WORKING PAPER 05/2

CREATING THE NATIONAL CLASSIFICATION OF CENSUS OUTPUT AREAS: DATA, METHODS AND RESULTS

Daniel Vickers, Philip Rees and Mark Birkin

Version 1.0

PUBLISHED JULY 2005

School of Geography, University of Leeds
Leeds LS2 9JT, United Kingdom

This Working Paper is an online publication and so may be revised in the light of comments of readers and users of the OA classification. We have given a Version number to the Working Paper in anticipation of such feedback, in the spirit of the very successful online encyclopedia, the Wikipedia (http://en.wikipedia.org/wiki/Main_Page).

Our full contact details are:

Mail address:
School of Geography
University of Leeds
Leeds LS2 9JT
United Kingdom
Fax: +44 (0)113 343 3308

Email and telephone:
Daniel Vickers
d.vickers@geog.leeds.ac.uk
+44 (0)113 343 3348

Philip Rees
p.h.rees@leeds.ac.uk
+44 (0)113 343 3341

Mark Birkin
m.h.birkin@leeds.ac.uk
+44 (0)113 343 6838

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The purpose of this paper 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 project was carried out on behalf of the Office for National Statistics (ONS) by Daniel Vickers of the School of Geography, University of Leeds as part of his PhD. thesis. The paper describes the creation of the classification: selection of the variables, assembly of the classification database, the methods of standardisation and the clustering procedures, 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.

The OA Classification fits into the ONS suite of area classifications complementing published classifications at Local Authority, Health Authority and Ward levels. The classification is freely available, and can be downloaded from the ONS Neighbourhood Statistics website at www.statistics.gov.uk.

ACKNOWLEDGEMENTS

The authors wish to thank the Economic and Social Research Council (ESRC) and the Office for National Statistics (ONS) who provided funding for this project in the form of a CASE PhD studentship for Daniel Vickers, supervised by Phil Rees and Mark Birkin (ESRC Ref: PTA-033-2002-00067). We are grateful to those who have offered help and support throughout the duration of the project. Graham Clarke and John Stillwell of the School of Geography, University of Leeds provided many valuable comments and suggestions on the project in their roles as Research Support Group members. Many thanks must go to the team at ONS, especially John Charlton and Andrew Botterill. Samantha Poole, Nick Ormiston-Smith, Susan Williams, Nargis Rahman, Gregg Philpotts, Simon Compton, Alison Whitworth and Chris Denham also provided valuable input into the project. We are grateful to Martin Callingham of the School of Geography, Birkbeck College London for help and advice and to Keith Dugmore of Demographic Decisions for all his support and enthusiasm for the project. We are grateful to Malcolm Brown of Cornwall County Council for his knowledge of housing in Cornwall and to everyone who took part in the classification consultation exercise so willingly. We hope that the final product does justice to their enthusiasm.

The 2001 Census Key Statistics data used to create the OA classification are produced by the Office for National Statistics, the General Register Office for Scotland (GROS) and the Northern Ireland Statistics and Research Agency (NISRA) and are Crown Copyright. The School of Geography at the University of Leeds holds a “click-use” licence from Her Majesty’s Stationery Office for the use of census data in value added products. The 2001 Census data were extracted using the CASWEB interface of the Census Dissemination Unit, Manchester Computing, University of Manchester, which is supported through the ESRC/JISC Census of Population Programme. Some of the data used were checked with the equivalent holdings at the ONS Census 2001 data library (http://www.statistics.gov.uk/census/), GROS Scotland’s Census Results OnLine (SCROL) data library (http://www.scrol.gov.uk/scrol/common/home.jsp) and NISRA’s Northern Ireland Census of Population data library (http://www.nisra.gov.uk/census/start.html).

Ordnance Survey mapping was accessed via the EDINA service of the Edinburgh University Data Library and used under the Digimap-Ordnance Survey sub licence agreement (see http://edina.ac.uk/digimap/osterms.html). SPOT satellite image was extracted from the database maintained by Manchester Information and Associated Services (MIMAS), Manchester Computing and used under licence via the Combined Higher Education Software Team (CHEST) agreement (see http://www.mimas.ac.uk/spatial/satellite/spot/licensing.html). Collins Bartholomew GB 1:200,000

raster map was extracted from the database maintained by Manchester Information and Associated Services (MIMAS), Manchester Computing and used under licence via the Combined Higher Education Software Team (CHEST) agreement (see http://www.eduserv.org.uk/chest/datasets/barts/).

EXECUTIVE SUMMARY

The rich and varied social geography of the United Kingdom has been captured through the creation of an area classification using variables from the 2001 Census of Population in a joint project carried out by the School of Geography at the University of Leeds and the Methods Division of the Office for National Statistics. The work was supported by the Economic and Social Research Council through a CASE studentship award.

The typology classifies all of the Output Areas (OAs) in the 2001 Census of the UK into a set of 7 super-groups, 21 groups and 52 sub-groups, which are linked in a hierarchy. Readers can access the OA classification on the National Statistics website from August 2005 via
http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/default.asp.
You can investigate in which OA cluster you live. You can download spreadsheets containing the cluster memberships of OAs in geographical sections of the country or databases containing the full classification for the UK and the accompanying input data.

This paper provides an account of the data and methods used to create the classification.

- *Section 2* (pages 3-5) tells you what **output areas** are and how they were created.
- *Section 3* (pages 6-21) describes the **variables** that were used to capture the socio-demographic character of output areas and how they were chosen.
- *Section 4* (pages 22-27) provides an account of how the data for OAs were accessed and extracted and how this **database assembly** was thoroughly checked.
- *Section 5* (pages 28-33) reviews briefly the methods that can be used for **clustering processes**, covering both hierarchical and optimizing methods and explaining why a k-means algorithm was preferred to a hierarchical.
- *Section 6* (pages 34-45) explains the **key methodological innovation** of the classification: how a k-means algorithm was used to create a hierarchical classification.
- *Section 7* (pages 46-49) describes the **creation of the classification** using this innovation.
- *Section 8* (pages 50-60) discusses the **naming and describing of the clusters**. Names are provided for the upper and middle tier clusters and codes for the lower. Only the codes are given for clusters by National Statistics. **Table 16 (page 53)** links the codes to names produced by the Leeds team.
- *Section 9* (pages 61-68) shows ways in which the output area classification can be **mapped** to maximize its utility and to avoid some of the well known problems of census mapping.

1. INTRODUCTION

The purpose of this paper 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. The 2001 Area Classifications place each geographic area into a group according to key characteristics of the people who live in each area. The groups are created using a statistical technique known as cluster analysis. This classification of output areas fits into the ONS suite of area classifications and follows the publication of classifications at local authority, health board and ward levels (ONS 2004). Clustering more than two hundred thousand output areas into 7, 21 and 52 clusters, using 41 attributes achieves a massive simplification of the original data. This simplification aids the recognition of patterns and relationships in many ways, which can be explained in more detail. The classification has the ability to be used by anyone and in many different ways. It can be used to answer simple questions or to aid more complex analysis. The classification has a general purpose and may not fit specific research aims. However, it should be a useful starting point for many specific research investigations.

The placing of areas into groups based on the socio-economic characteristics is not a new idea and can be traced as far back as Charles Booth's survey of the life and labour of the people of London (carried out 1886-1903); The maps he created can be found on the Charles Booth Online Archive housed at the London School of Economics (see 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. Booth actually found that the problem was worse than anyone had thought (Simey and Simey 1960). So over 100 years on what has changed? Some may say little has changed. Excluding the significant redevelopment of docklands, the poorest people of London still live in the same areas as they did when Booth undertook his survey over 100 years ago (Orford *et al.* 2002).

Despite the relatively static positions of areas in the socio-economic hierarchy, change has occurred. The classification provides a picture of the social geography of the UK at the start of the twenty first century just as Booth did for London at the beginning of the previous century. This classification is a unique and exciting development in the investigation of the social make up of the UK. It is the first whole UK social classification to be freely available and fully documented. It is hoped that the classification will aid understanding of the socio-geographic make up of our society and provide a research tool for analysing the inequalities that are still present in the modern United Kingdom.

This form of socio-geographic analysis is now more relevant than ever as we start a new millennium. The UK is a different place from the country that Charles Booth surveyed. We now live in a multicultural society with changing and evolving social patterns. Area classification, often called 'Geodemographics' by producers of commercial systems such as ACORN, Mosaic or Cameo, has found new users in social research. Several research projects are ongoing at both the University of Leeds and University College London, making use of Area Classification to understand such issues as crime, community safety and access to higher education. A conference entitled; *New Representations: The use of Geodemographic Classifications in Research and Public Service Delivery*, took place in London in March 2003 and this year 2005,has seen the release of a major new book on the subject, *Geodemographics: GIS and Neighbourhood Targeting* by Richard Harris, Peter Sleight and Richard Webber.

The paper introduces output areas (OAs) and their geography in section 2. Section 3 reviews the processes of variable selection and the selection criteria for the inclusion and exclusion of variables from the classification process. Section 4 describes the assembly of the database to be clustered and the quality assurance checks that were made. Section 5 gives a brief review of some of the clustering processes that were considered for use. Methods of both standardisation and clustering techniques are described here. Section 6 describes the methodology that was followed in the clustering procedure. We describes both the original methodology which was subsequently rejected and the alternative final methodology that was used. Section 7 describes the creation of the classification and how the selection of clusters was achieved. Section 8 names and describes the clusters for interpretation. Section 9 displays maps and visualisations of the final classification. Section 10 concludes and reflects on the success of the classification.

2. INTRODUCING OUTPUT AREAS

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

2.1. Output Area design

Output Areas (OAs) 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 Enumeration Districts (EDs) from the 1981 Census so that comparisons could easily be made. This work involved converting a set of addresses which constituted the Royal Mail's unit postcode into a territory on the map. A layer showing the main topographic features (roads, railways, rivers, fences, walls 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) foe England and Wales but it was felt to be too labour intensive and too expensive for implementation in a country with ten times the population (ONS 2005a).

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™) using a Thiessen polygon algorithm available in the GIS (ESRI's ARC). Thiessen polygons allocate territory to the nearest defined set of points. These generated straight line boundaries, which were improved by linking (clipping) to other ONS boundaries (e.g. EDs) which followed more natural landscape features (Martin 1997 and 2002b).

Martin's innovation was to adopt a zone design algorithm developed by Stan Openshaw (Openshaw and Gillard 1978 and 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 1997, 1998, 2000a, 2000b, 2002a, 2002b, Martin *el 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 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. Table 1 shows how these different methodologies have affected the number and size of OAs that have been produced in each country (ONS 2005a).

Table 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

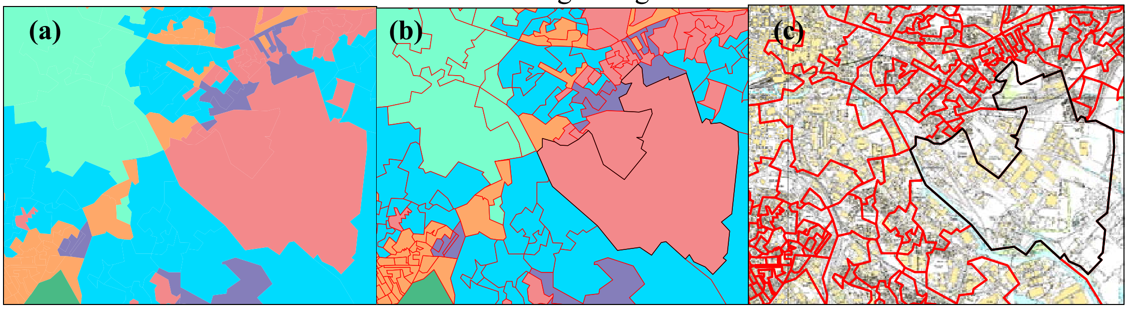
There are many issues relating to how OA boundaries divide up the country. Should the whole of a small settlement be included in one OA or should they bb split and combined with a hinterland of dispersed farmsteads? The first solution tends to create doughnut OAs out of the rural hinterland, while the 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 that is produced. The tower block is given its own OA regardless of the social make up of its inhabitants and creates OAs which have vary high population densities. These large multi-storey dwellings OAs often appear as outliers in the classification.

