged90+ (the highlighting in this list is explained below).

This is a more than comprehensive list of age groups. Some of them even overlap (e.g. Aged 85+, Aged 85-89 highlighted in bold), therefore covering the same information twice. How strongly are these two variables related? How much new information does including them both give us? By running a simple Pearson’s Correlation on the two variables the level of redundancy can be established. The analysis produced a statistically significant correlation of 0.941, a very strong relationship. By multiplying the result by one hundred then squaring that the percentage of one variable which is associated with the other can be calculated, that’s $(0.941\times100)^2$ = 88.5%. This shows that by having just one of the variables in the selection we also get 88.5% of the other or another way of looking at it by adding the second variable the amount of new information that is gained is only 11.5% of the information that is given by the first variable. This is not a surprising result as the two variables overlapped so shared information, but they were both included in the Mosaic system.

How much of a relationship do two contiguous variables show? To test this Aged 85-89 will be correlated with Aged 75-84 (*Italics*). The analysis produced a statistically significant correlation of 0.682, not as strong a relationship as last time which is not unexpected as the variables are no longer overlapping, but the relationship is still statistically significant. How much new information does the variable Aged 75-84 give? $(0.682\times100)^2$ = 46.5%, this shows that only just over 50% of the variable is new information. The rest is associated with the other variable. It may seem intuitive that these variables are as highly correlated as they are similar age groups will have similar residential needs/preferences. An important point to note is that in datasets with a very large number of observations such as this, even very low correlations will be significant simply because of the large number of data points involved.

Is there then any relationship with a variable that is not as similar? Aged 0-4 (underlined) is at the other end of the age scale and therefore the reasons of high correlation because of similarity should not be present. By correlating Aged 85-89 and Aged 0-4 the result gives a statistically significant correlation of -0.305, with a variance of $(-0.305\times100)^2$ = 9.3%. The correlation is much less than those previously seen but it is still statistically significant. We have almost 10% redundancy, why is this? The reason for the 10% data redundancy between these two variables is for a different reason from the previous variables. The variables do not overlap and they are not contiguous, so why does one explain 10% of the other? The clue is the negative nature of their relationship. The previous relationships were positive; this is because the two variables were inherently very similar. These two variables are, however, not very similar and are at different ends of the life course. However, the reason for the redundancy is a fairly simple one, in that they are both age variables. The redundancy is caused by each person only being able to be in one age category all the age variables will show a certain level of inter-redundancy because of this.

Negative redundancy between categories within the same variable will always be experienced and is something that cannot be avoided. However, there are some things that can be done to reduce its effect. If age variables are going to be used in the classification and we have *n* different age groups to get all the information about age into the classification we only need use *n*-1 groups. Why is this? For the same reason that 10% of the prevalence of 0-4 year olds is explained by 85-89 year olds' negative inter-correlations. By using *n*-1 all the information is still being used even though all the groups are not present, because of inter-dependency. Despite dropping any one of the groups within a variable the data are still there. A simple example of this is households with access to a car and households who do not have access to a car. The inter-dependency here should be clear to see: by having a car, you cannot also not have a car and can therefore only be in one group. Adding both variables into the classification does not give any more information than just using one. In fact, a variable has been inadvertently double weighted as the same data would have been used twice. If you look back at the list of age variables used in the Mosaic classification you can see that there is no gap in the age categories each one has been used, they have used *n* not *n*-1. This is unnecessary as it does not yield any more information.

The examples given show the type of considerations that need to be addressed when selecting variables for the classification. Inter-correlation and inter-dependency between variables are to be avoided. It is harder to interpret and understand the patterns that are produced and the reasons behind them.

3.2.7 Correlation Analysis

Correlation techniques have some idiosyncrasies that are important to keep in mind before using such techniques. Variables which share the same denominator (i.e. calculated as a percentage) have a natural tendency to produce a negative correlation (Miles and Shevlin 2001). There is for example a strong inverse relationship between the married and single population. In some cases this effect can be difficult to untangle from a genuine negative correlation. The relationship between the people who have no car and those who have numerous cars has not only a technical negative correlation, but one that also shows social importance (Voas and Williamson 2001a). Variables that don't share the same denominator can also show a close relationship. The number of married men shows significant relationship to the number of women who are married. This can be traced back to the fact that they are derived from the 'marital status' question on the original census form, which has then been sub divided by gender (Voas and Williamson 2001a).

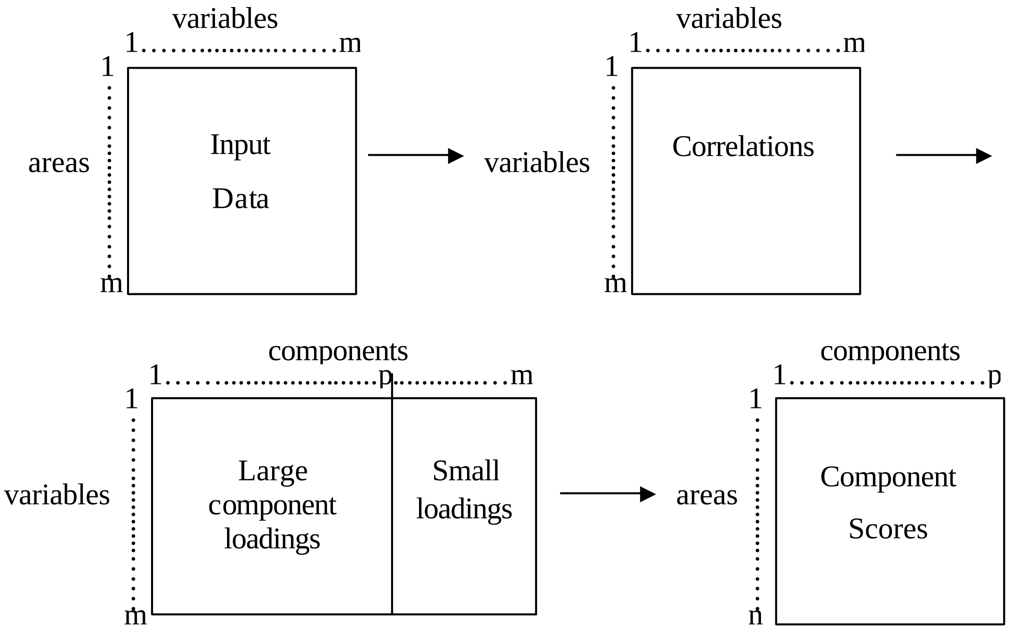
3.2.8 Data Reduction

In the past, mainly due to computational limitations, data reduction techniques have been employed on the variable list prior to clustering. Principal Components Analysis (PCA) is not a method of classification, but a preparatory technique used to remove redundancy from the variable list. By calculating the correlation of each variable with all others, redundancy can be reduced by removing one of a pair of variables that are highly correlated (Voas and Williamson 2001a). By removing redundancy from the dataset this not only makes classification techniques quicker and easier to run, but also enhances the effect of the less correlated variables on the classification.

The basic assumption made by PCA is that a few underlying components or factors within the data can be used to explain the complex relationships within the whole dataset (Norusis 1985). Correlations between the data show that the correlated variables share a dimension of commonality. The prime aim of PCA is to identify these non-directly observable factors based on a set of chosen observed variables. PCA has been widely used by geographers since the 1960's although its complexity is rarely appreciated by researchers who make use of it via off the shelf computing statistics packages (Robinson 1998). It is based on the application of Pearson's product-moment correlation to a standard geographical matrix of places versus a series of observed statistics about those places (Rummel 1970).

Starting with a matrix of *n* areas by *m* variables, the aim is to reduce this to a matrix of *n* areas by *p* important components, where the *p* components are combinations of the original *m* variables. The number of *p* is less than *m* because the components that are associated with only a small amount of the variance of the input dataset are ignored. The matrix scores of the *p* components are then inputted into the cluster analysis as a substitute for the original variable set. This sequence is outlined in Figure 3.1.

Figure 3.1: The sequence of Principal Component Analysis



The advantages of using PCA are that it removes variable redundancy from the dataset, focusing on the main patterns. The disadvantage is that not using the original variable set makes the resulting cluster profiles difficult to interpret as the component scores are composites. This results in the additional labelling problem of naming each of the component. It is sensitive to the magnitude of correlations between the variables. It is sensitive to outliers, missing data, and poor correlations between variables. Outliers must be screened out due to their influence upon the calculation coefficients, which in turn has a strong influence on the calculation of factors and components. In effect, it is a variable reduction technique, equivalent to reducing the number of variables based on correlations between them. However, examining and assessing the correlations between variables is a much more transparent way of reducing redundancy within the dataset.

With the continued increase in computational power, it is no longer essential to run data reduction techniques such as PCA on the variable list prior to clustering. So does PCA have any intrinsic value on top of its function data reduction or does it no longer need to be used? PCA still has value as a useful tool for assessing the predictive power of variables prior to clustering. The values of each component can be used to assess the likely discriminatory power of each variable prior to clustering. The variables which have high values for the early principal components represent those that are likely to have the most discrimination within the clustering process. PCA can therefore be used to make an assessment of the predictive power of variables prior to clustering. However, clustering on principal components rather than the variables themselves is an outdated and unnecessary course of action.

3.3 Processes

The processes involved in the creation of an area classification are more than just the procedure of clustering itself. The data must be prepared for clustering. Clustering algorithms are sensitive to difference in scale (e.g. tens and thousands) and types (e.g. ordinal, ratio or interval). If these issues are not attended to before clustering begins then it is likely that any clusters produced will be a feature of the format of the data rather than the actual data values.

3.3.1 Methods of Standardisation

Before any clustering can be done the variables need to be standardised . This ensures that each variable has the same weighting in the classificatio n. This is especially important when there are different types of data e.g. *population density* will give number of people per unit area, whereas *detached housing* is a percentage of all households. The range of the *population density* is only limited by the number of people who can fit into a specified area. In the UK at OA scale population density ranges from just above 0 to 12,715 people per hectare for OAs whereas housing type can only range between 0 and 100%. These variables are not on the same scale. If left un-standardised the population density would completely control the classification because of the larger range over which the data are stretched. This would also create a large number of outliers based solely on the population density variable. Therefore if these variables were clustered without being standardised it would add bias to the clusters.

All clustering techniques are based on the similarity or dissimilarity of the cases to be clustered. This is measured by constructing a distance matrix reflecting all the variables in the data set for each case. It is clear that problems will occur if there are differing scales or magnitudes among the variables. In general, variables with larger values and greater variation will have more impact on the final similarity measure. It is necessary to therefore make each variable equally represented in the distance measure by standardising the data. The process involved in calculating each type of standardisation is outlined in the following sub-sections.

3.3.2 Z-Score Standardisation

This is the most common form of standardisation. To create z-scores or ‘standard normal variate’ the standard deviation is calculated. The z-score is then calculated by taking the mean value of the variable away from the value for that variable for each area, squaring the difference, adding over all areas, square root the result and then dividing them by the standard deviation of the variable across all areas. This should be repeated for all variables to standardise them over the same range. Let x_i be the value of a variable for area i and x_{mean} the average value of the variable across all n areas.

The standard deviation is defined as:

$$S_x = \frac{\sqrt{\sum_i (x_i - x_{mean})^2}}{n} \qquad (3.1)$$

The standard normal variate or z-score is defined as:

$$Z_i = \frac{x_i - x_{mean}}{S_x} \qquad (3.2)$$

3.3.3 Range Standardisation

This method was implemented in the ONS 1991 classification of Local Authorities; see Wallace and Denham (1996). The data were standardised by the method of range standardisation between 0 and 1 for each variable. The range standardisation method is defined as:

$$R_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad (3.3)$$

where $x_{\max}$ is the maximum value of x and $x_{\min}$ the minimum value of x and R_i is the range standard variate. After the data have been standardised as above each variable has a range of 1 with the maximum value being 1 and minimum value being 0.

3.3.4 Inter-decile Range Standardisation

This method is a slight variation of the range standardisation method, standardising the data over a smaller range.

This method is defined as:

$$D_i = \frac{x_i - x_{med}}{x_{90^{th}} - x_{10^{th}}} \qquad (3.4)$$

The inter-decile range standardised variate D_i compares each value of a variable, x_i to the median, x_{med} which is then divided by the distance between the 90th percentile, $x_{90^{th}}$ and the 10th percentile $x_{10^{th}}$.

3.3.5 Weighting of Variables

Unintentional weighting of variables was touched upon in § 3.2.6. The weights being referred to here are intentionally given to certain variables. In the Mosaic system all variables are given weights defined by the creators of the classification. What is meant by weighting variables? If you have two variables *a* and *b* to put into a simple cluster analysis, but you think that, although both variables should be used, *b* is twice as important as *a*. If this is so the variables can be weighted by multiplying the value of *b* by 2 (after standardisation), it will have twice the effect on the clustering procedure than variable *a*.

The variable choice in itself is the start of the weighting process as in effect all the rejected variables are simply being weighted zero in the clustering process, and those that are chosen are given a weighting of one (Everitt *et al.* 2001). However, weighting goes much further and is much more complex than this. There are many different opinions on how variables should be weighted for cluster analysis. There are those who would simply use their knowledge and experience to weight variables. This maybe a satisfactory way of getting a good result, but is something that is hard to explain and therefore difficult to pass on to others. Others have experimented with weighting algorithms that are designed to reduce the influence of variables which are irrelevant to the clusters present within the data (Milligan 1989). As well as reducing the effect of variables that have little effect on the cluster structure it is also considered advantageous to weight the contributory variables to enhance the cluster structure that is present in the dataset (DeSarbo *et al.* 1984). Investigations into the success of weighting schemes have shown that, weighting schemes based upon carefully chosen estimates of within-cluster and between cluster variability are generally more effective than weightings based on standard deviation or range (Gnanadesikan *et al.* 1995).

Nearly all of the research into the success of different forms of weighting has been carried out on task specific classifications, where it is simple to assess how well the weighting has improved the classification by the partitioning of the clusters for that purpose. However, for a general purpose classification such as an area classification the task is all the more difficult, as the success of the partition cannot be compared against another specific application.

The weight given to a variable reflects the investigator's view of the importance of that variable to the task of the classification (Everitt *et al.* 2001). Therefore a question that arises from this is: is it sensible to weight variables in a general purpose classification? Applications of the classification are not known at the time the classification is created. Weighting variables and testing how well they perform against another dataset can not give a good indication of how the weighting has affected the performance of the classification, as an improvement of discrimination against one dataset could reduce its discriminatory powers against another.

The difficulty, impracticality and doubt over the benefit of weighting were considerations that were not lost on Romesberg (2004) who suggests that the weighting of variables for cluster analysis only makes sense when the research goal is clearly defined. For a general purpose classification the research goal is fairly loose as it will be used for multiple applications.

There have only been limited suggestions as to how to weight variables for a general purpose classification. Harris *et al.* (2005) suggests that variables should be weighted so each domain receives the same total weight when the weights of variables within each domain are summed together, but do all domains deserve or require equal weighting? Some domains may have more relevance than others. The domains with the least variables would receive the highest weightings, but exactly which of the variables should be weighted the highest? It is difficult to give an answer to any of these questions. Variables with strange or unreliable distributions could be weighted lowly (Everitt *et al.* 2001). However, it could be argued that they should not be included at all unless they are absolutely vital variables. It has been suggested than those variables which show the most discrimination should be weighted highest (Everitt *et al.* 2001). Is this really necessary though as these variables are already the most discriminatory? This would be fine in a specific purpose classification, but in a general purpose classification as much of a case could be made for weighting the least discriminatory variables higher to try and get more discrimination out of them.

Makarenkov and Legendre (2001) suggest that weighting procedures should be used to eliminate noisy variables that do not contribute relevant information to the classification structure. However, can any variable be seen as noisy or masking in a general purpose classification when the application is not known and therefore the value of each variable cannot be assessed for all possible uses? The noise that seems to be produced by a certain variable may actually improve discrimination for certain applications of the classification.

It would seem that for a general purpose classification the weighting of variables is as likely to confuse as it is to improve the classification. The main problem being as the classification is not task specific there is no way of knowing if the weightings chosen are any better than an alternative selection as all possible uses that the classification will be put to cannot be known. It is probably more sensible to spend extra time and effort in selecting the list of variables to go into the classification, which is in itself a form of weighting.

3.3.6 Methods of Clustering

The process of classifying information is one that many people have made attempts at redesigning and reinventing. There are positives and negatives to most of the procedures, from the more traditional clustering algorithms to more sophisticated techniques such as neural networks. This section will briefly review alternative clustering methodologies and then a more detailed description of the clustering methods that were used in the project is given in the subsequent sections.

There are many different clustering algorithms. However, many algorithms are either very similar to each other or unusual or designed for a specific purpose, and therefore rarely used. There are a few commonly used types that are favoured mainly due to their reliability and transparency. The most commonly used can be grouped into two broad types: hierarchical agglomerative and iterative relocation (Harris *et al.* 2005). Everitt *et al.* (2001) and Gordon (1999) give excellent descriptions of many different clustering methods.

Hierarchical agglomerative or stepwise clustering methods are top-down approaches to clustering. This is one of the conceptually simpler approaches, where each object starts separately and is joined together one at a time creating a cluster hierarchy, of every cluster number from *n* to 2 (Harris *et al.* 2005). The main advantage of this method being that it produces multiple cluster solutions with just a single running of the algorithm. The hierarchical nature of the system enables more than one solution to be selected and used simultaneously without contradiction (Everitt *et al.* 2001). The major disadvantage of this is that the top down approach takes a long time to compute and is consequently difficult to implement on datasets containing more than about one thousand objects (Harris *et al.* 2005). Although multiple solutions are created, because of the hierarchical nature of the classification one optimal solution is unlikely to be produced. Each new cluster level is created by the merging of two clusters from the previous level. The hierarchy does not allow objects to move between clusters with the increase in the number of clusters (Romesburg 2004). There are multiple forms of agglomerative clustering; available alternatives in the SPSS system are between-groups linkage, within-groups linkage, nearest neighbour, furthest neighbour, centroid clustering, median clustering, and Ward's method (SPSS Inc. 2001). Ward’s method is the most commonly used of these methods and appears to work well, although can sometimes impose a general spherical cluster where one does not necessarily exist (Everitt *et al.* 2001).

Divisive or ‘de-agglomerative’ methods work in the opposite way to agglomerative clustering methods, with all objects starting in one large cluster and successively splitting into more and more clusters (Everitt *et al.* 2001). Often used with binary data for which the method is efficient on a simple presence/absence basis. Less commonly used than agglomerative methods, their main advantage being that the main structure of the dataset is revealed from the start of the clustering rather than towards the end of the clustering as in agglomerative methods (Kaufman and Rousseeuw 2005). The method is computationally demanding if all possible sub-divisions are considered at each stage of division (Everitt *et al.* 2001).

The k-means iterative relocation algorithm is the most commonly used method of classification. The primary benefit of this method is that the clusters it produces retain a high proportion of the variance of the input variables. K-means also produces clusters that are relatively even in terms of membership, especially with a large number of objects and a small number of clusters (Harris *et al.* 2005). The main drawback of the method is that the number of clusters has to be specified before the process is run. Although this does provide a saving in terms of computational processing, it means that if the ideal number of clusters is not known before clustering. The process has to be run many times and a choice has to be made between solutions (Gordon 1999).

There has been a great deal of attention paid to artificial neural networks as a method of clustering in recent years (Everitt *et al.* 2001). Artificial neural networks are computing algorithms that attempt to emulate the capabilities of large networks of simple elements, originally introduced as models of neural activity in the brain (Openshaw and Wymer 1995). A neural network contains three main features: the neurons or basic computing elements, the design of connections between computing units and the training algorithm used to establish the parameters for performing the set task (Everitt *et al.* 2001). Openshaw (1994) describes how an artificial intelligence technique, know as a Self Organising Map developed by Kohonen (1984) was used to create the GB profiles geodemographic system, which clustered the EDs from the 1991 Census. An excellent overview of 'neural networks for clustering' is provided by Murtagh (1996). Artificial neural networks have been shown to successfully classify data especially unsupervised versions such as the Self Organising Map that do not require the number of output clusters to be pre-specified (Kohonen 1998). However, neural network methods have a major drawback. The black box nature of their hidden layer(s) makes the operations that take place within the system difficult to understand, repeat and describe. It is essential to understand exactly what is happening during the clustering process and using neural network methods makes this almost impossible.

The vast majority of clustering methods are based around mean values, although this is not essential. Kaufman and Rousseeuw (2005) describe a method called 'Partitioning Around Medoids' (PAM), which clusters data based on median values. The method is generally robust, but has a number of drawbacks, not least that a two equally valid mediods can be calculated around which a partition can be made and a tendency for atypical objects to produce singleton clusters even when a relatively small number of clusters are specified (Kaufman and Rousseeuw 2005).

A further clustering procedure is included in the SPSS statistical package. The TwoStep Cluster Analysis procedure is an exploratory tool. The algorithm employed by this procedure the ability to handle both categorical and continuous variables (SPSS Inc. 2001). The ability to incorporate categorical data into the clustering process is very useful in certain instances, but is not necessary for this study.

The partitioning of objects into classes is considered by many as an oversimplification of the structure of many complex datasets (Gordon 1999). This is especially relevant to objects that appear towards the edge of clusters, or objects that display attributes of more than one cluster. Fuzzy clustering is a method that is put forward to provide additional details about the properties of each object; they give proportional membership of a number of clusters rather than total membership of one (Everitt *et al*. 2001). Fuzzy versions of most clustering methods have now been developed, and have many advocates such as Feng and Flowerdew (1998). However, there is one major drawback to using a fuzzy classifier. The whole point of clustering data is to simplify a complex system to aid understanding and a fuzzy classifier adds more complexity to what was once simple. Despite undoubtedly more accurately representing reality, this should not be the main objective of a classification. The simplicity of a classification system has been its greatest asset and therefore fuzzy classifications will not be considered for use as the main output from this project.

Ward's and k-means algorithms were chosen for the methodology for this project (as described in Chapters 4 and 5). More details about how Ward's and k-means work are outlined in § 3.3.7 and § 3.3.8. Although Ward's and k-means have been chosen for use all the methods reviewed above are valid for this form of analysis. Some methods have been used in the creation of previous classifications; for others there is no recorded evidence of their use for area classification. The choice of clustering method can be a point of great debate as each has its strengths and weaknesses. The most important factor is that the researcher is both comfortable with and confident in the algorithm they are using and that they have a good understanding of how they work.

3.3.7 Ward's Hierarchical Clustering Algorithm

Developed by and named after Joe H. Ward of the Aerospace Medical Division, Lockland Air Force Base, Ward's hierarchical clustering algorithm was first published in the Journal of the American Statistical Association in 1963. It was developed as a method *"to cluster large numbers of objects, symbols or persons into smaller numbers of mutually exclusive groups, each having members that are as much alike as possible"* (Ward 1963 p236). The aim was to join objects together into ever increasing sizes of cluster using a measure of similarity of distance. At the start of the process each object is in a class by itself. Then in small steps the criterion by which the objects are clustered is relaxed to produce fewer but larger clusters at the next step up the hierarchy. This process continues until all the objects being clustered fall within a single cluster. The process of linking more and more objects together means that they are amalgamated into larger and larger clusters of increasing dissimilarity (Ward 1963). The number of clusters does not have to be pre-specified. The technique produces *n* clusters to 1 cluster inclusive, giving the user the ability to choose the most suitable number of clusters after the clustering process.

The process of hierarchical clustering is an agglomerative or stepwise approach beginning with n groups each containing 1 object then after merging them together ending with 1 group containing *n* objects. The process of getting from *n* to 1 groups can be summarised as below (following Ward 1963):

1. Place each object into its own cluster C, creating the cluster file *f* :

$$f = C_1, C_2, C_3, ..., C_{n-2}, C_{n-1}, C_n \qquad (3.6)$$

2. Compute a measure of similarity between every pair of clusters in the cluster file f to find the closest cluster to each cluster $\{C_i, C_j\}$
3. Remove C_i and C_j from f
4. Merge C_i and C_j to create a new cluster C_{ij} which will be the parent of C_i and C_j in the hierarchical cluster tree.
5. Return to step 2 until there is only one cluster left.

Methods of hierarchical clustering have been incorporated into the statistical packages for the social sciences and are frequently used to cluster census type information. There are several different distance formulae that can be used as the criterion in a hierarchical grouping procedure. The most common are Euclidean or Squared Euclidean measures, although others are used (discussed in § 3.3.9).

3.3.8 K-means Classification

The k-means algorithm is a simple non-parametric clustering method, where k stands for the number of clusters created. The objective of the k-means algorithm is to minimize the within cluster variability. If the number of clusters within the dataset has already been pre-specified, a k–means classifier can be used, for example, to form five clusters that are as distinct from each other as possible. The k-means clustering function in a statistical package such as SPSS will move objects between clusters with two specific purposes, firstly to minimise variation within clusters, and secondly to maximise variation between clusters. K-means is one of the most commonly used methods in the geodemographics industry (Harris *et al.* 2005). It is an iterative relocation algorithm based on an error sum of squares measure. The basic premise of the algorithm is to move a case from one cluster to another to see if the move would improve the sum of squared deviations within each cluster (Aldenderfer and Blashfield 1984). The case will then be assigned/re-allocated to the cluster to which it brings the greatest improvement. The next iteration occurs when all the cases have been processed. A stable classification is therefore reached when no moves occur during a complete iteration of the data. After clustering is complete, it is then possible to examine the means of each cluster for each dimension (variable) in order to assess how distinctiveness of the clusters (Everitt *et al.* 2001). The k-means clustering algorithm is comparatively simple and works as follows in its SPSS implementation (Everitt *et al*. 2001, pp. 99-100 and SPSS Inc.1999):

1. Choose an initial grouping of objects into the desired *k* clusters; compute the means for the groups over all variables and the sums of squared deviations of objects from group means.
2. Move each object from its own group to each other group and re-compute the sums of squared deviations (the clustering criterion).
3. Choose the change which leads to the greatest improvement in the clustering criterion.
4. Repeat steps 2 and 3 for all objects until no transfer of an object to a new group results in improvement in the clustering criterion.

The clustering criterion is to minimize the Euclidean sums of squared deviations of objects from the cluster mean, E_c which is defined as:

$$E_c = \sum_{i=1}^{n_c} \sum_{j=1}^{m} (Z_{ij} - Z_{cj})^2 \qquad (3.6)$$

where Z_{cj} is the mean value for cluster c of variable j and Z_{ij} I s the value for object i of variable j.

3.3.9 Distance Measures

The way in which clusters are identified is by a measurement of how close objects are in multidimensional space. This can be calculated by either a similarity or dissimilarity measure. A similarity measure (proximity) will report the largest value for the two objects that are closest together and the smallest value for the two objects that are furthest apart. Conversely a dissimilarity measure (distance) will report the smallest value for the two objects that are closest together and the largest value for the two objects that are furthest apart (Everitt *et al.* 2001). There are many different measures of both similarity and dissimilarity that can be used within cluster analysis. The suitability for use of a measure of similarity or dissimilarity depends on the specificities of the individual dataset. For example, different measures are more suited for different types of data. It would not be sensible to use the same measure for continuous, discrete and categorical datasets as these different types of data need to be treated in different ways. Things can get more complicated than this as there is no reason why continuous, discrete and categorical data cannot be used together in the same cluster analysis.

The remainder of this section will give a brief description of the distance measures considered for use in the project. However, there are many others available. Everitt *et al.* (2001) or Gordon (1999), give excellent overviews and descriptions of many different distance measures. Like clustering algorithms there are numerous distance measures, but only a few are commonly used and most are similar. Others are particular to specific applications. Probably the most commonly used distance measure is the Euclidean distance measure (Aldenderfer and Blasfield 1984). The Euclidean distance function measures the 'as-the-crow-flies' distance between a point $x(x_1x_2...x_n)$ and a point $y(y_1y_2...y_n)$. Calculating the Euclidean distance between two data points involves computing the square root of the sum of the squares of the differences between corresponding values. This is simply an extension of the Pythagoras theorem which give the distance between two points in n-dimensional space (Gordon 1999):

$$\text{distance}(x,y) = \{_i \sum (x_i - y_i)\}^{1/2} \qquad (3.8)$$

Squared Euclidean distance uses the same equation as the Euclidean distance metric, but does not take the square root. As a result, clustering with the Euclidean Squared distance metric is faster than clustering with the regular Euclidean distance. The distance scores are squared which enables the use of Increase in Sum of Squares, which minimizes the Euclidean Sum of Squares. The squared Euclidean distance measure is therefore better at handling larger and increasing numbers of objects (Everitt *et al.* 2001). Squared Euclidean distance helps convergence in large datasets. Euclidean distance does not always converge despite reaching the maximum number of iterations allowed within a clustering package or algorithm.

$$\text{distance}(x,y) =_i \sum (x_i - y_i)^2 \qquad (3.9)$$

K-Means clustering is not affected if Euclidean distance is replaced with Euclidean distance squared. However, the output of hierarchical clustering is likely to change (Everitt *et al.* 2001). It is therefore beneficial to use a Squared Euclidean distance measure when using k-means clustering especially when clustering large datasets as this will increase the speed of the analysis and increase the chances of the analysis reaching convergence. However, a Euclidean distance measure is preferable when using a hierarchical method of clustering. The classifications created in this project require only straight line distances to be measured between points; therefore a choice will be made between Euclidean distance and Squared Euclidean distance, dependent upon method of clustering.

3.4 Outputs

The outputs of an area classification are not just which cluster each area belongs to, but also a large amount of descriptive and explanative information, that is also required to produce a useful classification.

3.4.1 Selecting the Cluster Numbers and Classification Structure

One of the most difficult tasks in creating a classification is deciding what number of clusters will be the most suitable for use. This is especially difficult if there is no specific target number of clusters to be created, or if little or no information about the number of clusters expected to be present in the dataset. However, before the task of discerning how many clusters are present within the data it is important to consider the possibility that there are no naturally occurring clusters within the dataset (Milligan 1996).

There are several different rules of thumb that have been formulated to select the most suitable number of clusters. However, these can contradict each other within the same cluster analysis. Examples include:

- If you can’t choose between two solutions then the larger number of clusters should be selected.
- Select the cluster which shows the greatest reduction in the average distance from the solution with one fewer clusters, in a non-hierarchical system.
- Select the solution that shows the greatest increase in the average distance between the most dissimilar objects within merged clusters, in a hierarchical system.

- Select the solution which has the most suitable number of clusters for purpose.
- Select the solution which is most homogeneous in terms of the number of objects within each cluster, for example the solution which has the smallest difference between the number of objects in the smallest and largest clusters.

There is no right answer to the selection of the number of clusters; the selected solution will just be one of a number of possible representations. Therefore the solution selected should be judged as much on its usefulness in terms of cluster numbers, as being a correct representation of the patterns within the dataset. Hierarchical systems will require a structure to be given to the classification where the number of groups and the places where clusters are split to create another level of the hierarchy needs to be decided upon, as well as the initial number of clusters.

3.4.2 Naming the Clusters

The next step in the clustering process is to profile and name the clusters. The naming of the clusters is a near impossible task and one that always provokes much debate. However, it is a very important job, as if it is done wrongly it can give a false impression of the areas within a cluster.

Names and descriptions are a very contentious issue in geodemographic classifications. They can become an increasingly sensitive subject as the scale gets smaller and the classifications appear to be more person than area based. The names could and maybe should be seen as very much a side issue to the whole classification process as no matter what each cluster is called it does not alter the variable values of the cluster. However, many users of classifications use only the name to get an idea of what the clusters are like ignoring any additional information that is provided. The cluster names have also be easily picked up on by the media as they provide striking, but not always accurate headlines. Much of the criticism of geodemographics has been focused on the names of the groups. Make the name too specific and they only represent those areas very close to the centre of the same cluster. One could think of this as a form of the ecological fallacy. Users would think of the classification as being wrong as they find the very specific descriptions unrepresentative of the areas they are studying. Alternatively make the names too broad in an attempt to represent all of the areas that fall within a cluster and they become to vague and start to sound alike; a healthy balance needs to be found.

The commercial classifications available in the UK were slower than their American counter parts in giving their clusters catchy names. However, some systems have now embraced the use of snazzy eye catching names while others still have a very British way of naming their clusters. This can be seen clearly in the difference between the names in the Mosaic and Cameo systems.

Mosaic's names include such things as: *Global Connections, Fledgling Nurseries, Coronation Street, University Challenge* and *Pastoral Symphony* (Experían 2005). The Cameo names include the following: *Affluent Singles in Quality Rented Flats, Well off School Age Families in Semi-detached Properties, Younger Couples in Smaller Terraced Housing* and *Young Student Areas* (EuroDirect 2005). The distinction between the two in terms of their approach to naming clusters is clear. The Mosaic profiles (Experían) are designed to be creative, provocative and are perhaps a little inaccurate. The Cameo (EuroDirect) are more factual and duller. The names suggest little about the quality of the product, but they are indicative of the market each company is targeting. While the Mosaic names will be loved by a more style than substance advertising executive, the Cameo names would appeal to the more analytical minded spatial analyst. Whether this is a deliberate tactic from the two companies to target opposite ends of the market is unclear. What is clear is that the names matter and the two different approaches taken by Experían and EuroDirect in naming their clusters reflects not only on individual products, but on their businesses as a whole.

3.4.3 Pen Portraits, Maps, Photos and Visual Descriptions

The idea behind Pen Portraits is to create a short description, using text and variable information, which significantly expand on the names but can be understood simply in only a few minutes. Pen Portraits are intended to significantly expand the users understanding of the group without them having to trawl through the large amounts of variable information for each cluster. The profiles often include graphs, photos of typical homes or neighbourhoods and some statistical information along with an extended description of the clusters. Some of the recent releases of commercial systems have interactive portraits with sound as well as visuals and the ability for the user to find out an almost limitless number of statistics about each cluster. The most recent release of Experían's Mosaic system is particularly impressive in this regard see Experian (2004).

An area classification is a representation of areas and places and therefore the mapping of area classification should be seen as an essential output of the classification process. A vital part of understanding an area classification is mapping it, to see how the classification looks in reality. Mapping the classification brings it to life and the geography of the area classification can be truly fascinating. Unfortunately the geography of geodemographic classifications is often overlooked when put to such uses as profiling consumer records.

3.5 Conclusions

This chapter has provided a sequential overview of inputs, processes and outputs that form the backbone of cluster analysis, taking the researcher from census data in its rawest form to an area classification. Any classification, areal or otherwise, is never the correct answer nor is it the incorrect answer. There is no correct answer in creating a classification just a near infinite number of possible outcomes, based on decisions made during its creation.

Making a classification is a thankless task; criticisms are offered from many directions, by people who see the classification as a rival to their own product or somebody who has little understanding of the processes involved, or a misunderstanding of what the classification represents. However, to create a classification is of far more value than criticising an existing one. Milligan (1996) formulated a seven step approach to cluster analysis, which provides an excellent overview of the main decision points that have to be made in cluster analysis. Milligan's seven steps can be further summarised into a three stage sequence of inputs, processes and outputs as used by Harris (1999).

The discussion of inputs into a classification system covered many issues, discussing the availability and quality issues of both census data and other sources of demographic data. The contrasting theories of how much data to use in a classification system were discussed. The view portrayed by commercial geodemographics firms that the greater number of variables used in a classification system the better, this contrasts with the view portrayed in much of the cluster analysis literature that the fewest possible number of variables should be used, as adding more variables can cloud important patterns within the dataset. The different methods for comparing and reducing the number of variables in the dataset were described and assessed.

The processes involved in creating a classification were reviewed in detail. The descriptions began with the different forms of variable standardisation and the issue of variable weighting. The methods of classification used in the project were described in full, along with a review of other clustering methodologies not utilised. The distance measures that were used in the classification are reviewed and explained.

The chapter closed with a discussion of the outputs from an area classification. The importance of selecting a representative and practical number of clusters was stressed. The significance of the naming of the cluster solutions and the effect that they can have on the image of the classification were identified as important issues. The value of additional outputs such as pen portraits, maps and photographs to the user's understanding of what each of the clusters represent were drawn out. After understanding all the issues described in this chapter, a researcher is ready to start classifying!

Chapter Four - A Classification of the UK's Local Authorities

4.1 Introduction

The aim of this chapter is not to produce a local authority level classification for publication as a 'National Statistic'. However, the classification will be published as it will have value for further research. The main motivation behind the creation of this classification is as a precursor to the vastly more complex output area level classification (the creation of which is described in Chapter Five). Starting with an output area classification would be very much trying to run before having learnt to walk. By creating a local authority level classification, which is simpler and requires a smaller number of data points, enables an understanding of variables, techniques and issues relating to the clustering process to be identified. This will enable any difficulties or uncertainties to be identified and ironed out before attempting to create the output area classification. With ONS creating the official national classification of local authorities there is little point in using the same list of variables or using exactly the same methodology as the resulting classifications would be very similar. It would make sense to take where appropriate, decisions that make the classification as different as possible to the ONS classification and thus provides users with two alternatives to choose between or even use together.

Section 4.2 introduces the local authorities of the UK that are clustered to create the classification. Section 4.3 consists of a comparative evaluation of variables that have been used in previous classification systems, for which variable lists have been published. Section 4.4 outlines the variable choices that were made, reporting on the creation of an initial list of variables and how it was reduced by the merging and eliminating of variables. Section 4.5 describes the clustering process, the selection of the number of clusters and the hierarchy created. Section 4.6 describes the process of describing, rationalising and naming the clusters. Section 4.7 maps the clusters, showing the distinctive geography of each cluster type. Section 4.8 compares the classification created in this chapter with the official ONS classification of local authorities describing how the differing methodologies have produced different classifications. The strengths and weaknesses of both are also reflected upon. Section 4.9 concludes and evaluates the success of the classification and what has been learnt from the exercise. The classification created in this chapter was published as a working paper with detailed descriptions of the membership of each cluster and additional information (Vickers *et al*. 2003).

4.2 The Areas to be Clustered

The UK consists of 434 Local Authorities (LAs): 354 in England comprising: London Boroughs, Metropolitan Districts, Non-metropolitan Districts and Unitary Authorities, 22 in Wales which are all Unitary Authorities; 32 in Scotland which are all Council Areas and 26 in Northern Ireland which are all District council areas. These are the highest level of geography at which local government operates. There are slight differences in the administrative powers of the different types of authorities, but for the purposes of this investigation they will all be treated the same under the heading of Local Authorities.

LAs can vary greatly in population size and area as shown in Table 4.1, although the majority are of a similar size. The most populous LA in the UK is Birmingham, with a population of just under one million, the least populous being the Isles of Scilly with a population of just over two thousand, some five hundred times smaller. In terms of area the largest area is Highland, at just over two and a half million hectares, the smallest is the City of London at just under three hundred hectares, almost nine thousand times smaller. So despite being at the same scale of administrative geography there are significant differences of scale between some of the authorities.

Table 4.1: The variation in size of the UK's LAs in terms of population and area

Rank	LA Name	Population	Rank	LA Name	Area (hectares)
1	Birmingham	977,087	1	Highland	2,565,934
2	Leeds	715,402	2	Argyll & Bute	690,899
3	Glasgow City	577,869	3	Dumfries & Galloway	642,601
4	Sheffield	513,234	4	Aberdeenshire	631,259
5	Bradford	467,665	5	Perth & Kinross	528,581
430	Shetland Islands	21,988	430	Hammersmith & Fulham	1,640
431	Orkney Islands	19,245	431	Isles of Scilly	1,637
432	Moyle	15,933	432	Islington	1,486
433	City of London	7,185	433	Kensington and Chelsea	1,213
434	Isles of Scilly	2,153	434	City of London	290

Source: 2001 Census

LAs are comparatively large areas so there are several things that need to be taken into consideration when producing a classification. The scale and areal extent at which a classification system is produced is very important to allow it to be used, appropriately and accurately. A study that uses areal units will suffer from ecological fallacy and the MAUP, as discussed in § 2.7.1 and §2.7.2. The classifications produced will represent clustering based on the average value of each LA and the results of the classification should not be projected onto individuals within an area, but only used as a description of the area as a whole.

4.3 What went before? A comparative evaluation of variables used in previous classification systems

In the development of a new classification system it is important not to ignore what previous classifications have done, and decided works best. By examining the variables that have been used in previous classification systems, a base can be laid on which to build the foundations of the new classification system. In this section the variables used in several previous area classification systems (at various scales) will be explored and examined with the intention of creating a list of possible variables for inclusion in this new classification.

An evaluation of variables used in previous classifications is valuable, as it represents a conveyance of intellectual and conceptual thinking. Many variables in the 2001 Census were included in many previous censuses, and are therefore likely to have been used in for previous classifications. Unfortunately many commercial firms do not publish lists of variables from which their classification are created. Studies that have published the variables that make up their classifications rarely provide a detailed explanation or an audit trail detailing how the variables were chosen (Blake and Openshaw 1995).

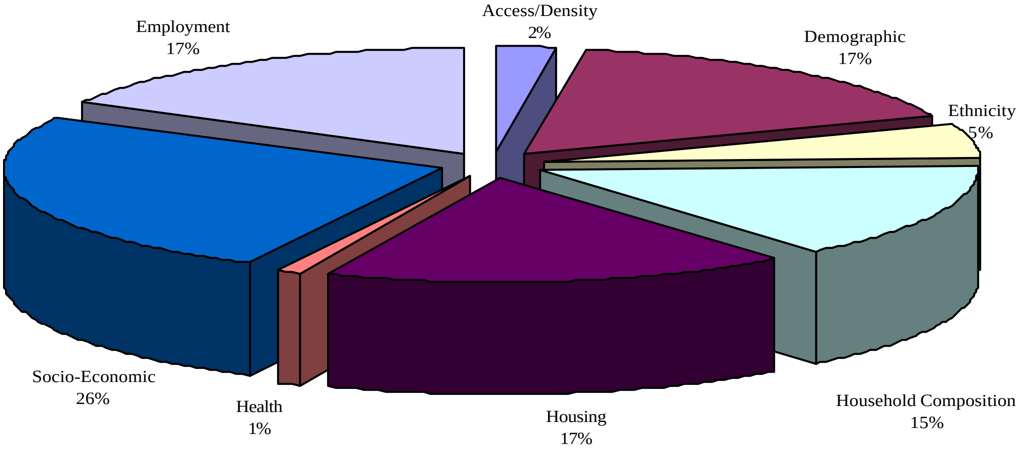
Ten classifications or reviews with variable lists from different dates and sources were identified. The classifications are listed in Table 4.2. They are from a variety of sources, commercial, public sector and academic. The classifications showed a great deal of variation in the number of variables they used, ranging from the OPCS 1981 classification, which contains just 35 variables to the Mosaic neighbourhood classification, which contains 137 variables.

Table 4.2: Sources and dates of the ten reviewed classifications

Name of Classification	Producer	Source of Data	Documentation	Variables
OPCS 1984	OPCS	1981 Census	Craig 1984	35
ONS 1996	ONS	1991 Census	Wallace and Denham. 1996	37
ONS 1999	ONS	1991 Census	Bailey *et al.* 1999b	37
ONS 2003	ONS	2001 Census	Higgs *et al.* 2002	62
GB Profiles 1995	University of Leeds	1991 Census	Blake and Openshaw 1995	84
Super Profiles 1994	Claritas	1991 Census /others	Batey and Brown 1995	118
Debenham 2001	University of Leeds	1991 Census	Debenham 2002	51
Mosaic (household)	Experían Ltd.	1991 Census/others	Experían 2001	59
Mosaic (neighbourhood)	Experían Ltd.	1991 Census/others	Experían 2001	137
Voas & Williamson 2001	University of Liverpool	1991 Census	Voas and Williamson 2001a	53

The classifications also vary greatly in the variables that are used. As there are so many different variables that have been used in the classifications, it was essential to group the variables to enable a meaningful comparison to be made between them. The purpose of the variable selection is to capture the complete spectrum of people’s lives, living arrangements and problems. Therefore the classification can be seen as a ‘socio-economic life course’ of the people, in which each person experiences a sequence of several parallel ‘careers’ during their lifetime. The variables used in the classifications can be split into separate domains each representing a different 'career' within the 'socio-economic life course'. Eight domains were identified, which represent different types of variables Some variables are not obviously exclusive to one domain and could be argued to belong to several domains, e.g. a variable named 'Indian, Pakistani, Bangladeshi and council rented' could be placed in Ethnicity or Housing. Figure 4.1 shows the distribution of the 673 variables used in the ten reviewed classifications split between the eight domains.

Figure 4.1: The percentage of each domain type used in the reviewed classifications



The domains vary significantly in size from the smallest, Health, which on average accounts for only 1% of the variables used in the ten reviewed classifications, to Socio-Economic, which accounts for 26% of the variables. However, the average number of variables in each domain from all ten classifications only shows part the picture, there is also significant variation in the proportion that each domain represents in each of the ten reviewed classifications. Figure 4.2 shows how certain domains dominate some of the classifications and how other domains don't appear in all the classifications. The Mosaic household classification comprises of only three domains (Demographic, Household Composition and Socio-Economic). The Health domain only appears in four of the ten classifications and the Access/Density domain only appears in Mosaic neighbourhood classification.

Figure 4.2: The percentage of each domain type used in each of the reviewed classifications

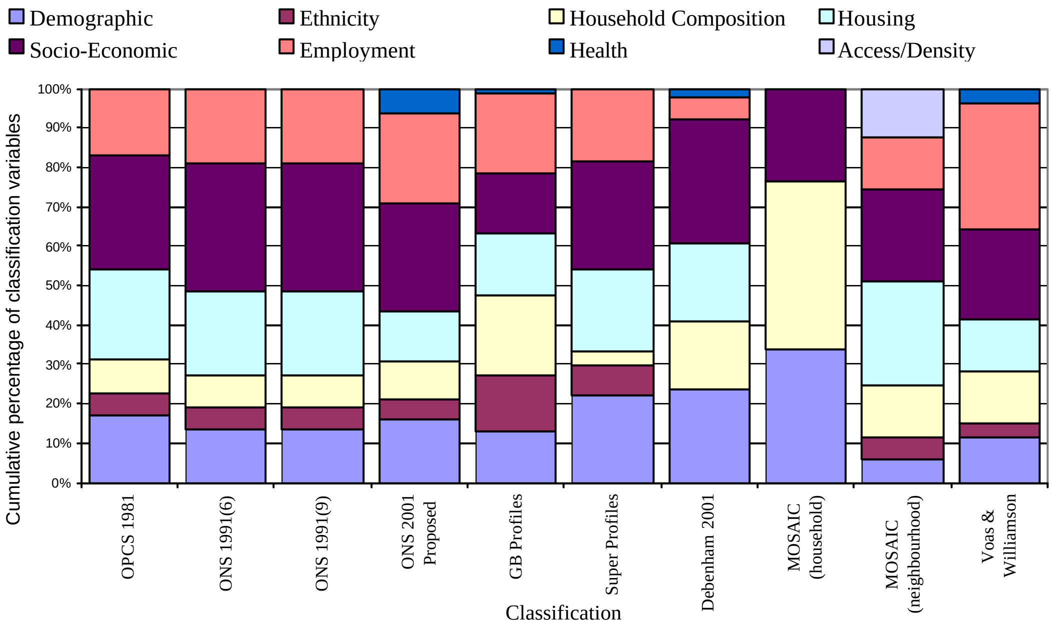


Figure 4.3 shows the range in the amount that each classification consists of variables from each domain. The blue blocks represent the least amount that each domain goes to making up the variable list of a classification. Figure 4.3 shows that only three domains (Demographic, Household Composition and Socio-Economic) appear in all ten classifications. At 39% Household Composition shows the biggest range in importance to the make-up of the classifications from 3.4% at the least to 42.4% at most. At 6.5% Health is the domain that shows the least variation ranging from 0% to 6.5%. The Socio-Economic domain shows itself to be regarded as important in all classifications because it has the highest minimum amount to make up a classification at 10.7% being its least contribution to any of the classifications.

Figure 4.3: The range of each domain type used in the reviewed classifications

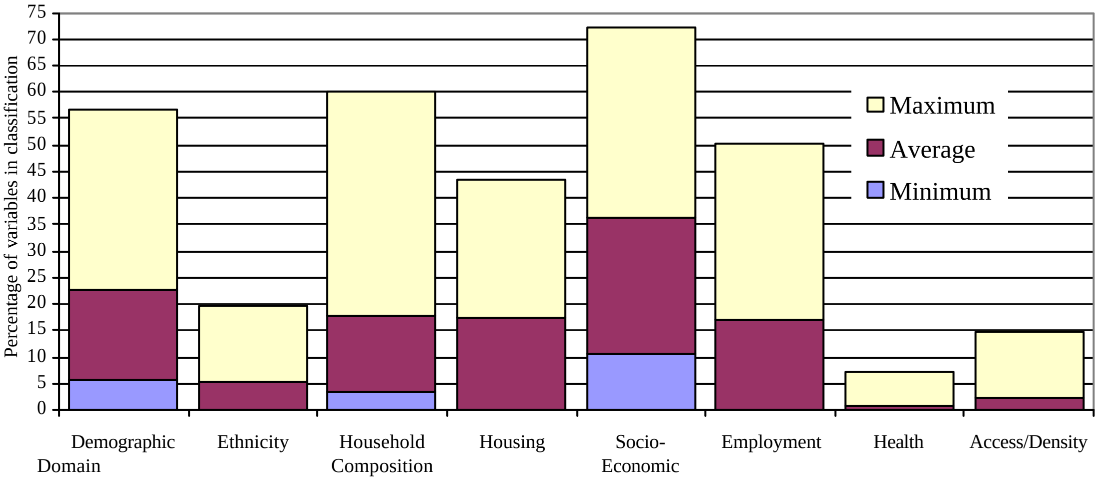


Table 4.3 lists the variables that appear in three or more of the reviewed classifications. No variable appeared in all ten classifications. However, seven different variables appear in nine of the ten classifications reviewed, of these seven variables five are demographic variables, one is

housing and one is socio-economic. A further seven variables are contained in eight out of the ten classifications. Ten variables are contained within seven classifications, five variables appear in six of the classifications, eight appear in five, eight appear in four, six appear in three, and 28 appear in two of the classifications. A further 308 (46%) variables only appear in one out of the ten reviewed classifications.

Table 4.3: Variables which are used in at least three of the ten reviewed classifications

Domain	Variable	Incidence
Demographic	Aged 0 - 4	9
Demographic	Aged 5 - 14	9
Demographic	Aged 25 - 44	9
Demographic	Aged 45 - 64	9
Demographic	Aged 65 +	9
Housing	LA Rented	9
Socio-Economic	Households with 2+ cars	9
Housing	purpose-built flats	8
Housing	Owner occupiers	8
Housing	Terraced houses	8
Socio-Economic	Public transport to work	8
Socio-Economic	No car households	8
Employment	Unemployment	8
Employment	Students	8
Demographic	Aged 15 - 24	7
Ethnicity	Black minority ethnic groups	7
Ethnicity	Indian, Pakistani or Bangladeshi	7
Household Composition	Households with 6/7+ Rooms.	7
Housing	Private Rented	7
Housing	No central heating	7
Socio-Economic	Moved in Last Year	7
Socio-Economic	HE qualification	7
Employment	Working Women ft	7
Employment	Agricultural employment	7
Housing	Detached housing	6
Housing	Semi-detached Housing	6
Employment	Mining and manufacturing employment	6
Employment	Services, government & defence employment	6
Health	Limiting long-term illness	6
Demographic	Couple	5
Household Composition	Household size	5
Household Composition	Large Families	5
Household Composition	Households >1.5 persons per room	5
Housing	Rooms per person	5
Socio-Economic	Professional households / Social Class I	5
Socio-Economic	Non manual households / Social Class II	5
Socio-Economic	Unskilled households / Social Class V	5
Demographic	Single	4
Ethnicity	Chinese	4
Household Composition	One-person no-pensioner households	4
Household Composition	Households with children / dependents	4
Housing	Bed sits	4
Socio-Economic	Skilled manual households/Social Class III	4
Socio-Economic	Semi-skilled households / Social Class IV	4
Employment	Self-employed	4
Household Composition	Single pensioner households	3
Housing	Mortgaged	3
Housing	Lacking bath and shower	3
Socio-Economic	Migrants	3
Socio-Economic	Lone parent households	3
Socio-Economic	Pensioner migrants	3

The classifications contain differing numbers of variables from 35 at the lowest end up to 137 at the highest end. As more variables are added to the classification the prevalence of the domains changes. Using the variables employed in the ten reviewed classifications as an example Figure

4.4 shows that by adding variables in accordance with their prevalence in the reviewed classifications the percentage make-up of a theoretical ‘amalgamated classification’ changes. Point ‘a’ consists of the seven variables that are in nine of the ten reviewed classifications. Point ‘b’ consists of the variables from point ‘a’, plus the variables that appear in eight of the ten reviewed classifications. The number of variables in the classification increases in this way until point ‘i’ where the 308 variables that appear in just one of the ten reviewed classifications are added.

Figure 4.4: The changing prevalence of each domain type with increasing number of variables

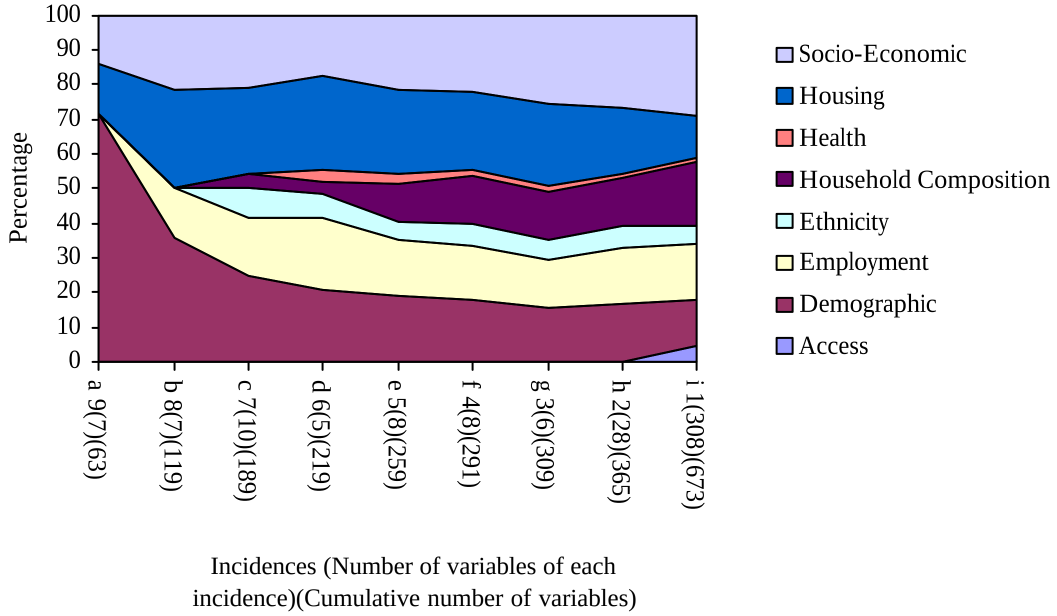


Figure 4.4 shows that demographic variables represent 71% of the information at point ‘a’ but by the time all the variables are added at point ‘i’ they represent only 12% of the data. Employment shows the opposite trend to Socio-Economic representing 0% of the data at point ‘a’ it then leaps up to 14% at point ‘b’ reaches a maximum of 21% at point ‘e’ it finishes with 14% at point ‘i’. This can be explained by the fact that most of the demographic variables used in the classifications are the same; they are always present no matter how few variables are in the classification. Apart from the demographic variables that are in nearly all the classifications, comparatively few other demographic variables are present in the classifications. The more variables that are added to a classification, the less the effect that the demographic variables have. As the number of variables in the classification increases, so does the type of domains covered. The increasing diversity of variables will affect the nature of the classification.

Reviewing variables used in previous classifications has proved to be an extremely useful exercise. It has provided invaluable background knowledge as to the sorts of variables that have produced successful classifications in the past, thus providing an indication as to which variables and variable domains are the most suitable for inclusion in a new classification.

4.4 Inputs

There is no standard method for the selection of variables and it is far from an exact science. Methods range from the most unscientific, which involve the minimum amount of statistical investigation to detailed statistical investigations. Variables can be selected based on the factors that are thought to be important and chosen on the basis of which best represent those factors. A better approach would be to use a series of statistical methods to aid the selection of variables. The knowledge gained by reviewing variables used in previous classification systems will also be very useful when selecting variables to be included in the classification.

4.4.1 Potential Variables to be used in the Classification

Before selecting variables to be used in the classification a list of possible variables needs to be compiled, this is a comprehensive list of variables containing far more than will eventually be used in the classification. From the comprehensive list, a final list of variables will be selected. The variables comprising the potential variables list are those used in two or more of the ten reviewed classifications. Added to this were variables from the census Key Statistics tables that covered areas that were missed by the reviewed variable lists. The additional variables included those on religion, which were added to the census in England and Wales for the first time in 2001 (since 1851), plus other variables that were included in the list of from the ONS 2001 classification list, as outlined in Higgs *et al.* (2002)

The variables that are used in a classification are vitally important because the results that the classification produces will be determined by the variables which were included (Blake and Openshaw 1995). For a classification to be comprehensive it should include variables from all domains (Demographic, Ethnicity, Household Composition, Housing, Socio-Economic, Employment and Health) as discussed in the previous section. What needs to be decided is how many variables of each domain should be included, and what those variables should be. A representative set of census based variable indicators needs to be created. The importance of each domain should be a general reflection of the original census questionnaire rather than that of the cross-tabulated counts.

After an intensive review a comprehensive list of 129 variables was complied, representing the majority of information within the census available at local authority scale. Table 4.4 displays a list of all 129 and the domain that each represents.

Table 4.4: The comprehensive list of 129 variables considered for use in the LA Classification

	Variable	Domain
1	Population Density	Demographic
2	Male	Demographic
3	Female	Demographic
4	Communal Establishments	Demographic
5	People aged: 0 - 4	Demographic
6	People aged: 5 - 7	Demographic
7	People aged: 8 - 9	Demographic
8	People aged: 10 - 14	Demographic
9	People aged: 15	Demographic
10	People aged: 16 - 17	Demographic
11	People aged: 18 - 19	Demographic
12	People aged: 20 - 24	Demographic
13	People aged: 25 - 29	Demographic
14	People aged: 30 - 44	Demographic
15	People aged: 45 - 59	Demographic
16	People aged: 60 - 64	Demographic
17	People aged: 65 - 74	Demographic
18	People aged: 75 - 84	Demographic
19	People aged: 85 - 89	Demographic
20	People aged: 90 & over	Demographic
21	Married (Living in Couple)	Demographic
22	Cohabiting	Demographic
23	Single (Never Married)	Demographic
24	Married (Not living in Couple)	Demographic
25	Separated	Demographic
26	Divorced	Demographic
27	Widowed	Demographic
28	Born in: England	Ethnicity & Religion
29	Born in: Scotland	Ethnicity & Religion
30	Born in: Wales	Ethnicity & Religion
31	Born in: Northern Ireland	Ethnicity & Religion
32	Born in: Republic of Ireland	Ethnicity & Religion
33	Born in: Other EU Countries	Ethnicity & Religion
34	Born Rest of the World (Outside EU)	Ethnicity & Religion
35	Black minority ethnic groups	Ethnicity & Religion
36	Indian, Pakistani or Bangladeshi	Ethnicity & Religion
37	Chinese	Ethnicity & Religion
38	White	Ethnicity & Religion
39	Christian	Ethnicity & Religion
40	Other Religion	Ethnicity & Religion
41	Not Stated or No Religion	Ethnicity & Religion
42	Limiting long-term illness	Health
43	Residents whose health is good	Health
44	Residents whose health is fairly good	Health
45	Residents whose health is not good	Health
46	Residents who provide unpaid care	Health
47	Unemployment	Employment
48	Self-employed	Employment
49	Economically active residents 16+	Employment
50	Male Unemployment	Employment
51	Working Women ft	Employment
52	Women who work part-time	Employment
53	Agriculture; hunting; forestry and fishing employment	Employment
54	Mining, quarrying and construction employment	Employment
55	Manufacturing employment	Employment
56	Electricity; gas and water supply employment	Employment
57	Wholesale & retail trade; repair of motor vehicles employment	Employment
58	Hotels and catering employment	Employment
59	Transport, storage and communication employment	Employment
60	Financial intermediation employment	Employment

	Variable	Domain
61	Real estate; renting and business activities employment	Employment
62	Public administration and defence employment	Employment
63	Education employment	Employment
64	Health and social work employment	Employment
65	Managers and senior officials employment	Employment
66	Professional occupations employment	Employment
67	Associate professional and technical occupations employment	Employment
68	Administrative and secretarial occupations employment	Employment
69	Skilled trades occupations employment	Employment
70	Personal service occupations employment	Employment
71	Sales and customer service occupations employment	Employment
72	Process; plant and machine operatives employment	Employment
73	Elementary occupations employment	Employment
74	No qualifications	Employment
75	Highest qualification attained level 1	Employment
76	Highest qualification attained level 2	Employment
77	Highest qualification attained level 3	Employment
78	Highest qualification attained level 4/5	Employment
79	Full time Students	Employment
80	Large employers and higher managerial occupations employment	Employment
81	Higher professional occupations employment	Employment
82	Lower managerial and professional occupations employment	Employment
83	Intermediate occupations employment	Employment
84	Small employers and own account workers employment	Employment
85	Lower supervisory and technical occupations employment	Employment
86	Semi-routine occupations employment	Employment
87	Routine occupations employment	Employment
88	Never worked	Employment
89	Long-term unemployed	Employment
90	Train to work	Socio-Economic
91	Bus, Mini Bus or Coach to work	Socio-Economic
92	Car to work	Socio-Economic
93	Motorcycle, Scooter or Moped to work	Socio-Economic
94	Walk to work	Socio-Economic
95	Bike to work	Socio-Economic
96	Work mainly from home	Socio-Economic
97	Purpose-built flats	Housing
98	Terraced houses	Housing
99	Detached housing	Housing
100	Semi-detached Housing	Housing
101	Bed sits	Housing
102	Households With no residents: Vacant	Housing
103	Households With no residents: Second residence / holiday home	Housing
104	Caravan or other mobile or temporary structure	Housing
105	Households with 3+ cars	Socio-Economic
106	Households with 2 cars	Socio-Economic
107	Households with 1 car	Socio-Economic
108	No car households	Socio-Economic
109	Average number of cars per household	Socio-Economic
110	LA Rented	Housing
111	Owner occupiers	Housing
112	Private Rented	Housing
113	Mortgaged	Housing
114	Household size	Housing
115	Rooms per household	Housing
116	No central heating	Housing
117	Lacking bath, shower and toilet	Housing
118	Households: with an occupancy rating of -1 or less (Overcrowding)	Household Composition
119	One-person no-pensioner households	Household Composition
120	Single pensioner households	Household Composition
121	Wholly student households	Household Composition
122	2 adults no children	Household Composition
123	Only Pensioner households	Household Composition
124	Households with dependent children	Household Composition
125	Lone Parent Families	Household Composition
126	Households: With one or more person with a limiting long-term illness	Household Composition
127	Households: No adults in employment :with dependent children	Household Composition
128	Male lone parents	Household Composition
129	Population change 1991 - 2001	Demographic

Migration data could not be used, because it had not yet been published for Northern Ireland. However, at the local authority level in Great Britain migration shows a significant positive correlation with population change. Therefore, a proportion of the information that the migration data contains will be represented in the population change variable. The classification may be updated in the future to include migration variables when they have been published for Northern Ireland. The 129 variables in Table 4.4 need to be assessed in terms of how much information they contain about the areas and the inter correlations within them.

4.4.2 Variable Selection Procedure

Several different methods were used to investigate the suitability of each variable. Principal components analysis (PCA) as described in § 3.2.8 was used to establish which variables had the strongest influence on the dataset. A correlation matrix was used to identify and remove high levels of correlation within the dataset. The variance for each variable was also examined to establish the extent to which the value of each variable varies across space.

The component loadings matrix produced by the PCA was studied first; this is a matrix showing how much of the variance of a variable was accounted for by each component. Variables that have a large amount of their variance covered by the early components will be those variables that are likely to have the most significance within the data and hence drive the classification. The component loadings of the first five principal components for the variables that have the greatest amount of their variance accounted for by component I are shown in Table 4.5. The component loading is the correlation between a variable and a component. Variable 13 (People aged 25 -29), is the variable that has the largest amount of variance explained by principal component I, the correlation of variable 13 with principal component I is 0.89, therefore 79% $((0.89^2)\times100)$ of the variance of variable 13 is explained by component I. This suggests that variable 13 has a strong influence on component I and hence a strong influence on the classification. Table 4.6 shows the component loadings for the second principal component. The variable that has the greatest amount of its variance associated with component II is variable 65 (Managers and senior officials employment), which has a loading of -0.88 so 77% of the variance of variable 65 is explained by component II. Table 4.7 shows the component loadings for the third principal component. The variable that has the greatest amount of its variance explained by component III is variable 124 (Households which contain dependent children), which has a loading of -0.89 therefore 79% of the variance of variable 124 is explained by component III. Although the first few variables with the highest loadings for components I, II and III are all of similar value after this the loadings for component III decline quicker than I and II. This is because the later components are weaker than the early ones, with the first component being the strongest.

Table 4.5: First ten Rows of component I

Variable Number	Variable Name	loading
13	People aged: 25 - 29	0.89
118	Households: with an occupancy rating of -1 or less	0.88
37	Chinese	0.88
119	One-person no-pensioner households	0.87
34	Born Rest of the World (Outside EU)	0.86
1	Population Density	0.86
21	Married (Living in Couple)	-0.86
92	Car to work	-0.85
23	Single (Never Married)	0.84
24	Married (Not living in Couple)	0.82

Table 4.6: First ten Rows of component II

Variable Number	Variable Name	loading
65	Managers and senior officials employment	-0.88
126	Households: With one or more person with LLTI	0.86
45	people who's health is not good	0.86
127	Households: No adults in employment with dependent children	0.86
74	residents age 16 - 74 with no qualifications	0.85
50	residents who are 16+ and male and unemployed	0.85
47	Unemployment	0.84
125	one parent households as a percentage of households containing children	0.84
89	Long-term unemployed)	0.83
80	Large employers and higher managerial occupations	-0.82

Table 4.7: First ten Rows of component III

Variable Number	Variable Name	loading
124	Households which contain dependent children	-0.89
114	The Average Number of people per household	-0.86
120	Single pensioner households	0.82
6	people aged 5 - 7	-0.81
122	Households which contain 2 adults no children	0.79
19	people aged 85 - 89	0.74
18	people aged 75 - 84	0.74
7	people aged 8 - 9	-0.74
5	people aged 0 - 4	-0.74
20	people aged 90+	0.71

As well as establishing which variables power the dataset it is important to consider the correlations between variables. There is no sense in having two highly correlated variables as they provide the much of same information. There are two different types of correlation between variables. Variables that are positively correlated represent characteristics of people that are likely to be present due to the type of person that they are, e.g. a student is likely to be in their late teens or early twenties, therefore, the full time student variable will be positively correlated with variable aged 18-24. Negative correlations occur between variables that represent characteristics that are unlikely to be present in a person. For example, people over 65 years of age are highly unlikely to be full time students, and therefore these two variables will have a high negative correlation. Negative correlations can also appear between variables within the same domain. An example of this is age groups. Age groups at opposite extremes i.e. young and old will be negatively correlated, as an individual can only be of one age and therefore can only be in one of the groups. An area with a high number of old people is therefore likely to have a low number of young people. Figure 4.5 shows a section of the correlation matrix, which correlates all 129 variables against each other. The selection shows the 16 age variables considered for inclusion in the classification. Highlighted are two examples of different types of correlation within the dataset. Highlighted in yellow is a very high positive correlation of 0.97 between people aged 85-89 and people aged 90+. This suggests that people in these age groups are likely to live in the same areas, with one able to explain 94% of the variance within the other. Highlighted in green is a high negative correlation of -0.80 between people aged 60-64 and people aged 30-44 this suggests that people in these age groups are unlikely to live in the same areas. In spite of this they can explain 64% of the variance within each other, but this time based on absence rather than presence. Highlighted in blue are two sets of correlations between people aged 18-19 and people aged 5-7, and between people aged 18-19 and people aged 8-9.In both cases the correlation is very weak (0.05). This shows that virtually none of the variance of one variable can be explained by the other. These simple examples can be easily identified. However, understanding the complexities of the relationships between 129 inter-correlated variables is a much greater task and requires much consideration, careful examination, checking and rechecking.

Figure 4.5: Correlation matrix of age variables

0-4	5-7	8-9	10-14	15	16-17	18-19	20-24	25-29	30-44	45-59	60-64	65-74	75-84	85-89	90+	Age
X	0.84	0.68	0.55	0.45	0.47	0.15	0.27	0.49	0.56	-0.60	-0.75	-0.77	-0.72	-0.66	-0.61	0-4
	X	0.91	0.81	0.70	0.67	0.05	-0.04	0.07	0.29	-0.31	-0.45	-0.51	-0.55	-0.57	-0.53	5-7
		X	0.90	0.79	0.74	0.05	-0.13	-0.10	0.14	-0.18	-0.29	-0.37	-0.44	-0.52	-0.48	8-9
			X	0.92	0.88	0.09	-0.20	-0.25	-0.08	-0.07	-0.13	-0.22	-0.31	-0.41	-0.39	10-14
				X	0.92	0.16	-0.15	-0.24	-0.14	-0.07	-0.09	-0.18	-0.28	-0.38	-0.37	15
					X	0.21	-0.09	-0.17	-0.13	-0.11	-0.14	-0.23	-0.32	-0.39	-0.37	16-17
						X	0.81	0.29	-0.08	-0.54	-0.43	-0.36	-0.30	-0.29	-0.29	18-19
							X	0.72	0.27	-0.78	-0.69	-0.58	-0.45	-0.37	-0.37	20-24
								X	0.69	-0.78	-0.80	-0.76	-0.64	-0.52	-0.49	25-29
									X	-0.42	-0.70	-0.78	-0.78	-0.68	-0.62	30-44
										X	0.79	0.61	0.47	0.41	0.39	45-59
											X	0.91	0.74	0.62	0.57	60-64
												X	0.91	0.78	0.73	65-74
													X	0.93	0.88	75-84
														X	0.97	85-89
															X	90+

In addition to considering which variables power the dataset and the correlation between the variables, the variance of the variable across all local authorities can also be considered. One way of doing this is to compare the standard deviation of each variable, this will reveal the variables that show the biggest differences between the LAs. A variable that shows a large variation of values is more useful than a variable that shows little variation because the clustering should be based on values that show a larger distinction between areas. The variables with the highest and lowest standard deviation can be seen in Table 4.8, this shows how different the standard deviation can be for each variable ranging from 31.54 to 0.14.

The variable that shows the greatest variation is people born in England, with the equivalent variables for the other three countries in of UK not far behind. This suggests that there is a great deal of geographic variation within these variables. This is perhaps not surprising, because although the internal borders within the UK are little more than lines on a map people are still in general more likely to live in the country in which they were born rather than another country of the UK. Other variables that show high levels of variation are predominately related to housing and cars. The variables that show the smallest variation within them include household size, Chinese and several of the age groups. However, values especially at this end of the scale need to be treated with caution as the variation within a group is also affected by its prevalence across that UK not just the variation within it. This is a good thing in terms of selecting variables for the classification, because variables that include a large proportion of the classification are more representative and safer to work with than variables that only contain a small amount of the population.

Table 4.8: Variables with the highest and lowest standard deviation across all LAs

Largest Std. Deviation			Smallest Std. Deviation		
Rank	Variable	S.D.	Rank	Variable	S.D.
1	Born in: England	31.54	129	Household size	0.14
2	Born in: Scotland	22.45	128	People aged: 15	0.16
3	Average cars per household	22.28	127	People aged: 90+	0.22
4	Born in: Northern Ireland	21.63	126	People aged: 8 - 9	0.25
5	Population Density	18.74	125	People aged: 16 - 17	0.30
6	Born in: Wales	16.37	124	Chinese	0.34
7	Detached housing	13.87	123	Lacking bath, shower and toilet	0.36
8	Purpose-built flats	10.84	122	People aged: 85 - 89	0.36
9	Car to work	10.80	121	People aged: 5 - 7	0.37
10	Terraced houses	9.63	120	Motorcycle/scooter/moped to work	0.39
11	No car households	9.41	119	Electricity/gas/ water supply employ	0.41
12	Owner occupiers	9.01	118	Rooms per household	0.44
13	White	8.70	117	Long-term unemployed	0.49
14	Christian	8.48	116	People aged: 18 - 19	0.49
15	Semi-detached Housing	8.43	115	Caravan/mobile/temporary home	0.51

When selecting variables for input into a classification, it is sensible to use information from all of the different methods of selection outlined, because using just one method you can make a case for the inclusion of most variables. For example Chinese has 88% of its variance represented by Principal Component I suggesting that it could be an important predictive variable, but it has the 6th lowest standard deviation showing that it varies very little between local authorities. Chinese is unlikely to add significant value to the classification in terms of separating authority areas into clusters of similarity.

It is also important to consider which variable domains are covered by the variables selected. The variables within the classification were split in seven domains which represent different types of variables. The seven domains covered by the variables have been named: Demographic, Employment, Ethnicity & Religion, Household Composition, Health, Housing and Socio-Economic. Variables from each domain need to be included in the final variable list, to ensure that different types of data, representing different characteristics of the people who live within each local authority are included.

The methods discussed are all valuable ways of evaluating the potential of variables to be included in the classification. However, the results of these evaluations should not be accepted without considering what the variables actually represent, a mean value. The variables that perform best in the various forms of analysis are not guaranteed to be the most relevant variables. These variables may only be representative of minor trends, although variables representing more important trends may have not been shown to be as predictive in the analysis. Variables representing more important social trends should be included and are likely to add more relevance to the classification.

The process of selecting the final set of variables is therefore one that is far from straightforward and the cause of much debate. If two people were asked to construct a list of variables for creating an area classification, although they are likely to have broad overlap, they are unlikely to be the same. Great consideration was given to the selection or rejection of all the variables considered for use in the classification. It is impossible to convey the entirety of the information that was used and the reasons behind all the decisions made. However, an attempt has been made to provide a brief account of whether a variable was: selected for the classification, was merged with other variables and then changed in some way before being included, or was rejected and not used in the classification.

After all the criteria for reducing the variable list had been considered a final list of 56 variables was produced. The variables along with the reason behind their inclusion or non inclusion are listed in Table 4.9. The table has a traffic light colour coding system where green = go = included, amber = wait = merged and red = stop = rejected. The decisions made and the reasons behind them are described in Table 4.9. There are some general trends and patterns in the decisions can be seen. For example, many of the age variables were merged because individually they do not contain a large proportion of the population, and many of them were shown to be highly correlated. It therefore makes sense to merge these variables so that the data are kept, but without increasing the number of variables in the cluster analysis.

Table 4.9: The list of inclusion, rejection or merger of variables

	Variable	Reason for inclusion, rejection or merger
1	Population Density	Included - As it is unlike any other variable giving a good in indication of the rural/urban variation of the country. It also has a very large variance.
2	Male	Rejected - No variation across the dataset
3	Female	Rejected - No variation across the dataset
4	Communal Establishments	Rejected - Their location is sporadic and not indicative of the population of the area.
5	People aged: 0 – 4	Merged - With 6 & 7 due to high positive correlation
6	People aged: 5 – 7	Merged - With 5 & 7 due to high positive correlation
7	People aged: 8 – 9	Merged - With 5 & 6 due to high positive correlation
8	People aged: 10 – 14	Merged - With 9 & 10 due to high positive correlation
9	People aged: 15	Merged - With 8 & 10 due to high positive correlation
10	People aged: 16 – 17	Merged - With 8 & 10 due to high positive correlation
11	People aged: 18 – 19	Merged - With 12 due to high positive correlation
12	People aged: 20 – 24	Merged - With 11 due to high positive correlation
13	People aged: 25 – 29	Included - A good indicative group, representing first time buyers.

	Variable	Reason for inclusion, rejection or merger
14	People aged: 30 – 44	Rejected - Little variation across the dataset. However, pseudo included as the rest of the variance in the age category is included
15	People aged: 45 – 59	Merged - With 16 due to high positive correlation
16	People aged: 60 – 64	Merged - With 15 due to high positive correlation
17	People aged: 65 – 74	Merged - With 18,19 & 20 due to high positive correlation
18	People aged: 75 – 84	Merged - With 17,19 & 20 due to high positive correlation
19	People aged: 85 – 89	Merged - With 17,18 & 20 due to high positive correlation
20	People aged: 90 & over	Merged - With 17,18 & 19 due to high positive correlation
21	Married (Living in Couple)	Merged - With 24
22	Cohabiting	Rejected – Indicates little, small variance across areas
23	Single (Never Married)	Included – Indicative of a mobile population
24	Married (Not living in Couple)	Merged - With 21
25	Separated	Rejected - Indicates little, small variance across areas
26	Divorced	Rejected - Indicates little, small variance across areas
27	Widowed	Rejected - Indicates little, small variance across areas
28	Born in: England	Rejected - Does little except split countries of the UK
29	Born in: Scotland	Rejected - Does little except split countries of the UK
30	Born in: Wales	Rejected - Does little except split countries of the UK
31	Born in: Northern Ireland	Rejected - Does little except split countries of the UK
32	Born in: Republic of Ireland	Merged - With 33 & 34
33	Born in: Other EU Countries	Merged - With 32 & 34
34	Born Rest of the World (Outside EU)	Merged - With 32 & 33
35	Black minority ethnic groups	Included - High variance, strong distinction in numbers between rural and urban areas
36	Indian, Pakistani or Bangladeshi	Included - High variance, strong distinction in numbers between rural and urban areas
37	Chinese	Rejected - Little variation across the dataset
38	White	Rejected - Pseudo Included as the rest of the variance in the ethnicity category is included
39	Christian	Included - Considered important to include as it is the first time the religion question was asked in the census. Also shows some significant regional differences.
40	Other Religion	Included - Considered important to include as it is the first time the religion question was asked in the census. Also shows some significant regional differences.
41	Not Stated or No Religion	Rejected - Pseudo Included as the rest of the variance in the religion category is included
42	Limiting long-term illness	Included - Considered important as a measure of the health
43	Residents whose health is good	Included - Considered important as a measure of the health of the nation. Also the other extreme to LLTI giving a fuller picture of the health of the nation.
44	Residents whose health is fairly good	Rejected - Vague in its nature, however pseudo included as the extremes of the variance in the health category is included.
45	Residents whose health is not good	Rejected - Vague in its nature, however pseudo included as the extremes of the variance in the health category is included.
46	Residents who provide unpaid care	Included - An alternative measure of the nation's health
47	Unemployment	Included - An important measure in the employment domain
48	Self-employed	Rejected - Vary Similar to 84
49	Economically active residents 16+	Included - A good indication of the size of the workforce in an area taking into account all factors.
50	Male Unemployment	Included - Indicative of a more extreme problem than total unemployment as men are more likely to be the sole or main wage earner in a household.
51	Working Women ft	Included - An indication of the changing employment structure of the UK as more women continue to join the workforce.
52	Women who work part-time	Included - An indication of the changing employment structure of the UK as more women continue to join the workforce.
53	Agriculture; hunting; forestry and fishing employment	Included - High distinction between rural and urban areas
54	Mining, quarrying and construction employment	Rejected -Too specific
55	Manufacturing employment	Rejected - Too specific
56	Electricity; gas and water supply employment	Rejected - Too specific
57	Wholesale & retail trade; repair of motor vehicles employment	Rejected - Too specific
58	Hotels and catering employment	Rejected -Too specific

	Variable	Reason for inclusion, rejection or merger
59	Transport, storage and communication employment	Rejected -Too specific
60	Financial intermediation employment	Rejected -Too specific
61	Real estate; renting and business activities employment	Included - Indicative of areas of business and a buoyant housing market.
62	Public administration and defence employment	Rejected - Too specific
63	Education employment	Rejected - Too specific
64	Health and social work employment	Rejected - Too specific
65	Managers and senior officials employment	Included - Indicative of the wealthiest people within society
66	Professional occupations employment	Rejected - Too specific
67	Associate professional and technical occupations employment	Rejected - Too specific
68	Administrative and secretarial occupations employment	Rejected - Too specific
69	Skilled trades occupations employment	Rejected - Too specific
70	Personal service occupations employment	Rejected - Too specific
71	Sales and customer service occupations employment	Rejected - Too specific
72	Process; plant and machine operatives employment	Rejected - Too specific
73	Elementary occupations employment	Rejected - Too specific
74	No qualifications	Included - Indicative of poorer areas, and people with a poor education
75	Highest qualification attained level 1	Rejected - Indicates little, but Pseudo Included as the extremes of the variance in the education category is included.
76	Highest qualification attained level 2	Rejected - Indicates little, but Pseudo Included as the extremes of the variance in the education category is included.
77	Highest qualification attained level 3	Rejected - Indicates little, but Pseudo Included as the extremes of the variance in the education category is included.
78	Highest qualification attained level 4/5	Included - Indicative of the richest areas, and people with a very good education
79	Full time Students	Included - A large and important group within the modern society
80	Large employers and higher managerial occupations employment	Included - Indicative of the top end of the employment ladder.
81	Higher professional occupations employment	Included - Indicative of the top end of the employment ladder.
82	Lower managerial and professional occupations employment	Included - Indicative of the top end of the employment ladder.
83	Intermediate occupations employment	Rejected - The middle rung on the employment ladder, little variance and indicates little.
84	Small employers and own account workers employment	Included - Self employed a significant proportion of the workforce as yet not included.
85	Lower supervisory and technical occupations employment	Rejected - The lower middle rung on the employment ladder, little variance and indicates little.
86	Semi-routine occupations employment	Rejected - The lower middle rung on the employment ladder, little variance and indicates little.
87	Routine occupations employment	Included - Indicative of the bottom end of the employment ladder.
88	Never worked	Included - Indicative of a more serious unemployment problem, picks out deprived areas with a significant lack of employment.
89	Long-term unemployed	Included - Indicative of a more serious unemployment problem, picks out deprived areas with a significant lack of employment.
90	Train to work	Rejected - Small numbers in some areas
91	Bus, Mini Bus or Coach to work	Rejected - Small numbers in some areas
92	Car to work	Included - Indicative of the commuter, high variance
93	Motorcycle, Scooter or Moped to work	Rejected - Small numbers in some areas, little variation
94	Walk to work	Included - A contrast to 92
95	Bike to work	Rejected - Small numbers in some areas
96	Work mainly from home	Rejected - Small numbers in some areas

Variable		Reason for inclusion, rejection or merger
97	Purpose-built flats	Included - Housing type is indicative of the type and standing of people who live in an area
98	Terraced houses	Included - Housing type is indicative of the type and standing of people who live in an area
99	Detached housing	Included - Housing type is indicative of the type and standing of people who live in an area
100	Semi-detached Housing	Rejected - Pseudo Included as the rest of the variance in the housing category is included
101	Bed sits	Included - Housing type is indicative of the type and standing of people who live in an area
102	Households With no residents: Vacant	Rejected - Very small numbers in some areas
103	Households With no residents: Second residence / holiday home	Included - Indicative of areas where tourism is an important industry. An industry which is of increasing importance to the UK economy.
104	Caravan or other mobile or temporary structure	Rejected - Little variance across areas.
105	Households with 3+ cars	Merged - With 106, Indicative of wealth
106	Households with 2 cars	Merged - With 105, Indicative of wealth
107	Households with 1 car	Rejected - Pseudo Included as the rest of the variance in the car category is included
108	No car households	Included - Indicative of deprivation
109	Average number of cars per household	Rejected - Covered by previous variables, highly correlated with 105-108.
110	LA Rented	Included - Shows areas with a large amount of council renting, indicative of the poorer end of society.
111	Owner occupiers	Rejected - Little variance, Pseudo Included as if it is not rented it must be owner occupied
112	Private Rented	Included - Indicative of a young mobile population
113	Mortgaged	Rejected - Little variance
114	Household size	Included - Gives a good
115	Rooms per household	Rejected - Covers the information in 119 plus a bit more
116	No central heating	Included - Variation between regions especially urban/rural
117	Lacking bath, shower and toilet	Rejected - Small numbers, little variance.
118	Households: with an occupancy rating of -1 or less (Overcrowding)	Included - An indication of poverty
119	One-person no-pensioner households	Rejected - Covered to a large extent by 119
120	Single pensioner households	Included - Shows areas with a lot of elderly residents, especially coastal resorts.
121	Wholly student households	Rejected - Highly correlated with 79
122	2 adults no children	Included - The opposite to single parent families an indicator of wealth.
123	Only Pensioner households	Rejected - Highly correlated with 120 and age groups
124	Households with dependent children	Included - Gives a distinction between the number of children in an area. An indication as to the make up of the population structure of an area.
125	Lone Parent Families	Included - An indication of lower levels of wealth and a changing family structure.
126	Households: With one or more person with a limiting long-term illness	Rejected - Highly correlated with 42
127	Households: No adults in employment :with dependent children	Included - Indicative of poverty, especially within children.
128	Male lone parents	Rejected - Too Specific
129	Population change 1991 – 2001	Included - An indication of the growth of an area. Also highly correlated with migration, Information that was not available at the time that the classification work was done for the whole of the UK

4.4.3 Final Variable List

In general an attempt was made to reduce the list of 129 variables as much as possible whist losing as little of the information as possible. To do this variables that included larger percentages of the population have been treated as the most important variables. The final list of 56 variables used can be found in Table 4.10; this table represents the final set of variables to be used in the classification and a reduction of 57% on the original list of 129.

