

Creating the UK National Statistics 2001 output area classification

Dan Vickers

University of Sheffield, UK

and Phil Rees

University of Leeds, UK

[Received April 2006. Final revision November 2006]

Summary. The paper describes the creation of the Office for National Statistics 2001 output area classification, which was created in collaboration with the authors. The classification places each 2001 census output area into one of seven clusters based on the socio-economic attributes of the residents of each area. The classification uses cluster analysis to reduce 41 census variables to a single socio-economic indicator. The classification was made available with a host of supporting and descriptive information as a National Statistic via National Statistics on line. The classification forms part of a suite of area classifications that were produced by the Office for National Statistics from 2001 census data. Classifications of local authorities, statistical wards and health areas are also available.

Keywords: Area classification; Cluster analysis; Geodemographics; Output areas

1. Introduction

The Office for National Statistics (ONS) 2001 output area (OA) classification was published on July 29th, 2005, via the ONS Web site. The classification groups geographic areas according to key characteristics that are common to the population in that grouping. These groupings, or clusters, are generated from 2001 census data (Office for National Statistics, 2005a). The classification is the result of a partnership between the ONS and the authors and is published by the ONS as a ‘National Statistic’ (Vickers and Rees, 2006). The classification forms part of a suite of area classifications that were produced by the ONS from the 2001 census. Classifications of local authorities, statistical wards and health areas are also available (Office for National Statistics, 2005a).

The aim of this paper is to provide a transparent account of the procedures that are used in creating the OA classification. We begin by discussing the nature of area classification (Section 2) and cluster analysis (Section 3). Then we review the principles underpinning variable selection, transformation (if needed) and standardization in Section 4, whereas in Section 5 we describe how the final choice of variables was made. Section 6 gives an account of data extraction and assembly and the vital process of verification. Section 7 reviews clustering methods and selects one to use. How cluster levels and numbers were chosen is described in Section 8. The penultimate Section 9 presents the diagnostic cluster profiles and justifies the cluster names that were chosen. Section 10 then concludes.

Address for correspondence: Dan Vickers, Department of Geography, University of Sheffield, Sheffield, S10 2TN, UK.
E-mail: d.vickers@sheffield.ac.uk

2. What is area classification?

Area classification is the classifying of areas into groups on the basis of the similarity of characteristics of selected features within them. One of the most commonly used area classifications is geodemographic classifications. Geodemographics is ‘the analysis of people by where they live’ (Sleight (2004), page 18). Geodemographics works on the principle that the place and population are inextricably linked. Knowing where somebody lives can reveal information about that person. Geodemographics can be said to work because ‘birds of a feather flock together’, i.e. similar people and households cluster spatially. Information on age, ethnicity, education, employment and type of housing etc. is used to paint a picture of the type of people who live in an area. If similar people live in similar places then knowing information about one person enables information about others in that locality to be broadly inferred (Sleight, 2004; Weiss, 2000).

Area classifications provide a unique way of bringing together spatial patterns from a range of variables, and identifying similarities and dissimilarities between areas (Webber and Craig, 1978). In its widest sense, a scheme of classification represents a convenient technique for the organization of a large data set to enhance the efficiency of recovery of information. Class labels describing arrangements of differences and similarities between objects provide a convenient summary of the data (Everitt *et al.*, 2001).

The complexities of 223060 individual and different census OAs from the 2001 census of the UK are too much information for the human mind to process at one time. However, by clustering these areas into a handful of groups which share similar properties, our understanding of the areas is greatly increased. The reduction in the amount of data makes it much easier for our brains to process the information; we can begin to see patterns in the distribution of the different types of area.

OAs are the smallest geographical units for which data from the 2001 census of the UK are released. The three census agencies, the ONS for England and Wales, the General Register Office for Scotland for Scotland and the Northern Ireland Statistics and Research Agency 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 (Office for National Statistics, 2005b). The ONS and Northern Ireland Statistics and Research Agency 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 (Office for National Statistics, 2005b). Table 1 shows how these different methodologies have affected the number and size of OAs that have been produced in each country. Clear and detailed descriptions of the methods that were used to create 2001 census OAs are given in Martin (2002a, b) and Martin *et al.* (2001).

Table 1. Average size of OAs in the constituent countries of the UK

Country	OAs	Population	Households	Average population per OA	Average households per OA
UK	223060	58789194	24479439	264	110
England and Wales	175434	52041916	21660475	297	124
Scotland	42604	5062011	2192246	119	52
Northern Ireland	5022	1685267	626718	336	125

3. What is cluster analysis?

Area classifications are created by the clustering of geographical entities by using specific methods. The process of cluster analysis, although based largely around clustering algorithms, is much wider than just the clustering of the objects themselves. To run a cluster analysis and therefore to create an area classification requires a series of steps, with multiple decisions to be made at each stage (Milligan and Cooper, 1987). There is no right or wrong answer to the majority of decisions that must be made. Each decision simply produces an alternative result, of what is one of an infinite number of parallel classifications. Consequently different decisions could be more or less suitable dependent on the purpose of the classification that is to be created (Lorr, 1983).