Buildings with empty tenancy or non residential function can be a problem in the creation of OAs as they can take up a large area even in an urban setting but do not represent many more people. When data are mapped to represent each OA, the geographically larger OAs dominate the map even though they can 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 as in rural areas. Good local knowledge of the area being looked at is often required.

Figure 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 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) then reveals that the majority of the area of the large OA is in fact made up of a non-residential/industrial area. Even though the industrial area contains no people it is assigned to an OA as the OA are designated to provide 100% coverage of the UK. Therefore, it is often the geographically largest OA which represent the fewest number of people, they simply had to be stretched that far to reach the minimum size threshold. Conversely it is the smallest OAs (in terms of area) which often represent the most people as they live in large residential dwellings than cannot be split. The most populous OA in the UK, the University of Lancaster campus containing 4,156 people on census day, could not be split as it has a single postcode.

Figure 1: OA boundaries for an area of Leeds overlain on Ordnance Survey 1:10,000 mapping. (a) with OA boundaries dissolved, (b)with OA boundaries marked, (c) with OA boundaries and street/building background.



©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 there boundaries and therefore constitutes part of an OA.

3. VARIABLE SELECTION

The purpose of this section is to discuss the choice of variables for analysis in the OA level 2001 Area Classifications. The results of any classification exercise will of course depend on the variables selected.

The variables were chosen 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 that is available in the UK (Rees *et al.* 2002): no other dataset contains anything like as much information with such a comprehensive geographic coverage. Another important factor is the scale of the data. At present the only data available at OA level derive from the 2001 Census so that to use data from other sources would require aggregating the data from other scales to OA scale. Linking of datasets at different spatial scales would create all kinds of reliability issues. For a discussion of the dangers of linking differently aggregated datasets Vickers (2003).

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. The underlying objective in variable choice is to select the minimum number of variables that will adequately represent the main dimensions in the census data (Bailey *et al.* 1999 and 2000).

3.1. Initial set of variables considered

Five main domains were identified within the Census with the intention to represent these as fully as possible within the classification. The five identified domains are: Demographic Structure, Household Composition, Housing, Socio-Economic, and Employment. The variables will be discussed within these groups in the rest of this section.

The Key Statistics have already been identified as being the most important variables so the initial data set 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 aids 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 to represent the main domains of the Census, with the intention that it would be reduced significantly following detailed assessment of each variable. Variables from all Key Statistics tables were considered for use. Some variables were merged to create composite variables; for example, the variable. Indian, Pakistani and Bangladeshi' represents people identifying as Indian, Pakistani or Bangladeshi. The initial set of variables considered is listed in Table 2.

Table 2: The initial set of variables considered for inclusion in the classification

No: Variable	No: 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 employ
4: Population Density	51: % Process, plant and machine operatives employ
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 and 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 and 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 employ	79: % LA Rented
33: % Mining and 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 and retail trade, repair of vehicles employ	83: % With an occupancy rating of -1 or less
37: % Hotels and catering employment	84: % No Central heating
38: % Transport, storage and communication employ	85: % No Bath or shower
39: % Financial intermediation employment	86: % Lowest floor level above the ground
40: % Real estate, renting and business activities employ	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. and technical occupations employ	93: % All Pensioner households (other)
47: % Administrative and secretarial occupations employ	94: % No adult in employment with dependant children

Note: employment shortened to employ in some cases

3.2. Reducing the initial set of variables

Variable selection for the OA classification was done in conjunction with that for the ward level 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 were the same (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. A number of reasons for inclusion were formulated which resulted in the guidelines set out in sections 3.2.1 – 3.2.8

3.2.1. Highly correlated variables

Strong correlations within a dataset are undesirable for cluster analysis, as they represent data redundancy. Each set of highly correlated variables repeats a lot of the information that is contained within just one variable. Including highly correlated variables makes it very hard to gauge the effect of any individual variable on the clustering process. A number of strong correlations were found in the initial set of variables. Table 3 shows a list of variable pairs from the original list that are correlated at 0.7 or above (i.e. redundancy of 49%+). Looking down the list at the variables which are highly correlated, the pairs of variables are perhaps not surprising. However, there are actually three different types of correlation visible. The first are pairs of variables which share the same denominator, so 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 the 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 the pattern will still be seen but not to the same extent.

Table 3: Highly correlated variables from the original variable list (ordered by data redundancy)

Variable	Variable	Correlation	Redundancy
1: % Male	2: % Female	-1.00	100%
15: % of people Indian, Pakistani and 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 who's 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 who's health is not good	0.89	79.21%
22: % of people who's health is good	23: % of people who's health is fairly good	-0.88	77.44%
45: % Professional occupations employment	56: % Higher professional occupations	0.88	77.44%
22: % of people who's health is good	24: % of people who's 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 and 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 who's 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 and 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 second type of correlation consist of those that are inherently connected due to causality i.e. one is fundamentally a property of the other, but they don't share the same denominator. An example of this is the pair of variables that is third in the list on Table 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.

The third type of correlation is made up of correlations between variables where the presence of one 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 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 of 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. These 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. Some correlations of this kind can be more surprising than others.

Table 3 shows several correlations of all three types. Common sense would suggest that one of each pair of highly correlated variables should be removed as much of the information held by the second is redundant as it can be inferred from the first. However, 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 to be put in the classification (Voas and Williamson, 2001). Evidence that they can predict the value of other variables suggests that if they were included in the classification they 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 some inclination to retain a high proportion of highly correlated variables as they are 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 a high level of data redundancy. Correlations between variables must be carefully examined; highly correlated variables that share the same denominator e.g. male and female must see at least one of the variables dropped as this is not evidence of predictive power but of a closed system in which the correlation is inevitable. Highly correlated variables that do not share the same denominator 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.

3.2.2. Variables with badly behaved distributions

Methods of clustering and standardisation work reliably with data that have a normal distribution. This is not a problem for the majority of variables as they tend to be normally distributed, but highly skewed distributions can create problems in the 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 are when a variable only identifies very small sectors of the population. Another way of 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.

So what is the reason that these distributions are so troublesome? This can be shown with the use of some census data. Let us 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 having a value above this being 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 everything with a value of 0 (88%) in one group and everything else (12%) in another. The important point to remember here is that 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. However there are other ways of splitting the data as 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 size groups as 99.7% of the data is below 50. So what would happen if this variable were used in a classification using the k-means procedure? The easiest way to test this out was to run a simple cluster analysis, which 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 (97.7%) in one group and 5,165 (2.3%) in the other, the point at which they have been split is above 15.2 which does not reflect any actual split in the data there is no reason why 15.2 and 15.3 should be in different groups. 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 as the data already suggests how it should be split.

By increasing the number of groups to be produced the extent of the problem will be come apparent. For purposes of illustration the data will be clustered into 50 clusters (a number that is not unusual for the lower level of a classification). If the data were normally distributed we would expect to find about 2% of the areas in each cluster but remember that this variable has 88% 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 groups. What we find is what we expected, one group contains 88% of the data and the rest are spread about with 27 of the groups containing less than 100 areas, split evenly each group should have around 4,500 members. Groups with small memberships are the real problem here, as this is how outliers are formed. If several highly skewed variables are included in the classification and if 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 would be of little practical use or value.

One solution to this kind of problem is to of 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).

3.2.3. Composite variables

Composite variables can be formed from two related variables which show comparable patterns. These variables would be those which 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 represent only 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 one single all flats variable. This increases the sample size of people on which the variable is based and increases the reliability of the data. This is especially important when working with areas as small as OAs as the numbers can be small and affected by disclosure controls put on the data. For an explanation of what disclosure control entails and the effect it has on the data see ONS (2005b)

3.2.4. Geographic consistency 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 having an 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 introduced to the Census in England and Wales for the first time in 2001 whereas it has previously been asked in Scotland and Northern Ireland. However it was only introduced in England and Wales as a voluntary question. Consequently 7.71% of the population of England and Wales did not answer the question. As it is a compulsory question in Scotland and Northern Ireland, this makes it difficult to compare the variables across the UK as some high rates of religious affiliation in observed in Scotland and Northern Ireland would be attributable to the voluntary nature of the question in England and Wales. Although some interesting patterns maybe visible, if data are not available for all parts of the UK, it is not possible to include the variable in the classification.

3.2.5. Vague or uncertain variables

It would seem common sense to assume that all the census variables are collated in the same way i.e. from the answers written on each census form. However, that is not the case for all variables. Examples are the 'household spaces with no residents' variables on table KS16 which 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 as 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 of what they saw. It is widely accepted that 'Second residence/holiday accommodation' was under recorded by 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 figure fell from 11,550 in 1991 to 10,500 in 2001 which seems highly unlikely with the continuing trend for people to buy second homes in the area over that period. 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 census (Brown 2005). The posting back of census forms could account for some of this as forms delivered to second homes would only be sent back if the own happened to be there at the time. The homes may have been imputed as permanent residences. Brown cautiously suggests that there are at least 50% more second homes in Cornwall than were picked up by the 2001 Census.

3.2.6. Uninteresting geographic distributions

For variables to work in the classification they need to show variation over space; otherwise a distinction between areas cannot be made. If we take as an example the ethnic group variables, 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?' Peach investigates to what extent different ethnic groups are dispersed throughout Britain. Table 4 shows the percentage of each ethnic group present in the major metropolitan areas of England.

Table 4: The percentage of each ethnic group present in the major metropolitan areas of England (London, W Midlands, G Manchester and 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 from Peach 1996 p219 source: OPCS

Table 4 shows that for every group apart from White and Chinese over 60% of that group are found in the four major urban centres, showing that these ethnic groups have a distinct pattern to their distributions. Whereas the 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 as their distributions vary more over space. This research is brought up to date by Stillwell 2005 who investigated the segregation of ethnic groups in Britain using data from the 2001 Census. The segregation indexes that were calculated showed that the Chinese were 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.

3.2.7. 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 as an ONS product until a new classification is produced, which is unlikely to be until 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 when the data was created or the changing membership of the EU.

3.2.8. Standardisation of illness by age

The limiting long term illness rate (percent of the population suffering limiting long-term illness) as provided in the Key Statistics could have been used in its raw form. However, it was considered that this was unsatisfactory as crude rates are greatly affected by the age structure of the population. 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 of this will be greater for OAs than for higher level geographies, because of their relatively small size increasing 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 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 or SIR. SIR works by comparing the expected illness count for an area with the observed count. The expected count is the created from age-specific illness rates for the whole UK population. By doing this for all areas and summing them we can then see if the illness rate is higher or lower than expected.

The SIR for an area is defined as follows:

$$SIR^i = 100 \times (I^i/\sum_a r_a^n P_a^i) \qquad (1)$$

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, that OA experiences 50% more illness than it would have if the age-specific rates for the standard population (the UK) applied to its local age distribution. 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.

3.3. The process of variable selection

The aim, as discussed previously is to represent the main dimensions of the census with the minimum number of variables. 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. 94 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 in 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 marketed 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. Within the process of variable selection some sacrifices were made at one scale to aid comparability with another scale. 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 1. 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 several different types of activity, including care homes, hostels, prisons and university residences. These are very different types of people with little in common who would be grouped together with the inclusion of this variable.