Table 4.10: The final list of 56 variables to be used in the classification

	Variable	Domain
1	Population Density	Demographic
2	People aged: 0 - 9	Demographic
3	People aged: 10 - 17	Demographic
4	People aged: 18 - 24	Demographic
5	People aged: 25 - 29	Demographic
6	People aged: 45 - 64	Demographic
7	People aged: 65+	Demographic
8	Married	Demographic
9	Single (Never Married)	Demographic
10	Born outside UK	Ethnicity & Religion
11	Black minority ethnic groups	Ethnicity & Religion
12	Indian, Pakistani or Bangladeshi	Ethnicity & Religion
13	Christian	Ethnicity & Religion
14	Other Religion	Ethnicity & Religion
15	Limiting long-term illness	Health
16	Residents whose health is good	Health
17	Residents who provide unpaid care	Health
18	Unemployment	Employment
19	Economically active residents 16+	Employment
20	Male Unemployment	Employment
21	Women who work Full-time	Employment
22	Women who work Part -time	Employment
23	Agriculture; hunting; forestry and fishing employment	Employment
24	Real estate; renting and business activities employment	Employment
25	Managers and senior officials employment	Employment
26	No qualifications	Employment
27	Highest qualification attained degree level or above	Employment
28	Full time Students	Employment
29	Large employers and higher managerial occupations employment	Employment
30	Higher professional occupat ions employment	Employment
31	Lower managerial and professional occupations employment	Employment
32	Small employers and own account workers employment	Employment
33	Routine occupations employment	Employment
34	Never worked	Employment
35	Long-term unemployed	Employment
36	Car to work	Socio-Economic
37	Walk to work	Socio-Economic
38	purpose-built flats	Housing
39	Terraced houses	Housing
40	Detached housing	Housing
41	Bed sits	Housing
42	Households With no residents: Second residence / holiday home	Socio-Economic
43	Households with 2+ cars	Socio-Economic
44	No car households	Socio-Economic
45	LA Rented	Housing
46	Private Rented	Housing
47	Household size	Household Composition
48	No central heating	Housing
49	Households: with an occupancy rating of -1 or less (overcrowding)	Household Composition
50	One-person no-pensioner households	Household Composition
51	Single pensioner households	Household Composition
52	2 adults no children	Household Composition
53	Households with dependent children	Household Composition
54	Lone Parent Families	Household Composition
55	Households: No adults in employment :with dependent children	Household Composition
56	Population change 1991 - 2001	Demographic

The ONS LA classification contains only 42 variables; these will be discussed further in § 4.8 where the two classified are compared.

4.5 Processes

After the final variable list had been constructed the process of creating the classification could begin. This firstly involved standardising the variables to account for difference in scale between the variables. Then clustering the data using cluster analysis techniques to produce the structure of the classification and split the areas into groups of similarity.

4.5.1 Variable Standardisation

The method chosen for standardising the variables was to transform them into z-scores. Z-score standardisation is described and explained in § 3.3.2. Z-scores are based on standard deviation away from mean so they provide a good measure of variance, while allowing some of the extent of extreme values to remain in the dataset, which may be needed to differentiate between clusters at this coarse spatial scale. Another reason for selecting z-scores as the standardisation method for the LA classification was that the ONS chose to use a different method for their equivalent classification. One of the aims of the classification is to produce a classification to complement rather than to match the ONS classification, so the use of a different method of standardisation will help with this.

4.5.2 Clustering Method

The method that was used for clustering the variables was Ward's Hierarchical Grouping Procedure. The classification is to take the form of a three tier hierarchy to provide not just one, but three integrated classifications which fit inside each other. Ward's method has been shown to work very successfully in previous classifications at this scale such as Bailey *et al.* (1999). This method, although difficult to run on larger datasets, will provide a quick and easy clustering method at the local authority scale, where there are only 434 units to be clustered. Ward's method gives the user a great deal of information about the clusters, which helps in choosing the most suitable number of clusters. Ward's classification is explained in full in § 3.3.7. The cluster analysis was run using SPSS statistical package, which provides an easy to use interface with the ability to choose between a number of clustering methods and preset methods of standardisation. A detailed description of how the clustering algorithms in the SPSS system work can be found in SPSS Inc. (2001). The distance measure used for the LA classification using Ward's method was the Euclidean distance measures as recommended by Kaufman and Rousseeuw (2005) and Everitt *et al.* (2001). The Euclidean distance measure and others are explained in full in § 3.3.9.

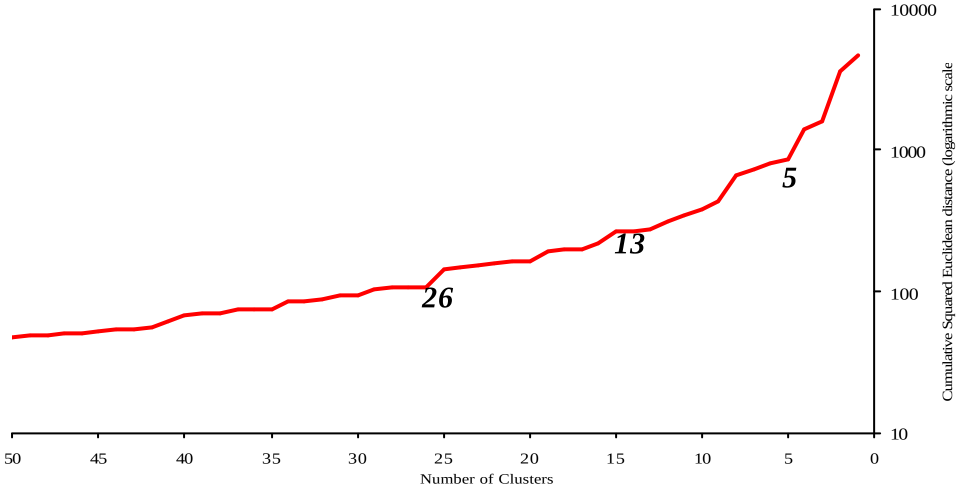
Once the variables have been clustered the next decision that has to be made is how many clusters to split the LAs into. Unlike other methods of clustering such as k-means, the Ward's method of clustering does not have to be provided with a number of clusters before the process of clustering. Instead a range of solutions are produced, from 434 clusters where all LAs are in separate groups, to just two clusters. In total this gives 433 different classifications of the LAs therefore a method of selecting the most suitable number of clusters to use is needed. It is also important to remember that the clustering procedure is hierarchical so a multiple level classification system can be produced.

The ONS classification of Great Britain's local authorities using 1991 data produced a three tier hierarchy of 27, 15 and 7 clusters (Bailey *et al.* 1999). Using the ONS classification as a guide to a suitable solution. The aim will be to produce a three tier hierarchy with the number of clusters more or less doubling with each tier hopefully ending in the tier with between 25 – 30 clusters (e.g. 28, 14 and 7).

Knowing the structure of classification and the number of clusters that would work best theoretically is one thing. However, this does not mean that this will be the must suitable number of clusters in reality. The method used to choose the number of clusters was to examine the relative increase in the sum of squares. The sum of squares increases as every object is joined to a cluster, reducing the number of clusters from *n* to *n*-1. The clusters that are suitable for selection are those that where the sum of squares shows a sharp increase between that level, and the next level of clusters, therefore being the most compact clusters relative to the number of clusters that have been formed (Everitt *et. al.* 2001).

Figure 4.6 shows how the three tiers of the classification were chosen; the graph shows the increase in the sums of squares with the reduction of the number of clusters in the solution. Moving left to right across the x axis of the graph the first sharp rise comes at 26 clusters. This falls within the 25-30 cluster target, so 26 was chosen as the first level of the hierarchy. Moving on there are rises at 20 and 17, but these do not represent a big enough difference to the higher tier of 26 clusters. There is another rise at 13 which is exactly half the number of clusters chosen for the first tier so would be an ideal selection for the middle tier of the hierarchy. Moving on there is an increase at 9, but again this is too close to the previous tier (13 clusters) and would not provide enough distinction between the two tiers. The next significant is rise at 5, this provides a much better final tier of the hierarchy and was therefore selected.

Figure 4.6: Increase in distance between the most dissimilar LA within merged clusters



An approximate doubling in the number of clusters has been achieved, with each tier; 5 to 13 show an increase of 2.6 times, and 13 to 26 doubles exactly. Both the number of clusters produced and the increase in the number of clusters between tiers fit within the framework that was identified as being appropriate prior to the clustering process.

4.6 Outputs

It is essential that any classification produces good and easy to understand outputs. The quality of the classification produced is irrelevant if the information that accompanies it does not provide a quick and easy way of understanding it. This section describes the outputs from the local authority classification, including the structure of the classification, naming, describing the classification and mapping the classification.

4.6.1 Structure and Naming

A three tier hierarchy of clusters has been created and will be referred to in the following way: the tier with 5 clusters as Families, the tier with 13 clusters as Groups, the tier with 26 clusters as Classes, as used in Wallace *et al.* (1999). Table 4.11 shows how the Families, Groups and Classes fit together and the way in which they have been labelled and named. Each cluster has a code and a name, Table 4.11 shows what proportion the UK population falls within each cluster. The membership of each cluster can be seen in Appendix A.

Table 4.11: The Structure of Families, Groups and Classes

5 Families	13 Groups	26 Classes
A: Urban UK (103 LAs 35.8% population)	A1: Industrial Legacy (38 LAs 9.4% population)	A1a: Industrial Legacy (38 LAs 9.4% population)
	A2: Established Urban Centres (43 LAs 17.7% population)	A2a: Struggling Urban Manufacturing (14 LAs 5.6% population)
		A2b: Regional Centres (6 LAs 3.0% population)
		A2c: Multicultural England (13 LAs 6.1% population)
		A2d: M8 Corridor (10 LAs 3.0% population)
	A3: Young & Vibrant Cities (22 LAs 8.7% population)	A3a: Redeveloping Urban Centres (14 LAs 6.7% population)
		A3b: Young Multicultural (5 LAs 2.0% population)
B: Rural UK (205 LAs 36.2% population)	B1: Rural Britain (93 LAs 14.7% population)	B1a: Rural Extremes (24 LAs 2.7% population)
		B1b: Agricultural Fringe (35 LAs 5.8% population)
		B1c: Rural Fringe (39 LAs 6.2% population)
	B2: Coastal Britain (44 LAs 7.6% population)	B2a: Coastal Resorts (8 LAs 1.7% population)
		B2b: Aged Coastal Extremities (28 LAs 4.6% population)
		B2c: Aged Coastal Resorts (8 LAs 3.0% population)
	B3: Averageville (67 LAs 14.0% population)	B3a: Mixed Urban (41 LAs 8.8% population)
		B3b: Typical Towns (26 LAs 5.2% population)
	B4: Isles of Scilly (1 LA 0.0037% population)	B4a: Isles of Scilly (1 LA 0.0037% population)
C: Prosperous Britain (77 LAs 16.3% population)	C1: Prosperous Urbanites (23 LAs 5.4% population)	C1a: Historic Cities (3 LAs 2.7% population)
		C1b: Thriving outer London (10 LAs 2.7% population)
	C2: Commuter Belt (54 LAs 10.9% population)	C2a: the Commuter Belt (54 LAs 10.9% population)
D: Urban London (26 LAs 9.6% population)	D1: Multicultural Outer London (11 LAs 4.4% population)	D1a: Multicultural Outer London (11 LAs 4.4% population)
	D2: Mercantile Inner London (7 LAs 2.0% population)	D2a: Central London (6 LAs 1.9% population)
		D2b: City of London (1 LA 0.01% population)
	D3: Cosmopolitan Inner London (8 LAs 3.2% population)	D3a: Afro-Caribbean Ethnic Borough (5 LAs 2.0% population)
		D3b: Multicultural Inner London (3 LAs 1.2% population)
E: Northern Irish Heartlands (23 LAs 2.2% population)	E1: Northern Irish Heartlands (23 LAs 2.2% population)	E1a: Northern Irish Urban Growth (10 LAs 1.1% population)
		E1b: Rural Northern Ireland (13 LAs 1.1% population)

Although the clusters can be easily named Family A, Group A3, Class A3a etc. this tells one nothing about the nature of the Local Authorities within the clusters. There is no indication of where the areas are, or the characteristics that the areas have. Therefore, each Family, Group and Class requires an informative name.

Naming the five families was not a difficult process as they are uncomplicated and reflect the underlying geography of the UK. Naming the groups and classes is a little trickier. The increased number of clusters makes geography less of an indicator as to why each LA has been placed into that individual cluster, although a good knowledge of the geography of the UK and the likely social characteristics of people in each area is invaluable. By finding the average value of each variable in each cluster, it can be established which variables have the most effect on each cluster. By knowing which variables have the most effect on shaping the character of each cluster a suitable name can be given to the cluster. For example, if the most distinct characteristic of a cluster is a very low value for population density, it is likely the area is rural and then the cluster may be labelled as rural areas.

For each cluster the variables with the most extreme values were selected to explain the characteristics of the cluster. By examining these variables it is now possible to see which have been the most important variables in terms of the creation of each cluster. By using this information along with any useful geographic information that the names and locations of each LA within the cluster may provide, each cluster can be given a suitable name.

It is important when naming the clusters not give them derogatory names. The purpose of giving the clusters names is not so it can be easily assessed whether one area is better than another, but to quickly get some idea of where the area is likely to be and the characteristics of the people who live there. It is all too easy to let personal preference or prejudice about an area cloud one's judgement when naming clusters. Bill Bryson expressed the view that *"Bradford's role in life is to make every place else in the World look better in comparison"* (Bryson 1995 p196). Taking Bryson's view as inspiration, class A2c containing Bradford could be named 'the worst places in the UK'. However, this would import serious prejudice to the classification system and would seriously offend anyone who lives in an area that falls within cluster A2c.

4.6.2 Pen Portraits

The information that has been gathered as to which are the most extreme values in each cluster can also be used to create pen portraits; these are short descriptions (or a simple list) of the characteristics of each cluster. Pen portraits can be referred to by the user of the classification system after they have established to which cluster the area that they are interested in belongs. Users can then read the pen portrait for the relevant cluster to get more information about the areas in that cluster.

Figures 4.7-4.11 show the pen portraits for the five families. They consist of a short description of where they are likely to be found, and a graph showing the variable values. The numbers on the x axis on the graphs in Figures 4.7- 4.11 refer to the final list of 56 variables used in the classification and the various strengths of each variable with each cluster. Table 4.10 or Table 4.11 can be used as a key to relate the numbers to the variable names. Another point to note is that the scale of each graph varies between clusters so they should be studied carefully. Cluster profiles are given for just the five families for brevity, but profiles were also created for both groups and classes.

Figure 4.7: Cluster Profile of Family A: Urban UK

Family A: Urban UK 103 Local Authorities containing 35.8% of the population are in this family. This family contains the UK's most urban Local Authorities (excluding London Boroughs). These Authorities can be found mainly in the English Midlands, North, North West and North East as well as South Wales and the urban corridor between Glasgow and Edinburgh. The Family is characterised by poor health (15, 16), high unemployment (18, 20), low economic activity (19), low car ownership (43, 44) and a negative population change (56).

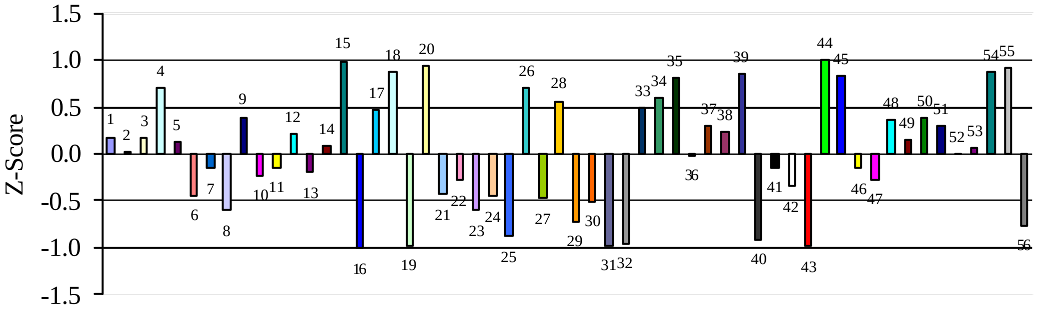


Figure 4.8: Cluster Profile of Family B: Rural UK

Family B: Rural UK 205 Local Authorities containing 36.2% of the population are in this cluster. This Family contains UK's most rural Local Authorities. They are spread throughout the country, are comparatively large in area and are located away from areas of high population. The Family is characterised by a low population density (1), a lot of employment in agriculture, hunting, forestry and fishing (23), detached housing (40) and second / holiday homes (42).

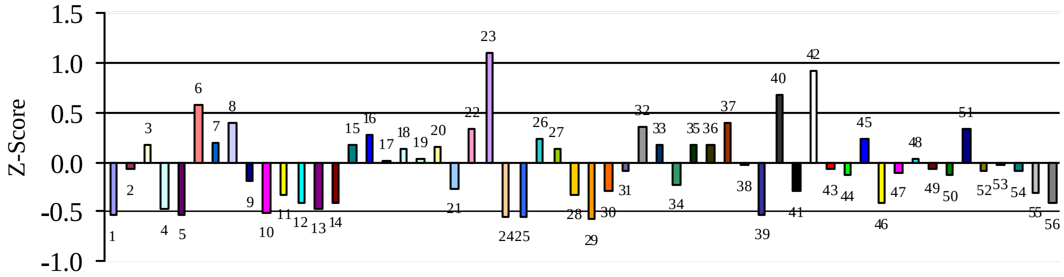


Figure 4.9: Cluster Profile of Family C: Prosperous Britain

Family C: Prosperous Britain 77 Local Authorities containing 16.3% of the population are in this cluster. This Family contains Britain's most prosperous Local Authorities. Typical local authorities in this family include the commuter zone around London and some other large cities, plus some of the Britain's smaller historic cities. The Family is characterised by Good health (15, 16), Low unemployment (18, 20), an economically active community (19), highly qualified (27) mobile people, high car ownership (43, 44) and traditional family values (54).

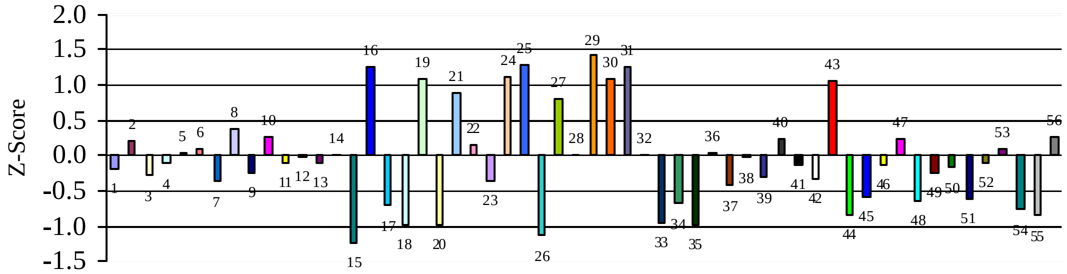


Figure 4.10: Cluster Profile of Family D: Urban London

Family D: Urban London 26 Local Authorities containing 9.6% of the population are in this cluster. This Family contains the densely populated area of London and some of their satellite towns. No local authorities in this family are outside the area immediately around London. The Family is characterised by extreme values for a large number of variables. Trends include high population density (1) and overcrowding (49), a young single population (9), ethnic and religious diversity (11, 12, 14) and low car ownership (43, 44).

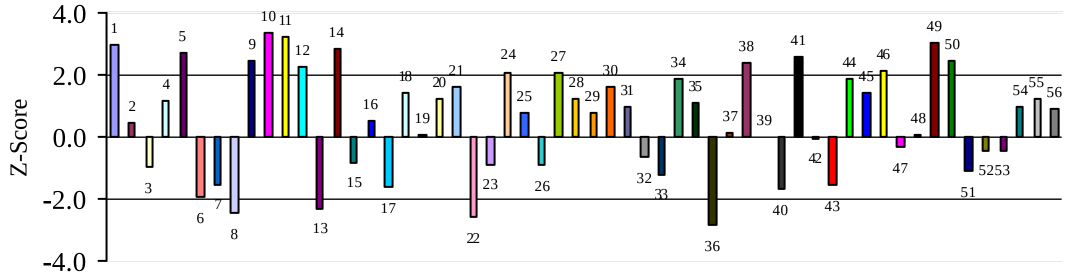
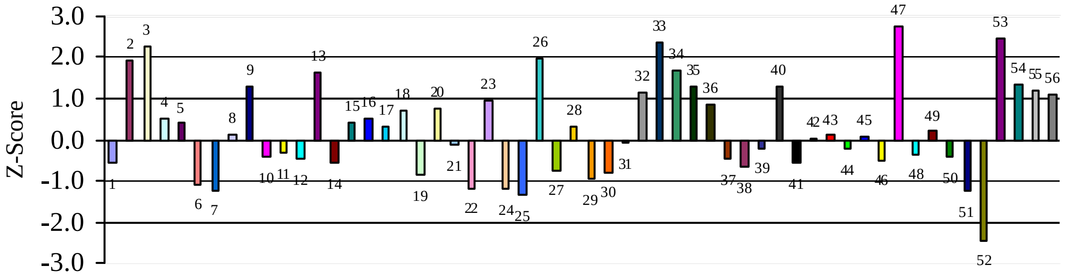


Figure 4.11: Cluster Profile of Family E: Northern Irish Heartlands

Family E: Northern Irish Heartlands 23 Local Authorities containing 2.2% of the population are in this cluster. This Family contains all the Local Authorities in Northern Ireland except Belfast, Castlereagh and North Down. The Family is characterised by extreme values for many variables, a very young (2, 3) growing population (56) with a large number of dependent children (53) and little ethnic and religious diversity (10, 11, 12). Significant numbers of people with no qualifications (26) and routine occupations (33). The Catholic/Protestant divide cannot be seen because the data were not available for the whole UK so could not be used. If variables that only appeared in Northern Ireland census were used more variation would be seen.



The cluster profiles shown in Figures 4.7 – 4.11 make the clusters easier to understand. They give a quick and easy to understand overview of the variable values that have created each cluster. By being able to pick out the variables that have the most effect on each cluster, can add a great deal of contextual information about how and why each of the clusters were formed and the differences between them. After looking at the name given to each cluster, the cluster profiles is the next place the user should look in order to further understand the classification.

4.6.3 Additional Information

In addition to knowing which are the extreme variables for each cluster it is also useful to view the data the other way round. For example you may want to know which LA has the highest or lowest rate of unemployment. Table 4.12 enables this to be done by listing the class that shows the most extreme positive and negative values for each variable. The cluster that has the highest value for variable 1 (population density) is D2a *Central London*; the cluster with the lowest is B1a *Rural Extremes*. This is as one would expect because central London is very densely populated and the more rural areas of the country are most sparsely populated. The cluster that has the highest value for variable 15 (Limiting long-term illness) is A1a *Industrial Legacy*; the cluster with the lowest is B4a *Isles of Scilly*. Again this is not a surprising result. *Industrial Legacy* contains areas with a long tradition of mining and heavy industry, which are known to cause health problems. The reason why *Isles of Scilly* returns the lowest level of LLTI is initially not so obvious. However, there are several possible reasons for this: firstly with it only has a population of 2,153 people and is therefore more likely than any other cluster to return an extreme result either high or low. There is also a possible social reason for the low level of LLTI in the *Isles of Scilly* not that it is necessarily a healthier place to live, but that it is a difficult place to live for anyone with a long term illness, mainly down to poor transport links, needing to get a boat or a helicopter to visit a major hospital. The cluster that has the highest value for variable 40 (Detached Housing) is E1b *Rural Northern Ireland*, the cluster with the lowest is D2b *City of London*. Again this does not seem an unreasonable result; *Rural Northern Ireland* has a great deal of space for large homes, whereas *City of London*, the smallest LA by area, has more pressure on space.

Table 4.12: Classes with the highest positive and negative values for each variable

	Variable	Class with the highest value	
		Positive	Negative
1	Population Density	D2a	B1a
2	People aged: 0 - 9	E1b	D2b
3	People aged: 10 - 17	E1b	D2b
4	People aged: 18 - 24	A2b	B2c
5	People aged: 25 - 29	D2a	B2c
6	People aged: 45 - 64	B4a	D3b
7	People aged: 65+	B2c	D3b
8	Married	B4a	D2b
9	Single (Never Married)	D2a	B4a
10	Born outside UK	D3b	A1a
11	Black minority ethnic groups	D3a	B4a
12	Indian, Pakistani or Bangladeshi	D3b	B4a
13	Christian	E1b	D3b
14	Other Religion	D3b	E1b
15	Limiting long-term illness	A1a	B4a
16	Residents whose health is good	B4a	A1a
17	Residents who provide unpaid care	A1a	D2a
18	Unemployment	D3a	B4a
19	Economically active residents 16+	B4a	A2b
20	Male Unemployment	D3b	B4a
21	Women who work Full-time	D2b	B2c
22	Women who work Part-time	B1c	D2b
23	Agriculture; hunting; forestry and fishing employment	B1a	D2b
24	Real estate; renting and business activities employment	D2b	E1b
25	Managers and senior officials employment	D2b	E1b
26	No qualifications	E1b	D2b
27	Highest qualification attained degree level or above	D2b	A2a
28	Full time Students	A2b	B4a
29	Large employers and higher managerial occupations employment	D2b	B4a
30	Higher professional occupations employment	D2b	B4a
31	Lower managerial and professional occupations employment	D2b	A2a
32	Small employers and own account workers employment	B4a	A2b
33	Routine occupations employment	E1b	D2b
34	Never worked	D3b	B4a
35	Long-term unemployed	D3a	B4a
36	Car to work	E1a	D2b
37	Walk to work	D2b	D1a
38	purpose-built flats	D2b	E1b
39	Terraced houses	A2c	D2b
40	Detached housing	E1b	D2b
41	Bed sits	D2a	E1a
42	Households With no residents: Second residence / holiday home	B4a	A2a
43	Households with 2+ cars	C2a	D2b
44	No car households	D2b	C2a
45	LA Rented	D3a	B2c
46	Private Rented	B4a	A2d
47	Household size	E1b	D2b
48	No central heating	B4a	A2d
49	Households: with an occupancy rating of -1 or less (overcrowding)	D2b	B1c
50	One-person no-pensioner households	D2b	B2c
51	Single pensioner households	B2c	D3b
52	2 adults no children	B2c	E1b
53	Households with dependent children	E1b	D2b
54	Lone Parent Families	D3a	B4a
55	Households: No adults in employment :with dependent children	D3b	B4a
56	Population change 1991 - 2001	D2b	A2a

The classification puts the LAs into clusters of similarity, but each LA does not have to be in the same cluster as the LA to which it is most similar. Data from the classification can be used to see which LA is most similar to each of the 434 LAs in the UK. Table 4.13 shows which local authorities are similar to a selected list of local authorities. This provides useful extra information in addition to the classification. The clusters are not used in any way to create this information. The information created can be seen as a new set of clusters each one specific to each local authority. Although it is likely that LAs will be similar to those in their own cluster, this does not have to be the case. LAs located at the edge of a cluster may very well be similar to LAs in other clusters. If a researcher is interested at looking at just one authority, knowing which authorities are most like that LA may be more useful than the classification itself. The selection of authorities shown in Table 4.13 show the five most similar LAs for each of the selected LAs. The LA names in italics indicate that the LA is in a different cluster (class, 26 level) from the LA that it is being compared to, values represent Euclidian distance in multidimensional space. There are clear differences between the authorities in terms of the distance to their most similar LA and whether their most similar LAs are in the same cluster as them. *Bath and North East Somerset* and *Blackburn with Darwen* are in the same cluster as the five LAs they show most similarity to. In contrast *Birmingham* and *Blaby* are in the same cluster as only two of the five LAs they show similarity to. Therefore, if a researcher is for example interested in comparing the migration rate of *Birmingham* with the five most similar authorities, it is better to use the information from Table 4.13 rather than just select five LAs from the cluster that contains *Birmingham*.

Table 4.13: Distance to five most similar local authorities, for selected local authorities

LA	Most similar	2nd most similar	3rd most similar	4th most similar	5th most similar
Bath and North East Somerset	York UA 2.966	Cheltenham LA 3.09	Chester LA 3.359	Warwick LA 3.451	Colchester LA 3.988
Bedford	Colchester LA 3.262	*Northampton LA* *3.609*	*Hillingdon LB* *3.751*	Peterborough UA 3.865	Dartford LA 3.902
Belfast	Middlesbrough 6.653	Liverpool LA 7.359	Sunderland LA 7.853	Knowsley LA 7.965	*Hartlepool UA* *7.967*
Berwick-upon-Tweed	*Scarborough LA* *4.416*	Alnwick LA 4.595	Dumfries & Galloway 5.054	*North Devon LA* *5.076*	Teesdale LA 5.211
Bexley LB	Havering LB 2.381	Stockport LA 3.546	*Bury LA* *3.57*	*Basildon LA* *3.572*	*Dartford LA* *3.576*
Birmingham	Bradford LA 5.046	*Wolverhampton* *5.317*	*Sandwell LA* *5.537*	Blackburn with Darwen UA 5.924	*Leicester UA* *6.034*
Blaby	*Hinckley and Bosworth LA* *2.783*	South Derbyshire 3.01	*South Gloucestershire* *3.089*	*Eastleigh LA* *3.105*	Selby LA 3.309
Blackburn with Darwen	Bradford LA 3.462	Oldham LA 4.551	Pendle LA 4.621	Rochdale LA 4.809	Burnley LA 5.718

4.6.4 Mapping out the Clusters

Mapping is very important to area classifications. It is essential to have good visualisations of the classification to bring it to life and get the most out of it. By mapping the classification the user is able to recognise geographical patterns that are not apparent from just looking at a list of members of each cluster. As the local authorities are generally large areas it is possible to pick most of them out at a national scale. Therefore, maps of the UK showing the distribution of each cluster type are very useful as they enable any geographic patterns within the clusters to be seen and easily interpreted. Figures 4.12 – 4.14 display maps of all Families, Groups and Classes throughout the UK.

Figure 4.12 demonstrates that maps can really bring the classification to life. Clear geographic patterns can be seen in the distribution of the five families around the UK. The grouping of family C *Prosperous Britain* around London gives a good representation of the commuter zone into London. Family E Northern Irish Heartlands really exemplifies how different the majority of Northern Ireland is to the rest of the UK. Family D *Urban London* shows how different London is from the rest of the country. Family A *Urban UK* clearly highlights all the urban centres (excluding London) of the UK. Family B *Rural UK* clearly highlights the non-urban areas of UK.

Although there are clear and sensible geographic patterns to the five families, there are several interesting cases that can be clearly seen on the maps that warrant further discussion. For example, why is Ceredigion classified as a city when it is in the middle of rural Wales? At first this seems odd because it is a large rural area with a small population of just under seventy five thousand. The reason is in fact quite simple; it is because of the number of students in the LA. Despite its location and its small population Ceredigion actually contains two universities at Aberystwyth and Lampeter. The students in these universities account for almost twelve and a half thousand or 17% of the population. It is easy to see how socially this would make the area more like an urban area than the surrounding rural areas.

Why are Luton and Slough in *Urban London* even though they are not in London itself? The simple answer to this question is that, although they are not in Greater London (but very close to it), they have a social structure that is more like that within London than those authorities outside London. Labelling the group 'London, Luton and Slough' was not a very attractive proposition so *Urban London* was used because it is indicative of the areas within the cluster, despite Luton and Slough not being within the bounds of Greater London.

Stirling stands out as the only member of *Prospering Britain* not in England. The reason for the inclusion of Stirling in this group is that it acts as a commuter zone into Glasgow to the south, much in the same way as the Home Counties do for London, but to a lesser degree.

Figure 4.12: Map of the five Families of the Local Authority classification

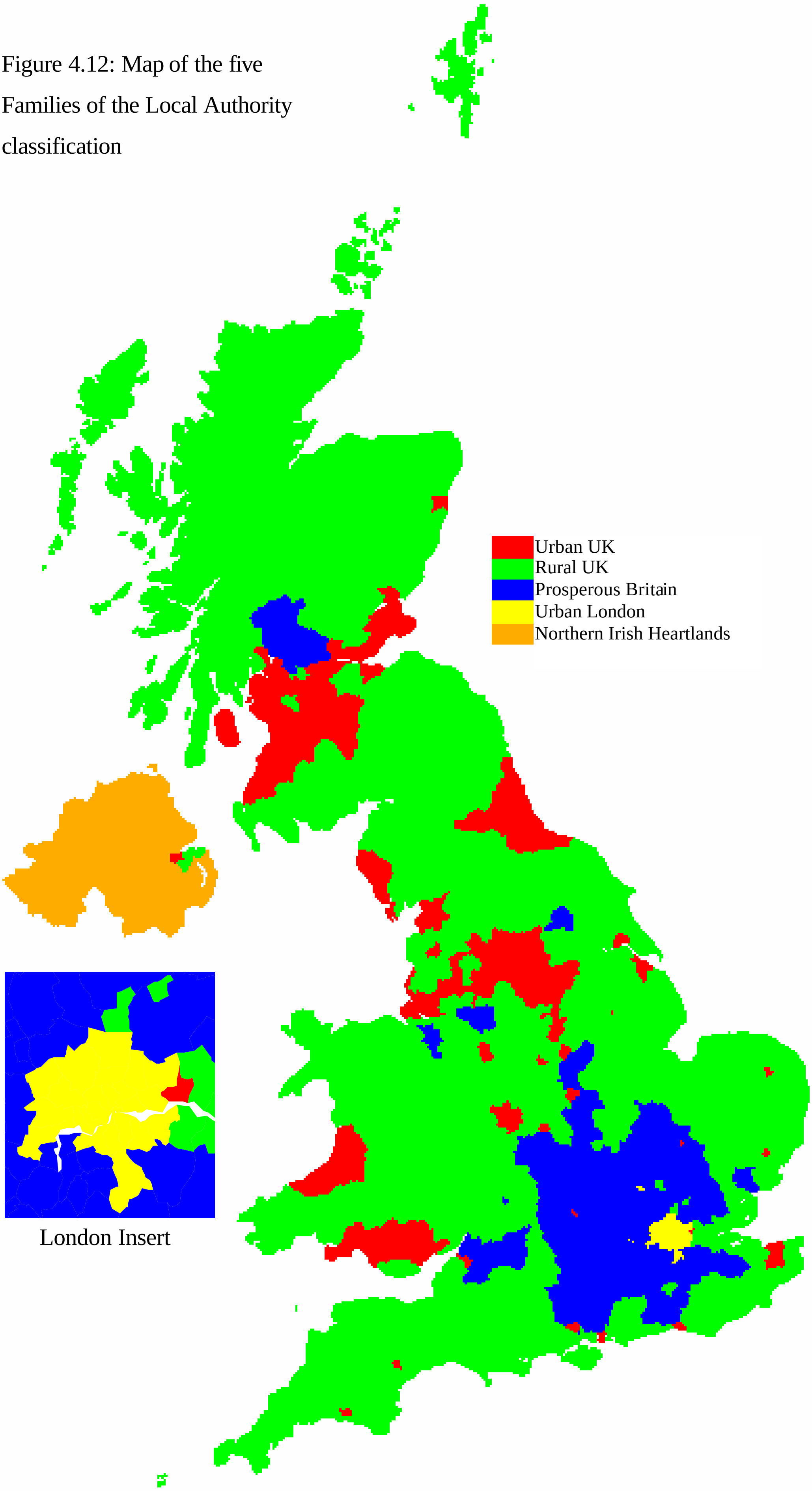
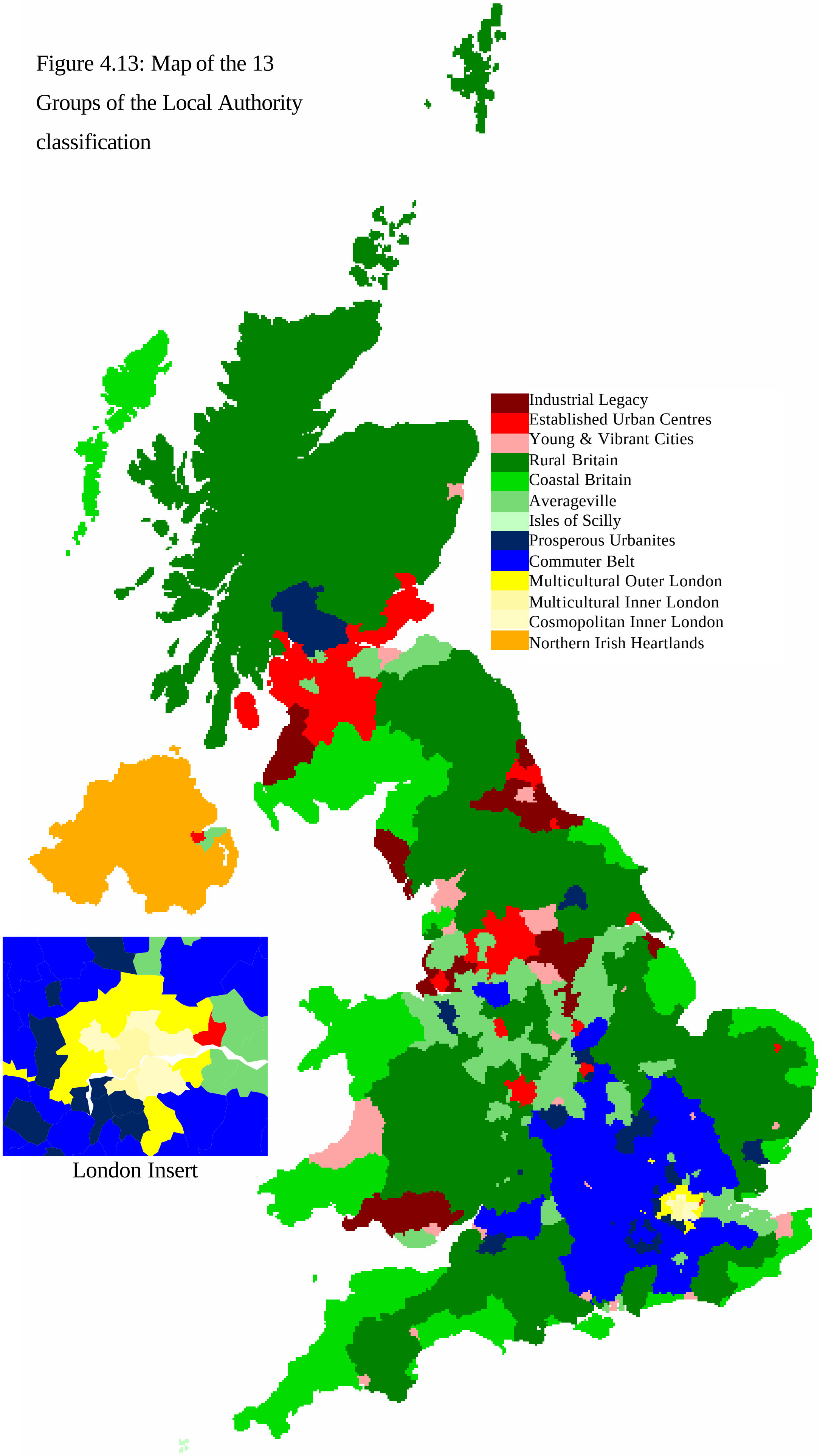


Figure 4.13: Map of the 13 Groups of the Local Authority classification



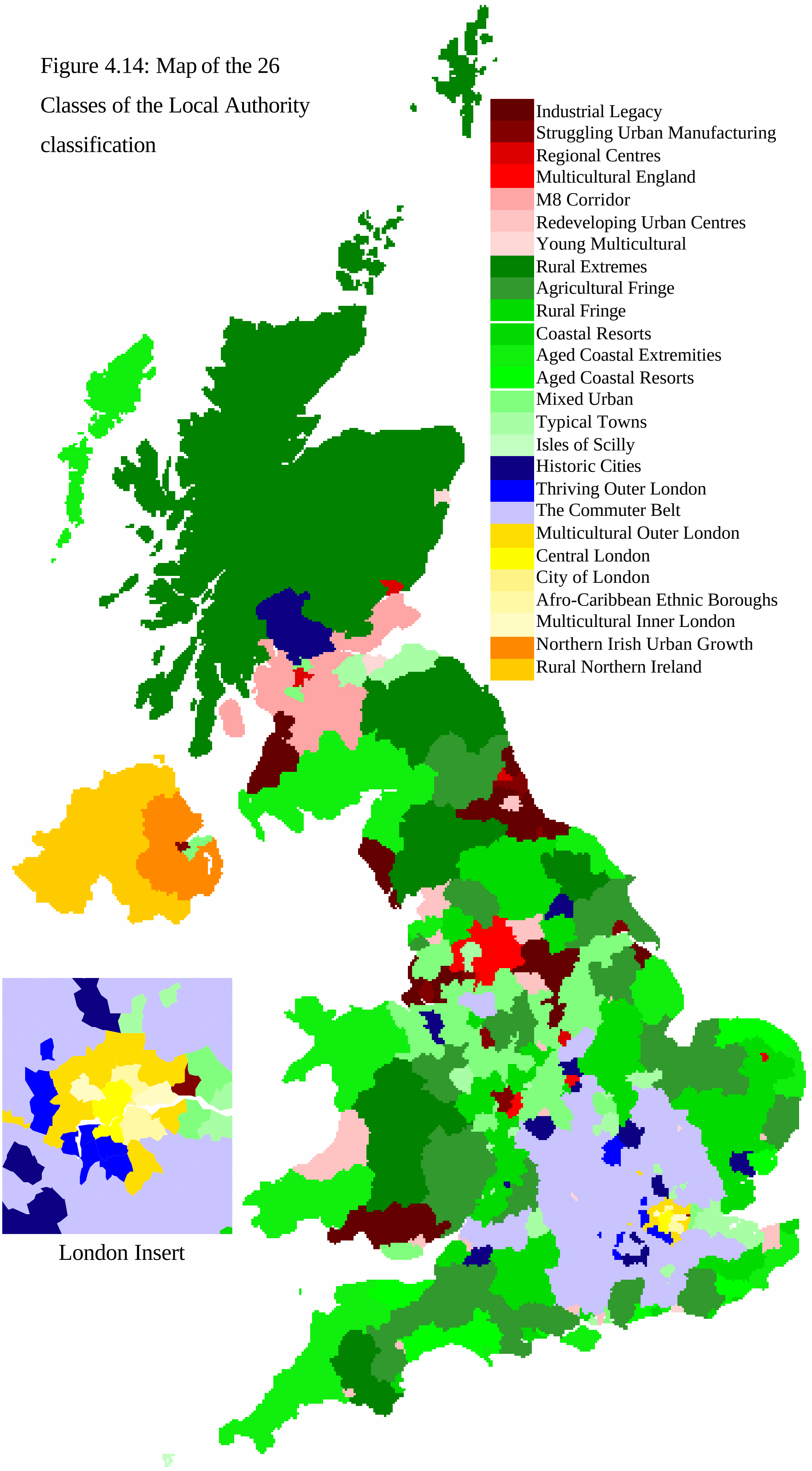


Figure 4.14: Map of the 26 Classes of the Local Authority classification

4.7 A Comparison with the 2001 ONS Classification of Local Authorities

The classification described in this study was developed in parallel with an ONS project to create a national classification of local authorities for release through their website. This section will compare and contrast the two LA classifications created. This will show what extent of the different variable choices and differing methodologies had on the final result, and to what extent the two classifications match despite their different approaches. This will also provide a means of further analysing and scrutinising the outputs from this system. The classification developed in this project will be referred to as the 'Leeds classification'; the classification developed by the ONS will be referred to as the 'ONS classification'.

4.7.1 Variable Choices Comparison

The main difference in the variable choice between the two classifications is that the ONS classification only contains 42 variables whereas the Leeds classification system contains 56. Variables that are included in the Leeds classification, but are not in the ONS classification include: Religion variables, which ONS chose not to include, but shows high levels of Christianity in Northern Ireland and high levels of other religions in areas which have high non-white ethnicity. Second/holiday homes that can have a significant effect especially in remote areas where the population is relatively small. No Cars that provides an extra indication of lack of wealth and poor access to some services. 'Population Change 1991-2001', the ONS classification only used static variables; the population change variable added an extra dimension to this. Good health was used in the Leeds classification, whereas the ONS only reported on poor health. 'No Qualifications' was used in the Leeds classification, only a highly qualified variable was included in the ONS classification. A full list of variables used in the ONS classification can be found in ONS (2004b). The differences in the two sets of variables used will undoubtedly have had an effect on the two classifications produced.

4.7.2 Methodology Comparison

Both classifications used the same clustering methodology, namely Ward's hierarchical clustering procedure. However, different methods of standardisation were employed, ONS used range standardisation rather than z-scores. The main difference between the two is that z-scores allow different ranges between the variables, therefore keeping some representation of extreme values within the dataset, whereas range standardisation gives all variables a range of 1. Both methods are described in § 3.3.1.

The ONS merged the two local authorities with the smallest population with their nearest neighbours, therefore reducing the number of objects clustered from 434 to 432. The Isles of

Scilly was merged with Penwith on the southern tip of Cornwall, and the City of London was merged with Westminster. Whether this was the best thing to do is a point that can be argued either way. It makes things easier but not necessarily better, because although in close proximity to the areas with which they have been grouped, it does not necessarily follow that they would naturally be in the same cluster. This is especially true for the Isles of Scilly because the border for the authority is not arbitrary it is a physical constraint.

4.7.3 Results Comparison

First impressions of the ONS classification suggest that it is more London-centric than the Leeds classification. The 2001 Census figures state that London represents 11.4% of the UK population. However, in the ONS classification, 3 out of 8 (38%) clusters are wholly within London at family level, 4 out of 14 (29%) clusters are wholly London clusters at group level, 7 out of 25 (28%) clusters are wholly London clusters at class level. In the Leeds classification system 1 out of 5 (20%) are wholly London clusters at family level, 3 out of 13 (23%) are wholly London clusters at group level and 5 out of 26 (19%) are wholly London clusters at class level. Both classifications over-represent London in terms of the number of clusters that represent it compared with the size of its population. This is perhaps expected to a certain extent due to the diversity within London. London boroughs show a lot of extreme values that can put them in very small clusters. However, even taking these factors into account the ONS over classify London at all levels, especially at the family level where detail is not needed. As the two classifications have very similar number of clusters at group and class level this means that the rest of the UK could be under-represented in the ONS classification. Northern Ireland also seems to be over represented. This has also happened in the Leeds classification and seems unavoidable as it is quite different from the rest of the UK.

4.7.4 Cluster Numbers Comparison

The number of clusters produced in the two classifications is fairly similar apart from at the family level. It is questionable whether as many as eight families are needed in the ONS classification as the difference between the groups and families is limited. Table 4.14 shows that the increase between the tiers in the Leeds classification is greater than in the ONS classification. Therefore the Leeds classification shows three tiers of more diverse information in comparison to the ONS classification.

Table 4.14: The distance in scale between the Families, Groups and Classes in the two classifications

Leeds			ONS		
Family	5		Family	8	
		×2.6			×1.75
Group	13		Group	14	
		×2			×1.8
Class	26		Class	25	

The ONS used a different method of standardisation to that used in this classification. They used the range standardisation rather than Z-Scores. Z-Score standardise the range over a smaller set of values and therefore reduce outliers further, this would probably have resulted in less London only clusters in the ONS classification if they had used it. This is likely to be one of the main reasons for some of the differences between the two classifications.

The two classifications are two different representations of the same thing. It would be possible to argue that one is better or more useful that the other. However, a more sensible way of looking at the difference between the two classifications is that because of their differences, they are each better at representing specific things. The ONS classification has more wholly London clusters; it is therefore more suitable for investigating diversity within Greater London. Conversely the Leeds classification has more clusters representing non-metropolitan areas, it would therefore be sensible to choose the Leeds classification to investigate diversity between non-metropolitan areas.

4.7.5 Matching the Clusters

By cross-tabulating the two classifications against each other, it can be established to what extent the cluster membership of the two classifications correspond. Figures 4.15 - 4.17 show each of the three levels of the hierarchy in the classification cross-tabulated against the equivalent level of the LA classification.

Figure 4.15 shows the families from the Leeds classification cross-tabulated with the highest level of the ONS classification, termed ‘super-groups’. Significant overlaps are highlighted with a blue background, less significant overlaps are not highlighted. There are a different number of clusters in the two classifications, this will obviously affect the extent to which the two classifications correspond. However, there is a clear association between certain clusters within the two classifications.

Figure 4.15: Crosstabulation of the Leeds families with the ONS super-groups

ONS	Leeds					
	A: Urban UK	B: Rural UK	C: Prosperous Britain	D: Urban London	E: Northern Irish Heartlands	
1: Cities & Services	43	6	7			56
2: London Suburbs			1	11		12
3: London Centre				9		9
4: London Cosmopolitan				6		6
5: Prospering UK	2	113	69			184
6: Coastal & Countryside	1	63				64
7: Mining & Manufacturing	57	23			10	90
8: Northern Ireland Countryside					13	13
	103	205	77	26	23	434

The Leeds family *Urban UK* is split between two of the ONS super-groups, *Cities & Services* and *Mining & Manufacturing*. The two classifications clearly match up significantly, but *Urban UK* does not clearly map on to any single ONS super-group. This suggests that the difference in the variables and the methodology has made a difference between the classifications.

Rural UK is split between three of the ONS super-groups, 55% of the family corresponds to *Prospering UK*, and a further 30% to *Coastal & Countryside* and 11% *Mining & Manufacturing*. Although one ONS super-group type makes up over 50% of this family, it is comprised of significant amounts of three different super-group types.

Prosperous Britain is made up by 90% of just one ONS super-group *Prospering UK*. This shows a near to perfect match between the two clusters, suggesting that this is a very well defined and natural cluster because it largely consists of members of just one ONS super-group.

Family D, *Urban London* consists of three ONS super-groups: *London Suburbs*, *London Centre* and *London Cosmopolitan*. Although this family is made up of three different super-groups, it comprises all but one member of them. The cross-tabulation shows how the one wholly London group in the Leeds classification is comprised of the three wholly London groups from the ONS classification. Evidence suggests that London does not need to be represented by so many groups at the highest level of the hierarchy, if the three groups can be almost exactly represented by one single group. Although this is sound in theory, the intricacies of Ward's hierarchical clustering method may not make the selection of a single London group possible without manual intervention in the clustering process.

Northern Irish Heartlands is comprised of two ONS super-groups, *Mining & Manufacturing* and *Northern Ireland Countryside*. Like *Urban UK*, *Northern Irish Heartlands* has been split virtually down the middle by two of the ONS super-groups. There is a clear match up, but it is not clearly to one cluster or the other.

Overall, when crosstabulated the two classifications appear to match fairly well. Firstly and most importantly there does not appear to be any great contradictions between the two classifications. Having different variable lists and differing methodologies there are bound to be differences in membership between the two classifications. An example of a contradiction would be an LA that is in the *Urban London* family being part of the *Northern Ireland Countryside* super-group. These two clusters are completely opposed to each other, in terms of location and urbanisation. Irrespective of methodology and variable choice, no crossover in groups such as these should be expected. Of the 40 cells in Figure 4.15 only 16 (40%) have numbers in them which represent where the memberships of the two classifications meet. The minimum possible number would be eight and would suggest the classifications were almost identical despite the different numbers of clusters in each. If all 40 cells were filled in this would suggest randomness of membership in one or both of the classifications and that the classifications were very different or unrelated.

Figure 4.16 shows a crosstabulation of the second level of the hierarchy, the Leeds groups against the ONS groups. Of the 182 squares in the cross tabulation only 41 (23%) contain a number and therefore represent a crossover between clusters in the two classifications. in terms of percentage this is almost half as many as at the higher level suggesting that the two classifications match up even better at this level.

Figure 4.16: Crosstabulation of the Leeds groups with the ONS groups

		Leeds													
		A1	A2	A3	B1	B2	B3	B4	C1	C2	D1	D2	D3	E1	
ONS	1		6	14		5		1							26
	2		18	3											21
	3			2	1			5	1						9
	4							1		11					12
	5										7	1			8
	6											7			7
	7			2	64	1	26	9	13						115
	8				1		18	5	2						26
	9				3			2	38						43
	10			1	15	36							1		53
	11				9	2									11
	12	24	19				3								46
	13	14					20							10	44
	14													13	13
		38	43	22	93	44	67	23	54	11	7	8	1	23	434

Figure 4.17 shows a crosstabulation of the third level of the hierarchy, the Leeds classes against the ONS sub-groups. Of the 650 cells only (12%) are filled. However, at this point it becomes almost impossible to draw any conclusions from this as there are only 434 local authorities so only two thirds of the cells could actually have a number in them, even if all numbers were 1.

What is apparent though from Figures 4.16 and 4.17 is that there is still a great deal of agreement between the two classifications. Those instances where there are only one or a few LAs in common tend to come from clusters with smaller memberships or are in clusters that are similar to those which contain more of that type of LA.

Figure 4.17: Crosstabulation of the Leeds classes with the ONS sub-groups

Leeds

ONS	A1a	A2a	A2b	A2c	A2d	A3a	A3b	B1a	B1b	B1c	B2a	B2b	B2c	B3a	B3b	B4a	C1a	C1b	C2a	D1a	D2a	D2b	D3a	D3b	E1a	E1b	
1		2	11			2		5		1																	21
2						2										3											5
3		1	3				10																				14
4		3				2	2																				7
5																2											2
6											1	1						5									7
7															7			1									8
8															4												4
9																			6	1							7
10																						1					1
11																					5						5
12																						2					2
13			2						11	8	1	1	1	1													25
14					28				2		14	2						1									47
15				1	2				11		19	10															43
16									2	2	1	2		16				3									26
17										2	3	38															43
18			1	10	5								6				8										30
19								3					19										1				23
20				8					1				2														11
21	22	8					1																				31
22	2													3										10			15
23	14								14					6													34
24																									9	1	10
25																									1	12	13
	38	14	17	19	35	6	13	8	41	13	39	54	28	26	11	5	8	10	6	1	5	3	1	10	10	13	434

Figures 4.15 - 4.17 show that by crosstabulating the two classifications against each other it is possible to ascertain, which clusters in each classification correspond most closely to each other. By doing this it becomes apparent that the two classifications are different, which would be expected as they have differences in both their variable lists and methodologies. However, they display the same general patterns and there is a lot of agreement between the two systems. The most important point is that the classifications do not contradict each other by putting any LA into very different clusters.

4.8 Conclusions

A local authority classification has been successfully created, providing a comprehensive picture of the socio-economic differences and similarities that exist between local authorities in the UK. The classification displays sensible, coherent and distinctive geographic patterns that accurately represent social variations within the UK. The created classification complements and contrasts with ONS classification. Broadly the same patterns are visible, but with interesting differences between the two classifications. They do not contradict each other, but provide different perspectives on the same question.

As well as the creation of the LA classification the aim of this chapter was to gain the skills, knowledge and experience to move on from the LA classification to create the more complicated and more demanding output area classification. The creation of the LA classification has proved invaluable in terms of providing knowledge about variable selection, standardisation and the running of cluster analysis. The LA classification was created in the SPSS statistical package that did the job very smoothly and should be more than adequate to do a similar job at the output area scale.

So, what lessons that have been learnt from the creation of the LA classification can be taken forward to aid in the creation of the OA classification? It is important to recognise that the change in scale makes the OAs a completely different proposition to LAs, and that some of the things that affected the LAs won't affect the OAs, but also that new problems will emerge with the OA classification that were not problems in the LA classification. However, there are several things learnt in creating the LA classification that should not be forgotten in the creation of the OA classification. A careful examination and comparison of a comprehensive list of variables is required to ensure that all possible useful information is considered for inclusion in the classification. All available information should be taken into account when deciding whether or not to include any variable. Using just one piece of information could result in poor choices being made. The information collated in the review of the ten previous classification systems can also be used to aid the variable choice for the OA classification, which is described in Chapter 5.

The number of clusters to be created needs to be selected based on a number of factors. The perceived ideal number of clusters prior to clustering may not match the number that the statistics imply. However, it should always be remembered that it is most important that the classification is useful, and that the number of clusters created should reflect a number that will be useful to the user of the classification.

Accurately naming and describing of the clusters is vital to explain the classification and enable users to interpret what each cluster represents. Mapping the classification greatly aids the understanding and interpretation of the clusters. The geographic patterns displayed by the classification can reveal what the classification tells us about the social make-up of the country.

Different, but similar classifications can be produced using a different set of variables, methods of standardisation or clustering method. No one result is right or wrong, they are just different representations of the same thing. However, it is the job of the creator of a classification to justify the decisions they have made in the creation of the classification, because a different choice would produce a different classification.

Chapter Five - A Classification of 2001 Census Output Areas

5.1 Introduction

The 2001 Area Classifications place each geographic area into a group according to key characteristics of the people who live in that area. The groups are created using statistical techniques known as cluster analysis. This classification of Output Areas fits into the ONS suite of area classification and follows the publication of classifications at local authority, health board and ward level (ONS 2004b), the hierarchy of which, will be reviewed in Chapter 8.

Agreement with ONS was needed on what form the classification should take, because the classification was to be jointly published under their name. The author made variable and methodological decisions that then went before an ONS project board for approval. Before publication as a 'National Statistic' the classification was approved by the then Director of National Statistics, Mr Len Cook.

In section 5.2 the concepts underpinning Output Areas (OAs) and their geography are explained, in addition some of their idiosyncrasies are identified. In section 5.3 the inputs to the classification are discussed. Variable selection and the thinking behind them are explained as are the processes involved in the assembly of the classification database; there is also an elucidation of methods of data checking of the input database. Section 5.4 describes the processes involved in the creation of the classification, including the method of standardisation of the variables. The original methodology for the creation of the classification is described. The section moves on to describe why the original methodology was rejected due to inherent problems. The section discusses how the problems were over-come using a revised methodology. The reasons for the use of this new methodology, and how it is more reliable than the original methodology are outlined. Section 5.5 discusses the outputs from the classification, outlining the often problematic task of naming and describing the clusters, then adding the geography back into the classification by examining different ways of mapping the cluster membership. Section 5.6 will conclude the chapter by summarising its findings.

5.2 Introducing Output Areas: the Objects to be Clustered

The purpose of this section is to briefly introduce output areas (OAs), the smallest geography at which demographic data are released from the 2001 Census. OAs are available for the whole UK, but differ slightly in development and definition, as described later. OAs replaced the previously used enumeration districts (EDs), the difference between the two being that EDs were created for the purpose of data collection (enumeration) rather than for the publication of outputs. The new OAs were principally created for data dissemination. They were built after the collection of the census data using the data collected in their design, in order to produce a new output geography independent of the data collection areas.

5.2.1 Output area design

Output areas were pioneered by the General Register Office for Scotland (GROS) for the publication of small area statistics from the 1991 Census. These were built from postcodes using a geographical information system (GIS) by GROS staff. The aim was to create OAs that matched the EDs from the 1981 Census so that comparisons could be made easily. This involved converting a set of addresses that constituted the Royal Mail's unit postcode into a territory on the map. A layer showing the main topographical features (roads, railways, rivers, fences, walls and buildings) was used to enable staff using the GIS to choose sensible OA boundaries. Such a system was considered by the Office for Population, Censuses and Surveys (OPCS) for England and Wales in 1991, but it was felt to be too labour intensive and expensive for implementation in a country with ten times the population (ONS 2005b).

In the 1990s this problem was overcome through an innovative project piloted by David Martin (Department of Geography, University of Southampton) while on study leave at the Office for National Statistics. He developed an automatic method for generating postcode territories using georeferences for addresses (Ordnance Survey's Address Point™) and a Thiessen polygon algorithm available in the GIS (ESRI's ARC). Thiessen polygons allocate territory to the nearest defined set of points. Thiessen polygons are made up of straight line boundaries, which were improved by linking (clipping) to other ONS boundaries (e.g. EDs) that followed more natural landscape features (Martin 2002b).

Martin's innovation was to adopt a zone design algorithm developed by Stan Openshaw (Openshaw and Gillard 1978; Openshaw and Rao 1995) for the task of constructing *n* OAs from *N* unit postcode territories in a way that met a set of constraints (having above threshold numbers of people and households; being contiguous) and that optimised OA properties such as population size homogeneity, socioeconomic homogeneity and shape (as close to circles as possible). For detailed descriptions of the creation of the 2001 Census Output areas see the papers by David Martin which give a very good and clear description of how they were created (Martin 1998, 2000a, 2000b, 2002a, 2002b, Martin *et al*. 2001).

The three census agencies, ONS for England and Wales, GROS for Scotland and NISRA for Northern Ireland were all individually responsible for the creation of OAs in their countries. There were some differences in the methodology between the agencies (ONS 2005b). ONS and NISRA followed the ONS design methodology with a minimum OA size of 100 residents and 40 households. In Scotland OAs were matched as closely as possible to 1991 OAs, retaining a smaller minimum size of 50 residents and 20 households (ONS 2005b). Table 5.1 shows how these different methodologies have affected the number and size of OAs that have been produced in each country.

Table 5.1: The average size of OAs in the constituent countries of the UK

Country	OAs	Population	Households	Average Population per OA	Average Households per OA
UK	223,060	58,789,194	24,479,439	264	110
England and Wales	175,434	52,041,916	21,660,475	297	124
Scotland	42,604	5,062,011	2,192,246	119	52
Northern Ireland	5,022	1,685,267	626,718	336	125

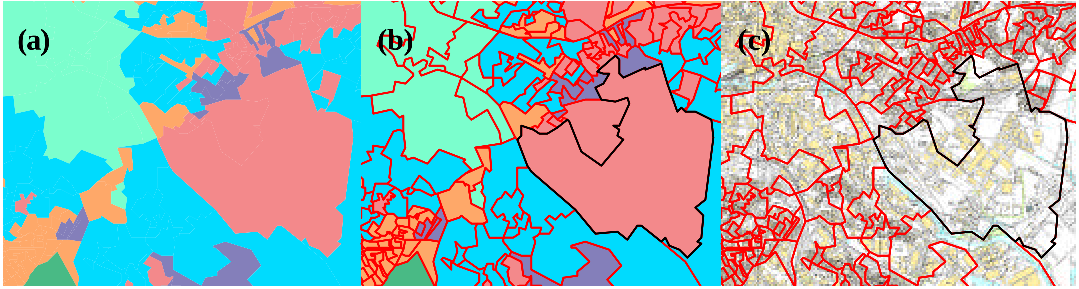
Source: 2001 Census

There are many issues relating to how OA boundaries divide up the country. Should the whole of a small settlement be included in one OA or should they be split and combined with a hinterland of dispersed farmsteads? The first solution tends to create doughnut shaped OAs from the rural hinterland, while the second solution divides up what is a single community. Examples of both solutions can be found among the rural OAs. Another issue is that of stacked postcodes (tower blocks) in urban areas. These are dwellings that cannot be split for the purposes of census mapping, as they occupy the same space on the ground. This has two effects on the output produced. The tower block is given its own OA regardless of the social make up of its inhabitants and thus creates OAs which have very high population density. These large multi-storey dwellings OAs often appear as outliers in the classification (Martin 2002b).

Buildings with empty tenancy or non-residential function can be a problem in the creation of OAs because they can take up a large area, even in an urban setting, but do not represent many people. When data are mapped to represent each OA, the geographically larger OAs dominate the map even though they are likely to have fewer residents than a smaller OA. This is of course a long standing issue in cartography. This problem is most troublesome in urban areas where non-residential areas are not as obvious. Good local knowledge of the area is often required to identify them.

Figure 5.1 shows three maps of an area of Leeds containing both residential and non-residential areas. Looking at (a) you would naturally assume that cluster represented by the red colour is the most prominent in the area. Looking at map (b) where the OA boundaries have been added, you will probably start to have some doubts as you will see that most of the red area is made up of one OA outlined in black. Map (c) reveals that the majority of the area of the large OA is infact made up of a non-residential/industrial area. Even though the industrial area contains no people it is assigned to an OA because the OAs are designed to provide 100% geographic coverage of the UK. The OAs had to be stretched that far to reach the minimum size threshold. Conversely it is the smallest OAs (in terms of area) that often represent the most people as they live in large residential dwellings that cannot be split. The most populous OA in the UK, the University of Lancaster campus containing 4,156, it could not be split because it has a single postcode.

Figure 5.1: shows OA boundaries for an area of Leeds overlain on OS 1:10,000 mapping



©Ordnance Survey Crown Copyright

Large bodies of water represent a similar problem to non-residential buildings as they have to be included within an OA, the most sizable example being Lough Neagh in Northern Ireland. Another unusual example is how the city of Bristol extends into the Seven Estuary presumably as Bristol City Council have the responsibility to maintain it, so it falls within their boundaries and therefore constitutes part of an OA.

The previous chapter outlined the creation of a classification at local authority level. There are 434 LAs in the UK that were each represented by 56 variables which works out at 24,304 data points. Output Areas are a much finer geography than local authority districts. There are 223,060 output areas in the UK. If the same 56 variables were used in an output area classification, then 12,491,360 data points would be produced. Classifying OAs is clearly a more complex process than classifying local authorities, therefore computational and technological limitations need to be considered.

5.3 Inputs

The aim of this section is to discuss the variable inputs into the classification, the choice of variables for analysis in the OA level 2001 Area Classifications, the reduction of the initial variable list to a smaller final list and the reasons behind the selections. In addition the assembly of the variable database and data checking procedures applied to the database are explained. The results of any classification will of course depend on the variables selected to create it. All issues arising from the review of previous classification systems in § 4.3, were also considered in the creation of the OA classification.

5.3.1 Variable Selection

The goal of the variable choice for this classification was to select the minimum possible number of variables that satisfactorily represent the main dimensions of the 2001 Census (Bailey *et al.* 1999 and 2000). The variables will be selected solely from the 2001 Census. There are several reasons why it was felt that using non-census data would be inappropriate. The Census is the most complete and reliable socio-economic dataset available in the UK (Rees and Martin e*t al.* 2002). No other dataset has the same amount of data with such a comprehensive geographic coverage. Another important factor is the scale of the data. At present the only substantial body of official statistics available at OA level are data from the 2001 Census, so using data from other sources would require converting the data from other scales. Linking of datasets at different spatial scales would create all kinds of reliability issues, as discussed by Vickers (2003). Complete confidence in all data included in the classification is needed if the classification is to be published as a 'National Statistic'.

5.3.2 Initial Set of Variables Considered

By reviewing the census data available at OA level, five main domains have been identified: Demographic Structure, Household Composition, Housing, Socio-Economic and Employment. The variables will be discussed within these groups in the rest of this section of the chapter.

The Key Statistics have already been identified as being the most important variables by ONS in the creation of the data, so the initial data set examined in this study included all variables from the OA Level Key Statistics Tables. The Key Statistics represent both the most important variables within the published data from the census, and have a comparatively simple data structure that will aid data extraction. They were also the first data to be released at OA level from the 2001 Census and so presented the earliest opportunity to start the project.

An initial selection of variables was made with the intention of representing the main domains of the census; this list would then be reduced significantly following detailed assessment of each variable. Variables from the Key Statistics tables were considered for use. Variables were merged to create composite variables; for example, the variable Indian, Pakistani and Bangladeshi represents people identifying themselves as Indian, Pakistani or Bangladeshi separately. Having previously created a classification system of local authorities as discussed in Chapter 4, knowledge has been built up about which variables were likely to perform well within a classification system. The review of variables in previous classification systems was also useful information in the formation of the initial variable list. The initial set of variables considered are listed in Table 5.2

Table 5.2: The initial set of variables considered for inclusion in the classification

Number: Variable	Number: Variable
1: % Male	48: % Skilled trades occupations employment
2: % Female	49: % Personal service occupations employment
3: % Living in communal establishments	50: % Sales and customer service occupations emp.
4: Population Density	51: % Process, plant and machine operatives emp.
5: % Aged 0-15	52: % Elementary occupations employment
6: % Aged 16-24	53: % No Qualifications
7: % Aged 25-44	54: % Qualification level 4 or 5
8: % Aged 45-64	55: % Large employers & higher managerial occupations
9: % Aged 65+	56: % Higher professional occupations
10: % Married	57: % Lower managerial and professional occupations
11: % Cohabiting	58: % Intermediate occupations
12: % Single	59: % Small employers and own account workers
13: % Divorced	60: % Lower supervisory and technical occupations
14: % of people born outside UK	61: % Semi-routine occupations
15: % of people Indian, Pakistani & Bangladeshi	62: % Routine occupations
16: % of people Black	63: % Never worked
17: % of people Chinese	64: % Long-term unemployed
18: % Christian	65: % Work from home
19: % other Religion	66: % Car or Van to work
20: % No Religion or Religion not stated	67: % Public transport to work
21: % of people with LLTI	68: % Walk to work
22: % of people whose health is good	69: % Second residence/ holiday accommodation
23: % of people whose health is fairly good	70: % Detached house or bungalow
24: % of people whose health is not good	71: % Semi-detached house or bungalow
25: % of people who provide unpaid care	72: % Terraced house or bungalow
26: % of people employed part time	73: % Purpose built block of flats or tenement
27: % of people employed full time	74: % Part of a converted or shared
28: % of people self employed	75: % In commercial building
29: % of people unemployed	76: % Caravan or other mobile or temporary structure
30: % of people full time students	77: % No Car
31: % of people look after family/home	78: % 2+ Cars
32: % Agriculture, hunting, forestry and fishing emp.	79: % LA Rented
33: % Mining & quarrying and construction employment	80: % Private Rented
34: % Manufacturing employment	81: Average household size
35: % Electricity, gas and water supply employment	82: Average number of rooms per household
36: % Wholesale & retail trade, repair of vehicles emp.	83: % With an occupancy rating of -1 or less
37: % Hotels and catering employment	84: % No Central heating
38: % Transport, storage and communication emp.	85: % No Bath or shower
39: % Financial intermediation employment	86: % lowest floor level above the ground
40: % Real estate, renting and business activities emp.	87: % Single pensioner household
41: % Public administration and defence employment	88: % Single person non-pensioner household
42: % Education employment	89: % All pensioner household (family)
43: % Health and social work employment	90: % Two Adults no Children
44: % Managers and senior officials employment	91: % Lone parents
45: % Professional occupations employment	92: % All Student households
46: % Associate prof. & technical occupations emp.	93: % All Pensioner households (other)
47: % Administrative and secretarial occupations emp.	94: % No adult in employment with dependant children

Note: employment shortened to emp. in some cases

5.3.3 Reducing the Initial Set of Variables

Variable selection for the OA classification followed a different process to that used in the LA classification; it was done in conjunction with that for the electoral ward classification. It was decided by the ONS Project Board and the School of Geography team that it would aid the understanding of the user if the two sets of variables could be the same, minimising confusion when comparing the two classifications (allowing for some differences that are unavoidable due to the change of scale). In all cases, the decision to include or exclude a variable also involved using the judgement of the members of the team.

The reduction in scale from LAs to OAs makes the classification more sensitive to variable selection. This raises a number of issues. A number of reasons for inclusion were formulated which resulted in the guidelines as set out in § 5.3.4 – 5.3.11 were followed in order to assess the value of including any particular variable in the classification, the guidelines are as follows.

5.3.4 Highly Correlated Variables

As discussed in Chapter 3 strong correlations within a dataset are undesirable for cluster analysis, because they represent data redundancy. In order to look for redundancy within the dataset a correlation matrix of all 94 variables against each other was constructed. Including highly correlated variables makes it very hard to assess the effect of any individual variable on the clustering process. A number of strong correlations were found in the initial set of variables. Table 5.3 shows a list of variable pairs from the original list that are correlated at 0.7 or above (i.e. redundancy of 49%>). The pairs of variables that are highly correlated, are perhaps not surprising.

There are three different types of correlation visible. The first are pairs of variables that share the same denominator, hence that the correlations will have a natural propensity to be negative. For example males (1) and females (2) show a perfect negative correlation. This is not surprising as being one rules out someone from being the other. As they share the same denominator and each person can only be present in one of the categories. If there are only two possible categories (such as male or female or yes or no) a perfect negative correlation will be produced. If there are more categories a high negative correlation will still be seen, but not to the same extent.

The second type of correlation consists of those variables that are inherently connected due to causality i.e. one variable is a fundamental property of the other, but they don't share the same denominator. An example of this is the pair of variables third in the list in Table 5.3, % Purpose built block of flats or tenement (73) and % lowest floor level above the ground (86). These variables are linked as the majority of flats are found above ground level, but they don’t share the same denominator.

Table 5.3: Highly Correlated Variables from Original Variable List (ordered by redundancy)

Variable	Variable	correlation	redundancy
1: % Male	2: % Female	-1.00	100
15: % of people Indian, Pakistani & Bangladeshi	19: % other Religion	0.93	86.49
73: % Purpose built block of flats or tenement	86: % lowest floor level above the ground	0.92	84.64
21: % of people with LLTI	22: % of people whose health is good	-0.90	81.00
28: % of people self employed	59: % Small employers and own account workers	0.90	81.00
21: % of people with LLTI	24: % of people whose health is not good	0.89	79.21
22: % of people whose health is good	23: % of people whose health is fairly good	-0.88	77.44
45: % Professional occupations employment	56: % Higher professional occupations	0.88	77.44
22: % of people whose health is good	24: % of people whose health is not good	-0.87	75.69
45: % Professional occupations employment	54: % Qualification level 4 or 5	0.86	73.96
54: % Qualification level 4 or 5	56: % Higher professional occupations	0.86	73.96
77: % No Car	78: % 2+ Cars	-0.86	73.96
91: % Lone parents	94: % No adult in employment with dependant children	0.83	68.89
9: % Aged 65+	87: % Single pensioner household	0.82	67.24
53: % No Qualifications	57: % Lower managerial and professional occupations	-0.81	65.61
10: % Married	12: % Single	-0.80	64.00
10: % Married	77: % No Car	-0.80	64.00
81: Average household size	82: Average number of rooms per household	0.79	62.41
29: % of people unemployed	64: % Long-term unemployed	0.77	59.29
53: % No Qualifications	54: % Qualification level 4 or 5	-0.77	59.29
66: % Car or Van to work	67: % Public transport to work	-0.77	59.29
70: % Detached house or bungalow	78: % 2+ Cars	0.77	59.29
10: % Married	78: % 2+ Cars	0.76	57.76
14: % of people born outside UK	19: % other Religion	0.76	57.76
44: % Managers & senior officials employment	55: % Large employers and higher managerial occupations	0.76	57.76
52: % Elementary occupations employment	57: % Lower managerial and professional occupations	-0.74	54.76
52: % Elementary occupations employment	62: % Routine occupations	0.74	54.76
10: % Married	88: % Single person non-pensioner household	-0.73	53.29
22: % of people whose health is good	53: % No Qualifications	-0.73	53.29
28: % of people self employed	65: % Work from home	0.73	53.29
54: % Qualification level 4 or 5	57: % Lower managerial & professional occupations	0.72	51.84
31: % of people look after family/home	94: % No adult in employment with dependant children	0.71	50.41
51: % Process, plant and machine operatives	62: % Routine occupations	0.71	50.41
53: % No Qualifications	62: % Routine occupations	0.71	50.41
77: % No Car	79: % LA Rented	0.70	49.00

The third type of correlation is between variables where the presence of one variable indicates the presence or absence of another, but does not fundamentally cause it to be so. The pair of variables that are second on the list in Table 5.3 are an example of such a relationship. The % of people Indian, Pakistani and Bangladeshi (15) and % other Religion (19) are highly correlated, and have a strong power of prediction over each other. Somebody who answers yes to one of these questions is more than likely to answer yes to the other, because the socio-cultural make up of that type of person is that they generally have both characteristics, even though having one doesn’t force the other to be true. This third type of correlation are the most interesting type of correlation within a dataset, because their relationship is not preordained, even though a small amount of knowledge could suggest that they would be highly correlated.

Table 5.3 shows several correlations of all three types. Common sense would suggest that one of each pair of highly correlated variables should be removed because much of the information is redundant, but there is another way of looking at highly correlated variables. The predictive and descriptive power of the highly correlated variables is exactly what we are looking for in variables for use in the classification (Voas and Williamson, 2001a). Evidence suggests that variables that can predict the value of other variables would enable the classification to predict other behaviours as the data within it would be proven to be highly predictive of something else. Therefore, there is an inclination to retain a high proportion of highly correlated variables as they can be seen as powerful predictors. This view needs to be balanced against a desire to drop at least one of each pair of highly correlated variables due to data redundancy. Correlations between variables must be carefully examined. Highly correlated variables must be judged on the individual merits of each variable against every other variable and not just rejected because of high correlation with one variable.

5.3.5 Variables with Badly Behaved Distributions

Methods of clustering and standardisation (as described in Chapter 3) work reliably with data that have a normal distribution. However, highly skewed distributions can create problems in both the standardisation and clustering procedures. The skew observed most often in census data and the one that causes the most problems when clustering is a positive skew. That is to say the majority of the data are found at the lower end of a 0-100 scale. The most common form of this in census data is when a variable only identifies very small sectors of the population. Another way to look at this is that the majority of areas have an absence of a particular feature leading to a large number of zeroes within a specific variable. These problems become more acute as the scale reduces because the likelihood of extreme values becomes more likely as the population of each area reduces.

What is the reason that these distributions are problematic? Let’s take as an example variable from the 2001 Census, the percentage of people living in communal establishments. Some 88% of OAs have a value of 0. This suggests that the important fact about this variable is whether or not its value is 0. Areas with a value greater than zero are inherently different, as they have a presence of something that the majority of the areas lack. If this variable was to be split into two groups the most obvious place to split it would be OAs with a value of 0 (88%) in one group and everything else (12%) in another, the important point to remember here is that when working with one variable is that areas with the same value have to be in the same group. Therefore, the most evenly sized groups that can be produced in this case are those already suggested. Nevertheless there are other ways of splitting the data because the range is 100 (the minimum being 0 and the maximum being 100) by splitting the range in half e.g. above and below 50 this would result in two very unequally sized groups as 99.7% of the data is below 50. What would happen if this variable were used in a classification using the k-means procedure? The easiest way to test this out is to run a simple cluster analysis, this was done on just two clusters to aid simplicity and comparability to previous reflection on how it should be split. The results of the clustering put 217,895 OAs (97.7%) in one group and 5,165 OAs (2.3%) in the other, the point at which they have been split is above 15.2, but this does not reflect any actual split in the data. This is not to say that arbitrary splits do not occur in all variables just that the extreme nature of the skew in this variable makes this an especially acute example.

By increasing the number of groups to be produced the extent of the problem will become apparent. For purposes of illustration the data was clustered into 50 clusters, a number that is not unusual for the lower level of a classification. If the data were normally distributed one would expect to find about 2% of the areas in each cluster, but this variable has 88% of data that cannot be split further as they all have the same value, so what is actually being classified is the remaining 12% of the data into 49 clusters. The result is, one group that contains 88% of the data and the rest are spread about with 27 of the groups containing less than 100 areas. If split evenly each group should have around 4,500 members. Groups with small memberships are problematic because this is how outliers are formed. If several highly skewed variables are included in the classification and an area appears in a small group for more than one of those variables, it is easy to see how micro clusters of single figure membership can be formed. With 223,060 OAs to cluster, producing such small clusters is of little practical value.

One solution to this kind of problem is to transform the data. Common transformations that are used to combat this type of problem are: logarithmic transformations, square rooting the data or converting the data to ranks (Harris *et al*. 2005). Variables with highly skewed distributions or large numbers of zeros were therefore frowned upon in the variable selection.

5.3.6 Composite Variables

Composite variables can be formed from two related variables which show comparable patterns. These variables have to share the same denominator (otherwise the proportion of people relating to that variable could exceed 100%). This method can be used to group together highly correlated variables or variables which only represents a small proportion of society. Examples of variables for which this method has been used are grouping separated and divorced people together, and combining all the different varieties of flats into a single all flats variable. This increases the sample size on which the variable is based and increases the reliability of the data. This is especially important when working with OAs because the numbers can be small and affected by disclosure controls imposed on the data. For an explanation of what disclosure control entails and the effect it has on the data see ONS (2003b).

5.3.7 Geographic Constancy of Variables

Some variables that show interesting geographic variations were not available in all four countries of the UK. For example, the Knowledge of the Welsh language question was only asked of residents of Wales. Some questions were asked in all countries but their results were reported in different ways. A good example of this is the religion question, where in Northern Ireland the results were reported by splitting the data into several different categories of Christian and Other Religion variable, in which all other religions were combined. In England and Wales the situation was reported in the opposite way round by reporting all types of Christians in a single variable and reporting other religions separately e.g. Buddhists, Hindus, Jews, Muslims and Sikhs each as a separate variable.