As a 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. Because of the inter dependency within the age variables, ages 25-44, 45-64 and 65+ were all kept.

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. 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 1 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 1 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.

3.4. The final list of 41 variables that were used in the classification

Table 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 is results from the implementation of the decisions made in Section 3.3. Variables will often be referred to in the text only by number (for brevity); Table 5 can be used as a look up in these cases.

Table 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 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-dependent 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 resident in public sector rented accommodation
v17	Rent (Private): Percentage of households that are resident in 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 household spaces 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 and 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

3.5. 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 previously 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.

4. 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, 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 to big too 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 6. The columns in Table 6 represent as follows: 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. E and W Table refers to the name of the table in England and Wales. E and 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 and W Table but for Scotland and Northern Ireland respectively. Scot Ref and NI Ref represent the same as E and W Ref but for Scotland and Northern Ireland respectively. England and Wales, Scotland and Northern Ireland have to be done separately in this way, 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 6: Full variable definitions and sources (specified by key statistic table and column number)

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

4.1. 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 which is able to handle the number of data rows required for the whole UK to be held in one file.

A extraction program was written in FORTRAN 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 was 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 SPSS database.

4.2. 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 checks were carried out as follows. 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, some 492 (12*41) separate checks must be made to ensure that the data were entered correctly. As the data were extracted automatically by the extraction program 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 by hand 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 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 that 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 had 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 was also noted by the ONS team who were building the ward level classification, that there are some internal rounding processes which take place within SPSS that it is difficult to assess accurately. The difficulty being 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 7: Example of the data checking results across the UK

Variable Number	OA Code	OA Order Code	GOR / Country	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 the 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 8 shows an example of the results of this data checking.

Table 8: Example of the data checking results (for the North East GOR)

Variable	Observed	Expected	Difference	Difference, people /houses
V1	5.50	5.50	0.003	5
V2	12.92	12.92	-0.005	-15
V3	28.01	28.01	0.004	29
V4	24.54	24.54	0.003	19
V5	16.55	16.56	-0.013	-54
V6	1.21	1.21	0.000	0
V7	0.16	0.16	-0.003	0
V8	2.93	2.94	-0.014	-11
V9	2.93	2.93	-0.002	n/a
V10	10.93	10.93	-0.005	-10
V11	15.10	15.10	0.002	3
V12	15.64	15.64	-0.004	-6
V13	10.75	10.76	-0.009	-10
V14	16.87	16.87	0.002	3
V15	10.61	10.63	-0.023	-26
V16	27.65	27.64	0.015	44
V17	6.28	6.28	-0.002	-2
V18	32.10	32.10	-0.001	-3
V19	14.50	14.50	-0.002	-4
V20	13.92	13.92	0.001	2
V21	3.95	3.94	0.006	2
V22	5.19	5.19	0.003	n/a
V23	0.45	0.45	-0.003	n/a
V24	14.97	14.97	0.002	6
V25	23.90	23.89	0.005	22
V26	20.98	20.98	0.000	-1
V27	14.69	14.69	0.004	6
V28	7.68	7.68	-0.002	-2
V29	22.73	22.73	-0.003	-15
V30	11.00	11.00	0.002	6
V31	7.01	7.01	-0.005	-6
V32	4.53	4.53	0.004	3
V33	11.87	11.87	0.004	9
V34	6.58	6.58	-0.004	-5
V35	1.16	1.17	-0.013	-2
V36	7.87	7.88	-0.013	-11
V37	16.99	16.99	-0.001	-2
V38	5.10	5.10	-0.004	-2
V39	12.74	12.74	0.001	2
V40	3.04	3.04	-0.002	-1
V41	16.19	16.19	-0.001	-2

Table 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. The error across the three GORs was 500,000 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 9 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' websites for known errors and download and replace the relevant data when they becomes 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 9 million data points in the clustering database this would be difficult to achieve. However, the data checks do provide proof that it is highly 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. However, 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. CLUSTERING PROCESSES

This section gives a brief overview of some of the methods for standardisation and clustering that were considered, tested or used in this project. We outline of the differences between the methods and their good and bad points. The description and explanation is brief so, for a fuller account, you may wish to consult some of the texts referenced in this section.

5.1. Methods of 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 in the classification. 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 this case it 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 if these variables were clustered without being standardised it would add bias to the dataset.

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. It was felt that using the z-score standardisation 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 in the ward or local authority data. Z-scores do not set an absolute limit as to what the maximum value of each variable can reach and therefore, do not limit the effect of extreme values. This also means that different variables can have different maximum values. The process involved in calculating each type of standardisation is outlined in the following sub-sections.

5.1.1. Z-score standardisation

This is the most common form of standardisation. To create z-scores, firstly 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 in turn 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 $\overline{x}$ the average value of the variable across all *n* areas.

The Standard deviation is defined as:

$$\sigma_x = \frac{\sqrt{\sum_i (x_i - \overline{x})^2}}{n} \qquad (2)$$

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

$$Z_i = \frac{x_i - \overline{x}}{\sigma_x} \qquad (3)$$

Z-score standardisation works well when the data are normally distributed, but data may not always be normally distributed.

5.1.2. Range standardisation (0-1)

This method was implemented in the 1991 classification; 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:

$$Z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad (4)$$

where $x_{\max}$ is the maximum value of x and $x_{\min}$ the minimum value of x.

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.

5.1.3. Inter-decile range standardisation

This method is a slight variation of the range standardisation method that overcomes the problems associated with outliers.

This method is defined as:

$$\frac{x_i - x_{med}}{x_{90^{th}} - x_{10^{th}}} \qquad (5)$$

It 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, $x_{10^{th}}$

Skewed variables were given too much weight by the inter-decile range standardisation. This problem was resolved when the data were standardised using the range standardisation method. Therefore, the range standardisation method was used to standardise both the ward level data and the output area level data.

5.2. 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 and allocation-relocation algorithms, there are several different ways in which classifications can be constructed. A brief description of the procedures used is given here. The first method discussed is Ward's clustering method, which is a hierarchical clustering procedure. The second method reviewed is k-means clustering which is an iterative, non-hierarchical method. Further alternative methods are also briefly reviewed.

5.2.1. Hierarchical (Ward's method)

Developed by and named after Joe H. Ward of the Aerospace Medical Division, Lockland Air Force Base, it was first published in the Journal of the American Statistical Association in 1963. 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 must suitable number of clusters after the clustering process.

The process of hierarchical clustering is a agglomerative or bottom-up 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:
$$f = C_1, C_2, C_3, \ldots, C_{n-2}, C_{n-1}, C_n \qquad (6)$$
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.

Euclidean distance:
$$\text{distance}(x, y) = \{\sum_i (x_i - y_i)^2\}^{\frac{1}{2}} \qquad (7)$$

Squared Euclidean distance:
$$\text{distance}(x, y) = \sum_i (x_i - y_i)^2 \qquad (8)$$

5.2.2. Non-hierarchical (k-means)

The k-means algorithm is a simple non-parametric clustering method. 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 distinct the clusters are (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):

- 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.
- Step 1: Move each object from its own group to each other group and recompute the sums of squared deviations (the clustering criterion).
- Step 2: Choose the change which leads to the greatest improvement in the clustering criterion.
- Repeat steps 1 and 2 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:

$$E_c = \sum_{i=1}^{n_c} \sum_{j=1}^{m} (Z_{ij} - Z_{cj})^2 \qquad (9)$$

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

5.2.3. Alternative clustering methods

Although k-means and Ward's method were chosen for the methodology for the project (as described in section 6) many other clustering methods are available. Some have been used in the creation of previous classifications; for others there is no recorded evidence of their use for area classification. However, all the methods are valid for this form of analysis.

Openshaw (1994) describes how an artificial intelligence technique, know as a Self Organising Map developed by Kohonen (1984) was used to develop the GB profiles geodemographic system which clustered the Enumeration districts from the 1991 Census. This system is still available to use and a full description of how it was created can be found at:
http://www.geog.leeds.ac.uk/software/gbprofiles/.

Another useful technique is TwoStep clustering, available in the SPSS statistical package. The benefit of this method is that it has the ability to incorporate categorical data into the clustering process. Kaufman and Rousseeuw (2005) provide an excellent summary of various clustering methods, some of which are not to be found anywhere else. Some of the methods they use have freely available software that is pointed to in the book. The bible for cluster analysis is Everitt *et al.* (2001), which provides excellent descriptions of all the common forms of clustering.

Towards the start of the 1990s, something called 'Fuzzy Thinking' or 'Fuzzy Theory' was starting to develop within the sciences (Kosko 1994). The easiest way to think of Fuzziness in terms of classification is that everything is on a grey scale. In conventional clustering everything is black and white; something is either a member of a cluster or it is not. In fuzzy clustering everything is a member of every cluster but to a different extent.

It is known that classifications in their nature have points of uncertainty towards the outer reaches of each class. At the point furthest from the cluster centre the objects which have been clustered are often more similar objects on the edge of other clusters rather than the objects in the centre of the cluster to which they have been assigned. A fuzzy classification system uses this property of the classification process by classifying each point as having a proportional membership to several classes, as opposed to being strictly a member of one or other class (Voas and Williamson 2001). An excellent description of a method of fuzzy classification is given in Feng and Flowerdew (1998).

6. CONSTRUCTING A USEABLE METHODOLOGY

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 are two requirements which are not always compatible. All these issues need to be considered during the design and implementation of the clustering methodology.

6.1. The hierarchy to be created

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 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 solution is most suitable. 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 9.

Table 9: The number of clusters in the most widely used classification systems

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

Table 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 experience as to what he has found most useful could provide excellent guidance in this matter. His’s views are quoted below.

*“**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.”* (Callingham 2003) emphasis added.

*“**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.”* (Callingham 2003) emphasis added.

***“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.

6.2. The original methodology

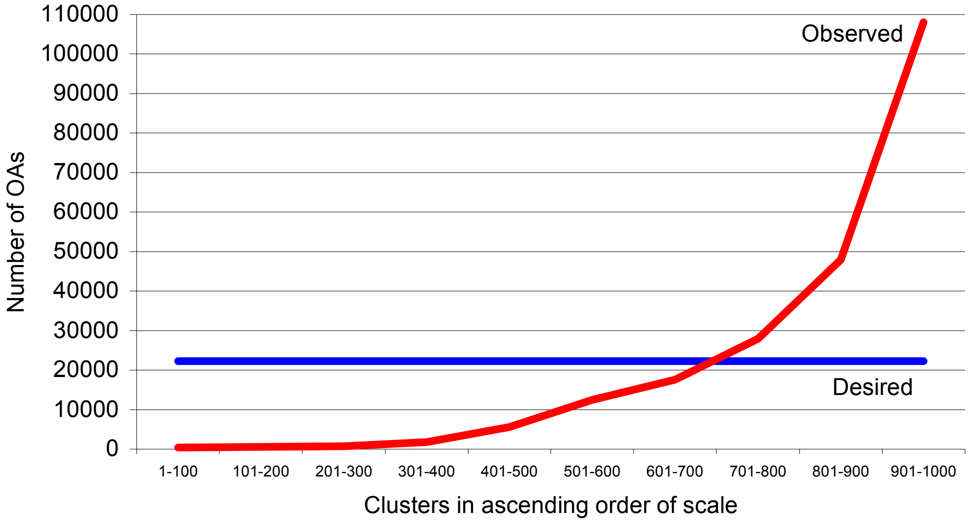
The objective is to create a three tier hierarchy to complement that created by the ONS for the ward and local authority level classifications. It was therefore planned that Ward's hierarchical clustering algorithm would be used to create the hierarchy within the classification. However, Ward's algorithm can only run on relatively small datasets of approximately 1000 or fewer data points, 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 1000. This was necessary as hierarchical clustering procedures cannot handle very large datasets. 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 1000 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 both clustering algorithms working together. The first and biggest problem is created when 1000 clusters are created using the k-means algorithm, the problem being that the within the data there are areas that have unusually extreme values these outliers get clustered into groups of small or single membership. Figure 2 shows how this affects the size of membership of the clusters, the clusters have been split into deciles in ascending order (1-100 representing the 100 clusters with the smallest membership and 901-1000 representing the 100 clusters with the largest membership). The red line on the graph represents the distribution if all clusters were the same size (223 members). The blue line 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 2: The distribution of observed and desired cluster sizes (1000 clusters using k-means)



The problem is then compounded when Ward's algorithm is run on the k-means centres. Table 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. How has this severely skewed distribution of membership come about? It becomes a little clearer by looking at the original 1000 k-means clusters from which the smaller number of clusters is formed. Of the original 1000 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 1000 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 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 10.