Another geographic inconsistency in the Religion table is that it has only just been reintroduced to the Census in England and Wales in 2001 (last included in 1851), whereas it has previously been asked in Scotland and Northern Ireland. However, it was only introduced on a voluntary basis in England and Wales. Consequently 7.71% of the population of England and Wales did not answer the religion question. Whereas it is a compulsory question in Scotland and Northern Ireland, this makes it difficult to compare the variables across the UK because high rates of religious affiliation observed in Scotland and Northern Ireland would be attributable to the voluntary nature of the question in England and Wales. Some interesting patterns maybe visible, but if data are not available for all parts of the UK it is not possible to include the variable in the classification.

5.3.8 Vague or Uncertain Variables

It would seem sensible to assume that all census variables are collated in the same way i.e. from the answers written on each census form. However, this is not the case for all variables. Examples are the ‘household spaces with no residents’ variables on table KS16 that are coded as either ‘Vacant’ or ‘Second residence/holiday accommodation’. Unlike other census variables there was no-one to fill in a form for these variables because all the properties were empty on census day. The variables were imputed by the census enumerator making a deduction of whether the property was ‘Vacant’ or ‘Second residence/holiday accommodation’ based on their own judgement. It is widely accepted that ‘Second residence/holiday accommodation’ was under recorded using this method, especially in the more rural parts of the country.

Brown (2005) doubts the reliability of the number of second homes in the 2001 Census for Cornwall. According to the census the number fell from 11,550 in 1991 to 10,500 in 2001 which seems unlikely with the continuing trend for people to buy second homes in the area over that period. Office of the Deputy Prime Minister (ODPM) tax register figures for the number of second homes in the county suggest the real number is over three times that given in the 2001 Census (Brown 2005). The posting back of census forms could account for some of this because forms delivered to second homes would only be sent back if the owner happened to be there at the time. The homes may have been imputed as permanent residences. Brown (2005) cautiously suggests that there are at least 50% more second homes in Cornwall than were identified by the 2001 Census.

5.3.9 Uninteresting Geographic Distribution of Data

For variables to work in the classification they need to show variation over space; otherwise a distinction between areas cannot be made. Not all ethnic groups show the same distribution over space. Some are distributed fairly evenly others show a more ghettoised population. Peach (1996) explores this phenomenon by asking the question; 'Does Britain have ghettos?' to investigate to what extent different ethnic groups are dispersed throughout Britain. Table 5.4 shows the percentage of each ethnic group present in the major metropolitan areas of England.

Table 5.4: Percentage of ethnic group in London, W Midlands, G Manchester & West Yorkshire

Ethnicity	Percentage of group present
White	22.6
Black Caribbean	79.0
Black African	82.7
Black Other	62.8
Indian	65.8
Pakistani	64.2
Bangladeshi	74.5
Chinese	47.7
Total Population	25.0

Adapted form Peach 1996 p219 source: OPCS

Table 5.4 shows that for all groups apart from White and Chinese, over 60% of that group are found in the four major urban centres, these ethnic groups have a distinct urban pattern to their distributions in contrast, White and Chinese populations vary less significantly over space. Black Caribbean, Black African, Black Other, Indian, Pakistani and Bangladeshi variables would add more to the classification than White or Chinese variables because their distributions vary more over space. Segregation research is brought up to date by Stillwell (2005) who investigated the segregation of ethnic groups in Britain using data from the 2001 Census. Stillwell (2005) calculated segregation indices that showed that the Chinese to be the most integrated ethnic group in the UK with a segregation index of 0.32 in comparison to White 0.52, Indian 0.57, Pakistani 0.56, Black 0.65 and other 0.43. This suggests that Chinese is not a good variable to use as comparatively the percentage of Chinese people in an output area gives little information about an area because they are well integrated within the population as a whole, and therefore does not act as a good predictor of other attributes of that area.

5.3.10 Consistency of the Variable for the Life Time of the Classification

The longevity of the classification has to be considered as the classification is likely to remain the most current ONS area classification until after the release of the 2011 Census results. Any variable whose understanding by the user may change over the life course of the classification should not be included as it may cause confusion. What does this mean? A variable that was considered for use in the classification was born in other European Union (EU) (excluding UK and Republic of Ireland). On Census day April 29th 2001 there were 15 members of the EU; on the first of May 2004 Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia and Slovenia joined increasing membership to 25 countries. The consequence of this is that the Born in other EU variable in the census no longer reflects the current membership of the EU. There are also applications in to join from Bulgaria, Croatia, Romania, Turkey and Macedonia. If and when these countries join the number of member countries of the EU will have doubled from the time of the 2001 census. It is therefore easy to see how the inclusion of this variable would cause increasing confusion over time, as the user maybe unaware of either the time at which the data was created or the changing membership of the EU.

5.3.11 Standardisation of Limiting Long Term Illness

The percentage of the population suffering from limiting long-term illness (LLTI) as provided in the Key Statistics table KS08 could have been used in its raw form, as it was in the LA classification. However, it was considered that this was unsatisfactory as crude rates are greatly affected by the age structure of the population at fine geographic scales. This would therefore result in an area which has a high proportion of older people (taking all other things to be equal) to have a much higher illness rate than an area with a younger population. The effect will be greater for OAs than for higher level geographies, because their relatively small size increases the likelihood of there being OAs with a very old age structure. Such areas will without standardisation be classed as being areas of above average illness based as much on their age structure as the intensity of ill health.

It is therefore necessary to standardise the LLTI data by age to counteract for the influence that age structure has over the crude illness rate. Only when this is done will the relationship of illness with other variables become clear. The technique used to do this is the indirectly Standardised Illness Ratio (SIR). SIR works by comparing the expected illness count for an area with the observed count. The expected count is created by multiplying age-specific illness rates for the whole UK population by the OA population by age. We can then see if the illness rate is higher or lower than the national norm.

The SIR for an area is defined as follows:

$$SIR^i = 100 \times (I^i / \Sigma_a r_a^n P_a^i) \quad (5.1)$$

Where I^i = observed count of ill people in area i, r_a^n = rate of illness for age group a in the national population and P_a^i = population in area i of age group a.

The SIR is a relative measure. The national illness rate always has the value 100, a value of 150 means that an OA experiences 50% more illness than it would have if the age-specific rates for the standard population. A value of 50 means the OA experiences 50% less illness than the UK population. There is substantial variation between the OAs with values ranging from 0 to 505. The healthiest areas are OAs with SIRs below 70 and the least healthy are OAs with SIRs exceeding 130.

5.3.12 The Process of Variable Selection

The previous sections discussed reasons why variables may be dropped from the initial list; the following section discusses some of the decisions that were made in reducing the initial variable list to create the final list. Ninety four variables were included in the initial set of variables for consideration. The final list is composed of just 41 so a large number of variables have been rejected or combined with others. This section outlines what decisions have been taken in the reduction of the variable list by 53 variables. An attempt will be made to account for all the decisions made. However, these decisions are very complex: the decision as to which variables to include was made by comparing all variables to all variables. For many of the variables it is not as simple as giving a single reason (such as a high correlation with another variable). In many cases a variable may have a significant relationship with tens of other variables. All these relationships were examined to assess a variable’s suitability for selection. It is impossible to report on all the relationships within the dataset which account for the decisions made. However, an effort will be made to give reasoning behind all decisions made.

A further point to take into account is that the variable choice was done in conjunction with the team from the ONS who were creating the ward level classification. This joint effort was intended to match as closely as possible the variable selections at both scales. This was done with the intention of making the classifications as simple and comparable as possible for users to understand. The comparability across scales is an important part of the project, the area classification systems that are being created are to be disseminated as a suite of systems to be used together or from which one is selected that an individual feels is the most appropriate for their use. Chapter 8 outlines how these systems have been linked together to create a multi-scale system. Within the process of variable selection some sacrifices were made at one scale to aid comparability with the other. This is an issue that needs to be considered when reviewing the reasons for certain decisions.

The reasons for variable selection will be reviewed in the order in which they appear in Table 5.2. Both male and female variables were rejected as it was felt that gender told us very little about an area. Looking at the data it was found that the majority of areas had very similar numbers in terms of gender mix. It was very unusual for an OA to be dominated by one or other gender.

It was decided not include the proportion of people who live in a communal establishments as there are a lot of areas with a zero value for this variable. Inclusion could lead to things being grouped together because of an absence of something rather than a presence. Some areas did have very high proportions of people living in communal establishments, e.g. student residences. "Communal establishments" is a vague term that covers residences for several different population groups, including care homes, hostels, prisons and university residences. These house very different types of people with little in common who would be grouped together with the inclusion of this variable.

As an Urban/Rural indicator was not available at the time of classification, Population Density was used as a proxy. Density has the advantage of being a continuous scale variable. It was decided that this should be kept as there is little else in the list of variables which gives such a distinction between urban and rural areas.

Some changes were made in the age variables: the youngest age group (0-15) was spit into two variables 0-4 and 5-14 to pick up the difference between younger and older children, 16-24 was changed to 15-24 to match the ward level classification but was then dropped as it was highly correlated with students. The age variables, ages 25-44, 45-64 and 65+ were all retained.

Married, cohabiting and single were not included as variables as they had a strong relationship with other family variables such as single person households and two adults with no children. Divorced was combined with separated, which brought more detail into the variable but also covered the problem of divorce not being allowed in certain religions (e.g. Catholic). These people will report their marital status as separated rather than divorced, by combining the variables these people would be included.

Percentage of people born outside the UK was kept as a variable as it gave an indication of international migration. Indian, Pakistani and Bangladeshi was kept as was percentage Black as they showed an interesting geographic distribution and identified significant minority populations within the UK. Chinese was not included as their geographic distribution showed much less variation across the UK in comparison to other ethnic groups. All of the religion variables were dropped due to a high correlation with ethnicity and the voluntary nature of the question in England and Wales.

Two of the health variables that were considered were included. Limiting Long Term Illness (LLTI) was included but it was standardised by age creating a Standardised Illness Ratio (SIR), rather than using percentage of working age population. This enabled 100% of the population to be used which is important as the OAs are small areas. As age distribution of some areas may be mainly outside the working age population, using percentage of working age population may not be reliable for some areas with a high elderly population, although this was considered suitable for the ward level classification as the areas and therefore the population are significantly larger. People whose health is good, fairly good and not good were all found to be highly correlated with LLTI. The other health variable that was included was percentage of people who provide unpaid care as this gave an indication not only of the general health of the area but combined with the LLTI variable would give an indication of how well people are cared for.

People working part-time and people unemployed were included; those working full time were not due to a correlation with other employment variables; self employed was dropped as it was highly correlated with people who work from home which was considered to be a more distinct group. The full time students variable and economically inactive looking after the family and home were included as they represent two distinct groups in society.

Of the twelve industry sector groups in the original list seven (Agriculture, Hunting, Forestry and Fishing employment; Mining, Quarrying and Construction employment; Manufacturing employment; Hotel and Catering employment; Health and Social Work employment; Financial Intermediation employment; and Wholesale and Retail trade employment) were included as they showed interesting geographic patterns. The other five (Electricity, Gas and Water supply employment; Transport, Storage and Communication employment; Real Estate, Renting and Business Activities employment; Public Administration and Defence employment; and Education employment) were rejected for less distinctive geographic distributions, inter correlations and limited representation in terms of numbers.

The nine occupation groups, numbers 44-52 in Table 5.2 were not selected as they were correlated with the industry sector variables and the education and the socio-economic classification variables. Of the education variables people with qualification level 4 and 5 (degree level and above) were included; no qualification was not, as it was correlated with other indicators of deprivation and low social standing such as unemployment.

Most of the data in the socio-economic class domain, numbers 45-62 in Table 5.2 were highly correlated with other variables such as employment, qualifications, ethnicity and health especially at the higher end of the scale. The only two variables from the original list that were included were semi-routine occupations and routine occupations which were combined together to give an extra variable indicating lower social standing.

Never worked and long-term unemployed were not included as they only identified small sections of the population and were highly correlated with unemployment. Work from home was included as it represents an increasing trend within society. Public transport to work was included as it showed some interesting geographic patterns; walk to work, and car or van to work were not selected as they were correlated with public transport and showed less interesting patterns.

Renters from both the private and public sector are included as they give indicators of several things including stage of the life course, transitoriness and wealth. The second residence/holiday accommodation variable was not kept as this was not an actual question on the census form. These data were created from the enumerator's assessment of each household. It is generally recognised that these data are unreliable, especially at such a small scale.

Detached and terraced housing variables were included; semi-detached housing was not included as it was highly correlated with other housing types and was less descriptive. It also does not represent such a distinct group as terraced or detached. Purpose built flats, converted flats and flats in commercial buildings were combined to create the all flats variables. Caravan or temporary structure accommodation was rejected as it only accounted for a very small part of the population.

The variable 2+ cars was included in preference to no car households because the two variables are very highly correlated, but 2+ cars was selected to add additional information on affluence.

Average household size was rejected as it did not reveal information about a distinct type of household; the average number of rooms per household was included as it gave a good indication of the affluence. OAs with an occupancy rating of -1 or less was rejected in favour of a new variable people per room. No central heating was included as it is a good indicator of poor living conditions, but no bath or shower was rejected as the numbers are very small. Lowest floor above ground level was not included as was highly correlated with flats.

Single pensioner households and single person non-pensioner households were both included as they identify a housing situation which is of increasing prevalence. All pensioner households (family) this was rejected as it was highly correlated with single pensioner households and age 65+. Two adults no children and lone parent households were both included as they show fascinating opposing residential situations. All student households was rejected as it is highly correlated with students. All pensioner (other) household was rejected due to correlation with similar variables. No adult in employment with dependent children was not included as it was highly correlated with lone parents. A new variable households with non-dependent children was included, to identify a new and increasing section of society which sees children living with their parents for longer because of the difficulty they experience trying to get on to the housing ladder.

The percentage of unemployed who are long-term unemployed, was used in the ward classification, but could not be used in the OA classification due to the effect of disclosure control on the data. Several OAs reported values of over 100% when values for this variable were calculated. The reduction in scale from wards to OAs makes this kind of effect much more likely as the population numbers are much smaller. Because of the obvious errors present in the variable created we could have little confidence in this variable and it was therefore not used in the OA classification.

The decisions on the variables to include in the classification were made by the ONS/Leeds group comparing statistical information and working within a theoretical framework of which types of variables were should be included in the classification. However, like with any choice of variables for a classification a different group of people may have made different decisions resulting in a different variable list.

5.3.13 The Final List of 41 Variables Used in the Classification

Table 5.5 lists the 41 variables selected for input to the classification, gives them a short definition and a longer verbal description. This final list of variables results from the implementation of the decisions made. Variables will often be referred to in the text only by number for brevity; Table 5.5 can be used as a look up in these cases.

Table 5.5: Full list of 41 variables selected for input to the classification,

Demographic	
v1	Age 0-4: *Percentage of resident population aged 0-4*
v2	Age 5-14: *Percentage of resident population aged 5-14*
v3	Age 25-44: *Percentage of resident population aged 25-44*
v4	Age 45-64: *Percentage of resident population aged 45-64*
v5	Age 65+: *Percentage of resident population aged 65+*
v6	Indian, Pakistani or Bangladeshi: *Percentage of people identifying as Indian, Pakistani or Bangladeshi*
v7	Black African, Black Caribbean or Other Black: *Percentage of people identifying as Black African, Black Caribbean or Other Black*
v8	Born outside the UK: *Percentage of people not born in the UK*
v9	Population Density: *Population Density (number of people per hectare)*
Household Composition	
v10	Separated/Divorced: *Percentage of residents 16+ who are not living in a couple and are separated/divorced*
v11	Single person household (not pensioner): *Percentage of households with one person who is not a pensioner*
v12	Single pensioner household: *Percentage of households which are single pensioner households*
v13	Lone Parent household: *Percentage of households which are lone parent households with dependent children*
v14	Two adults no children: *Percentage of households which are cohabiting or married couple households with no children*
v15	Households with non-dependant children: *Percentage of households comprising one family and no others with non-dependent children living with their parents*
Housing	
v16	Rent (Public) : *Percentage of households that are public sector rented accommodation*
v17	Rent (Private): *Percent of households that are private/other rented accommodation*
v18	Terraced Housing: *Percentage of all household spaces which are terraced*
v19	Detached Housing: *Percentage of all household spaces which are detached*
v20	All Flats: *Percentage of households which are Flats*
v21	No central heating: *Percentage of occupied household spaces without central heating*
v22	Average house Size: *average house size (rooms per household)*
v23	People per room: *The average number of people per room*
Socio-Economic	
v24	HE Qualification: *Percentage of people aged between 16 - 74 with a higher education qualification*
v25	Routine/Semi-Routine Occupation: *Percentage of people aged 16-74 in employment working in routine or semi-routine occupations*
v26	2+ Car household: *Percentage of households with 2 or more cars*
v27	Public Transport to work: *Percentage of people aged 16-74 in employment usually travel to work by public transport*
v28	Work from home: *Percentage of people aged 16-74 in employment who work mainly from home*
v29	LLTI (SIR): *percentage of people who reported suffering from a Limiting Long Term Illness (Standardised Illness Ratio, standardised by age)*
v30	Provide unpaid care: *Percentage of people who provide unpaid care*
Employment	
v31	Students (full-time): *Percentage of people aged 16-74 who are students*
v32	Unemployed: *Percentage of economically active people aged 16-74 who are unemployed*
v33	Working part-time: *Percentage of economically active people aged 16-74 who work part time*
v34	Economically inactive looking after family: *Percentage of economically inactive people aged 16-74 who are looking after the home*
v35	Agriculture/Fishing Employment: *Percentage of all people aged 16-74 in employment working in agriculture and fishing*
v36	Mining/Quarrying/Construction Employment: *Percentage of all people aged 16-74 in employment working in mining, quarrying and construction*
v37	Manufacturing Employment: *Percentage of all people aged 16-74 in employment working in manufacturing*
v38	Hotel & Catering Employment: *Percentage of all people aged 16-74 in employment working in hotel and catering*
v39	Health and Social work Employment: *Percentage of all people aged 16-74 in employment working in health and social work*
v40	Financial intermediation Employment: *Percentage of all people aged 16-74 in employment working in financial intermediation*
v41	Wholesale/retail trade Employment: *Percentage of all people aged 16-74 in employment working in wholesale/retail trade*

5.3.14 Weighting of Variables

The role of weighting variables in the current classification is simple; they will all be set to 1 (equal weighting for all variables). There are several reasons for this. The classification is for general purpose use. By weighting a variable higher than another, this could make the classification more suitable for one purpose than another. As discussed in § 3.3.5 there are all sorts of weightings going on within the data due to inter-correlation that are difficult to quantify. By adding weightings to some or all variable it is difficult to predict what the effect may be. There is no perfect solution and there is no reliable way of telling if adding one set of weights or another set of weights has improved the classification. By not using weights but rather being more selective in the variable choice the process of classification can be made much simpler. The classification could be reproduced in a different form, targeted at a more specific purpose by weighting some variables higher than others.

5.3.15 Database Assembly

To be able to cluster the OAs into groups the data about them all needs to be in one database. This sounds sensible and simple enough. However, for each Key Statistics table there are twelve separate tables that need joining together: nine representing the English Government Office Regions (GORs), one for Wales, one for Scotland and one for Northern Ireland. The data were published in this way because to put the data into one file would have made it too big to be opened in the most commonly used statistical package Microsoft Excel. Also few users would require the use of data at such a fine scale for the whole country. The tables could not simply be joined one on top of the other because in some cases the formats of the tables were different in each of the countries of the UK. So to do this data extraction, a computer program was built so that the data needed could be extracted from each table and output to a single file.

Before this can be done the exact source of the data to create each variable must be carefully recorded. The full list of table and references for the 41 variables used in the classification is given in Table 5.6. The columns in Table 5.6 contain the following entries: Variable Number is a number that has been given to each variable for the purposes of classification as a quick reference they can be related back to the names and descriptions in Table 5.5. E & W Table refers to the name of the table in England and Wales. E & W Ref is the reference calculation to extract the data from the tables for England and Wales, the numbers refer to the columns of data within the original census table. Scot Table and NI Table represent the same as E & W Table but for Scotland and Northern Ireland respectively. Scot Ref and NI Ref represent the same as E & W Ref, but for Scotland and Northern Ireland respectively. England and Wales, Scotland and Northern Ireland have to be done separately in this way, because there are differences between the layout and design of the tables in the three censuses. Anybody working with the census for the whole of the UK will find they have this problem. It is a very time consuming process to standardise across all areas, but it is vital to ensure the same data are used for all constituent parts of the UK.

Table 5.6: Full variable definitions and sources
(specified in terms of Key Statistics Table and column number)

Variable Number	Table	E&W Table	E&W Ref	Scot Table	Scot Ref	NI Table	NI Ref
v1	KS02	e00201a,b,d,e,f,g,h,j,k,w	(2/1)*100	KS02	2	KS02OA	2
v2	KS02	e00201a,b,d,e,f,g,h,j,k,w	((3+4+5)/1)*100	KS02	3+4+5	KS02OA	3+4+5
v3	KS02	e00201a,b,d,e,f,g,h,j,k,w	((10+11)/1)*100	KS02	10+11	KS02OA	10+11
v4	KS02	e00201a,b,d,e,f,g,h,j,k,w	((12+13)/1)*100	KS02	12+13	KS02OA	12+13
v5	KS02	e00201a,b,d,e,f,g,h,j,k,w	((14+15+16+17)/1)*100	KS02	14+15+16+17	KS02OA	14+15+16+17
v6	KS06	e00601a,b,d,e,f,g,h,j,k,w	((9+10+11)/1)*100	KS06	6+7+8	KS06OA	5+6+7
v7	KS06	e00601a,b,d,e,f,g,h,j,k,w	((13+14+15)/1)*100	KS06	11+12+13	KS06OA	9+10+11
v8	KS05	e00501a,b,d,e,f,g,h,j,k,w	((6+7+8)/1)*100	KS05	6+7+8	KS05OA	6+7+8
v9	KS01	e00101a,b,d,e,f,g,h,j,k,w	Area From shape files	KS01	11	KS01OA	7
v10	KS04	e00401a,b,d,e,f,g,h,j,k,w	(5+6/1)*100	KS04	5+6	KS04OA	5+6
v11	KS20	e02001a,b,d,e,f,g,h,j,k,w	(3/1)*100	KS20	3	KS20OA	3
v12	KS20	e02001a,b,d,e,f,g,h,j,k,w	(2/1)*100	KS20	2	KS20OA	2
v13	KS20	e02001a,b,d,e,f,g,h,j,k,w	((11+12)/1)*100	KS20	11+12	KS20OA	11+12
v14	KS20	e02001a,b,d,e,f,g,h,j,k,w	((5+8)/1)*100	KS20	5+8	KS20OA	5+8
v15	KS20	e02001a,b,d,e,f,g,h,j,k,w	((7+10+12)/1)*100	KS20	7+10+12	KS20OA	7+10+12
v16	KS18	e01801a,b, d,e,f,g,h,j,k,w	(5+6/1)*100	KS18	5+6	KS18OA	5+6
v17	KS18	e01801a,b,d,e,f,g,h,j,k,w	(7/1)*100	KS18	7+8	KS18OA	7
v18	KS16	e01601a,b,d,e,f,g,h,j,k,w	(6/(3+4+5+6+7+8+9+10))*100	KS16	10	KS16OA	10
v19	KS16	e01601a,b,d,e,f,g,h,j,k,w	(4/(3+4+5+6+7+8+9+10))*100	KS16	8	KS16OA	8
v20	KS16	e01601a,b,d,e,f,g,h,j,k,w	((7+8+9)/(3+4+5+6+7+8+9+10))*100	KS16	11+12+13	KS16OA	11+12+13
v21	KS19	e01901a,b,d,e,f,g,h,j,k,w	((6+7)/1)*100	KS19	6+7	KS19OA	6+7
v22	KS19	e01901a,b,d,e,f,g,h,j,k,w	3	KS19	3	KS19OA	3
v23	KS19	e01901a,b,d,e,f,g,h,j,k,w	2/3	KS19	2/3	KS19OA	2/3
v24	KS13	e01301a,b,d,e,f,g,h,j,k,w	(6/1)*100	KS13	6	KS13OA	6+7
v25	KS14	e01401a,b,d,e,f,g,h,j,k,w	((8+9)/1)*100	KS14	8+9	KS14OA	8+9
v26	KS17	e01701a,b,d,e,f,g,h,j,k,w	((4+5+6)/1)*100	KS17	4+5+6	KS17OA	4+5+6
v27	KS15	e01501a,b,d,e,f,g,h,j,k,w	((3+4+5+9)/1)*100	KS15	3+4+5+9	KS15OA	3+4+8
v28	KS15	e01501a,b,d,e,f,g,h,j,k,w	(2/1)*100	KS15	2	KS15OA	2
v29	KS08	e00801a,b,d,e,f,g,h,j,k,w	(2/1)*100	KS08	2	KS08OA	2
v30	KS08	e00801a,b,d,e,f,g,h,j,k,w	(7/1)*100	KS08	(7/1)*100	KS08OA	(7/1)*100
v31	KS09	e00901a,b,d,e,f,g,h,j,k,w	(6+8/1)*100	KS09	6+8	KS09OA	6+8
v32	KS09	e00901a,b,d,e,f,g,h,j,k,w	(5/1)*100	KS09	5	KS09OA	5
v33	KS09	e00901a,b,d,e,f,g,h,j,k,w	(2/1)*100	KS09	2	KS09OA	3
v34	KS09	e00901a,b,d,e,f,g,h,j,k,w	(9/1)*100	KS09	9	KS09OA	9
v35	KS11	e01101a,b,d,e,f,g,h,j,k,w	((2+3)/1)*100	KS11	2+3	KS11OA	2
v36	KS11	e01101a,b,d,e,f,g,h,j,k,w	((4+7)/1)*100	KS11	4+7	KS11OA	5
v37	KS11	e01101a,b,d,e,f,g,h,j,k,w	(5/1)*100	KS11	5	KS11OA	3
v38	KS11	e01101a,b,d,e,f,g,h,j,k,w	(9/1)*100	KS11	9	KS11OA	7
v39	KS11	e01101a,b,d,e,f,g,h,j,k,w	(15/1)*100	KS11	15	KS11OA	13
v40	KS11	e01101a,b,d,e,f,g,h,j,k,w	(11/1)*100	KS11	11	KS11OA	9
v41	KS11	e01101a,b,d,e,f,g,h,j,k,w	(8/1)*100	KS11	8	KS11OA	6

The England & Wales data differed from the Scotland and Northern Ireland data as they had to be converted to percentages, whereas the Scotland and Northern Ireland data were available as percentages so this calculation was not necessary. Whilst constructing the database some variables were more problematic than others; v33 working part time presented a particular oddity. The variable is in the second column of table KS09 for England & Wales and Scotland, but is in the third column in Northern Ireland. In the Northern Ireland table working part time and working full time are in the opposite order to the order for England & Wales and Scotland tables. There appears to be no reason or explanation for this. Looking down Table 5.6 there are relatively few variables that have the same table reference for all three censuses making the construction of the database a tricky and time consuming process that required constant checking and rechecking to ensure the correct data had been selected.

5.3.16 A Program Used to Extract the Variables

Before any analysis can take place the data need to be agglomerated into one file which can be opened in the SPSS statistical package. SPSS is able to handle the number of data rows required for the whole UK to be held in one file.

An extraction program; written in FORTRAN; was developed with Mark Birkin to automate this process. This was done for two reasons: firstly, to vastly speed up the process of creating the database and secondly, to remove human error from the process which would have been potentially a problem if the data were copied and pasted into the database. The program went thought several versions before the fifth version successfully handled the intricacies of the census tables. The design of the program was made more difficult by the differences between the formats and design some of the tables between the three census agencies. The program reads in data from raw data key statistic table files in comma separated variable format, performs any necessary automatic calculations and writes out the subset of variables needed to output files. A separate text file is created for each variable. These files are then inputted to the SPSS package and merged into a single database. The FORTRAN code for the census data extraction program can be found in Appendix B.

5.3.17 Data Checking

Data checking is a vital part of the creation of the database; if the data are entered into the database incorrectly everything that is done subsequently will therefore be incorrect. A great effort was made to identify any errors in the database. The nature of the creation of the classification means that a mistake at any point in its creation means that everything after that point will contain errors and will need to be redone, causing a great deal of time to be lost. Two different forms of data checking were conducted on the database to ensure that the correct values were being used.

The first form of data checking was to test variable values for individual OAs, to establish if the data in the database matched the data in the original census tables. This check essentially tested the reliability of the data extraction program and its ability to extract the correct data in the correct order. The database is assembled from 12 tables (as they are split by GOR) and 41 variables. Therefore to test that each table was extracted and re-assembled correctly, a check on data for each GOR for each variable must be done, a minimum of 492 (12×41) separate checks must be made to ensure that the data were entered correctly. As the data were extracted automatically it can be assumed that if one item is wrong then everything extracted from that table is wrong. However, to add more rigour to the test the same OA was not selected each time. Every two thousandth OA was selected (including the first and last) to form a list of 112 OAs from which one from each GOR would be selected to test for each variable. For each of the checks the calculation done by the extraction program was redone manually by locating the relevant OA and variable from the original census tables and then comparing its value to the value in the database for the same OA for the same variable.

Table 5.7 shows a selection of the results of the data checking procedure. The results show that 446 of the 492 variables checked showed a difference of 0.0 and 46 of the 492 showed a difference of plus or minus 0.1 when rounded to 1 decimal place. The differences of 0.1 are not because of errors, but due to the fact that during the calculations in the extraction program it worked to only 1 decimal place and that when the data were checked variables were often represented using more than one decimal place, accounting for small differences between the two sets of figures. It has also been noticed during calculations in this project and by the ONS team who were building the ward level classification, that some internal rounding processes that take place within SPSS are difficult to assess accurately. The difficulty is that the SPSS program does not always make calculations using the number of decimal places that could be expected (the number displayed on the screen). It was therefore concluded that the small differences the data checking process showed could not be attributed to errors in assigning the data from the original census tables to the database.

Table 5.7: Example of the Data checking results

Variable Number	OA Code	OA Order Code	GOR	Data Check Code	Value in Database	Checked Value	Difference
V15	35UDHH0001	8001	North East	5	6.6	6.6	0.0
V15	00BMFR0013	16001	North West	9	4.9	4.9	0.0
V15	00CZFP0032	42001	Yorkshire and The Humber	22	9.5	9.5	0.0
V15	00FYNH0022	50001	East Midlands	26	5.9	5.9	0.0
V15	41UKFR0016	76001	West Midlands	39	9.3	9.3	0.0
V15	26UCHJ0009	90001	East of England	46	6.6	6.6	0.0
V15	00APGB0037	100001	London	51	10.0	10.0	0.0
V15	00MGPA0001	126001	South East	64	13.0	13.0	0.0
V15	00HBPJ0023	150001	South West	76	8.2	8.1	0.1
V15	00PRMX0009	174001	Wales	87	8.7	8.8	-0.1
V15	60QU000273	204001	Scotland	102	12.5	12.5	0.0
V15	95ZZ160009	223060	Northern Ireland	112	12.2	12.1	0.1

The second form of data checking involved the entire database. The aim was to compare the values in the database with the values for higher levels of geography. It was decided that the level of geography to compare the data to should be GORs in England plus Wales, Scotland and Northern Ireland. This check tested both the ability of the extraction program to reproduce the data in the correct order and this provided a check of the OA data against a different level of geography. This set of data checks involved multiplying out the data in the database (in percentages) by the population of each OA (e.g. total population, number of households, people of working age etc.), then summing all the OAs in each GOR/Country and then checking the value against that of the GOR/Country to ensure the numbers correspond to a reasonable level of accuracy to the value given for the GOR/Country in the census table. Some error is unavoidable due to rounding when multiplying out the data and the effects of disclosure control. Table 5.8 shows an example of the results of this data checking.

Table 5.8: Example of the Data checking results (for the North East GOR)

Variable	Observed	Expected	Difference	Difference, people /houses	Variable	Observed	Expected	Difference	Difference, people /houses
v1	5.50	5.50	0.003	5	v22	5.19	5.19	0.003	n/a
v2	12.92	12.92	-0.005	-15	v23	0.45	0.45	-0.003	n/a
v3	28.01	28.01	0.004	29	v24	14.97	14.97	0.002	6
v4	24.54	24.54	0.003	19	v25	23.90	23.89	0.005	22
v5	16.55	16.56	-0.013	-54	v26	20.98	20.98	0.000	-1
v6	1.21	1.21	0.000	0	v27	14.69	14.69	0.004	6
v7	0.16	0.16	-0.003	0	v28	7.68	7.68	-0.002	-2
v8	2.93	2.94	-0.014	-11	v29	22.73	22.73	-0.003	-15
v9	2.93	2.93	-0.002	n/a	v30	11.00	11.00	0.002	6
v10	10.93	10.93	-0.005	-10	v31	7.01	7.01	-0.005	-6
v11	15.10	15.10	0.002	3	v32	4.53	4.53	0.004	3
v12	15.64	15.64	-0.004	-6	v33	11.87	11.87	0.004	9
v13	10.75	10.76	-0.009	-10	v34	6.58	6.58	-0.004	-5
v14	16.87	16.87	0.002	3	v35	1.16	1.17	-0.013	-2
v15	10.61	10.63	-0.023	-26	v36	7.87	7.88	-0.013	-11
v16	27.65	27.64	0.015	44	v37	16.99	16.99	-0.001	-2
v17	6.28	6.28	-0.002	-2	v38	5.10	5.10	-0.004	-2
v18	32.10	32.10	-0.001	-3	v39	12.74	12.74	0.001	2
v19	14.50	14.50	-0.002	-4	v40	3.04	3.04	-0.002	-1
v20	13.92	13.92	0.001	2	v41	16.19	16.19	-0.001	-2
v21	3.95	3.94	0.006	2					

Table 5.8 shows only very small errors which can be explained by rounding or disclosure controls. However, three of the GORs (Eastern, South East and London) showed very large differences for one variable, v30 percentage of people who provide unpaid care. Each GOR was found to have approximately 500,000 people missing from the OA data compared to the GOR/country data. At this point much checking was done of the tables. It was found that the differences were not in the database, but between the original census tables at the two different scales. But which was wrong? Which was right? This was fairly simple to deduce that the GOR tables showed a similar level for the variable across all GORs whereas in the OA data the level was significantly lower in the three GORs in which the discrepancies were found in comparison to the other nine GORs. It was therefore safe to conclude that the errors were contained in the original published census data at OA level. The errors were reported to the ONS who supplied new corrected tables. The new data were added to the database and checked again. This time no significant differences were found between the data at the two different geographic scales. An exercise that had been designed to find errors in the inputting of data into the database for

classification had found that the only errors in the database were not down to input errors during the creation of the database, but errors in the original census data.

This brought about an issue that had previously not been discussed: does all census data need to be checked against another level of geography before it is used? This is a problem that will reduce with time as errors in the data are found and new data issued. However, if you downloaded the original release of census data no errors within the dataset will have been corrected. It would therefore be sensible for any intensive user to keep checking the census agencies and dissemination units' websites for known errors and download and replace the relevant data when they become available. By doing this the chances of errors in the data are significantly reduced. It may also be worth reordering data a year or so after its original release by which time errors are likely to have been found and corrected.

These data checks are not 100% fool proof, but without checking all nine million data points in the clustering database this would be difficult to achieve. However, the data checks do provide proof that it is unlikely that any errors remain in the dataset. The checks were designed to find errors both by checking back to the original OA data and against data at another scale to see if the values were consistent. The error that was picked up shows that the data checking worked, in terms of finding a major error in the dataset. However, tiny individual errors could slip though, but would be almost undetectable. The data extraction program was an automated process and worked very smoothly. The spot checks did not find any errors produced by the data extraction procedure so the likelihood of any errors is small.

5.4 Processes

Now that the final variable list has been constructed the process of creating the classification can begin. This firstly involves standardising the variables to account for difference in scale between the variables. Then cluster the data using cluster analysis techniques to produce the structure of the classification and split the areas into their groups of similarity. This section outlines the methodological problems that were experienced and how they were eventually overcome.

5.4.1 Variable Standardisation

Before any clustering can be done the variables need to be standardised over the same range. This ensures that each variable has the same weighting on the classification. This is especially important when there are different types of data e.g. population density will give number of people per an area, whereas detached housing is a percentage of all households. The range of the population density is only limited by the number of people who can fit into a specified area. For OAs in the UK ranges from just above 0 to 12,715 people per hectare whereas housing type can only range between 0 and 100%. These variables are not on the same scale. If left un-standardised the population density would completely control the classification because of the larger range of which the data are stretched over. This would also create a large number of outliers based solely on the population density variable. Therefore f these variables were clustered without being standardised it would add bias to the dataset. Methods of standardisation were discussed in § 3.3.1.

All clustering techniques are based on the similarity or dissimilarity of the cases to be clustered. This is measured by constructing a distance matrix reflecting all the variables in the data set for each case. It is clear that problems will occur if there are differing scales or magnitudes among the variables. In general, variables with larger values and greater variation will have more impact on the final similarity measure. It is necessary to therefore make each variable equally represented in the distance measure by standardising the data. The preferred method of standardisation for the OA classification is range standardisation, outlined in § 3.3.3. It was felt that using the z-score standardisation that was used in the LA classification in Chapter 4 was not suitable to be used at the OA scale because it does not cope as well with extreme outliers which are more prevalent in the OA data than the LA data. Z-scores do not set an absolute limit as to what the maximum value of each variable can reach therefore, not limiting the effect of extreme values. This also means that different variables can have different maximum values. By using range standardisation an absolute limit is put on the value of each variable, therefore reducing the effect an extreme value for one variable can have on the clustering process.

5.4.2 The Hierarchy that is to be Created

Creating a classification is not as simple as just running a set of data through a clustering algorithm. There are many considerations to be taken into account such as the number of clusters to be produced, the number of layers in the classification and the minimum membership size of each cluster. A careful balance must also be struck between creating a classification that reflects the real world and one that is both usable and user friendly. These two requirements are not always compatible. All these issues need to be considered during the design and implementation of the clustering methodology.

The classification was built as a three tier hierarchy to fit in with the already published ward and local authority district level classifications. This also gives the classification scope to tackle an increased number of problems as different numbers of clusters are useful for different purposes, as will be explained later. When choosing the number of clusters to have in the classification there were three main issues.

1. Analysis of the average distance from cluster centre for each cluster number option. The ideal solution would be the number of clusters which gives smallest average distance from the cluster centre across all clusters.

2. Analysis of cluster size homogeneity for each cluster number option. It would be useful, where possible, to have clusters of as similar size as possible in terms of the number of members within each. This makes the clusters more comparable with each other.

3. The number of clusters produced should be as close to the perceived ideal as possible. This means that the number of clusters needs to be of a size that is useful for further analysis.

These first two issues can both be quantitatively measured and it is fairly simple to measure if one solution is better than another or not. However, the third issue is not so clear cut and cannot be said to have a right or wrong answer. Neither can the suitability of a solution be easily assessed quantitatively as to which is most suitable solution. There are different views on what is the best number of clusters to produce. As a guide, the number of clusters in the five most commonly used small scale area classifications in the UK are listed in Table 5.9.

Table 5.9: The number of clusters in the most commonly used classification systems

Classification System	Clusters in Level 1	Clusters in Level 2	Clusters in Level 3
Mosaic	11	N/A	61
Cameo	10	N/A	58
ACORN	5	18	57
PRiZM	N/A	16	60
Super Profiles	10	40	160

Table 5.9 shows that there is considerable difference in existing systems not only between number of clusters at each level, but also how many levels are present in the classification system. There seems to be little or no agreement as to how many clusters there are within the UK. It may have been expected that over time a number of clusters may have become accepted as being the most representative, but this does not seem to be the case. It would seem that the only way to select the number of clusters that are to be used in a classification is to select the number of clusters that work best for that individual system.

There is another way of considering what the best number of clusters to select is. That is to consider if a certain number of clusters will be more useful to a user than another number of clusters. Communication has taken place with potential users and members of the area classification advisory board. Martin Callingham (Birkbeck College) supplied an opinion about which would be the most suitable number of clusters for users. He has many years of experience in using classification systems in both commercial and academic contexts, his views as to what he has found most useful could provide excellent guidance in this matter.

*"**At the highest level of aggregation, the cluster groups should be about 6** in number to enable good visualisation and these clusters should also be given descriptive names.*

***At the next level of aggregation, the number of groups should be about 20**. This would be good for conceptual customer profiling (that is, when one wants to gain some conceptual understanding of one's customer base) and would also allow market propensity measures to be established with comparatively small surveys (for, example, two waves of an omnibus). This level could also be used for setting up sampling points for some market research surveys and would ideally also have descriptive names.*

***At the next level of aggregation, the number of groups should be about 50**. This can be used for market propensity measures from the larger commercial surveys such as TGI and the readership surveys. This level would probably also be good for use with the current government surveys. These clusters do not need names."* (Callingham 2003) emphasis added.

The above comments give good guidance as to the suitability of use of different numbers of clusters in the solution. Each level has a different purpose. The three tiers aren't created just for the sake of creating an extra dataset; rather, the number of clusters at each level dictates what the classification can be used for. Although there is no recognised ideal number of clusters that represent UK small areas, certain numbers of clusters are more useful than others. The classification needs to be fit for purpose so a great deal of attention needs to be paid to the number of clusters created during the classification process.

5.4.3 The Original Methodology

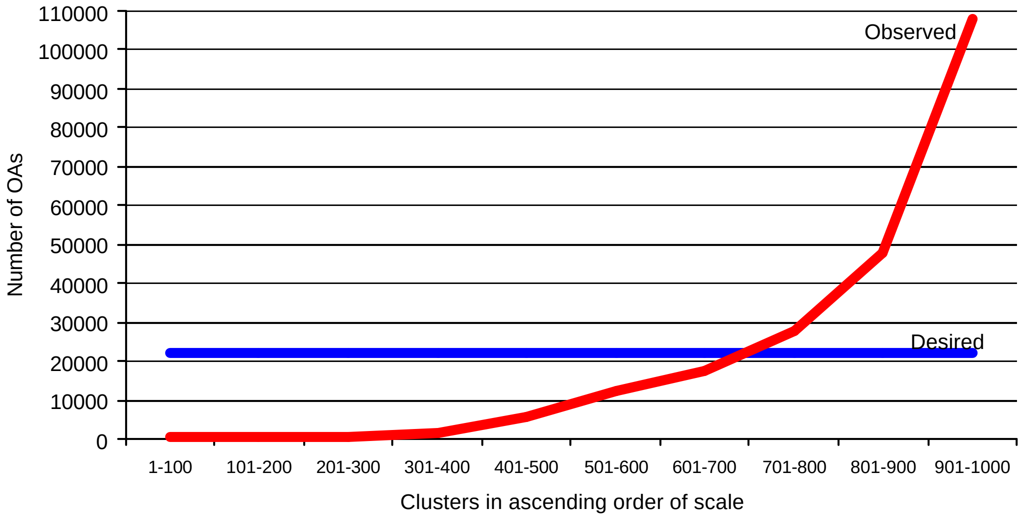
The objective is to create a three tier hierarchy to complement that created by the ONS for the ward and local authority level classifications. It was therefore planned that Ward's hierarchical clustering algorithm would be used to create the hierarchy within the classification. However, Ward's algorithm can only run on relatively small datasets of approximately 1,000 objects or fewer, not the 223,060 that are contained in the OA dataset. Therefore something needed to be done to enable Ward's algorithm to be run on the dataset.

The initial intention for the clustering method was going to be as used in the ward level classification. The procedure used was to first cluster the data using the k-means clustering procedure setting the number of clusters to be produced as 1,000. Ward's hierarchical clustering procedure was then run to be run on the cluster centres produced by the k-means procedure, and therefore adding the hierarchy to the classification.

It soon became apparent that at the OA scale this method did not work as well as had been experienced when working at the ward scale. When Ward's hierarchical clustering procedure was run on the 1,000 cluster centres produced by the k-means procedure, clusters were being produced that were several factors different in scale. Clusters that were produced ranged in size from 125,000 OAs to 3. This was caused by outliers within the dataset that were still having a significant effect despite standardisation. Even at the top level where the target size was between five and ten groups this problem was experienced.

This problem is caused by the two clustering algorithms working together. The first and biggest problem is created when 1,000 clusters are created using the k-means algorithm, within the data there are areas that have unusually extreme values, these outliers get clustered into groups of small or single membership. Figure 5.2 shows how this affects the size of membership of the clusters. The clusters have been split into deciles in ascending order with 1-100 representing the 100 clusters with the smallest membership and 901-1,000 representing the 100 clusters with the largest membership. The blue line (desired) on the graph represents the distribution if all clusters were the same size (223 members). The red line (observed) represents what we have in reality with the 30% of the clusters with the highest membership containing 85% of the OAs and the other 70% containing only 15%.

Figure 5.2: The size of observed and expected cluster sizes
(when creating 1000 clusters using the k-means procedure)



The problem is then compounded when Ward's algorithm is run on the k-means centres. Table 5.10 shows how the first attempt of clustering using the original methodology; in this seven cluster solution 98.6% of OAs are in just two of the seven groups, obviously an unsatisfactory outcome.

Table 5.10: Number of OAs in each cluster based on the original methodology: Attempt 1

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Range
OAs	125,364	94,602	1,067	1,536	213	275	3	125,361
OA %	56.2	42.4	0.5	0.7	0.1	0.1	0.0	56.2

How has this severely skewed distribution of membership come about? It becomes a little clearer by looking at the original 1,000 k-means clusters from which the smaller number of clusters is formed. Of the original 1,000 k-means clusters 124 had only 1 member; 263 had single figure membership, only 300 had above average membership, with the highest number of OAs in a cluster being 2,212. Of the original 1,000 clusters, the top 250 (25%) contained 174,694 (78%) of the OAs, the bottom (25%) contained 591 (0.3%) of the OAs. Why is this a problem? Each cluster is weighted equally and treated as one object to cluster whether it contains 2,000 OAs or only 1 OA. The reason the problem gets even worse when the data are re-clustered using Ward's algorithm is that the k-means clusters that contain only 1 OA are outliers on the edge of the dataset and the clusters with large membership are those from the centre of the data set. When the data gets re-clustered the clusters with large membership are likely to be clustered together and the outliers with small membership are likely to be clustered together producing the extreme results observed in Table 5.10.

Several different methods of data transformation were tried to make the methodology work for the OA classification. Transformation in this context means making alterations to the data before standardisation to reduce the effect of outliers in the clustering process. The different methods of transformation that were tried are listed below:

- Capping the data at the top and bottom 1%.
- Capping the data at the top and bottom 3%.
- Capping the data at the top and bottom 5%.
- Capping the data at the top and bottom 10%.
- Capping of extreme values at differing levels for each variable.
- Converting the data into ranks (1 to 223,060) for each variable.
- Converting to logarithm values.

All the transformation methods reduced the extreme range in cluster membership that was experienced when the clustering algorithm was first run. A transformation method needs to be judged in two different ways. Firstly, how much does it improve the distribution of the data? Secondly, how much has the transformation affected the integrity of the original dataset?

The method of transformation that improved the distribution of the dataset the most was converting the data to ranks. Table 5.11 shows the impact that converting the data to ranks made on the final result. By converting the data to ranks based on their value e.g. the OA with the highest value would become rank 1, and the OA with the lowest value would be rank 223,060 for each variable. The data would be in the same order but the distance between the OAs would alter, reducing distances at the extremes and increasing distance in the centre of the dataset therefore reducing the effect of the outliers.

Table 5.11: Number of OAs in each cluster based on the original methodology (ranks)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Range
OAs	21,190	43,500	30,567	38,427	69,619	12,809	6,948	62,671
OA %	9.5	19.5	13.7	17.2	31.2	5.7	3.1	28.1

Table 5.11 shows that the difference in size between the clusters produced has dramatically reduced when the converting to ranks is implemented. This difference is also visible in the in the original 1000 k-means, with only 5 of the clusters having single OA membership (compared to 124 previously). Of the original 1000 clusters, the top 250 (25%) contained 89,864 (40%) (previously 174,694, 78%) of the OAs the bottom 250 (25%) contained 26,210 (12%) (previously 591, 0.3%). The conversion into ranks has reduced the differences in the data values to a more acceptable level and looks as if it could be a usable methodology. However, there are concerns about doing this: the original integrity of the data maybe compromised by subjecting it to such extreme transformations. The data have become more usable to create a classification because of the transformation, but the transformation has also removed some of the detail from the dataset. Therefore the clusters produced would not be completely representative of the original data. The method which was felt upheld the integrity of the original data the most was transforming the data onto a logarithm scale, but as shown in Table 5.12, the log transformation does not reduce the difference in size between the clusters as much as converting the data into ranks.

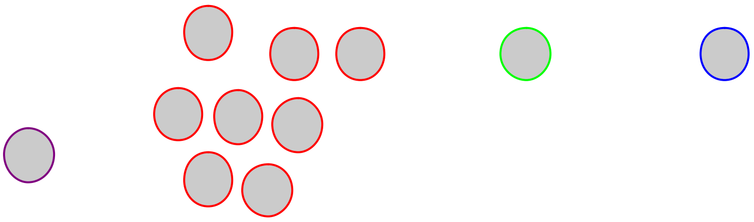
Table 5.12: Number of OAs in each cluster based on the original methodology (logs)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Range
OAs	45,041	2,694	90,837	75,785	1,473	1,938	5,292	89,364
OA %	20.2	1.2	40.7	34.0	0.7	0.9	2.4	40.0

Therefore if the one of these transformation methods is going to be used on the data a decision has to be made. Should we use a method that reduces the difference between the sizes of the cluster memberships or is it more important to keep the integrity of the original data? However, there are further concerns about Ward's method which may cause its use at this very fine spatial scale to be reconsidered.

The intricacies of Ward's method also seem to have been a contributing factor in the differences in cluster sizes experienced using this methodology. Ward's method works by grouping the nearest two OAs together and then repeating the process again at the next run but it treats the two OAs clustered on its first run as an unsplitable whole. This tends to increase the likelihood that unevenly sized groups are produced, especially in a very large data set. An OA that is an outlier on several variables will be clustered last and left on a group on its own, even though there maybe OAs clustered together that are further apart. Figure 5.3 shows how this can happen.

Figure 5.3: The intricacies of Ward's Hierarchical Clustering Procedure



The red dots in Figure 5.3 are clearly a cluster so they are grouped together in the first seven runs in of the Ward's clustering algorithm. What happens next is what can cause a problem. The purple and then the green dots are grouped with the reds in the eighth and ninth runs. Even though the purple and the green dots are twice as far apart as the green and the blue, green and purple end up in the same cluster and blue is left on its own in a ten/one split. If the same data are clustered using the k-means algorithm, green and blue would form a group, as would the purple and the reds.

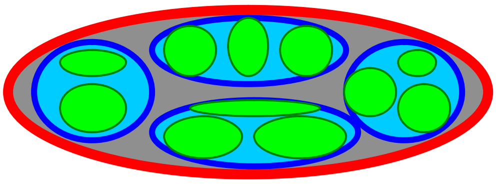
If the problem is scaled up to from 11 dots in 2 dimensions to 223,060 OAs in 41 dimensions and the number of clusters increase, it becomes apparent that Ward's method cannot cope with extreme data points. The nature of the OA data means that there are many extreme values in many dimensions. Using Ward's clustering algorithm on the OA data produces a few large clusters (e.g. 95,000 OAs) and then very small clusters (e.g. 3 OAs). This in an inherent problem of using this technique on this large amount of data. It would seem that the larger the dataset the more likely Ward's method is going to produce uneven cluster sizes.

These experiments with the OA database have shown that when a hierarchical clustering procedure is used, it will inherently produce clusters of uneven size. There are therefore serious doubts about the reliability and quality of result. The use of this methodology was therefore rejected on the basis that it could not be made to work without transforming the data to a much greater level than we were comfortable with. It was therefore decided to investigate the possibility of using a new methodology solely based on the k-means algorithm. This brings up the problem of how to create a hierarchy using a non-hierarchical approach.

5.4.4 The Final Methodology: Creating a Hierarchical System Using K-means

The solution to the problems found with the original methodology was be to adapt the k-means clustering procedure (a non-hierarchical procedure) to produce a hierarchical classification. This can be done by artificially adding the hierarchy during the clustering procedure. There are two possible ways in which this could be done. The idea is basically very simple and is represented graphically in Figure 5.4.

Figure 5.4: The creation of a hierarchical system using the k-means algorithm



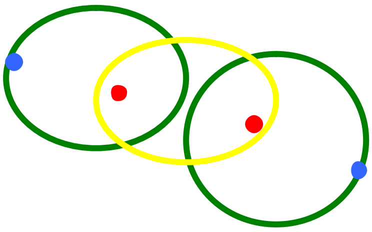
The first way is a top down approach and works as follows: the k-means algorithm is run on the dataset and *n* clusters are produced. The original dataset is then split into *n* separate datasets (representing the highest level of the hierarchy) of which one is represented by the red area in Figure 5.4. Each of the new datasets then has the k-means algorithm run on them separately to create the second level of the hierarchy (as represented by the blue areas in Figure 5.4). The second level of the hierarchy is then separated into *m* separate datasets and each one has the k-means algorithm run on them to create the lowest level of the hierarchy (as represented by the green areas in Figure 5.4).

The second way in which this could be done is a bottom up approach and works the opposite way round. The lowest level of the classification is created first (as represented by the green areas in Figure 5.4); about 50 clusters are generated using the k-means algorithm. The centres of the 50 clusters produced are then re-clustered to produce the middle level of the hierarchy (as represented by the blue areas in Figure 5.4). Then in turn the same would be done on these to create the highest level (as represented by the red area in Figure 5.4).

The top down procedure, tier by tier, was chosen as it was believed that this method is fundamentally better than the bottom up approach. With this method the objects to be classified were always a set of OAs rather than a set of cluster centres. Bottom up would have meant using sets of cluster centres throughout.

There are inherent problems in clustering using the cluster centres as found with the original methodology which applied Ward’s algorithm to cluster centres produced by the k-means algorithm and produced clusters of very uneven size. The cluster centres are not necessarily representative of the whole cluster. Not only that, but the cluster centre used is not adequately representative of all of its members. The two most dissimilar OAs can quickly be clustered together using the bottom up approach; they can be on opposite sides of the two most similar cluster centres, but totally unlike each other, as shown in the Figure 5.5. The two green circles represent two clusters formed using the bottom up approach the red dots represent their cluster centres, and the blue dots represent an outlier within each cluster. The yellow circle shows how the second level of clustering in the bottom up approach clusters the two groups together based only on their centres creating a cluster based on the values of the two centres. However, the cluster actually includes everything in both green circles including both blue dots which bear little resemblance to each other.

Figure 5.5: An illustration of the inherent problem of clustering cluster centres



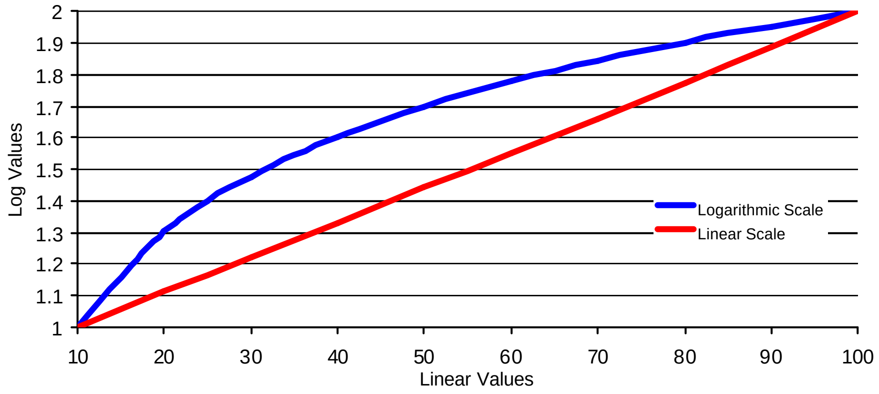
There is also the issue of which level of the hierarchy is seen to be the most important. The first level was seen as the most important level (and likely to be the most used). Therefore it was decided that the lower two levels should be made up from the top level not vice versa. There is a trade-off here: to create a hierarchy it is not possible to have the perfect solution at all levels. This is an inherent problem with any form of hierarchy. The first tier determines to a certain extent what is in the later tiers.

5.4.5 Elucidating Log Transformation

Before standardisation the data were transformed to a log scale. This was done because of the effect of a large number of outliers at the high end of the value scale. Population density was a particular problem here. By transforming the data to log scales the problem of very high value outliers was greatly reduced as the differences between values at the extremities of the data set are reduced by more than those more in the centre of the dataset. Using logs is one of several ways in which the effect of outliers can be reduced (Harris *et al*. 2005). Other methods to reduce the effect of outliers on the classification include capping the data to a specified value or percentage of cases, down weighting of variables with problematic values. Many different methods were tried to reduce the effect of outliers. Transforming the data to a log scale was the preferred method as it kept the data in the same order as opposed to other methods such as capping that grouped the data at the top and bottom of the scale.

A log (logarithm) is the exponent of the power to which a base number must be raised to equal a given number. The logarithm to the base 10 of 100 is 2 because 100 = 10^2. A log is a constant ratio scale where equal distances on the scale are represent equal ratios of increase. The sum of the logarithms of any two or more numbers is the log of their product. Therefore the effect that the log transformation will have on the data set is to reduce the effect of large gaps between variable values, which were typically found at the higher end of the range of values. The log transformation of the data squashes the ends of the data series and expands the middle. This can be seen graphically by examining the differences between the two lines in Figure 5.6.

Figure 5.6: the effect of logarithmic transformation on a dataset



Linear graphs are scaled so that equal vertical distances represent the same absolute (e.g. a drop from 100 to 99 is represented in the same way as a drop from 10 to 9. A logarithmic scale reveals percentage change so a drop from 100 to 99 is represented as being ten times less severe as a drop from 10 to 9, which therefore is represented in the same way as a drop from 100 to 90. See Figure 5.6.

Before the data were converted to a log scale, all the values had 1 added to them. This was because of zeros (of which there are many in the data). The logarithm of zero returns no result. Any value between 0 and 1 produces a negative value, which would have confused the dataset. By adding 1 to every data point this problem was resolved. The new value of the dataset can therefore be summarised by the statement below.

Log(X+1) = new value to be range standardised (5.2)

Logging the data not only reduces the effect of individual outliers but also greatly reduces the likelihood of a highly skewed distribution within a variable. This is imperative because highly skewed variables create uneven cluster sizes. Clustering algorithms work best on normally distributed data. If variables are skewed this would affect the clustering procedure as the skewed variables could have an undesirable effect on the calculations within the algorithm. Table 5.13

outlines how logging the data reduces the skew of a variable. The table shows the difference between the mean value for each variable after standardisation and 0.5, for two sets of variables, one logged and one not. It is clear from the table that in all but 3 cases the mean of the logged data is closer to 0.5 than that of the non-logged data, therefore suggesting that the logged data has more of a normal distribution than the non-logged data, in turn suggesting that the logged data will be less skewed and will contain fewer outliers. The average for all variables at the bottom of the table shows a significant difference between the two. It is vital when clustering such a large number of objects that very small groups do not emerge.

Table 5.13: The effect of logging data on the distribution of the data

Variable	Difference of mean value from 0.5 after standardisation		Difference	Variable	Difference of mean value from 0.5 after standardisation		Difference
	Not Logged	Logged			Not Logged	Logged	
V1	0.31	0.03	0.28	V22	0.09	0.09	0.00
V2	0.27	0.12	0.15	V23	0.28	0.22	0.06
V3	0.16	0.25	-0.09	V24	0.28	0.12	0.16
V4	0.12	0.26	-0.15	V25	0.18	0.21	-0.04
V5	0.33	0.09	0.24	V26	0.22	0.18	0.04
V6	0.47	0.36	0.11	V27	0.34	0.04	0.30
V7	0.48	0.40	0.08	V28	0.41	0.05	0.36
V8	0.42	0.12	0.30	V29	0.29	0.24	0.05
V9	0.50	0.14	0.36	V30	0.36	0.04	0.32
V10	0.32	0.08	0.24	V31	0.43	0.10	0.33
V11	0.34	0.07	0.27	V32	0.41	0.14	0.27
V12	0.35	0.05	0.30	V33	0.21	0.17	0.04
V13	0.36	0.01	0.35	V34	0.32	0.02	0.30
V14	0.24	0.17	0.08	V35	0.47	0.36	0.11
V15	0.29	0.08	0.22	V36	0.43	0.07	0.36
V16	0.29	0.03	0.27	V37	0.35	0.07	0.28
V17	0.42	0.13	0.29	V38	0.45	0.15	0.30
V18	0.25	0.05	0.20	V39	0.39	0.03	0.36
V19	0.27	0.01	0.26	V40	0.43	0.17	0.27
V20	0.28	0.05	0.24	V41	0.33	0.11	0.23
V21	0.42	0.12	0.29	Mean	0.33	0.13	0.20

5.4.6 The Creation of the Classification

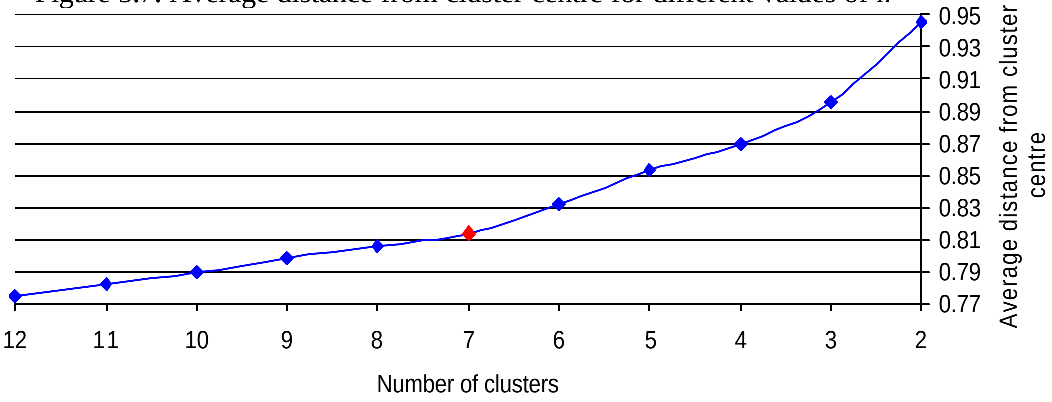
This section describes the implementation of the final methodology as described in § 5.4.4. The descriptions, the cluster size choices that were made and the reasons behind the decisions are outlined. The decisions were made based upon a plethora of information that can be outputted from the clustering process. Although it is impractical to report all of the data on which the decisions were made, an attempt has been made to give a flavour of the reasons behind the decisions that have been made.

The hierarchy was created by first clustering the whole dataset to create the super-group level. Then the dataset was split up so the data for each super-group is stored in a separate file. Each data file is then re-clustered separately. This would then be done again on the groups (middle tier) to create the sub-groups (lowest level tier).

Another problem that needed to be overcome using this method was that with k-means clustering k must be specified before running the clustering algorithm. This problem was solved by running the algorithm several times specifying different values of k each time and selecting the k which showed the most dramatic decrease in the average distance to cluster centre in comparison to k-1 (the previous cluster), in the approximate region of number of clusters that would be suitable.

It had been suggested that the most useful number of clusters In the first level would be around 6 (Callingham 2003). Taking this as a starting point clusters from 2 - 12 were examined to see how the average within cluster distance from centre changed. Figure 5.7 shows how the average distance to cluster centre increases as the number of clusters is reduced. The target was a number of clusters around 6. This was then narrowed to an expectable range of 4 - 8. Within this range it was not evident that there is any significant difference in the increase in the average distance from cluster centre, although there appears to be a peak at 5 which leaves a choice between 4, 6, 7 and 8.

Figure 5.7: Average distance from cluster centre for different values of k



Another factor that has to be taken into consideration when choosing the number of clusters to use in a classification is the relative size of the clusters (in terms of number of members). It is preferable to have the clusters as closely sized to each other as possible. For example , if creating two clusters from 10 objects, 2 clusters both containing 5 members would be the optimal solution. Oppositely , a solution of one cluster with 9 members and another with only 1 member would be the worst solution. This would not have actually created two clusters, but only removed an outlier from the original dataset. An explanation using ten data points and two clusters is fairly simple, but the same principle is true with any number of data points and clusters. The choice of a solution that produces a small cluster is even more of a problem when it is the first level of a hierarchy (as is being created here). As clusters are broken down to create the next level of the hierarchy the membership the size of the clusters get smaller. If the cluster was small to start with, this greatly increases the chances of creating a very small cluster at a lower level.

To make sure that the classification did not fall foul of this problem, a method of comparing the range of cluster sizes (with a different number of clusters) was devised. By calculating the average difference between the number of members in each cluster from the mean (the mean is the optimal solution as all clusters will have the same number of members), it is possible to ascertain which is the best solution in terms of the number of members in each cluster. The simple example in Table 5.14 shows three possible solutions from clustering 12 data points into 2, 3 or 4 clusters. The 2 cluster solution has an average difference from the mean (in this case 6) of 2. The 3 cluster solution has a smaller distance form the mean (in this case 4) at just 1.33. The 4 cluster solution is an average of 1.5 from its mean of 3 making it the second best solution. From this example, if the choice of the number of clusters was based solely on how homogenous they are in terms of number of members, the 3 cluster solution would be selected as the optimal solution.

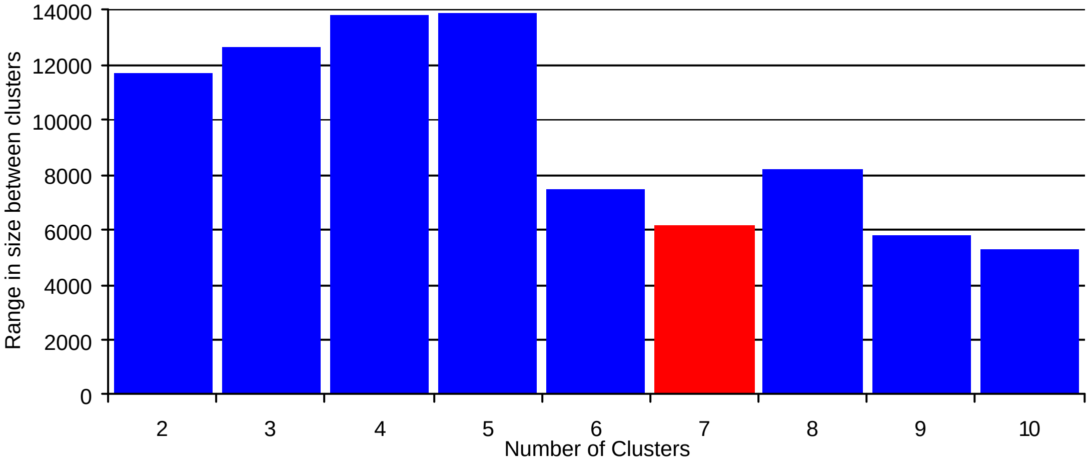
Table 5.14: Example of method of calculating which solution is most homogenous in terms number of members in each cluster

	2 Cluster Solution	3 Cluster Solution	4 Cluster Solution
Number of members in each cluster	8	4	2
	4	2	4
	N/A	6	1
	N/A	N/A	5
Average distance from the mean	2	1.33	1.5

Table 5.14 shows how the method works on a small data set, but what results were produced using this method on the possible solutions for the OA classification? Figure 5.8 shows the average distance from the mean cluster membership for solutions of cluster numbers 2 to 10. The best solution based on this criterion is ten clusters, followed by nine, then seven clusters. The worst solution is a virtual tie between four and five clusters.

A minimum cluster membership target of 50% of the average membership size for each cluster levels was set. Therefore if the first level contains 6 clusters the minimum size would be (223,060/6) ×0.5 = 18,588. If the middle layer consisted of say 25 clusters the minimum target would be (223,060/25) ×0.5 = 4,461. This target was put in place to try and get groups of fairly even sizes. However, it was viewed flexibly and if a sensible group formed that was within about 10% of the target it would be acceptable. Also smaller groups were allowed if it meant that their non-formation would have prevented the splitting of a cluster into a lower level.

Figure 5.8: The range in the size of clusters when choosing the number of clusters



Two separate forms of analysis have been run on the clusters to establish which cluster solution is most suitable to represent the first level of the hierarchy. The choice is based on the solution which performs well on both tests. The choice of solution will be made from solutions of cluster numbers of 4 to 8.

The 4 cluster solution performs well in Figure 5.7 but poorly in Figure 5.8. The 5 cluster solution performs poorly in both tests. The 6 cluster solution performs reasonably in both tests; the 7 cluster solution performs reasonably in Figure 5.7 and well in Figure 5.8; the 8 cluster solution performs reasonably in both tests. Therefore solutions 4 and 5 can be rejected for performing badly in one or both of the tests. This leaves cluster solutions 6, 7 and 8 which all performed equally well in Figure 5.7, but in Figure 5.8 the 7 cluster solution out performs 6 and 8 suggesting that it is the best solution. Therefore cluster solution 7 has been selected as the solution for the first level of the hierarchy.

Once the first level of the classification (to be known as super-groups) had been decided upon as containing seven clusters, this then needed to be broken down to create the second level of the hierarchy. This was done in a similar way to the first level, by examining the average within cluster distance. However, at this level only two, three or four clusters were considered to ensure that the number of clusters reflected as closely as possible the target number of clusters of around 20, and that the super-groups were broken down into a broadly similar number of groups. Also taken into consideration was the number of OAs in each cluster, with the intention of keeping the clusters as similar in size as possible. A second level of 21 clusters was created splitting cluster 1 into 1a, 1b and 1c, cluster 2 into 2a and 2b etc. The second level (to be known as groups) then needed to be split down again to create the third level of the hierarchy with a target size of around 50 clusters. To create the third level the clusters in the second level were spilt into two, three or four clusters, again considering the within cluster difference and the number of OAs in each cluster. The third level of the hierarchy (to be known as sub-groups) numbers 52 clusters by splitting cluster 1a into 1a1, 1a2 and 1a3, cluster 1b into 1b1 and 1b2 etc. Table 5.16 shows the structure of the classification, indicating into how many groups each cluster was split.

Table 5.15 shows that the clusters produced are of a much more even size than even the best results obtained using the original methodology with the range in size between the largest and smallest clusters halved, the range reducing from 62,671 using the most compact solution from the original methodology, to 30,613 with the use of the new methodology.

Table 5.15: Number of OAs in each cluster based on the final methodology

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Range
OAs	35,837	16,638	27,743	47,251	33,166	40,769	21,721	30,613
OA %	16.0	7.5	12.4	21.2	14.8	18.3	9.7	13.7

5.5 Outputs

It is essential that any classification produces good and easy to understand outputs. The quality of the classification produced is irrelevant if the information that accompanies it does not provide a quick and easy way of understanding it. This section will describe and discuss the outputs from the OA classification, which include the structure of the classification, naming and describing the classification, preparing the classification for use and mapping the classification.

5.5.1 Naming and Describing the Clusters

One of the world's most underrated art forms must be the naming and labelling of the clusters of area classifications. The process can be long and drawn out, as everybody will have a different opinion of what to call each group. Like much of the rest of the classification process there is no right or wrong answer. The objective is to come up with something that is thought to be the most accurate and acceptable name to describe each cluster. However, it is a very important job; as if it is done wrongly it can give a false impression of the areas within a cluster.

Names and descriptions are a very contentious issue in geodemographic classifications. They can become an increasingly sensitive subject as the scale gets smaller and the classifications appear to be more person than area based. The names could and maybe should be seen as very much a side issue to the whole classification process as no matter what each cluster is called it does not alter the variable values of the cluster. Names can also be easily pilloried by the media as they provide good headlines. Much of the criticism of geodemographics has been focused on the names of the groups. Make the name too specific and they only represent those areas very close to the centre of the same cluster. One could think of this as a form of the ecological fallacy. Alternatively make the names too broad in an attempt to represent all of the areas that fall within a cluster and they become too vague and start to sound alike; a healthy balance needs to be found.

The commercial classifications available in the UK were slower than their American counter parts in giving their clusters catchy names. However, some systems have now embraced the use of “snazzy” eye catching names while others still have a very British way of naming their clusters. This can be seen clearly in the difference between the names in the Mosaic and Cameo systems. Mosaic’s names include such titles as: Global Connections, Fledgling Nurseries, Coronation Street, University Challenge and Pastoral Symphony (Experían 2005); while the Cameo names include the following: Affluent Singles in Quality Rented Flats, Well off School Age Families in Semi-detached Properties, Younger Couples in Smaller Terraced Housing and Young Student Areas (EuroDirect 2005). The distinction between the two in terms of their approach to naming clusters is clear. The Mosaic names (Experían) are designed to be creative, provocative (and are perhaps a little inaccurate). The Cameo names (EuroDirect) are more factual (and are duller). The names suggest little about the quality of the product. They are, however, indicative of the market each company is targeting. While the Mosaic names will be loved by a more style than substance advertising executive, the Cameo names would appeal to the more analytical minded spatial analyst. Whether this is a deliberate tactic of the two companies to target opposite ends of the market is unclear. What is clear is that the names matter and the two different approaches taken by Experían and EuroDirect in naming their clusters reflects not only on their individual products but on their businesses as a whole.

5.5.2 Cluster Names

It was decided (after discussion between the Leeds and ONS teams) that the first two levels of the hierarchy would be named and the third level would receive a subcategorised name from the second level. It was thought that the time taken to develop a set of 52 names for the third tier was not justified by the value that they would give to the classification. This therefore meant that 28 names needed to be developed to represent the first two layers of the classification, 7 for the first layer and 21 for the second.

Two general principles, were followed in the naming of the clusters: they must not offend residents and they must not contradict other official classifications or use already established names. Coming up with descriptive, inoffensive names for some areas is easier than for others. For a pleasant area it is not such an arduous task as for areas where in general few would choose to live. “Rural” and “urban” were not to be used as they could cause confusion as the government have produced an urban/rural classification at OA scale (ONS 2005c). “Prosperous” and “affluent” were rejected as giving too much of a stigma of wealth or indeed non-wealth to areas. “Elderly” was also a word that was not allowed to be used as it was said to portray old age in a negative sense.

Some comments and suggestions on names was received from people who took part in a consultation exercise about the classification (described in § 6.7), but much of this advice was in the form *“I don’t like this name but I have no suggestions for a better one”*. The names have gone though several revisions and names have moved from one group to another as it became apparent that a name already given to a group was more suitable for an as yet unnamed group. The names were reviewed, developed and approved by a group of ONS Neighbourhood Statistics and geography specialists.

The names (as displayed in Table 5.16) were created by firstly examining the variable values for each cluster to establish which variables have high and low values for each cluster to establish what kind of areas were represented by each cluster. The names given to the previous classifications (LA and Ward level) and several commercial systems were examined to see what kind of names had been used previously. This was done to give guidance and to make sure that names were not selected that had already been used in another classification. Repeating names from another classification system would have implications beyond simply being seen to steal someone else’s names. Someone who was comparing two classification systems and found that two groups had the same name would intuitively assume that the two groups were intended to represent the same set of areas/people when this is not necessarily the case. Armed with a dictionary and a thesaurus the task was then addressed with an open mind. The results are shown in Table 5.16.

Table 5.16: The Cluster Names

1: Blue Collar Communities	1a: Terraced Blue Collar	*1a1: Terraced Blue Collar (1)*
		1a2: Terraced Blue Collar (2)
		1a3: Terraced Blue Collar (3)
	1b: Younger Blue Collar	*1b1: Younger Blue Collar (1)*
		1b2: Younger Blue Collar (2)
	1c: Older Blue Collar	*1c1: Older Blue Collar (1)*
		1c2: Older Blue Collar (2)
		1c3: Older Blue Collar (3)
2: City Living	2a: Transient Communities	*2a1: Transient Communities (1)*
		2a2: Transient Communities (2)
	2b: Settled in the City	*2b1: Settled in the City (1)*
		2b2: Settled in the City (2)
3: Countryside	3a: Village Life	*3a1: Village Life (1)*
		3a2: Village Life (2)
	3b: Agricultural	*3b1: Agricultural (1)*
		3b2: Agricultural (2)
	3c: Accessible Countryside	*3c1: Accessible Countryside (1)*
		3c2: Accessible Countryside (2)
4: Prospering Suburbs	4a: Prospering Younger Families	*4a1: Prospering Younger Families (1)*
		4a2: Prospering Younger Families (2)
	4b: Prospering Older Families	*4b1: Prospering Older Families (1)*
		4b2: Prospering Older Families (2)
		4b3: Prospering Older Families (3)
		4b4: Prospering Older Families (4)
	4c: Prospering Semis	*4c1: Prospering Semis (1)*
		4c2: Prospering Semis (2)
		4c3: Prospering Semis (3)
	4d: Thriving Suburbs	*4d1: Thriving Suburbs (1)*
		4d2: Thriving Suburbs (2)
5: Constrained by Circumstances	5a: Senior Communities	*5a1: Senior Communities (1)*
		5a2: Senior Communities (2)
	5b: Older Workers	*5b1: Older Workers (1)*
		5b2: Older Workers (2)
		5b3: Older Workers (3)
		5b4: Older Workers (4)
	5c: Public Housing	*5c1: Public Housing (1)*
		5c2: Public Housing (2)
		5c3: Public Housing (3)
6: Typical Traits	6a: Settled Households	*6a1: Settled Households (1)*
		6a2: Settled Households (2)
	6b: Least Divergent	*6b1: Least Divergent (1)*
		6b2: Least Divergent (2)
		6b3: Least Divergent (3)
	6c: Young Families in Terraced Homes	*6c1: Young Families in Terraced Homes (1)*
		6c2: Young Families in Terraced Homes (2)
	6d: Aspiring Households	*6d1: Aspiring Households (1)*
		6d2: Aspiring Households (2)
7: Multicultural	7a: Asian Communities	*7a1: Asian Communities (1)*
		7a2: Asian Communities (2)
		7a3: Asian Communities (3)
	7b: Afro-Caribbean Communities	*7b1: Afro-Caribbean Communities (1)*
		7b2: Afro-Caribbean Communities (2)

5.5.3 Cluster Profiles

The idea behind cluster profiles is to create a short description, using text and visuals, which expands on the cluster names, only takes a few seconds to read, but significantly expands the users', understanding of the group. The cluster profiles include graphs, photos of typical homes or neighbourhoods and some statistical information along with an extended description of the clusters.

Like the names, the cluster profiles were not easy to produce, especially for the sub-group level where the clusters are more numerous and in some cases not easy to distinguish from each other. However, at the sub-group level there are more extreme values. Therefore for many sub-groups it is easier to get a handle on which variables are distinguishing that cluster from other sub-groups. Clusters that show extreme values for one or more variables are easier to describe than groups which have average values for all variables. This is perhaps not surprising as researchers tend to focus on exploring extremes, whether it is poverty of affluence; averageness is not generally studied. The non-interest in situations of an average nature has led to there being almost a stigma about being average, to the extent where people would rather be rated as poor for something than average. It is likely that at some point in your life you have heard somebody say at least I am not average. This preference to be poor rather than average is not such a hard concept to understand. The benefit system illustrates the notion; those who are rich don't need them, those who are poor receive them, but those who are average would perhaps benefit from them, but are not eligible to receive benefits. The descriptions also, where appropriate, contain information about the geographical distribution of the groups whether the group is found in a particular geographical milieu, in particular parts of towns and cities or only in rural areas. Specific place names are avoided, because these have resulted in geographical mislabelling in past classifications.

Cluster profiles are given for each of the seven super-groups in Figures 5.9-5.15 (other levels are not shown due to limitations of space). Each portrait has a radial plot which represents the values for each variable. The numbers on the scale represent the difference from the mean value for that variable; therefore the mean for all variables is 0. The mean is represented by the middle ring at 0, the value of each variable for that super-group can then be seen by the amount that the blue line differs from the mean for each variable.

Figure 5.9: Cluster summary of super-group 1: Blue Collar Communities

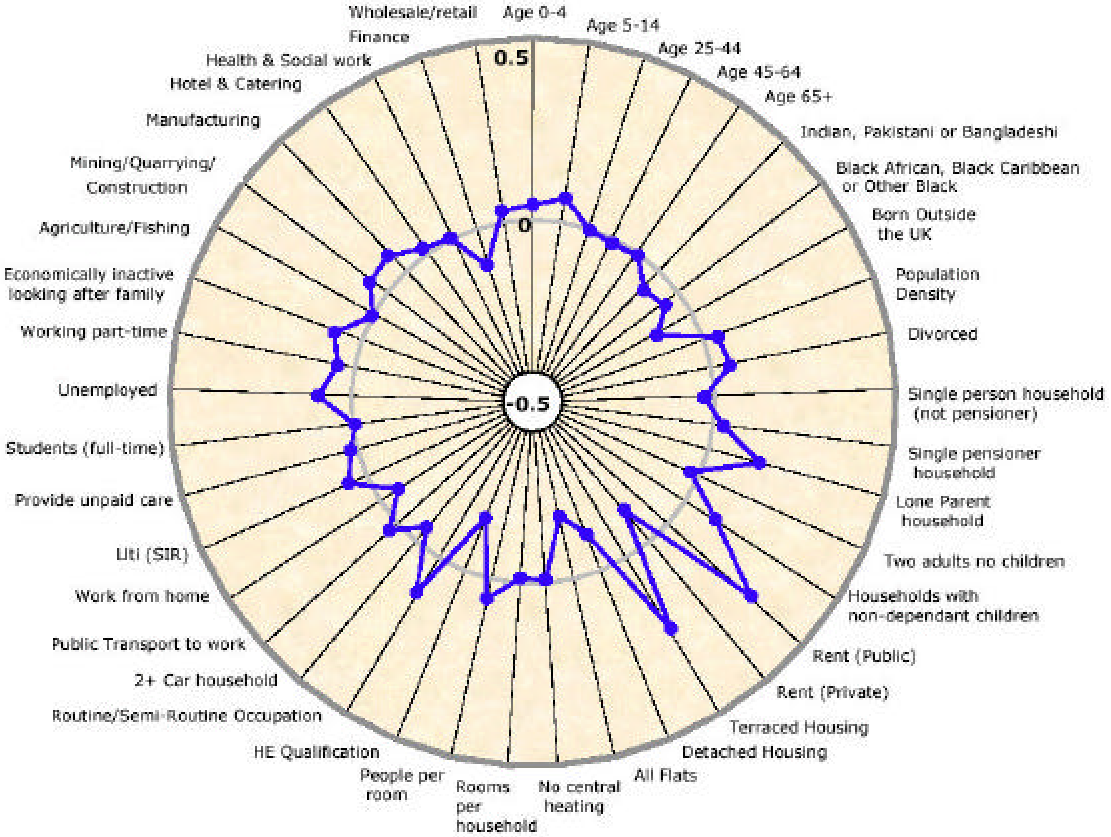


Figure 5.10: Cluster summary of super-group 2: City Living

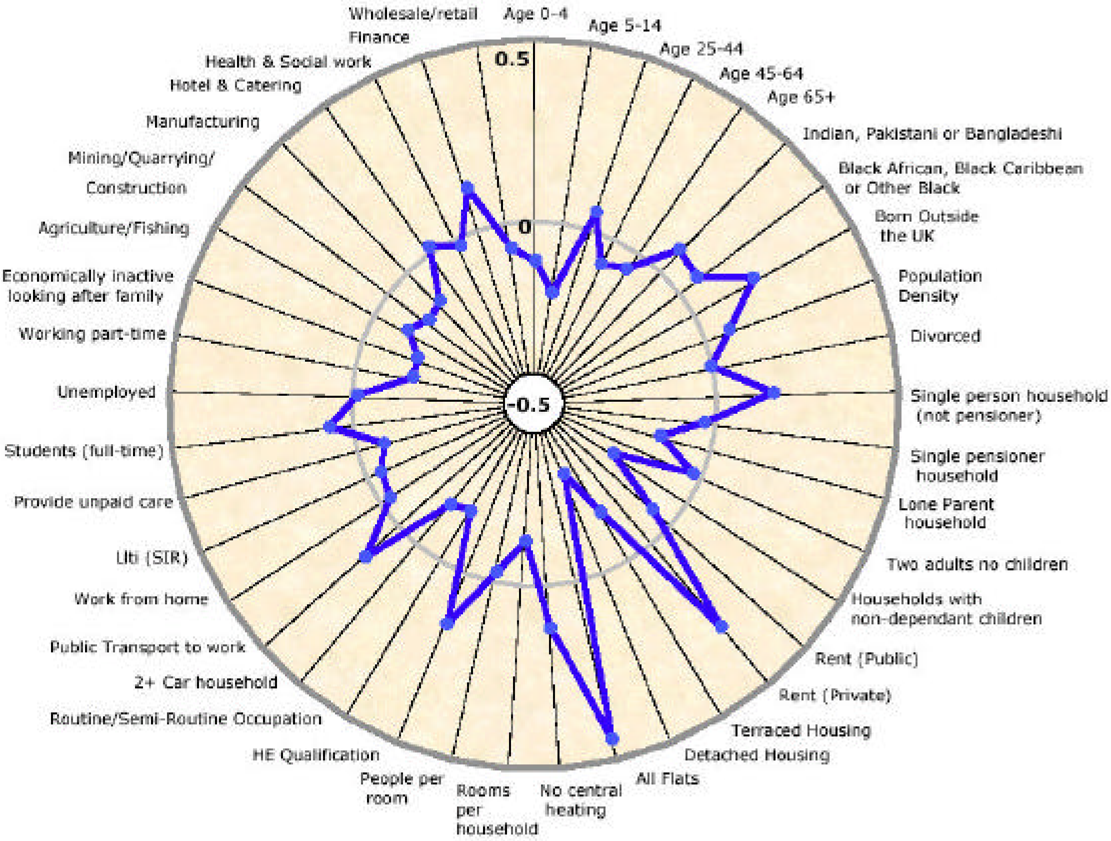


Figure 5.11: Cluster summary of super-group 3: Countryside

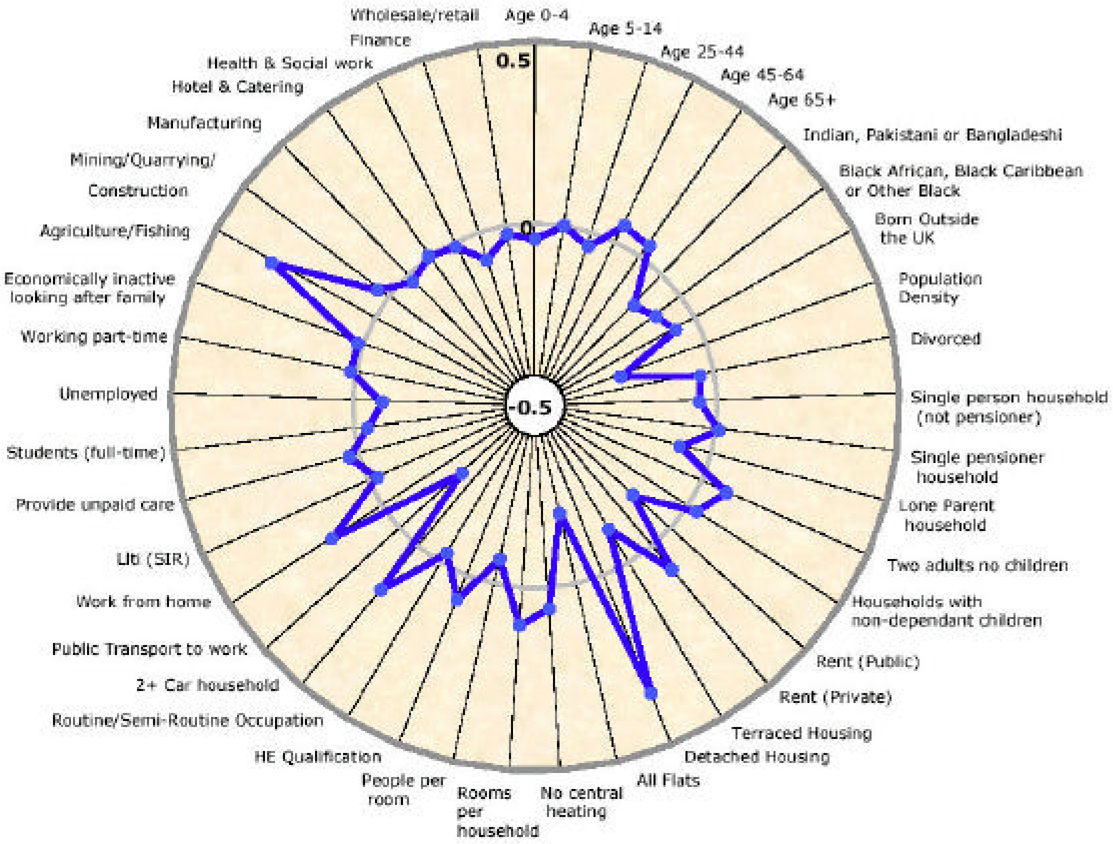


Figure 5.12: Cluster summary of super-group 4: Prospering Suburbs

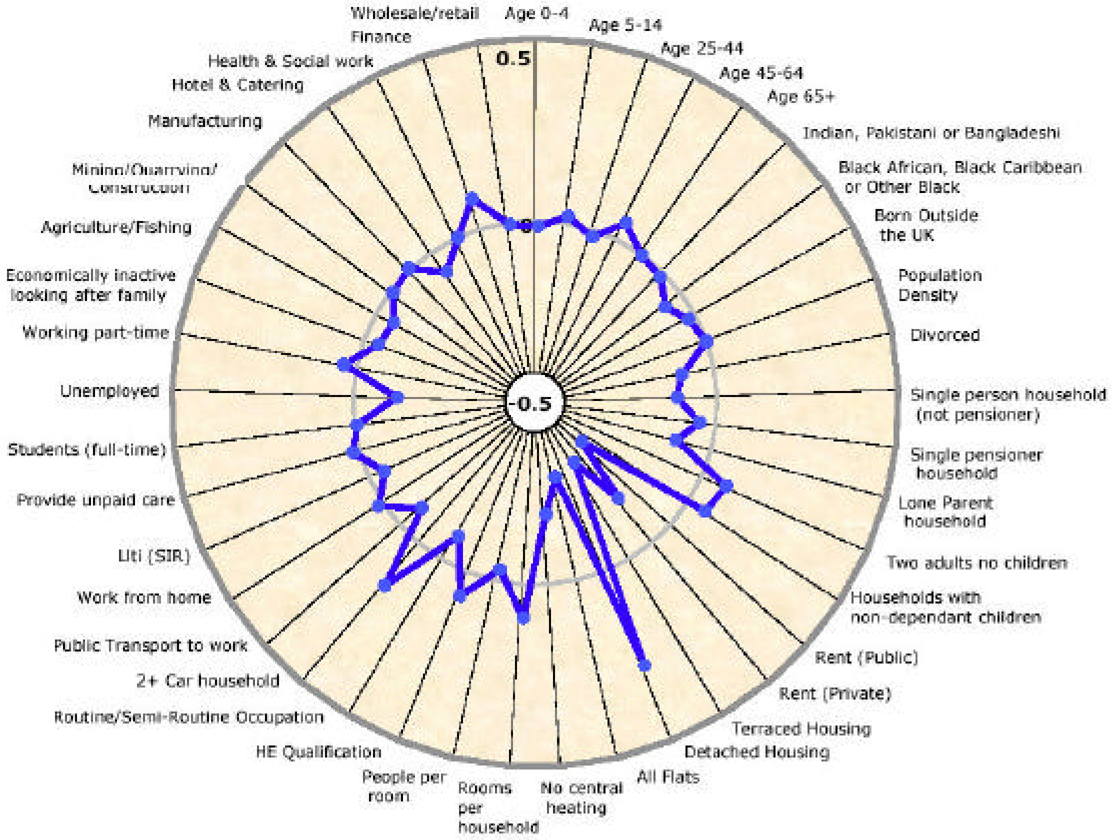


Figure 5.13: Cluster summary of super-group 5: Constrained by Circumstances

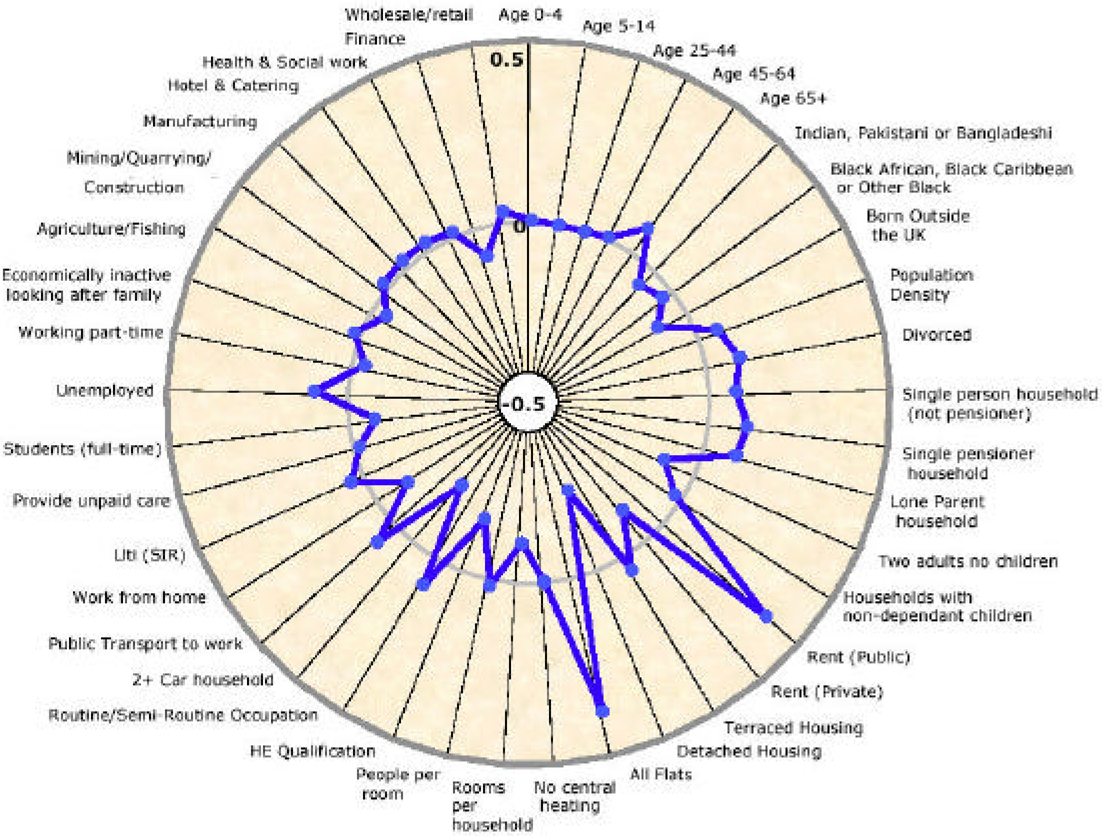


Figure 5.14: Cluster summary of super-group 6: Typical Traits

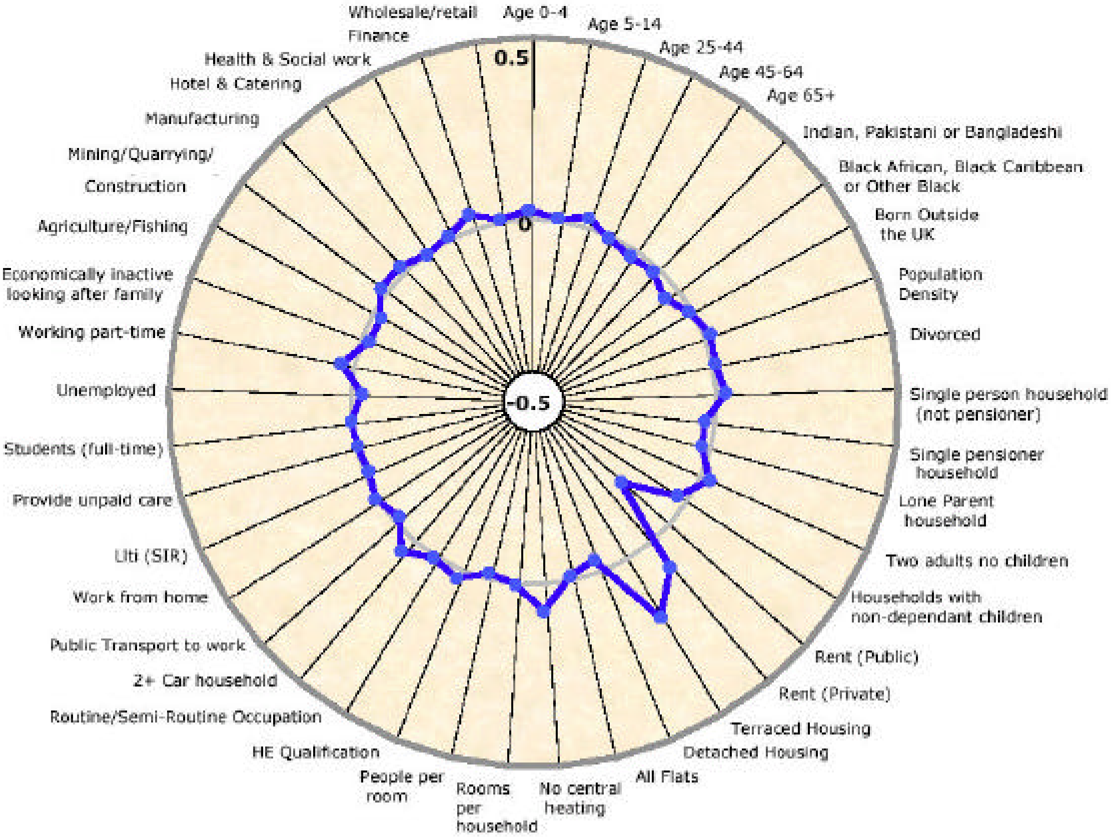
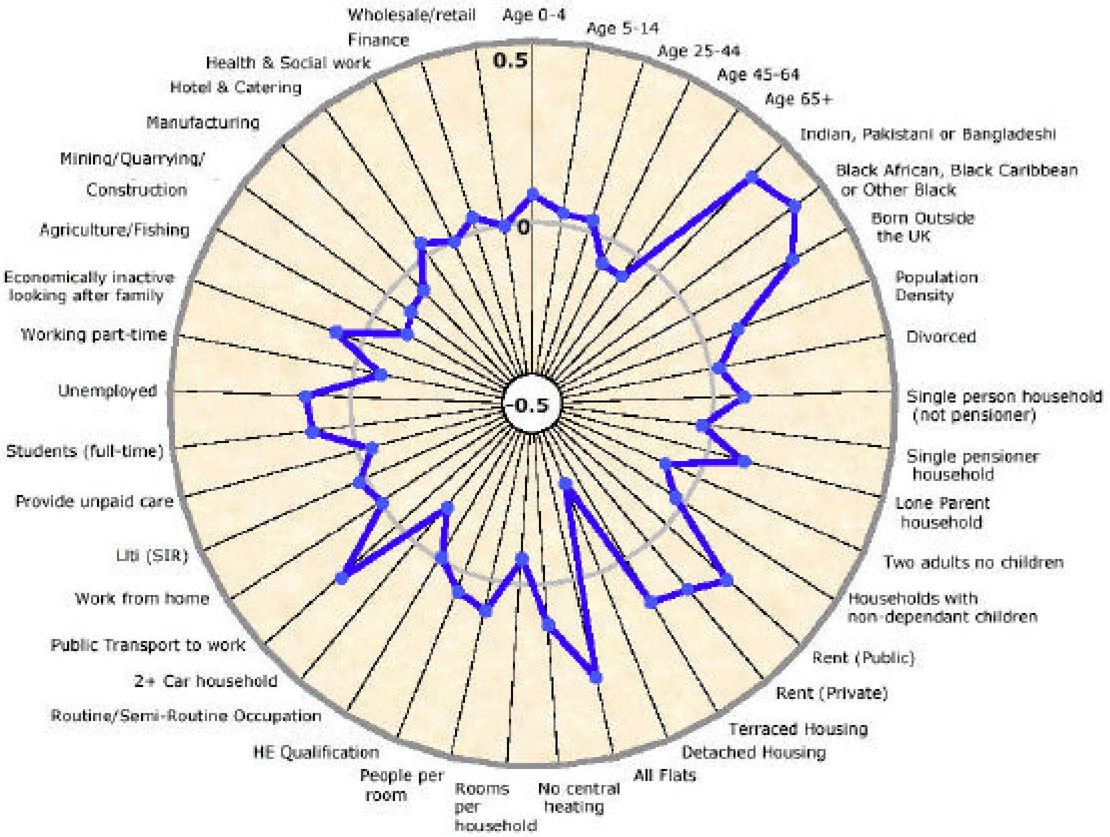


Figure 5.15: Cluster summary of super-group 7: Multicultural

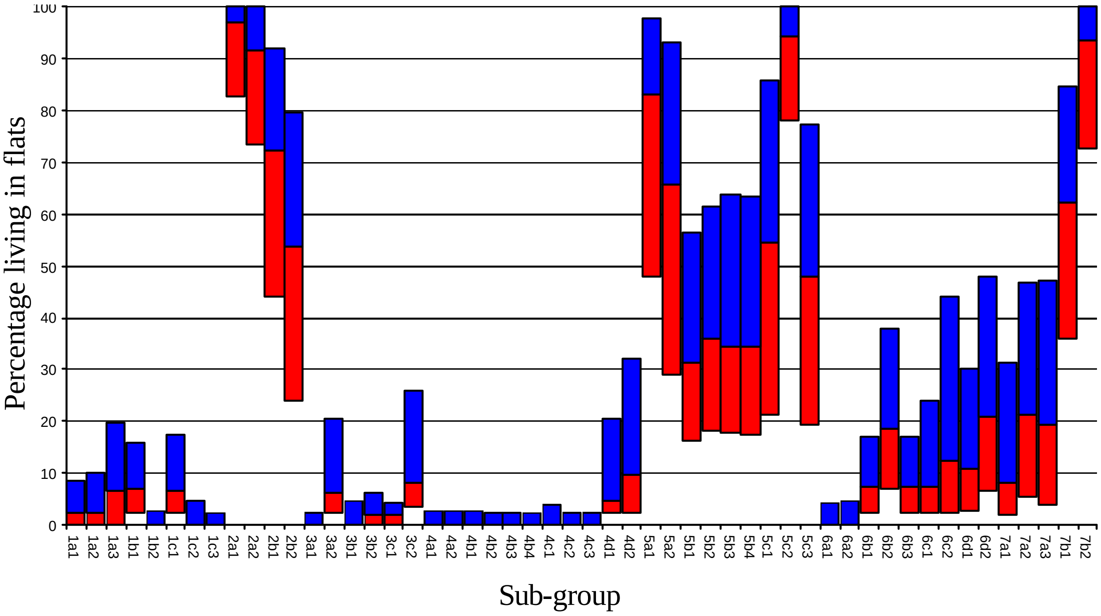


5.5.4 Other Outputs

As well as the traditional cluster profiles shown in § 5.5.3, where the strength of each variable within a cluster group can be seen, the data can be displayed in an alternative and perhaps a more revealing way. The values for any one particular variable can be given for all super-groups, groups or sub-groups. This alternative way of looking at the data allows the user to establish which group(s) have the most or extreme values for any particular variable. Figure 5.16 shows variable 20 (percentage of households which are flats) for sub-groups. The graphs don't just give the mean value, but give added context by giving an indication of the range of values represented. The top of the of the bar of the graph is the 90th percentile of the data range, the point at which the two colours meet is the mean, and the bottom of the bar is the 10th percentile.

Figure 5.16 shows some extreme values at both ends of the scale from 2a1, 5c1 and 7b2 which are dominated by people living in flats to, 1b2, 1c3, 3a1, 4a1-4b4, 4c2 and 4c3 where flats are somewhat of a novelty. The indication of the range given by the length of the bars also gives much information about each cluster. For example, compare 5c1 and 5c2. 5c2 is more homogeneous in terms of its housing type in comparison to 5c1. It is therefore possible to gauge differences between clusters not just on average variable values which attempt to represent the whole cluster, but also on the range of values contained within that cluster, giving an indication of diversity or homogeneity for each variable within each cluster.

Figure 5.16: Variable by cluster graph using the original data for variable 20 (All Flats)



5.5.5 Mapping the Classification

It is easy to forget, especially for those who are not used to dealing with geographic information, that each piece of data represents the attributes of a number of people and each output area code represents a real place on the ground containing real people, their homes and their lives. These are not insignificant numbers; they represent the way people live and where they choose to live their lives.

The final step of the classification, but perhaps the most important, is to map it and thus bring it to life. To give the location back to the output areas to see how they are spread across the country, within the towns and cities and look for patterns that emerge. If the location and distribution of the different clusters is not known the attributes of the people who live inside them becomes just an act of statistical manipulation rather than a useful piece of information. By mapping the classification the real essence of the classification can be brought out, it comes alive and really starts to mean something, displaying the rich tapestry of the social geography of the UK at the start of the twenty first century.

All the mapping in this document uses just the super-group level of the hierarchy, for simplicity. The seven clusters at the super-group level constitute a handy number to be mapped. There are enough of them to show the differences between the areas, but few enough so that there are not too many colours that they start making the map confusing or that some of the colours start to look similar to others.

The best place to view the classification is in a Geographic Information System (GIS) such as Arc Map or MapInfo. This gives the user the ability to zoom in and out and look at the data at a variety of scales plus the ability of adding many different forms of background mapping and contextual information to aid understanding.

5.5.6 Visualising the Classification in Alternative Ways

There are problems with the mapping of output areas (discussed in § 5.2.1). Mapping at such a small scale has inherent scaling problems, problems wrapped up in the design of the OAs (see Figure 5.1) and problems in adding locational information to aid the identification of places along with the information about the classification membership. This section displays a variety of different ways of mapping and visualising the information from the classification.

Enabling good visualisation of the classification does not necessarily mean mapping the classification in the most accurate way. The best example of someone who found that taking a step back from reality produced the most usable map or graphical representation was Harry Beck; Beck devised possibly the most famous map in Britain, the London Underground map. The underground map works because it depicts a complicated network by displaying only the information that the user requires, rather than producing a true depiction of the network. It is not really a map but a travel aid. It is not to scale, but does not need to be to fulfil its purpose (Garland 1994). So what is the connection between the map of the London Underground and a good visualisation of the OA Classification? The answer is that we need to look at the geography in the same way that Beck did. The only information that needs to be put on the map is that which is to be conveyed to its user. If the intricacies of the OA boundaries are what makes the map difficult to interpret then the way to make the map easier to understand is not to map the OAs and their boundaries, but simply display something which represents each area. This can be done by using the centroid of the OA (preferably the population weighted centroid) as the location for a symbol to represent each OA. This therefore removes the problem of the variability in areal size between the OAs despite there relative similarity in population size.

Figure 5.17 shows the whole UK mapped at OA scale for Super-groups using OA centroids (the centroids for England and Wales are population weighted centroids whereas for Scotland and Northern Ireland they are simple geographic centroids as population weighted centroids are not available). The advantage of mapping using centroids rather than using the geographic extent of all the OAs is that the sparsely populated areas (the largest OAs in terms of area) do not dominate the map, this also serves to make those OAs which only cover a small geographic area more visible. What is obvious from the map is that super-group 3 *Countryside* is perhaps unsurprisingly located outside the large urban centres. Some variation can be seen within urban centres at this scale for example *Multi-Cultural Blend* and *City Living* can be seen in London, while in Tyne and Wear and South Wales *Blue Collar Communities* can be more easily identified. It is vital to be able to view the classification of the UK for the whole of the UK at once. This gives a good form of comparison between all places but to get a real idea of what is going on the classification must be viewed for a much smaller area.

Figure 5.17: Mapping the OA Classification at the super-group level for the whole UK, using centroids

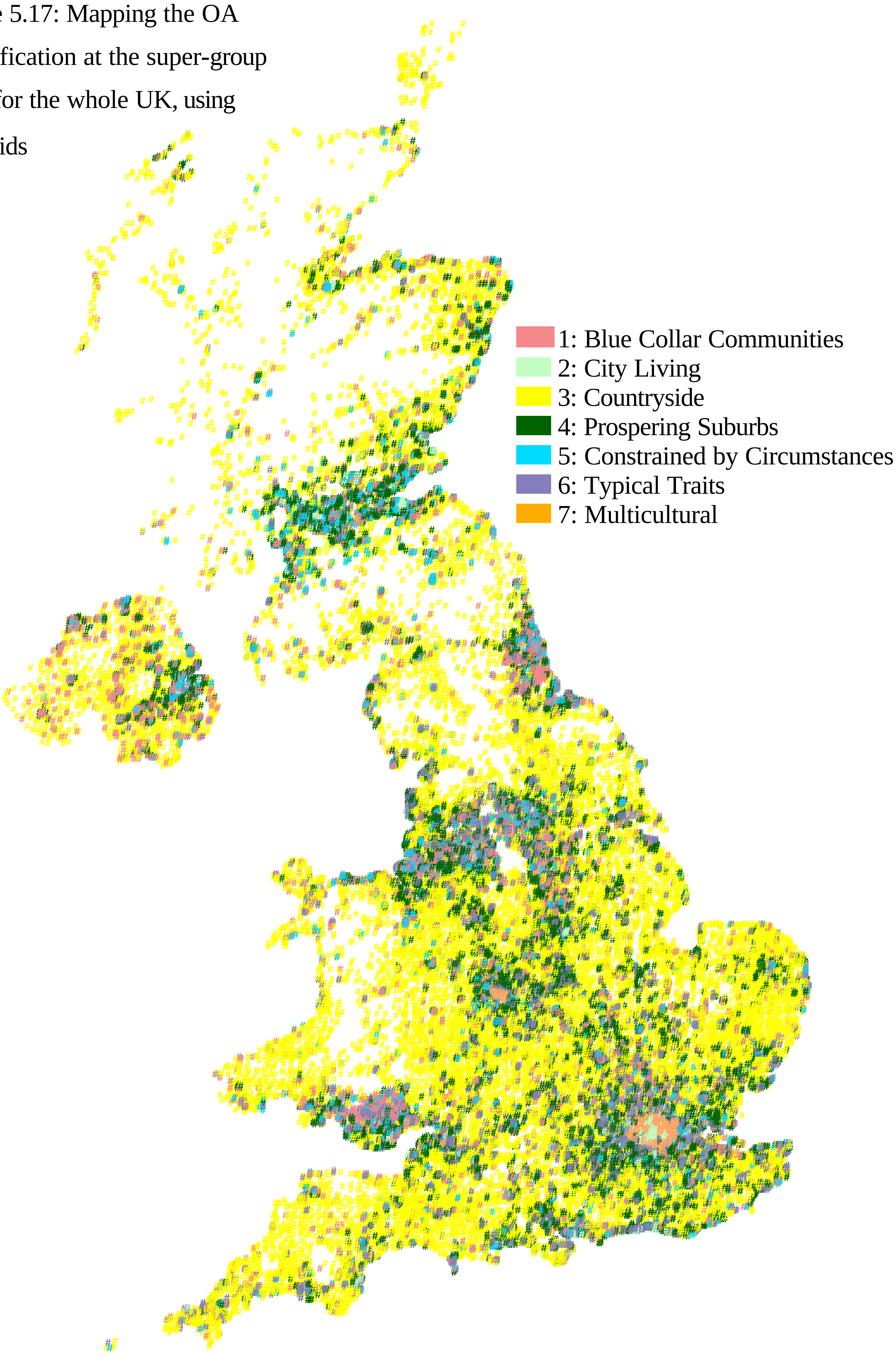
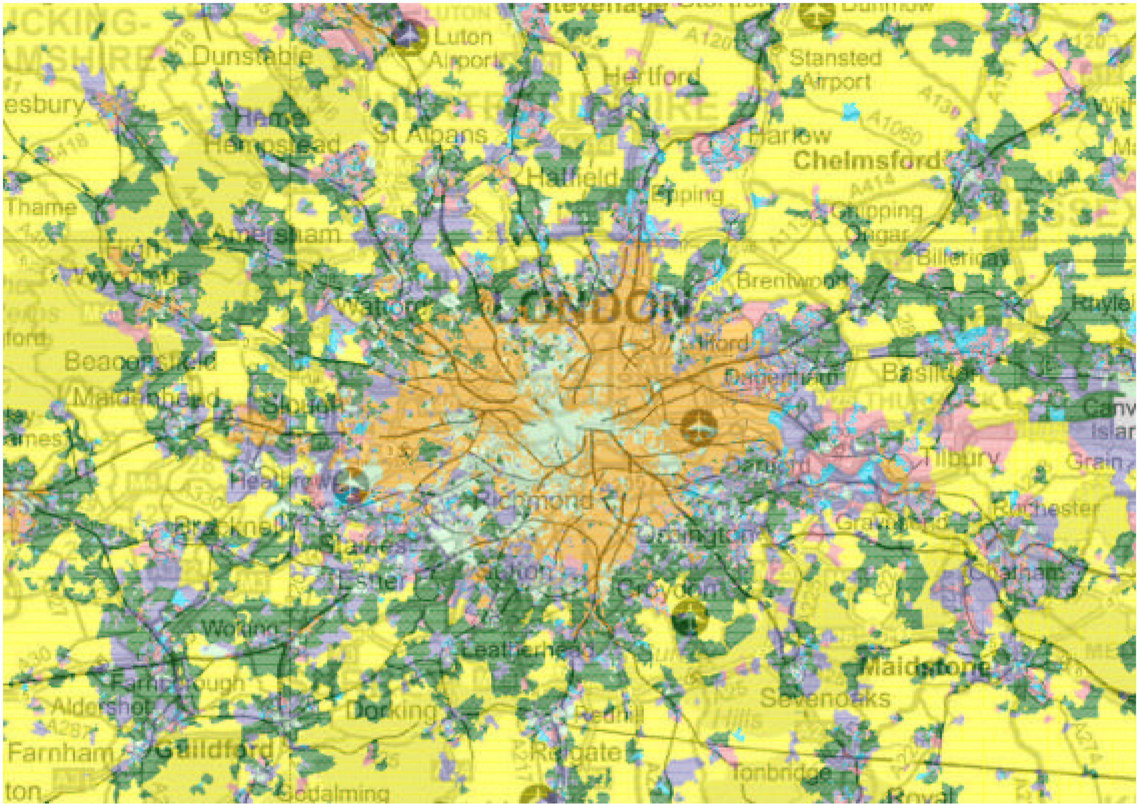


Figure 5.18 shows the classification for London and its surrounding area. Figure 5.18a shows a map using the full boundaries of the OAs whereas Figure 5.18b shows a map using just the OA centroids. Both maps give a good impression of the distribution of the different groups within London clearly showing the dominance of the *City Living* group in the very centre of the city and the pattern of *Multicultural* group surrounding it. However, it is away from the metropolitan area where the difference between the two maps becomes apparent. The *Countryside* group is dominant in one map, but not in the other, much greater diversity can be seen on the centroid map as it is not dominated by one colour, which enables smaller areas of other colour to be viewed more easily. Both these figures are overlaid on maps showing these urban centres of the UK. This can provide a great deal of detail and information when visualising the classification. Different things can be used to overlay the classification onto, satellite images or aerial photographs can also be used to add information to the classification.

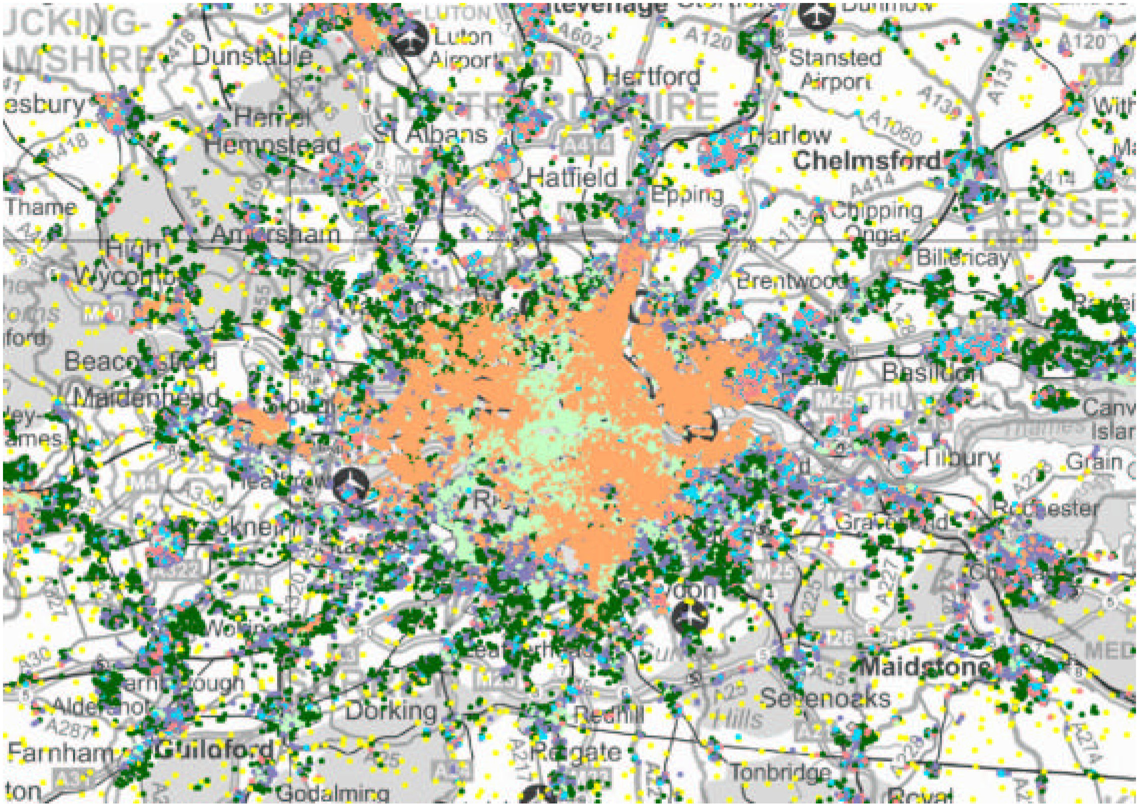
Figure 5.18a: Mapping the OA classification at super-group level using boundaries for London and surrounding area overlaid on a UK settlement map



© Collins Bartholomew

🟥1: Blue Collar Communities 🟩2: City Living 🟨3: Countryside 🟩4: Prospering Suburbs 🟦5: Constrained by Circumstances 🟪6: Typical Traits 🟧7: Multicultural

Figure 5.18b: Mapping the OA classification at super-group level using centroids for London and surrounding area overlaid on a UK settlement map



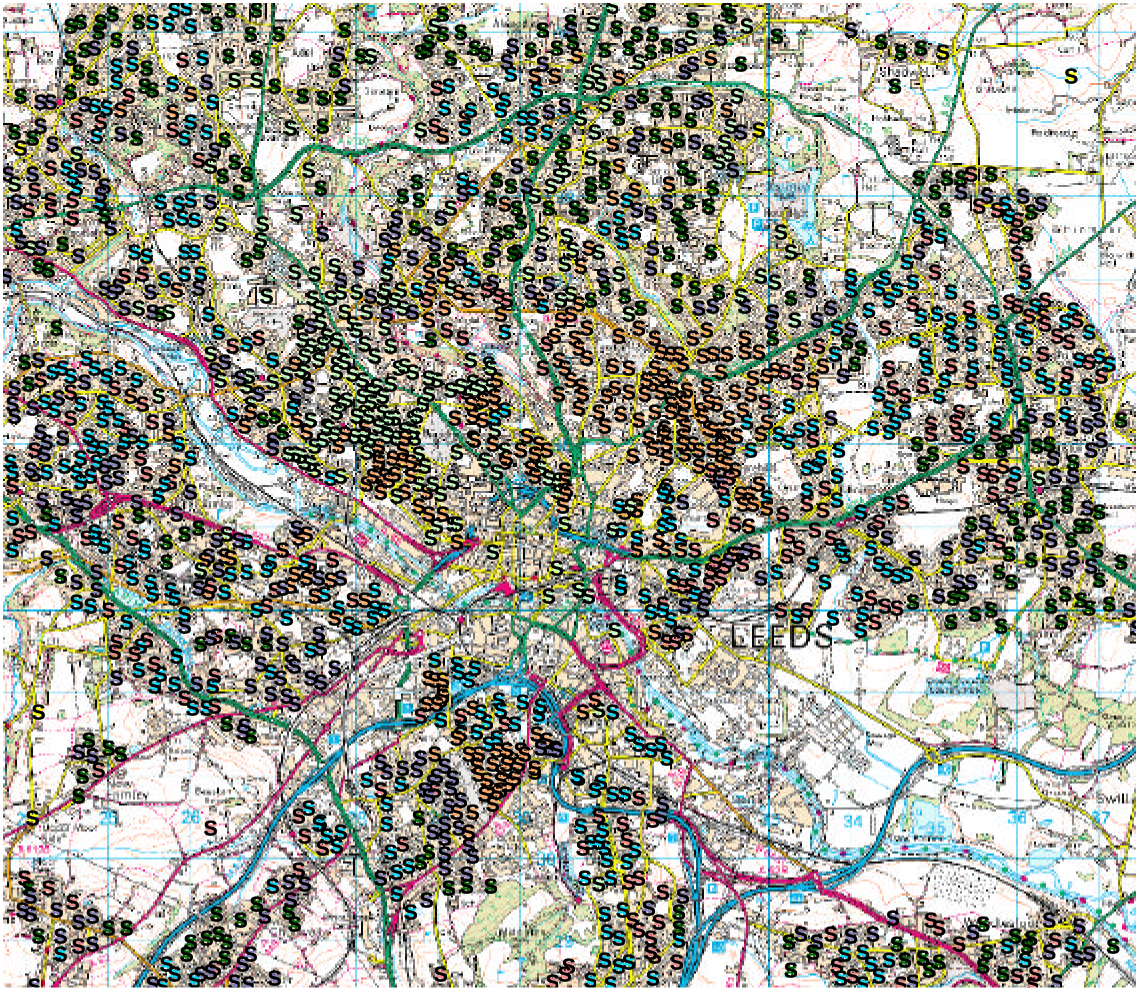
© Collins Bartholomew

🟥1: Blue Collar Communities 🟩2: City Living 🟨3: Countryside 🟩4: Prospering Suburbs 🟦5: Constrained by Circumstances 🟪6: Typical Traits 🟧7: Multicultural

Figure 5.18a accurately represents the area that is covered by each super-group type. However it is misleading in terms of the number of people who live in each super-group type. Figure 5.18b more accurately represents the population within each super-group type. Each coloured dot represents one OA (although their populations are not identical, they are broadly similar). By visualising the classification in this way it is possible to get a much better idea of the number of OAs of each type that are in the area. The *Countryside* super-group no longer dominates the map like in Figure 5.18a and this allows other information to be drawn out.

Figure 5.19 shows the population weighted centroids for the OA classification at super-group level overlaid on Ordnance Survey 1:50,000 scale mapping for the city of Leeds. This shows much more detail than any of the previous maps. The road network and the extent of the built up area can clearly be seen underneath the coloured points representing the super-groups. This helps to give more context to the classification; it gives a really good idea of how the classification maps on to the underlying geography of the streets and the buildings. Things that can be clearly seen are the homogeneity of some areas especially the *City Living* and *Multicultural* areas which can be found close to the city centre. The city centre itself can be identified from the sparsity of points due to the lack of residential properties in the very centre of the city. A north-south divide within Leeds is also noticeable. The North of Leeds has always been more prosperous than the south and this can be seen from the relative number of *Prospering Suburbs'* which are far more prevalent in the north than the south.

Figure 5.19: Mapping the OA classification at super-group level using centroids for Leeds and surrounding area overlaid on 1:50,000 Ordnance survey Mapping

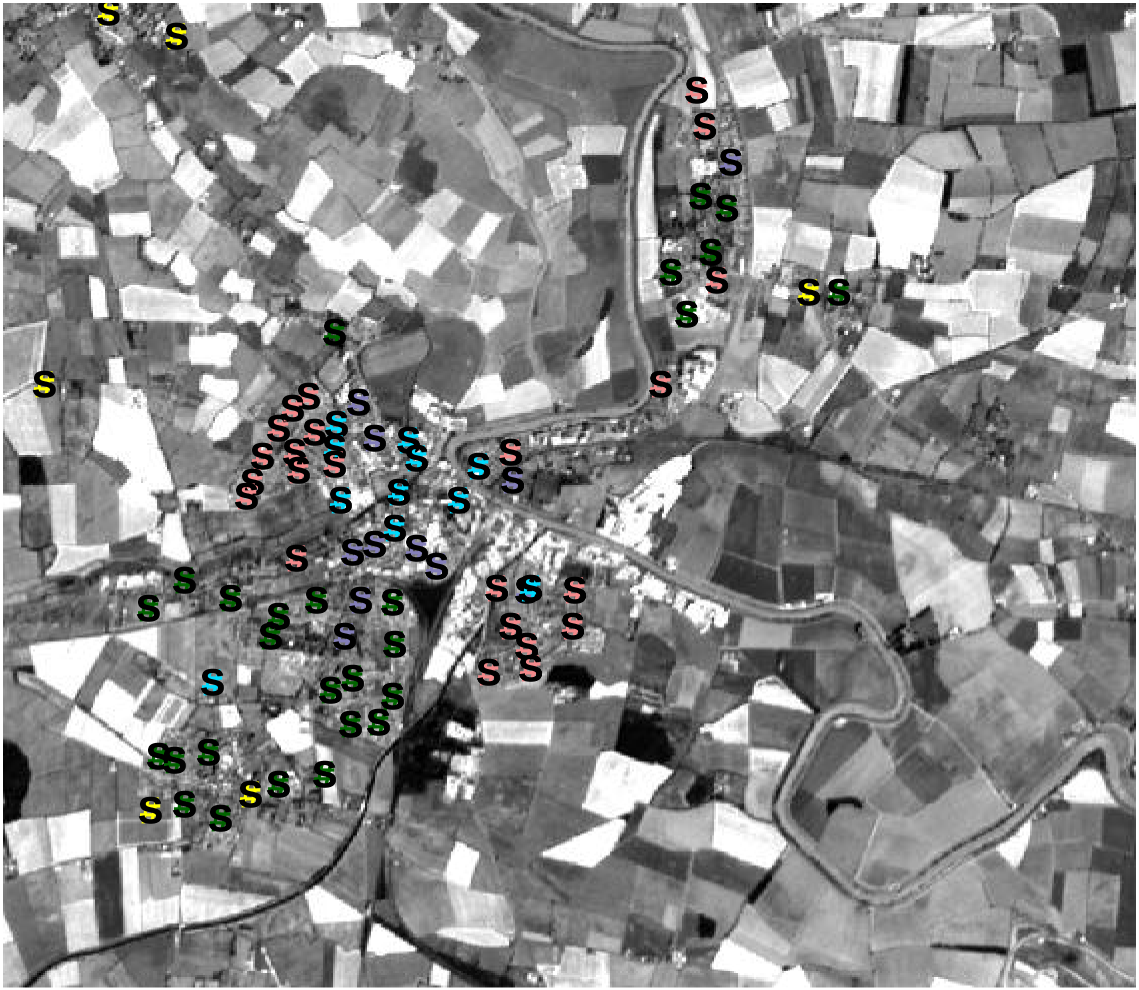


Ordnance Survey Crown Copyright

🟥1: Blue Collar Communities 🟩2: City Living 🟨3: Countryside 🟩4: Prospering Suburbs 🟦5: Constrained by Circumstances 🟪6: Typical Traits 🟧7: Multicultural

Figure 5.20 shows a SPOT satellite image (resolution 5-20m) of the town of Selby in North Yorkshire. Clear physical and man-made features can be seen on the image. Using a satellite image to add context to the classification works in a similar way to using a map, only with a satellite image the topography of the area becomes more apparent. Selby is a small market town built on a bend in the River Ouse. To the south of the town is the village of Brayton and the main roads to Leeds and Doncaster. This is the most prosperous part of town and is dominated by the *Prospering Suburbs* super-group. Clear clustering of the other super-group types can also be seen. *Typical Traits* and *Constrained by Circumstances* areas are located in the centre of town and two estates of 'Blue Collar Communities' are found to the east and west of the town. To the north of the town over the river is the village of Barlby, which is the first stop on the way to York 12 miles up the road. Barlby has a mixed residential picture with significant numbers of older residents but there is also a significant amount of new build that has attracted some young families to the area. Between Selby and Barlby is a non-residential area that is occupied by a large cattle feed factory, this can be seen on the image between the two river bends where there is no dot. The classification gives an accurate representation of Selby's social make-up and clearly demarcates the social areas within the town.

Figure 5.20: Mapping the OA classification at super-group level using centroids for Selby and surrounding area overlaid on SPOT satellite image



©SPOT Source: Satellite Image Data Service http://www.jisc.ac.uk/coll_landmap.html

🟥1: Blue Collar Communities 🟩2: City Living 🟨3: Countryside 🟩4: Prospering Suburbs 🟦5: Constrained by Circumstances 🟪6: Typical Traits 🟧7: Multicultural

Traditional maps of the OA classification at a national scale can be seen in Appendix E.

5.6 Conclusions

What can be concluded from the creation of the classification? Has what was set out to be created been achieved? Well an Output Area classification has been successfully created; it clearly and accurately splits the population of the UK into a hierarchy of 7, 21 and 52 types based on their residence. Associated data have been produced to go with the classification to aid understanding and assist in the use of the classification.

This chapter has discussed of all the decisions that were made during the creation of the classification and the reasons behind them. The chapter discusses the inclusion and exclusion of variables from the classification, it elucidates the building of the classification database and the careful data checks that were performed on it. The chapter explains clustering process and the creation of the classification and the thought processes behind it. The clusters have been named and explained through a careful and considered process. Then the classification was brought to life by adding the reality back into the classification with the use of a variety of mapping and visualisation techniques.

The classification was published by the ONS as an official national statistic on the 29th July 2005 and is available via the ONS website: http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/oa/default.asp or can be ordered on CD (Appendix F) from the ONS. An alternative source for the classification including additional information can be found on the University of Leeds website: http://www.geog.leeds.ac.uk/people/d.vickers/OAclassinfo.html

Explaining the creation of the classification is not the end of the story. To fully understand the classification that has been created it must be fully investigated. This will be done over the next two chapters. Chapter 6 will discus, the quality assurance investigations that were performed on the classification and examine some of its idiosyncrasies. Chapter 7 will then use the OA classification to help investigate and explain a selection of case studies based on current socio-demographic issues.

Chapter Six: Quality Assuring and Adding Value to the OA Classification

6.1 Introduction

It is one thing to create a classification at a small scale; it is another to create a classification which accurately represents the reality on the ground. The purpose of this chapter is to investigate the quality of the OA classification. A variety of methods will be used to investigate several different aspects of the classification and how well it represents the real world. The methods include a consultation exercise, which compares the classification to people's perceptions of the area in which they live.

The structure of the chapter is as follows. Section 6.2 concentrates on the importance of the variable choice by examining how sensitive the classification is to the removal of each variable. The reduction in variability within the dataset, which can be attributed to each variable, is assessed. Section 6.3 examines the change in variability both between and within clusters with movement down the hierarchy of the classification. Section 6.4 looks at those OAs which are atypical of their clusters, and establishes reasons why some areas do not fit the classification as well as others. Section 6.5 displays the results of a ground truthing exercise where photos of areas are related to the classes in which those areas have been placed. Section 6.6 compares each OA with all seven cluster centres at the super-group level to create an *ad hoc* fuzzy classification system. Section 6.7 outlines the implementation of and the results produced by a consultation exercise, designed to use the expertise of colleagues to validate the classification in an area which they know. Section 6.8 concludes the chapter commenting on the success of the different forms of quality assurance and what the results suggest about the classification.

6.2 Variability Reduction from Clustering and the Power of Each Variable

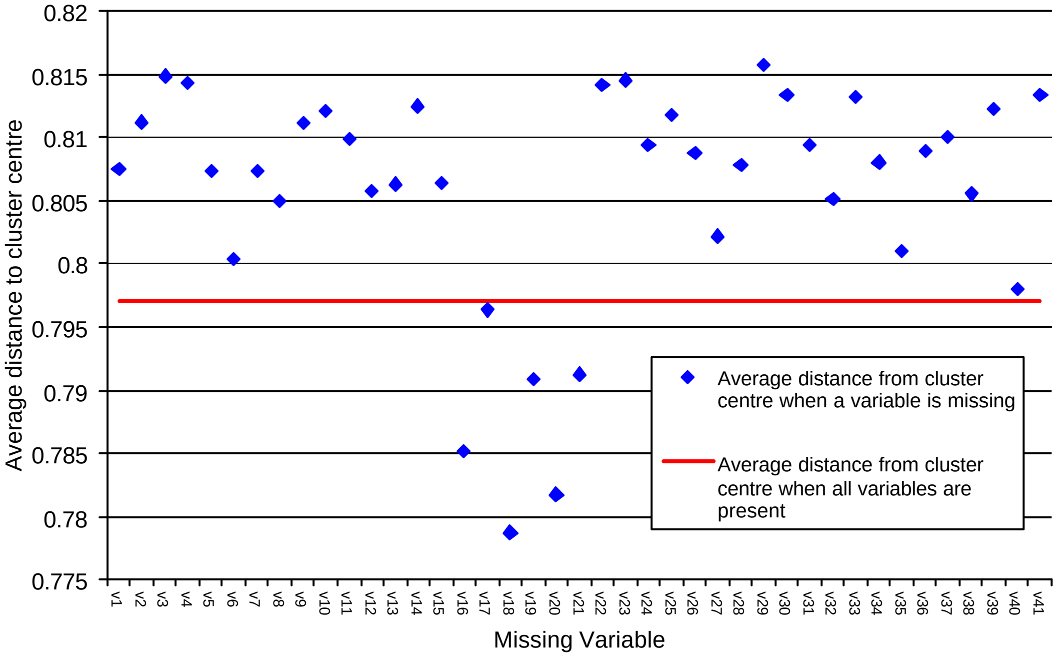
This section comprises two forms of analysis: firstly, sensitivity analysis will show what effect each variable has on the classification by establishing what would happen if each variable was removed from the classification. The second form of analysis will be to test the reduction in variability that the classification produces for each variable. This is important for several reasons: first, it exemplifies the importance of the variable selection in general and second, it gives an indication of the extent to which each variable affects the classification. Although all variables were equally weighted this does not mean that they have the same influence in the classification. The distribution of values within each variable and the skew of the data in each variable will affect the grouping procedure in different ways, and therefore alter the amount that each variable affects the classification. The examination of the reduction in variability of each variable can be compared with a previous study that will give an indication as to how the OA classification compares to other systems.

6.2.1 Sensitivity Analysis

The method that is used to test the sensitivity of each variable is to systematically remove each variable from the database and run the classification again, recreating the super-group level. This was done 41 times each time with one of the original set of variables missing. The average distance of OAs away from their cluster centre was recorded for each rerunning of the classification. These distances were then compared to each other and the average distance of OAs from their cluster centre for the actual classification. To enable comparison to the actual classification the average distance from cluster centre had to be multiplied by 0.9756 (40/41) to account for the fact that it contains one more variable. By examining the difference that removing each variable from the classification makes to the average difference from cluster centre, an assessment can be made as to how much effect each variable has on the classification. Variables that cause the greatest change in the average difference from cluster centre have the greatest effect on the classification, and those variables that cause the smallest change in the average difference from cluster centre having the least effect on the classification. This can also be compared with other indicators of the effect that variables are having on the classifications such as the values shown in the pen portrait profiles in Figures 5.9 -5.15.

Figure 6.1 shows the results of the sensitivity analysis. The variable numbers on the x axis represent each of the 41 variables. Refer to Table 5.5 for the name of each variable. The most obvious effect of the sensitivity analysis is the difference between the variables whose removal causes an increase in the average distance from cluster centre (above the red line) and those variables whose removal causes a decrease in the average distance from cluster centre (below the red line). The removal of six of the forty one variables cause a reduction in the average distance from cluster centre, Variables 16-21 are in fact the tenure and housing type variables: v16 Rent (public), v17 Rent (private), v18 Detached Housing, v19 Terraced Housing, v20 All Flats and v21 No Central Heating. If the removal each of one of these variables causes a reduction in the average distance from cluster centre then their inclusion in the classification causes the average distance from cluster centre to be greater than it would otherwise. However, it would be foolish to suggest that by removing these variables the classification would be improved. These variables are all from the same domain (housing), to remove these variables from the classification would deprive the classification of a lot of important information.

Figure 6.1: Sensitivity analysis results: showing the effect of removing each variable from the classification on the average distance from cluster centre.



The increase in the distance from cluster centre caused by the inclusion of these variables in the classification is not caused by these variables being irrelevant and unrepresentative. If they were, these variables would not be so distinctive within the classification. If these variables were unimportant we would expect to see OAs that have similar values for these variables to be in different clusters, but that is simply not the case. The cluster profiles as seen in § 5.9.2, show these variables to be among the most distinctive within the classification. What is actually happening with the inclusion of these variables is that they have such a powerful effect on the classification the average distance from cluster centre is increased due to the effect that the housing variables have on the other variables in the classification. The housing variables make OAs move between clusters based on their values because of the strength they have, this has the effect of increasing the overall distance from cluster centre because variables other than the housing variables have increased in distance from the cluster centre. There are several reasons why housing type has such a strong affect on the classification. Firstly, most streets and therefore most OAs consist of a single housing type or very similar housing types, meaning that housing variables are almost always heavily skewed in one way or another for all housing variables. Secondly, housing type was one of the variables that was use to define the OA boundaries in their creation so OAs are likely to be fairly homogeneous in terms of housing type, again increasing the likelihood of extreme values for these variables and therefore their effect on the classification.

It is clear that the housing variables have a great effect on the classification, but which other variables have a significant effect? No other variables cause the distance from cluster centre to increase in the same way as the housing domain, but the inclusion of each variable has a different effect on the classification. The removal of variable 29 LLTI causes the greatest increase in the distance from cluster centre and therefore its inclusion reduces the average distance from cluster centre within the classification. Other variables, the removal of which causes large increases in the average distance to cluster centre are v3 Age 25-44, v23 People per room, v4 Age 45-64 and v22 Average house Size. The reason why these variables have such a big effect on the classification is hard to interpret due to inter-correlation and colinearity within the dataset. It would easy to just assume that these variables are the most representative of underlying social trends. However, it is more likely that these variables are having a strong effect within the dataset because they pick up things that other variables within the dataset don't, rather than representing an important social trend. The most important social trends are more likely to be represented by more than one variable. Removing a variable that represents an important social trend may not always have the greatest effect statistically, but the small amount it adds to the classification could be more important than the statistics suggests.

The removal of v40 financial intermediation employment causes the least change in the average distance from cluster centre. The change that is caused to the average distance to cluster centre by this variable is significantly less than all other variables, over three times less than the next least, v6 Indian, Pakistani or Bangladeshi. Other variables that have a relatively small effect on the average distance to cluster centre are v35 Agriculture/Fishing Employment, v27 Public Transport to work and v8 Born outside the UK. These are quite a diverse set of variables. What the variables do have in common is that they have comparatively extreme values for one or at most two super-groups, but relatively low values for the rest of the super-groups. V40 Financial intermediation employment shows an extreme value for super-group 2 City Living; v6 Indian, Pakistani or Bangladeshi shows an extreme value for super-group 7 Multicultural; v35 Agriculture/Fishing Employment shows an extreme value for super-group 3 Countryside, v27 Public Transport to work shows an extreme values for super-group 2 City Living and for super-group 7 Multicultural, v8 Born outside the UK shows an extreme value for super-group 7 Multicultural. Although the removal of these variables has little effect on the average distance from cluster centre it is likely that the clusters for which these variables have extreme values will be greatly affected by their removal.

The sensitivity analysis has shown that by systematically removing each variable and rerunning the classification a different result will be produced. The removal of each variable affects the classification in a different way and to a different extent. However, it is difficult to assess the importance of a variable to the classification by the change in the average distance to cluster caused by its removal from the classification. The variables need to be looked at as more than just statistics. A variable can't just be removed because they seem to only have a small effect on the classification. A variable that only seems to have a small effect overall could be vital to the formation of an individual cluster. It is vital to consider what that variable represents and why it was included in the classification in the first place. It is reckless to remove variables from a classification after clustering as it is near impossible to gauge whether or not it has improved the classification and as long as the reasoning for the original variable selection was sound, removing a variable from the analysis cannot really be justified.

There are several difficulties in conducting such assessment on a small scale classification for the whole UK. Changing the variables in the classification will improve the classification for some areas, but will have a negative effect in others. To fully evaluate the extent of the effect of the removal of each variable, the movement of each individual OA between clusters and its changing distance from cluster centre would need to be examined. However, assessing the effect of the removal of each variable from the classification on each and every OA and then making an assessment as to which has the greatest effect on the classification is literally an impossible task.

6.2.2 Variability Reduction from the Clustering Process

Voas and Williamson (2001a) argue that *"much of the modest reduction in total variance from classification can be reproduced by a relatively simple approach"* (p73). This assertion was based on analysis carried out comparing the GB Profiles (99 clusters, 85 variables) and Super Profiles (128 clusters) systems with a simple *ad hoc* system of 96 clusters made from just 6 variables and a Townsend index split into 100 classes.

By examining the differences along each individual variable axis they calculated the average reduction in dispersion for each variable achieved by the classification systems. They showed that GB profiles achieved a reduction in variance of 33.9% across a list of 54 variables selected by the authors, whereas their *ad hoc* system achieved a reduction of 25.4% over the same variables, a reduction that represents three quarters of the more complex system. Although acknowledging that the selection of variables was in their favour, an assertion was made that *"they* [Geodemographic Classifications] *do not necessarily have any special advantage in reducing within-group heterogeneity* "(Voas and Williamson 2001a p73).

It is important to remember that, splitting a single variable in the middle will halve the within cluster difference and a 100 clusters can reduce the variance by 99% in a single variable system

(Flowerdew 1990). However, it is unreasonable to expect a system based on multiple variables to show a reduction in variability of a level anywhere near this for all variables. By having a system of multiple variables it impossible for large reductions in variance to take place for all variables as the relationship between the variables will mean that a reduction in the variance of one will group together dissimilar values for another variable. This process is exacerbated by every extra variable that is added to the classification. This provides further evidence to the view expressed in Chapter 5, that the traditional approach of many commercial firms to include as many variables as possible is not necessarily the best solution. Although adding more variables to the classification can add to its predictive power, the reduction in variability for the other variables can be restricted. The advantage of classifications is not that they can necessarily account for any more variability of data than one individual variable, but that the same classification can be used reliably on any dataset. The results that Voas and Williamson (2001a) found and the conclusions that they drew from them made it essential to put the OA classification through the same test which is outlined below:

(a) Calculate the variance of each variable for each cluster.

(b) Weight the values of (a) by the population of each cluster for each variable (i.e. if a cluster contains 6.3% of the population, it will receive a weight of 0.063).

(c) Sum the values from (b) to give a mean within class difference.

(d) Divide the value from (c) by the overall mean absolute difference (i.e. as if all OAs were in the same cluster). Subtract (d) from 100 to give the reduction in variability for each variable.

Adapted from Voas and Williamson 2001a p 70 & 72

The results of the reduction in variability test are shown in Table 6.1. At the super-group level (7 clusters) the average reduction in variability was 31.0%, 39.8% at the group level (21 clusters) and 44.5% at the sub-group level (52 clusters). The variables that show the greatest reduction in variability are perhaps not surprising; they are variable that the Cluster Portraits (§ 5.5.3) show to have extreme value, either high presence or absence of those variables in each cluster. The variable that shows the highest reduction in variability is all flats at 83.5% at the sub-group level and 79.0% at the group level. Three more variables: Rent (Public), detached housing and 2+ Car households show a variability reduction of over 70% at both the sub-group and group levels. The variable that shows the lowest reduction in variability is Health and Social work employment, which shows a reduction of just 2.2% at the super-group level. A further four variables (Hotel & Catering employment, Wholesale/retail trade employment, Provide unpaid care and Mining/Quarrying/Construction employment) have a reduction in variability of between 10% and 20% at the sub-group level. Perhaps the most surprising result is the comparatively small reduction in variability seen for the Students (All) variable, just 21.3% at the sub-group level. This is probably because the student variable in the census also includes further as well as higher education students, many of whom are likely to still live with parents therefore reducing the effect of the clustering of higher education students who tend to live in a more concentrated pattern.

Table 6.1: Reduction in variability for each variable in the classification (ordered by largest reduction at sub-group level)

Variable	Super -group	Group	Sub -group
V20: All Flats	63.7	79.0	83.5
V16: Rent (Public)	61.7	75.5	77.9
V19: Detached Housing	62.5	72.2	76.8
V26: 2+ Car households	63.0	71.6	74.1
V18: Terraced Housing	57.9	60.5	69.9
V22: Rooms per household	56.4	65.8	68.4
V7: Black African, Black Caribbean or Other Black	47.8	60.5	63.7
V9: Population Density	41.0	57.3	60.5
V24: HE Qualification	32.7	39.8	60.3
V8: Born Outside the UK	36.8	53.2	60.0
V6: Indian, Pakistani or Bangladeshi	43.0	53.1	59.6
V17: Rent (Private)	44.1	53.3	57.4
V25: Routine/Semi-Routine Occupation	37.2	51.9	55.6
V27: Public Transport to work	41.4	49.2	55.0
V23: People per room	43.0	50.7	53.9
V11: Single person household (not pensioner)	42.3	47.3	52.0
V29: LlTI (SIR)	40.5	47.6	51.2
V21: No central heating	34.3	44.3	49.5
V35: Agriculture/Fishing employment	34.0	42.0	45.5
V13: Lone Parent household	32.3	39.3	44.5
V2: Age 5-14	30.2	40.0	44.2
V14: Two adults no children	37.9	41.9	44.2
V10: Separated/Divorced	30.5	39.5	42.2
V32: Unemployed	21.6	37.2	41.7
V4: Age 45-64	15.5	32.5	39.3
V3: Age 25-44	9.3	30.2	38.6
V5: Age 65+	21.1	32.7	37.2
V12: Single pensioner household	18.5	28.2	35.8
V33: Working part-time	11.9	25.5	33.6
V34: Economically inactive looking after family	26.4	30.8	32.0
V15: Households with non-dependant children	10.4	24.8	29.9
V28: Work from home	19.4	24.7	28.2
V1: Age 0-4	18.1	24.5	28.1
V40: Financial intermediation employment	17.7	21.1	27.2
V37: Manufacturing employment	14.6	19.4	24.1
V31: Students (All)	16.2	17.7	21.3
V36: Mining/Quarrying/Construction employment	11.6	13.3	15.4
V30: Provide unpaid care	7.3	12.1	14.6
V41: Wholesale/retail trade employment	9.2	12.5	14.2
V38: Hotel & Catering employment	6.9	8.7	12.7
V39: Health and Social work employment	0.8	1.6	2.2
Average	31.0	39.8	44.5

The results of the reduction in variability test results compare more than favourably with the results of the classifications tested by Voas and Williamson (2001a). Voas and Williamson (2001a) examined systems split into 96, 99, 100 and 128 clusters. The reduction in variability that they found for these systems produced an average reduction in variability of 33.9% (GB Profiles), 32.0 (Super Profiles), 25.4 (ad hoc) and 13.7 (Townsend), compared to the OA classification that showed average reduction in variability of 44.5% at the sub-group level, 39.8% at the group level and 31.0% at the super-group level. Both the sub-group and group levels of the OA classification easily out perform all the systems examined by Voas and

Williamson (2001a) despite only having roughly half and five times fewer clusters respectively. Only GB Profiles and Super Profiles performed (slightly) better than the super-group level despite it having over fourteen times the number of clusters. As an increase in the number of clusters should increase the reduction in variability, these figures are very positive for the OA Classification.

Those variables that were tested by Voas and Williamson (2001a) that are comparable to variables used in the OA classification are shown in Table 6.2, 25 were identified. For only two variables (figures highlighted in **bold**) is the reduction in any of the other systems greater than it is for the sub-group level of the OA Classification. The variability reduction in the Townsend system for the Unemployment variable is greater than the OA classification, but the Townsend system is a measure of deprivation and is focused on identifying the most deprived area that are likely to show a distinction on unemployment (Townsend et *al*. 1988). The Townsend index contains just four variables one of which is unemployment, the fewer variables used would make it likely that it should perform poorly overall, but very well for those variables included within it. GB Profiles, Super Profiles and Voas and Williamson's *ad hoc* system out perform the OA Classification on the pensioners variable, this is perhaps a little surprising, but there is a reason for this. LLTI was age standardised in the OA classification to counteract the problem of older people generally having poorer health. The documentation of the other systems indicate that the variable was included in unstandardised form, therefore effectively adding extra weight to the pensioners variable. LLTI is not a variable included in the Townsend index the same effect is not seen there.

Whilst for eleven variables such as Flats, all four of the other classifications are out performed by all three levels of the OA classification (figures highlighted in *italics*). This is quite an impressive achievement, as the super-group level of the OA classification only has 7 clusters so to have a variability reduction for any variable greater than a system that has 128 clusters shows the great discriminatory power that the OA Classification has. It is important to remember that with the difference in scale in the number of cluster produced, the OA classification should really be out performed by all of the other systems.

Table 6.2: Reduction in variability for each variable in previous classification systems (ordered by largest reduction for GB Profiles)

Variable	GB Profiles	Super Profiles	_Ad hoc_	Townsend
Detached House	68.7	52.4	41.3	34.7
Public Rental	_63.5_	_53.3_	_38.6_	_41.4_
Two or more cars	61.2	56.6	53.2	46.0
Terraced Housing	59.9	34.2	17.1	13.5
Non-white residents*	50.4	50.1	17.1	12.6
Flats	_49.2_	_37.6_	_26.2_	_16.9_
Born Outside UK	_44.4_	_48.5_	_16.8_	_7.6_
Unemployed Males*	44.0	43.6	39.8	**54.7**
One Person in household	43.0	42.2	53.5	13.1
Pensioners*	**41.8**	**45.6**	**44.9**	2.6
Residents with long-term illness	40.7	39.0	38.1	16.0
No Central Heating	39.9	24.9	13.9	13.2
Private Rental	_37.7_	_39.4_	_19.2_	_5.8_
Adults with qualifications*	_36.3_	_37.0_	_21.7_	_14.0_
Head in manual occupation*	_35.1_	_31.3_	_17.7_	_11.4_
Over 0.5 persons/room*	_34.0_	_36.5_	_28.9_	_11.1_
Travel to work by public transport	_34.2_	_29.5_	_16.4_	_13.8_
One Pensioner household	34.0	38.6	42.3	6.0
Child under five in household*	32.4	38.0	26.8	6.0
Part-time females*	_28.4_	_30.9_	_0.4_	_14.2_
Inactive females*	27.8	29.2	46.1	15.8
Males in finance & other services*	19.5	19.1	8.0	2.4
Student males & females (average)	18.1	20.4	9.7	2.9
Lone parent with dependent child	_18.6_	_17.6_	_16.0_	_14.9_
Males in manufacturing*	_10.3_	_13.4_	_4.6_	_0.8_

Source: adapted from Voas and Williamson 2001a p71

*Denotes that these variables are not 100% comparable to those used in the OA Classification, but were considered similar and are included with recognition that there could be some matching error (i.e. Part-time and Part-time Females or Pensioners and Aged 65+)

In the interests of fairness and consistency the average reduction in variability was computed for all three levels of the OA classification and the four systems tested by Voas and Williamson (2001a), but this time just on the 25 comparable variables. The results can be seen in Table 6.3. The average reduction in variability for the OA classification is 51.0% at the sub-group level, 45.4% at the group level and 35.5% at the super-group level. The average values for the comparable variables are higher than those for the entire variable list used in the OA Classification. GB Profiles and Super Groups have also seen a slight increase in variability reduction, while the values for the *ad hoc* and Townsend systems have remained comparatively stable. The results give approval to the OA classification with the Sub-group and Group levels having the greatest reduction in variability. When the number of clusters produced by each system is taken into account the OA classification performs even better, especially the sub-group level. When weighting the variability reduction by the number of clusters that caused the reduction, the OA Classification sub group level performs almost 13 times better than GB Profiles, around 18 times better than Super Profiles or the *ad hoc* system and a staggering 32 times better than the Townsend system.

Table 6.3: The average reduction in variability for comparable variables

System	Average Variability Reduction for comparable variables	Number of Clusters	Weighted Reduction (reduction per cluster)
OA Classification Sub-groups	51.0	52	0.98
OA Classification Groups	45.4	21	2.16
GB Profiles	39.4	99	0.40
Super Profiles	36.9	128	0.29
OA Classification Super-groups	35.5	7	5.07
Ad hoc	26.0	96	0.27
Townsend	15.5	100	0.16

The results of this comparison not only show the quality of the OA Classification, but also challenge Voas and Williamson’s view that general purpose classifications do not offer enough reduction in variability to be useful. The above examples have shown that for comparable variables the OA classification can achieve a variability reduction of 2 or 3 times more than a simpler system even though the other systems are split into twice as many clusters.

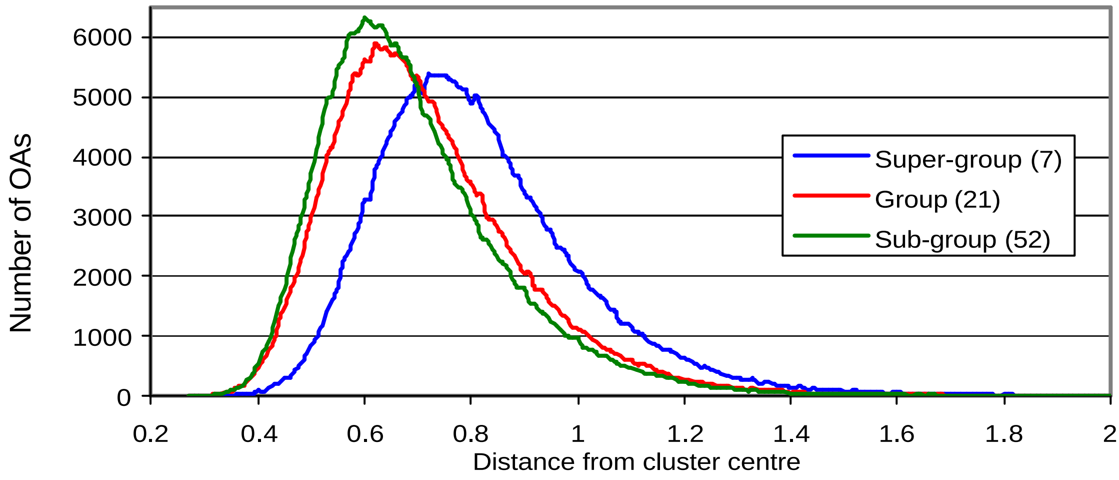
6.3 Variability Reduction Within the Cluster Hierarchy

Variability reduction goes hand in hand with the creation of a hierarchical classification system. The hierarchical system allows the user to select the amount of focus and detail they want by selecting all or part of the classification at different levels of the hierarchy. The hierarchy works by reducing the variability within the clusters and increasing distance between cluster centres, with movement down the hierarchy. There are two ways in which the reduction in variability within the hierarchy can be measured. By either looking at the average distance from the cluster centre within each cluster; alternatively, the between cluster centre differences can be examined.

6.3.1 Within Cluster Distance

A good classification should aim to have the smallest possible within cluster distances, making each cluster as compact as possible. It would follow that the within cluster distance should reduce as the number of clusters increases. This works perfectly in theory, but does it work in reality? By examining the average distance to cluster centre at each of the three levels of the classification it is clear that as the number of clusters within the classification increases the average within cluster distance decreases. The average distance to cluster centre at the Super-group level is 0.82, the average distance to cluster centre at the Group level is 0.75 and the average distance to cluster centre at the Sub-group level is 0.70. Figure 6.2 shows the frequency distribution of the distances from cluster centres. All three levels of the classification show a positive skew with a long tail. As the number of clusters increases down the hierarchy, the skew increases as the number of OAs with smaller distances from cluster centre becomes greater. The biggest change comes with the movement from the sub-group to the group level. There is also a clear shift between the group and super-group levels, but this is not as pronounced.

Figure 6.2: The frequency distribution of the distances from cluster centres



Across all clusters the average distance to cluster centre reduces with the increase in the number of clusters. However, this does not mean that the average distance to cluster centre is reduced for all clusters. Table 6.4 shows that the within cluster difference reduces in most cases with movement down the hierarchy (an increase in the number of groups). However, this does not always happen for all clusters; the reduction of within cluster distance for one cluster can cause an increase in the within cluster distance for another (this is especially likely in a hierarchical system where additional restraints are put upon the cluster formation). An example of this is super-group 5 which has a within cluster distance of 0.89. This super-group splits down into three groups, two of which show a reduction in within cluster distance (5b 0.75 and 5c 0.86). However, group 5a shows an increase in the within cluster distance to 1.03. Why has this happened? The answer is fairly simple. Super-group 5 contains a comparatively compact formation of OAs around the cluster centre and it also contains a slightly less compact set of OAs slightly further from the cluster centre. When super-group 5 is split into its three constituent groups, two of the groups are formed from the relatively compact core of the cluster and therefore have a lower within cluster distance than the super-group. The third group (5a) is created from the slightly less compact grouping and therefore its average distance from cluster centre is greater than that of the whole of super-group 5.

Table 6.4: The reduction of within cluster distance
(figures refer to average Squared Euclidian distance of OAs from the cluster centre)

Super -group	Group	Sub -group	Super -group	Group	Sub -group	Super -group	Group	Sub -group
1: 0.77	1a: 0.75	1a1: 0.68	4: 0.79	4a: 0.74	4a1: 0.73	6: 0.79	6a: 0.71	6a1: 0.65
		1a2: 0.72			4a2: 0.67			6a2: 0.71
		1a3: 0.71		4b: 0.69	4b1: 0.68		6b: 0.69	6b1: 0.68
	1b: 0.69	1b1: 0.65			4b2: 0.61			6b2: 0.66
		1b2: 0.69			4b3: 0.65			6b3: 0.60
	1c: 0.71	1c1: 0.66			4b4: 0.61		6c: 0.69	6c1: 0.64
		1c2: 0.68		4c: 0.70	4c1: 0.66			6c2: 0.63
		1c3: 0.66			4c2: 0.63		6d: 0.76	6d1: 0.76
2: 0.98	2a: 0.99	2a1: 1.02			4c3: 0.67			6d2: 0.71
		2a2: 0.83		4d: 0.75	4d1: 0.72	7: 0.82	7a: 0.77	7a1: 0.70
	2b: 0.89	2b1: 0.91			4d2: 0.72			7a2: 0.69
		2b2: 0.80	5: 0.89	5a: 1.03	5a1: 0.95			7a3: 0.68
3: 0.76	3a: 0.68	3a1: 0.63			5a2: 1.09		7b: 0.74	7b1: 0.67
		3a2: 0.66		5b: 0.76	5b1: 0.71			7b2: 0.72
	3b: 0.72	3b1: 0.70			5b2: 0.73			
		3b2: 0.67			5b3: 0.66			
	3c: 0.70	3c1: 0.63			5b4: 0.70			
		3c2: 0.72		5c: 0.86	5c1: 0.78			
					5c2: 0.90			
					5c3: 0.73			

By following any of the clusters through the hierarchy they become increasingly distinctive as their membership reduces in size. Even for those clusters which see an increase in average distance from cluster centre when split it would be misleading to say that the lower level is less representative than the higher level. Even though the average distance to cluster centre is greater at the lower level, the cluster centre at that level is more representative of the areas within that cluster that the cluster centre at the higher level.

6.3.2 Between Cluster Centre Distance

In contrast to the within cluster distance, the between cluster centre distance needs to be as large as possible, making each group as different as possible from each of the others. As the number of clusters increases with the hierarchy, a broadening out of distances between cluster centres can be seen. Groups which are formed from the same super-group can show comparatively small distances between their cluster centres and the distances between sub-groups formed from the same group can be even smaller. Figures 6.3 - 6.5 show three matrices (one for each level) that show the distance between the cluster centres of each cluster with every other cluster. The matrices are colour coded to make the values easier to interpret (white <0.75, light blue 0.75<1.25, medium blue 1.25<1.75, dark blue 1.75=). These distances represent the Euclidian distance between the centre of each cluster, they are not in themselves meaningful unless compared to each other.

Figure 6.3 shows a matrix of the distances between the cluster centres at the super-group level. The largest differences are shown between clusters that represent very different types of area, for example between super-groups 3 *Countryside* and 7 *Multicultural* with difference (Euclidian) of 1.55. The two super-groups with the least distance between their cluster centres and therefore most similar to each other are 1 *Blue Collar Communities* and 5 *Constrained by Circumstances* with a distance of 0.68. The average distance between any two cluster centres at the super-group level is 1.09, therefore the two clusters that are furthest apart are 42% further apart than average and the two clusters that are closest together are 38% closer together than average.

Figure 6.3: The difference between cluster centres at the Super-group level

1	2	3	4	5	6	7	
-	1.34	1.07	1.12	0.68	0.70	1.11	1
	-	1.42	1.41	1.05	0.95	0.88	2
		-	0.72	1.34	0.81	1.55	3
			-	1.39	0.85	1.55	4
				-	0.90	0.98	5
					-	1.01	6
						-	7

Figure 6.4 shows a matrix of the distances between the cluster centres at the group level. The average distance between any two cluster centres at the group level is 1.11. The largest difference between two clusters at this level is between 3b *Agricultural* and 7b *Afro-Caribbean Communities* with a distance of 1.99 (80% above the average distance). The two most similar clusters at the group level are 3a *Village Life* and 3c *Accessible Countryside*, perhaps unsurprisingly as these two groups belong to the same super group the distance between them is just 0.47 (58% below the average distance).

Figure 6.5 shows a matrix of the distances between the cluster centres at the sub-group level. The average distance between any two cluster centres at the sub-group level is 1.11. The largest difference between two clusters at this level is between 3b2 *Agricultural (2)* and 7b2 *Afro-Caribbean Communities (2)* and is 2.2 (92% above the average distance). Interestingly these are members of the two groups which are furthest apart. The two most similar clusters at the sub-group level are 1b1 *Younger Blue Collar (1)* and 1b2 *Younger Blue Collar (2)*, like the two most similar types at the group level; these two types are members of the same cluster type at the next level up the hierarchy, the distance between them is just 0.4 (56% below the average distance).