Table 10: Number of OAs in each cluster based on the original methodology

	1	2	3	4	5	6	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%

Several different methods of data transformation were 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 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 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 11: Number of OAs in each cluster based on the original methodology (ranks)

	1	2	3	4	5	6	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 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 12, the log transformation does not reduce the difference in size between the clusters as much as converting the data into ranks.

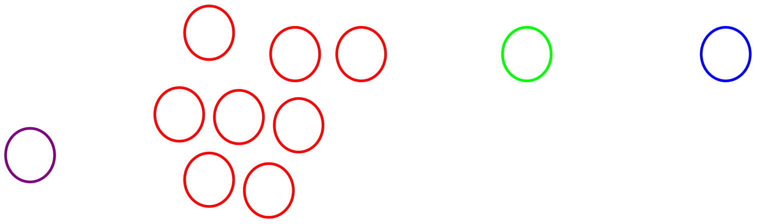
Table 12: Number of OAs in each cluster based on the original methodology (logs)

	1	2	3	4	5	6	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, even in a very large data set. An OA which 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 which are further apart. Figure 3 shows how this can happen.

Figure 3: The intricacies of Ward's Hierarchical Clustering Procedure



The red dots in Figure 3 are clearly a cluster so they are grouped together in the first six 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 seventh and eighth 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 is 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 using the 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.

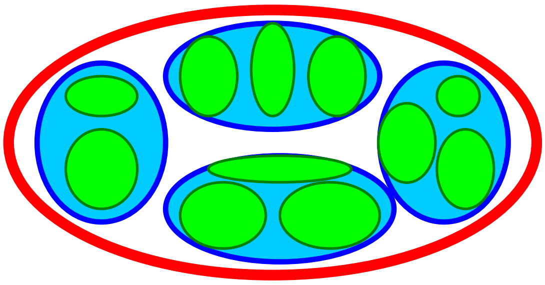
These experiments with the OA database have shown that when a hierarchical clustering procedure is used, it will inherently produce clusters of uneven size. There are therefore serious doubts about the reliability and quality of result. The use of this methodology was therefore rejected on the basis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

6.3. The final methodology: creating a hierarchical system using the k-means algorithm

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 4.

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 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 4). The second level of the hierarchy is then 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 4).

Figure 4: The creation of a hierarchical system using the k-means algorithm

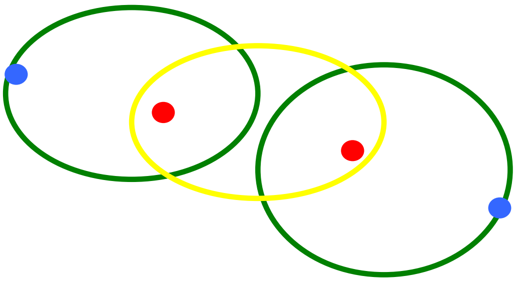


The second way in which this could be done is a bottom up approach and works the opposite way round. The lowest level of the classification is created first (as represented by the green areas in Figure 4); about 50 clusters are generated using the k-means algorithm. The centres of the 50 clusters produced are then re-clustered to produce the middle level of the hierarchy (as represented by the blue areas in Figure 4). Then in turn the same would be done on these to create the highest level (as represented by the red area in Figure 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. 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: 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.

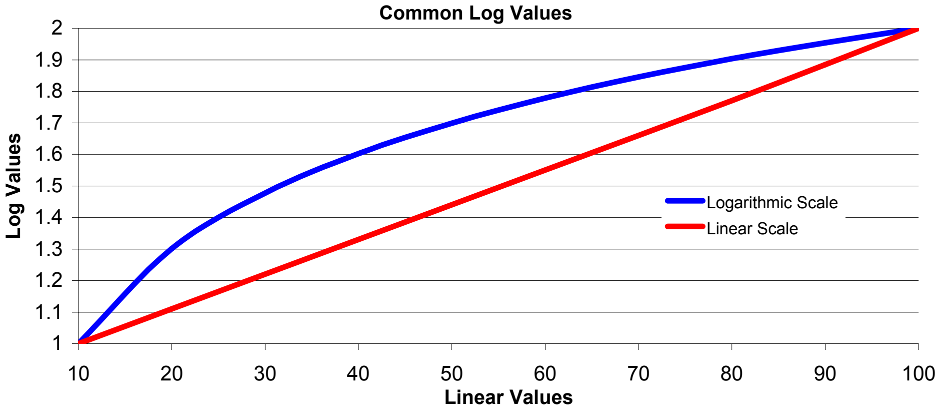
6.3.1. Elucidating the logarithmic transformation

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 6.

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 6.

Figure 6: the effect of logarithmic transformation on a dataset



Before the data were converted to a log scale, all the vales 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.

$$\text{Log}(X+1) = \text{new value to be range standardised} \qquad (10)$$

You may ask whether a log transformation is necessary considering that the variables will all be ranged standardised. 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 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 which in turn suggests 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. Transforming of the data onto a logarithmic scale is one way of reducing the likelihood of this.

Table 13: The effect of logging data on the distribution of the data

Variable	Difference of mean value from 0.5 after standardisation		Difference
	Not Logged	Logged	
S1	0.31	0.03	0.28
S2	0.27	0.12	0.15
S3	0.16	0.25	-0.09
S4	0.12	0.26	-0.15
S5	0.33	0.09	0.24
S6	0.47	0.36	0.11
S7	0.48	0.40	0.08
S8	0.42	0.12	0.30
S9	0.50	0.14	0.36
S10	0.32	0.08	0.24
S11	0.34	0.07	0.27
S12	0.35	0.05	0.30
S13	0.36	0.01	0.35
S14	0.24	0.17	0.08
S15	0.29	0.08	0.22
S16	0.29	0.03	0.27
S17	0.42	0.13	0.29
S18	0.25	0.05	0.20
S19	0.27	0.01	0.26
S20	0.28	0.05	0.24
S21	0.42	0.12	0.29
S22	0.09	0.09	0.00
S23	0.28	0.22	0.06
S24	0.28	0.12	0.16
S25	0.18	0.21	-0.04
S26	0.22	0.18	0.04
S27	0.34	0.04	0.30
S28	0.41	0.05	0.36
S29	0.29	0.24	0.05
S30	0.36	0.04	0.32
S31	0.43	0.10	0.33
S32	0.41	0.14	0.27
S33	0.21	0.17	0.04
S34	0.32	0.02	0.30
S35	0.47	0.36	0.11
S36	0.43	0.07	0.36
S37	0.35	0.07	0.28
S38	0.45	0.15	0.30
S39	0.39	0.03	0.36
S40	0.43	0.17	0.27
S41	0.33	0.11	0.23
Mean	0.33	0.13	0.20

7. THE CREATION OF THE CLASSIFICATION

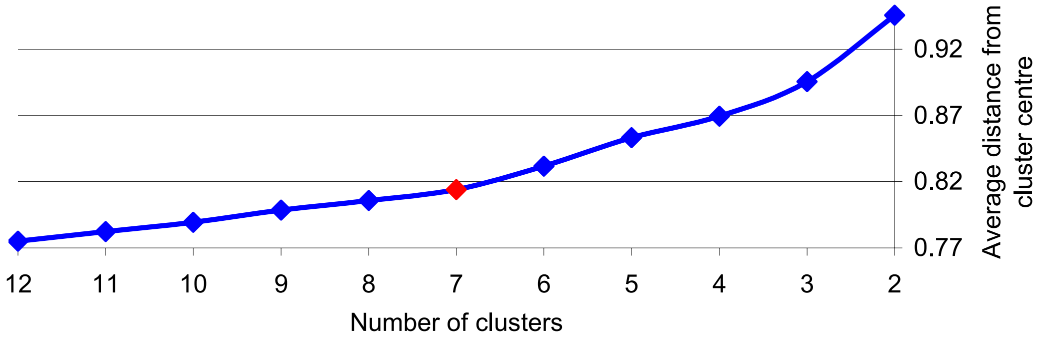
This section describes the implementation of the final methodology as described in section 6.3. The descriptions, the cluster size choices that were made and outline the reasons behind the decisions. The decisions were made based upon a plethora of information that can be outputted from the clustering 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 one 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 at that level.

It had been suggested that the most useful number of clusters In the first level would be around 6, taking this as a starting point clusters from 2 - 12 were examined to see how the average within cluster distance from centre changed. Figure 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 5 which leaves a choice between 4, 6, 7 and 8.

Figure 7: Average distance from cluster centre for different values of k, using k-means clustering



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. Very small clusters are of little use and may represent nothing more than outliers within the original dataset rather than a small set of unusual areas.

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 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 14: Calculation of which solution is most homogenous in terms of cluster membership size

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

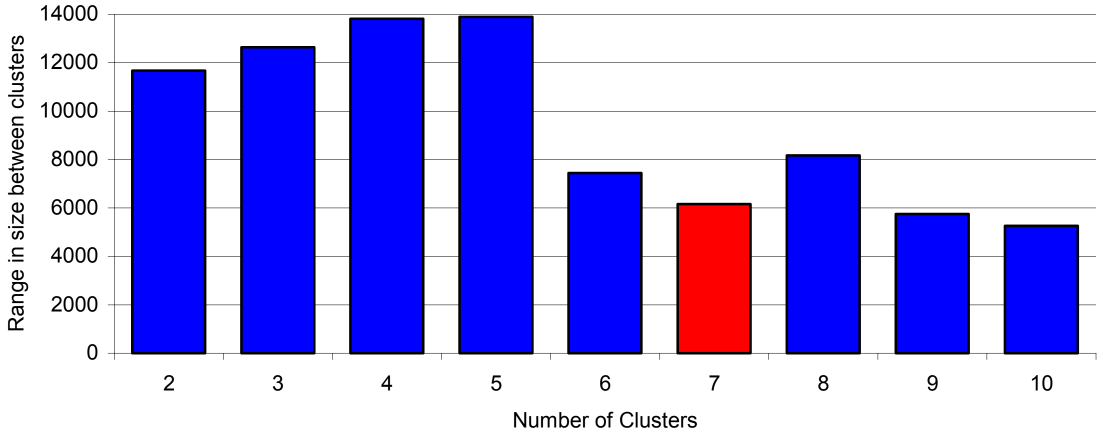
Table 14 shows how the method works on a small data set, but what results were produced using this method on the possible solutions for the OA classification? Figure 8 shows the average distance from the mean cluster membership for solutions of cluster numbers 2 to 10. The best solution based on this solution 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 there non-formation would have prevented the splitting of a cluster into a lower level.

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 7 but poorly in Figure 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 7 and well in Figure 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 7, but in Figure 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.

Figure 8: The range in the size of clusters as the number of clusters increases



Once the first level of the classification (to be know 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 2, 3 or 4 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 16 (see later) shows the structure of the classification, indicating into how many groups each cluster was split.

Table 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 being halved, falling from 62,671 using the most compact solution from the original methodology, to 30,613 with the use of the new methodology.

Table 15: Number of OAs in each cluster based on the final methodology

	1	2	3	4	5	6	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%

8. NAMING AND DESCRIBING THE CLUSTERS

One of the world's most underrated art forms must be the naming and labelling of the clusters of a geodemographic (area) classifications. The process can be long and drawn out, 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. 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 wrong 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 a 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. Names can also be easily pilloried on 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. 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 to 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; 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. The distinction between the two in terms of their approach to naming clusters is clear. The Mosaic profiles (Experían http://www.experianbs.com/Content.asp?ArticleID=566) are designed to be creative, provocative (and are perhaps a little inaccurate). The Cameo (EuroDirect http://www.eurodirect.co.uk/) 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.

8.1. Cluster names

Before discussing in detail the names adopted for the clusters, it is useful to review the naming process. The first author made proposals which were commented upon by the second and third authors. These names were presented to the ONS Steering Group, led first by John Charlton and later by Simon Compton and Andrew Botterill. The Steering Group made many suggestions, based on the experiences of naming the ward level clusters, and the names were revised. The names were then agreed upon and approved by the authors and the members of the ONS Steering Group. The names then went forward to the Director of National Statistics, Len Cook, for final review as the OA classification was to become a "National Statistic". He took the decision that naming of the clusters at any level was inappropriate because of the danger that the residents in any cluster would feel affronted by the name. Labelling of areas might have adverse effects on those who lived there. So the National Statistics version of the OA classification uses simply the number-letter-number identification system.

However, the authors took a different view. We consider that the British population is intelligent enough to know that the cluster names are approximate and average labels and that users of the classification will feel more comfortable with using a set of names rather than codes. An agreement was reached that ONS would publish the codes for output areas and refer users to this online publication for a set of names, as set out in Table 16.

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.

With this classification to be used by the ONS as the official national classification of output areas the names needed to follow two general principles: they must not offend residents and they must not contradict other 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 comment and suggestion on names was received from people who took part in a consultation exercise about the classification, 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 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 16.

Table 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: Aspiring Households (1)
		6b2: Aspiring Households (2)
		6b3: Aspiring Households (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)

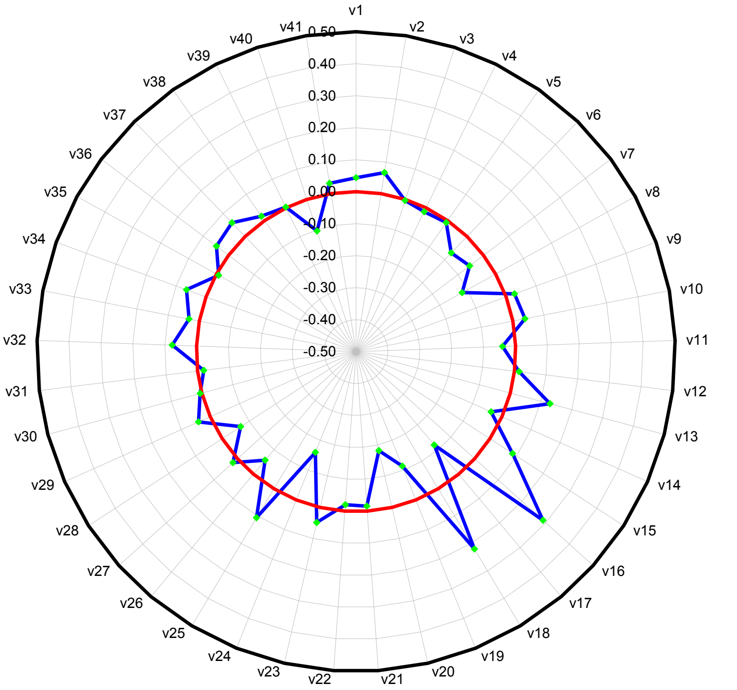
8.2. Cluster profiles

The idea behind cluster profiles is to create a short description, using text and visuals, which expands on the cluster names but, only takes a few seconds to read but which significantly expands the user's, 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. Groups which 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 them. 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. We avoid specific place names, however, because these have resulted in geographical mislabelling in past classifications.

Cluster profiles are given for each of the seven super-groups in Figures 9-15 (other levels 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 red ring at 0, the value of each variable for that super-group can then be seen by the amount that the blue line (showing the difference from the mean for each value) is above or below the red one. The variable codes (v1 - v41) relate back to Table 5. A list of the most distinctive variables for each group is also given in the box beside the radial plot.

Figure 9: Cluster profile for Cluster 1 Blue Collar Communities



Distinctive Variables

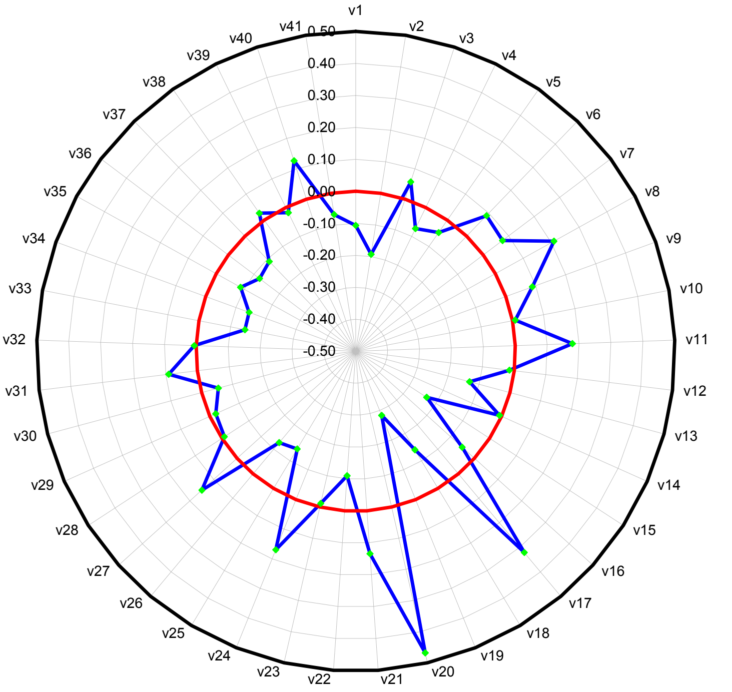
High:

- Age 5-14
- Lone Parent Households
- Households with non-dependent children
- Terraced Housing
- Routine/Semi-routine employment
- Mining/Quarrying/Construction Employment
- Manufacturing Employment
- Retail Trade Employment

Low:

- Indian, Pakistani and Bangladeshi
- Black
- Born Outside the UK
- Rent (Private)
- Flats
- HE Qualifications
- Financial Intermediation employment

Figure 10: Cluster profile for Cluster 2 City Living



Distinctive Variables

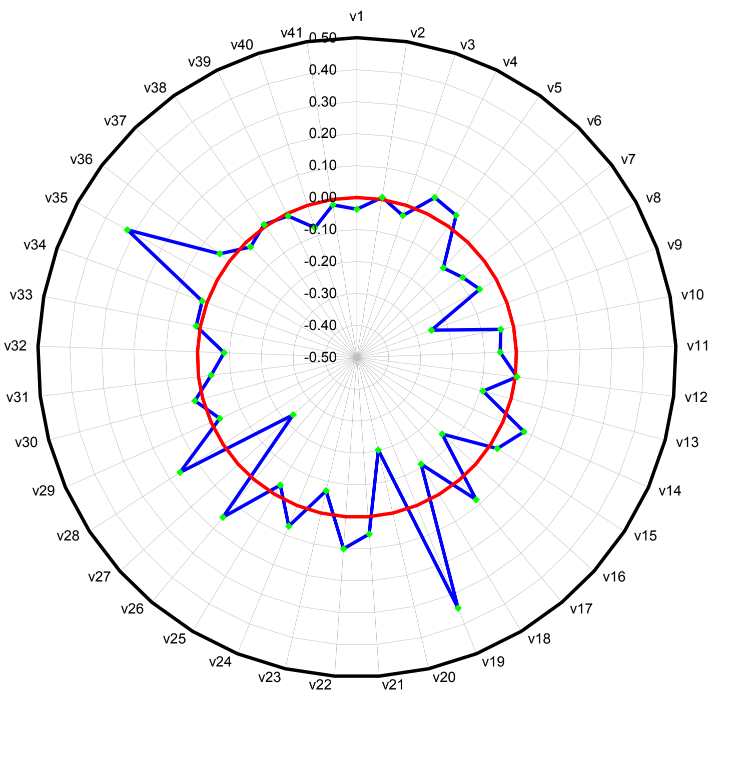
High:

- Age 25-44
- Born Outside UK
- Population Density
- Single Person household
- Rent (Private)
- Flats
- No Central Heating
- HE Qualification
- Students
- Financial Intermediation Employment

Low:

- Ages 0-4, 5-14, 25-44 and 65+
- Single Parent Household
- Households with non-dependant children
- Room per Household
- Provide unpaid Care
- Economically inactive Looking after Family
- General Employment

Figure 11: Cluster profile for Cluster 3 Countryside



Distinctive Variables

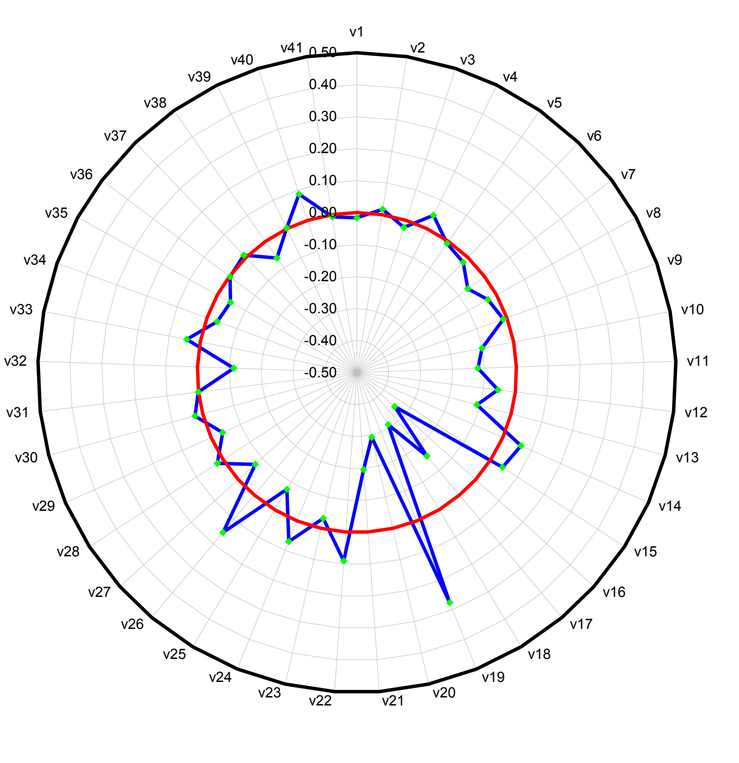
High:

- Ages 45-64 and 65+
- Detached Housing
- Rooms per Household
- 2+ Car Households
- Work from Home
- Provide Unpaid Care
- Agricultural Employment

Low:

- Indian, Pakistani and Bangladeshi
- Black
- Population Density
- Single Person Household
- Flats
- People per Room
- Public Transport to Work
- Unemployment