Figure 6.4: the difference between cluster centres at the Group level

1a	1b	1c	2a	2b	3a	3b	3c	4a	4b	4c	4d	5a	5b	5c	6a	6b	6c	6d	7a	7b	
-	0.50	0.50	1.69	1.35	1.05	1.54	1.34	1.29	1.45	1.08	1.31	1.14	0.68	0.84	0.89	0.90	0.96	1.12	1.16	1.40	1a
	-	0.51	1.55	1.19	0.98	1.45	1.27	1.36	1.53	1.11	1.32	1.17	0.60	0.82	0.78	0.77	0.68	1.04	0.92	1.24	1b
		-	1.60	1.19	0.59	1.11	0.93	1.05	1.12	0.76	0.99	1.13	0.65	1.01	0.65	0.57	0.82	0.88	1.11	1.43	1c
			-	0.61	1.67	1.88	1.65	1.78	1.86	1.63	1.46	1.20	1.27	1.36	1.50	1.27	1.21	1.19	1.16	1.00	2a
				-	1.20	1.46	1.15	1.39	1.48	1.24	1.02	1.07	0.91	1.19	1.04	0.74	0.77	0.65	0.87	0.98	2b
					-	0.59	0.47	0.94	0.93	0.77	0.82	1.39	1.00	1.40	0.79	0.59	1.00	0.84	1.31	1.65	3a
						-	0.48	1.03	0.90	0.92	0.94	1.78	1.48	1.82	1.10	1.04	1.35	1.12	1.66	1.99	3b
							-	0.82	0.76	0.73	0.64	1.57	1.24	1.64	0.80	0.71	1.05	0.72	1.41	1.78	3c
								-	0.59	0.61	0.69	1.79	1.45	1.71	0.81	1.03	1.27	0.86	1.49	1.85	4a
									-	0.50	0.54	1.73	1.52	1.82	1.02	1.11	1.46	1.03	1.67	1.99	4b
										-	0.58	1.46	1.18	1.48	0.61	0.83	1.06	0.81	1.36	1.71	4c
											-	1.43	1.19	1.52	0.88	0.76	1.15	0.63	1.30	1.58	4d
												-	0.70	0.81	1.43	1.06	1.24	1.31	1.37	1.27	5a
													-	0.55	0.98	0.60	0.76	0.93	0.96	1.08	5b
														-	1.34	1.07	1.10	1.30	1.16	1.06	5c
															-	0.63	0.58	0.58	1.03	1.50	6a
																-	0.59	0.47	0.96	1.27	6b
																	-	0.64	0.81	1.21	6c
																		-	0.94	1.30	6d
																			-	0.69	7a
																				-	7b

Figure 6.5: the difference between cluster centres at the Sub-group level

1a1	1a2	1a3	1b1	1b2	1c1	1c2	1c3	2a1	2a2	2b1	2b2	3a1	3a2	3b1	3b2	3c1	3c2	4a1	4a2	4b1	4b2	4b3	4b4	4c1	4c2	4c3	4d1	4d2	5a1	5a2	5b1	5b2	5b3	5b4	5c1	5c2	5c3	6a1	6a2	6b1	6b2	6b3	6c1	6c2	6d1	6d2	7a1	7a2	7a3	7b1	7b2	
-	0.5	0.5	0.6	0.6	0.6	0.6	0.6	1.9	1.8	1.5	1.5	1.1	1.2	1.6	1.6	1.4	1.5	1.5	1.3	1.5	1.5	1.7	1.4	1.2	1.2	1.3	1.5	1.4	1.2	1.3	0.7	0.9	0.9	0.8	0.7	1.1	0.8	1.1	1.1	1.2	1.0	1.0	1.1	1.1	1.2	1.3	1.4	1.1	1.4	1.5	1.5	1a1
	-	0.5	0.6	0.6	0.6	0.7	0.6	1.7	1.7	1.4	1.4	1.1	1.1	1.6	1.6	1.3	1.4	1.5	1.2	1.5	1.4	1.6	1.3	1.1	1.0	1.1	1.4	1.3	1.1	1.2	0.6	0.9	0.8	0.9	0.8	1.3	0.9	0.9	0.8	1.1	0.9	0.9	1.0	1.0	1.1	1.2	1.4	1.0	1.3	1.4	1.6	1a2
		-	0.5	0.7	0.6	0.8	0.7	1.7	1.6	1.3	1.3	1.1	1.1	1.6	1.6	1.3	1.4	1.4	1.2	1.6	1.5	1.6	1.4	1.2	1.1	1.2	1.4	1.3	1.2	1.3	0.6	0.8	0.7	0.8	0.7	1.2	0.8	0.9	0.9	1.1	0.8	0.9	0.9	1.0	1.0	1.1	1.2	0.8	1.2	1.2	1.4	1a3
			-	0.4	0.5	0.7	0.6	1.6	1.5	1.2	1.2	1.0	1.0	1.5	1.5	1.3	1.3	1.6	1.3	1.6	1.6	1.7	1.5	1.2	1.1	1.2	1.4	1.3	1.1	1.3	0.6	0.7	0.5	0.5	0.7	1.0	0.5	0.9	0.9	0.9	0.7	0.7	0.7	0.7	1.0	1.1	1.0	0.7	1.1	1.1	1.3	1b1
				-	0.7	0.7	0.5	1.7	1.6	1.3	1.3	1.0	1.1	1.5	1.4	1.3	1.3	1.5	1.2	1.5	1.5	1.6	1.4	1.2	1.1	1.1	1.4	1.4	1.3	1.4	0.9	1.0	0.8	0.8	1.0	1.3	0.9	0.8	0.7	1.0	0.9	0.8	0.7	0.8	1.1	1.2	1.1	0.8	1.2	1.3	1.5	1b2
					-	0.5	0.4	1.6	1.6	1.1	1.2	0.7	0.6	1.2	1.1	1.0	0.9	1.3	1.0	1.2	1.3	1.4	1.1	0.9	0.8	1.1	1.2	1.0	1.1	1.1	0.6	0.5	0.7	0.6	0.9	1.2	0.8	0.7	0.8	0.7	0.5	0.5	0.9	0.8	0.9	0.9	1.3	0.9	1.2	1.3	1.5	1c1
						-	0.5	1.7	1.7	1.2	1.4	0.7	0.9	1.1	1.2	1.1	1.1	1.2	1.1	1.1	1.1	1.3	1.1	0.7	0.8	0.8	1.2	1.0	1.1	1.2	0.8	0.7	0.9	0.9	1.0	1.2	1.0	0.8	0.9	1.0	0.8	0.7	1.0	1.0	1.1	1.1	1.4	1.1	1.4	1.5	1.6	1c2
							-	1.8	1.7	1.3	1.3	0.6	0.8	1.1	1.1	0.9	1.0	1.2	0.9	1.1	1.1	1.3	1.0	0.8	0.7	0.9	1.2	1.0	1.3	1.3	0.9	0.8	0.9	0.9	1.1	1.4	1.1	0.6	0.7	0.9	0.7	0.6	0.9	0.9	0.9	1.0	1.3	1.0	1.3	1.5	1.7	1c3
								-	0.6	0.7	0.8	1.8	1.6	1.9	1.9	1.8	1.5	1.8	1.8	1.8	1.9	1.9	1.9	1.6	1.7	1.6	1.5	1.5	1.2	1.4	1.5	1.3	1.3	1.4	1.6	1.3	1.5	1.6	1.5	1.2	1.3	1.4	1.2	1.3	1.3	1.2	1.6	1.3	1.3	1.2	1.4	2a1
									-	0.8	0.6	1.8	1.6	2.0	1.9	1.8	1.6	1.9	1.8	2.0	2.0	1.9	1.9	1.7	1.8	1.6	1.5	1.5	1.2	1.4	1.4	1.3	1.2	1.4	1.5	1.3	1.4	1.6	1.5	1.3	1.3	1.5	1.2	1.3	1.2	1.3	1.3	1.1	1.0	0.8	0.9	2a2
										-	0.6	1.3	1.0	1.4	1.4	1.2	0.9	1.5	1.3	1.4	1.4	1.5	1.4	1.2	1.2	1.2	1.1	0.9	1.0	1.0	1.1	0.8	0.9	1.0	1.3	1.2	1.2	1.1	1.1	0.6	0.7	0.9	0.9	0.9	0.8	0.7	1.3	1.0	1.0	1.1	1.3	2b1
											-	1.4	1.2	1.6	1.6	1.4	1.2	1.6	1.4	1.6	1.7	1.6	1.5	1.4	1.4	1.3	1.2	1.2	1.2	1.3	1.2	1.0	0.9	1.1	1.4	1.3	1.2	1.1	1.0	0.8	0.8	1.0	0.7	0.9	0.7	0.8	1.1	0.8	0.7	0.8	1.2	2b2
												-	0.4	0.6	0.6	0.5	0.6	1.1	0.8	0.8	0.9	1.0	0.7	0.7	0.6	0.9	1.0	0.8	1.6	1.5	1.2	0.9	1.2	1.2	1.4	1.7	1.4	0.7	0.9	0.9	0.8	0.6	1.1	1.0	0.9	0.9	1.5	1.3	1.4	1.7	1.9	3a1
													-	0.7	0.6	0.6	0.5	1.2	1.0	1.0	1.1	1.2	0.9	0.9	0.8	1.1	1.0	0.8	1.4	1.3	1.1	0.7	1.0	1.0	1.3	1.5	1.3	0.8	1.0	0.6	0.6	0.5	1.1	0.9	0.9	0.8	1.5	1.2	1.4	1.5	1.7	3a2
														-	0.4	0.5	0.6	1.0	1.0	0.7	0.9	0.9	0.8	0.8	0.9	1.0	1.0	0.9	1.8	1.7	1.6	1.3	1.6	1.6	1.8	1.9	1.8	1.1	1.3	1.2	1.2	1.0	1.5	1.4	1.2	1.1	1.9	1.7	1.7	2.0	2.1	3b1
															-	0.5	0.6	1.3	1.1	1.0	1.2	1.2	1.0	1.0	1.0	1.2	1.2	1.0	1.9	1.7	1.6	1.3	1.6	1.5	1.8	2.0	1.8	1.0	1.3	1.0	1.1	1.0	1.4	1.3	1.2	1.1	1.8	1.6	1.7	1.9	2.2	3b2
																--	0.4	1.0	0.7	0.7	0.9	0.9	0.6	0.8	0.6	0.9	0.9	0.7	1.7	1.6	1.4	1.1	1.4	1.4	1.7	1.9	1.7	0.7	0.9	0.9	0.9	0.7	1.2	1.1	0.9	0.8	1.6	1.4	1.5	1.8	2.0	3c1
																	-	1.1	0.9	0.8	1.0	1.0	0.8	0.9	0.8	1.0	0.8	0.6	1.5	1.4	1.3	1.0	1.3	1.2	1.6	1.7	1.5	0.8	1.1	0.7	0.8	0.6	1.1	1.0	0.8	0.6	1.6	1.3	1.4	1.6	1.9	3c2
																		-	0.5	0.8	0.5	0.7	0.7	0.8	0.8	0.8	0.8	0.8	1.9	1.9	1.6	1.4	1.7	1.8	1.7	1.9	1.8	1.0	1.2	1.4	1.2	1.2	1.5	1.5	1.1	1.1	1.8	1.7	1.6	1.9	2.1	4a1
																			-	0.8	0.7	0.7	0.5	0.7	0.5	0.8	0.8	0.7	1.8	1.7	1.4	1.2	1.4	1.5	1.6	1.9	1.6	0.6	0.8	1.1	0.9	0.8	1.1	1.2	0.7	0.8	1.6	1.4	1.4	1.7	2.0	4a2
																				-	0.4	0.5	0.4	0.5	0.6	0.8	0.8	0.6	1.8	1.6	1.6	1.3	1.6	1.6	1.8	1.9	1.8	1.0	1.2	1.3	1.1	1.0	1.5	1.5	1.2	1.1	1.9	1.7	1.7	2.0	2.1	4b1
																					-	0.4	0.4	0.5	0.6	0.6	0.7	0.6	1.8	1.7	1.6	1.3	1.7	1.7	1.7	1.9	1.8	1.0	1.1	1.4	1.2	1.1	1.5	1.5	1.2	1.1	1.9	1.7	1.7	2.0	2.1	4b2
																						-	0.5	0.7	0.7	0.8	0.5	0.7	1.9	1.8	1.7	1.4	1.7	1.8	1.9	2.1	2.0	1.1	1.2	1.4	1.2	1.2	1.6	1.6	1.1	1.1	1.9	1.7	1.6	1.9	2.1	4b3
																							-	0.6	0.4	0.7	0.7	0.6	1.8	1.6	1.5	1.2	1.6	1.6	1.7	2.0	1.8	0.8	1.0	1.2	1.0	0.9	1.4	1.4	1.0	0.9	1.8	1.6	1.6	1.9	2.1	4b4
																								-	0.4	0.4	0.8	0.6	1.5	1.4	1.3	1.0	1.3	1.3	1.5	1.6	1.5	0.8	0.9	1.1	0.9	0.8	1.2	1.2	1.0	0.9	1.7	1.4	1.5	1.7	1.9	4c1
																									-	0.6	0.8	0.6	1.6	1.5	1.2	1.0	1.3	1.3	1.5	1.7	1.5	0.5	0.7	1.0	0.8	0.6	1.1	1.1	0.8	0.8	1.6	1.3	1.4	1.7	1.9	4c2
																										-	0.7	0.7	1.5	1.5	1.3	1.1	1.3	1.4	1.5	1.6	1.5	0.8	0.7	1.2	1.0	0.9	1.1	1.2	0.9	0.9	1.5	1.3	1.3	1.6	1.8	4c3
																											-	0.5	1.6	1.6	1.4	1.1	1.4	1.5	1.6	1.7	1.6	1.0	1.0	1.1	0.9	0.9	1.2	1.3	0.8	0.8	1.4	1.3	1.1	1.5	1.7	4d1
																												-	1.4	1.3	1.2	0.9	1.2	1.3	1.5	1.6	1.5	0.9	1.0	0.9	0.7	0.7	1.2	1.1	0.8	0.6	1.6	1.3	1.3	1.6	1.8	4d2
																													-	0.6	0.7	0.8	0.8	0.8	0.8	0.7	0.8	1.5	1.4	1.1	1.1	1.2	1.3	1.2	1.4	1.3	1.6	1.1	1.4	1.2	1.2	5a1
																														-	0.9	0.8	1.0	1.0	1.1	1.2	1.2	1.5	1.5	1.1	1.0	1.2	1.4	1.4	1.4	1.3	1.8	1.3	1.6	1.5	1.6	5a2
																															-	0.5	0.5	0.6	0.5	0.9	0.6	1.1	1.1	0.9	0.7	0.9	1.0	1.0	1.1	1.1	1.4	0.9	1.3	1.2	1.3	5b1
																																-	0.5	0.6	0.8	1.0	0.8	1.0	1.0	0.6	0.4	0.6	1.0	0.9	0.9	0.8	1.4	0.9	1.2	1.2	1.3	5b2
																																	-	0.4	0.7	0.8	0.5	1.0	1.0	0.7	0.6	0.8	0.7	0.7	1.0	1.0	1.1	0.6	1.0	0.9	1.1	5b3
																																		-	0.7	0.9	0.5	1.1	1.1	0.7	0.8	0.8	0.9	0.7	1.1	1.1	1.2	0.8	1.2	1.1	1.3	5b4
																																			-	0.6	0.4	1.4	1.3	1.2	1.0	1.1	1.2	1.2	1.3	1.3	1.4	1.0	1.4	1.2	1.2	5c1
																																				-	0.6	1.6	1.5	1.3	1.2	1.4	1.3	1.3	1.5	1.5	1.5	1.1	1.4	1.2	1.0	5c2
																																					-	1.3	1.2	1.1	1.0	1.1	1.0	1.0	1.2	1.3	1.2	0.8	1.2	1.0	1.1	5c3
																																						-	0.4	0.8	0.7	0.4	0.7	0.7	0.6	0.6	1.3	1.1	1.2	1.5	1.8	6a1
																																							-	0.9	0.8	0.7	0.5	0.7	0.6	0.7	1.1	0.9	1.0	1.3	1.6	6a2
																																								-	0.5	0.4	0.7	0.5	0.7	0.5	1.3	0.9	1.1	1.2	1.5	6b1
																																									-	0.4	0.7	0.7	0.5	0.5	1.2	0.8	1.0	1.1	1.4	6b2
																																										-	0.7	0.6	0.6	0.5	1.3	0.9	1.1	1.3	1.6	6b3
																																											-	0.4	0.6	0.7	0.9	0.7	0.8	1.0	1.4	6c1
																																												-	0.8	0.7	1.1	0.8	1.0	1.2	1.5	6c2
																																													-	0.4	1.1	0.9	0.8	1.1	1.5	6d1
																																														-	1.3	1.0	1.0	1.3	1.6	6d2
																																															-	0.6	0.7	0.9	1.1	7a1
																																																-	0.5	0.6	0.9	7a2
																																																	-	0.6	1.0	7a3
																																																		-	0.5	7b1

6.4 Atypicality

The clusters at all levels of the hierarchy are defined by their cluster centres. However, the cluster centres do not reflect every area within each cluster to the same degree. As discussed previously not all areas in each cluster are equally spaced from the centre. While some areas in a cluster are very close to the centre of a cluster, others are so far away they would sit equally as well in one of the other clusters. They are atypical of the cluster they are in, but still a member of it. These areas on the periphery of clusters could be seen as a failure of method, technique or process, but this would be an unjust given the complexities of the real world that are being clustered together. Whatever is being clustered something has to be at the periphery or the edge of each cluster this is not a failure of method, just a fact of life.

Are atypical areas a bad thing? Atypicality is acceptable, but only if there is a good reason for it. Some areas are very different from all others and therefore are bound to appear towards the periphery of a cluster. If it is these areas which are atypical or outliers, providing the data and their location can explain why they appear as they do, there are no problems. However, if the outliers do not seem out of the ordinary and the reason for them being peripheral to their cluster cannot be easily established then this is likely to be more indicative of some form of problem within the clustering process.

What can be said about the atypical areas in the OA classification? A sample area was selected in which atypical areas were identified and an explanation of their atypicality attempted. The Leeds area was selected because of the local knowledge of the author which, aided in the process of identifying reasons behind the presence of atypical areas. For the purposes of this investigation OAs with a squared Euclidian distance of 1.5 or greater from their cluster centre were selected as being atypical. This is a value just over double the average distance to cluster centre, there are 752 OAs that are a distance of 1.5 or more from cluster centre. This provides a large enough sample for analysis, but also small enough to examine them in detail. There are 18 such OAs in the area covered by the Leeds metropolitan district.

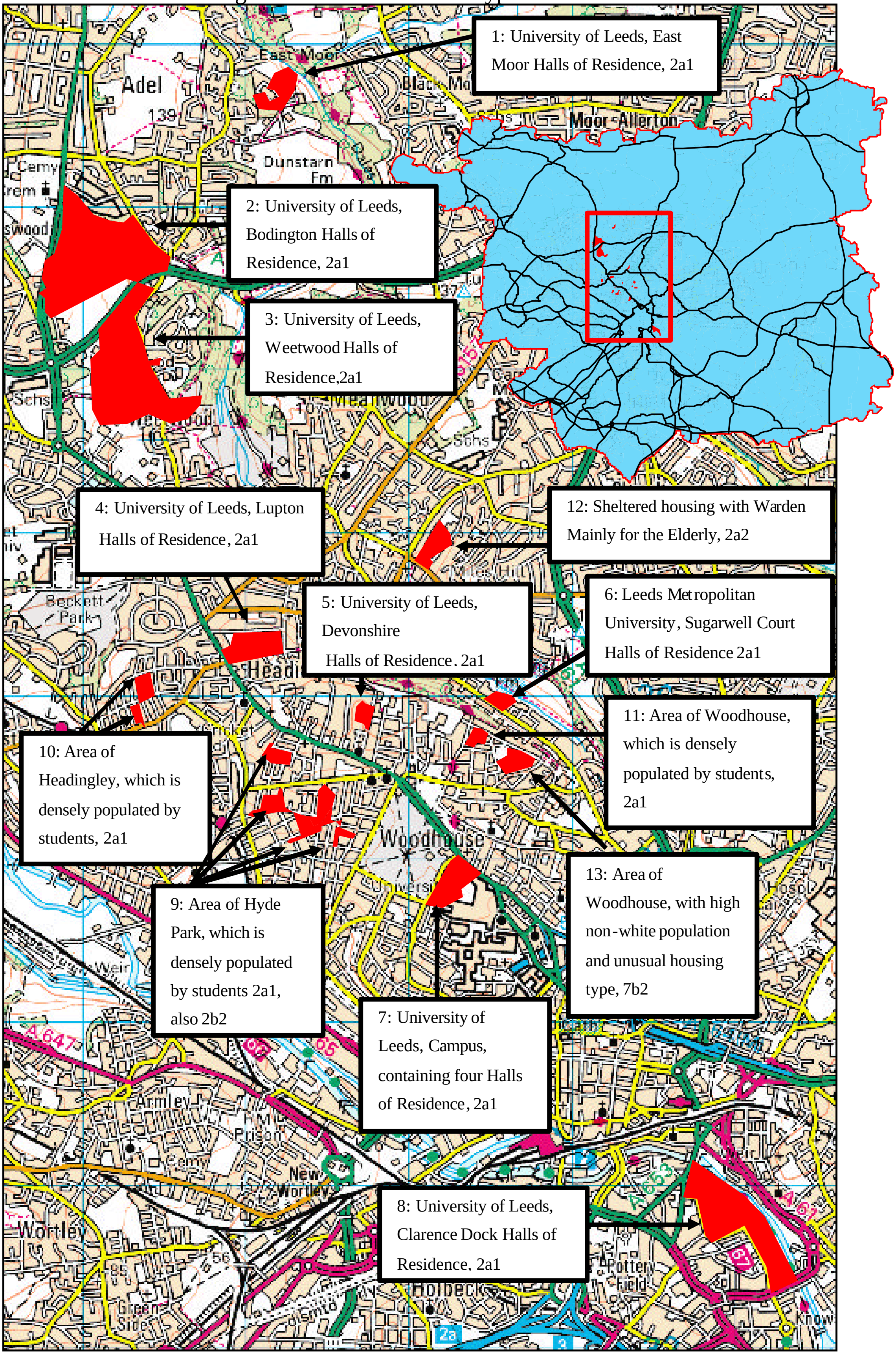
Can the reason for the atypicality of these OAs in Leeds be explained? Figure 6.6 shows the location of and the reasons behind the atypical areas of Leeds; some 16 of the 18 OAs are areas which show extremely large proportions of students (boxes 1-11). Many are halls of residence for either Leeds University or Leeds Metropolitan University. Other OAs include areas of Headingley and Hyde Park that contain streets of almost 100% student housing. These densely populated student areas are contained within Super-group 2 (City Living) which is characterised by having the highest proportion of students of any group. However, such a dense concentration of students has made these areas atypical of their cluster because they have such a high concentration of one section of society. These areas with a high density of student population are not uncommon in the larger university cities of the UK. However, they are not common enough to form a more compact cluster based on the distinctive variable that they display.

This explains the majority of the outliers in Leeds but there are still two left. One of these (box 12) is a complex of sheltered housing for the elderly. Although the elderly represent a completely different section of society to the students the reason for the appearance of this OA as atypical is the same as the students. It is the concentration of such a large number of elderly people in a small area in virtually identical forms of housing.

This leaves just 1 of the 18 atypical areas in Leeds (box 13) to be accounted for. The reasons for the atypicality of this OA are not quite as obvious as for the others. There is no single obvious reason as to why the OA is atypical unlike the others it does not contain a communal establishment or have a concentration of a certain type of population. The atypicality of this OAs is produced by several factors: firstly, it has an unusual housing type for its social composition because the area has gone through some redevelopment in recent years so where the expected housing type is a mixture of terraced housing and flats, there is actually a mixture of terraced, semi-detached and detached housing and an absence of flats. A consequence of this is that there are then slightly more young people than expected as the large homes attract families rather than single people. The ethnic make-up of the area puts the OA firmly in the Multicultural group however, the development of housing that has taken place in that area makes it an atypical residence for the Multicultural Super-group.

The investigation of atypical areas in Leeds has shown that they can be understood by examining their location, and social and environmental make-up. The areas are atypical because they exhibit extreme cases of social phenomena and therefore have a greater distance from their cluster centre.

Figure 6.6: The location of atypical areas in Leeds



1: University of Leeds, East Moor Halls of Residence, 2a1

2: University of Leeds, Bodington Halls of Residence, 2a1

3: University of Leeds, Weetwood Halls of Residence,2a1

4: University of Leeds, Lupton Halls of Residence, 2a1

5: University of Leeds, Devonshire Halls of Residence. 2a1

6: Leeds Metropolitan University, Sugarwell Court Halls of Residence 2a1

7: University of Leeds, Campus, containing four Halls of Residence, 2a1

8: University of Leeds, Clarence Dock Halls of Residence, 2a1

9: Area of Hyde Park, which is densely populated by students 2a1, also 2b2

10: Area of Headingley, which is densely populated by students, 2a1

11: Area of Woodhouse, which is densely populated by students, 2a1

12: Sheltered housing with Warden Mainly for the Elderly, 2a2

13: Area of Woodhouse, with high non-white population and unusual housing type, 7b2

The detailed examination of atypical areas in Leeds has shown that they can be understood as extreme cases of social phenomena and not as a breakdown in technique or method. However, across the UK there are 752 output areas which at sub-group level are 1.5 or more from their cluster centre. Where are they located and which groups are they in? Is atypicality more prevalent in certain parts of the UK or are you likely to find extreme cases in all regions? Do some clusters exhibit more cases of atypicality than others or are all clusters (sections of society) as likely to exhibit extreme values as each other? These are questions which can be answered by examining the distribution of these atypical areas by location and cluster, as shown in Table 6.6.

Table 6.5 shows that almost half of the outliers in the UK are in Scotland with 1 in every 116 OAs in Scotland being an outlier. London has the smallest percentage of outliers with only 1 in every 1,250 OAs in the capital defined as an outlier. This was unexpected, outliers are caused by extremes, think of extremes in the UK and you think of London.

Table 6.5 Number of Outliers by Government Office Region or Country
(defined as Squared Euclidian distance of 1.5 or more away from cluster centre)

Government Office Region (G) or Country(C)	Outliers	Percentage of Outliers	No. of OAs in Area	Average number of People per OA	Percentage of OAs that are outliers	Average distance from Super-group centre	Average distance from Group centre	Average distance from Sub-group centre
Scotland (C)	365	48.5%	42,604	118.8	0.86%	.96	.89	.84
South East (G)	82	10.9%	26,645	300.3	0.31%	.78	.71	.67
South West (G)	52	6.9%	17,016	289.6	0.31%	.77	.71	.67
Yorkshire & Humber (G)	50	6.6%	16,792	295.7	0.30%	.79	.72	.67
East of England (G)	46	6.1%	18,200	296.1	0.25%	.77	.71	.67
East Midlands (G)	31	4.1%	14,105	295.8	0.22%	.78	.71	.67
North West (C)	41	5.4%	22,710	296.3	0.18%	.78	.70	.66
Northern Ireland (G)	9	1.2%	5,022	335.6	0.18%	.79	.73	.69
West Midlands (G)	29	3.8%	17,458	301.7	0.17%	.78	.71	.67
Wales (C)	15	2.0%	9,769	297.2	0.15%	.75	.69	.65
North East (G)	12	1.6%	8,599	292.1	0.14%	.78	.71	.67
London (G)	20	2.7%	24,140	297.1	0.08%	.83	.76	.69
UK Total	752	100.0%	223,060	263.6	0.34%	.82	.75	.70

So why are there so few outliers in London and so many in Scotland? The first and most obvious reason lies in the different design of output areas in Scotland compared to the rest of the UK (outlined in §5.2.1). The OAs in Scotland are less than half the size of those in the rest of the UK in terms of their population (see Table 6.5 for a comparison of figures). Therefore, the possibility of a Scottish OA having an extreme value for a variable is more likely than the other three countries of the UK. To produce a result where 100% of people in an OA report the same answer to a specific question may require only 50 people to report the same thing in Scotland but over 100 in all other parts of the UK. To look at it in another way imagine two streets containing 100 people , one in Glasgow and one in London. The street in Glasgow will form two OAs each with 100% student population. The street in London will only form one OA but with the same 100% student population. All three OAs are likely to be outliers because of the high concentration of students that they display. However, in Glasgow two outliers have been formed while in London there has only been one. If the OAs were made half the size again the likelihood is that there would be even more outliers. The simple fact is that the smaller the areal unit that is being used the more extreme the values that will be observed.

We can therefore understand why Scotland has got so many outliers, but why has London got so few? It is a city of extremes. That London is a city of extremes explains why there are so few outliers. When there are so many extreme values, but all based on the same variables these values no longer become extreme. They form their own cluster. The centre of London is dominated by two Super-groups *City Living* and *Multicultural.* These two clusters are very much powered by those extreme London values. The ‘Multicultural’ nature of central London is repeated to a much lesser extent in other English cities. However, even in the largest cities of Scotland such as Edinburgh and Glasgow, this feature is barely visible.

So we know where the outliers are, but which clusters are they in? It may not be the first thing that is apparent by looking at Table 6.6, but it is worth pointing out that there are very few outliers in the *Multicultural* Super-group. There are few examples of this super-group in Scotland, where there are a lot of outliers. Is there a link here? It would not be true to say that one has caused the other, but rather that by elimination one thing makes the other more likely.

Table 6.6: Number of Outliers by Super-group
(defined as Squared Euclidian distance of 1.5 or more away from cluster centre)

Super-group	Outliers	Percentage of Outliers	No. of OAs in Type	Average number of People per OA	Percentage of OAs that are outliers	Average distance from cluster centre
2: City Living	335	44.6%	16,637	216.6	2.01%	.98
5: Constrained by Circumstances	175	23.3%	33,165	193.4	0.53%	.89
6: Typical Traits	84	11.2%	40,769	278.2	0.21%	.79
4: Prospering Suburbs	67	8.9%	47,250	286.9	0.14%	.79
1: Blue Collar Communities	42	5.6%	35,837	274.1	0.12%	.77
3: Countryside	33	4.4%	27,681	264.5	0.12%	.76
7: Multicultural	16	2.1%	21,721	310.1	0.07%	.82
UK Total	752	100.0%	223,060	263.6	0.34%	.82

City Living and *Constrained by Circumstances* account for two thirds of all outliers. This is perhaps not surprising as Table 6.4 shows these are the two least compact clusters. Also these two super groups have the fewest number of people per OA, as discussed earlier with the case of Scotland this will increase the number of outliers. From the evidence of the investigation of atypicality in Leeds it is likely that many of the *City Living* outliers are caused by large concentrations of students. Over 70% of the *Constrained by Circumstances* and 45% of the *City Living* outliers are in Scotland, representing 75% of the Scottish outliers. This was calculated by running a cross-tabulation on the two previous tables to create Table 6.7. Several patterns can be seen for example, 16 of London's 20 outliers are *'City Living'* and half of the *'Multicultural*' outliers are in Scotland.

Table 6.7 Cross-tabulation of Outliers by Super-group and Government Office Region/Country (defined as Squared Euclidian distance of 1.5 or more away from cluster centre)

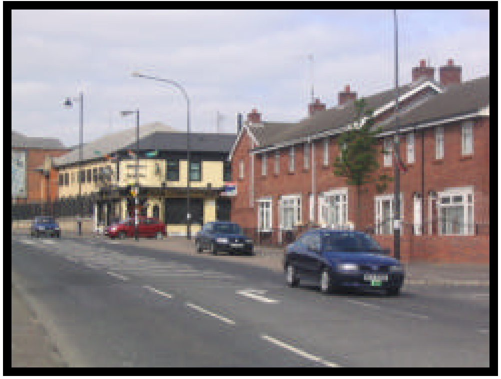
	Government Office Region (G) or Country(C)												
Super-Group	East of England (G)	East Midlands (G)	London (G)	North East (G)	Northern Ireland (C)	North West(G)	Scotland (C)	South East (G)	South West (G)	Wales (C)	West Midlands (G)	Yorkshire and the Humber (G)	Total
1:Blue Collar Communities	7	3	0	1	0	0	22	2	3	0	0	4	42
2:City Living	9	12	16	8	1	26	151	39	17	11	17	28	335
3:Countryside	7	3	0	0	0	3	12	3	4	0	0	1	33
4:Propering Suburbs	6	6	0	0	1	2	25	9	10	0	5	3	67
5:Constrained by Circumstances	4	5	3	2	0	8	126	6	4	4	4	9	175
6:Typical Traits	11	0	0	1	7	1	21	23	14	0	2	4	84
7:Multicultural	2	2	1	0	0	1	8	0	0	0	1	1	16
Total	46	31	20	12	9	41	365	82	52	15	29	50	752

The extreme number of atypical OAs that have been observed in Scotland could suggest that Scotland is atypical to the rest of the UK and should be classified on its own. However, it is likely that the majority of the reason behind the large number of outliers observed in Scotland is the difference in how OAs were defined in Scotland compared with the rest of the UK.

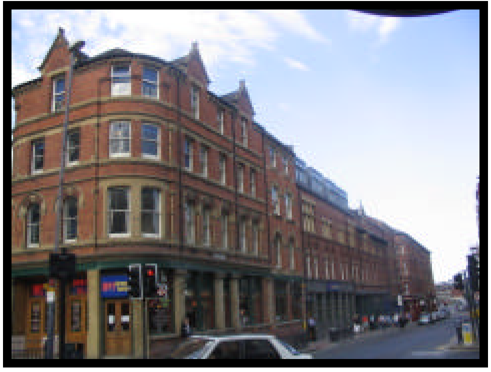
6.5 Ground Truthing

When creating a classification it is useful to build up as a clear picture as possible of what life is like in each of the different clusters. To do this there is a need to go beyond just looking at the statistics and creating short verbal profiles. A great way to add a sense of reality to a classification is to get out of the office and have a look at how the classification compares with the real world. This process is called 'ground truthing' essentially checking that what the data tells you about an area is reflected in what can be seen in that area. Ground truthing can be used as a check on the accuracy of the product, but if photographs are taken during the ground truthing they can be used as an output of the classification to aid users' understanding of each of the clusters. Ground Truthing was carried out in locations across the UK. Streets were selected from 1:10,000 Ordnance Survey map prior to carrying out the ground truthing. This was done to ensure that areas representing all clusters were visited; photos were taken at all locations. Example photos of selected areas representing the Super-group level can be seen in Figure 6.7.

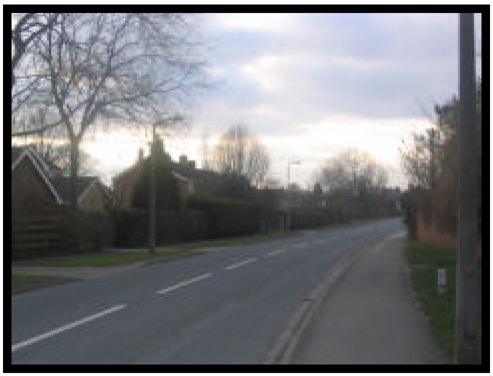
Figure 6.7: Photographs representing the classification at super-group level



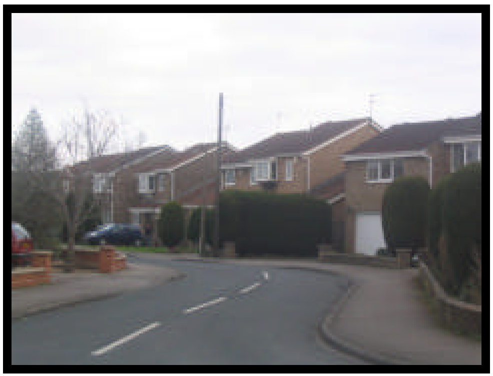
1: Blue Collar Communities, (1a1), Falls Road, Belfast, BT12 4PD



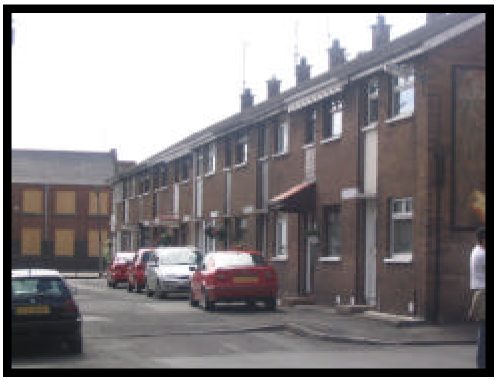
2: City Living, (2a1), New York Street Leeds LS2 7DT



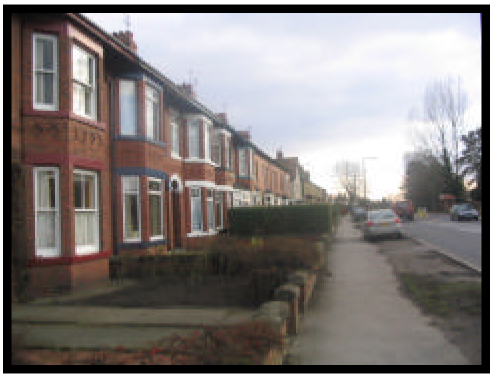
3: Countryside , (3c1) Brayton Lane, Brayton, Selby, YO8 9DZ



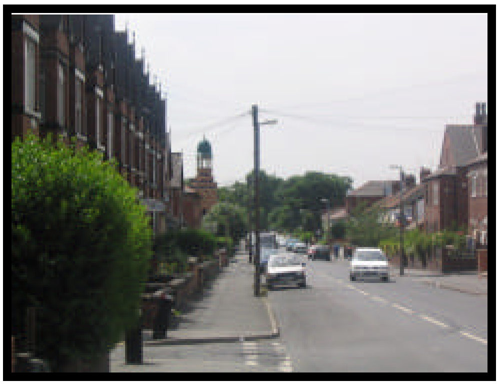
4: Prospering Suburbs, (4B2) Baffam Gardens, Selby, YO8 9AY



5: Constrained by Circumstance (5b1) Bombay Street, Belfast.



6: Typical Traits, (6a1) Doncaster Road, Selby, YO8 9BU



7: Multicultural (7a3) Brudenell Road, Leeds LS6 1LS

Interesting results were experienced when a ground truthing exercise was run using first year undergraduate students on a field trip to Bangor north Wales. The students were sent out with a map of the classification and asked to make an assessment as to how well the classification represented what they saw on the ground. All, but one group returned to the field centre stating that they thought that the classification was a more than adequate representation of what they had seen on the ground. However, one group said that the classification was nothing like what they had seen whilst out groundtruthing. When they were asked where they had been and this was compared to the map they were given it was discovered that the error lay in the map reading ability of the students and not the classification. The students had visited a completely different part of the town to the area on their classification map. The classification had not only been verified by the majority of the students, but was also able to identify those students who had gone wrong and enabled them to understand the reasons why (Rees 2004).

6.6 Fuzzyfying the Classification: Distance to All Cluster Centres

Although each OA is a member of one cluster at each level, reality is never as black and white as this. As discussed in chapters 2 and 3 many practitioners advocate using a fuzzy classification system. It is hard to envisage fuzzy classifications replacing traditional systems for one reason, the need for simplicity. The reason for the continued and increasing popularity of geodemographic systems is that they are simple to understand, use and interpret. For the average user a fuzzy classification contains more information that they could possibly need. However, a high end user or some body investigating only a small area may find value in the additional information. Giving each area a relative membership of all clusters rather than a binary membership of one would give a great deal more detail about the classification.

The OA classification can be turned into an *ad hoc* fuzzy classification by measuring the distance from each OA to all of the cluster centres. This enables the examination of troublesome OAs that appear on the periphery of a cluster (like those examined in § 6.4). It can be investigated if these OAs fall on the edge of more than one cluster or whether they are outliers at the whole UK level. Extension or reduction of each cluster to the *n* nearest to any of the cluster centres, for example, identification of the 50,000 OAs nearest to the City Living cluster centre or the 10,000 OAs furthest from the Prospering Suburbs cluster centre. Like the rest of the information about the classification, the data can be easily mapped for analysis and show some intricate and intriguing patterns. Figures 6.8 – 6.14 show maps of the distance of each OA to each cluster centre. Maps cover the whole of the UK and the Leeds Metropolitan District.

Figure 6.8: Distance from super-group1 Blue Collar Communities (Top map depicts whole

Figure 6.9: Distance from super-group 2 City Living (Top map depicts whole UK, bottom

UK, bottom map depicts Leeds Metropolitan District)

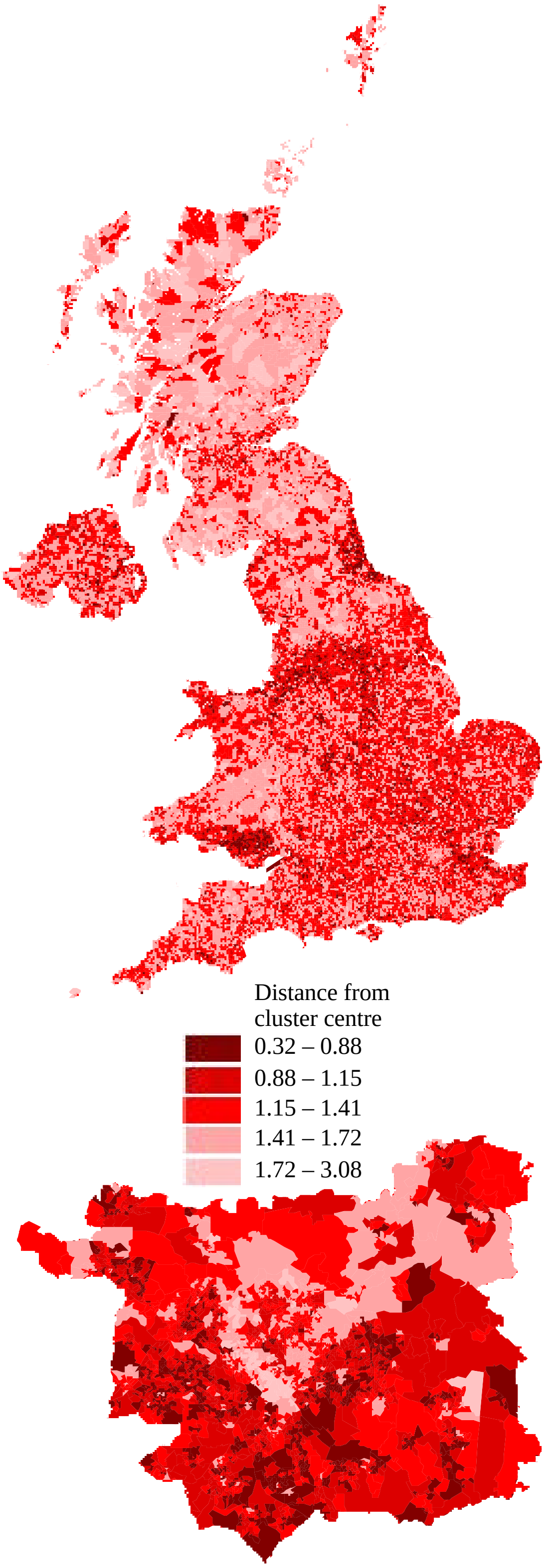


Figure 6.10: Distance from super-group 3

map depicts Leeds Metropolitan District)

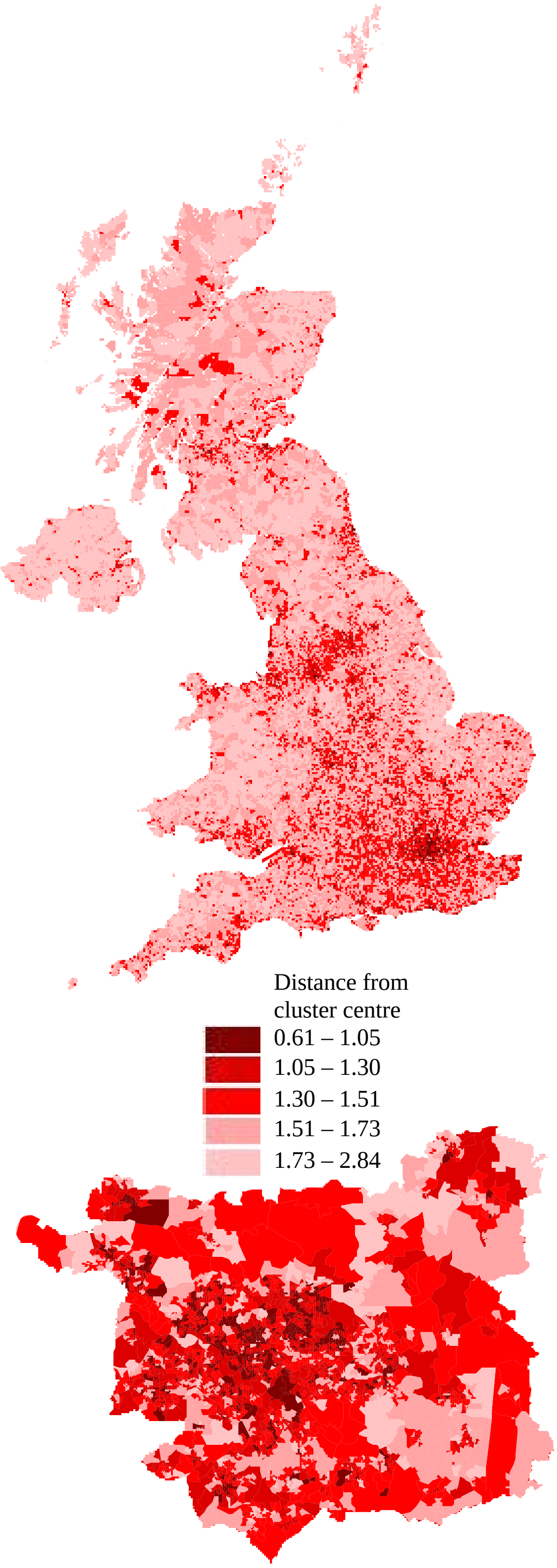


Figure 6.11: Distance from super-group 4

Countryside (Top map depicts whole UK, bottom map depicts Leeds Metropolitan District)

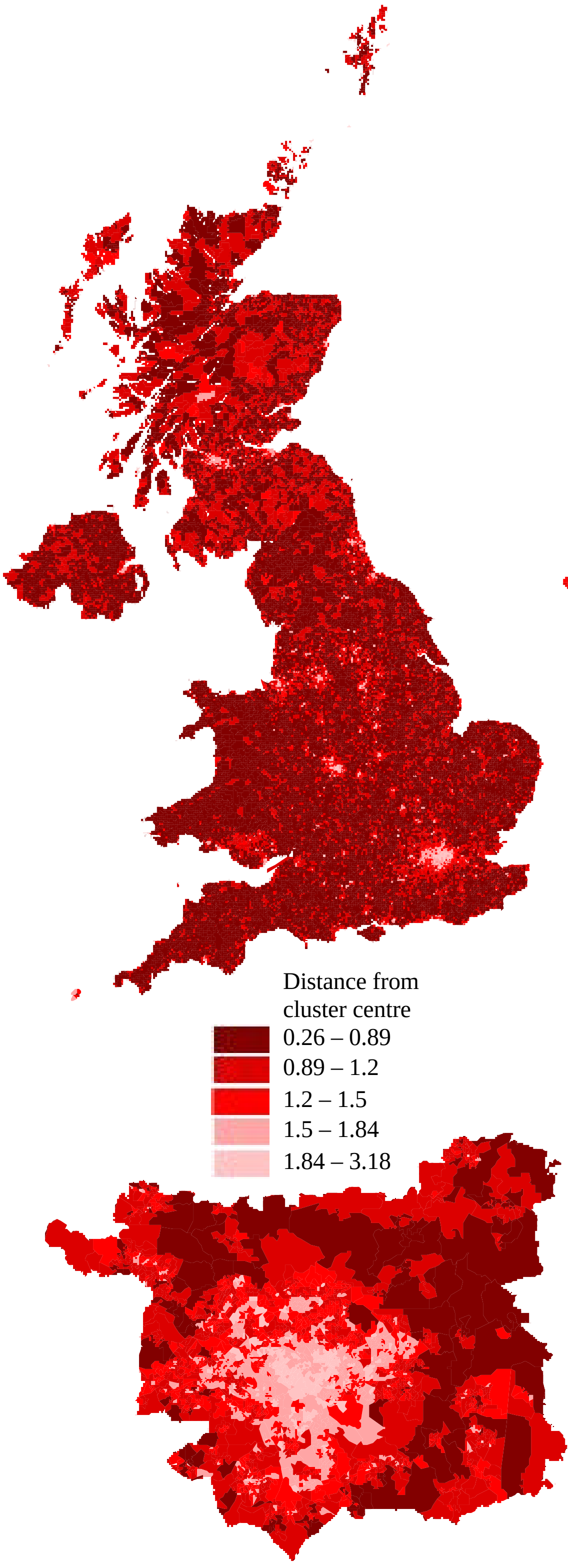


Figure 6.12: Distance from super-group 5

Prospering Suburbs (Top map depicts whole UK, bottom map depicts Leeds Metropolitan District)

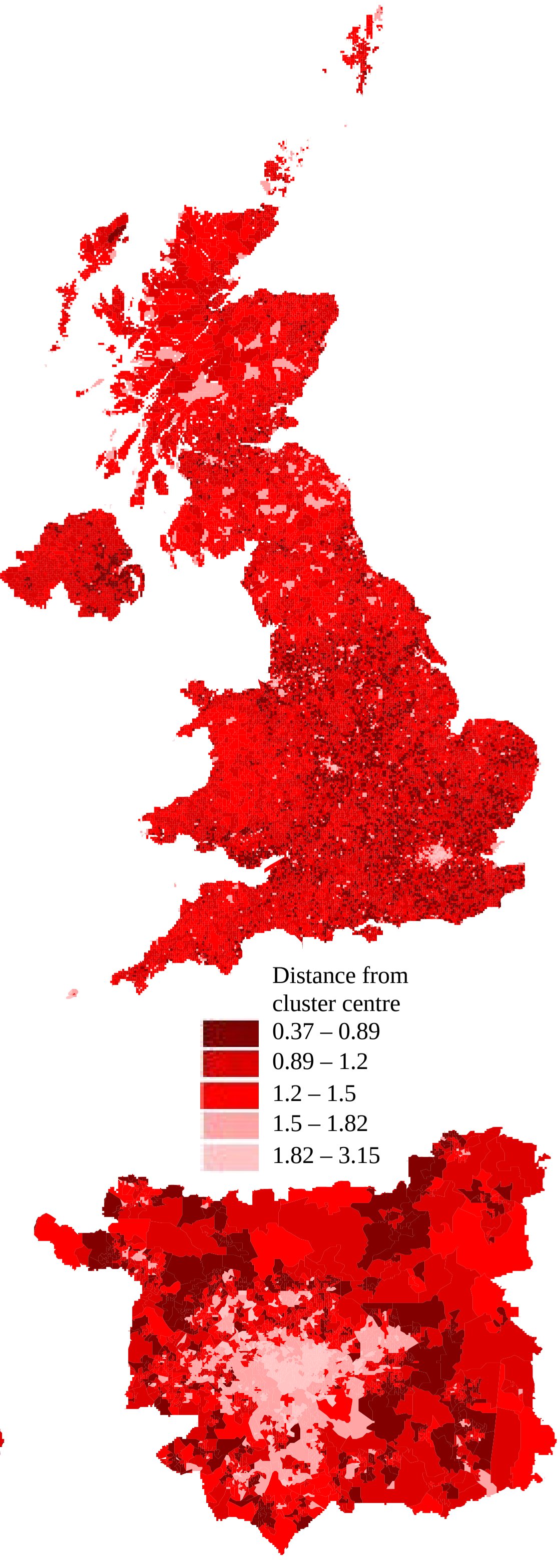
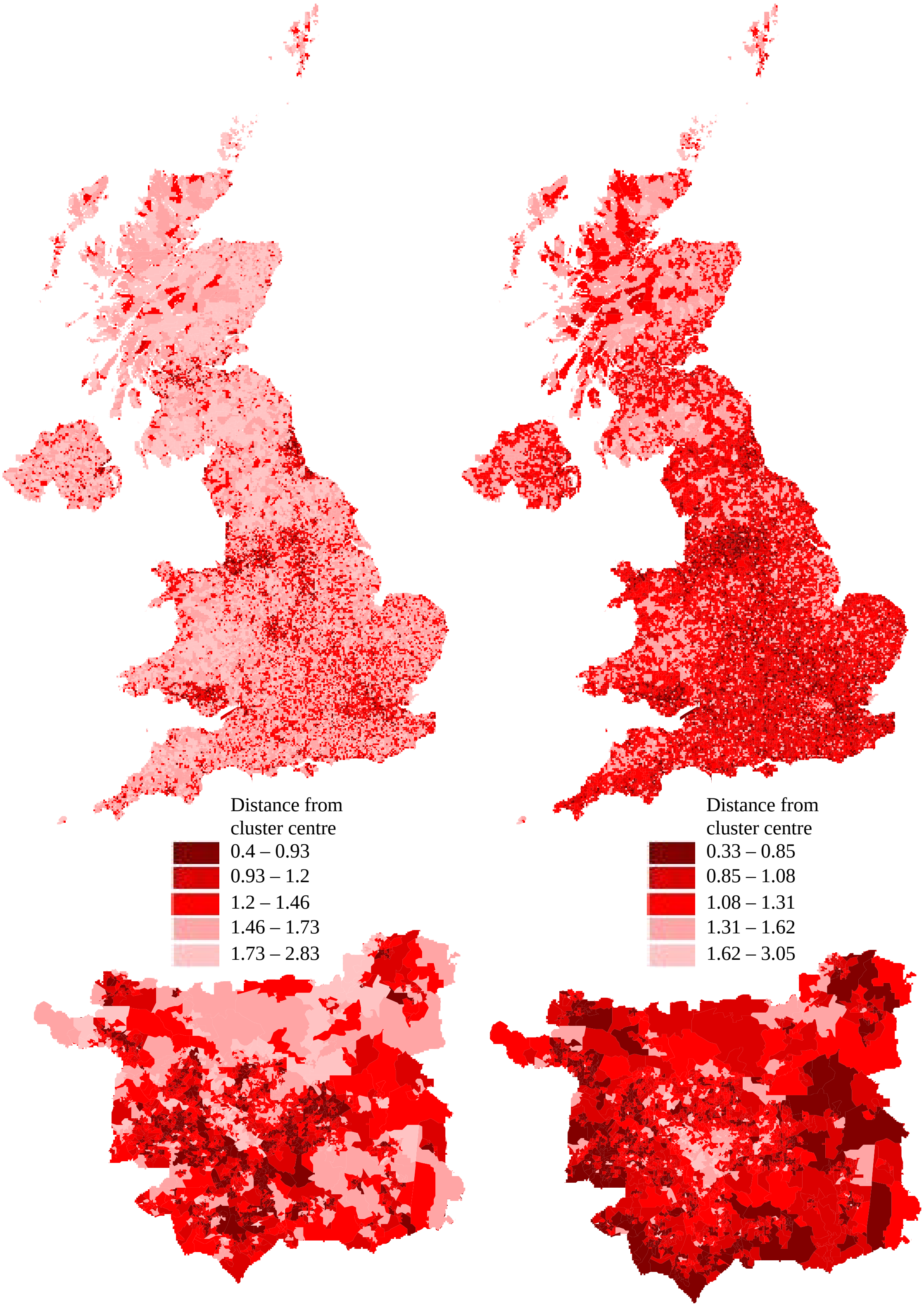


Figure 6.13: Distance from super-group 6



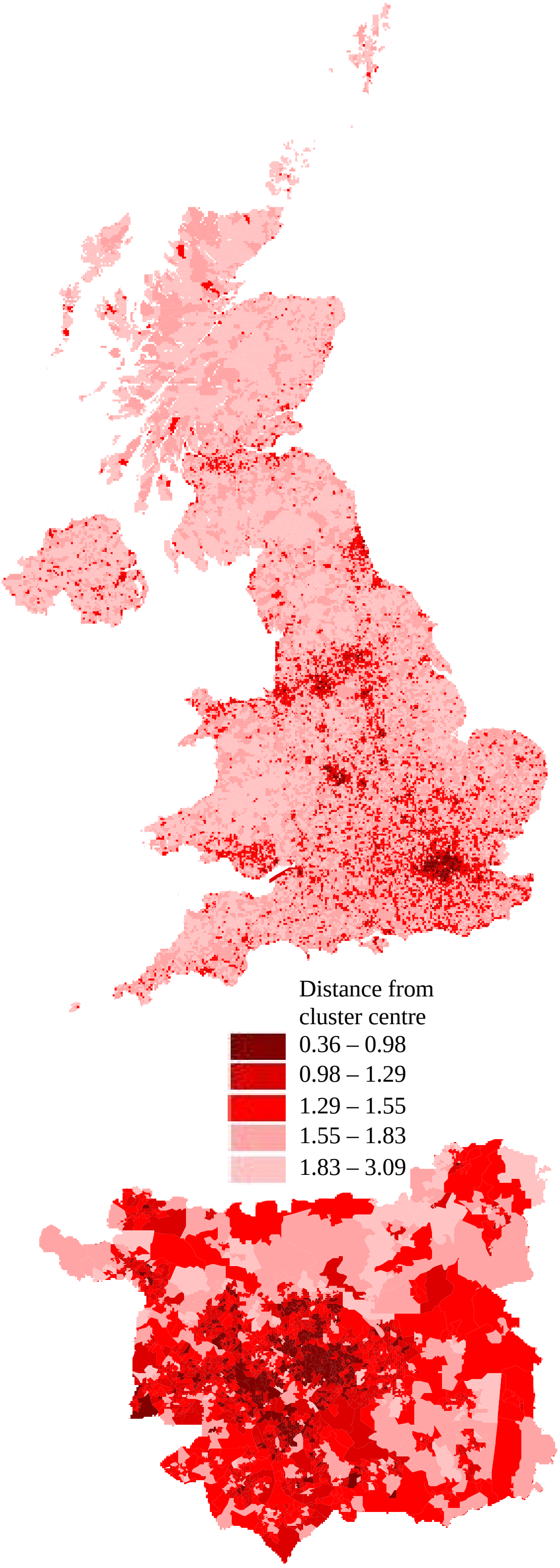
Constrained by Circumstances (Top map depicts whole UK, bottom map depicts Leeds Metropolitan District)

Typical Traits (Top map depicts whole UK, bottom map depicts Leeds Metropolitan District)

Figure 6.14: Distance from super-group 7

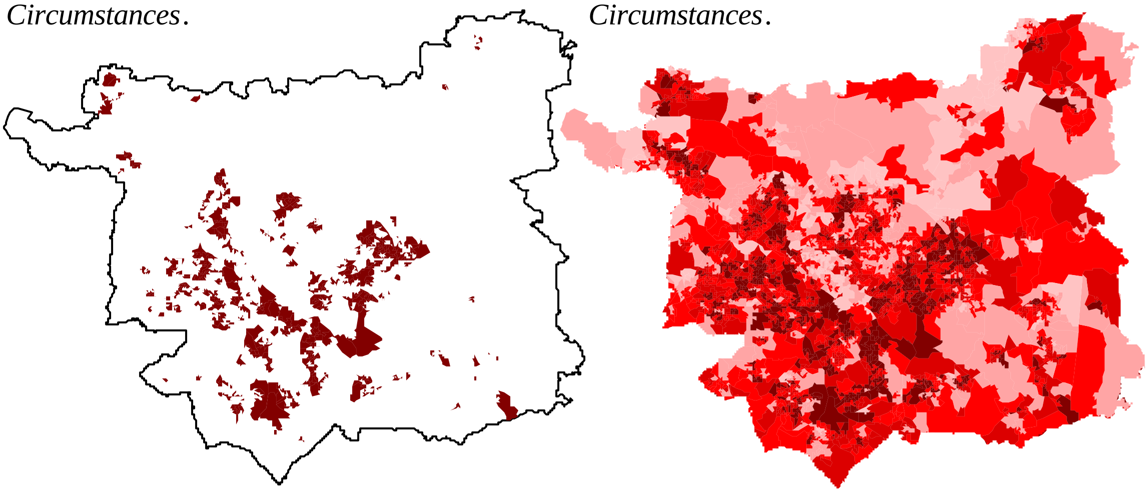
Each of the maps of the distances to each cluster centre (in Figures 6.8 - 6.14) shows a distinctive and

Multicultural (Top map depicts whole UK, bottom map depicts Leeds Metropolitan District)



The fuzzy classification shows that by mapping the distances from cluster centre for any one of the super groups, it is possible to see clear geographic differences between the OAs which are closest to the centre of a cluster and those that are further away. But how much more information does the fuzzy version of the classification give above and beyond the single membership classification. Figure 6.15 compares the fuzzy classification against the single membership classification for super-group 5 Constrained by Circumstances.

Figure 6.15: A comparison of the *ad hoc* fuzzy classification with actual membership of the cluster for super-group 5 *Constrained by Circumstances*, for the Leeds Metropolitan District.



Membership of super-group 5 *Constrained by Circumstances*.

Distance from super-group 5 *Constrained by Circumstances*.

The two maps in Figure 6.15 shows that the OAs that are a member of super-group 5 can be picked out very clearly on the map of membership, but this is the only information that can be obtained from the map. In contrast the map of distances from the centre of super-group 5 makes it a little more difficult to make a distinction as to which OAs are in that cluster. However, the distance from cluster centre map contains a wealth of extra information suggesting areas that are somewhat like those in super-group 5 and areas that are not at all like *Constrained by Circumstances*.

It is clear that the fuzzyfication of the classification adds a great deal of valuable information that gives increased knowledge about the areas that it covers. However, the two versions of the classification should not be seen as rival forms of classification that need to be chosen between. The two classifications complement each other and the greatest value will be found when they are used in conjunction with each other. There will be instances when one form of the classification is better than the other and then a choice can be made, but it would be unwise to suggest that one or other of the forms of the classification is better than the other for all instances. The form of the classification that is chosen should be dependent upon use.

6.7 The Consultation Exercise

One of the biggest problems faced by someone who creates a classification is the age old question, how do I know it is right? The simple answer is you don't know if it is right or not. You certainly can't tell from looking at the numbers. Careful validation of the classification must be done by examining maps of places which the creators of the classification are familiar with (Harris et al. 2005). This is fine to check the classification for a small number of places, but unless the creator of the classification has lived in most cities of the UK they maybe could do with a bit of help.

At the time of classification there was little administrative data at the OA scale to compare the classification against. This left a problem as to how to assess the quality of the classification. A methodology was formulated to carry out a consultation exercise, where people with suitable demographic experience (academics, local government officers, central government officers) would be asked to take part to validate the classification against their knowledge of an area. The consultation exercise was devised to take advantage of the knowledge of places of people all round the country. By doing this not only does the number of places that can be checked increase significantly, but it is also a very independent way of assuring the quality of the classification. A consultation of peers is something that is under used in the academic community. The views and opinions of others can be of great value and can throw up ideas or issues which may have otherwise gone undiscovered. An example of a consultation exercise which gained significant knowledge from people's opinions on how to move a project forward is the Rees (1998) consultation as to what Census users wanted to see from the then forthcoming 2001 Census. The paper outlines the views of some 140 respondents, who expressed views on such things as, the religion question, the one number census methodology, postcode-based outputs and categories of ethnicity (Rees 1998). This form of qualititative review of quantitative research is something that is under used in academic research and unprecedented with respect to area classification.

6.7.1 Method of Consultation

The target population at which the consultation was aimed was to be academics, local government officers and other professionals with a background in population and mapping of areal data, so as to ensure that they had enough knowledge to complete the exercise easily and independently. The first step that needed to be taken was to sign up some people to take part in the consultation. This was done by making use of a list of colleagues and contacts that were considered suitable for the task. The people chosen had all had involvement with the ESRC/JISC Census Programme or known by the author in some other capacity and deemed to be suitable. The participants were contacted by e-mail, in which the project and the consultation exercise was explained and a request was made for the postcode(s) of a place(s) they knew well (e.g. their home address, parents home, work place etc.).

A map was created centred on the postcode (that had been specified by each participant in the consultation) covering the surrounding town, city or countryside. The maps consisted of two layers: an Ordnance Survey 1:50,000 feature map and the OA classification represented as shaded colours without containing boundaries so that the OS background is clearly visible. No key to the shaded class was provided. Each respondent had to link the map colours to the list of classes. Each map was then sent to the relevant person, along with information about the make-up of the clusters at the super-group level, including the names of the groups, pen portrait descriptions and graphs giving values for selected variables (Age, Ethnicity, Population Density, Employment Status, Car Ownership, Tenure, Housing Type) for each super-group. The consultation was run at the super-group level, to keep the task as simple as possible for the participants (matching 7 groups being easier than 21 or 52). Mapping at the lower levels also causes a problem as there are not enough easily distinguishable colours. The e-mail contacts with the participants and the documents they were sent can be seen in Appendix C.

The participants were asked to match the colours shown on the map to the super-group names based on their knowledge of the demographics of the area, compared with the demographic information given with the names of the super-groups. Note, that the names used in the consultation exercise were not the list of final names as they appear in chapter 5. Changes, some of which were suggested by the consultation took place after the completion of the consultation exercise . They were also asked for additional specific comments and questions pertaining to the quality of the classification, the good and bad points, how well they thought it represented their area and the values and uses of the classification. A typical map is shown in Figure 6.16; the red triangle marks the postcode given by the participant.

It was important to get a good spread of locations across the UK. Representation was needed in each of the nine Government Office Regions in England plus, Wales, Scotland and Northern Ireland. People were contacted across the UK and the response was good though some areas are better represented than others in the results. Figure 6.17 shows the location of the people who took part in the consultation exercise. The respondents are marked by red dots.

Figure 6.16: Example map used in the consultation

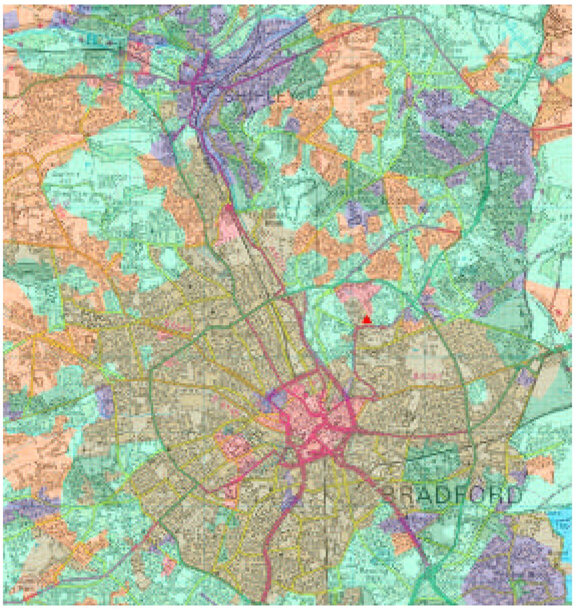
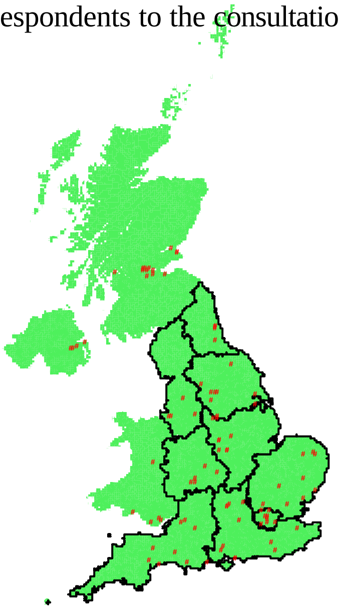


Figure 6.17: The location of the maps looked at by the respondents to the consultation



6.7.2 Consultation Results

The consultation produced two different types of information: firstly, the statistical data for judging the success of matching the super-group names to super-group type represented by colours on the map, and secondly, significant qualitative data in the form of comments and suggestions made about the classification.

The names and numbers used to represent the super-groups in the consultation exercise are not the same as those used in the rest of the document. Changes were made after the consultation was run (several as a result of comments made in the consultation. see § 6.7.9). The differences between the two sets of names are outlined in Table 6.8.

Table 6.8: The differences between the final set of names and those used in the consultation

Consultation Names	Final Names
1: City Centre Melting Pot	2: City Living
2: Typical Traits	6: Typical traits
3: Inner City Multicultural Blend	7: Multicultural
4: Blue Collar Communities	1: Blue Collar Communities
5: Idyllic Countryside	3: Countryside
6: Constraints of Circumstance	5: Constrained by Circumstances
7: Comfortable Suburban Estates	4: Prospering Suburbs

6.7.3 Statistical Results

Table 6.9 outlines the percentage of respondents who successfully matched the colours on the map to the clusters. Each map contained upto seven colours (and as few as two depending on the area chosen by the respondent) representing the different Super-group types of the classification. Each of the colours could be matched to any of the seven names of the super-groups according to the respondent's opinion about which colour represented which group. Therefore on a map which contained all seven colours, there are 5,040 (7×6×5×4×3×2×1 = 5,040) possible permutation of answers.

Table 6.9: Percentage of respondents who identified the super groups correctly

Super-group	1: City Centre Melting Pot	2: Typical Traits	3: Inner City Multicultural Blend	4: Blue Collar Communities	5: Idyllic Countryside	6: Constraints of Circumstance	7: Comfortable Suburban Estates	Total
Responses	51	55	38	33	54	47	58	336
Correct	43	41	33	25	54	34	54	286
% Correct	84%	75%	87%	76%	100%	72%	93%	85%

In total 61 people returned their completed questionnaire by the 4th October 2004 deadline. 85% of super groups were matched to the correct name. Matching percentages ranged from 72% for super group 4 *Blue Collar Communities* to 100% for group 5 'Idyllic Countryside'. The most common mistake that was made was confusing super group 4 *Blue Collar Communities* with super group 6 'Constraints of Circumstance'. Super group 2 Typical Traits was also sometimes mistaken for super group 4 *Blue Collar Communities*. There are several reasons for this: firstly, the two clusters are not as dissimilar as other pairs of clusters. From comments received it is clear that people's opinion of an area is formed by very few and only visible characteristics, primarily housing type. Super groups 2, 4 and 6 are hard to distinguish by housing type as they are broadly similar; these groups have been formed predominantly by the values of other less visual variable variables such as illness and tenure.

Another reason connects to the places where the people who took part live. The maps they were given were usually centred on their home location. Only a quick glance is needed at Table 6.10 showing which super-group their home location is in, to see that nobody lives in super group 4 and only one person lives in super group 6. There is poor first hand knowledge of super groups 4 and 6 and good first hand knowledge of super groups 5 and 7, where two thirds of respondents live. There are few mistakes for super groups 5 and 7 with 108 of 112 matched correctly.

Table 6.10: The super groups in which respondents live

Super-group	1: City Centre Melting Pot	2: Typical Traits	3: Inner City Multicultural Blend	4: Blue Collar Communities	5: Idyllic Countryside	6: Constraints of Circumstance	7: Comfortable Suburban Estates	Total
Percentage of Respondents who gave a location in	18%	13%	4%	0%	19%	1%	45%	100%

It would have been good to have surveyed a more even socioeconomic spread of people in order to achieve a representative sample. However, what has actually happened here is that the classification has worked on a test that it has been entered into completely by accident. The people who took part in the consultation were of approximately the same standing either academics or people who worked as planners or demographers in local government. It would be hoped then that if you classified this unsuspecting group of people that many of them would come out in the same super-group and that is exactly what has happened. Some 45% of people fall into the top super-group, 64% of people fall into the top two groups, 82% of people fall in the top three groups and 95% of people fall in the top four super-groups with only 5% of people living in the other three super-groups. The classification has successfully distinguished the characteristics of the super-group of people who responded to the consultation exercise.

6.7.4 Feedback on the Classification

Several themes were identified from the feedback on the classification. They demonstrate the value of local knowledge when looking at geographic information. The feedback will be summarised under theme headings where the views and comments of the respondents will be drawn out with the use of quotations (*in italics*) and discussion of the relevance of the comments and the formation of action points or necessary changes that come out of each theme. Respondents are not identified by name as confidentiality of the consultation was guaranteed. They are given reference codes so that their quoted responses can be tracked back to the right survey form.

6.7.5 General Comments

In general the response to both the classification exercise was very positive *"overall- it's great – it makes sense"* (Respondent qa13). The majority of the respondents not only were very encouraging about the classification but were impressed with the *"innovative and impressive consultation exercise"* (Respondent qa5) They seemed genuinely happy and willing to have been asked to take part in the exercise: a respondent said *"I enjoyed the exercise (hope it corresponds to your classification!) and look forward to seeing the finished product with groups and sub-groups"* (Respondent qa57). Even a sense of competition was built up between some people when undertaking the exercise *"It would be nice to know how I did, not least because I have placed 3 bets on the uncertain ones with my partner"* (Respondent qa1). This can only be seen as not only a reflection on the quality of the work, but also its great relevance and the interest in the subsequent publication of the classification.

6.7.6 Teaching Resource

The educational values of the classification were noted upon by several people who are involved in teaching geography in the university sector *"There's probably a lecture example on geodemographics and lifestyle classifiers in there somewhere..."* (Respondent qa38). It was mentioned that it would make a good case study of residential patterns. *"I wonder if it would be possible to use this technique as a teaching tool, once the classifications are finalised? I am sure it would make a great teaching resource for CHCC* [Collection of Historical and Contemporary Census Data and Related Materials] *if you could automate the map production once a postcode is entered?"* (Respondent qa1). The respondent was excited about the prospect of an automated way of students specifying a map of an area by entering a postcode. It was mentioned that this could be done on the existing CHCC website that is already widely used as a teaching resource.

6.7.7 Life Course and Change Over Time

One of the respondents who had asked for several maps of different places they had lived at different times in their life commented on how the classification worked as a narrative of their life course. *"An interesting exercise, which tells my own life history - I grew up in 'typical traits'; went off to be a student; as a postgrad I rented a room also in 'typical traits' before as a young academic buying a small terraced house in 'city centre melting pot'. After a few years I moved on to a house in 'typical traits' and a few years ago finally arrived in 'comfortable suburban estates'. You could probably classify life-histories according to transition through these profiles!"* (Respondent qa38). This shows how the respondent has used the exercise to classify their own life at different stages of the life course, exemplifying that the classification does not only have to work in a static time frame but also shows that people do not live in the same kinds of places their entire life and people can move through the classification. This leaves the door open for a great deal of investigation of people's life course by using their movement through the classification with time. This would be a fascinating and innovative research project.

There were further suggestions as to using the classification over time. While the respondent who was following their life course was tracking people through time, another respondent wanted to be able to follow areas through time by creating a classification from data from the previous census *"Can you create the same for 1991 and show us change?!"* (Respondent qa13). this shows that there is great interest in being able to follow patterns through time and if and when a classification is created for the 2011 Census comparability with the previous version should be taken into account.

6.7.8 Fuzzy Classification

There was interest in the possibility of a fuzzy form of the classification being available. *"Suggestion: I assume you have some statistic which says how sure you are of a classification for a particular OA. If you are not very sure and that OA does not border another OA of the same classification *and* it does border OAs of the 2nd mostly likely classification it fits into – I would allocate it to that (the 2nd most likely cluster it is in) – "intelligent spatial smoothing". You'd produce maps which were simpler and more likely to be true. Otherwise you need a big health warning about OA data and how marginal it can be that changing 1 number in a table moves an area from group X to group Y"* (Respondent qa13). Although the main classification will take the form of a very traditional crisp classification, due to the advantage of simplicity and preference from the industrial partner for that form of classification, there is great scope for developing a form of fuzzy classification (or at least fuzzy information) for the classification giving distances to all cluster centres as well as their own.

6.7.9 Names of Clusters

Perhaps the most expected comments to come out of the consultation exercise were the criticism of some of the names of the super-groups *"I thought some of the labels and descriptions were a bit stereotyped."* (Respondent qa19). The disappointing aspect of this was not the criticism as the consultation was intended to find any problems as they are easier to solve before publication rather than after, but the fact that nobody suggested any alternative names. The most criticised name was 'idyllic countryside' *"'Idyllic' is unnecessary & value-laden."* (Respondent qa9). Several people thought this label unrepresentative of many people's experiences of rural life. Respondent qa19 remarked *"Deep countryside isn't idyllic for everybody"* and Respondent qa49 said *"Idyllic – not sure rurality is many people's ideal even if the grass is greener."* Other names that received a negative reaction were 'City Centre Melting Pot' and 'Inner City Multicultural Blend'. *"The labels are not very distinguishable from each other, e.g. City Centre melting pot and inner city multicultural blend conjure up the same image to me."* (Respondent qa49). The following quotation shows perhaps more than any other how important the names are to the classification. *"There were some areas that would have been best described as 'student areas,' however, there was no option for this" (*Respondent qa31). The word "student" is so important to the respondent's view of the area that to not find the word student within any of the names this has made it difficult for them to link that cluster to a name. This dissatisfaction occurred even though the names given to each cluster do not affect in any way the intrinsic values on which the each cluster is based. The evidence from the consultation exercise suggests that the names are the primary source of information on which the users of the classification base their opinion of each cluster. These comments need to be taken on board as these are people who may want to use the classification. If they think that certain names are unsuitable, they should not be used.

6.7.10 Preconceptions and Idealisation of Home

It would be reasonable to expect that respondents come to the consultation exercise with preconceptions of which clusters they would expect to live in. It is the knowledge of these areas on which the preconceptions are based that is essential for the successful implementation of the consultation exercise. However, the preconceptions that came across manifested themselves more as an idealisation of home. *"Both areas on my piece of map should be group E as they are both 'Idyllic Countryside',* "(Respondent qa36). In contrast to many respondents (who showed objection to the name Idyllic Countryside) this quotation shows how this respondent's positive opinion of their own area manifests itself as wanting their area to be in the group with perhaps the most positive sounding name. In contrast the following quotation shows a respondent who appears to have been perturbed by their area being described as 'Typical' or "*average*". *"OA* [withheld] *is classified as A and should really be E This OA is just over the Bristol Suspension Bridge and is definitely not average!"* (Respondent qa39). It is intriguing that being described as average can cause offence while for others who suffer from deprived circumstances being average is a measure of safety and affluence which they can only hope to reach.

6.7.11 More Difficult than First Thought

A common theme from the consultation exercise was that it was much more difficult than the majority of respondents thought that it was going to be. Here are just a few comments exemplifying the difficulties respondents found: *"I found this exercise very difficult*" (respondent qa44) *"This was **very** hard to do"* (Respondent qa25), and *"I found this exercise a real challenge and would not be surprised if my answers are entirely wrong!"* (Respondent qa34). So why did many of the respondents find the exercise difficult to do? The earlier statistical analysis suggested that the exercise was not fundamentally a difficult task. What is clear is that respondents thought the exercise was going to be easier that it was *"More difficult than I was expecting!"* (Respondent qa14) and *"This was harder than I thought it would be*" (Respondent qa53). Several respondents gave the impression that they thought the exercise would be a fairly simple one. It would seem that a number of respondents underestimated the complexity of residential patterns.

In retrospect this result is not necessarily surprising as on the surface residential patterns look fairly simple. In the main streets seem to be composed of fairly similar housing types. From this it is then assumed that the people within the home are of a similar level of affluence and social characteristics However, it is not until someone gets deep into the data that a much more complicated pattern emerges. If the complexity of residential patterns are underestimated this would then manifest it self as respondents viewing the exercise as difficult or at least harder than they thought it was going to be. The complexity of these patterns is illustrated by how one respondent described their area to show why it was a particularly difficult one for which to completing the exercise. *"The part of East Oxford that I'm commenting on is a very complicated little area. It consists mainly of terraced Victorian houses, varying in size - from 2 up 2 down, to large 4 storey houses, often in the same street. These houses also vary in terms of gentrification – some are very run down multiple occupancy houses and others are multi-car-owning professional families"* (Respondent qa41). The respondent has described the diversity which is present within a small area. It is therefore not surprising that to place all this variety into one group is a difficult task. This level of diversity manifests itself in the data as described by one respondent. *"The values in one graph may be representative of your area, but the values in the others may be very different."* (Respondent qa54). Great diversity within an area leads to a complicated set of data values relating to it. If the diversity within an area is not recognised many of these values could be seen as almost contradictory.

What is becoming abundantly clear is that no matter how well a set of areas is clustered together diversity can be seen to exist within that cluster. This is an argument which was put forward by Voas and Williamson (2001a); they discuss how diversity remains at most scales even after the implementation of a geodemographic classification. The reason for this is summed up well by the following quotation *"Two households filing identical* [Census] *returns would from our point of view count as identical, even if one is composed of short, vegetarian, non-smokers and the other of tall carnivorous, nicotine addicts"* (Voas and Williamson 2001a p63). This point is relevant not because that there is concern about accurately identifying *"carnivorous, nicotine addicts"* but because no matter how much information is put into a classification, some things are left out. Therefore diversity will still exist within the clusters no matter what scale the classification is created at or the number of clusters that are formed.

6.7.12 OA Boundaries and Disclosure Control

There was recognition from the participants that the OA boundaries could be responsible for some of the unexpected contrasts that they saw between broadly similar areas, although nobody mentioned the problems of the MAUP (Modifiable Areal Unit Problem). *"There are some interesting boundary problems in the map which relate, I suppose, to the underlying postal geography."* (Respondent qa26). Participants singled out specific instances within their specified areas where they felt the OA boundaries had caused an unexpected divide between adjacent areas. *"Unconvinced by the sharp boundary cutting 'The Ryde' into A & E; but don't know anything about the ward/OA boundaries! The estate should be pretty homogenous."* (Respondent qa23).

The disclosure control procedures run on the data were also picked up on as a danger of working at this very small scale. *"Sometimes using OAs works – you correctly identify the tiny council estate built at the back up my suburb. Sometimes OAs don't work – you identify enclaves which are not enclaves – probably results of ONS random number generation (0's and 3's) – or you split an estate into two groups along an OA boundary where there is no real difference on that line."* (Respondent qa13). There are obviously some reservations about using statistics at such a small scale. Many of the participants were concerned that they were seeing artificial differences between areas caused by the MAUP or statistical disclosure control. However, there was a general feeling that these perceived problems did not greatly devalue the integrity of the product. In fact participants who pointed out fundamental problems of this kind were often the greatest advocates of the product asking for more information, how could they get hold of the finished product and when would it be available for use. It is good that a number of respondents have shown an understanding of some of these problems. It shows that they have good knowledge of issues involved in a classification project and suggests that they were the correct people to involve in the consultation.

6.7.13 Map Colours

Several people commented that they found it hard to distinguish the colours representing the super-groups from the underlying maps. *"Maps are very impressive, except that there is already shading on the underlying map that mixes with your overlay shading, and can make the map difficult to interpret"* (Respondent qa2). This was a hindrance to some people, especially those who later revealed they were colour blind. It was appreciated that it was unlikely that the maps could have been produced in any other way and still retain all the information on the map. No suggestions were received as to how the data could have been displayed in a more useful manner.

6.7.14 Distinguishing the Function of an Area from its Residential Pattern

A difficulty when creating a classification based on residence with 100% geographic coverage is that not all areas are residential. Some areas contain industrial or commercial premises. In more rural areas there are large areas of open countryside between areas of residence. Why is this a problem? Despite the classification being based on residential patterns, the user may assume that, for example all factories would appear in the same or similar clusters. The truth is however, that they could appear in any cluster as they will take on the value of the residences around them that fall within the same areal unit on which the classification is being built.

The following quotation shows surprise at the clustering of an industrial area with an area they consider to be fairly affluent. *"While at such a level no mapping is going to be perfect, the map provided put very different land covers together into single clusters. For example, on the map provided, large industrial areas were clustered together with comfortable suburban developments."* (respondent qa54). The presence of an industrial development in an area is not reflected in the census data and therefore has no effect on the classification. The clusters are dependent on two factors that are completely unconnected to the function of the industrial area. Firstly the creation of the OA boundaries, which again were designed around residences, secondly the properties of the residences in the same OA. Historically industrial areas were likely to be found only in less affluent areas, but with the movement of some industry to the edge of towns and cities, this is no longer such a safe assumption.

It is not always easy to look past function and focus solely on residence. To most the term city centre conjures up images of shops and commerce, but people also live in city centres (Jones *et al*. 2004). It is much harder to get a feel for the type of people who live in the city centre as the places of residence are not the main function of the area and can often be well hidden. It is impossible to get an idea of who lives in the city centre by looking at the people who are present in the city centre streets because unlike the people who may be seen in the streets of a housing estate who are likely to have some connection to that area, the people in the city centre may have no more connection to the area than going to buy a new toothbrush. This is illustrated by the following quotation *"I'm surprised that Headingley* [area of Leeds heavily populated by students] *appears to be in the same category as the city centre"* (Respondent qa44). The function of these two areas is different but both have very young and transient demographic profiles that live in smaller dwellings such as flats or converted houses. The places are different but the residents are similar. The commercial built environment and social make up, residential environment need to be separated to avoid confusion. This needs to be made clear to avoid misuse of the classification.

6.7.15 Conclusions and Action Points Resulting from the Consultation

The consultation exercise has brought out many interesting and valuable aspects of the classification and its use, the design of the OAs and the people taking part. The overall response to the consultation exercise has been very positive both in terms of the results of the matching exercise and the comments that have been received. Many of respondents expressed their delight that this work is taking place and stressed what a valuable resource it will be when it is released. The comments on the classification also contained many points on which action can be taken to improve the classification and widen its use.

The points below summarise what was brought out by the comments given:

- Even experts have only limited spatial awareness of what is around them.
- There are unrealistic expectations about what a classification can show, possibly as a result of how commercial classifications are marketed as a perfect representation for everything.
- The name given to a cluster is vitally important as users of the classification look to this first and if it does not make sense they may not look any further.
- An individual's view is sometimes based on only one or two variables, when an area is put into a group based on another variable they are flummoxed.
- The OA boundaries can have a great effect on the map view especially on the edges of towns and in the countryside.
- People find it hard to distinguish the economic function of an area e.g. city centre shopping, factories from the people who actually live in those areas e.g. shop workers do not all live in the town centre.
- The residential patterns of the UK are more complicated than many people appreciate.
- Non-conformist OAs e.g. University Halls, Prisons and other Communal Establishments have an impact.
- There were a few comments about how the classification does not pick out the very rich areas from areas of lower wealth.
- Mean figures are not adequate enough to show the value of a variable for each type. Some measure of the variance of each variable for each type needs to be used. This is more important at this scale as the values are more extreme than those of the wards and the LAs.