Figure 12: Cluster profile for Cluster 4 Prospering Suburbs



Distinctive Variables

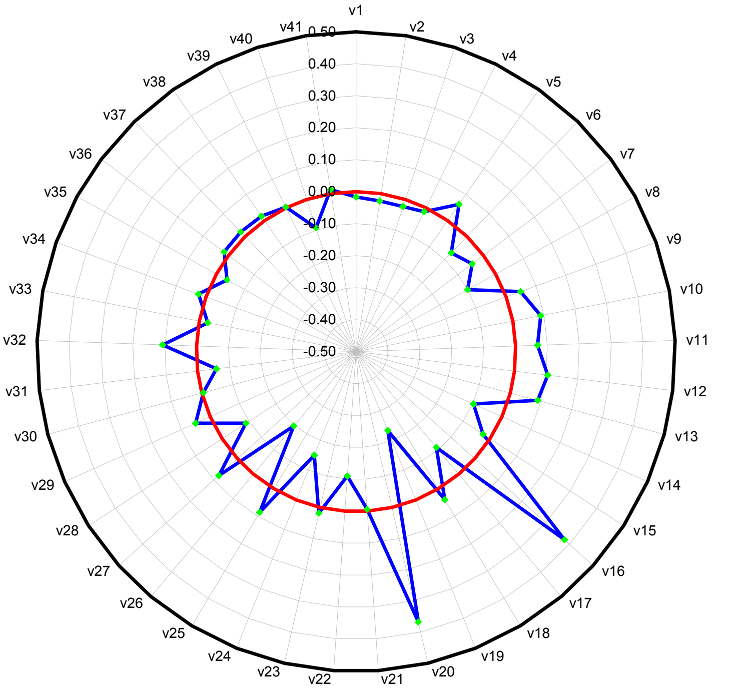
High:

- Age 45-64
- Two adults no children
- Households with non-dependant children
- Detached housing
- Rooms per Household
- 2+ Car households
- Provide unpaid care

Low:

- Indian, Pakistani and Bangladeshi
- Black
- Divorced/Separated
- Single Person Household
- Single Pensioner Households
- Renting Public and Private
- Terraced housing
- Flats
- No Central Heating
- LLTI
- Unemployment

Figure 13: Cluster profile for Cluster 5 Constrained by Circumstances



Distinctive Variables

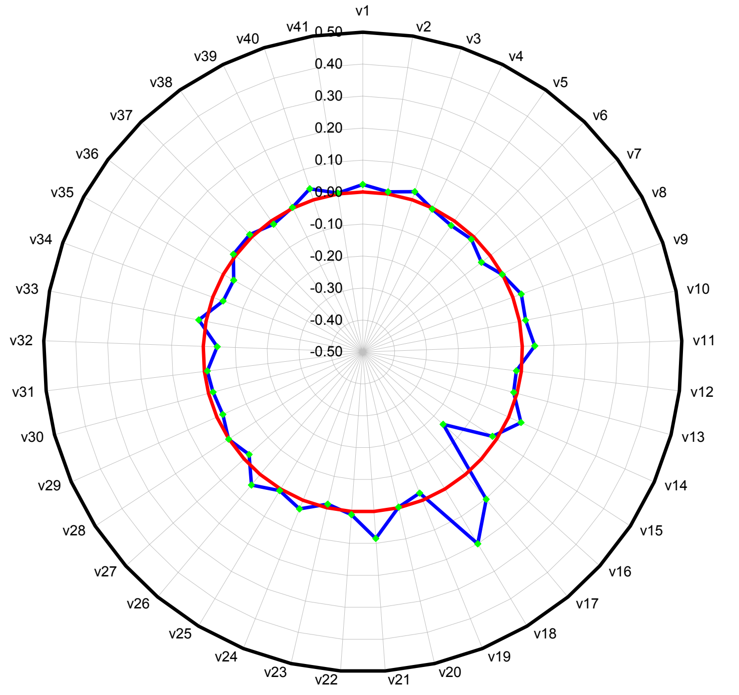
High:

- Age 65+
- Divorced/Separated
- Single Pensioner households
- Lone Parent Households
- Rent (Public)
- Flats
- People per Room
- Routine/Semi-Routine employment
- LLTI
- Unemployment

Low:

- Two Adults no Children
- Rent (Private)
- Detached Housing
- Rooms per Household
- HE Qualifications
- 2+ Car Households
- Work from home

Figure 14: Cluster profile for Cluster 6 Typical Traits



Distinctive Variables

High:

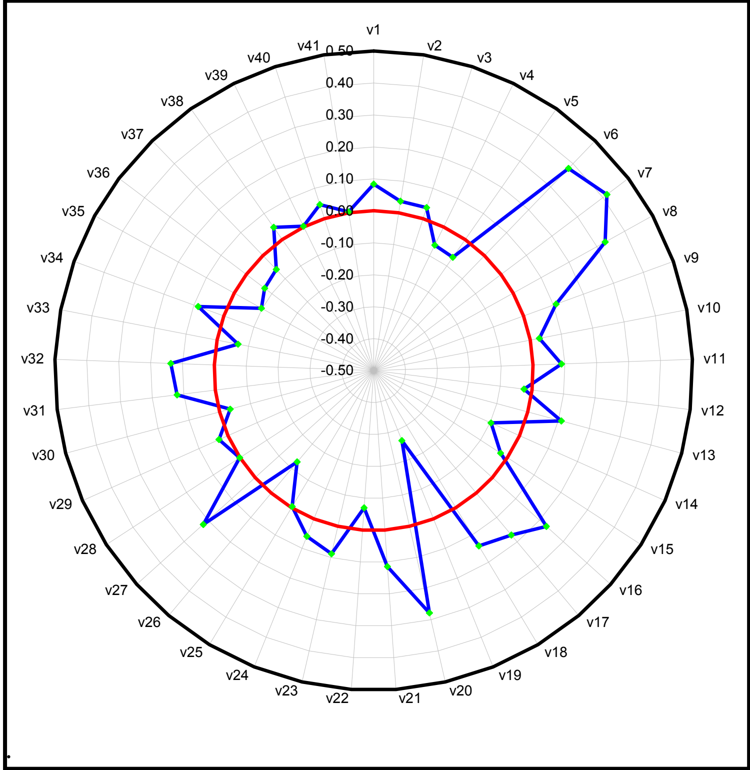
- Work Part Time
- Terraced Housing

Low:

- Age 65+
- Rent (Public)

Characterised by its averageness, this group has few values which are high or low in comparison to the other groups.

Figure 15: Cluster profile for Cluster 7 Multicultural



Distinctive Variables

High:

- Ages 0-4 and 5-15
- Indian, Pakistani and Bangladeshi
- Black
- Born Outside UK
- Population Density
- No Central Heating
- People per Room
- Public Transport to Work
- Students
- Unemployment

Low:

- Ages 45-64 and 65+
- Single Pensioner Households
- Two Adults No Children
- Economically Inactive/ Looking after Family or Home

8.3. Additional outputs

As well as the traditional pen portraits shown in section 8.2, 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 enables the data to be looked at in the opposite way to the cluster portraits for which the most significant variables within a given cluster can be seen. 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. Figures 16-18, show variable 20 (percentage of households which are flats) for all three tiers of the hierarchy. 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 of the data range.

Figure 16 shows that 'City Living' is the place to be if you are looking for flats, whereas there are not particularly rich pickings in 'Prospering Suburbs'. Figures 17 and 18 show how the hierarchy affects the values. The 'City Living' super-group splits into 2a 'Transient Communities' and 2b 'Settled in the City'. 'Transient Communities' shows an increase on the value of 'City Living' whereas 'Settled in the City' has a rate which is not as high as that of its super-group. As the super-groups split into groups the effect of the hierarchy can be seen on the values. For example 4d 'Thriving Suburbs' shows a value which indicates the presence of significantly more flats than the other groups within its super-group.

Figure 16: Variable by super-group graph using the original data for variable 20 (All Flats)

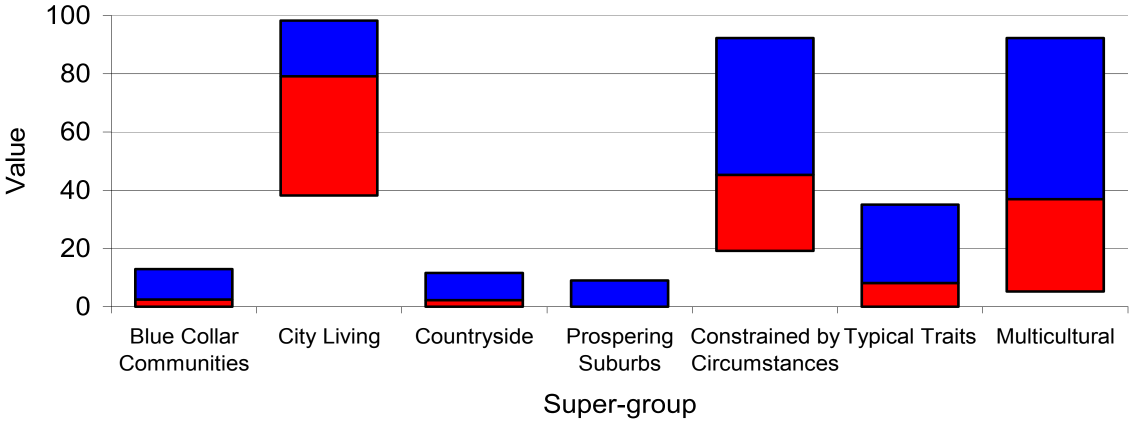


Figure 17: Variable by group graph using the original data for variable 20 (All Flats)

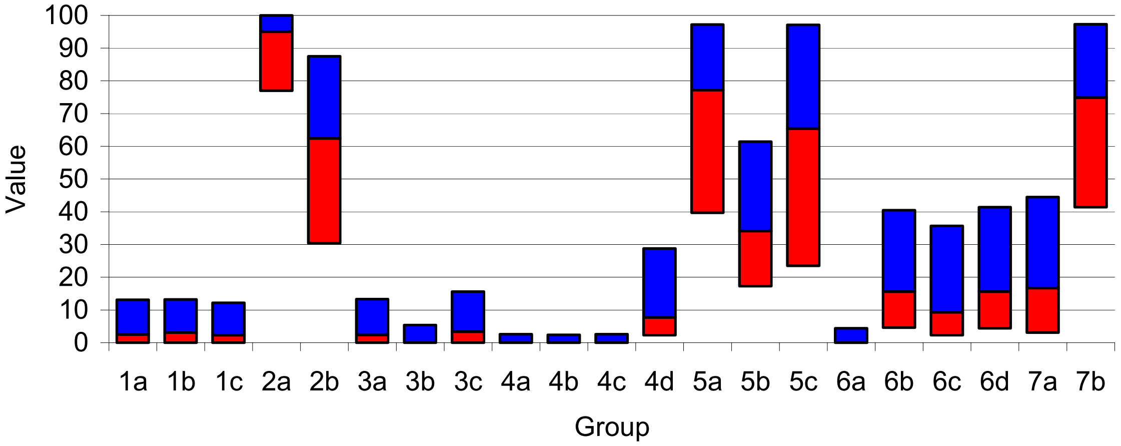
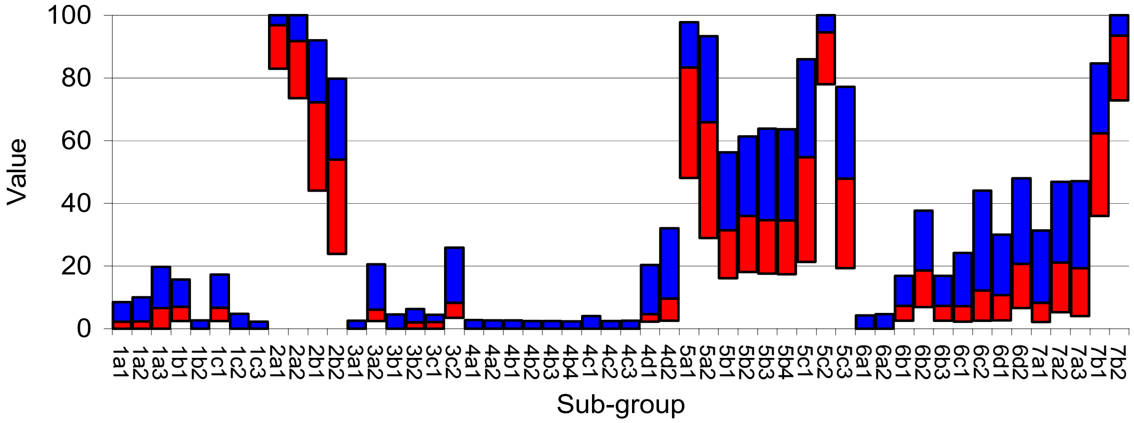


Figure 18 shows some extreme values at both ends of the scale from 2a1, 5c1 and 7b2 which are dominated by people living in flats to, 1b2, 1c3, 3a1, 4a1-4b4, 4c2 and 4c3 where flats are somewhat of a novelty. The indication of the range given by the length of the bars also gives much information about each cluster. For example, compare 5c1 and 5c2. 5c2 is much 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 18: Variable by sub-group graph using the original data for variable 20 (All Flats)



9. 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. Without mapping the classification the most important part of the classification will be lost. 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 (as discussed in section 6.1). 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.