Action points resulting form the consultation exercise:

- Redesign the graphical profiles of the classification to show the range covered by each variable in each cluster rather than just using the mean figure which can be unrepresentative.
- Investigate the possibility of using the classification as a teaching resource and add a postcode to OA class look-up via CHCC.
- Review names given to clusters: *Idyllic Countryside* seems to cause a lot of problems and should not be used.
- Using the classification to look at people's life course was one of the most interesting suggestions that came out of the consultation exercise. The OA classification could be added to existing longitudinal studies such as the Census Longitudinal Study (England & Wales), the Scottish Longitudinal Study, the British Household Panel Study or the Millennium Cohort Survey. In each case there will be the challenge of linking historic postcodes of respondents to 2001 Census OAs.
- Change over time is also a very interesting suggestion; investigate the possibility of using 1991 data to create a comparison for an earlier period.
- Investigate adaptation of classification to have a fuzzy component at super-group level to start with.
- Expand explanations of the creation of the OA boundaries and the difference between residential and economic functions of areas.
- Clearly state that the classification represents residence and not economic function.

6.8 Conclusions

The investigations in this chapter have shown that the quality of the classification is in no doubt. Although sensitive to the removal of variables to varying degrees, the classification is robust and each variable adds valuable information to the system. The system has been shown to reduce the overall variability to a greater extent than previous comparable systems. The within cluster variability reduces and the between cluster distances increase with movement down the hierarchy from the super-group to the sub-group level.

The sensitivity analysis conducted on the dataset revealed that the removal of any variable from the cluster analysis will produce a different result. Each variable has a unique effect on the classification that is produced. Some variables have more of an effect on the classification than others, but each variable included is vital to the classification. Even a variable that seems to have little effect on the classification will be a vital component for a number of OAs. The analysis of variability reduction within each variable as used by Voas and Williamson (2001a) shows that the OA classification significantly reduces the variability within the dataset. The OA classification reduces variability considerably more than for comparable previous systems despite the OA classification containing many fewer clusters than the other systems.

The within cluster distance reduces as the number of clusters in the classification increases, showing that the hierarchical nature of the classification doesn't just increase the number of clusters in the classification system, but creates clusters that are more focused and compact. Between cluster distances vary between clusters and increase as the number of the clusters in the hierarchy grows.

The classification does have outliers and atypical areas which do not fit comfortably into the hierarchy appearing to towards the edge of a cluster. Investigations have shown that these clusters or examples of unusual residence are often communal establishments which do not fit easily into prescribed groups. They are not examples of mistakes in the classification process. Atypical OAs are not equally distributed throughout the UK, almost 50% of atypical areas located in Scotland. The ground truthing exercise has shown that the names and descriptions of the clusters were representative of what could be seen there in reality.

By "fuzzyfying" the classification it is possible not only to see which cluster each OA belongs to but their distance from all cluster centres. This enables analysis of more detailed patterns of each individual cluster type. This is especially useful when examining OAs that fall towards the edge of a cluster and may be as easily attributed to another cluster as to the one they are in.

The consultation exercise was both innovative and informative and provided a wealth of quality assurance data and comments about the classification. The classification exercise results were very positive with respondents achieving an 85% accuracy rate of matching the areas on the maps to the group names. The majority of comments about the classification had praise for the quality of the product and there was a lot of interest in using the classification. Comments also provided several action points and ideas for future projects. The consultation also proved to be a good advert for the release of the classification as many of those who were involved in the consultation stated that they would be interested in using the classification.

Chapter Seven - Testing the OA Classification: Accounting for behaviours and patterns

7.1 Introduction

As well as the consultation exercise (outlined in Chapter 6) a number of investigations of utility of the OA Classification have been carried out. These investigations were undertaken with two intentions: firstly, as a further test and therefore quality assurance of the classification and secondly, to illustrate just a few of the many possible uses that the classification has, thus underlining its value.

The use of area classification and geodemographics for social policy driven applications is essential to place them in a proper socio-geographic context. Geodemographics should be viewed as more than a marketer's tool and the reason behind a lot of unwanted mail and phone calls. By exemplifying the wide range of potential political, social and culturally relevant uses of the classification, it is hoped that its use in these kinds of applications will increase. It will hopefully become clear that geodemographics can be used in a more socially motivated and less market driven way.

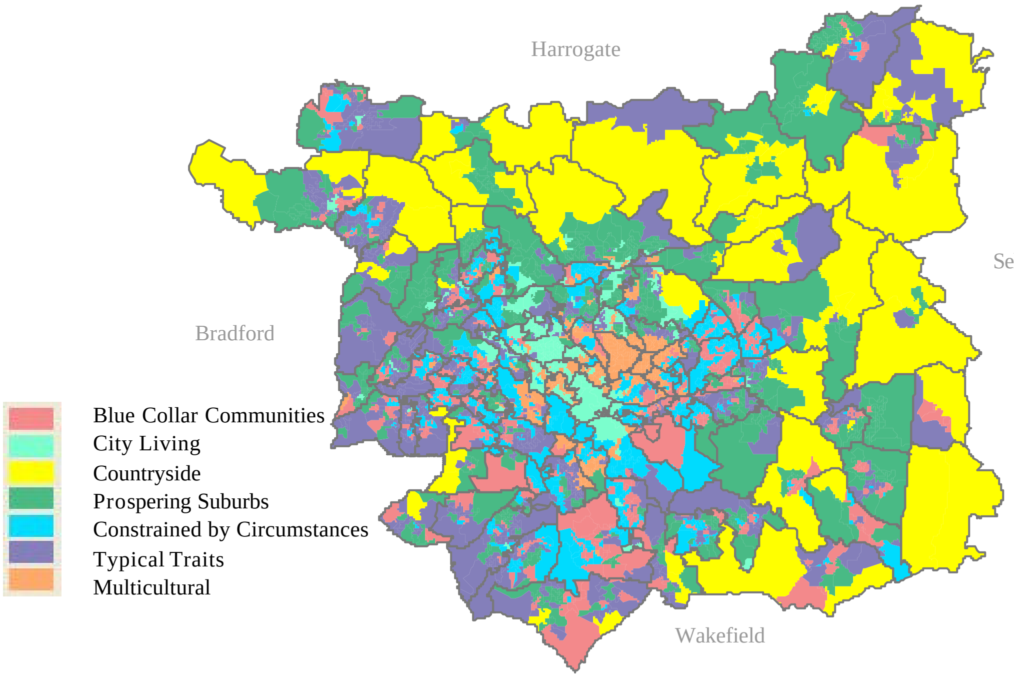
This chapter illustrates the value of the classification with a series of case study examples. Section 7.2 shows a simple example of how the classification can be used to map the residential zones of the city of Leeds. Section 7.3 uses the classification to compare and contrast the socio-demographic make up of the eight members of the core cities group. Section 7.4 shows how the classification can be used to account for changes over time by comparing voting patterns for the 2001 and 2005 general elections, using the classification to see which people changed their vote. Section 7.5 profiles the classification against the ODPM classification of rural areas for England and Wales enabling an assessment of diversity within rural areas to be made. Section 7.6 profiles the classification against the ODPM's Indices of Multiple Deprivation and assesses the extent to which the classes vary in deprivation status. Section 7.7 computes the percentage of people who speak Welsh for the classified areas of Wales, extracting socio-economic reasons for the observed differences. Section 7.8 profiles religion in Northern Ireland against the classification to investigate if socio-economic differences can still be seen either side of the religious divide. Section 7.9 profiles the classification against migration data from the 2001 Census to build a picture of who is migrating at the start of the 21st century. Section 7.10 examines whether the north-south divide exists today using the classification. Section 7.11 concludes the chapter and discusses how successful the classification has been in accounting for or predicting differences in the examples used.

7.2 Focus on Leeds: the Geography of a Regional City

Leeds is a large city in the north of England with a population of 715,402 people (2001 Census). Leeds is a good proxy for the UK as demographically it closely matches the national average and locally it is often known as 'Average Leeds'. A recent investigation also shows the Aireborough ward of Leeds as the most average in England and Wales (Dilley 2003). The boundary of Leeds stretches well out of the city into more rural areas so the administrative city covers a wide range of urban and rural residential areas. An investigation into the spread of the super groups in Leeds reveals some interesting patterns. Figure 7.1 shows Leeds split into community areas as defined by Peter Shepherd of the School of Geography, University of Leeds and used as the geographic base for the maps in a major new book about the city (Unsworth and Stillwell 2004). The community areas were designed to group similar areas together and are built from 2001 Census Output Areas so they can be directly compared. 106 community areas were created with the intention of forming spatially coherent areas whose names are meaningful to the people of Leeds. The 106 community areas were derived from previous work by Leeds City Council who identified 100 communities. The 100 communities were overlaid on the OA boundaries using a GIS. Each OA was assigned to the community in which it fell or the majority of its area fell. New community area boundaries were defined by merging the boundaries of the OAs that were assigned to each community. In some of the more rural part of Leeds larger community areas were subdivided using roads and parish boundaries increasing the number of community areas to 106 (Unsworth and Stillwell 2004). As the community areas are built from OA boundaries, the OA classification can be used to examine the homogeneity and diversity within them.

It is possible to distinguish patterns that could be assigned to several models of urban structure within the distribution of super-groups around Leeds. The super-groups seem to be spread around the city in a fairly regular pattern, which approximately resembles the concentric ring model of a city as proposed by Burgess and Park of the Chicago School in the 1920s (Park & Burgess 1925). However, the city also shows Hoyt-like sectors with south and southwest Leeds having very different mixes of OA types than the northwest, north and east (Hoyt 1939). The Burgess and Hoyt models are characterisations of map patterns of the socio-economic status of neighbourhoods in cities. Historical studies associated with each model have established the processes leading to the patterns that they display. The patterns are similar to those shown in Chicago by Rees (1970) who found both ring and sector patterns as important dimensions in the distribution of neighbourhood types.

Figure 7.1: A Map of Leeds showing the distribution of Super-groups by Community Areas



The distribution of the super-groups is also dependent upon the geography of the individual city and its surroundings (Carter 1995). Leeds is no different from any other city in this matter; the geographical surroundings of the city have without doubt affected the distribution of residential types within the city. The west and the south of the city are bordered by other large urban areas (Bradford, Wakefield), whereas the north and east of the city are bordered by much smaller conurbations that are surrounded by countryside (Harrogate, Selby). Figure 7.1 clearly shows the difference in the residential patterns between the southwest and northeast of the city, there is little doubt that this has developed because of the difference in function of the land that surrounds Leeds.

The distribution of super-groups throughout the city can be summarised as follows. *Blue Collar Communities* are well spread across the city, but mainly concentrated in the south of the city. *City Living* is seen mainly in the centre of the city and running into the main student areas of around Hyde Park and Headingley. The *Countryside* super-group surrounds the city's urban extent to the north and the east; it covers an area which is outside urban Leeds, but still within the local authority boundary. *Prospering Suburbs* represents the urban extent of the city to the north and the east, also with outliers in Aireborough and Wetherby. Beyond this group is the *Countryside* super-group. In the main the *Constrained by Circumstances* super-group surrounds the city centre enclosing *City Living* and *Multicultural. Typical Traits* can be seen in suburban areas of the city mainly to the south, but there are also significant extent in Aireborough (north west corner) and Wetherby (north east corner). This super-group can be predominantly seen to the north east of the city centre, neighbouring the *City Living* super-group. Leeds contains the highest proportion of city living of the LAs in the Yorkshire and the Humber region (Vickers and Stillwell 2005)

So how homogeneous are the community areas? By looking at the distribution of OA Classification types within the community areas an assessment as to how homogenous the community areas are. Many of the community areas are dominated by one or two of the groups suggesting that they are homogenous in their make up. Analysis shows 85 of the 106 community areas are made up of 40% or more by a single super-group type. No community area contains all seven super-group types, ten community areas contain six super-group types, thirty one community areas contain five super-group types, twenty nine community areas contain four super-group types, eighteen community areas contain three super-group types, sixteen community areas contain two super-group types and two community areas contain only one super-group type. Table 7.1 shows the five most diverse and the five most homogenous community areas. The most diverse community area based on the standard deviation of the number of OAs of each super-group type within the community area is *Ireland Wood*. The lower the standard deviation the more diversity there is in a community area, the higher the standard deviation the less diversity. In the north west of the city, it contains 14 OAs the most numerous super-group type is *Constrained by Circumstances* with 4 OAs. All super-group types apart from *Countryside* are represented in *Ireland Wood*. The most homogenous community area is *Harehills Triangle*, which is located to the east of the city centre; this community area contains 23 OAs all of which are in the Multicultural super-group. *Ledston & Ledsham* in the south east corner of the Leeds district also contains 100% of one super-group type *Countryside*, although this community area only contains 2 OAs.

Table 7.1: The diversity of OA types within selected community areas

Community Area	s.d.	Community Area	s.d.
Ireland Wood	8.54	Harehills Triangle	34.99
Roundhay	10.87	Ledston & Ledsham	34.99
Fearnville	11.78	Chapeltown	33.81
Upper Armley	12.07	Garforth East	31.50
Scott Hall & Miles Hill	12.49	Horsforth West End	31.44

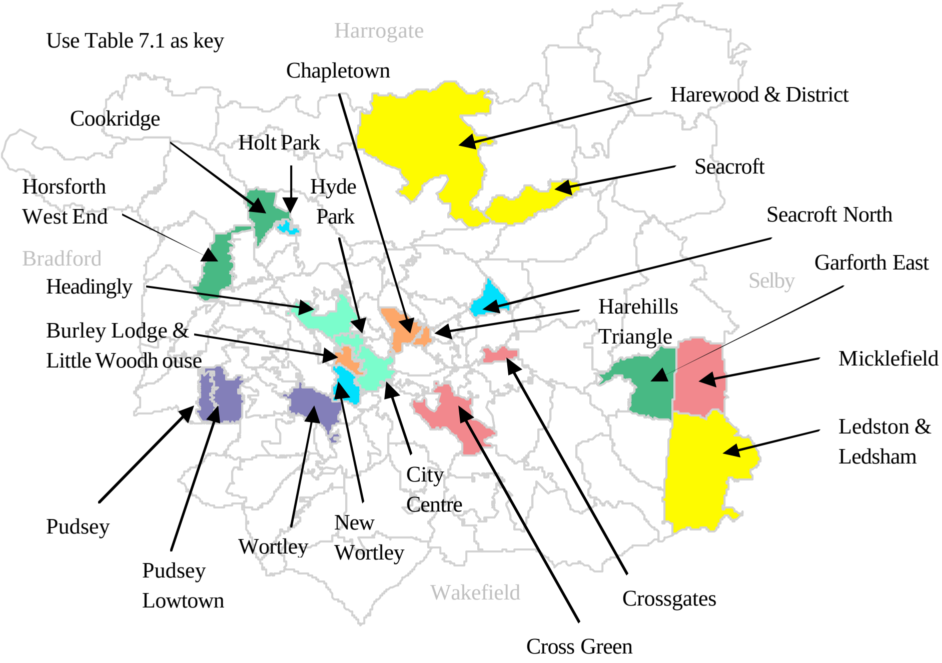
Table 7.2 shows the Community Areas that contain the highest proportion of each super-group. The area with the highest proportion of *Blue Collar Communities* is Micklefield, a former mining village to the east of the city about halfway between Leeds and Selby. The other two areas are much closer to the centre of the city and can be seen in Figure 7.2. The three areas that have the highest proportion of the *City Living* Super-group are contiguous and near the centre of the city running past two universities. The three Community Areas that contain the highest proportion of the *Countryside* super-group are unsurprisingly towards the edge of the local

authority boundary. The Community areas that contain the highest proportion of *Prospering Suburbs* are found between the built-up area around the centre of the city and the surrounding rural fringe. Two of the areas are in the prosperous north west of the city. The other Garforth East is to the east of the city and is a commuter settlement that has grown significantly in recent years its location on the main train line between Leeds and Selby means that that it is an ideal location to commute into the city.

Table 7.2: The Community Areas that contain the highest proportion of each super-group

Super-group	1st	2nd	3rd
Blue Collar Communities	Micklefield 67%	Crossgates 60%	Cross Green 50%
City Living	City Centre 82%	Hyde Park 78%	Headingley 77%
Countryside	Ledston & Ledsham 100%	Scarcroft 75%	Harewood &District 50%
Prospering Suburbs	Garforth East 91%	Horsforth West End 91%	Cookridge 83%
Constrained by Circumstances	Holt Park 78%	New Wortley 78%	Seacroft North 72%
Typical Traits	Pudsey 72%	Pudsey Lowtown 57%	Wortley 56%
Multicultural	Harehills Triangle 100%	Chapeltown 97%	Burley Lodge & Little Woodhouse 80%

Figure 7.2: The location of the Community Areas that contain the highest proportion of each Super-group



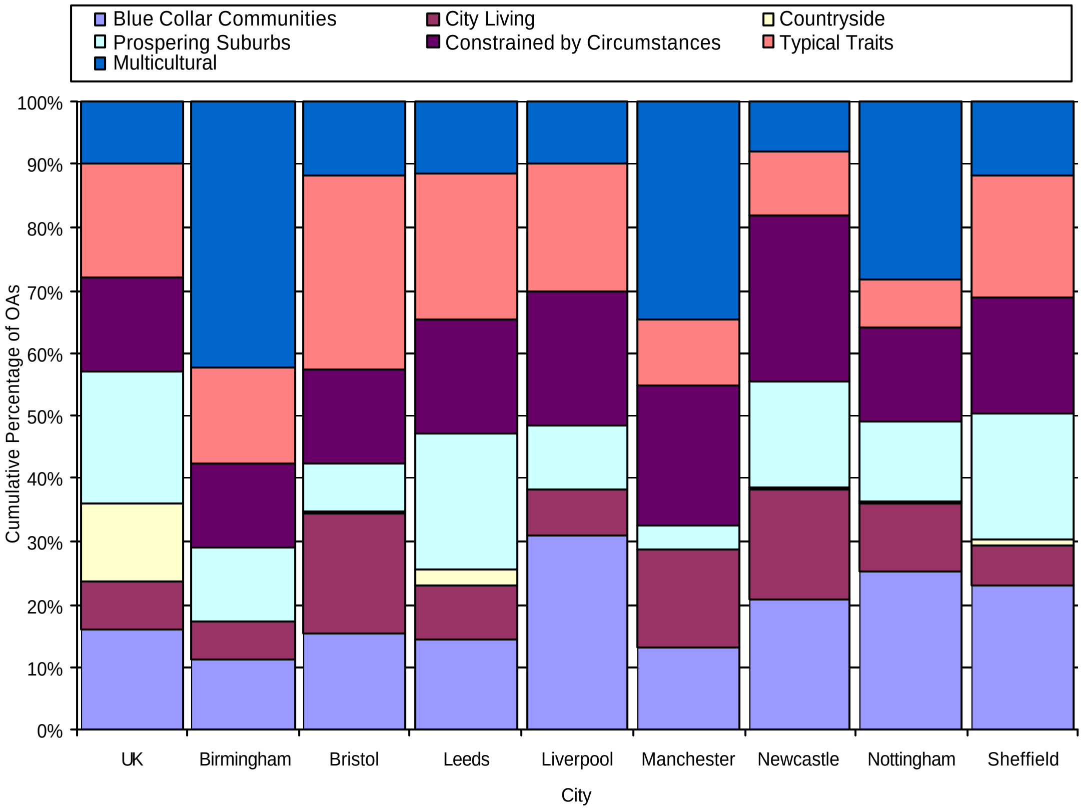
The areas that contain the highest proportion of *Constrained by Circumstances* OAs are located on the periphery of the built up area of the city. The three areas that contain the highest proportion of the *Typical Traits* super-group are located to the west of the city centres on the boundary with Bradford. The Community Areas that contain the highest proportion of *Multicultural* are areas around the centre of the city.

7.3 A Tale of Eight Cities: Profiling England's Core Cities Group

The study of Leeds can be broadened out to look at all eight Core Cities of England established in 1995 (Core Cities Group 2005). The Core Cities Group includes Birmingham, Bristol, Leeds, Liverpool, Manchester, Newcastle, Nottingham and Sheffield. The group was formed to push forward the economies of England's eight primary regional centres (outside London). The Core Cities' mission statement states the role of the Core Cities as: *"To work in partnership with Government and other key stakeholders to promote and strengthen Core Cities as drivers of regional and national competitiveness and prosperity with the aim of creating internationally competitive regions"* (Core Cities Group 2005).

As eight large regional centres that have joined together to strengthen their power and influence it would be logical to assume that the Core Cities would be broadly similar in their social make up. This is a theory that can be tested with use of the OA Classification. Figure 7.3 shows the distribution of super-group types within each of the cities (the extent of each city is taken to be its local authority district).

Figure 7.3: The percentage of each Super-Group in the eight Core Cities



What becomes immediately apparent is that despite being grouped together as *"a group of cities which represent England's largest City-regions"* (Core Cities Group 2005), there are significant differences in the social make-up between the cities. The clearest division is in the *Multicultural* super-group which dominates Birmingham, Manchester and Nottingham. These three cities have a percentage membership of the *Multicultural* super-group which is close to that of London while the other five cities are much closer to the UK average. This is the greatest social division within the Core Cities and they could be grouped into two separate types on that basis.

Liverpool contains the greatest proportion of *Blue Collar Communities* at just under 31%. Bristol contains the largest percentage of *City Living* (19.1%) and *Typical Traits* (30.7%). Leeds contains the highest proportion of *Countryside* (2.5%) and *Prospering Suburbs* (21.7%). Newcastle contains the largest percentage of *Constrained by Circumstances* which represent 26.3% of the city's OAs. These figures can be compared to the distribution of the whole UK with the use of Appendix D.

It is clear that although the Core Cities can be thought of as being broadly similar there are significant differences between them. These differences are real and there to be seen, but the boundaries of the local authority districts (LAs) also need to be considered, as they could go some way to explaining why these differences exist. A classic example is a comparison of Leeds and Manchester. The boundary of the Leeds LA goes far beyond the urban extent of the city and takes in small towns, villages and countryside. In contrast, the boundary of Manchester takes in a lot smaller area than people would consider as the City of Manchester. Much of what many would consider to be part of the Manchester region is in fact in neighbouring LAs such as Trafford, Tameside or Salford.

7.4 The Swingometer: Change in British Electoral Patterns 1997 – 2005

Norris and Evans (1999) set out a framework for the study of election change, outlining five differing election types based on the change in voting patterns that is experienced. The framework is outlined as follows:

- Maintaining Elections: Characterised by electoral flux where only a few voters shift between parties but the balance of power is not significantly.
- De-aligning - Deviating Elections: A temporary sharp reversal in the 'normal' share of the vote caused by particular personalities issues or events.
- De-aligning - Secular Elections: A long-term cumulative and incremental progressive weakening of a party's vote.

- Re-aligning - Secular Elections: An evolutionary, cumulative strengthening of a party's vote over a series of elections.
- Re-aligning - Critical Elections: These are exceptional elections that produce an abrupt and significant re-alignment within the electorate, which has long term consequences for the party order.

(Adapted from Norris and Evans 1999)

By examining the change in voting patterns between the 2001 and 2005 general elections it will be possible to ascertain which type of election took place for each of the parties. However, with the use of the OA classification much more than this can be done. Webber (2004) has previously shown how classes in geodemographic systems display different voting patterns. By splitting the electorate by the OA classification it is possible to look at the election separately for each super-group. These elections within the elections can be examined to establish whether the OA classification can bring out multiple election patterns within a single election.

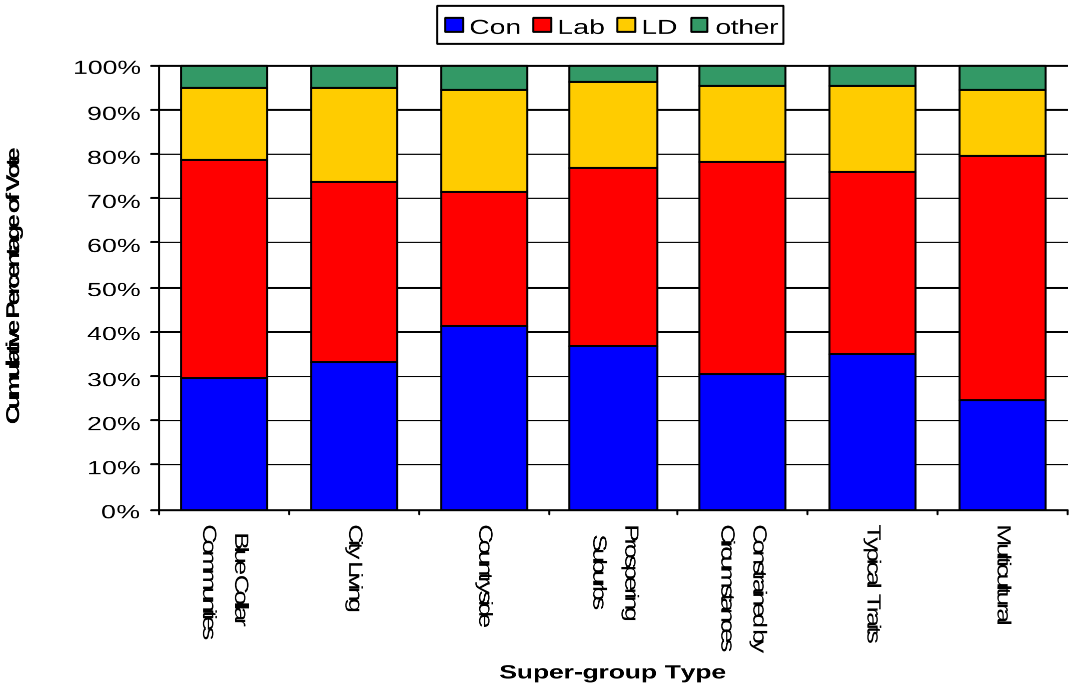
The study was confined to England and Wales, due to boundary changes in Scotland and the main parties not standing in Northern Ireland. This analysis was done by assigning the percentage vote of the three main political parties from each parliamentary constituency to each constituent OA. OAs were linked to the parliamentary constituency in which their population weighted centroid fell. This is an approximate procedure, which is restricted by the lack of election results for any lower level of geography.

The first thing that has to be established is the state of play before the 2005 election. The 1997 General Election was without doubt a *Re-aligning: Critical Election*. Labour swept into power with a 'landslide' majority of 177 seats. This was a large and significant change after 18 years of Conservative government, brought about by a young and dynamic Labour leader Tony Blair. Blair defined the party "New Labour" and positioned it towards the centre of the political field away from the traditional Labour left and the trade unions.

In the 2001 UK General Election (held on the 7th June 2001, 6 weeks after the 2001 Census) the Labour Party of Prime Minister Tony Blair won a second term in office by returning 413 MPs to Westminster with a second 'landslide' majority of 167, a reduction in majority by just ten votes. Labour's popularity remained high and the other two main parties failed to make significant inroads into Labour's territory. The Conservative Party led by William Hague was second with 166 MPs and the Liberal Democrats led by Charles Kennedy returned 52 MPs. Hague was forced to resign soon after the election. This election can be seen as a *Maintaining Election*, little had changed in the mind of the electorate between 1997 and 2001.

Figure 7.4 shows the 2001 vote split by super-group. Labour were the most popular party in six of the seven super-groups. Only in the *Countryside* did the Conservative party prove to be more popular. The Liberal Democrats trailed in third place for all super groups.

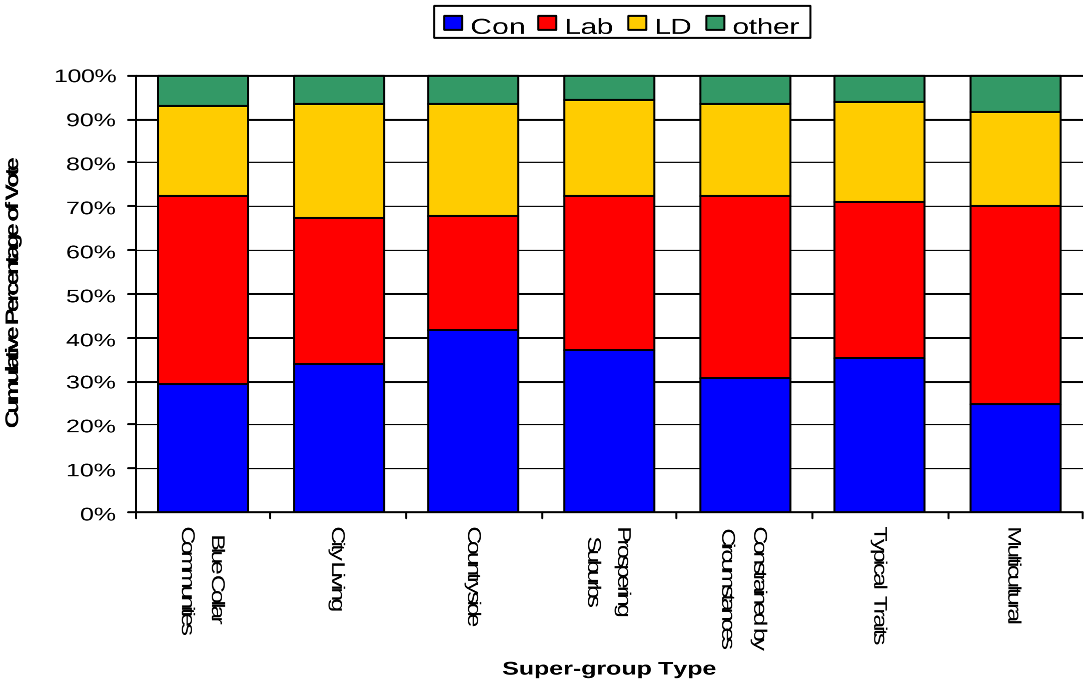
Figure 7.4: The 2001 General Election results for England and Wales assigned to OAs and split by Super-groups.



Following an unpopular war in Iraq and several ministerial scandals by the time of the 5th May 2005 General Election, Labour's popularity had dropped significantly. Political commentators were suggesting that the Conservative Party (led by Michael Howard) had an outside chance of taking power. The result of the 2005 General Election was to once again elect the Labour Party to government, but with a much reduced majority of 66. Labour returned 356 MPs, the Conservatives 197, and the Liberal Democrats 62.

Figure 7.5 shows the results of the 2005 general election by Super-group. Labour were the most popular party in four of the seven super-groups, with the Conservative party now being the most popular in three of the seven groups. The Liberal Democrats trailed in third place for all super groups, despite their vote increasing across the board.

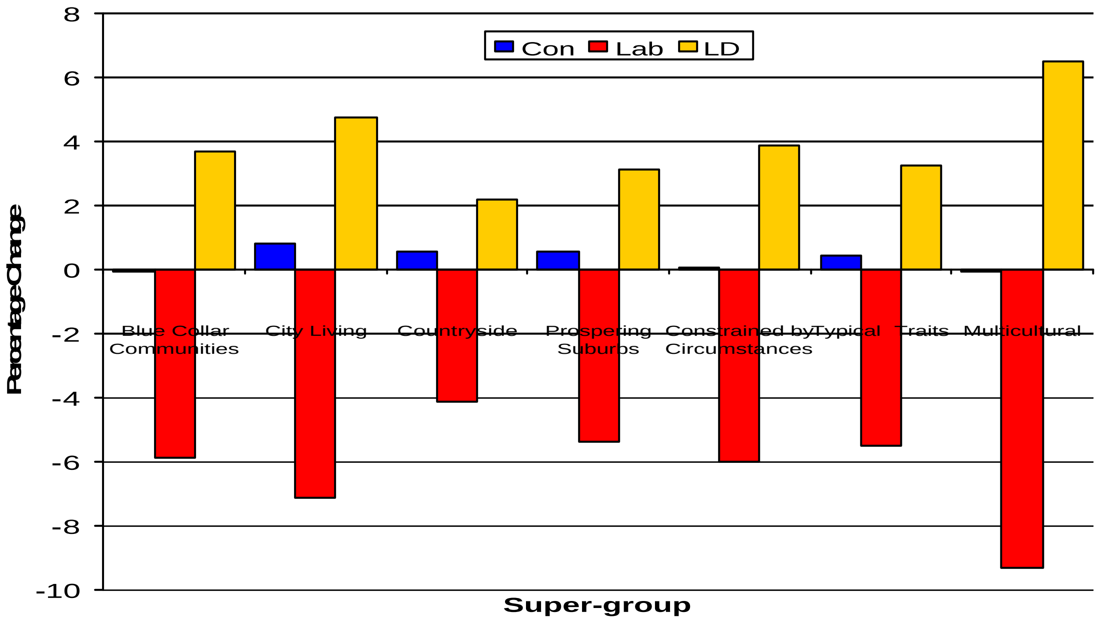
Figure 7.5: The 2005 General Election results for England and Wales assigned to OAs and split by Super-groups.



First let's consider the overall picture. Labour lost 47 seats which it held in 2001, 31 to the Conservatives, 12 to the Liberal Democrats and 4 to others. The Labour vote dropped by 5.4%, while the Conservatives increased by 0.6% and the Liberal Democrats increased by 3.8%. So which of Norris and Evans' election types did each party experience in 2005? Labour experienced what was possibly the start of a process of a *De-aligning: Secular Election*, whereas the Conservatives are probably starting to experience the opposite a *Re-aligning: Secular Election*. The Liberal Democrats seem to be going through the same processes as the Conservatives, but they could be experiencing a *De-aligning: Deviating Elections* along with a *Re-aligning: Secular Election*. The emergence of the Liberal Democrats as the third main party in British politics has made election results much harder to interpret although a lot more interesting. The Liberal Democrats undoubtedly benefited from their opposition to the Iraq war, which would cause their vote to be seen as a *De-aligning: Deviating Election*, despite the fact it follows a trend of increasing share of the vote from previous elections.

By examining the swing within each super-group, it should be possible to disaggregate the vote and examine patterns within patterns. Figure 7.6 shows how the majority of the changes in the voting patterns were caused by a swing from Labour to the Liberal Democrats. Although the swing can be seen in all super-groups the pattern is not uniform with the *Multicultural* super-group turning their back on Labour to the greatest extent, where their share of the vote was reduced by over nine percent. In other super-groups the swing was not so extreme in the *Countryside* and *Prospering Suburbs* the reduction in the Labour vote was only half of that of *Multicultural*. This could perhaps be interpreted as an Iraq war effect, the *Multicultural* super-group is the cluster in which people of Middle Eastern origin or Muslim faith are most likely to live.

Figure 7.6: The 2001-2005 change in the voting pattern in the UK General Election in England and Wales assigned to OAs and split by Super-groups.



Do the changes that can be seen match any of the hypotheses suggested by Norris and Evans? On first glance there appears to be little difference between the super-groups in terms of the swing experienced. There is a difference in magnitude of the swings, but the trend in all super-groups is the same. However, this difference in magnitude is important and can be explained by different forms of election using Norris and Evans' election change framework. For the purposes of explanation the seven super-groups can be seen as containing three patterns between them.

Firstly, the *Countryside*, *Prospering Suburbs* and *Typical Traits* super-groups could be said to have experienced *Secular Elections* with the Labour vote *de-aligning* due to general apathy towards the party in power and the Conservatives and the Liberal Democrats *re-aligning* as much because of the fall in popularity for Labour as any real appetite for the other two parties.

Secondly, in the *Multicultural* and *City Living* super-groups, there is more going on than can be explained by a Secular Election. The Iraq war was a big issue for the people living in these super-groups; there was vocal condemnation of the government's support of the US led invasion of Iraq by many in these communities. There is no doubt that this was an issue for many when casting their vote in the May 5th ballot. The larger swing from Labour to the Liberal Democrats (the only major party which opposed the war) is therefore without doubt linked to this. The swing in these super-groups can therefore be seen as a *De-aligning: Deviating Election*, based on opposition to the Iraq war.

Thirdly the swing in the *Blue Collar Communities* and *Constrained by Circumstances* super-groups falls somewhere in the middle of the previous two examples, it is likely that the majority of the swing in these super-groups is due to a *Secular Election*, but it is also likely that some form of opposition to the Iraq war played a part in increasing the swing.

The key to understanding the change in 2005 was the performance of the Liberal Democrats; they gained in terms of the percentage of the vote, but not to the same extent in terms of the number of seats. This is due to Britain's first past the post system. Effectively the move away from Labour towards the Liberal Democrats enabled the Conservatives to win back a number of seats despite only a marginal increase in their overall share of the vote.

Although this example is fairly crude due to the imputation of voting patterns from large political constituencies to much smaller OAs, it clearly shows that the OA classification can bring out differences in voting patterns. The classification shows not only the traditional breakdown of the vote by social class, with Labour strong in the urban and least wealthy areas and the Conservatives strong in the rural and most affluent areas, but can also enable an analysis of the reasons behind different degrees of swing in different areas. It can account for the changes in voting behaviour between elections and demonstrate which clusters were responsible for the changing patterns and perhaps a changed government.

7.5 Out in the Country Air: Disaggregating the Urban-Rural Classification

The answer to the question of what is "rural" has always been one that has provoked a great deal of argument and debate. A succession of definitions has been produced in recent years, but they have been limited by the availability of data rather than any theoretical perspective or scientific analysis (Select Committee on Environment, Food and Rural Affairs 2002).

It is difficult to differentiate 'rural' from 'urban' areas, however they are defined there will always be disagreement on how it should be done and the result that is produced. Many would suggest that the 'rural' could not be defined as it is a concept of the mind that means something different to each person. However, it is necessary to have a geographical definition of rurality on which to base the rural policy and funding distribution (DEFRA 2004).

The definition produced in 1993 for the Rural Development Commission in the 'Tarling Report' is the widest used previous definition of rurality; the classification is based on local authority districts. Under the heading of 'urban' the Tarling report identifies metropolitan, major urban and 'coalfield' local authority districts and the remainder, termed 'rural', are sub-divided into 'remote rural' and 'accessible-rural' according to accessibility to the metropolitan areas of England (DEFRA 2004).

There are acknowledged weaknesses in this classification. Most of the non-metropolitan districts are classified as 'rural', despite containing sizeable towns within them. All contain different patterns of the small towns, villages, hamlets and isolated dwellings that make up their 'rural' settlement pattern (DEFRA 2004). The local authority level geography on which the Tarling report is based is clearly too broad to provide adequate gradations of 'rural'. To provide a settlement oriented and more policy relevant definition of 'rural' a consortium of government organisations was commissioned to create a completely new and innovative definition of 'rural' for England and Wales (DEFRA 2004).

The increasing availability of data at finer scales has opened up real choices in terms of the way that rural definitions can be constructed. The Office of the Deputy Prime Minister (ODPM) commissioned a classification of urban and rural areas for England and Wales that was released in 2004 (DEFRA 2004). The new classification is based on Output Areas. Each OA is initially defined as urban or rural depending on whether the majority of the population falls inside a settlement with a population 10,000 or more. The overall classification is based on a settlement approach and builds upon the identification of rural towns, villages and scattered dwellings within a grid framework of cell size 1 hectare (100x100 metre squares) (Bibby & Shepherd 2004). The grid was used as the basis for the classification of Output Areas in terms of settlement context and settlement form. The OAs classified as rural are then sub-classified into *Town and Fringe*, *Village*, and *Hamlet & Isolated Dwellings* (ONS 2005c).

The intention of the project was to come up with a single classification to replace and to harmonise previous multiple definitions of rurality (ONS 2005c). The stated objective of the project was to identify policies that required or would benefit from a definition of urban and rural areas. This would aid the development of the production of a set of definitions to meet a wide range of policy needs and also new techniques that would better meet both established and anticipated needs (Bibby & Shepherd 2004). As the urban-rural classification was created at the Output Area scale it is geographically compatible with the general purpose Output Area Classification created by this project. By profiling the Classification of Rural Areas against the OA Classification it is possible to both check the two classifications against each other and investigate the diversity within the defined rural areas.

Figure 7.7 shows the percentage of each OA Classification group type which is in each urban-rural classification type. What this essentially tells us is the make-up of the OA Classification groups. By examining this cross-tabulation the urban/ruralness of each OA Classification Group can be established. The most striking thing about the groups by urban-rural class is that for 18 of the 21 groups 70% or more of that group type are located within the 'Urban >10k' class. This means that the majority of OAs in these groups are to be found in a non-rural setting. This leaves three groups (3a, 3b & 3c) where over 75% of OAs fall into non-urban settings. What this suggests is that urban and rural areas are not just different in terms of the amount of grass or concrete that can be seen in the respective areas but that they represent a very different combination of residential social groups. The most polarised OA classification group types are 7a *Asian Communities* and 7b *Afro-Caribbean Communities*, 100% of both are found in Urban >10k. The most diverse in terms of urban rural types is group 3c *Accessible Countryside*, it is made up by 20.7% *Hamlet & Isolated Dwellings*, 38.5% *Village*, 18% *Town and Fringe* and 22.7% *Urban >10k*.

Figure 7.7: The percentage of each OA Classification group type which is in each urban-rural classification type

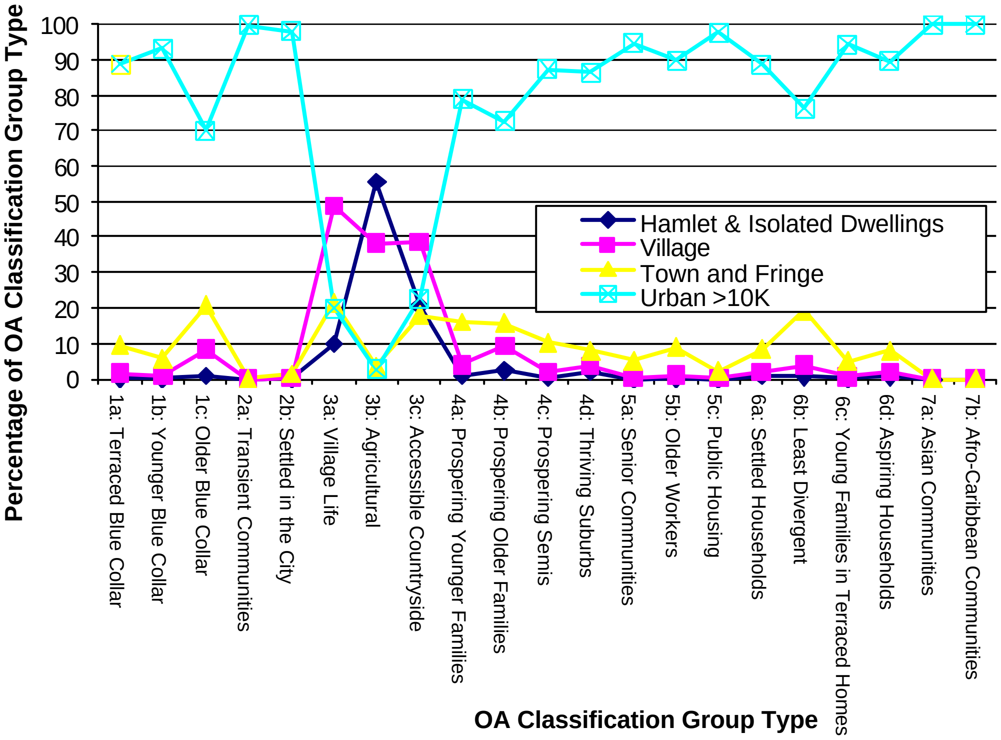


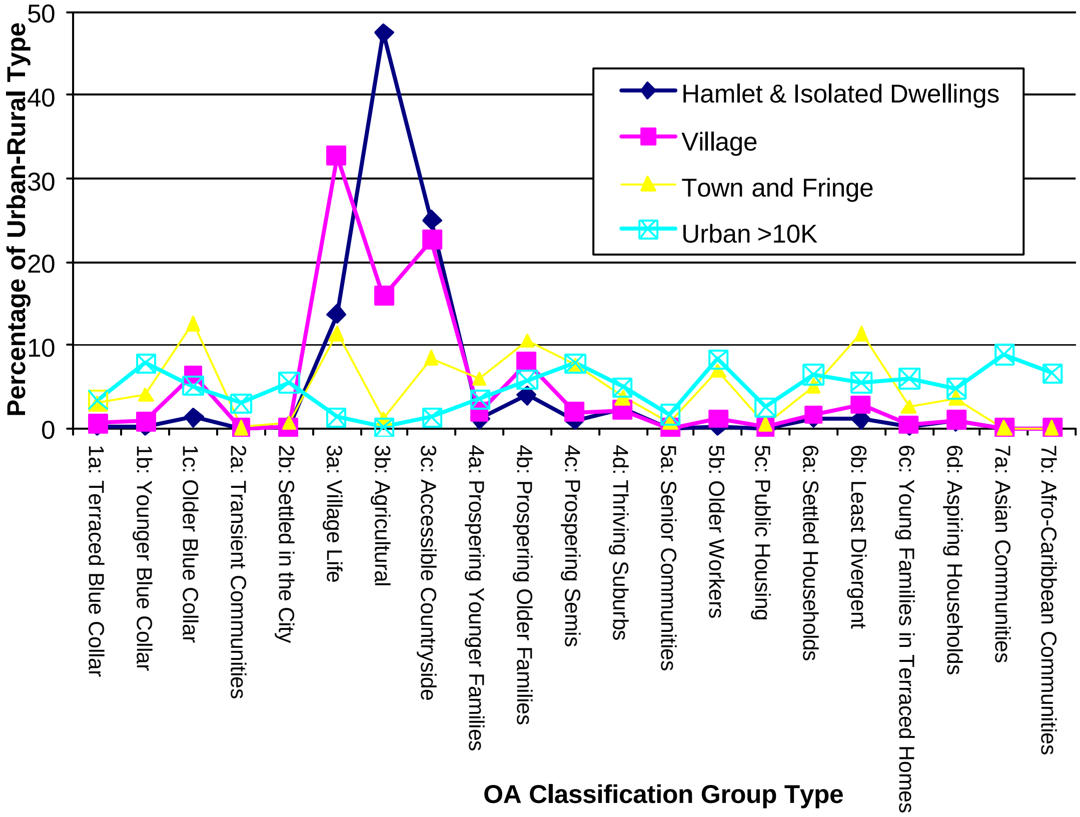
Figure 7.8 shows the percentage of each urban-rural classification type that is in each OA Classification group type; this shows the social diversity within the urban-rural classification. The *Urban >10k* class shows the greatest diversity with the highest representation of any single group being 8.9% 7a Asian *Communities* with all the groups present.

Town and Fringe is slightly less diverse than *Urban >10k* with four groups (*Older Blue Collar, Village Life, Prospering Older Families and Least Divergent*) each accounting for over 10% of the OAs within this class and two groups (*Asian Communities and Afro-Caribbean Communities*) unrepresented.

The *Village* class shows a real switch from diversity to more homogeneous communities. Four groups (*Transient Communities*, *Senior Communities*, *Asian Communities and Afro-Caribbean Communities*) are unrepresented. 'Village Life' makes up just under a third of the 'Village' class with *Agricultural* and *Accessible Countryside* making up a further 38.5% between them, making these three groups responsible for 71.2% of the OAs in this class.

Hamlet & Isolated Dwellings shows real homogeneity with *Village Life* (13.7%) *Agricultural* (47.5%) and *Accessible Countryside* (25.0%) making up 86.2% of all OAs in the class. Five groups (*Transient Communities*, *Senior Communities*, *Public Housing*, *Asian Communities and Afro-Caribbean Communities*) are unrepresented in this class.

Figure 7.8: The percentage of each urban-rural classification type that is in each OA Classification group type



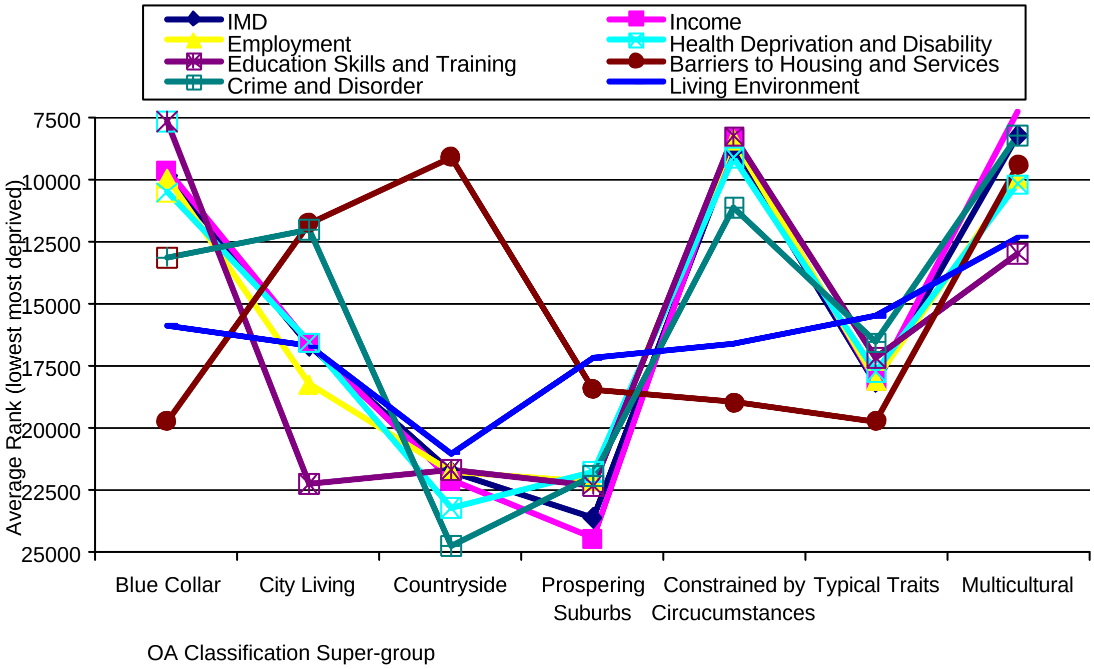
By cross-classifying the OA Classification with the urban-rural Classification (OA level) it becomes possible to combine the power of both classifications to add further context to both classifications. The exercise also shows how the prevalence of different social groups changes with increased rurality or urbanity. Great social diversity can be seen in the most urban group, whereas the more rural groups contain only a few social area types. Social groups that are traditionally associated with urban settings, such as people of non-white ethnicity and students are not found in the rural extremes.

7.6 How do the OA Classes Score in Terms of Deprivation?

The Indices of Multiple Deprivation 2000 (IMD) have been widely used despite receiving some criticism. It was considered a successful method and the Oxford Team was re-commissioned by the Office of the Deputy Prime Minister (ODPM) to produce a 2004 version. The analysis was once again carried out by Social Disadvantage Research Centre (SDRC) at the Department of Social Policy and Social Research at the University of Oxford. The 2004 IMD was carried out at the new Super Output Area Level One (SOA1) For more information about SOAs geography see ONS (2005e), whereas the previous version had been conducted using electoral wards (ODPM 2004).

The IMD is an important tool for identifying the most disadvantaged areas, for the purpose of feeding into policy and resource allocation. The Index of Multiple Deprivation 2004 contains an overall rating and ranking of deprivation for all SOA1s in England and Wales. It also contains seven separate domains representing different forms of deprivation: Income deprivation, Employment deprivation, Health deprivation and disability; Education, skills and training deprivation, Barriers to Housing and Services, Living environment deprivation and Crime (ODPM 2004). The IMD uses data from a wide variety of sources only a few of which are from Census 2001. Therefore the indices are ideal to profile and test against the OA Classification as they contain different data. Figure 7.9 shows the average rankings the for the overall IMD 2004 deprivation average rank and for each of the individual indices aggregated by the OA Classification at super-group level.

Figure 7.9: The IMD 2004 ranks (SOA) aggregated by the OA Classification at super-group level (ranks are used rather than scores to allow comparability between the different domains)



The overall IMD shows that there is a significant difference between the OA Classification super-groups in terms of the level their deprivation. The most deprived super-group type is *Multicultural* closely followed by *Constrained by Circumstances* and third most deprived overall is *Blue Collar Communities*. The least deprived super-group type overall is *Prospering Suburbs* followed by Countryside then *Typical Traits*. Most of the indices follow the same general pattern as the overall index, but there are some exceptions. *Barriers to housing and Services* deprivation is worst in the *Countryside* super-group. This is in contrast to the other dimensions of the index where the *Countryside* is either the first or second least deprived super-group type. *Crime and Disorder* index is much higher in the *City Living* super-group than would be expected when looking at the overall IMD. The *Living Environment* index shows much less variation across the super-groups compared to the other indices. The range from the highest to the lowest average ranks is 8,723 compared to an average range of 14,634 for the other indices.

There are clear differences in the levels of deprivation by OA classification super-group. There is a general ranking of deprivation, which shows *Multicultural* to be the most deprived super-group type and *Prospering Suburbs* to be the least. However, some individual indices do differ from this pattern.

7.7 Beth yw hwn y Gymraeg? The Geography of the Welsh Language

A language that could be recognised as modern day Welsh did not really come into existence until A.D. 700. The building of Offa's Dyke along the border with England (778-796) cemented the cultural divide between the men of the hills on the Welsh side and those of the English lowlands and fraught with it the crystallisation of the Welsh language (Rees 1982). The Welsh language had a varied history up to the present day of which, Rees (1982) and Aitchison and Carter (1994) provide detailed accounts.

The principality of Wales has been under the rule of the English government since 1284, with Henry VIII establishing Welsh representation in the English parliament in 1536 though an Act of Union (Rees 1982). Despite only accounting for 4.9% (2001 Census) of the UK's population Wales has managed to retain its unique cultural identity and proud heritage. Perhaps most surprisingly Wales has managed to maintain its language. Take a trip over the border to Wales and you will find that the road signs are written in two languages, the Police are called Heddlu and they have a predominantly Welsh language TV station called Sianel Pedwar Cymru (Channel Four Wales).

To anyone who does not speak a minority language its continued existence can be seen as little more than a quaint link to the past. However, those who speak and use the language everyday recognise the language as their cultural heritage. Aitchison and Carter (1994) affirm that *"Language serves as both a symbol and a definer of culture. The well-being and integrity of a culture depend heavily on the strength and vitality of its associated language. It is for this reason that the process of language change within particular cultures needs to be monitored and scrutinised."* The fate of the Welsh language has not always been a certain one. In a BBC radio broadcast in 1962 Saunders Lewis (a writer and nationalist who advocated direct action in defence of Welsh- speaking communities and a founder of Plaid Cymru) predicted that Welsh would end as a 'living' language by the beginning of the twenty-first century (Aitchison and Carter 2004). Reflecting his nationalist views Lewis called for 'revolutionary action' to protect the language (Lewis 1962). The broadcast led to the creation of Cymdeithas Yr Iaith Gymraeg (Welsh Language Society) and is seen as by many as having saved the Welsh language from extinction (Aitchison and Carter 2004). Welsh devolution in 1997 led to the creation of a Welsh Assembly which has further safeguarded the future of the language.

In the 2001 Census Just over 28% of people in Wales claimed to have at least some knowledge of the Welsh language (i.e. they can write, speak or understand spoken Welsh). This represented a 2% increase on 1991, although as Higgs *et al.* (2004) point out this increase can be more than adequately be accounted for by a subtle change in the question asked from 'do you speak Welsh?' to 'can you speak Welsh?'. Welsh is not spoken in equal intensity throughout the country and has a very interesting and particular geography. Figure 7.10 shows that Welsh is spoken widely in the rural north of the country and on the west coast. It is much less prevalent in the east of the country (bordering England) and in the south of the country where the majority of the population lives.

Figure 7.10: Knowledge of the Welsh language according to the 2001 Census.

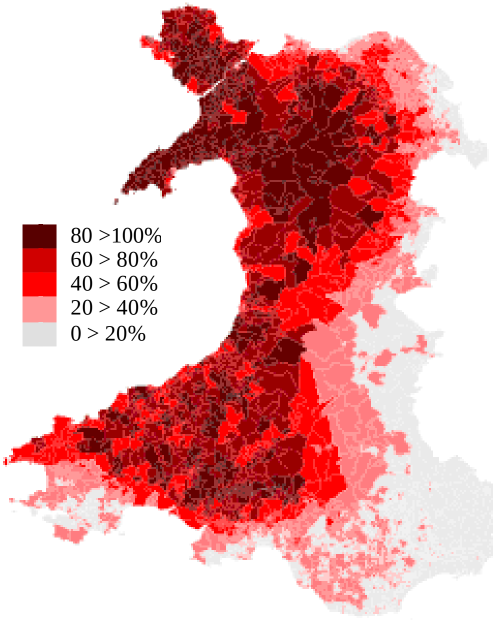


Figure 7.11: OA Classification at Super-group level for Wales

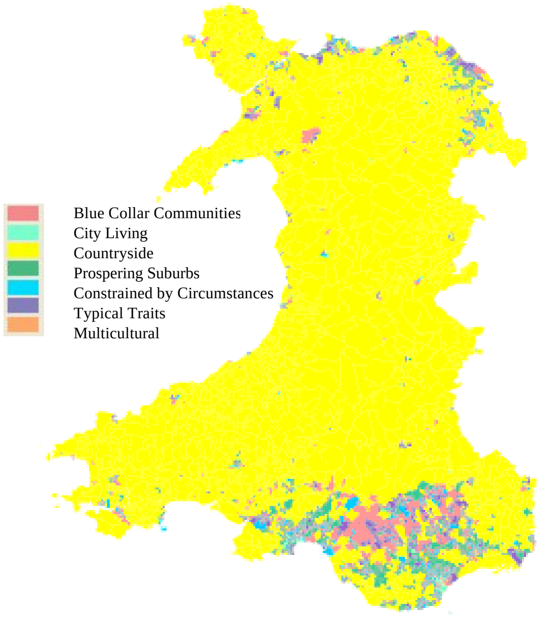
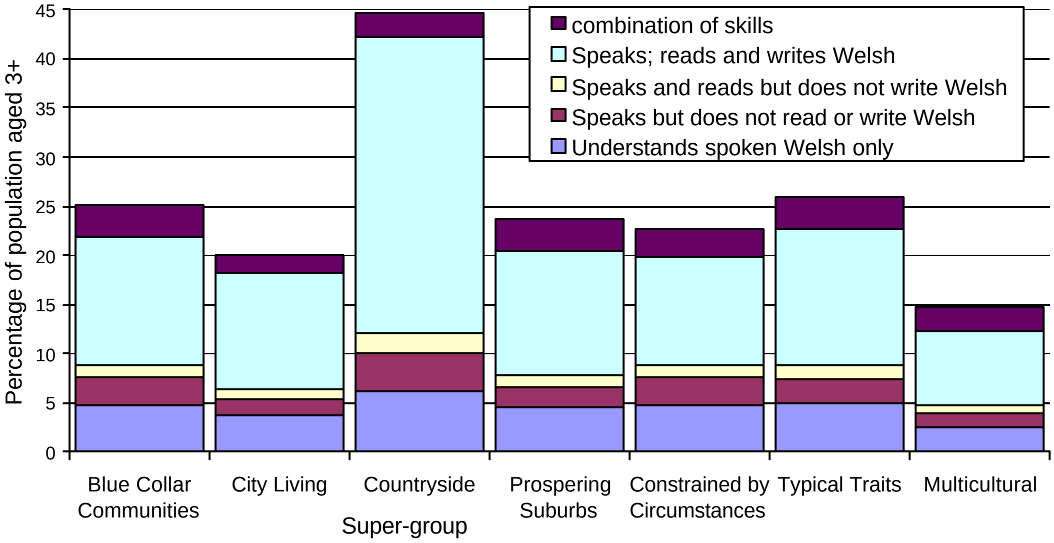


Figure 7.11 shows the distribution of the Super-group types across Wales. They do not obviously reflect the pattern shown in Figure 7.10. The map in Figure 7.10 is heavily influenced by migration into areas in Wales from England and by English speakers (returning, buying up second homes). This has led to the England/Wales borderlands being least Welsh speaking. The southern coast non-Welsh speaking areas (Vale of Glamorgan, South Pembrokeshire) are a legacy of the Norman Conquest. The South Wales coalfield low percentages are a product of mass migration by non-Welsh speakers in the nineteenth century to work in the mines and iron foundries. The consequence for the *Countryside* super-group in Wales is that it is still majority English speaking.

One thing that does seem to be clear is that Super-group 3 *Countryside* does seem to line up with the areas with high levels of Welsh speaking. It is commonly accepted that the prevalence of the Welsh language is higher in rural area than urban areas (Higgs *et al.* 2004). However, does the classification bear this out? Are there any other, subtler spatial patterns to the Welsh language that the classification can exemplify? By cross-tabulating the OA classification against the percentage of Welsh speakers in each OA, it will be possible to determine whether the OA classification can predict whether or not someone is more or less likely to speak Welsh by their membership of one or another OA classification cluster.

Figure 7.12 shows various levels of knowledge of the Welsh language profiled by the super-group level of the classification. Significant differences can be seen. The super-group that stands out is *Countryside* with 45% of people living in this super-group having some knowledge of Welsh. This contrasts with the *Multicultural* super-group where just fewer than 15% of residents have some knowledge of Welsh. It is therefore clear that there are significant geographic differences in the knowledge of Welsh based on the super-group level of the classification.

Figure 7.12: The distribution of Welsh speakers by Super-group



Unsurprisingly, as the maps in Figures 7.10 and 7.11 suggest, the *Countryside* super-group has the highest knowledge of the Welsh language. It also makes sense that the *Multicultural* super-group has the lowest knowledge of Welsh. The *Multicultural* super-group has the highest proportion of people born outside the UK therefore it is likely that a proportion of people resident in this group have neither English nor Welsh as a first language. If the knowledge of Welsh data is profiled against the group level of the classification it reveals diversity within the super-group. Group 3a *Village Life* has 49 % of its residents with some knowledge of Welsh; 3b *Agricultural* has 46% whereas only 34% of the population of 3c *Accessible Countryside* have some knowledge of Welsh. The highest knowledge of the Welsh language is in the more remote rural communities.

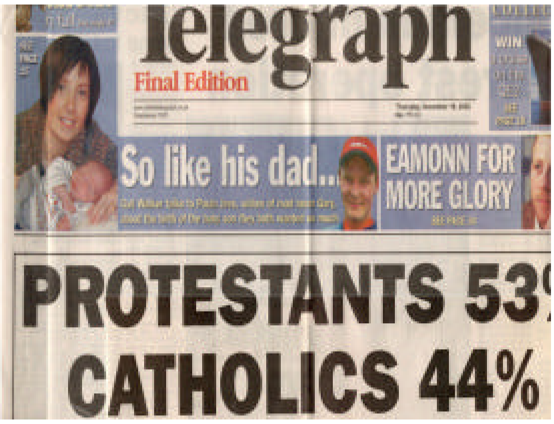
7.8 Demography Across the Divide: Religious Tensions in Northern Ireland

One of the issues that has made the news as much as any other in the UK over the past 30 years is undoubtedly the troubles in Northern Ireland. The picture is a complex one that is often simplified by the media. Tensions in Ireland can be traced back hundreds of years when English control was established on the island of Ireland in the 16th and 17th centuries. Migration from Britain to Ireland introduced Protestant Britons into Catholic Ireland resulting in tensions between the two ideologies throughout this period (O'Leary and McGarry 1993). The battle of the Boyne in 1690 is a clash which is still commemorated to this day. The island of Ireland was partitioned in 1921 creating Northern Ireland which contained predominantly Unionist Protestants who looked to London and Republic of Ireland to the south which contained predominantly Catholic Nationalists who looked to their own capital Dublin (Southern Ireland broke ties with the UK and became a republic in 1949). The partitioning of the island was hoped to be a temporary solution to the tensions, but time would prove that this would be an uneasy peace which in fifty years time would ignite into the troubles that were seen in Northern Ireland from 1971 to the mid 1990s The implementation of the Good Friday Agreement, has seen some disarmament and a peace (of sorts) across Northern Ireland (McKittrick & McVea 2001).

In its simplest form the Northern Ireland conflict can be viewed as follows: Catholic (mainly) Nationalists generally want Northern Ireland and the southern Republic of Ireland to form a united Ireland. In contrast the Protestant (mainly) Unionists want to remain part of the UK and continue their association with Great Britain. However, it is important not to see Protestants and Catholics as two opposing monoliths, as great diversity exists within the attitudes and values of both groups (Livingstone *et al*. 1998). The troubles in Northern Ireland have made violence and fear common place in the province. As the nationalist and unionist feelings are so entwined with religious belief (in fact the conflict is often misrepresented as a conflict over religion), a demographic fear based on religious belief was perpetuated. Religion is a good proxy for political/national identity; it is a pretty close match, but by no means 100%.

In contrast to the rest of the UK a religion question has historically been asked in the Northern Ireland Census. Demographic fear has risen to be a real issue in people's minds with coverage in the media, even though it had been an issue since the 17th century. Every ten years the numbers of Catholics and Protestants in the province are counted. With every census the proportion of Catholics rose and the proportion of Protestants fell. By the time of the 2001 Census rumours were rife that the number of Catholics would surpass the number of Protestants (McEldowney *et al.* 2004). The 2001 Census was awaited in Northern Ireland with just one question on people's minds: which group is in the majority, Catholics or Protestants? The Northern Ireland Census is a fundamental source of data as it provides information about religion beyond the theological (Anderson & Shuttleworth 1994; Macourt 1995). The politicians on either side of the divide were not shy in giving their opinion on the matter and media coverage was intense. Like in Eastern Europe post World War One, the partition of Northern Ireland was made on demographic grounds, so the reporting of demographic information especially on religion was very important politically.

Figure 7.13: Front page of the Belfast Telegraph 19th December 2002



Telegraph
Final Edition
WIN
So like his dad...
EAMONN FOR MORE GLORY
PROTESTANTS 53
CATHOLICS 44%

Figure 7.13 shows the coverage that the Belfast Telegraph gave to the results of the religion question from the 2001 Census. The relative numbers of Protestants and Catholics were headline news. The headline shows the results of the census reported as if it were a football or rugby score, rather than what in the rest of the UK maybe seen as just another piece of demographic information. However, it is intrinsic to the political discussion because of the basis of the foundation of Northern Ireland.

The differences between the religion question in Northern Ireland and in England and Wales shows the difference in the importance of the religion question in different parts of the UK. Figure 7.14 shows the religion question that was asked in England and Wales in 2001 (where it was asked for the first time since 1851 and was voluntary in nature). All forms of Christianity, including Protestants and Catholics are grouped together in the same option while other religious groups each have their own option. This is in sharp contrast to Figure 7.15 which shows the religion question asked in the 2001 Northern Ireland Census. Looking at question 8a it is apparent that the options are different from those in England and Wales. Each different form of Christianity is a separate option and all other philosophies are grouped together, the exact opposite to the question in England & Wales. There is now also a further religion question in the Northern Ireland Census that was introduced in 2001. Question 8b attempts to assess the religious background of the population who state they belong to no religion. What the question is essentially doing is giving a religion to those people who don't want to be identified as belonging to a religious group or who no longer practise.

Figure 7.14: England & Wales religion question

(10) What is your Religion?
This question is Voluntary

□ None
□ Christian (including Church of England, Catholic, Protestant and all other Christian denominations)
□ Buddhist
□ Hindu
□ Jewish
□ Muslim
□ Sikh
□ Any other religion, please write in

Figure 7.15: Northern Ireland religion question

(8) Do you regard yourself as belonging to any particular religion?

□ Yes go to 8a
□ No go to 8b

(8a) What religion, religious denomination or body do you belong to?

□ Roman Catholic
□ Presbyterian Church in Ireland
□ Church of Ireland
□ Methodist Church in Ireland
□ Other religion, please write in

(8b) What religion, religious denomination or body were you brought up in?

□ Roman Catholic
□ Presbyterian Church in Ireland
□ Church of Ireland
□ Methodist Church in Ireland
□ Other religion, please write in

□ None

The religious divide in Northern Ireland clearly stirs up a great deal of passion amongst the people of the province. How does the divide present itself geographically? There is a clear and distinct geography to Northern Ireland's religious divide; the east of the country is dominated mainly by Protestants while Catholics are found in greater numbers in the west. Figure 7.16 shows the distribution of Catholics in Northern Ireland at the time of the 2001 Census. There are clear sections of Protestant and Catholic territory. The map shows that most of the country is made up of areas which are either over seventy percent or less than thirty percent Catholic.

Mixed areas of between thirty and seventy percent Catholics are few. The Belfast insert is a microcosm of the picture that can be observed across the country. Catholic and Protestant dominated areas can be clearly seen. Areas where large numbers of Protestants border areas with large numbers of Catholics are where tensions have been at their highest, these include areas such as the Shankill and Falls Roads neighbourhoods of West Belfast.

Figure 7.16 shows the segregation between the two groups, but do they have anything to fear from each other? In some cases, at some times they do (McEldowney *et al.* 2004). Are these two groups of people demographically different? Or are they just split on religious grounds? Is the sectarianism seen in Northern Ireland the narcissism of small differences, rather than a significant demographic divide? The OA classification can give us an insight into this, by profiling the percentage of people of each religion by the classification. Demographic, social, economic and lifestyle differences between the members of the two faiths can be either shown or disproved (Pool and Boal 1973). The OA classification is ideal for doing this as it does not contain data about religion.

Figure 7.16: The distribution of Catholics in Northern Ireland, 2001 Census OA level

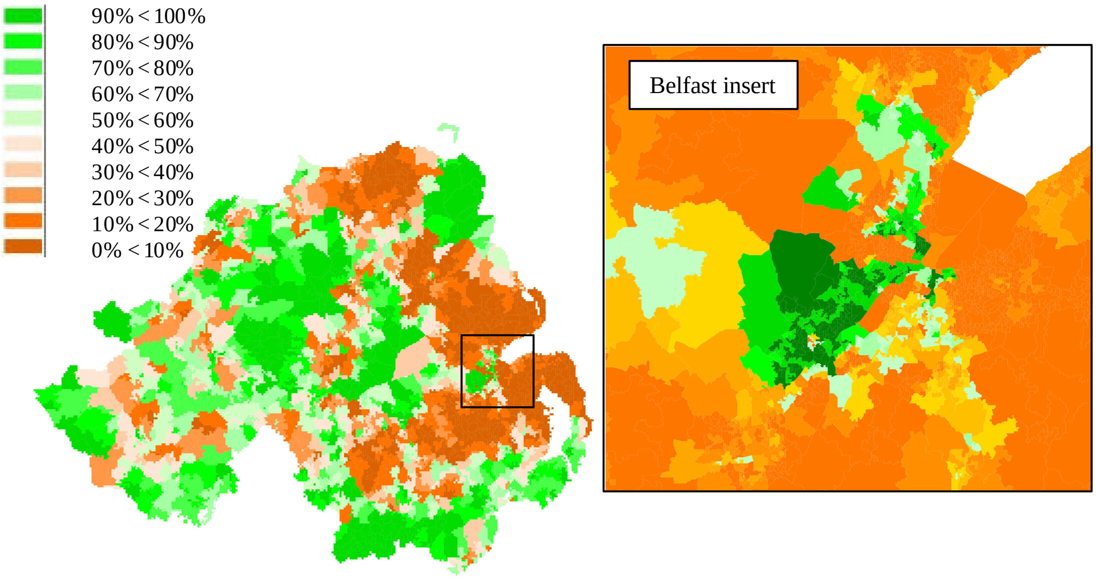


Figure 7.17 shows the distribution of the religious background of the people of Northern Ireland by Super-group type. At this level there are no real extremes where one group is dominated by people of either Catholic or Protestant background. Super-group 7 *Multicultural* does seem to show a significant relationship to people of Catholic background but because of the small number of OAs of this super-group type in Northern Ireland (only 3), it would be foolhardy to make claims that there is a definite relationship. However, there are some interesting deviations from the mean. Catholics are overly represented in *Blue Collar Communities* (+8.6%), *City Living* (+7.2%), *Countryside* (+3.5%), whereas Protestants are overly represented in *Prospering Suburbs* (+7.0%), *Constrained by Circumstances* (+5.8%) and *Typical Traits* (+5.2%). Catholics are overly represented in *City Living* (+7.2%) and *Multicultural* (+13.6%). Despite

these differences there does not appear to be a clear pattern. The evidence suggests that Catholics concentrate in both densely urban and intensely rural areas. This proposition is borne out by Figure 7.16, which shows Catholics to be prevalent in the more rural west of the country, but also in the centre of Belfast. Protestants are more likely to be prospering in the suburbs, which can be seen in Figure 7.16 as they dominate the suburbs and small towns around Belfast. However, Protestants are also more likely to be living within constrained circumstances. What this suggests is that, despite society being divided on religious grounds this divide does not propagate much further into society. The OA classification does show that there are some differences between Protestants and Catholics, but there is no evidence that these differences result in a vastly different social make-up within the two communities. The demographic differences between the two groups are perhaps easing from the position of the past (Anderson and Shuttleworth 1998).

Figure 7.17: The Distribution of Religious Background in Northern Ireland by Super-Group

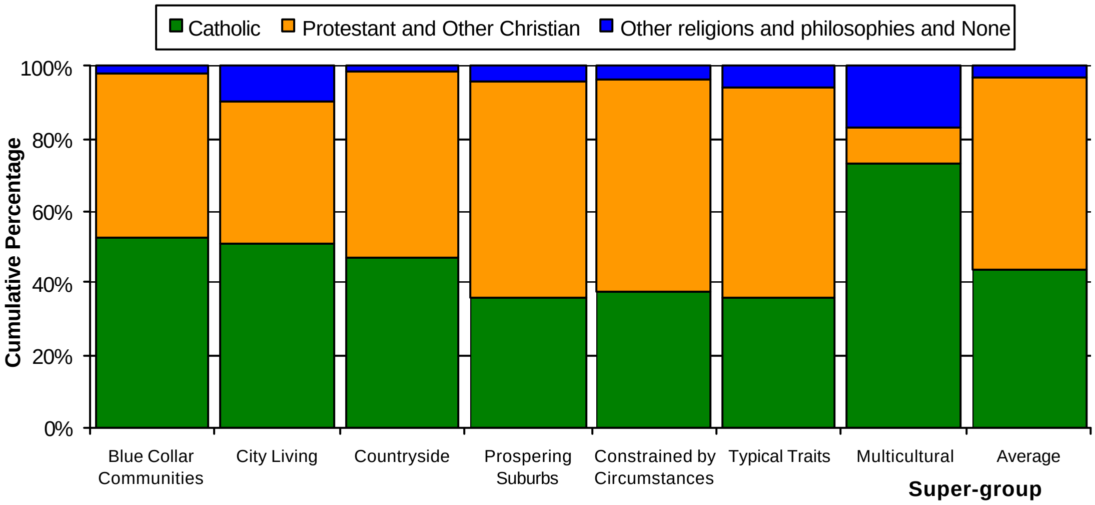
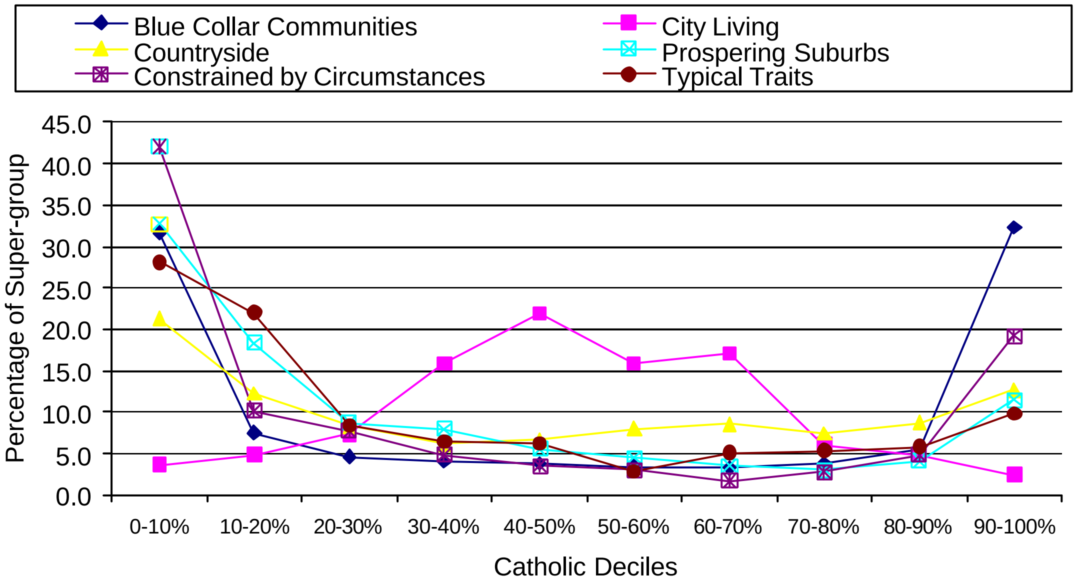


Figure 7.17 gives a very broad overview of which super-group types make up the Protestant and Catholic communities, but they can be examined in greater detail. Figures 7.18 and 7.19 show the percentage of each super-group type that is accounted for by each decile of the distribution of the Catholic (Figure 7.18) and Protestant (Figure 7.19) religions. The two religions appear to generally show a similar pattern, but with closer inspection some differences can be seen. Figure 7.18 shows a U shaped pattern for all, but one of the super-groups. *City Living* shows a more normal distribution. This is undoubtedly due to the more cosmopolitan nature and greater mixing that is present in central Belfast, where students and recent immigrants live. The U shaped nature of the graphs shows polarisation of the two sectarian groups. They are bunched at either end because they mainly live in areas with people of the same faith. In Figure 7.19, the values for the 0-10% decile suggest that these are Catholic areas and that in the 80-90% and 90-100% deciles these are Protestant areas. The two graphs suggest that the polarisation in the Catholic population is slightly greater than that in the Protestant population. This can be seen with the rise in numbers of each type visible as early as the 60-70% decile for the Protestant population, but not until the 90-100% decile for the Catholic population. Both social and

religious/communal polarisation is present in Northern Ireland (Graham and Shirlow 1998). The classification shows that these two forms of polarisation don't match up, but cut across each other.

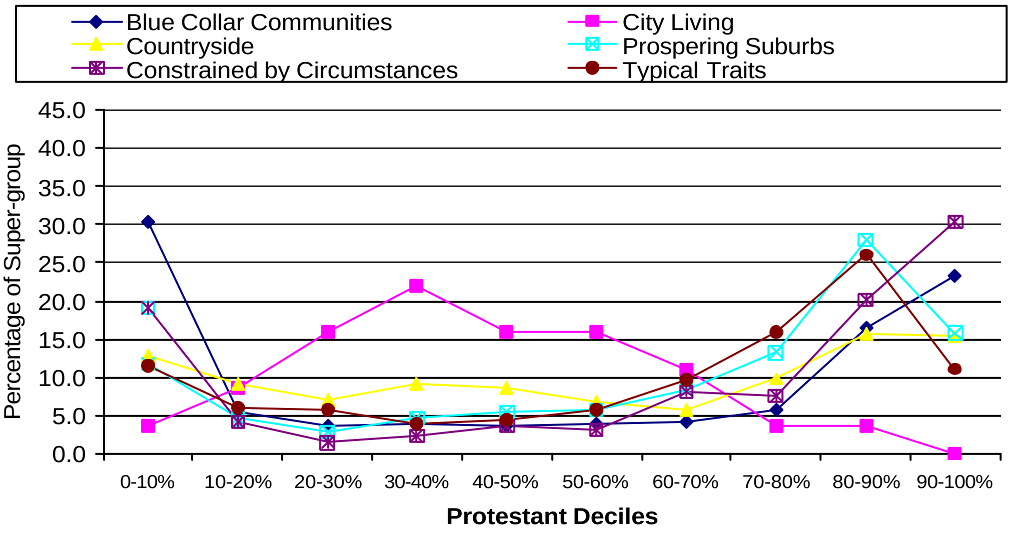
Even greater detail about the contrasting levels of polarisation in Northern Irish society can be drawn out. The super-groups show that the extent of the polarisation also differs by social grouping. Figure 7.19 shows those super-groups which could be traditionally described as working class (*Blue Collar Communities* and *Constrained by Circumstances*) have the highest values at the extremes of the scale and the lowest values at the centre of the scale. This suggests that these super-groups are the most polarised on religious grounds. This contrasts with the super-groups which can be seen to represent the traditional middle classes (*Typical Traits*, *Prospering Suburbs* and *Countryside*). These areas are not polarised to the same extent. They have lower values than the working class areas at the extremes of the scale and higher values towards the centre of the scale suggesting that on the whole they are slightly more mixed and less polarised; this pattern can also be seen for Catholics in Figure 7.18.

Figure 7.18: The percentage of each super-group type by Catholic deciles*



*Multicultural excluded due to small number of OAs (only 3 in Northern Ireland)

Figure 7.19: The percentage of each super-group type by Protestant deciles*



*Multicultural excluded due to small number of OAs (only 3 in Northern Ireland)

An integrated society on the grounds of religion would show a normal distribution like that of *City Living* for all of the super-groups. Although the classification cannot predict the religion of a person based on the super-group of the area in which they live, the classification has led to a much greater understanding about the intricacies of the distribution of the residential population in Northern Ireland. Figures 7.16, 7.18 and 7.19 give a very clear indication of how polarised residence of people is based on their faith group. Figures 7.18 and 7.19 show that there is one section of Northern Irish society that is not residentially polarised in terms of religion and that is the *City Living* super-group. This leaves a clear message that if greater integration is to be seen within the residential patterns in Northern Ireland the lesson that has to be learned is look to the *City Living* areas. If residential integration is happening there, can it happen in the rest of the country?

Mapping of Catholics and Protestants in Northern Ireland has shown that they display clear spatial divisions in terms of residential location. The OA classification shows that the extent of the polarisation is very much linked to status within society or class. Those super-groups representing working class groups (*Blue Collar Communities* and *Constrained by Circumstances*) are most polarised whereas middle class groups (*Prospering Suburbs* and *Countryside*) show social polarisation to a lesser extent. In contrast to other super-groups *City Living* shows itself to be representative of an integrated society.

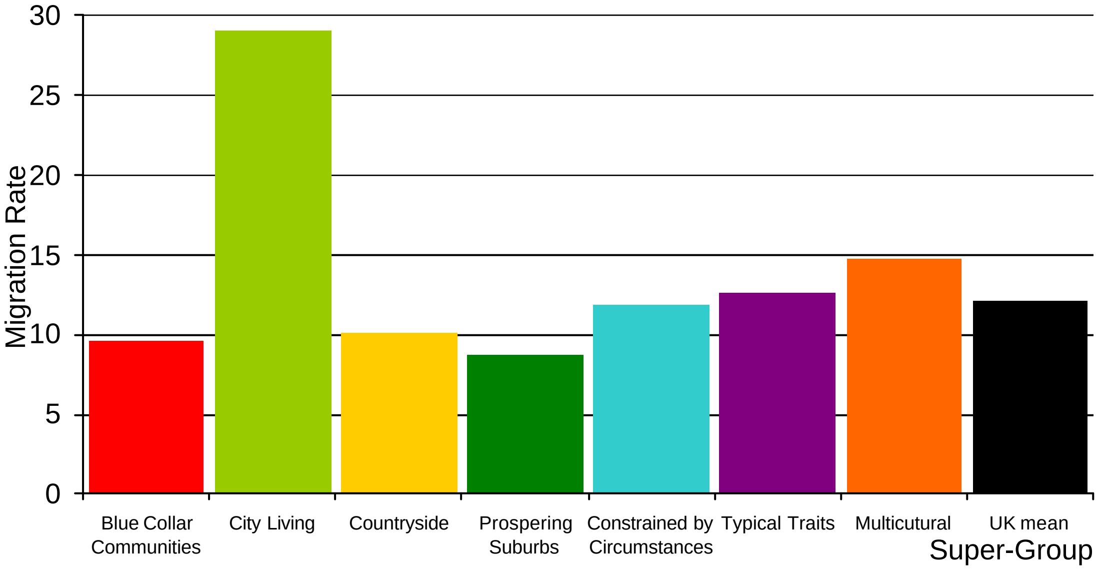
7.9 Accounting for Migration Rates Across the UK

Migration is the biggest component in population change in the UK (Champion *et al.* 1996) and is therefore vital in the understanding of the demographics of the country. Different types of people migrate and for different reasons. People's reasons to migrate can be many and varied and dependent upon a number of socio economic factors (Jones 1990). However, it has long been widely accepted that age is the most important characteristic for distinguishing migrants from non-migrants. In Thomas' (1938) review of migration differentials the only variable that consistently stood up in all contexts was that young adults were more migratory than all other groups. Although age can account for more variability than any other variable in a migration dataset that does not mean that other factors should be ignored (Boyle *et al.* 1998). The seminal work by Rossi (1955) linked migration to the life-cycle in an attempt to answer the question *Why Families Move?* Rossi concluded that migration of households was based around five reasons linked to a family's life cycle, stating that, the creation of a new household, circulation of existing households, mortality, household dissolution and moves related to work were the primary reason behind the movement of the family. Several people such as Hohn (1987) have added to the work of Rossi, by recognising the need to extend the explanation. Further work on the links between migration and life course transitions was put forward by Warnes (1992) who set out ten reasons for migration, as follows: Leaving parental home, sexual union, career position, first child, career promotion, divorce or cohabitation or second marriage, retirement, bereavement, income collapse and frailty or chronic ill health.

Modern studies of migration use a list of carefully selected determinant variables to predict the rate of migration for any given country or area (Champion *et al.* 1998; ODPM 2002). Studies such as Fotheringham *et al.* (2004) have shown that a list of socio-economic variables can be highly predictive of the likelihood of people from a particular area to migrate. As migration data were not included in the variables that went into the classification, it is possible to use the OA classification to assess the likelihood of someone from a particular area to migrate. By doing this a different rate of migration will be returned for each of the clusters, which will reveal to what extent the classification can reveal variation in rates of migration across the clusters. This can be done for all three levels of the classification. The migration indicator that will be used is the migration rate, that is, the percentage of the population of each OA that have moved in and out of each OA in the twelve months prior to the census. This is a measure of population turnover.

Figure 7.20 shows UK migration rates profiled against the classification at the Super-group level. The thing that stands out most clearly from this is the dominance of super-group 2 *City Living*; its migration rate in the year leading up to the 2001 Census was just short of 30%. Therefore just under 1 in 3 of the people who live in this super-group moved into their current home within the twelve months leading up to the 2001 Census enumeration. The rate in the *City Living* super-group is double that of the rate of the next highest which is super-group 7 *Multicultural* with a rate of just under 15% or one in every 6.5 people.

Figure 7.20: Profiling UK migration rates against the Super-group level of the classification



The super-group that has the lowest migration rate is 4 *Prospering Suburbs* with a rate of just under 9% or one person in eleven. If we relate back the results with those of Rossi (1955) it is clear to see why the *City Living* super-group has such a high rate of migration. The age structure in *City Living* is dominated by young adults, who have long been recognised as the most mobile group in society. The *Multicultural* super-group that has the second highest rate also has a relatively young age structure and is geographically the closest to *City Living* both having a dense urban setting

Figure 7.21 shows UK migration rates profiled against the classification at the Group level. Unsurprisingly 2a *Transient Communities* and 2b *Settled in the City*, that are sub-divisions of super-group 2 *City living*, show the highest rate. However, the effect of the disaggregation can be seen, *Transient Communities* shows an increase on the rate that was shown by *City Living* by 5% in contrast *Settled in the City* shows a reduction in the *City Living* rate by 5%. *Transient Communities* has the highest percentage of students who not only in the main are young adults, but are even more likely to move to be close to their place of study. This creates a constantly moving and changing population as they start and finish their studies. Other differences can be seen within the groups that make up each super-group. Super Group 4 *Prospering Suburbs* breaks down into 4a *Prospering Young Families*, 4b *Prospering Older Families*, 4c *Prospering Semis* and 4d *Thriving Suburbs*. Super-group 4 showed the lowest migration rate, but when broken down there are some clear differences between the constituent groups in terms of their migration rates. Group 4a has a rate of around 13% which is an increase on the 9% shown by its super-group. In contrast 4c has a rate of only 6.5% a reduction from the super-group and only half that of 4a.

Figure 7.21: Profiling UK migration rates against the Group level of the classification

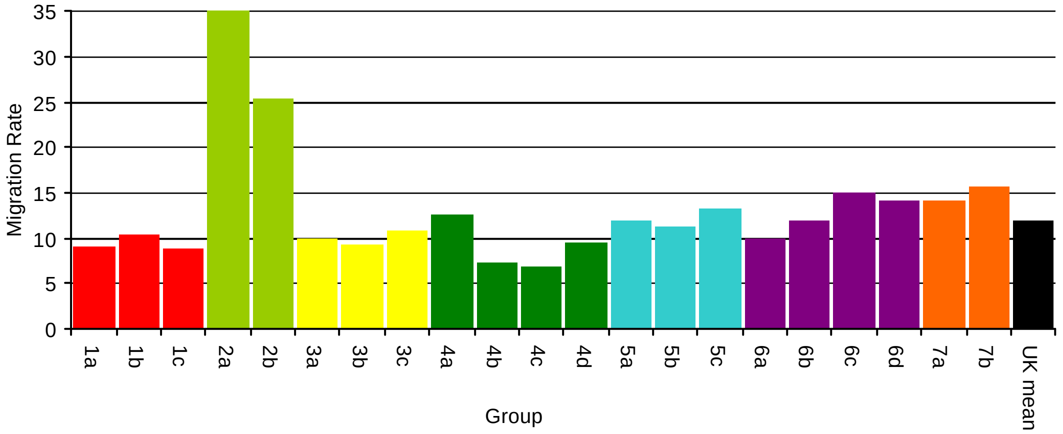
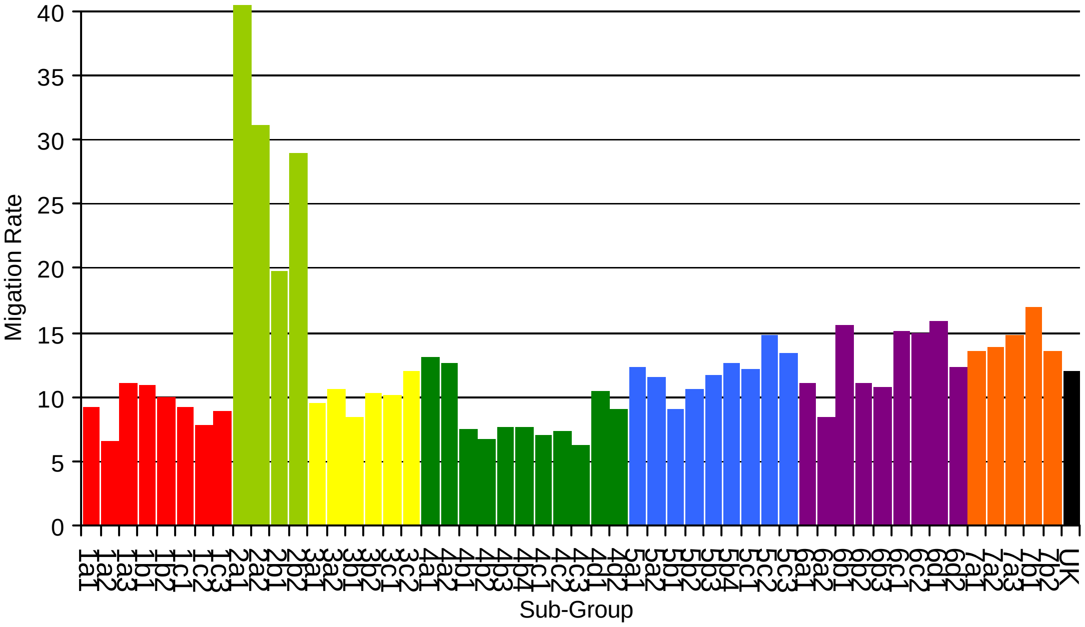


Figure 7.22 shows UK migration rates profiled against the classification at the Sub-group level. The most obvious feature at this level is still the dominance of super-group 2, at this level broken down into four sub-groups, 2a1 *Transient Communities (1)*, 2a2 *Transient Communities (2)*, 2b1 *Settled in the City (1)* and 2b2 *Settled in the City (2)*. The rate of 2a1 has seen an increase on the group level and moved up to 41% while the rate for 2a2 has fallen away from the rate of the *Transient Communities* group to 31%. 2b1 and 2b2 show a similar pattern one increasing and one decreasing at the lower end of the scale. This pattern can be seen to some extent with the breaking down of all the groups into sub-groups but it is more marked when the rates are greater. The rate for 2a1 is double that for 2b1, whereas at the super-group level they had the same value.

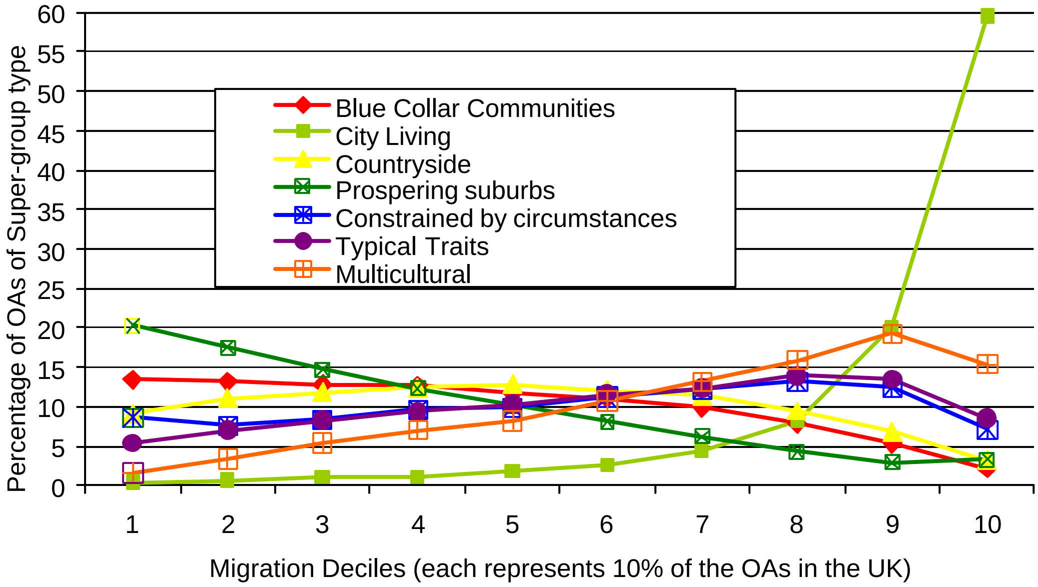
Figure 7.22: Profiling UK migration rates against the Sub-group level of the classification



How successful was the classification at accounting for rates of migration? It is clear that the classification showed a great deal of success in accounting for the variation in migration rates (remember the classification contains no migration data). There are two reasons for this assertion. Firstly, the classification shows significantly different migration rates for different clusters. Figure 7.20 shows a range of rates from 6% to 41%. Secondly, the rates changes with movement down the levels of the hierarchy, for example super-group 2 has a rate of 28%, group 2a has a rate of 35% and sub-group 2a1 has a rate of 41%. This shows clear evidence that the classification is discriminates between areas in terms of rates of migration.

It is possible to disentangle the migration rates by the OA classification even further. Figure 7.20 illustrates the differences in the rate of migration between OA super-group types, but what about within each super-group are they normally distributed or are the values skewed within each group? Figure 7.23 shows the distribution of migration rate deciles by OA classification super-group types. Each decile represents 10% of the OAs in the UK, the first decile represents the OAs with the lowest migration rate and the tenth deciles represents the OAs with the highest migration rates, each decile in between represents 10% of OAs with increasing rates of migration. Perhaps unsurprisingly the most striking thing about Figure 7.23 is the *City Living* super-group, 60% of the OAs in the are in the tenth decile of migration rates, 88% of *City Living* are in the top three deciles of migration rates. *City Living* is by far the most skewed of the super-groups in terms of migration rates with an extreme negative skew. *Prospering Suburbs* shows the opposite pattern to *City Living* showing a positive skew, but not to the same extent. 20% of the OAs in the *Prospering Suburbs* super-group are in the first decile of migration rates, 53% of *Prospering Suburbs* are in the first three deciles of migration rates. *Blue Collar Communities* and *Countryside* also display a slight positive skew, *Multicultural* and *Typical Traits* display a slight negative skew as does *Constrained by Circumstances*, which also has the most normally distributed set of values.

Figure 7.23: The distribution of migration rate deciles by OA classification super-group type



The OA classification has been shown to be able to discriminate between rates of migration, the classification agrees with Thomas (1938) who ascertained that age, especially related to young adults, was the best way of differentiate between migrant and non-migrants. Cluster types can be linked to migration life course transitions. The transitions made will depend on where in the social hierarchy you start, but some clusters represent certain movements more than others. Responses from consultation exercise said that the classification could be used to map people's life history and movement through the life course; this could also be applied to the Rossi (1955) and Warnes (1992) by linking of migration to the life-cycle. The classification could be used to track the migration of people through time and determine the nature and reason for the move.

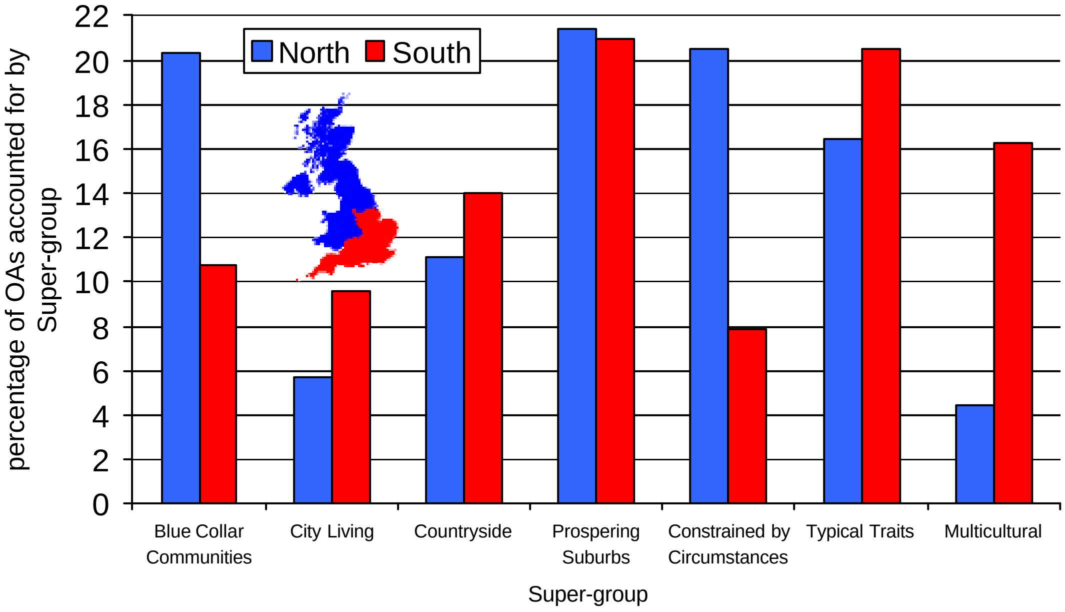
7.10 Reopening the North-South Divide

The north-south divide is popularly thought to originate with the Industrial Revolution, although it is in fact rooted in prehistory and is perhaps responsible for what can be seen as a northern consciousness within the population (Jewell 1994). The marked differences between the north and south of the country are not a new observation and as a consequence are often seen as unfashionable and out of date. However, it has been widely regarded that England was socially divided down simple geographical lines for most of the twentieth century (Dorling 2004). The north-south divide has relevance and can be shown to exist in many spheres. Blackaby and Manning (1990) found that long-term unemployed are disproportionately concentrated in the north, having an effect not only on those that are unemployed, but on the region as a whole. In some contexts divide can be shown to be widening. Johnston and Pattie (1989) identified increasing polarisation in terms of voting patterns in general elections throughout the 1980s. Despite election campaigns with an increasing national focus the population was focusing more and more on regional and local issues, resulting in increasing spatial variation in support for the major political parties (Johnston and Pattie 1989). Health inequalities can be shown to exist between the north and south; even among people in the same social class the risk of ill health varies greatly by where they live. A northwest-southeast divide can be seen in social class inequalities in health (Doran *et al.* 2004).

It is clear that the north-south divide can be shown to exist in many domains; inequalities persist between the north and the south. The OA classification can be used to examine whether the north and south divide persists in terms of social areas. By splitting the classification along a line separating the north and the south, disparities in the distribution of social areas can be examined. The north-south divide traditionally follows a line from the Bristol Channel to the Wash, but for the purposes of simplicity for this analysis, a pseudo north-south divide using GORs will be used. The north is considered to be made up of the West Midlands, Wales, Yorkshire and the Humber, the North West, the North East, Scotland and Northern Ireland. The South is considered to be composed of the South West, the South East, London, the East of England and the East Midlands.

Figure 7.24 shows the north and south stratified by the OA Classification, the most numerous super-group type on both sides of the divide is *Prospering Suburbs* representing about 21% of areas on both sides of the divide. However, this is where the similarities end. *Blue Collar Communities* are almost twice as prevalent in the north as the south, but it is in *Constrained by Circumstances* where real inequalities can be seen with the prevalence in the north almost three times that of the south. The prevalence of *Multicultural* is four times greater in the south than the north. *City Living*, *Countryside* and *Typical Traits* are also more prevalent in the south than the north.

Figure 7.24: The north and south stratified by the OA Classification

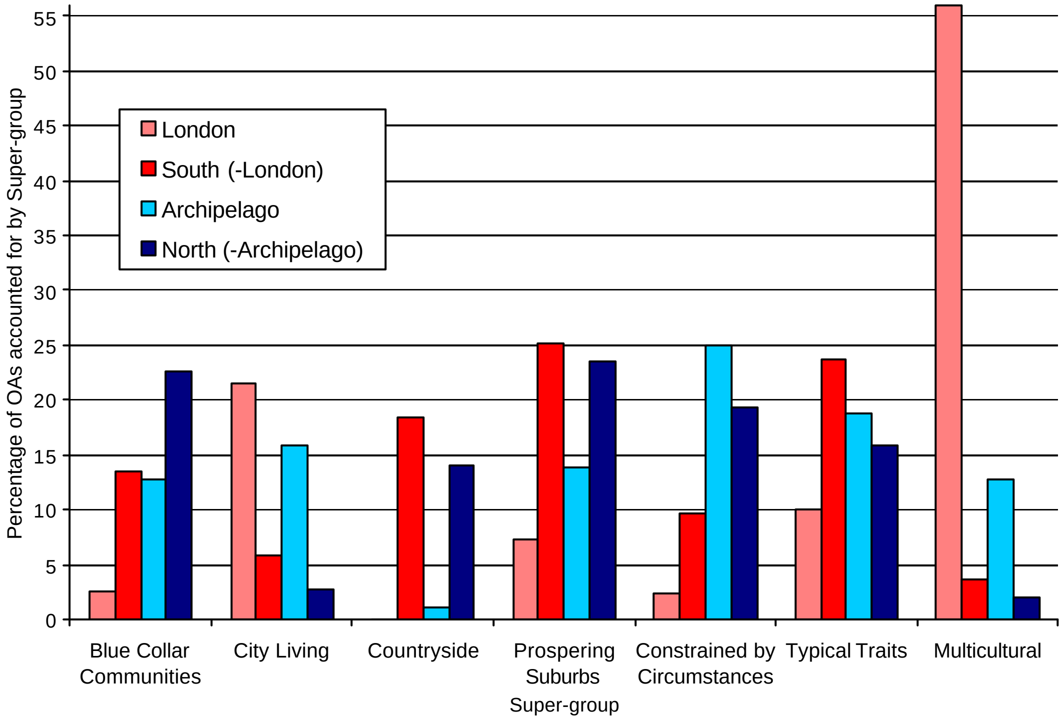


There are clear differences between the north and south in terms of social areas, with the north displaying a greater prevalence for the less affluent super-groups therefore showing a concentration of poorer communities north of the divide.

Dorling and Thomas (2004) finish their 2001 Census atlas of the UK 2001 with a figure titled ‘London and the Archipelago’ an alternative look at the increasing dominance of London in the south of England and the north-south divide within the UK, suggesting that the ‘country is becoming more and more divided’. Dorling and Thomas (2004) see the entire south of England as the suburbs of an extended Greater London; the north contains the Archipelago, the core of which is made up of the main urban centres (Edinburgh, Glasgow, Newcastle, Leeds, Sheffield, Kirklees, Calderdale, Bradford, Oldham, Tameside, Manchester, Salford, Liverpool, Birmingham, Belfast and Cardiff) above the divide. The core of the Archipelago roughly equate to Greater London in terms of population. By removing London and the Archipelago from the south and north respectively we find an even more divided society.

Figure 7.25 shows London and the Archipelago stratified by the OA Classification, what is immediately obvious is how the *Multicultural* super-group dominates London accounting for over 50% of the OAs in the capital. The rest of the south is dominated by the more affluent super-group types with *Countryside*, *Prospering Suburbs* and *Typical Traits* making up over two thirds of its OAs. The effect of disaggregating the Archipelago from the north is not as great as removing London from the south. However, there is a contrast between them most clearly seen in the *Countryside* and *Multicultural* super-groups. The Archipelago shows the greatest value for the least affluent super-group *Constrained by Circumstances*, but the rest of the north shows the highest value for the next worst off group *Blue Collar Communities*.

Figure 7.25: The London and the Archipelago stratified by the OA Classification



The OA classification has shown that there is a clear north-south divide in terms of social areas, with less affluent Super-group types over represented north of the divide. When London is considered separately to the rest of the south of England these inequalities become even greater. With the continued 'brain drain' from the north of the UK to South East England it is difficult to see how these disparities will be addressed.

7.11 Conclusions

The examples in this chapter outline the many and diverse uses of the OA classification, and its great value and relevance to many aspects of social science, public policy and geographic thought. The case study examples used in this chapter show how the OA classification can be used to account for differences between different groups of people, different areas. By splitting a dataset using the clusters in the OA classification, a large amount of variance within the dataset can be removed.

The focus on Leeds shows clear and distinctive residential patterns within the city of Leeds; the distribution of OA super-group types within the city closely replicates recognised patterns of residential structure (§ 7.2). The distribution of OA super-group types shows both diversity and homogeneity within a community specified geography of the city.

A tale of eight cities shows how the core cities of England (§ 7.3), show significant differences in their social make-up when analysed using the OA classification. Although the cities are of a similar size in terms of their populations the OA classification shows how each one is distinctive.

The Swingometer analysis (§ 7.4) shows clear differences can be seen in voting patterns and the changing nature of voting patterns by OA super-group. The reduction of the Labour vote between the 2001 and 2005 general elections is much more pronounced for the *Multicultural* super-group that the other clusters.

Out in the country air (§ 7.5) shows that perhaps the official urban-rural classification lacks a bit of context and the urban category needs to be broken down further. When cross-tabulated with the OA Classification it appears that little social diversity exists with the most rural areas and that urban areas contain great social diversity.

When the OA classification is cross-tabulated with the Indices of Multiple Deprivation (§ 7.6) it is able to replicate the general trend of deprivation show in the IMD. Most of the component indices show the same pattern, but there are some interesting deviations

The analysis of the Welsh language by the OA classification (§ 7.7) shows clear differences in the extent to which the Welsh language is spoken by super-group type. Those people living in the *Countryside* super-group are far more likely than those in other clusters to speak Welsh.

Demography across the divide (§ 7.8), an analysis of the sectarian divide in Northern Ireland has shown that the two groups are polarised in terms of their residential location, apart from the *City Living* super-group, which displays a more integrated section of society.

The migration rate analysis shows a clear difference between migration rates by cluster type (§ 7.9). The *City Living* super-group and *Transient Communities* group are shown to have the most mobile populations. This funding can be linked to the age structure of these clusters which is represented by significant numbers of young adults.

The classification shows that there is definitely a north-south divide in terms of social areas (§ 7.10), with an over representation of less affluent clusters in the north of the country and more affluent clusters in the south of the country.

The OA Classification enables busy researchers to investigate the socio-geographic context of their research questions easily. The classification summarises the essence of a very large and complex dataset and so saves the researcher from having to undertake their own detailed socio-geographic analysis. The relationships revealed through the use of the classification can be further investigated in detail. The classification can serve as a hypothesis generation method, particularly when combined with mapping.

Chapter Eight - A Multi-scale Integrated Classification System: Investigating Diversity within Area Classifications

8.1 Introduction

The National Statistics geodemographic project has created classifications of the UK at three different geographies. The ONS team have produced classifications at the Local Authority and Ward levels and this project has produced classifications at the Output Area level. These geographies are hierarchical and easily comparable as OAs fit within wards and wards fit within local authorities. Note that the LA classification used in this chapter is the official ONS classification and not the one described in Chapter 4. The aim is to create a hierarchy using all the official ONS classifications.

The final investigation that needs to be conducted after creating a hierarchy of multi-scale area classifications is to investigate the diversity within the cluster types at each level. The purpose of area classifications is to examine diversity within and between areas, so the use of different scales not only enables the examination of diversity between areas, but also within them. Areas which are homogenous at one scale can be seen to contain great diversity at another.

8.2 Why do we need a Multi-scale Classification System?

The effect of scale on the analysis of data pertaining to the human population is a phenomenon that has long been recognised. Gehlke and Biehl (1934) demonstrated the variability of statistical results from the use of data at different scales. It is no secret to those who regularly use multi-scale datasets that conclusions derived from analysis at one scale are specific to that scale and cannot necessarily be applied with any confidence to any other scale (McCarthy *et al.* 1956). Many real world phenomena are interesting precisely because they exhibit different behaviours at different scales. Webber (2004) recognises that different phenomena operate at different scales *"there is no optimal scale for classifying neighbourhoods. Consumer behaviour within some product categories is better predicted using demographic data for areas more* *geographically extensive than Census output areas, while for others the appropriate granularity is as low as unit postcodes"* p219. Thresholds can be identified that correspond to identifiable levels within a hierarchical system. When investigating complex phenomena it is essential to understand how processes operate at multiple spatial scales and how they can be linked. To observe and study a phenomenon most accurately, the scale of analysis must match the actual scale of the phenomenon. An understanding of the effects of scale and aggregation on statistical results is essential (Marceau 1999). There are several dangers in using data at one scale to make inferences about phenomena at other scales this includes a good understanding of the MAUP and ecological fallacy as outlined in Chapter 2.

Most work is conducted at what can be termed the 'available scale' these are the units that are present in the data, which are not necessarily fundamentally meaningful, but used because they are the only units for which data exist. Some work such as Alder *et al.* (2005) gets round this problem by using data at the household or person level allowing aggregation to any geography, but this is far from the norm. Using the 'available scale' can cause many problems of analysis and poor results can arise. However, this is generally unavoidable due to the lack of availability of data at any other scale.

Analysis using census data is already limited by the administrative boundaries at which the data is published. These boundaries are not necessarily meaningful and are affected by the MAUP (Openshaw 1984). A tria l and error approach to try and identify at which scale a phenomena should be analysed is often necessary. A multi-level integrated system allows a choice between scales to be made. Although none of the scales may be an ideal solution, providing a choice of scales offers an opportunity for a researcher to make a choice turning the available scale into available scales.

The title of this thesis is *"multi-level integrated classifications based on the 2001 Census"*. The multi-scale nature of the project is a key feature of the research. Previous chapters have shown the great diversity which exists within the UK and how geography plays an integral part to the diversity and character of an area. Examples of all cluster types can be found in most regions, but some are concentrated in particular regions. A classification at a single scale can show diversity between recognised and defined areas, but these classifications do not have to be used independently. They can be joined together in a multi scale classification system.

Voas and Williamson titled their 2001 critique of geodemographic classification "the diversity of diversity", questioning the value of general purpose classifications as diversity can be seen within clusters of supposed similarity. In the academic community it is not in doubt that these variations exist, a point made by Harris (2001) in response to Voas and Williamson (2001a).

There is a school of thought that says linking data together at different scales should not be done as the ecological fallacy will cause each scale to disprove the others and fatally flaw any findings from the research (Openshaw 1984). However, it is surely time to use the ecological fallacy to the researcher's advantage and view them as ecological truths that can add contextual information from one scale to another. Knowing the context or the mix of areas which surround an area or reveal themselves at a larger scale can actually aid our understanding of what is happening rather than being something to fear. Context is important in any form of analysis, but in geographical investigation it is especially so. Areas that fall within the same group will by default contain broadly similar characteristics. However, the context of the area or the properties of the surrounding areas can reveal significantly more information about the area in terms of form and function.

Why link classifications at a different geographic scale together? This is a question which is central to the way as geographers we understand how the world works. Appropriateness of scale is vital; some things are better classified at different geographic scales. By linking the classifications at different scales together this makes this decision easier to make for users. Too few studies are carried out at multiple scales although recently they have become more numerous.

The aim of this chapter is to demolish the misapprehension that classifications at different scales contradict each other. Researchers have often been fearful of the ecological fallacy. Reading Monmonier (1996), it is easy to see why, the ecological fallacy is an easy trap to fall into. However, a researcher who is fully aware of the problems that the ecological fallacy poses should not only be able to overcome it, but also be able to use it to their advantage. Different things shown at different scales does not mean that one or other scales has produced a wrong answer, only that different processes are visible. By conducting analysis at multiple scales a researcher not only ensures they are aware of the possibility of the ecological fallacy, but also that a far greater amount of information can be revealed. To get a true picture of what is happening, all scales must be examined, but not in isolation. To be used effectively they must be joined together and form part of the same system where the multiple levels can easily be compared and contrasted.

8.3 The Importance of Scale in Assessing Diversity

Because we can see different things at different scales, it is important to choose the most appropriate scale for each specific purpose (Monmonier 1996). This is true for all forms of areal analysis, but especially so for area classifications where areas are not only given values based on data, but they are also given names and descriptions based on those values. Therefore using an area classification which is at an inappropriate scale for purpose could not only produce poor analysis, but may also cause offence.

An example of this is the occurrence of wholly London groups in the ONS LA Classification. This does not occur in the OA Classification because as the scale changes things become less aggregated. An area such as Leeds at the LA scale is close to the national average as its boundary encloses both urban and rural areas. At the OA scale, parts of Leeds have mixes of OA types that resemble the mixes in Inner London. By joining the three classifications (LA, ward and OA) together several things can be examined.

1. What is the effect of scale on the area classification process?
    a. Do different things appear at different scales?
    b. Does the ecological fallacy affect the classification?

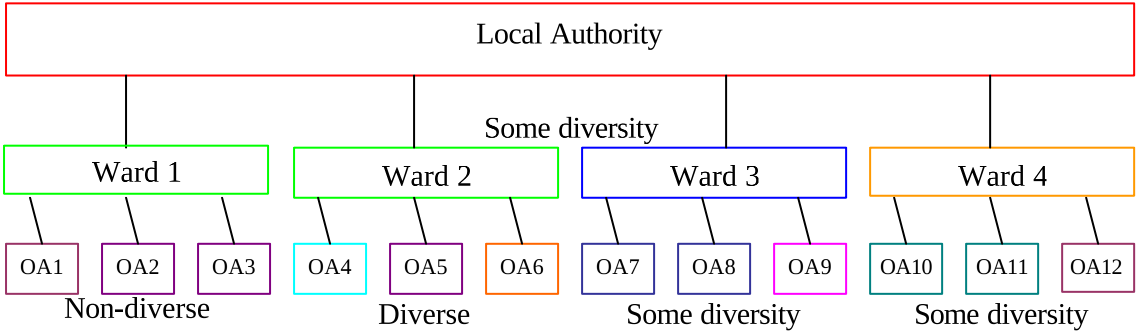
2. Are some cluster types more diverse than others?
    a. Do some cluster types show great diversity within them?
    b. Do some cluster types show great homogeneity within them?

3. What does the diversity within different types tell us about them?

4. What is the most appropriate scale to examine different types of areas?

It is easy to talk about assessing diversity, but can it be easily recognised if it is there? Figure 8.1 gives an indication of what may or may not be considered as a sign of diversity. A theoretical local authority is shown containing a variety of ward types, and in-turn the wards show a range of OA types within them, each different colour representing a different cluster type. The local authority shows some diversity with the four wards within it represented by three different cluster types. Ward 1 shows no diversity as it is made up of OAs of all the same type. Ward 2 shows diversity as all the OAs within it are of different cluster types. Wards 3 and 4 show some diversity with the three OAs within each made up of two different cluster types. This simple example shows how the diversity with an area can be examined using an area classification of a smaller geographic scale. To investigate the diversity within the classifications individual local authorities or wards will not be investigated, but the agglomeration of local authorities or wards that make up each cluster type.

Figure 8.1: What constitutes diversity within areas



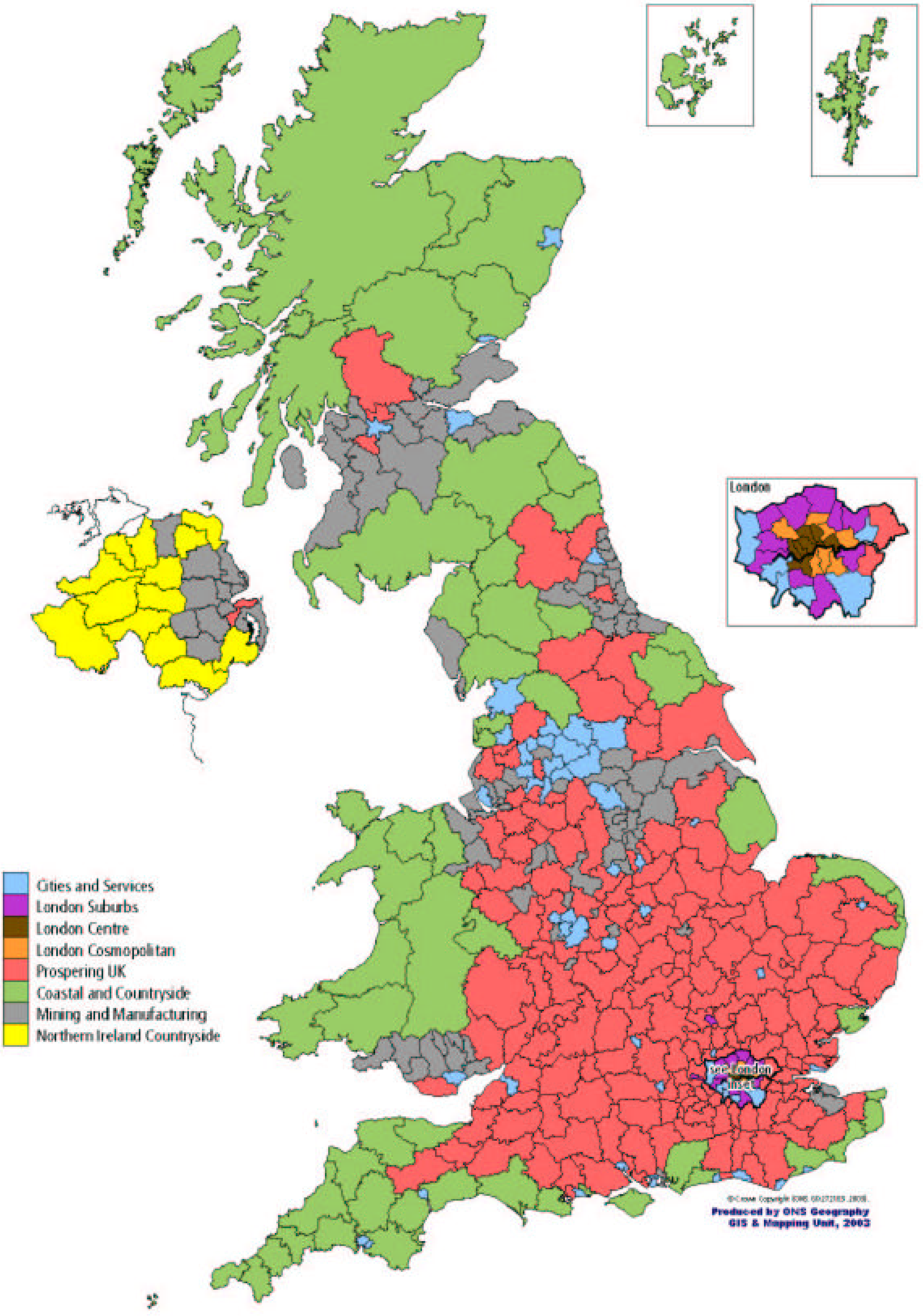
8.4 The ONS Local Authority and Ward Classifications

Along with the OA classification, as part of a wider project, classifications have been created for three other geographies by ONS. Classifications exist for Electoral Wards, Local Authority Districts and Health Board Areas. This chapter will see how the OA classification fits within and complements the ward and local authority level classifications. Before this can be done the classifications need to be described and understood.

8.4.1 The ONS Local Authority Classification

There are 434 local authorities in the UK which are classified into a three tier hierarchy of 8 super-groups, 13 groups and 24 sub-groups. The classification was constructed from 42 census variables and was produced using Ward's hierarchical clustering procedure. All data were extracted from the Key Statistics tables. One quirk of the classification is that it actually clustered only 432 local authorities, because the City of London was merged with City of Westminster and Isles of Scilly was merged with Penwith. City of London and Isles of Scilly were considered to have a too small population to be clustered on their own (ONS 2005d). The ONS LA Classification had a very similar methodology to the alternative LA classification described in Chapter 4. The two classifications were compared in § 4.7. Figure 8.2 shows how the Local Authority Classification maps at the super-group level.

Figure 8.2: A Map of the ONS Local Authority Classification at the super-group level



Source of map: http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/la/downloads/kmean8Supergroup.pdf

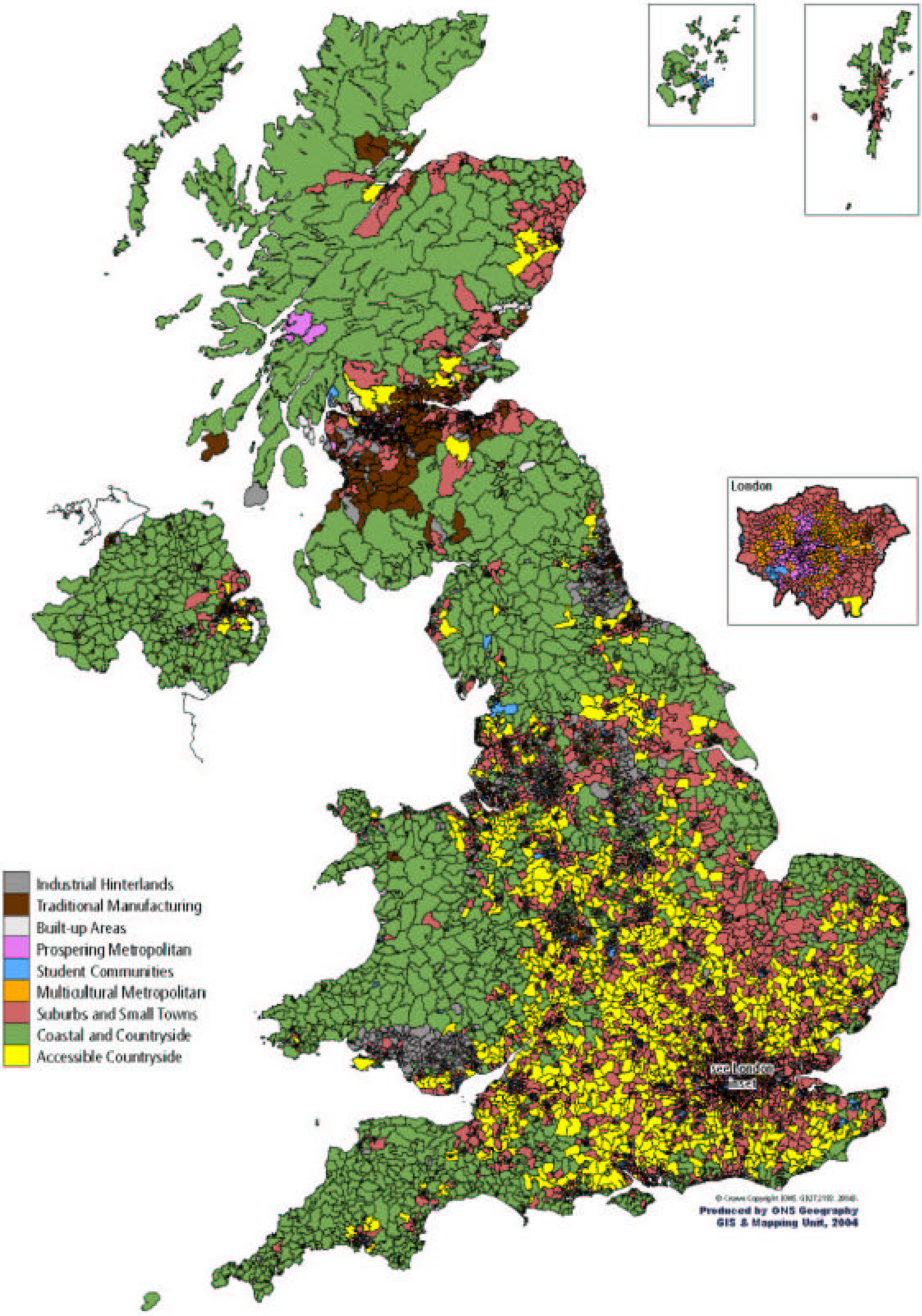
8.4.2 The ONS Statistical Ward Classification

A statistical ward is a ward where the minimum population is 1,000 people. Wards with a population of fewer than 1,000 people were merged with a neighbouring ward to create the statistical wards. The ward level classification covers the 8,800 statistical wards in England and Wales, 1,176 statistical wards in Scotland and 577 statistical wards in Northern Ireland (ONS 2005e).

There are 10,553 Statistical Wards in the UK which are classified into a three tier hierarchy of 9 super-groups, 17 groups and 26 sub-groups. The classification was constructed from 43 census variables extracted from the Key Statistics tables and was produced using the original methodology that was going to be used for the OA classification (outlined in Chapter 5). This involved creating 1,000 clusters using the k-means clustering procedure, and then reducing this number by running the Ward's clustering algorithm on the cluster centres that were produced by the k-means procedure (ONS 2005e). Figure 8.3 shows how the Statistical Ward Classification maps at the super-group level. The procedure consisted of the following steps:

- Generate a random classification of all wards into 1,000 clusters using the k-means method.
- Apply Ward's method to the resulting 1,000 cluster centres from k-means method.
- Determine the number of super-groups, groups and sub-groups by examining the agglomeration schedule.
- Refine the subgroups obtained from Ward's method using k-means to ensure that each ward was assigned to its correct subgroup (ONS 2005e).

Figure 8.3: A Map of the ONS Statistical Ward Classification at the super-group level

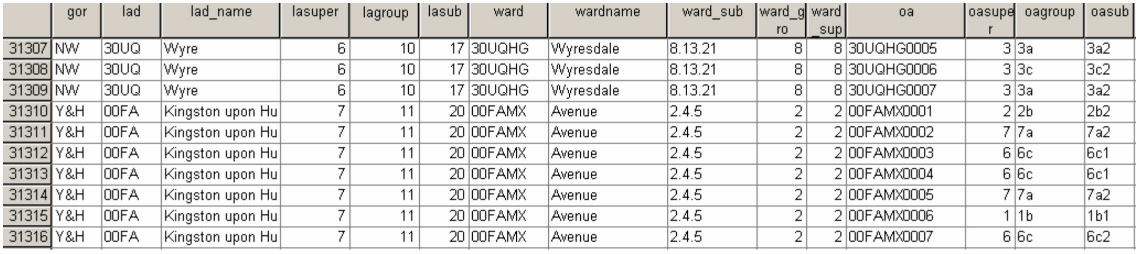


Source of map: http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/wards/downloads/wards_supergroups.pdf

8.5 Linking the Levels

The levels were joined together using linking table functions in a GIS. A database containing the membership of all areas at all levels was created. This was built from the smallest scale geography, OAs, and then the higher levels were added. The ward and local authority in which each OA is located was added to the database creating a geographical hierarchy of OAs in wards and OAs and wards in LAs. The cluster membership information for all three different geographies was added to the database. This enabled an examination of any of the geographies by a classification at another scale. For example, the ward group could be given for any OA or the percentage of each type of OA super-groups that make up a LA or LA type could be established. Figure 8.4 shows a sample section from the database containing the membership of all levels of the classification for all areas. The database contains a geographical reference to all areas and a reference to the cluster membership at all areas. Starting at the right hand side of the database we can see the reference to the OA classification membership and the OA geographical code. Moving left there are the Statistical Ward classification codes followed by the ward name and the ward's geographical code, of which the OA to the right is contained. Moving left again we have the LA classification code followed by the classification name and the LA geographical code, of which the ward to the right is contained within. To give further geographical information, the column on the far left of the database adds further geographic information by stating which GOR or country each of the smaller geographical areas are contained within.

Figure 8.4: Section from the database linking the classifications at different scales



	gor	lad	lad_name	lasuper	lagroup	lasub	ward	wardname	ward_sub	ward_gro	ward_sup	oa	oasuper	oagroup	oasub
31307	NW	30UQ	Wyre	6	10	17	30UQHG	Wyresdale	8.13.21	8	8	30UQHG0005	3	3a	3a2
31308	NW	30UQ	Wyre	6	10	17	30UQHG	Wyresdale	8.13.21	8	8	30UQHG0006	3	3c	3c2
31309	NW	30UQ	Wyre	6	10	17	30UQHG	Wyresdale	8.13.21	8	8	30UQHG0007	3	3a	3a2
31310	Y&H	00FA	Kingston upon Hu	7	11	20	00FAMX	Avenue	2.4.5	2	2	00FAMX0001	2	2b	2b2
31311	Y&H	00FA	Kingston upon Hu	7	11	20	00FAMX	Avenue	2.4.5	2	2	00FAMX0002	7	7a	7a2
31312	Y&H	00FA	Kingston upon Hu	7	11	20	00FAMX	Avenue	2.4.5	2	2	00FAMX0003	6	6c	6c1
31313	Y&H	00FA	Kingston upon Hu	7	11	20	00FAMX	Avenue	2.4.5	2	2	00FAMX0004	6	6c	6c1
31314	Y&H	00FA	Kingston upon Hu	7	11	20	00FAMX	Avenue	2.4.5	2	2	00FAMX0005	7	7a	7a2
31315	Y&H	00FA	Kingston upon Hu	7	11	20	00FAMX	Avenue	2.4.5	2	2	00FAMX0006	1	1b	1b1
31316	Y&H	00FA	Kingston upon Hu	7	11	20	00FAMX	Avenue	2.4.5	2	2	00FAMX0007	6	6c	6c2

8.6 Diversity within the Levels

The different levels of the classification were cross-tabulated to reveal the distribution of types within types. This enables an examination of how scale has affected the groups that were produced. We can examine which types are homogeneous and which show great diversity within them. Three different cross tabulations were produced: the distribution of OA super-groups within ward super-groups, the distribution of OA super-groups within LA super-groups and the distribution of ward super-groups within LA super-groups.

To help examine which clusters are the most diverse, a diversity index was calculated for each of the cross-tabulations. The Simpson index of diversity measure how diverse the mix of a set of groups are within an area (Simpson 1949). The diversity index calculates the chance of two randomly selected people differing in membership of the specified group types (in this case ward level super-group type). The index is calculated in three steps: firstly square the percent for each group, then sum the squares and finally subtract the sum from 1 (Brewer and Suchan 2001). This can be expressed as follows:

$$V_i = 1 - \sum_{e=1}^{m} (r_{ie}{}^2) \qquad (9.1)$$

where: V_i is the Index of Diversity for area *i*

r_{ie} is the row percentage for area *i* and group *e*

$e = 1, m$ indicates the start and end values of the index being summed, with all the intermediate values implicitly included.

High values of the index represent diversity within an area, while low values represent homogeneity. The minimum value of the index of diversity is 0; the maximum value depends on the number of groups in the analysis. With four groups each with values of 25% the maximum value will be 0.75 while with five groups each with a value of 20% the maximum value will be 80% and so on (Simpson 1949).

8.6.1 Wards within Local Authorities

This section investigates the diversity of ward super-group types within local authority super-groups types. This will indicate the diversity that is present within the each of the local authority super-groups. Table 8.1 shows a cross-tabulation of local authority and ward super-group types; if the super-group type is homogenous then it will be made up of wards which are largely of a single super-group type and will have a low value on the diversity index. However, a local authority super-group type that is made up of a variety of different ward super-group types and will have a high value of the diversity index. The table is coloured to aid quick analysis, the colours will be used to describe the diversity within the super-group types.

Table 8.1: The Ward Classification Super-group types that make up the LA Classification Families (percent of wards by type for LA types)

LA Classification Family types	Ward Classification Super-group types										
	Industrial Hinterlands	Built-up Areas	Student Communities	Prospering Metropolitan	Traditional Manufacturing	Multicultural Metropolitan	Suburbs and Small Towns	Coastal and Countryside	Accessible Countryside	Total	Diversity Indices
Cities and Services	19.5	18.0	11.4	8.2	12.4	6.0	18.5	5.8	0.3	100	0.85
Diverse Outer London	3.7	1.3	0.0	17.2	1.4	31.4	44.5	0.5	0.0	100	0.67
Central London	0.0	0.0	0.0	63.5	0.0	36.5	0.0	0.0	0.0	100	0.46
Cosmopolitan London	0.0	0.9	0.0	16.3	0.0	78.7	4.2	0.0	0.0	100	0.35
Prospering UK	13.6	3.1	1.2	0.2	4.2	0.3	47.9	17.0	12.6	100	0.70
Coastal and Remote Britain	8.7	8.5	8.9	0.3	1.1	0.0	13.4	56.8	2.2	100	0.64
Mining and Manufacturing	32.7	31.8	6.6	0.4	1.1	0.0	12.0	14.3	1.1	100	0.75
Rural Northern Ireland	24.2	21.6	1.4	0.0	0.4	0.0	11.5	39.8	1.1	100	0.72
UK	*16.4*	*10.2*	*3.6*	*2.2*	*3.4*	*3.0*	*26.9*	*25.1*	*9.1*	*100*	*0.81*

Not coloured = made up by less than ten percent of that type.
Coloured yellow = made up by between ten and twenty percent of that type.
Coloured orange = made up by thirty to forty percent of that type.
Coloured red = made up by over forty percent by that one type.

So what does Table 8.1 show about the diversity within local authority super-group types? It is important to remember before looking at diversity that everything is relative and areas that are the most diverse or homogeneous are only like that by comparison to other area types. It does not mean that no diversity exists within the most homogeneous types.

The diversity indices show that with a value of 0.85 *Cities and Services* is clearly the most diverse super-group type. In fact *Cities and Services* is more diverse than the UK as a whole. *Mining and Manufacturing* and *Rural Northern Ireland* also show themselves to be fairly diverse groups. The least diverse super-group is *Cosmopolitan London* with a diversity value of just 0.35. The next least diverse is *Central London* with a value of 0.46. These two wholly London clusters are significantly less diverse than all other super-groups with the next lowest value of 0.64 belonging to *Coastal and Remote Britain*. The diversity indices give a good overview of which super-groups are the most diverse, but to find out why each super-group has a different level of diversity the make-up of each needs to be examined.

The *Cities and Services* (0.85) LA type is made up of five yellow ward types, but does not contain any orange or red types. This suggests that there is great diversity within *Cities and Services*. The ward classification reveals more detail about the *Cities and Services* super-group than can be gauged from the local authority level classification.

Diverse Outer London (0.67) is made up of one yellow, one orange and one red ward type. Despite its name this suggests that there is limited diversity within *Diverse Outer London*, although it is not dominated by one ward super-group type. The ward classification reveals some extra detail and information about the *Diverse Outer London* super-group.

Central London (0.46) is made up of one orange and one red ward type, but does not contain any yellow types. This LA super-group type is made up of just two ward super group types. This suggests that there is limited diversity within *Central London*. The ward classification adds only limited extra information about the *Central London* super-group.

Cosmopolitan London (0.35) is made up of one yellow and one red ward type, but does not contain any orange types. One ward type (*Multicultural Metropolitan*) makes up over 78% of the wards in this LA type and the two most abundant ward types make up 95% of this local LA type suggesting that this type is comparatively fairly homogenous. Although the ward classification does give limited additional information about the make up of these types of areas. The local authority classification is a good representation as most of the diversity within the area is accounted for at that scale.

Prospering UK (0.70) is made up of three yellow and one red ward types, but does not contain any orange types. This suggests that there is some diversity within *Prospering UK*. The ward classification reveals more detail about the *Prospering UK* super-group than can be gauged from the local authority level classification.

Coastal and Remote Britain (0.64) is made up of one yellow and one red ward type, but does not contain any orange types. This suggests that there is some diversity within *Coastal and Remote Britain*. The OA classification reveals more information about the *Coastal and Remote Britain* super-group than can be ascertained from the local authority level classification.

Mining and Manufacturing (0.75) is made up of two yellow and two orange ward types, but does not contain any red types. This suggests that there is great diversity within *Mining and Manufacturing*, more than can be deduced from the local authority level classification.

Rural Northern Ireland (0.72) is made up of one yellow and three orange ward types, but does not contain any red types. This suggests that there is great diversity within *Rural Northern Ireland*, far more than can be gauged from the local authority level classification.

The eight local authority super-group types show a great deal of variation in terms of the diversity within them. Some super-groups such as *Cities and Services* (0.85), *Mining and Manufacturing* (0.75) and *Rural Northern Ireland* (0.72) show a great diversity of ward types within them. In contrast *Cosmopolitan London* (0.35) and *Central London* (0.46) show little diversity within, with the vast majority of wards that make up these ward super-group types being of a single ward super-group type.

8.6.2 OAs within Wards

This section examines the diversity of OA super-group types within the ward super-groups types. Table 8.2 shows a cross-tabulation of ward and OA super-group types.

Table 8.2: The OA Classification Super-group types that make up the Ward Classification Super-groups (percent of OAs by type for ward types)

	OA Classification Super-group types								
Ward Classification Super-group types	Blue Collar Communities	City Living	Countryside	Prospering Suburbs	Constrained by Circumstances	Typical Traits	Multicultural	Total	Diversity Indices
Industrial Hinterlands	30.8	1.1	3.5	22.4	18.7	20.9	2.6	100	0.77
Built-up Areas	32.9	2.2	1.8	9.2	31.0	16.0	7.0	100	0.76
Student Communities	7.6	14.9	2.3	5.4	53.3	14.1	2.4	100	0.66
Prospering Metropolitan	0.2	61.6	0.2	2.1	6.6	4.6	24.7	100	0.55
Traditional Manufacturing	3.5	33.3	1.3	9.6	7.3	29.6	15.5	100	0.76
Multicultural Metropolitan	1.0	6.9	0.0	1.2	1.6	1.8	87.4	100	0.23
Suburbs and Small Towns	9.4	4.4	12.3	36.9	7.1	24.6	5.4	100	0.77
Coastal and Countryside	15.8	1.2	32.6	22.5	9.1	18.5	0.2	100	0.78
Accessible Countryside	3.4	0.3	55.2	33.6	1.5	6.0	0.0	100	0.58
UK	*16.1*	*7.5*	*12.4*	*21.2*	*14.9*	*18.3*	*9.7*	*100*	*0.84*

Not coloured = made up by less than ten percent of that type.
Coloured yellow = made up by between ten and twenty percent of that type.
Coloured orange = made up by thirty to forty percent of that type.
Coloured red = made up by over forty percent by that one type.

The diversity indices show that with a value of 0.84 the UK displays more diversity than any individual super-group type. The most diverse super-group type is *Coastal and Countryside* with a value of 0.78, closely followed by *Industrial Hinterlands*, *Suburbs and Small Towns*, *Built-up Areas* and *Traditional Manufacturing* with values of between 0.76-0.77. By far the least diverse super-group is *Multicultural Metropolitan* with a value of just 0.23.

Industrial Hinterlands (0.77) is made up of three orange and one yellow OA types, but does not contain any red types. This suggests that there is significant diversity within *Industrial Hinterlands*, being well spaced between four main types. The OA classification reveals more detail and information about the *Industrial Hinterland* super-group than can be gauged from the ward level classification.

Built-up Areas (0.76) is made up of two orange and one yellow OA types, but does not contain any red types. This suggests that like *Industrial Hinterlands*, *Built-up Areas* have significant diversity within them.

Student Communities (0.66) is made up of one red and two yellow OA type, but does not contain any orange types. This suggests that Student *Communities* show less diversity than *Industrial Hinterlands* or *Built-up Areas*, being made up by over 50% by just one OA type (*Constrained by Circumstances*). However, *Student Communities* are far from homogeneous with two other OA types also represented prominently, each accounting for about 15 % of the OAs with the ward type. This suggests that the OA classification reveals some extra information about the *Student Communities* super-group than can be gauged from the ward level classification.

Prospering Metropolitan (0.55) is made up of one red and one orange OA type, but does not contain any yellow types. One OA type (*City Living*) makes up over 60% of the OAs in this ward type and the two must abundant OA types make up over 86% of this Ward type suggesting that this type is fairly homogenous. Although the OA classification does give additional information about the make up of these types of areas the Ward classification is a good representation as most of the diversity within the area is accounted for at that scale.

Traditional Manufacturing (0.76) is made up of two orange and one yellow OA type, but does not contain any red types. This suggests that like *Industrial Hinterlands* and *Built-up Areas, Traditional Manufacturing* wards have significant diversity within them. The OA classification reveals more detail and information about the *Built-up Areas* super-group.

Multicultural Metropolitan (0.23) is made up of one red OA type, but does not contain any yellow or orange types. In fact, 87.4% of the OAs within this ward type belong to just one OA type (*Multicultural*). This is the most homogeneous ward type. The ward classification accounts for the majority of the diversity within these types of areas. The OA classification only adds a small amount of information.

Suburbs and Small Towns (0.77) is made up of two orange and one yellow OA types, but does not contain any red types. This suggests that like *Industrial Hinterlands*, *Built-up Areas* and *Traditional Manufacturing, Suburbs and Small* Towns have significant diversity within them. This suggests that the OA classification reveals more detail and information about *the Built-up Areas* super-group than can be gauged from the ward level classification.

Coastal and Countryside (0.78) is made up of two orange and two yellow OA types, but does not contain any red types. It rivals *Industrial Hinterlands* as being the most diverse of the ward super-group types. This suggests that the OA classification reveals more detail and information about the *Industrial Hinterland* super-group than can be ascertained from the ward level classification.

Accessible Countryside (0.58) is made up of one red (*Countryside*) and one orange (*Prospering Suburbs*) OA type, but no yellow OA types. This super-group type is comparatively homogenous.

The nine ward super-group types show a great deal of variation in terms of the diversity within them. Some super-groups such as *Industrial Hinterlands* (0.77) and *Coastal and Countryside* (0.78) show a great diversity of OA types within them. In contrast *Multicultural Metropolitan* (0.23) shows little diversity with the vast majority of OAs that make up this ward super-group type being members of a single OA super-group type

8.6.3 OAs within Local Authorities

This investigation examines the diversity of OA super-group types within the local authority super-groups types. Table 8.3 shows a cross-tabulation of local authority and OA super-group types. It would be expected that this investigation should show greater diversity than the previous two investigations because there is a larger difference in scale between the two geographies in this investigation that there is in the previous two.

Table 8.3: The OA Classification Super-group types that make up the LA Classification Families (percent of OAs by type for LA types)

	OA Classification Super-group types								
LA Classification Family types	Blue Collar Communities	City Living	Countryside	Prospering Suburbs	Constrained by Circumstances	Typical Traits	Multicultural	Total	Diversity Indices
Cities and Services	13.1	15.3	1.2	15.5	20.7	20.7	13.5	100	0.83
Diverse Outer London	2.2	14.9	0.0	9.4	1.7	11.0	60.8	100	0.59
Central London	0.0	46.2	0.0	0.1	0.2	0.8	52.6	100	0.51
Cosmopolitan London	0.4	11.6	0.0	0.9	0.3	0.8	86.0	100	0.25
Prospering UK	13.1	3.4	19.6	30.0	9.3	22.8	1.8	100	0.79
Coastal and Remote Britain	14.6	3.1	35.9	16.0	12.6	17.7	0.1	100	0.78
Mining and Manufacturing	28.9	2.2	6.0	23.6	23.2	15.3	0.9	100	0.78
Rural Northern Ireland	28.1	0.4	27.9	25.8	11.2	6.5	0.0	100	0.76
UK	*16.1*	*7.5*	*12.4*	*21.2*	*14.9*	*18.3*	*9.7*	*100*	*0.84*

Not coloured = made up by less than ten percent of that type.
Coloured yellow = made up by between ten and twenty percent of that type.
Coloured orange = made up by thirty to forty percent of that type.
Coloured red = made up by over forty percent by that one type.

The diversity indices show that like in the previous example, that the UK displays more diversity than any individual super-group type with a value of 0.84. However, *Cities and Services* is only slightly behind with a value a value of 0.83. Four other super-groups *Prospering UK, Coastal and Remote Britain*, *Mining and Manufacturing* and *Rural Northern*

Ireland show significant diversity, with values between 0.79-0.76. By far the least diverse super-group type is *Cosmopolitan London* with a value of just 0.25.

Cities and Services (0.83) is made up of four yellow and two orange OA type, but does not contain any red types. This shows that there is great diversity within *Cities and Services*, and that the OA classification reveals more detail and information about the *Cities and Services* super-group than can be gauged from the local authority level classification.

Diverse Outer London (0.59) is made up of one yellow and one red OA type, but does not contain any orange types. Despite its name this suggests that there is limited diversity within *Diverse Outer London*, although it is not totally dominated by one OA super-group type. The OA classification reveals some extra detail and information about the *Diverse Outer London* super-group than can be obtained from the local authority level classification.

Central London (0.51) is made up of two red OA types, but does not contain any yellow or orange types. This LA super-group type is dominated by just two OA super group types. This suggests that there is some, but limited diversity within *Central London*. The OA classification adds some extra detail and information about the *Central London* super-group than can be inferred from the local authority level classification.

Cosmopolitan London (0.25) is made up of one yellow and one red OA type, but does not contain any orange types. One OA type (*Multicultural*) makes up over 86% of the OAs in this OA type and the two must abundant ward types make up 97.6% of this local authority type suggesting that this type is comparatively homogenous and although the OA classification does give limited additional information about the make up of these types of areas. The local authority classification is a good representation as most of the diversity within the area is accounted for at that scale.

Prospering UK (0.79) is made up of two yellow and two orange ward types, but does not contain any red types. This suggests that there is some diversity within *Prospering UK*. This suggests that the OA classification reveals more detail and information about the *Prospering UK* super-group than can be gauged from the local authority level classification.

Coastal and Remote Britain (0.78) is made up of four yellow and one orange OA types, but does not contain any red types. This suggests that there is great diversity within *Coastal and Remote Britain*, and that the OA classification reveals more detail and information about the *Coastal and Remote Britain* super-group than can be ascertained from the local authority level classification.

Mining and Manufacturing (0.78) is made up of one yellow and three orange OA types, but does not contain any red types. This suggests that there is great diversity within *Mining and Manufacturing*, and that the OA classification reveals more detail and information about the *Mining and Manufacturing* super-group that the can be gauged from the local authority level classification.

Rural Northern Ireland (0.76) is made up of one yellow and three orange OA types, but does not contain any red types. This suggests that there is great diversity within *Rural Northern Ireland*, and that the OA classification reveals more detail and information about the *Rural Northern Ireland* super-group than can be gauged from the local authority level classification.

The eight local authority super-group types show a great deal of variation in terms of the diversity within them. Some super-groups such as Cities *and Services* (0.83)*, Prospering UK* (0.79)*, Coastal and Remote Britain* (0.78)*, Mining and Manufacturing* (0.78) and *Rural Northern Ireland* (0.76) show a great diversity of OA types within them. In contrast *Cosmopolitan London* (0.25) and *Diverse Outer London* (0.51) show little diversity with the vast majority of OAs that make up these local authority super-group types being of a single OA super-group type.

8.6.4 Summarising Diversity within the Levels

It is generally regarded that large urban centres display great diversity and rural areas are more homogeneous. The evidence here on the diversity within the levels appears to suggest otherwise. The lists of the most homogeneous cluster types include names such as *Central London*, *Multicultural Metropolitan* and *Cosmopolitan London*. On initial inspection this would seem to suggest that urban areas area not as diverse as perhaps previously thought and that non-urban areas may contain more diversity than previously considered. However, things are more complicated than simply seeing areas as diverse or not diverse based on the distribution of area types within them. What the perceived lack of diversity in dense urban areas shows is that in these urban areas social patterns are operating at several different scales. What has been shown in this analysis is that the same patterns can be seen at the local authority, ward and the OA scale in the most densely populated urban areas especially the wholly London clusters (clusters which only contain areas that are located in London). This can be looked at in several different ways. In a positive way it can be considered that in these areas the multi-level classification system has not fallen foul of the ecological fallacy as a similar impression of these areas is given by the classifications at the local authority, ward and OA scales. However, if the ecological fallacy is to be embraced as a function of scale rather than a failure of analysis, then it is good to see greater diversity of area types within larger areas.

The reason for constructing a multi-scale system is so that greater information about areas can be learnt by using information from more than one scale. If all scales give the same information about an area, then the users of the classifications are not receiving full value from the multi-scale system. One of the reasons for this is that urban areas cover comparatively small geographic areas even at the local authority scale. Tobler's law (all things are similar, but nearer things are more similar than those further away) would dictate that ward and OAs within more urban local authorities are likely to be fairly similar to each other because the distance between them is comparatively small. However, people's lives are not confined to one spatial scale. They may reside in a small area, but they shop and socialise in a large area on a frequent basis they find work in an even bigger area. So in using classifications in studies it may be useful to use classes from more than one level.

Contrary to what the multi-scale system has shown in the densest urban areas, the less densely populated areas (areas which cover a larger geographic area, but perhaps contain fewer people) show a great deal of diversity. These include clusters such as *Cities and Services*, *Industrial Hinterlands* and *Prospering UK*. These are not deeply rural areas, but areas which are perhaps close to large urban settlements or areas which contain a reasonably sized urban centre themselves, but they are not the densely populated inner cities. These areas show diversity when they are examined at a smaller scale as they represent the fuzzy transition from urban to rural. Within these areas (unlike in the areas at the urban and rural extremes) are a variety of smaller cities, large towns, small towns, villages and hamlets.

8.7 Diversity within the Output Area Classification

The Output Area Classification has been used to examine diversity within the local authority and ward classifications. It reveals the diversity of the country at a very small scale, but it would be misleading to suggest that OAs are completely homogeneous in their make up. There are differences within OAs as there are within any areal unit. Many will be broadly similar and because of their comparatively small geographic size they are unlikely to contain extreme differences; others may contain a diverse population within their small area.

There are no smaller geographic areas for which census data are released for creating a classification and investigating the diversity within output areas. However, the diversity within OAs and therefore the OA classification can be investigated using the household composition variables in the census.

8.7.1 Investigation into the Diversity of Households Types within OAs

By looking at the distribution of selected household types within the OA Classification super-groups, insight into diversity within the OA Classification super-groups can be achieved. As a smaller scale area classification is not available this investigation of household diversity acts as a more than ample proxy. By using the information about household types, we can investigate diversity within the OA classification, and learn how diversity differs between the OA classification super-groups.

Table 8.4 shows an aggregation of selected household types by OA classification super-group, each column represents an OA super-group type and the colours that make up the each column represent the percentage make up of that super-group by each of the household types. A quick glance at Table 8.4 immediately reveals that there is diversity in terms of household type within the OA classification. Although 'Married couple/family households' account for the largest proportion of all the super-groups, their prevalence differs significantly between the super-groups.

Table 8.4: The household types that make up the OA Classification Super-groups (percent of household types by OA type)

	Household types								
OA Classification Super-group types	One pensioner households	One person (not pensioner) household	All pensioner households	Married couple/family households	Cohabiting couple/family households	Lone parent households	Other households	Total	Diversity Indices
Blue Collar Communities	5.7	4.6	6.7	46.7	11.0	16.9	8.3	100	0.72
City Living	8.1	18.0	4.9	28.4	12.7	6.6	21.2	100	0.81
Countryside	5.6	4.4	10.1	58.6	7.7	6.2	7.4	100	0.63
Prospering Suburbs	4.5	3.2	9.8	64.7	6.6	5.3	5.9	100	0.56
Constrained by Circumstances	11.9	10.3	7.8	34.6	10.3	17.2	8.0	100	0.80
Typical Traits	5.6	7.4	6.8	51.2	11.5	9.0	8.4	100	0.70
Multicultural	4.8	9.0	3.4	37.8	8.5	15.8	20.7	100	0.77
UK	*14.4*	*15.8*	*8.8*	*36.7*	*8.1*	*9.7*	*6.6*	*100*	*0.79*

Source: 2001 Census KS table 20 Household Composition

The most diverse super-group type in terms of its household type make up is *City Living* with a value of 0.81. Both *City Living* and *Constrained by Circumstances* (0.80) are more diverse than the UK as a whole (0.79). The least diverse super-group is *Prospering Suburbs* with a value of 0.56.

The *Blue Collar Communities* (0.72) super-group is made up of 47% 'Married couple/family households', 17% 'Lone parent households' and 11% 'Married couple/family households'. The other household types make up less that 10% of this super-group type. *Blue Collar Communities* do not have the highest percentage value for any of the household types, but are a close second to *Constrained by Circumstances* in terms of the proportion of ‘Lone parent households’.

The *City Living* (0.81) super-group is made up of 28% ‘Married couple/family households’, 21% ‘Other households’, 18% ‘One person (not pensioner) households’ and 13% ‘Cohabiting couple/family households’. The other household types account for less that 10% of this super-group type. *City Living* has the highest percentage make up of ‘One person (not pensioner) households’, ‘Cohabiting couple/family households’ and ‘Other households’. This super-group has the lowest percentage of ‘Married couple/family households’ the most prevalent household type across the UK. This makes *City Living* the most diverse OA Classification super-group in terms of the household types within it.

The *Countryside* (0.63) super-group is made up of 59% by ‘Married couple/family households’ and 10% ‘All pensioner households’. The other household types make up less that 10% of this super-group type. *Countryside* has the highest percentage of ‘All pensioner households’ and the second highest percentage of ‘Married couple/family households’. This suggests that the *Countryside* super-group is fairly homogeneous in terms of its household type mix.

The *Prospering Suburbs* (0.56) super-group is made up of 65% ‘Married couple/family households’. The other household types make up less that 10% of this super-group. The *Prospering Suburbs* super-group has the highest percentage of ‘Married couple/family households’ and the second highest percentage of ‘All pensioner households’. The *Prospering Suburbs* super-group is the least diverse super-group type. It is the only super-group that does not have two or more household types with a value of 10% or greater.

The *Constrained by Circumstances* (0.80) super-group is made up of 35% ‘Married couple/family households’, 17% ’Lone parent households’, 12% ‘One Pensioner Households’, 10% ‘One person (not pensioner) households’ and 10% ‘Cohabiting couple/family households’. The other types make up less that 10% of this super-group type. The *Constrained by Circumstances* super-group has the highest percentage make up of ’Lone parent households’ and ‘One Pensioner Households’. This super-group has the most household types with over 10% membership. In fact, 5 out of the 7 household types are have more than 10% representation in *Constrained by Circumstances*. This shows that in comparison to the other super-group types *Constrained by Circumstances* is one of the more diverse types.

The *Typical Traits* (0.70) super-group is made up of 51% ‘Married couple/family households’ and 12% ‘Cohabiting couple/family households’. The other household types make up less that 10% of this super-group type. *Typical Traits* does not have the highest percentage for any of the household types. This super-group type shows only a moderate amount of diversity in comparison to the others.

The *Multicultural* (0.77) super-group is composed of 39% 'Married couple/family households' and 21% 'Other households' and 16% Lone parent households. The other household types make up less that 10% of this super-group type. *Multicultural* does not have the highest percentage for any of the household types. However, it is a close second to *City Living* in terms of 'Other households' with just 0.5% fewer. The *Multicultural* super-group is fairly diverse but not as diverse as *City Living* or *Constrained by Circumstances*.

The OA super-group types vary considerably in terms of household composition. The *City Living* (0.81) super-group is made up of no more than 29% of one household type whereas *Prospering Suburbs* (0.56) 'Married couple/family households' make up almost two thirds of the households.

This suggests that to get a true picture of an area it is not enough to just look at the smallest level of geographic output but the diversity within them needs to be explored. It is easy to think that behind any row of terraced houses there is a selection of broadly similar people with similar family/cohabiting arrangements. However, if the diversity of even a small area is explored a a variety of different households and people are revealed.

There is clearly a need here for the creation of a classification at the household level which would be able to illustrate diversity within neighbourhoods at the very smallest scale. Computing power is now sufficient to carry out a classification of several million objects (e.g. households). It would be possible to investigate diversity at a household level using the recently released Sample of Annonomised Records (SARs). Although the SARs contain no geography through an arrangement with the ONS it would be possible to apply a classification created on the SARs to the full household database. This could then be aggregated to output areas to provide an indication of the diversity within OAs. This would be the obvious next step in this investigation of the diversity of the UK and would add a further level to the hierarchy of area classifications. However, this would not be a small project and would require significant funding in order for the project to be carried out effectively.

8.8 Conclusions: is there Diversity within Area Classifications?

Urban areas show homogeneity when the LA and ward classifications are investigated, but show diversity when investigated at OA scale. In contrast rural and suburban areas show diversity when investigated at ward and LA scales, but appear more homogenous when investigated at the OA scale.

It would be great to think that there is significant homogeneity within the clusters of area classifications. However, this is not only an unrealistic expectation but also a misunderstanding of what the classification is for and what the groupings created represent. The classifications represent groupings of areas based on the characteristics of the people and households who live within them; they do not represent groupings of people who are the same. Diversity exists within all clusters, even those considered homogeneous. However, everything is relative and some classes do show real diversity within them whereas others appear much more homogenous.

The investigations in this chapter show that there is diversity within the classification super-groups. However, there is a difference in the amount of diversity not just within the different super groups, but also between the different scales of classification. This illustrates the value of creating classifications at more than one geographic scale. These features do not conflict with each other, but better allow the user of the classification to choose the level of detail which they wish to use for their investigation and help them to make comparisons between different scales.

The investigation of the diversity of ward super-group types within the local authority classification super-groups shows that, at this scale, the areas that contain the most diversity are those super-groups that represent less densely populated areas such as large suburbs and small towns. Rural areas show themselves to be fairly diverse. However, by comparison, densely populated urban groups especially those which are located wholly in London, show little diversity within them. What this tells us is that the reduction in scale from local authorities to wards reveals much more detail in the non-urban areas in comparison to the urban areas. This is because in terms of urban areas local authorities and wards are coarse in scale. Reducing the scale from local authorities to wards does not reveal much more of the diversity within urban areas as diversity within them can operate at an even smaller scale. However, in less densely populated areas this change in scale can reveal a lot more detail. The reason for this is fairly straight forward. A non-urban local authority (not solely containing a large city) is made up of a mixture of a small city or large town, several smaller towns and numerous villages and hamlets surrounded by countryside. By splitting the local authority down into its constituent wards these distinct environments are now contained within separate areas. The larger towns are made up of several wards; the smaller towns will probably each comprise its own ward; the villages and countryside in between will also comprise a number of wards. The local authority has been broken up into significantly different types of places. It is therefore easy to see how the change in scale of the classification from local authorities to wards would reveal a great amount of added detail about these areas.

The diversity of OA super-group types within the ward classification super-groups shows a similar pattern to the distribution of wards within local authorities with the same kinds of areas showing the most diversity. However, some more urban areas have started to show some diversity within them. The ward super-group type *Built-up Areas* shows significant diversity. These areas are found in the centre of smaller cities or towards the outskirts of larger cities. Those areas which are located in the centre of the largest urban areas are the super-groups which show the least amount of difference when the scale is reduced from wards to OAs. This includes super-groups such as *Multicultural Metropolitan*, *Prospering Metropolitan* and *Student Communities*. There are several possible reasons for this. Firstly, these types of area could be operating at a spatial scale which is akin to wards. This is not an unreasonable assumption. However, for the pattern to appear as it does, these areas would also have to respect ward boundaries. This is not out of the question as areas within cities are often known by their ward name and these areas are regarded as having certain types of people living within them. It does suggest that there maybe something else that is making these ward super-group types appear comparatively homogeneous in terms of the OA super-group types within them. The reason for this could be diversity within the OA super-groups of the type found within this ward type.

The investigation of the diversity of OA super-group types within the local authority classification super-groups shows great diversity for the majority of local authority super-group types; this is perhaps expected as the difference between the two scales is greater than in the previous two investigations. The local authority super-groups that show the least amount of diversity are the wholly London groups, *Diverse Outer London*, *Central London* and *Cosmopolitan London*. The reason for this is undoubtedly because these local authorities cover the smallest geographic area.

The investigation of the diversity of household types within the OA classification super-groups shows greater diversity in the most urban super-group types. This is the reverse of what has been observed at the higher scales and reinforces the previous proposition that the reason why the most urban local authority and ward types did not show diversity in terms of the OA types within them is because they contain the OA types which have they most diversity within them.

What this shows is that different types of areas display their diversity at different scales. Rural and semi-rural areas show diversity at a large scale and changing from local authorities to wards reveals much greater detail about areas. However, in the most densely urban areas their diversity cannot be seen until the diversity within output areas is examined. What this exemplifies is the importance and relevance of the multi-scale system. By using multiple scales of classification in the same investigation not only does it enable the user to choose between the different scales of classification to establish which is the most appropriate for use, but also will give the user an enhanced understanding of the importance of scale and the effect it has on the analysis of data. An understanding of the importance and value of scale especially in terms of areal statistics is vital for the interpretation and comprehension of spatial data.

Chapter Nine - Conclusions: The Way Forward for a Newly Classified Nation

9.1 Introduction

This final chapter summarises the findings of the research, the aims of which were outlined in Chapter 1. Section 9.2 discusses how these aims were achieved though the implementation of the research objectives. The research findings will be summarised focussing on the successful creation and publication of the output area classification. Section 9.3 outlines the publication of the OA classification, what information is available and where it can be obtained. Section 9.4 reviews some of the uses of the classification has been put to since its publication. Section 9.5 discuses some of the limitations of the research with a view to a future research agenda set out in Section 9.6. Section 9.6 looks to the future, outlining a number of possible research project ideas, some of which have already received funding and others for which funding will be applied for in the future.

9.2 Summary of Research Findings

The aim of this thesis was to create a classification of 2001 Census output areas of the UK to fit into the ONS suite of area classifications. Chapter 1 established that in order to attain this research aim a number of ancillary research objectives needed to be accomplished. Table 1.1 presents each chapter and their related research objectives. The work of each chapter will be discussed in terms of its success at meeting its stated research objective.

The first objective was to introduce the concept of clustering and the idea of area classifications through a variety of examples of previous systems and applications; this was achieved in Chapter 2 The concept of clustering was illustrated in a variety of different ways, using a number of examples. It was shown that not only do similar things tend towards each other geographically, but in many cases they are synthetically constructed clusters by people who understand that clustering is mutually beneficial to all parties.

A summary was given of the development of area classifications as a way of understanding and accounting for the differences between sections of society from the very early work of Charles

Booth, to the modern geodemographics industry. Central to the understanding of the development of area classification, especially in terms of applying techniques to smaller spatial scales covering larger areas, is the continued development and improvements in computing power and technology. This is what has made possible the development of the geodemographics industry over the last twenty years, leading to a point at which it is now possible to construct a fine level system on nothing more powerful than a home PC. It was established that the biggest development in area classification came with the initial development of the geodemographics industry. This industry has thrived over the past twenty years including ever larger and new data sets although it has remained largely unchanged in terms of focus and methods.

There have been some developments in terms more sophisticated methods, such as the work of Openshaw (1995). However, there has been little to suggest that more sophisticated methods produce a better or even as good solution. This is undoubtedly a view which is shared by the main players in the geodemographics market who are still using their original methodologies. A more recent debate has been the emergence of ‘fuzzy geodemographics’, which looks at classifications using a grey scale of multiple memberships rather than a black and white membership assignment that traditional methods give. Although there is no doubt that fuzzy classification methods do have numerous applications, the whole ethos of area classification is to simplify complex patterns. A fuzzy classification maybe a more accurate representation of reality, but from a practical point of view geodemographics has got where it is today for two reasons firstly, because it is simple and secondly because it works. Although fuzzy classifications will find their niche in more complex and detailed analysis, it is difficult to envisage the sea change to fuzzy systems that some have called for.

Chapter 3 builds on the ideas and issues of clustering and classification that were introduced towards the end of Chapter 2 to give a detailed overview of the methods and procedures that are involved in creating an area classification. The outline of the methods involved in the creation of a classification builds on and incorporates both ‘seven steps of cluster analysis’ (Milligan 1996) and the framework of Harris (1999) who condenses the creation of a geodemographic classification into three stages of ‘Inputs’, ‘Processes’ and ‘Outputs’.

The investigation and review of methods in Chapter 3 fulfils the requirements of objective three (objective outlined in §1.2). The theoretical and practical issues of variable selection are reviewed. Debates are outlined such as the differing opinions on whether it is better to have as many variables as possible in the classification or whether adding more variables to the classification only serves to mask the more important variables in the classification. Issues surrounding data quality and methods of reducing an initial set of variables into a final set of variables for classification are also discussed.

The review of processes used in the creation of a classification system evaluates the methods of standardisation and clustering used within this project. The chapter outlined the pros and cons of the different forms of standardisation and explained the workings and implementation of both k-means and Ward's clustering methods.

The outputs from a classification are a vital part of the creation of a successful system. Without useful and easy to understand outputs the value of the classification will be lost. The discussion on the outputs from the classification covers selection of the number of clusters to go into the classification and bringing a classification to life with the use of cluster portraits, names and maps.

Chapter 4 outlines the creation of the first classification to come out of this project, a classification of UK local authorities. The intention of this classification as outlined in objective four was to be experimental and to gain an understanding of the practicalities of creating a classification using a relatively small data set before the more difficult task of creating a classification using more numerous output areas. The successful creation of a classification at the local authority scale is described from start to finish resulting in a fully documented classification published as a working paper via the School of Geography, University of Leeds website. The chapter concludes with a comparison of the LA classification created during this project with the LA classification created by the ONS.

Chapter 5 resolves objectives five, six and seven. The knowledge gained in the review of previous classifications and methods and in the creation of the LA classification was used to create a classification of UK output areas. The chapter follows the creation of the classification in sequence from the initial variable selection to the final steps of naming and mapping the clusters. The chapter provides a detailed description of the variable selection and standardisation of the data, through the changing methodology used to cluster the OAs into groups to detailed descriptive and visual outputs.