9.1. Visualising the classification in alternative ways

There are problems with the mapping of output areas (discussed in section 2.1). Mapping at such a small scale has inherent scaling problems, problems wrapped up in the design of the OAs (see Figure 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 find 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 19 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 currently 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 unsurprisingly located outside the large urban centres. Some variation can be seen within urban centres at this scale for example Multicultural 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 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. At this scale the very basics can be picked out, urban areas clearly contrast to areas which are more sparsely populated some detail can be seen within larger cities. A contrast can be made between the more cosmopolitan larger cities and smaller urban areas which show less evidence of the ‘City Living’ and ‘Multicultural’ super-groups.

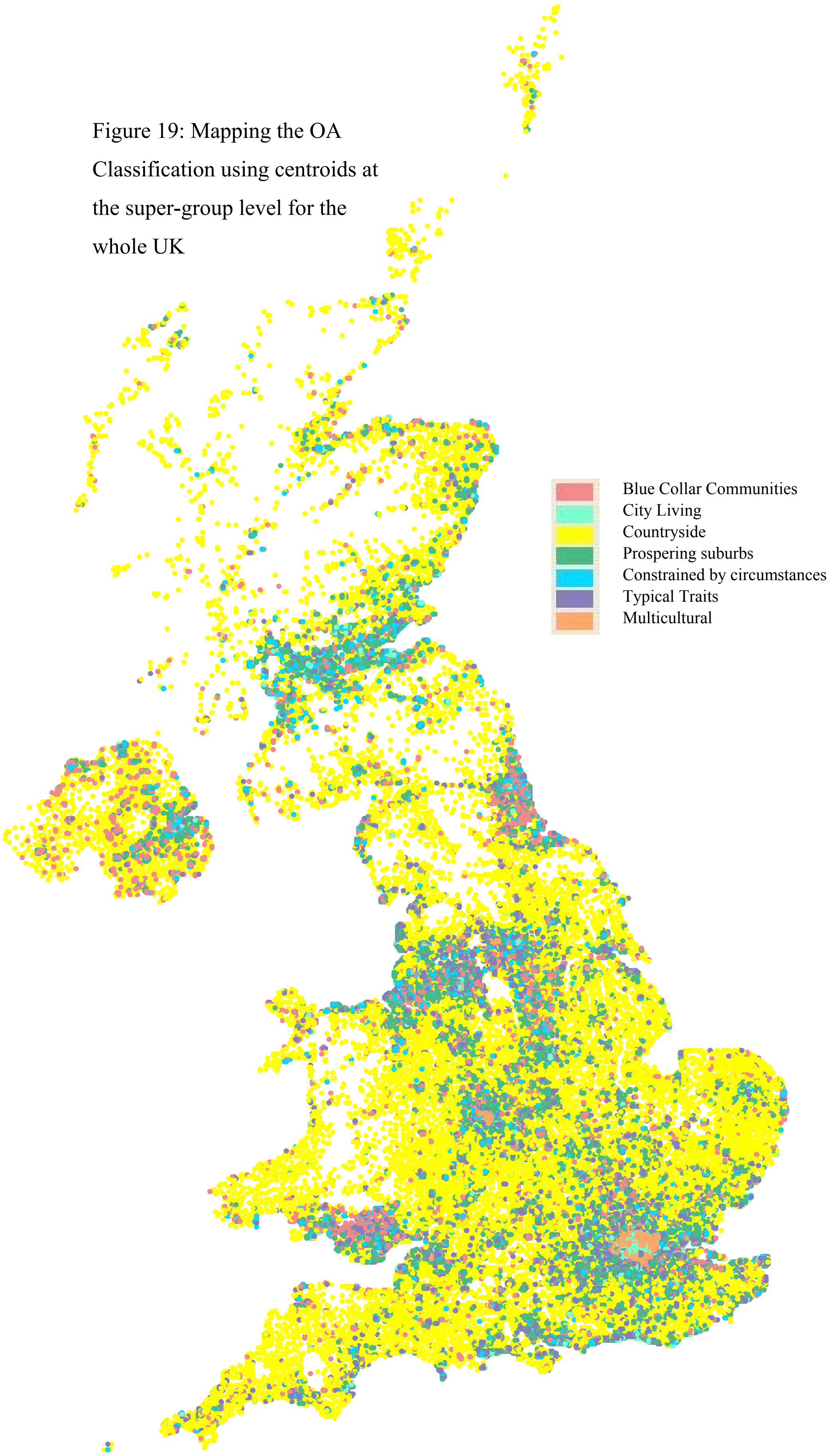
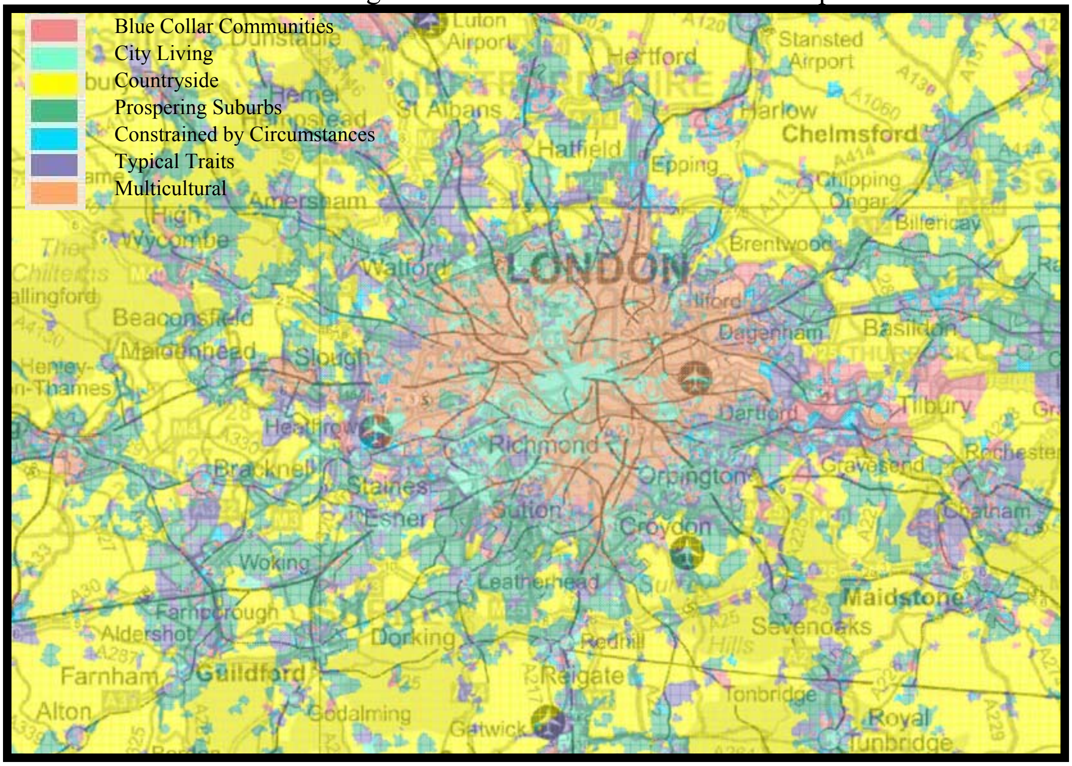


Figure 19: Mapping the OA Classification using centroids at the super-group level for the whole UK

By looking at the classification for much smaller areas settlement patterns become more apparent. Figures 20 and 21 show the classification for London and its surrounding area. Figure 20 shows a map using the full boundaries of the OAs whereas Figure 21 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 Multi-Cultural Blend 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 colours to be viewed more easily. Both these figures are overlain on maps showing the urban centres and transport networks of the UK, which adds information when visualising the classification.

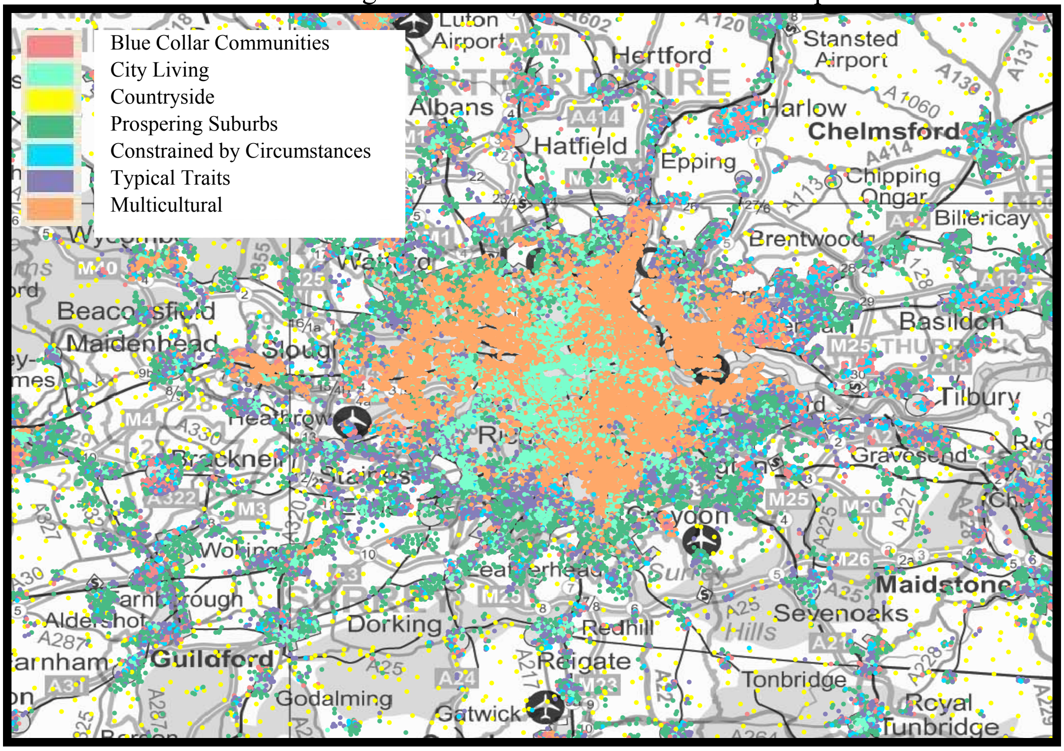
Figure 20: Mapping the OA classification at super-group level using boundaries for London and surrounding area overlaid on a UK settlement map



© Collins Bartholomew

Figure 20 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 21 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 20 and this allows other information to be drawn out.

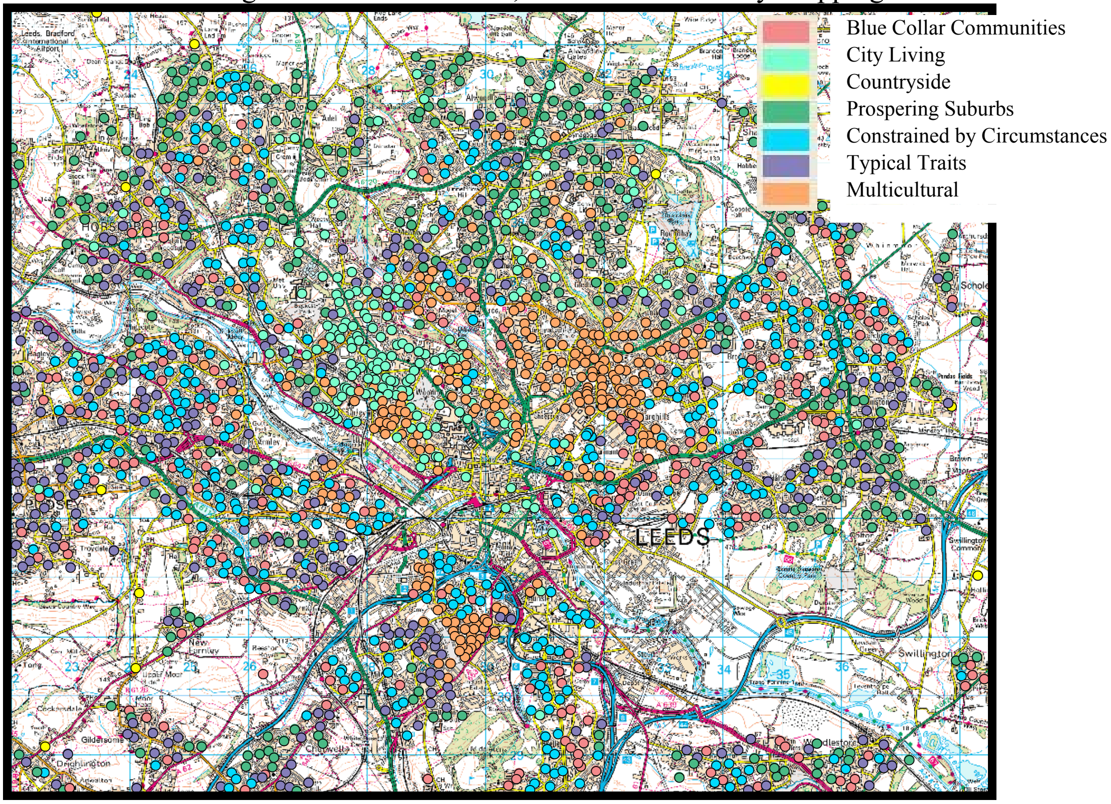
Figure 21: Mapping the OA classification at super-group level using centroids for London and surrounding area overlaid on a UK settlement map



© Collins Bartholomew

Figure 22 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 much 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 where there are many commercial properties. The 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.

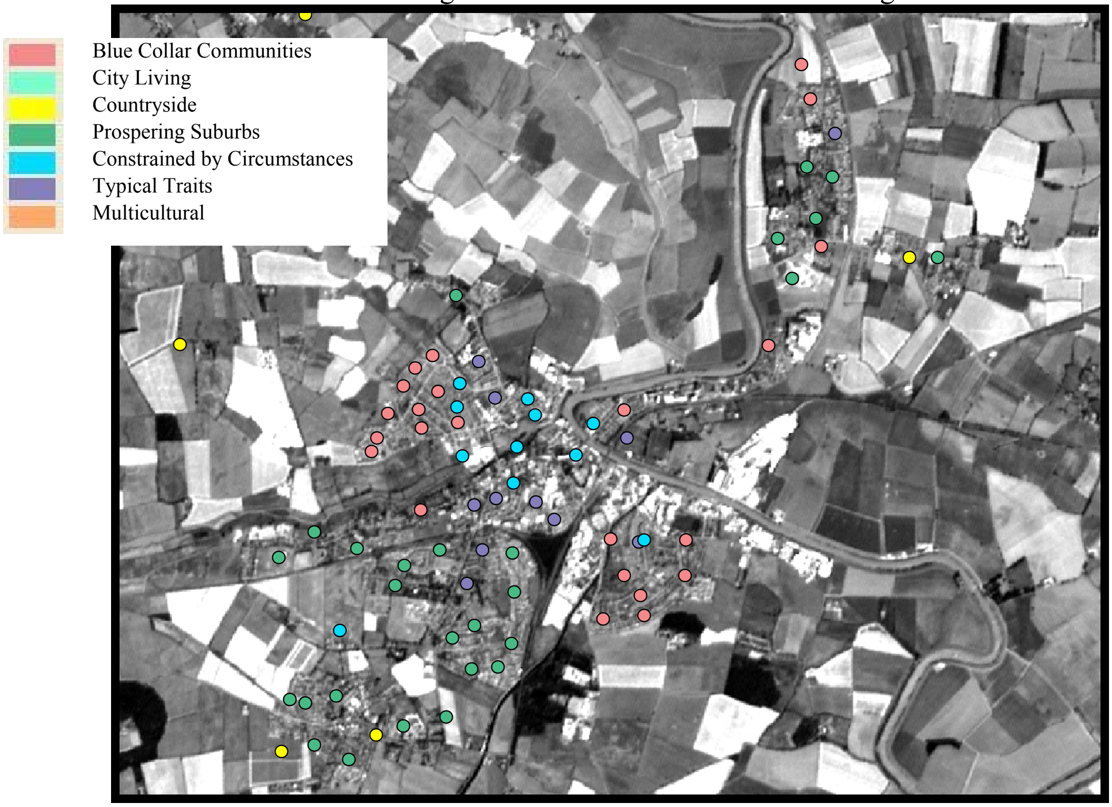
Figure 22: Mapping the OA classification at super-group level using centroids for Leeds and surrounding area overlaid on 1:50,000 Ordnance survey Mapping



©Ordnance Survey Crown Copyright

Figure 23 shows a SPOT satellite image (resolution 5-20m) of the town of Selby in north Yorkshire. Clear physical and synthetic 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 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 23: 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

10. CONCLUSIONS AND FURTHER WORK

What can be concluded from the creation of the classification? Has what was set out to be created been achieved? What has the report told us?

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 has been produced to go with the classification to aid understanding and assist in the using of the classification.

This report has discussed of all the decisions that were made during the creation of the classification and the reasons behind them. The report discuses 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 report 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 finally brought to life by adding the reality back into the classification with the use of a variety of mapping and visualisation techniques.

This report outlines and explains the creation of the classification. However there is a large amount of information about the classification that could not be covered in this document. This included; the consultation exercise that was carried out to gauge the opinion of the wider community about the classification, many of the tests and validation procedures that were carried out on the classification and many other pieces of data related to the classification. Further information associated with the classification includes:

- *Cluster profiles* for all clusters at all levels.
- *Additional outputs* for all clusters at all levels.
- A *fuzzy version* of the super-group level: This gives the distance of each OA to each cluster centre rather than just its own.
- A set of *photographs* taken across the country depicting areas representative of each cluster.
- A *Multi-Scale Classification Database* linking the OA classification to the Ward and Local Authority District classifications.
- Many more maps and visualisations.

Much of these data will be published over the next year on the completion of the project. If you are interested in using the additional information please feel free to contact the authors.

References

Aldenderfer, M. S. and Blashfield, R. K. (1984), *Cluster Analysis*, London, Sage.

Bailey, S., Charlton, J., Dollamore, G. and Fitzpatrick, J. (1999), *The ONS Classification of Local and Health Authorities of Great Britain: Revised for Authorities in 1999*, London, Office for 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.

Booth, C. (1969), *Charles Booth's London*, London, Hutchinson.

Brown, M. (2005), *Second homes in Cornwall*, [Personal Correspondence by e-mail] 8th March 2005.

Callingham (2003) *Current commercial sector use of geodemographics and the implications for the ONS area classification systems*, [Personal Correspondence by e-mail] 14th October 2003.

Everitt, B. S. Landau, S. and Leese, M. (2001), *Cluster Analysis 4th Ed.*, London, Edward Arnold.

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.

Garland, K. (1994), *Mr Beck's Underground Map*, London, Capital Transport Publishing.

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

Kaufman, L. and Rousseeuw, P. J. (2005), *Finding groups in data*, Chichester, Wiley.

Kohonen, T. (1998) The self-organising map, *Neurocomputing*. 21,1-6.

Kosko, B. (1994), *Fuzzy Thinking: The New Science of Fuzzy Logic*, London, Flamingo.

LSE (2005), *Charles Booth Online Archive*, [online] http://booth.lse.ac.uk/ accessed 14/7/05.

Martin, D. (1997), *From Enumeration Districts to Output Areas: experiments in the automated creation of census output geography*, Statistical Commission and Economic Commission for Europe, Conference on European Statistics, Brighton 22-25 September 1997.

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, in *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 conference on 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, Chapter 3, pp. 37-46 in Rees, P., Martin, D. and Williamson, P. (Eds.), *The Census Data System*, Chichester, Wiley.

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.

ONS (2004), *National Statistics 2001 Area Classification,* [online] http://www.statistics.gov.uk/geography/census_geog.asp accessed 14/12/2004

ONS (2005a), *Beginners' guide to UK geography: Census Geography* [online] http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/ accessed 19/07/05.

ONS (2005b), *Disclosure Protection Measures* [online] http://www.statistics.gov.uk/census2001/discloseprotect.asp accessed 19/07/05.

ONS (2005c), *Rural and urban classification 2004,* [online] http://www.statistics.gov.uk/geography/nrudp.asp accessed 27/7/2005

Openshaw, S., (1994), Developing smart and intelligent target marketing systems: part I, *Journal of Targeting, Measurement and Analysis for Marketing*, 2, 289-301.

Openshaw, S. and Gillard, A. A. (1978), On the stability of a spatial classification of census enumeration district data, in P. W. J. Batey (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 Wymer, C. (1995), Classifying and regionalizing census data, Chapter 8, pp. 239-270 in Openshaw, S. ed. *Census Users' Handbook*, Cambridge, GeoInformation International.

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.

Peach, C. (1996), Does Britain have ghettoes? *Transactions of the Institute of British Geographers*, 21, 216-235.

Rees, P., Denham, C., Charlton, J., Openshaw, S., Blake, M. and See, L. (2002), ONS classifications and GB Profiles: Census typologies for researchers, Chapter 11, pp. 119-170 in Rees, P., Martin, D. and Williamson, P. (Eds.), *The Census Data System*, Wiley, Chichester.

Simey, T. S. and Simey M. B. (1960), *Charles Booth: Social Scientist*, Oxford, Oxford University Press.

SPSS Inc. (1999) K-means cluster analysis, Chapter 29, pp. 333-339 in SPSS Inc., *SPSS Base 9.0, User's Guide.* Chicago, SPSS Inc.

Stillwell, J. (2005), *Ethnic Segregation Index*, Personal Correspondence, e-mail 7/03/2005.

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.

Voas, D. and Williamson, P. (2001), The diversity of diversity: a critique of geodemographic classification, *Area,* 33 (1), 63-76.

Wallace, M., Charlton, J. and Denham, C. (1995), The new OPCS area classifications, *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.