Chapter 6 addresses objective eight by quality assuring and adding value to the output area classification by putting it through a number of tests and examinations to ensure that the classification is both robust and usable. This chapter includes an analysis of the variability reduction and the power of each variable within the classification. The variability reduction within the classification hierarchy is examined, investigating the extent to which the hierarchy of the classification reduces the variability within clusters and increases the differences between clusters. Areas of atypicality, which are furthest from the centre of any cluster, are examined to establish the reasons behind their lack of convention. Photographs and experiences from ground truthing field trips are displayed and shared. By fuzzyfying the classification it is possible to see wider trends and patterns displayed by the classification. The chapter finishes by describing and outlining the results of an innovative consultation exercise, which gave people the option to assess and comment on the classification in terms of how well it reflects an area known to the respondent. Some of the respondents provided both interesting and useful comments and suggestions about the classification, some of which have led to ideas for further work, outlined in §9.6.

Chapter 7 confronts objective nine by providing evidence of the value of the OA classification by using it to predict and account for trends and patterns seen in a number of current socio-demographic issues. The classification was used to discover significant patterns in a series of phenomena with the use of a number of varied, topical case studies. The investigations included: an examination of the differences between England's 'Core Cities'; an analysis of the difference in voting patterns between the OA super-group and the differing amounts they have changed between elections; a comparison of the OA classification with both the urban-rural classification and the Indices of Multiple Deprivation; profiling of the differences between Welsh and non-Welsh speakers in Wales and Catholics and Protestants in Northern Ireland; an analysis of the difference in migration rates between clusters and using the classification to assess whether the north-south divide shows up in a geodemographic classification.

Objective ten is tackled in Chapter 8, this links the output area classification to the other classifications in the ONS suite of area classifications. The different scales of classification are joined together to create a multi-scale system. This enables the diversity within the higher level classifications to be examined and the importance of scale within area classification to be uncovered. The examination of diversity within each of the classifications shows that there is a great deal of variation between the clusters in terms of their diversity, with some clusters appearing to show a great deal of diversity and others looking to be almost homogeneous.

9.3 Publication of the Output Area Classification

The OA classification was subject to a series of rigorous quality assurance boards within the ONS before publication was approved. This involved producing a series of reports outlining how the classification was created and providing evidence of the quality of the classification in terms of the reliability of the methodology. It was necessary to demonstrate that no errors had been made during its creation and that the classification contained no errors. I was required to respond to all questions and queries that board members had about all aspects of the classification.

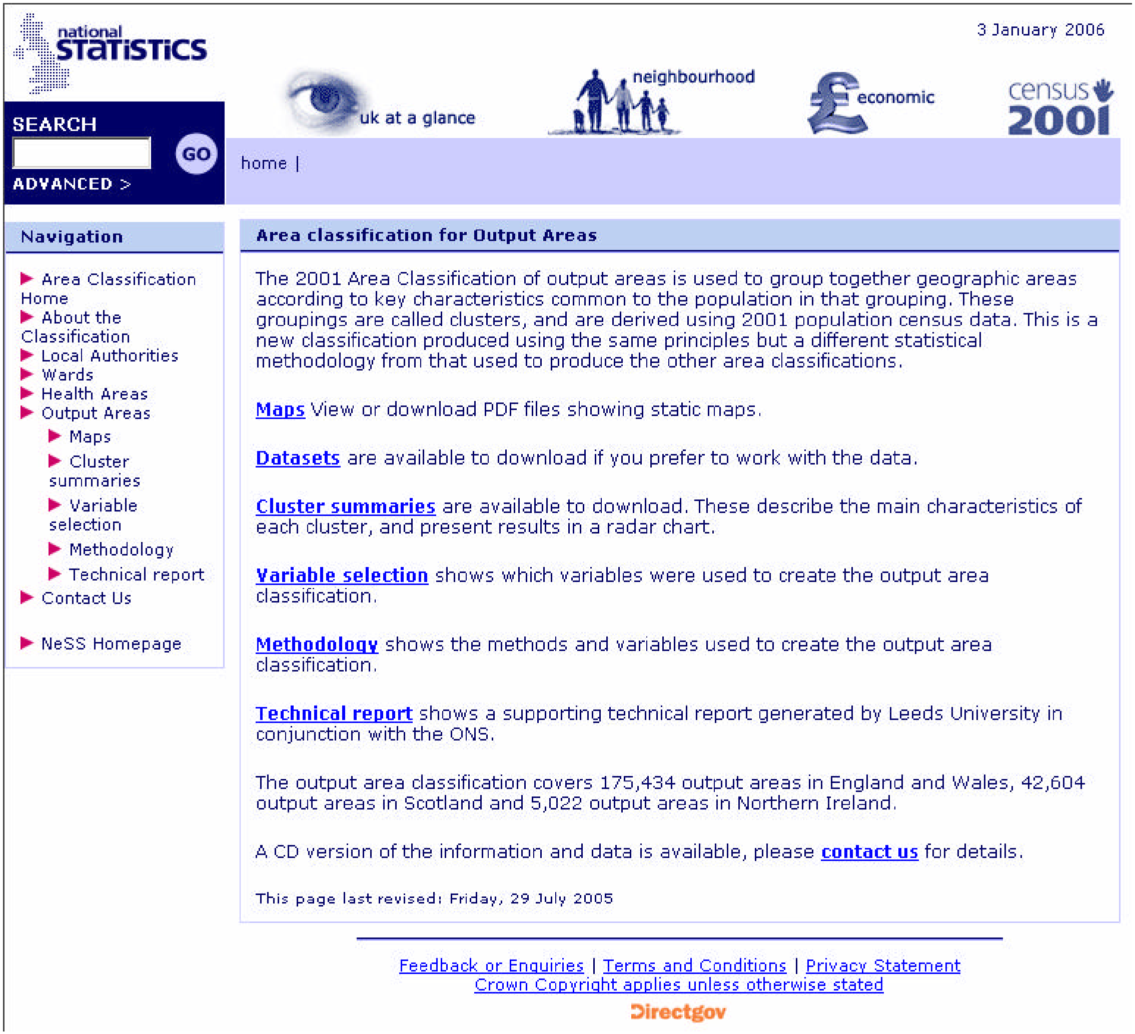
The OA Classification was published as a 'National Statistic' on the 29th July 2005, via National Statistics Online, joining the previously published LA and Ward level classifications. The classification can be found on the Office for National Statistics website, as shown in Figure 9.1 contains the following information:

- Maps
- Datasets
    - Cluster Membership
    - Distance from cluster centroid
    - Original data

- Cluster Summaries
- Variable Selection Report
- Methodology Report
- Technical Report – via the University of Leeds (Vickers *et al*. 2005)

The classification can also be ordered on CD (in Appendix F) from the ONS by e-mailing info@statistics.gov.uk

Figure 9.1: The OA Classification on National Statistics Online



Area classification for Output Areas

The 2001 Area Classification of output areas is used to group together geographic areas according to key characteristics common to the population in that grouping. These groupings are called clusters, and are derived using 2001 population census data. This is a new classification produced using the same principles but a different statistical methodology from that used to produce the other area classifications.

Maps View or download PDF files showing static maps.

Datasets are available to download if you prefer to work with the data.

Cluster summaries are available to download. These describe the main characteristics of each cluster, and present results in a radar chart.

Variable selection shows which variables were used to create the output area classification.

Methodology shows the methods and variables used to create the output area classification.

Technical report shows a supporting technical report generated by Leeds University in conjunction with the ONS.

The output area classification covers 175,434 output areas in England and Wales, 42,604 output areas in Scotland and 5,022 output areas in Northern Ireland.

A CD version of the information and data is available, please contact us for details.

This page last revised: Friday, 29 July 2005

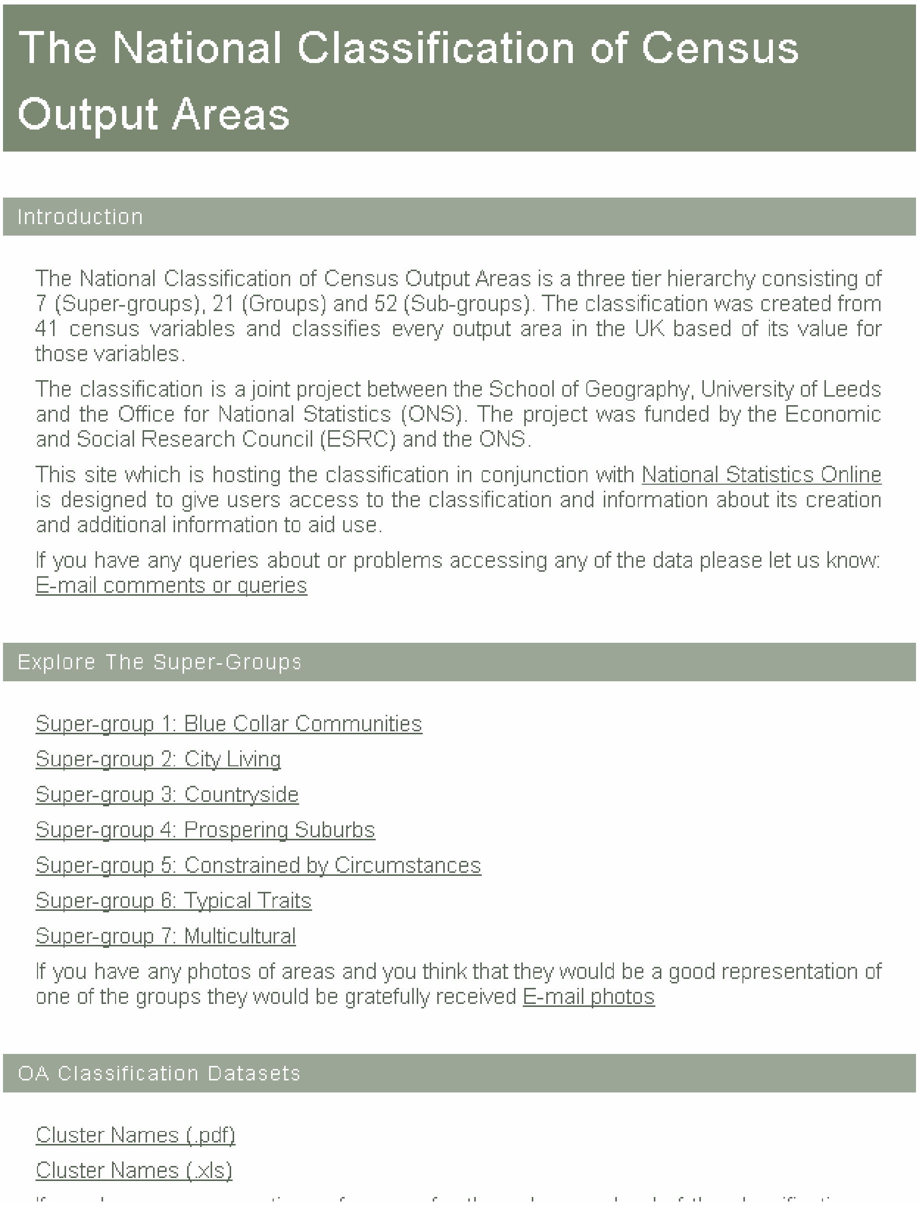
Feedback or Enquiries | Terms and Conditions | Privacy Statement
Crown Copyright applies unless otherwise stated
Directgov

http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/oa/default.asp

In conjunction with the publication on the ONS website the classification was made available via my homepage on the University of Leeds website shown in Figure 9.2. This was done to provide not only an alternative source from which the OA classification can be obtained, but also to enable the dissemination of additional information that is not available on the ONS website This additional material includes:

- Cluster Names
- Percentage of the population in each cluster type
- Alternative cluster profiles
- Hierarchy database of all three classifications
- Database of entire UK in one file
- Pseudo fuzzy classification
- Photographs of typical areas

Figure 9.2: The OA Classification on the University of Leeds website



The National Classification of Census Output Areas

Introduction

The National Classification of Census Output Areas is a three tier hierarchy consisting of 7 (Super-groups), 21 (Groups) and 52 (Sub-groups). The classification was created from 41 census variables and classifies every output area in the UK based of its value for those variables.

The classification is a joint project between the School of Geography, University of Leeds and the Office for National Statistics (ONS). The project was funded by the Economic and Social Research Council (ESRC) and the ONS.

This site which is hosting the classification in conjunction with National Statistics Online is designed to give users access to the classification and information about its creation and additional information to aid use.

If you have any queries about or problems accessing any of the data please let us know: E-mail comments or queries

Explore The Super-Groups

Super-group 1: Blue Collar Communities
Super-group 2: City Living
Super-group 3: Countryside
Super-group 4: Prospering Suburbs
Super-group 5: Constrained by Circumstances
Super-group 6: Typical Traits
Super-group 7: Multicultural

If you have any photos of areas and you think that they would be a good representation of one of the groups they would be gratefully received E-mail photos

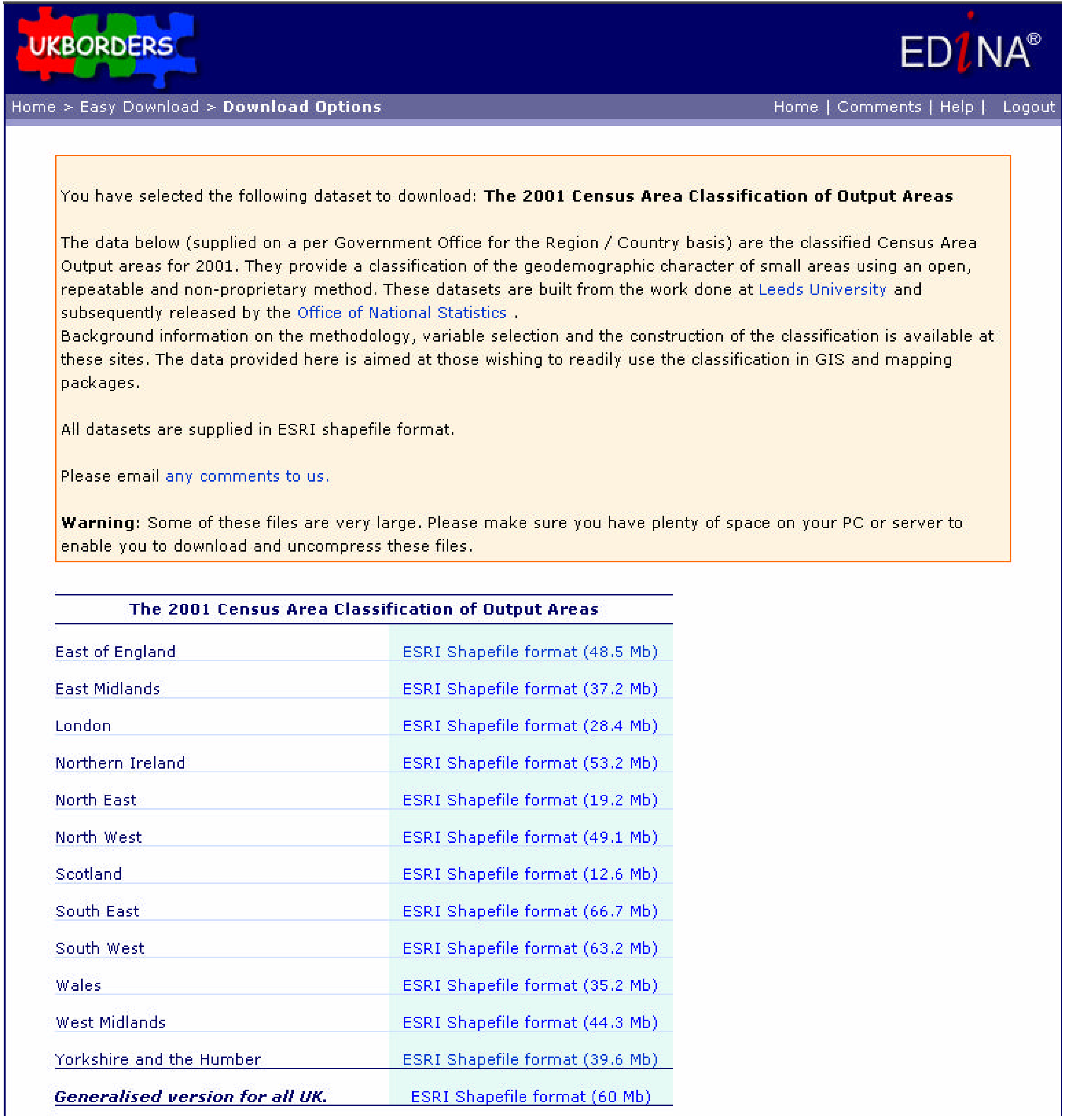
OA Classification Datasets

Cluster Names (.pdf)
Cluster Names (.xls)

http://www.geog.leeds.ac.uk/people/d.vickers/OAclassinfo.html

The classification has now also been made available via the UK BORDERS service of EDINA (part of the ESRC/JISC 2001 Census Programme), as displayed in Figure 9.3. The classification is supplied with digitised boundaries (ESRI shapefile format) each GOR/country can be downloaded as a separate file enabling registered users to visualise the classification very quickly with the use of GIS. Registered users (Athens authorisation required) can access the classification on the UK BORBERS website. It is hoped that the classification will also be made available via the CASWEB website.

Figure 9.3: The download page for the OA Classification on UK Borders


UKBORDERS
EDINA®
Home > Easy Download > **Download Options**
Home | Comments | Help | Logout

You have selected the following dataset to download: **The 2001 Census Area Classification of Output Areas**

The data below (supplied on a per Government Office for the Region / Country basis) are the classified Census Area Output areas for 2001. They provide a classification of the geodemographic character of small areas using an open, repeatable and non-proprietary method. These datasets are built from the work done at Leeds University and subsequently released by the Office of National Statistics .
Background information on the methodology, variable selection and the construction of the classification is available at these sites. The data provided here is aimed at those wishing to readily use the classification in GIS and mapping packages.

All datasets are supplied in ESRI shapefile format.

Please email any comments to us.

Warning: Some of these files are very large. Please make sure you have plenty of space on your PC or server to enable you to download and uncompress these files.

The 2001 Census Area Classification of Output Areas	
East of England	ESRI Shapefile format (48.5 Mb)
East Midlands	ESRI Shapefile format (37.2 Mb)
London	ESRI Shapefile format (28.4 Mb)
Northern Ireland	ESRI Shapefile format (53.2 Mb)
North East	ESRI Shapefile format (19.2 Mb)
North West	ESRI Shapefile format (49.1 Mb)
Scotland	ESRI Shapefile format (12.6 Mb)
South East	ESRI Shapefile format (66.7 Mb)
South West	ESRI Shapefile format (63.2 Mb)
Wales	ESRI Shapefile format (35.2 Mb)
West Midlands	ESRI Shapefile format (44.3 Mb)
Yorkshire and the Humber	ESRI Shapefile format (39.6 Mb)
Generalised version for all UK.	ESRI Shapefile format (60 Mb)

http://borders.edina.ac.uk/ukborders/restricted/easy_download/Classifications.html

The release of the classification was warmly received by researchers and policy makers from the public, private and academic sectors. Enquires have been received via phone and e-mail from people in all sectors. Download figures for the classification's first five months of releases have been calculated by the ONS, a total of 708 downloads were recorded in the period. The highest level of demand was in the first month with, 226 downloads in August, 137 in September, 141 in October, 125 in November and 79 in December.

The OA classification was awarded Demographic User Group Award 2005 for 'best new information from government'. This was presented to myself and Gregg Phillpotts (head of Regional and Local Division of ONS) representing the ONS at the Demographic User Group (DUG) Conference held at the Royal Society on the 10th November 2005. Figure 9.4 shows Keith Dugmore who runs DUG presenting me with the award. DUG was set up in 1996, with the objective of representing to government the needs of commercial users of its demographic statistics (Demographic Decisions 2005). Current members of DUG include: Abbey, Argos, Boots, Cabinet Office, Children's Mutual, Cornhill Insurance, HBoS, Landmark Information, Marks & Spencer, Nationwide, ONS, ODPM, Powergen, PPP Healthcare, RAC, Royal Bank of Scotland, Saga Group, Sainsbury's, Tesco, Thames Water, Whitbread, Woolworths and Yell. Members of the DUG have labelled the output area classification OAC, at fist glance little more than a standard acronym, but when we consider that the first small scale area classification system is named ACORN, OAC takes on a whole new meaning.

Figure 9.4: Presentation of award for the OA Classification from the Demographic User Group, titled 'best new information from government'

Demographic User Group Award 2005
"Best new information from Government"
University of Leeds (Daniel Vickers)
and the
Office for National Statistics team
"National Classification of Output Areas"

It is hoped that the release of the OA Classification can be rolled out even further. The ONS are currently engaged in fully incorporating the OA Classification into Neighbourhood Statistics, which would enable a user to enter their postcode into the website to return the OA Classification types, thus providing giving the user with an indication of the nature of that neighbourhood.

The hard work on the OA Classification has been done. However, it still has to cover the *"last 100 yards"* (Callingham 2005). For the OA Classification to reach its full potential in terms of the use it could be put to it needs to be supported by a user group preferably run by ONS, which they have recently agreed to create and to be put on all government surveys and the Target Group Index (TGI). The TGI is a continuous survey based on a sample size of approximately

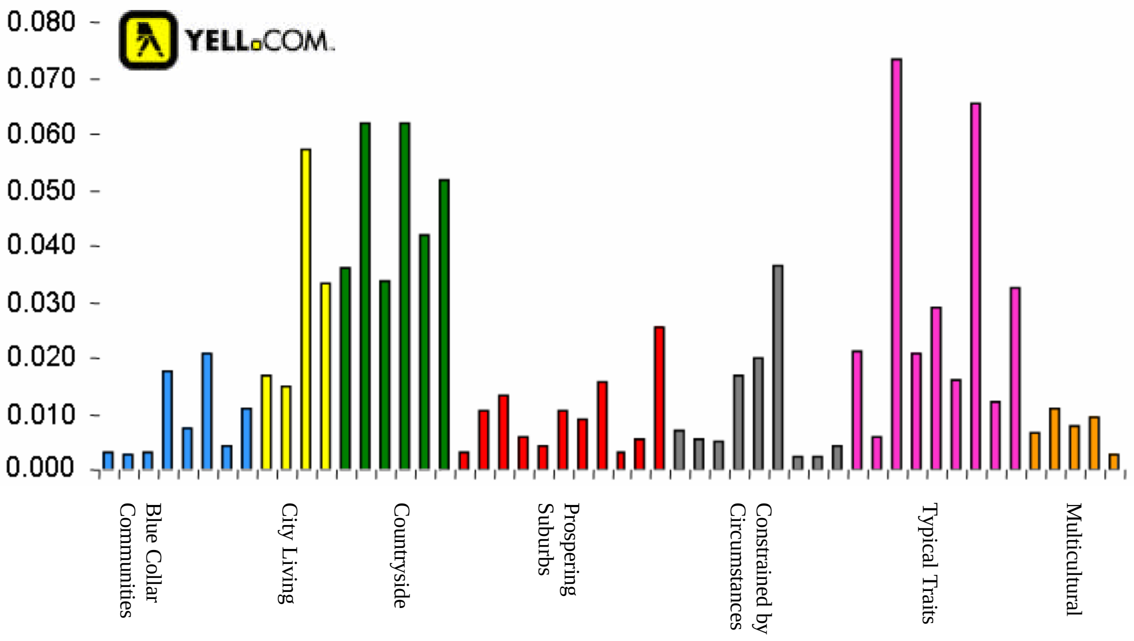
25,000 interviews per year. It provides information on the usage of over 4,000 brands in 500 product areas for those aged 15+. TGI is updated 4 times a year on a rolling quarterly basis (BRMB 2005). It would also be useful especially for academics to be able to obtain the classification attached to the SARs.

9.4 Applications of the Classification System

Chapter 7 outlined some of the many and diverse applications of the OA classification, using the classification to investigate and explain some current demographic issues and debates. The great value and relevance of the OA classification for many aspects of social science, public policy and geographic thought are clear to see, but there are many more uses for it. Some examples of applications of the OA classification in four months following publication (August to November 2005) are presented here.

One of the principal users of geodemographics and area classifications is the business community especially in connection with market research, database marketing and retail analysis. Several companies have already used the OA classification to profile against their customer databases or to examine the location of their stores. Several companies have found great value in the classification and have supplied evidence of how they have made use of the OA classification with their business. Yell.com produce the UK's largest business directory via both their website and their Yellow Pages directory delivered free to all addresses in the UK. Figure 9.5 shows how Yell described the distribution of the businesses in their directory using the OA classification.

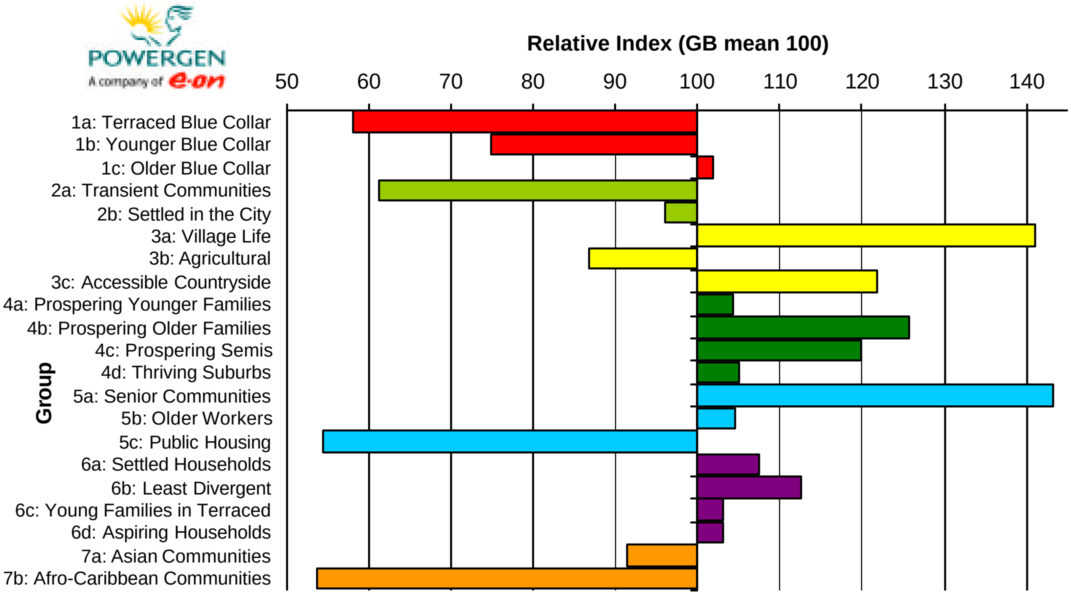
Figure 9.5: The distribution of YELL businesses by OA Classification sub-group



There are some clear differences in the propensity of businesses to locate in certain types of area. Businesses appear far more likely to locate in *City Living*, *Countryside* and *Typical Traits* than in the other super-group types and specifically within certain sub-group types within them. Within *Typical Traits*, *Least Divergent (1)* and *Young Families in Terraced Homes (2)*, show the greatest propensity to contain businesses and within *City Living*, *Settled in the City (1)* shows the greatest propensity to contain a business.

A second example of past publication use is by Powergen, an electricity and utility company. Powergen analysed the take up of their promotional energy efficiency measures by OA Classification group types. The results of their analysis are shown in Figure 9.6. The chart shows GB average as 100, therefore an index of over 100 indicates that this group is more prevalent than the GB average, less that 100 means less prevalent (Robbins 2005). The analysis shows a general trend for more affluent and rural areas to have greater take up of energy efficiency measures and also that older people are responding well to offers. However, there is some diversity within super-groups, for example within *Constrained by Circumstances*, *Senior Communities* has the highest take up rate of any group with an index value of 143 whereas *Public Housing* has the second worst take up, with an index value of just 55.

Figure 9.6: Take up of promotional energy efficiency measures OA Classification group



Other organisations including Abbey, Boots and Woolworths have also used the OA classification to profile against their customer databases or analyse their store locations (Callingham 2005). The classification has many varied and different applications. However, I would like to think that like Charles Booth's classification over 100 years ago, the OA classification will be put to uses that will help us account for and understand the geography of social inequalities within the UK.

9.5 Limitations of the Research

However successful a research project has been it is important to recognise the limitations of the research. This project is no different and there are several areas in which the project could be criticised. The majority of the limitations of this research relate to fundamental problems with area classification. In the majority of cases throughout this document area classification has been portrayed in a positive light. However, area classifications are not short of critics, although the proof of the pudding is in the eating and area classifications have proved to be pretty tasty dishes for many of their users who come back to them time and again. Many of the points made by critics of this form of analysis are valid.

The classification is reliant on the areal units that it is classifying being meaningful and representative of real world features. This is essentially the MAUP, the suggestion that the value of the unit is as much a feature of the drawing of the boundaries between the units as the underlying data. Although this is an unavoidable feature of this form of analysis, but it is highly unlikely that there has been a significant effect on the overall classification. However, some OAs which are very mixed in their make up or are towards the edge of a cluster could have been classified differently with a small movement of the OA boundary.

The classification was constructed wholly from census data as these were the only data available for output areas at the time of the project. It was decided not to incorporate data from other sources as they would not have provided the 100% geographic coverage of the census and would be at a different geographic aggregation which cannot be reliably linked to the census data. While these decisions mean that a lot of potential pitfalls have been avoided, it also meant that some additional data such as crime figures and information on wealth could not be included. This is a real catch 22 situation. Whichever decision is made the next person may have made the opposite decision. Yes the classification will have suffered from not including the additional information, but it is likely to have lost a lot of reliability from the inclusion of the less reliable data at different scales.

Another limitation of the research is the static nature of the census data. The census data represents a snapshot of the country on the 29th April 2001. No matter what date analysis is performed on the census data, the classification will remain applicable to the census date for its lifetime. There will undoubtedly be some socio-geographic change within the intervening period between the enumeration of the census and any analysis that is conducted using the classification. However, social change in the UK is not something that takes place rapidly over a short period of time, especially the geography of social groupings. Orford *et al.* (2002) discovered in their analysis of the changes in the social geography of a part of central London over the last one hundred years that the position of areas on the social hierarchy had changed little over the period with significant changes only noticeable in areas that have gone through large scale and well funded regeneration. What this tells us is that, although changes in society will date the classification and reduce its reliability through time, without large scale regeneration areas are unlikely to change significantly in their social make up and their position in the social hierarchy even over fairly long periods.

The crisp nature of classifications (an area in one class and not in other classes) has long been a criticism of classification and has led to the increasing popularity of fuzzy classifications that avoid the problem. However, fuzzy classifications do not have the simplicity of a traditional classification. The splitting of continuous datasets into deciles, groups or clusters is always going to create possible sources of error on or close to where the divisions are made. This may cause individual areas to be classified differently to very similar areas but the overall picture does not suffer from stratification.

Classifying areas in this way can be seen as much of an art as a science. A Criticism that has been levelled at this form of classification is that the classification is as much a result of the process that the data are put through as the data itself. If the analysis were conducted in a different way, a different classification would be produced. This is undoubtedly true, and is an easy attack on area classifications as the results are visual and easy to interpret. However, in all research (especially statistical analysis) the results of any research depend on how the research was conducted and performing the research in another way could alter the results.

Limitations of the original methodology to handle the complexity of the dataset meant that changes had to be made during the creation of the classification. The reason for this was not because the original methodology was unable to create a classification, but its inability to create a usable product. A limitation of the classification that has been created is that the way in which the classification reflects real world patterns had to be balanced against the need to produce a usable and interpretable classification. This will have slightly reduced the way in which the classification reflects reality. However, to produce a classification that was more representative would have created a very complicated and hard to use product.

9.6 (Endless) Possibilities for Future Research

Now the OA classification has been successfully created and published, it should not be allowed to dwell on the hard disks in the offices of public, private and academic researchers. The possibilities of research with the OA classification and research with other forms and scales of classification are endless. The OA classification is a great resource for research, policy and planning and the more work that is done with it the better. The case studies in Chapter 7 and the applications of the classification displayed in this chapter show just a few of the many research possibilities for the OA classification. This section outlines a series of research agendas and research projects relating to further research with the OA classification and the creation further classifications for different scales and geographies.

9.6.1 Mapping the Life Course through the OA Classification

Classifications are rarely tested against people's views of what the area is really like; an attempt was made to incorporate people's views of their areas into the OA classification via the consultation exercise outlined in § 6.7. By developing further qualitative work with the classification expanding well beyond the consultation exercise, it would become possible to attempt to answer several questions about the classification. How does the classification tally with the views of people about their neighbourhood? Can we account for the differences and or use them to improve the classification? Can this qualitative analysis be incorporated into the classification system?

A research theme that I am very keen and excited about taking forward is 'Mapping the life course through the OA Classification'. This is a really interesting research direction that came to the fore during the consolation exercise outlined in § 6.7. One of the participants in the consultation exercise who had requested several maps of different places in, which they had lived during their life described how they were able to track the stages of their life and their position in the life course via the OA classification super-group in which they lived at each stage of their life. *"An interesting exercise, which tells my own life history - I grew up in 'typical traits'; went off to be a student; as a postgrad I rented a room also in 'typical traits' before as a young academic buying a small terraced house in 'city centre melting pot'. After a few years I moved on to a house in 'typical traits' and a few years ago finally arrived in 'comfortable suburban estates'. You could probably classify life-histories according to transition through these profiles!"* (OA Classification consultation Respondent qa38). This complements work done by John Rex in the 1960s and 1970s, Rex recognised how people goes through a 'residential lifecycle' (Rex and Moore 1967). The lifecycle identified changes in the social and economic status of families or individuals as the go through chronological phases of their lives, based of the housing type in which people live.

The consultation exercise shows how the respondent has used the exercise to classify their own life at different stages of the life course, exemplifying that the classification does not only have to work in a static time frame but also shows that people do not live in the same kinds of places their entire life and people can move through the classification with their life stage. This could be broadened out into a research project where the life courses of a large number of people could be mapped through time and space with the use of the OA classification. Trends and patterns could be drawn out examining whether similar people go through similar stages in the life course at similar times and in what order they move through the groups in the classification. Do people move through the classification in the same order? Or is there much less of a pattern to how people move between classification types. This would be a fascinating research project and an innovative use of the OA classification. it would be useful, therefore, to add the classification to longitudinal analysis such as the Census Longitudinal Study or the British Household Panel Survey.

9.6.2 Changing Residential Patterns of the UK 1991 – 2001

I have received funding for an ESRC Postdoctoral Fellowship (PDF) to investigate 'Changing Residential Patterns of the UK 1991 - 2001' based on the OA classification. The changing nature of social trends never ceases to be of interest to social scientists. An appreciation of how social patterns change and develop over time is fundamental to understanding how society functions, during the consultation exercise I was asked *"Can you create the same for 1991 and show us change?!"* (Respondent qa13). This was a question that I had already asked myself. If the classification can be reproduced for different points in time, comparisons between them would reveal how the social geography of the UK has changed and is changing. The Fellowship will attempt to answer the following question. How have the residential patterns of the UK changed between the 1991 and 2001 Censuses of Population?

To find out if, and how, residential patterns have changed a means of comparison needs to be created from an earlier data source. Therefore a comparable classification will be created at the finest geography (enumeration districts) using data from the 1991 Census. The classification will be created using the same 41 variables (where available) used to create the Output Area Classification from the 2001 Census. A similar number of groups to the Output Area Classification will also be aimed for in order to make the two systems as comparable as possible.

When the classification of the 1991 enumeration districts is complete it will be mapped in the same way as the Output Area Classification and the residential patterns will be examined. The two classifications can be overlain with the use of GIS techniques and the patterns displayed by the two classifications can be systematically compared on a "best fit" basis. It will then be possible to assess if there have been any significant changes in the residential patterns of the UK between 1991 and 2001. The effect of processes and issues such as socio-economic inequalities, multiculturalism and gentrification on these residential patterns will also be highlighted.

The principal aim of this project is to develop a general purpose classification of 1991 Census Enumeration Districts that is comparable to the existing classification of 2001 Census Output Areas. The project's second aim will be to compare the two classifications in order to establish any changes in residential patterns over the period with a view to examining the trends and processes that have occurred.

The steps in the classification exercise are as follows:

- Assemble the database of variables, as close to those used in the 2001 Census Output Areas Classification as possible.
- Cluster the variables creating the same number of groups as in the OA classification.
- Prepare statistical and visual summaries of the classifications.
- Label the classifications with descriptions of varying lengths.
- Map the classification using GIS.
- Overlay the classifications of the two censuses and contrast and compare using GIS techniques, looking for differences and change over the period.

The methodology is now tried and tested and has been successfully implemented in the creation of the OA classification. There were fewer EDs in 1991 than OAs in 2001 so there are no issues in terms of the computing power or the ability of the software to cope with the data. There is vast scope to develop the project beyond the PDF. Further classifications could be developed using 1981 and possibly 1971 data to establish how far back demographic trends and socio-economic processes began and/or when changes started to occur.

Perhaps more importantly, the PDF would allow the development of a 1991 ED Classification that would be comparable not only with the 2001 OA Classification but also with Census 2011, because it will have the advantage of having a stable geography. Hence, there will be the potential to explore whether the demographic trends and socio-economic processes that the PDF project would reveal continued over at least a twenty year period, 1991-2011.

9.6.3 Household Classification

An application to create a household classification was made as part of an ESRC Postdoctoral Fellowship application, but unfortunately this did not receive funding. However, the household level classification is still something that I hope gets resurrected in the future as there is significant interest in its creation.

Originally planned in the CASE proposal to develop a household classification was to be part of this PhD thesis. The idea was to use a sample of households (the Household SAR from the 2001 Census) to explore methods and develop a multivariate classification. This classification could then be applied to the whole census. The investigation of diversity within the OA classification in § 8.7 exemplified how such a classification would be useful, to show diversity within output areas.

Availability on the researcher’s desktop of a public use sample from the 2001 Census was essential, given that extensive experiments would be needed and new solutions needed (e.g. how to use categorical as opposed to continuous variables). ONS publication plans available at the time of study design (autumn 2001) indicated that a Household SAR would be produced along the lines of the 1991 Census Household SAR, only minor changes in the specification were envisaged.

However, new concerns about the confidentiality and disclosure within National Statistics led to a review of the degree of detail that could be included in micro-datasets released for public use. The case for a Household SAR had been re-evaluated from scratch, despite intensive work by researchers at the Cathie Marsh Centre for Census and Survey Research (Tranmer *et al.* 2005).

As a consequence, there have been protracted negotiations about the content of the Household SAR from the 2001 Census. An announced release dates were continually moved into the future and was not released until after the conclusion of the CASE award period. The alternative offered by National Statistics for access to microdata has been the development of the Controlled Access Microdata Sample (CAMS). Such a dataset corresponding to the related Individual SAR is available in the CAMS suite, for example, if more detail on particular variables is required than in the public sample.

Examinations of the terms and conditions of the CAMS access arrangements (ONS 2005f) indicate that, even if a household microdata sample were made available, it would not be possible to extract and “bring home” the necessary microdata. The only feasible way to develop a household classification at present would be to work in a secure setting at the ONS for several months, but neither time nor resources were available for this *modus operandi*. The original plan envisaged doing all the experimental work with a public use sample and then spending about two weeks connecting the sample based classification to the full census database. A decision was taken at Research Support Group (RSG) meeting number 4 (29th June 2004) not to proceed with the Household Classification, but to carry out analysis of the diversity within the OA classification using the household classification already included in the Standard Area Statistics.

The creation of a household level classification would add the missing lowest level to the hierarchy and complete the original research objective. The household level classification would enable a greater understanding of what is happening at other scales. For example is an area with 50% non-white population made up of 50% non-white households or is it made up of mixed households? The development of the household classification would take the form of a classification of the household SARs, which are now available. A good overview of the 1991 SARs can be found in Dale *et al.* (2000).

Simple top down, rule based individual classifications (such as Wathan *et al.* 2004, Gordon 1995 and Goldthorpe 1987) have been created in the past. They enable some empirical research, but they often focus on a small number or single variables and are inherently inadequate. It is vital that the data are able to speak for themselves. I favour an inductive bottom up approach that would give scope for much wider use. Given the successful development of a household classification it will then be possible to "map" the distribution of households across neighbourhood or district types. The importance of doing this has been established in the context of deprivation by Fieldhouse (2000).

The principal aim of this project would be to develop a general purpose classification of households using the 2001 SAR. The project's second aim would be to compare existing classification methods and choose the most suitable method for household level data. Further aims include methodological issues of how to cluster household level data. For example, how to deal with households with more than one wage earner: should the household reference person be the only one used or should all earners be included?

The steps in the classification exercise are as follows:

- Review carefully the purpose of the classification and the demographic-social-economic-behaviour domains that should be covered.
- Develop a suitable set of variables that cover those domains, exploring the degree of colinearity and selecting variables that are independent.
- Decide on a method of indicator construction that treats chosen variables in a comparable way.

- Assemble the database of indicators for the units at each spatial level.
- Choose a general classification method (after review of the literature and assembly of suitable software).
- Decide on the characteristics desired in the classification (number of classes, degree of homogeneity within classes etc.).
- Experiment with the classification methods, selecting a variety of options.
- Prepare statistical and visual summaries of the classifications.
- Label the classifications with descriptions of varying lengths.

The methodology is now tried and tested and has been successfully implemented in the creation of the OA classification. The 1% sample in the households is approximately the same number as the numbers of OAs so there are no issues in terms of the computing power or the ability of the software to deal with the data.

The variables in the SARs classification will be based on the list of variables used in the OA classification. This will not be entirely possible as the data is based on households rather than areas, which makes some variables non-comparable. However, the SARs will cover the same domains as in the other levels of classification.

9.6.4 Super Output Area Classifications

Although a hierarchy of classifications of the UK using different geographies has been created. There are several other scales of geography that have not only, not been covered, but for which people have requested classifications. There has been interest and requests to produce more classifications at new scales and geographies.

Super Output Areas (SOAs) are a new geography roughly equating to the size of electoral wards. SOAs are aggregates of OAs there are three levels of SOAs, lower, middle and higher. Lower SOAs have a minimum population size of 1,000 people, middle SOAs has a minimum population size of 5,000 people and higher SOAs have a minimum population size of 25,000 people (ONS 2005g). Like the OAs, SOAs are planned to be a stable geography over time. The creation of SOA level classifications would complete the ONS classifications by including this new stable geography, which will be increasingly used as more data are made available at this scale. ONS have shown interest in producing an SOA level classification and I have been asked by them to tender to create the classification on their behalf. There has also been significant user interest in the creation of a SOA classification *"I would strongly urge that ONS consider the value of generating counterpart classifications for Super Output Areas"* (Hennell 2005). So the creation of classifications for SOAs is a distinct possibility.

9.6.5 Expansion beyond the UK

All the classifications that have been created cover only the UK. However it would be possible to create classifications that take in larger areas covering multiple nations. Comparing nations with nations or regions within countries with regions in other countries.

The first possibility would be to create a classification of Nomenclature of Units for Territorial Statistics (NUTS). NUTS are spatial units created by the European Office for Statistics (Eurostat) to report statistics about the countries of the European Union (ONS 2005h). There are three levels of NUTS units covering all 25 EU member states; level one has 89 units, level two has 214 units and level three has 1,221 units. Eurostat produces a range of information that is available for NUTS, available data includes: General Statistics, Economy and Finance, Population and Social Conditions, Industry, Trade and Services, Agriculture and Fisheries, Transport, Environment and Energy, Science and Technology (Eurostat 2005). Such a classification would be especially relevant at the current time because of the recent increase in membership of the EU from 15 to 25 countries. Not only has this increased the number of spatial units with which to work, but also the diversity within the countries of the EU.

A second possible expansion beyond the shores of the UK is a classification of the nations of the world. This is not uncommon especially in the field of economics, but less common in social studies with limited examples such as Russett (1967). Data for such a problem are available from both the CIA world fact book published annually and UN statistics statistical yearbook published annually (CIA 2005; UN 2005).

9.6.6 Specific Purpose Classifications

Voas and Williamson (2001a) suggested that specific purpose classifications should be created according to the demands of the application. Recently significant research has been conducted to create specific use classifications, classifications of crime and community safety (Ashby 2004; Ashby and Longley 2005; Shepherd *et al.* 2005) and linking geodemographics to access to higher education entry (Farr and Singleton 2004). It would be interesting not only to create specific purpose classifications for such things as education, crime, voting patterns, health etc. but to also see how the OA classification compares to these specific use classifications.

9.6.7 Regional Classifications

A final avenue of research would be to explore the effects of limiting the study area to subsets of the UK, such as regions. Can different things be seen in regional classifications in comparison to the whole UK classification? Can more extreme local and regional variations be seen? The analysis of atypical areas in § 6.4 raised issues of whether some areas of the country would benefit from regional classifications. The analysis of the north-south divide shows that there are inequalities between different regions of the UK. A regional level classification would allow extremes within each region that may not be identified at a national scale (due to more extreme values in other areas) to be more clearly identified.

9.7 Concluding statement

Classification is an important first step in all research areas (e.g. biological taxonomy-species & evolution). The simplification of a complex dataset can make the previously unfathomable, easy to understand. The OA classification simplifies the complexity of the census into as few as seven area types, providing a clear and easy to interpret picture of the socio-demographics of the country. There is a very high level of demand for area classification at the smallest geographic scale. Small scale area classifications have been done in the past by commercial firms and academics, but not officially. This project represents the fist small scale official classification of the UK, thanks to its publication as a 'National Statistic'. The resulting product produces a fascinating picture of UK residential populations which will be heavily used in the remainder of the 2001-11 intercensal period.

From every ACORN grows an OAC

References

Aitchison, J. W. and Carter, H. (1994), *A Geography of the Welsh Language 1961-1991*, Cardiff, University of Wales Press.

Aitchison, J. W. and Carter, H. (2004), *Spreading the Word: The Welsh Language and the 2001 Census*, Talybont, Y Lolfa.

Aldenderfer, M. S. and Blashfield, R. K. (1984), *Cluster Analysis*, London, Sage.

Alder, J., Mayhew, L., Moody, S., Morris, R. and Shah, R. (2005), *The Chronic Disease Burden – An Analysis of Health Risks and Health Care Usage*, London, Cass Business School City University.

Altman, I. and Low, S. M. (1992), *Place Attachment*, New York, Plenum.

Anderberg, M. R. (1973), *Cluster Analysis for Applications*, New York, Academic Press.

Anderson, J. and Shuttleworth, I. (1994), Sectarian readings of sectarianism: interpreting the Northern Ireland Census, *The Irish Review*, 74-93

Anderson, J. and Shuttleworth, I. (1998), Sectarian demography, territoriality and political development in Northern Ireland, *Political Geography*, 17(2), 187-208.

Ashby, D. I. (2004), Linking geodemographic classifications to crime data - examples from Camden and the British Crime Survey, paper presented at *New Representations: The Use of Geodemographic Classifications in Research and Public Service Delivery*, 18-19 February 2004, University College London, London.

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

Bailey, S., Charlton, J., Dollamore, G. and Fitzpatrick, J. (1999a), Which authorities are alike? *Population Trends*, 98, 29-41.

Bailey, S., Charlton, J., Dollamore, G. and Fitzpatrick, J. (1999b), The ONS classification of local and health authorities of Great Britain: revised for authorities in 1999, *Studies in Medical and Population Subjects* No. 63, London, Office of National Statistics.

Bailey, S., Charlton, J., Dollamore, G., and Fitzpatrick, J. (2000), Families, groups and clusters of local and health authorities of Great Britain: Revised for authorities in 1999, *Population Trends*, 99, 37-52.

Baker, K. (1997), The utility to market research of ACORN - Foreword, *Journal of the Market Research Society*, 39(1), 53-54.

Ballas, D., Rossiter, D., Thomas, B., Clarke, G. and Dorling, D. (2005), *Geography Matters: Simulating the Local Impacts of National Social Policies*, York, Joseph Rowntree Foundation.

Batey, P. W. J. and Brown, P. J. B. (1995), From human ecology to customer targeting: the evolution of geodemographics. In Longley, P. and Clarke, G. P. Eds., *GIS for Business and Service Planning*, Cambridge, GeoInformation International.

BBC Online (2002), *Postcode Lottery in GP Services*, [online] http://news.bbc.co.uk/1/hi/health/2116336.stm, accessed 30/11/2004.

Berry, B. J. L. and Horton, F. E. (1970), *Geographic Perspectives on Urban Systems*, Englewood Cliffs, Prentice Hall.

Bibby, P. and Shepherd, J. (2004), *Developing a New Classification of Urban and Rural Areas for policy Purposes – the Methodology*, [online] http://www.statistics.gov.uk/geography/downloads/Methodology_Report.pdf, accessed 21/03/2005

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

Blackaby D. H. and Manning, D. N. (1990), The north-south divide: earnings unemployment and cost of living differences in Great Britain, *Papers of the Regional Science Association*, 69, 43-55.

Blake, M. and Openshaw, S. (1995), Selecting Variables for Small Area Classifications of 1991 UK Census Data, *Working Paper 95/2, School of Geography, University of Leeds*, [Online] http://www.geog.leeds.ac.uk/papers/95-2/, accessed 12/3/2003.

Booth, C. (1969), *Charles Booth's London*, London, Hutchinson.

Borooah, V. K., McKee, P. M., Heaton, N. and Collins, G. (1995), Catholic-Protestant income differences in Northern Ireland, *Review of Income and Wealth*, 41, 41-56.

Boyle, P. and Dorling, D. (2004), Guest editorial: the 2001 UK census: remarkable resource or bygone legacy of the 'pencil and paper era'?, *Area*, 36(2), 101-110.

Boyle, P., Halfacree, K. and Robinson, V. (1998), *Exploring Contemporary Migration*, Harlow, Longman.

Brewer, C. and Suchan, T. (2001), *Mapping Census 2000: The Geography of US Diversity*, *US Census Bureau, Census Special Reports, Series CENSR/01-1*, Washington DC, US Government Printing Office.

BRMB (2005), *GB TGI (Adults 15+)*, [online] http://www.bmrb-tgi.co.uk/main.asp?p=130andr=2127.734, accessed 14/12/2005.

Brown, M. (2005), Second homes in Cornwall, [Personal Correspondence by e-mail] 8/3/2005.

Bryson, B. (1995), *Notes from a Small Island*, London, Doubleday.

Callingham, M. (2003), Current commercial sector use of geodemographics and the implications for the ONS area classification systems, [Personal Correspondence by e-mail] 14/10/2003.

Callingham, M. (2005), From areal classification to geodemographics, paper presented at the *Demographic User Group Conference*, Royal Society, London 10th November 2005.

Canter, D. (1977), *The Psychology of Place*, London, Architectural Press.

Carter, H. (1995), *The Study of Urban Geography* 4th Ed., London, Arnold.

CCG (2005), *GB Profiler User Guide*, [online] http://www.geog.leeds.ac.uk/software/gbprofiles/, accessed 7/9/2005.

Champion, A. G., Fotheringham, A. S., Boyle, P., Rees, P. and Stillwell J. (1998) *The Determinants of Migration Flows in England: A Review of Existing Data and Evidence***,** Report prepared for the Department of Environment, Transport and Regions, [online] http://www.gog.leeds.ac.uk/publications/DeterminantsOfMigration/report.pdf, accessed 17/7/2005.

Champion, A., Wong, C., Rooke, A., Dorling, D., Coombes, M. and Brunsdon, C. (1996), *The Population of Britain in the 1990s: A Social and Economic Atlas*, Oxford, Clarendon Press.

Champion, A. G., Green, A. E., Owen, D. W., Ellin, D. J. and Coombes, M. G. (1987) *Changing Places: Britain's Demographic, Economic and Social Complexion.* London, Edward Arnold.

CIA (2005), *The World Factbook*, [online] http://www.cia.gov/cia/publications/factbook/, accessed 16/12/2005.

Cliff, A. and Ord, J. (1973), *Spatial Autocorrelation*, London, Pion.

Cook, L. (2004), The quality and qualities of population statistics, and the place of the census *Area*, 36(2), 111-123.

Core Cities Group (2005), *About the Core Cities Group*, [online] http://www.corecities.com/coreDEV/about.html, accessed 24/5/05.

Craig, J. (1984), Which Local authorities were like in 1981?, *Population Trends*, 26, 15-39.

Curry, M. R. (1997), The digital individual and the private realm, *Annals of the Association of American Geographers*, 87(4), 681-699.

Curry, M. R. (1998), *Digital Places: Living with Geographic Information Technologies*, London, Routledge.

Dale, A. (1993), An Overview, in Dale, A. and Marsh, C. Eds., *the 1991 Census User's Guide*, London, HMSO, pp1-15.

Dale, A., Fieldhouse, E. and Holdsworth, C. (2000), *Analysing Census Microdata*, London, Arnold.

Debenham, J. E. (2002), Understanding Geodemographic Classification: Creating The Building Blocks For An Extension, *Working Paper 02/1 School of Geography, University of Leeds*, [online] http://www.geog.leeds.ac.uk/wpapers/02-1.pdf, accessed 14/11/2002

Debenham, J. E. (2003), *Extending Geodemographics: New Small Area Classifications for Yorkshire and the Humber*, Unpublished PhD thesis, school of Geography, University of Leeds.

Debenham, J., Clarke, G. and Stillwell, J. (2003), Extending geodemographic classification: a new regional prototype**,** *Environment and Planning A*, 35(6), 1025-1050.

DEFRA (2004), *Social and Economic Change and Diversity in Rural England*, A report by the Rural Evidence Research Centre Birkbeck College, London, Department for Environment, Food and Rural Affairs. [online] http://www.defra.gov.uk/rural/pdfs/rwpreview/rwp_review_birkbeck1.pdf, accessed 7/8/2005.

Demographic Decisions (2005), *The Demographics User Group*, [online] http://www.demographic.co.uk/dug.html, accessed 13/12/2005.

DeSarbo, W. S., Carroll J. D., Clark, L. A. and Green, P. E. (1984), Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables, *Psychometrika*, 49(1), 57-78.

Dilley, R. (2003), *Just an average day in Average Town*, [online] http://news.bbc.co.uk/1/hi/magazine/3163773.stm, accessed 21/8/03.

Doran, T., Drever, F. and Whitehead, M. (2004), Is there a north-south divide in social class inequalities in health in Great Britain? Cross sectional study using data from the 2001 Census, *British Medical Journal*, 328, 1043-1045.

Dorling, D. (2004), Distressed times and area: poverty polarisation and politics in England, 1918-171, in Barker, A. R. H. and Billinge, M. Eds., *Geographies of England: The North-South Divide Imagined and Material*, Cambridge, Cambridge University Press.

Dorling, D. and Thomas, B. (2004), People *and Places: A 2001 Census Atlas of the UK*, Bristol, The Policy Press.

Eason, G. (2002), *Change to bounty for poor students*, [online] http://news.bbc.co.uk/1/hi/education/1950631.stm, accessed 12/2/2003.

Eisenhauer, B., Kannich, R. and Blahna, D. (2000), Attachments to special places on public lands: An analysis of activities, reason for attachments, and community connections, *Society and Natural Resources*, 13, 421-440.

Engstrom, D. M. (1997), The economic determinants of ethnic segregation in Post-War Britain, *Discussion Papers in Economic and Social History (12) Oxford University*, [online] http://www.nuff.ox.ac.uk/economics/history/paper12/12paper.pdf, accessed 5/9/2003.

EuroDirect (2005), *Cameo UK*, [Online] http://www.eurodirect.co.uk/Pages/CAMEO_UK.html, accessed 14/4/2004.

EuroDirect (Unknown), *Cameo Classifications*, [Promotional Pamphlet], Bradford, EuroDirect.

Eurostat (2005), *Eurostat themes*, [online] http://www.eustatistics.gov.uk/themes/index.asp, accessed 16/12/2005.

Evans, M. (2004), Confidentiality, data protection and marketing, *Cardiff Centre for Ethics, Law and Society, Monthly Publication Series*, [online] http://www.ccels.cf.ac.uk/pubs/evanspaper.pdf, accessed 14/9/2005.

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

Exeter, D., Boyle, P., Feng, Z., Flowerdew, R. and Schierloh, N. (2005), The creation of 'Consistent Areas Through Time' (CATTs) in Scotland 1981-2001, *Population Trends*, 119, 28-36.

Experían (2001), *GB MOSAIC*, [online] http://mimas.ac.uk/docs/experian/gbmosaic.pdf, accessed 23/10/2002.

Experían (2004), *Mosaic United Kingdom: Multimedia Guide* [CD], Nottingham, Experían.

Experían (2005), *Mosaic United Kingdom: The Consumer Classification for the UK*, [online], http://www.business-strategies.co.uk/Content.asp?ArticleID=566, accessed 12/6/2005.

Farr, M. and Singleton, A. (2004), Linking geodemographic classification to government administrative datasets: extending access to higher education, paper presented at *New Representations: The Use of Geodemographic Classifications in Research and Public Service Delivery*, 18-19 February 2004, University College London, London.

Feng, Z. and Flowerdew, R. (1998), Fuzzy geodemographics: a contribution from fuzzy clustering methods. In Carver, S. Ed., *Innovations in GIS 5*, London, Taylor and Francis.

Fieldhouse, E. (2000), Deprived people or deprived places? Exploring the ecological fallacy in studies of deprivation using the Sample of Anonymised Records, in Dale, A., Fieldhouse, E. and Holdsworth, C. Eds., *Analysing Census Microdata*, London, Arnold.

Flowerdew, R. (1990), Classified residential area profiles and beyond, *North West Regional Research Laboratory Research Report 18*, Lancaster University.

Flowerdew, R. and Leventhal, B. (1998) Under the Microscope, *New Perspectives*, 18, 16-38.

Forrest, R. and Kearns, A. (2001), Social Cohesion, Social Capital and the Neighbourhood, *Urban Studies*, 38(12), 2125–2143.

Fotheringham, A. S., Rees, P., Champion, A., Kalogirou, S. and Tremayne, A. R. (2004), The Development of a migration model for England and Wales: overview and modelling out-migration, *Environment and Planning A*, 36, 1633 – 1672.

Fujita, M. and Thisse J. (2002), Economics of Agglomeration: cities, Industrial Location and Regional Growth, Cambridge, Cambridge University Press.

Garland, K. (1994), *Mr Beck's Underground Map*, London, Capital Transport Publishing.

Gehlke, C. E. and Biehl K. (1934), Certain effects of grouping upon the size of the correlation coefficient in census tract material, *Journal of the American Statistical Association*, 29, 169-170.

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

Gnanadesikan, R., Tsao, S. L. and Kettenring, J. R. (1995), Weighting and selection of variables for cluster analysis, *Journal of Classification*, 12, 113-136.

Goldthorpe, J. H. (1987), *Social Mobility and Class Structure in Modern Britain* 2nd Ed., Oxford,
Clarendon Press.

Gordon, A. D. (1999), *Classification* 2nd Ed., London, Chapman and Hall.

Gordon, D. (1995), Census-based deprivation indices: their weighting and validation, *Journal of Epidemiology and Community Health*, 49 (Suppl. 2), 39-44.

Goss, J. (1995), Marketing the new marketing: the strategic discourse of geodemographic information systems, in Pickles, J. Ed., *Ground Truth: the Social Implications of Geographic Information Systems*, New York, Guilford Press.

Graham, B. and Shirlow, P. (1998), An elusive agenda: the development of a middle ground in Northern Ireland, *Area*, 30(3), 245-254.

Greenlees, C. (2004), *Enjoying the Curry Mile*, [online] http://www.manchesteronline.co.uk/food/s/115/115318_enjoying_the_curry_mile.html, accessed 5/9/2005.

Haggett, P., Cliff, A. D. and Frey, A. (1977), *Locational Analysis in Human Geography*, London, Arnold.

Harrigan, K. R. (1985), An application of clustering for strategic group analysis, *Strategic Management Journal*, 6, 55-73.

Harris, R. (1999), *Geodemographics and the Analysis of Urban Lifestyles.*, Unpublished PhD. thesis. School of Geography, University of Bristol.

Harris, R. (2001), The diversity of diversity: is there still a place for small area classifications?, *Area*, 33(3), 329-336.

Harris, R. (2003), Population mapping by geodemographics and digital imagery, in Mesev, V. Ed., *Remotely Sensed Cities*, London, Taylor and Francis, pp223-42.

Harris, R., Sleight, P. and Webber, R. (2005), *Geodemographics, GIS and Neighbourhood Targeting*, London, Wiley.

Hennell, T. (2005), Area Classifications, Personal communication by e-mail 18/11/2005

Higgs, S., Whitworth, A. and Charlton, J. (2002), Proposed Variables for 2001 Area Classification, London, Office of National Statistics, Personal communication.

Higgs, G., Williams, C. and Dorling, D. (2004), Use of the census of population to discern trends in the Welsh language: an aggregate analysis, *Area*, 36(2), 187-201.

Hohn, C. (1987), The family life cycle: needed extensions of the concept, in Bongaarts, J., Burch, T. and Wachter, K. Eds., *Family demography: methods and their application*, Oxford, Clarendon Press, pp65-80.

Home Office (1997), *Fires in the Home in 1995: Results from the British Crime Survey*, [online] http://www.homeoffice.gov.uk/rds/pdfs2/hosb997.pdf, accessed 13/12/2005.

Home Office (2005), *Crime in England and Wales 2004/2005*, [online] http://www.homeoffice.gov.uk/rds/crimeew0405.html, accessed 13/12/2005.

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

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

Jackson, S. (1998), *Britain's Population: Demographic Issues in Contemporary Society*, London, Routledge.

Jewell, H. M. (1994), *The North-South Divide: the Origins of Northern Consciousness in England*, Manchester, Manchester University Press.

Johnston, R. J. and Pattie, C. J. (1989), A growing north-south divide in British voting patterns, 1979-1987, *Geoforum*, 20(1), 93-106.

Jones, H. (1990), *Population Geography*, London, Paul Chapman.

Jones, H., Kettle, J. and Unsworth, R. (2004), The roofs over our heads: Housing supply and demand, in Unsworth, R. and Stillwell, J. Eds., *Twenty-first century Leeds: Geographies of a regional city*, Leeds, University of Leeds Press.

Kaufman, L. and Rousseeuw, P. J. (2005), *Finding Groups in Data*, Chichester, Wiley.

Kelly, F. (1969), Classification of the London Boroughs, *Greater London Council Research and Intelligence Unit*, 9, 13-19.

Kelly, F. (1971), Classification of urban areas, *Quarterly Bulletin of the Greater London Research and Intelligence Unit report 9*, London, Greater London Council.

Kohonen, T. (1984), *Self-organisation and Associative Memory*, Berlin, Spring-Verlag.

Kohonen, T. (1998), The self-organising map, *Neurocomputing*, 21, 1-6.

Kosko, B. (1994), *Fuzzy Thinking: The New Science of Fuzzy Logic*, London, Flamingo.

Levene, T. (1999), Postcode persecution. *The Guardian*, 6/2/1999.

Lewis, S. (1962), *Tynged yr Iaith*, London, BBC.

Linnæus, C. (1737), *Genera Plantarum*, Leiden, Linnæus.

Liverpool City Council (1969), *Social Malaise in Liverpool: Interim Report on Social Problems and their Distribution*, Liverpool, Liverpool City Planning Department.

Livingstone, D. N., Keane, M. C. and Boal, F. W. (1998), Space for religion: a Belfast case study, *Political Geography*, 17(2), 145-170.

Longley, P. A. (2003), Geographical Information Systems: developments in socio-economic data infrastructures, *Progress in Human Geography*, 27(1), 114–121.

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

Longley, P. A. and Batty, M. (1996), *Spatial Analysis: Modelling in a GIS Environment*, Cambridge, GeoInformational International.

Lorr, M. (1983), *Cluster Analysis for the Social Sciences*, San Francisco, Jossey-Bass.

LSE (2005), *Charles Booth Online Archive*, [online] http://booth.lse.ac.uk/, accessed 14/7/05.

Macourt, M. P. A. (1995), Using Census Data: Religion as a key variable in studies of Northern Ireland, *Environment and Planning A*, 27, 593-614.

Macrone, M. (1998), *A Little Knowledge*, London, Ebury Press.

Makarenkov, V. and Legendre, P. (2001), Optimal variable weighting for ultrametric and additive trees and k-means partitioning: Methods and software, *Journal of Classification*, 18, 245-271.

Marceau, D. J. (1999), The scale issue in social and natural sciences, *Canadian Journal of Remote Sensing*, 25(4), 347-356.

Martin, D. (1998), Optimizing census geography: the separation of collection and output geographies, *International Journal of Geographical Information Science*, 12, 673-685.

Martin, D. (2000a), Towards the geographies of the 2001 UK Census of Population, *Transactions of the Institute of British Geographers*, 25, 321-332.

Martin, D. (2000b), Census 2001: making the best of zonal geographies, paper presented at the *The Census of Population 2001 and Beyond*, University of Manchester 22-23 June 2003.

Martin, D. (2002a), Geography for the 2001 Census in England and Wales, *Population Trends* 108, 7-15.

Martin, D. (2002b), Output Areas for 2001, in Rees, P., Martin, D. and Williamson, P. Eds., *The Census Data System*, Chichester, Wiley, pp 37-46.

Martin, D. (2004), Neighbourhoods and area statistics in the post 2001 census era, *Area*, 36(2), 136-145.

Martin, D., Nolan, A. and Tranmer, M. (2001), The application of zone-design methodology in the 2001 UK Census, *Environment and Planning A*, 33, 1949-1962.

Massey, D. (1995), The conceptualization of place, in Massey, D. and Jess, P. Eds., *A Place in the World? Places, Cultures and Globalization*, Oxford, Open University, pp87-132.

McCarthy, H. H., Hook, J. C. and Knos, D. S. (1956), *The Measurement of Association in Industrial Geography.* Iowa City, Department of Geography, State University of Iowa.

McEldowney, O., McNair, D. and Lavery, S. (2004), Demographic Fear in Northern Ireland: Politics, Society and Space, paper presented at *PopFest, School of Geography, University of Leeds, 22nd – 24th June 2004.* [PowerPoint presentation available online] http://www.geog.leeds.ac.uk/conferences/popfest2004/McEldowney,%20McNair%20and%20Lavery.ppt.

Mckittrick, D. and McVea, D. (2001), *Making Sense of the Troubles*, London, Penguin.

Miles, J. and Shevlin, M. (2001), *Applying Regression and Correlation*, London, Sage.

Milligan, G. W. (1989), A validation study of a variable weighting algorithm for cluster analysis, *Journal of Classification*, 6, 53-71.

Milligan, G. W. (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.

Milligan, G. W. and Cooper, M. C. (1987), Methodological review: Clustering methods, *Applied Psychological Measurement*, 11, 329-354.

Milligan, G. W. and Cooper, M. C. (1988), A study of standardisation of variables in cluster analysis, *Journal of Classification*, 5, 181-204.

Monmonier, M. (1996), *How to Lie with Maps 2nd Ed*, Chicago, University of Chicago Press.

Moser, C. A. and Scott, W. (1961), *British Towns: A Statistical Study of their Social and Economic Differences*, Edinburgh, Oliver and Boyd Ltd.

Murtagh, F. (1996), Neural networks for clustering, in Arabie, P., Hubert, L. J. and De Soete, G. Eds., *Clustering and Classification*, Singapore, World Scientific.

Noble, M., Wright, G., Dibben, C., Smith, G. A. N., McLennan, D., Anttila, C., Barnes, H., Mokhtar C., Noble, S., Avenell, D., Gardner, J., Covizzi, I. and Lloyd, M. (2004), *Indices of Deprivation 2004*, Report to the Office of the Deputy Prime Minister. London: Neighbourhood Renewal Unit, [online] www.odpm.gov.uk/embedded_object.asp?id=1128446, accessed 14/8/2004.

Norman-Butler, B. (1972), *Victorian Aspirations: the Life and Labour of Charles and Mary Booth*, London, George Allan and Unwin.

Norris, P. and Evans, G. (1999), Understanding Electoral Change, in Evans, G. and Norris, P. Eds., in *Critical Elections: British Parties and Elections in Long-term Perspective*, London, Sage.

Norusis, M.J. (1985), *SPSS-X Advanced Statistics Guide*, New York, McGraw-Hill.

O'Leary, B. and McGarry, J. (1993), *The Politics of Antagonism: Understanding Northern Ireland*, London, Athlone Press.

ODPM (2002), *Development of A Migration Model*, [online] http://www.odpm.gov.uk/stellent/groups/odpm_housing/documents/page/odpm_house_601865.pdf, accessed 21/11/2002.

ODPM (2003), *National Community Safety Fire Tool Box: What Statistics can tell you*, [online] http://www.firesafetytoolbox.org.uk/ncfsc/foundationstones/firefactsandstatistics/what+statistics+can+tell+you.htm, accessed 12/2/2003.

ODPM (2004), *The English Indices of Deprivation 2004 (revised)*, [online] http://www.odpm.gov.uk/stellent/groups/odpm_urbanpolicy/documents/page/odpm_urbpol_029534.pdf, accessed 2/12/2005.

ONS (2000), *Framework for National Statistics*, [online] http://www.statistic s.gov.uk/about/national_statistics/downloads/FrameDoc1.pdf, accessed 1/1/2006.

ONS (2003a), *Geography and National Statistics*, [online] http://www.statistics.gov.uk/geography/, accessed 24/1/2003.

ONS (2003b), *Edit and Imputation - Evaluation Report*, [online] http://www.statistics.gov.uk/census2001/editimputevrep.asp, accessed 12/4/2004.

ONS (2004a), *About National Statistics and ONS*, [online] http://www.statistics.gov.uk/about/national_statistics/introduction.asp, accessed 1/1/2006.

ONS (2004b), *National Statistics 2001 Area Classification*, [online] http://www.statistics.gov.uk/geography/census_geog.asp accessed 14/12/2004, accessed 21/10/2005.

ONS (2005a), *Census 2001: Quality Report for England and Wales*, [online] http://www.statistics.gov.uk/downloads/census2001/census_2001_quality_report.pdf, accessed 12/12/2005.

ONS (2005b), *Beginners' guide to UK geography: Census Geography* [online] http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/, accessed 19/07/05.

ONS (2005c), *Rural and Urban Classification 2004*, [online] http://www.statistics.gov.uk/geography/nrudp.asp, accessed 27/7/05.

ONS (2005d), *Methods for National Statistics 2001 Area Classification for Local Authorities*, [online] http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/la/downloads/Methods.pdf, accessed 10/08/2005.

ONS (2005e), *Area Classification for Statistical Wards - Methods*, [online] http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/wards/downloads/area_classification_for_statistical_wards_methods.pdf, accessed 11/08/2005

ONS (2005f), *Applying to Use the Controlled Access Microdata Sample (CAMS)*, [online] http://www.statistics.gov.uk/census2001/sar_cams.asp, accessed 13/12/2005.

ONS (2005g), *Super Output Areas*, [online] http://www.statistics.gov.uk/geography/soa.asp, accessed 13/12/2005.

ONS (2005h), *NUTS*, [online] http://www.statistics.gov.uk/geography/nuts.asp, accessed 13/12/2005.

ONS/GROS/NISRA (2001), *2001 Census. Census Area Statistics. Geographic Levels: Output Area to National*, Titchfield, Office for National Statistics.

Openshaw, S. and Wymer, C. (1995), Classifying and regionalizing census data, in Openshaw, S. Ed., *Census Users' Handbook*, Cambridge, GeoInformation International.

Openshaw, S. (1984a), Ecological fallacies and the analysis of areal census data, *Environment and Planning A*, 16, 17-31.

Openshaw, S. (1984b), *The Modifiable Areal Unit Problem: Concepts and Techniques in Modern Geography no. 38*, Norwich, Geo Books.

Openshaw, S. and Gillard, A. A. (1978), On the stability of a spatial classification of census enumeration district data, in Batey, P. W. J. Ed., *Theory and Method in Urban and Regional Analysis*, London, Pion.

Openshaw, S. and Rao, L. (1995), Algorithms for reengineering 1991 Census geography, *Environment and Planning A*, 27, 425-446.

Openshaw, S. and Taylor, P. J. (1991), The Modifiable Areal Unit Problem, in Wrigley, N. and Bennett, R. J. Eds., *Quantitative Geography a British View*, London, Routledge, pp 60 - 70.

Openshaw, S. and Wymer, C. (1995), Classifying and regionalizing census data, in Openshaw, S. Ed. *Census Users' Handbook*, Cambridge, GeoInformation International, pp 239-270.

Openshaw, S. (1994), Developing smart and intelligent target marketing systems: part I, *Journal of Targeting, Measurement and Analysis for Marketing*, 2, 289-301.

Orford, S., Dorling, D., Mitchell, R., Shaw, M. and Davey Smith, G. (2002), Life and death of the people of London: a historical GIS of Charles Booth's inquiry, *Health and Place*, 8, 25-35.

Park, R. E. and Burgess, E. W. (1925), *The City: Suggestions for Investigation of Human Behavior in the Urban Environment*, Chicago, University of Chicago Press.

Peach, C. (1996), Does Britain have ghettoes? *Transactions of the Institute of British Geographers*, 21, 216-235.

Pfautz, H. W. (1967), *Charles Booth on the City Physical Pattern and Social Structure*, Chicago, University of Chicago Press.

Pinker, S. (2004), *How the Mind Works,* London, Penguin.

Poole, M. A. and Boal, F. W. (1973), Religious residential segregation in Belfast in mid-1969: a multi-level analysis, in Clark, B. D. and Gleave, M. B. Eds., *Social Patterns in Cities*, Oxford, Alden and Mowbray.

Raper, J., Rhind, D. Shepherd, J. (1992), *Postcodes: The New Geography*, London, Longman.

Rees, P. (1970), Concepts of social space, in Berry, B. and Horton, F. Eds., *Geographic Perspectives on Urban Systems*, Englewood Cliffs: New Jersey, Prentice Hall, 306-394.

Rees, P. (1982), The Welsh language: a geographical description, *Working Paper 329, School of Geography, University of Leeds*, Leeds.

Rees, P. and Martin, D. (2002), The debate about census geography, in Rees, P., Martin, D. and Williamson, P. Eds., *The Census Data System*, Chichester, Wiley, pp1-24.

Rees, P., Denham, C., Charlton, J., Openshaw, S., Blake, M. and See, L. (2002), ONS classifications and GB Profiles: census typologies for researchers, in Rees, P. Martin, D. and Williamson, P. Eds., *The Census Data System*, Chichester, Wiley, pp149-170.

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

Rees, P. H. (2004), Bangor fieldtrip, personal communication, 14/10/2004.

Rees, P. H. (1998), What do you want from the 2001 Census? Results of an ESRC/JISC survey of user views, in *Environment and Planning A*, 30, 1775-1796.

Rees, P., Martin, D. and Williamson, P. (2002), Census data resources in the United Kingdom, in Rees, P. Martin, D. and Williamson, P. Eds. *The Census Data System*, Chichester, Wiley, pp28-36.

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*.

Rex, J. and Moore, R. (1967), *Race, Community and Conflict: A Study of Sparkbrook*, London, Oxford University Press.

Robbins, S. (2005), Targeting Energy Efficiency Measures, paper presented at *Demographic User Group Conference*, Royal Society, London 10th November 2005.

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

Robinson, V. (1981), The development of South Asian Settlements in Britain and the myth of return In C. Peach, V. Robinson and Smith, S. Eds., *Ethnic Segregation in Cities*, London, Croom Helm.

Robinson, W. S. (1950), Ecological Correlations and the Behavior of Individuals, *American Sociological Review*, 15 (3), 351–357.

Robson, B. T. (1971), *Urban Analysis: A Study of City Structure*, Cambridge, Cambridge University Press.

Romesberg, H. C. (2004), *Cluster Analysis for Researchers*, North Carolina, Lulu Press.

Rossi, P. (1955), *Why Families Move: A Study in the Social Psychology of Urban Residential Mobility*, Glencoe: Illinois, Free Press.

Rothman, J. (1989), Editorial, *Journal of the Market Research Society*, 31(1), 1-5.

Rose, G. (1995), Place and Identity: A Sense of Place, in Massey, D. and Jess, P. Eds., *A Place in the World? Places, Cultures and Globalization*, Oxford, Open University, pp87-132.

Rummel, R. J. (1970), *Applied Factor Analysis,* Evanston, Northwestern University Press.

Russett, B. M. (1967), *International Regions and the International System: A Study in Political Ecology*, Chicago, Rand and McNally.

Select Committee on Envir onment, Food and Rural Affairs (2002), *Position Statement on Rural Definitions*, [Online] http://www.publications.parliament.uk/pa/cm200102/cmselect/cmenvfru/386/1112114.htm, accessed 13/12/2005.

Shepherd, P., Stillwell, J. C. H. and Clarke, G. (2005), Neighbourhood profiling and classification for community safety. paper presented at *RGS-IBG Annual International Conference 2005*, London, 31 August to 2 September 2005.

Shevky, E. and Bell, W. (1955), *Social Area Analysis: Theory*, *Illustrative Application and Computational Procedures*, Stamford, Stamford University Press

Shevky, E. and Williams, M. (1949), *The Social Areas of Los Angeles: Analysis and Typology*, Berkeley, University of California Press.

Simey, T. S. and Simey M. B. (1960), *Charles Booth: Social Scientist*, Oxford, Oxford University Press.

Simpson, E. H. (1949). Measurement of diversity, *Nature*, 163:688.

Simpson, S. (2002), Dealing with the census undercount, in Rees, P., Martin, D. and Williamson, P. Eds., *The Census Data System*, Chichester, Wiley.

Sleight, P. (2004) *Targeting customers: How to Use Geodemographic and Lifestyle Data in Your Business*, Henley-on –Thames, World Advertising Research Centre.

SPSS Inc. (2001), *SPSS 11.0 Syntax Reference Guide: Volume 1*, Chicago, SPSS.

SPSS Inc. (1999), K-means cluster analysis, in SPSS Inc., *SPSS Base 9.0*, *User's Guide.* Chicago, SPSS Inc, pp. 333-339.

Stillwell, J. (2005), *Ethnic Segregation Index*, Personal Correspondence, [e-mail] 7/03/2005.

Thomas, D. S. (1938), *Research Memorandum on Migration Differentials*, New York, Social Science Research Council.

Timms, D. (1971), *The Urban Mosaic*, Cambridge, Cambridge University Press.

Tobler, W. (1970), A computer movie, *Economic Geography*, 46, 234-40.

Townsend, P., Phillimore, P. and Beattie, A. (1988), *Health and Deprivation: Inequality and the North*. London, Croom Helm.

Tranmer, M. and Steel, D. G. (1998) Using census data to investigate the causes of the ecological fallacy, *Environment and Planning A*, 30, 817-831.

Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M., Martin, D., Steel, D. and Gardiner, C. (2005), The case for small area microdata, *Journal of the Royal Statistical Society A*, 168(1), 29-49.

Tyron, R. C. (1939), *Cluster Analysis*. Ann Arbor, Edwards Brothers.

UN Statistics Division (2005), *Demographic Yearbook System*, [online] http://unstats.un.org/unsd/demographic/products/dyb/dyb2.htm, accessed 16/12/2005.

Unsworth, R. and Stillwell J. (2004), Leeds: Premier City, Regional Capital, in Unsworth, R. and Stillwell, J. Eds., *Twenty-first Century Leeds Geographies of a Regional City*, Leeds, University of Leeds Press.

Vickers, D. (2003), The difficulty of linking two differently aggregated spatial datasets: using a look-up table to link postal sectors and 1991 Census enumeration districts, *Working Paper 03/2, School of Geography, University of Leeds*. [online] http://www.geog.leeds.ac.uk/wpapers/03-2.pdf, accessed 14/8/2005.

Vickers, D. Rees, P. and Birkin, M. (2003), A New Classification Of UK Local Authorities Using 2001 Census Key Statistics, *Working Paper 03/3, School of Geography, University of Leeds*, [online] http://www.geog.leeds.ac.uk/wpapers/05-3.pdf, accessed 14/8/2005

Vickers, D. Rees, P. and Birkin, M. (2005), Creating the National Classification of Census Output Areas: data, methods and results, *Working Paper 05/2 School of Geography, University of Leeds*, [online] http://www.geog.leeds.ac.uk/wpapers/05-2.pdf, accessed 14/8/2005.

Vickers, D. and Stillwell, J. (2005), Area classification in Yorkshire and the Humber: a region of diversity, *The Yorkshire and Humber Regional Review*, 15(2), 10-12.

Voas, D. and Williamson, P. (2001a), The diversity of diversity: a critique of geodemographic classification, *Area*, 33(1), 63-76.

Voas, D. and Williamson, P. (2001b), Response (The diversity of diversity), *Area*, 33(3), 335-336.

Wallace, M. and Denham, C. (1996) The ONS Classification of Local and Health Authorities of Great Britain. *Studies on Medical and Population Subjects 59*, HMSO, London

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

Ward, J. H. (1963), Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 58, 236-244.

Warnes, A. (1992), Migration and the life course, in Champion, A. and Fielding, A. Eds., *Migration Processes and Patterns, Volume 1 Research Progress and Prospects*, London, Belhaven, 175-187.

Wathan, J., Holdsworth, C. and Leeser, R. (2004), Alternative household classifications for the 2001 Census, *Environment and Planning A*, 36, 1101-1123.

Webber, R. (1977), An introduction to the national classification of wards and parishes, *Planning Research Applications Group Technical Paper No. 23*, London, Centre for Environmental Studies.

Webber, R. and Craig, J. (1976), Which local authorities are alike?, *Population Trends*, 5, 13-19.

Webber, R. and Craig, J. (1978), Socio-economic classifications of local authority areas, *Studies in Medical and Population Subjects 35*, London, OPCS.

Webber, R. and Farr, M. (2001), MOSAIC-From an area classification system to household classification, *Journal of Targeting, Measurement and Analysis for Marketing*,10(1).

Webber, R. (2004) Reasons for non-uniform swing in British general elections 1992-2001: how important is demographics, paper presented at *New Representations: The Use of Geodemographic Classifications in Research and Public Service Delivery*, 18-19 February 2004, University College London, London.

Webber R. (2004), Designing geodemographics to meet contemporary business needs, *Interactive Marketing*, 5(3), 219-237.

Wiess, M. J. (1988), *The Clustering of America*, New York, Harper Row.

Wiess, M. J. (2000), *The Clustered World*, New York, Little Brown.

Williamson, P. (2005), Len Cook: hero or zero of the 2001 Census? A look at the impact of disclosure control on aggregate census outputs, paper presented at *National Statistics Disclosure Control: Now and the Future*, *BSPS day meeting*, 11/1/2005.

Winterman, D. (2004), *Targeting your letter box*, [online] http://news.bbc.co.uk/1/hi/magazine/3752956.stm, accessed 9/9/2005

Wrigley, N. (1995), Revisiting the Modifiable Areal Unit Problem and the Ecological Fallacy, in Cliff, A. D., Gould, P., Hoare, A. and Thrift, N. Eds., *Diffusing Geography: Essays for Peter Haggett*, Oxford, Blackwell.