The steps that are involved in cluster analysis were summarized by Milligan (1996), who outlined the ‘seven steps of cluster analysis’. Milligan’s seven steps were further simplified by Everitt *et al.* (2001), who added their own comments and ideas to Milligan’s framework. The steps are described as ‘fairly predictable’ (Milligan (1996), page 341). Each step represents a major or critical decision that must be taken to run a cluster analysis successfully. Milligan suggested that it is vital that the user recognizes the critical decisions that need to be made, and their effect on the final results. A clear distinction needs to be made between cluster analysis and clustering method. The clustering method is simply the method by which the clusters are formed, whereas cluster analysis refers to the much wider sequence of steps that must be followed to complete the whole analysis (Milligan, 1996).

It is essential for users of cluster analysis, especially those hoping that their classifications will be used by others, that they record and report decisions that are taken at each step of the cluster analysis and the reasoning behind each decision. This enables others not only to evaluate critically how the classification was created but also gives them the possibility of adding to or extending the results of the analysis (Milligan, 1996). There are many examples of researchers who have failed to provide significant information about the decisions that were taken. Milligan cited Harrigan (1985) who failed even to name the clustering method that was used in the study. No-one is more guilty of failing to provide information about the creation of classifications and the steps that are used in cluster analysis as the firms who create and license geodemographic classifications. Harris *et al.* (2005) recognized that little is known about how commercial geodemographic classifications are built or what information goes into them. The problem exists because of a need for commercial confidentiality, but anyone who wishes to use these potentially rich sources of data in an academic study needs to be aware of their lack of transparency.

Milligan’s seven steps are now paraphrased with further comments by Milligan (1996), pages 342–343, Everitt *et al.* (2001), page 179, and the authors, which relate more directly to area classification.

3.1. Step 1: clustering elements (objects to cluster, also known as ‘operational taxonomic units’)

- (a) These should, where possible, be defined to give a 100% geographical coverage.
- (b) These should be representative of the cluster structure that is believed to be present.
- (c) These should be sampled properly if generalization to a larger population is required.

3.2. Step 2: clustering variables (attributes of objects to be used)

- (a) The variables represent the measurements that are taken on each entity or area that is to be clustered.

- (b) Variables should only be included if there is a good reason for their presence such as adding definition to the clusters.
- (c) Irrelevant or masking variables should not be included as they can hide more significant patterns within the clusters.

3.3. *Step 3: variable standardization*

There is no requirement for standardization to be applied to any set of data. It is up to the researcher to decide whether standardization is necessary and if so which method should be used.

3.4. *Step 4: measure of association (proximity measure)*

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

3.5. *Step 5: clustering method*

- (a) The methods that are used should be those designed to recover the type of clusters that are suspected to be present. This is important as different types of clustering method are better at finding particular types of cluster structures.
- (b) The method should be robust and able to handle different amounts of data, and be sensitive to different kinds of data.

3.6. *Step 6: number of clusters*

- (a) This is the most difficult decision to be made in cluster analysis. It is especially troublesome if there is no prior information about the number of clusters that are expected to be in the data set.
- (b) There are several different rules that can be followed for the selection of the most suitable number of clusters. However, these can often be contradictory for the same application.
- (c) Also needing consideration is whether there are actually any clusters in the data. There may be no obvious difference between the different solutions that are produced.
- (d) There is no right answer to the selection of the number of clusters. The choice is not based on scientific theory and the solution that is selected should be judged on its usefulness rather than being a correct representation of the patterns within the data set.

3.7. *Step 7: interpretation, testing and replication*

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

- (b) Rerun the analysis, choosing a different starting- or 'seeding' point for the initial cluster centres, to ensure that the same solution is found on all runs of the procedure.
- (c) Test to determine whether there is a significant cluster structure within the data. Follow by cross-validation to investigate whether the clusters are representative of data that were not originally included in the analysis.
- (d) Test for sensitivity of the classification to variable choice: examine the difference that is caused by the removal of each of the variables that are included in the analysis.

4. Which variables and transformations should be used?

The goal of the choice of variables for this classification was to select the minimum possible number of variables that satisfactorily represent the main dimensions of the 2001 census (Bailey *et al.*, 1999, 2000). The variables were selected solely from the 2001 census. There are several reasons why it was felt that using non-census data would be inappropriate. The census is the most complete and reliable socio-economic data set available in the UK (Rees *et al.*, 2002b). No other data set has the same amount of data with such a comprehensive geographic coverage. Another important factor is the geography of the data. At present the only substantial body of official statistics that are available at OA geography are data from the 2001 census, so using data from other sources would require converting the data from other scales. The linking of data sets at different spatial scales would create all kinds of reliability issues (Vickers, 2003). Complete confidence in all data that are included in the classification is needed if the classification is to be published as a National Statistic.

Some data, such as credit reference information and life style data, are only available for inclusion in commercial classifications as they are collected by the companies who make the classification or companies with whom they have data sharing agreements. However, there are several dangers that should be taken into account when using different sources along with data from the census. Firstly, the accuracy of the data must be assured. Even data from official sources such as Government departments can contain errors. It is important to know how and when all the data were constructed. Few data sets are as well documented as the census in terms of the enumeration and processing methods and it is unusual to find significant data support for any of these other sources of data. The coverage of the data will not be the near 100% that is available in the census. Additionally these data are unlikely to be representative either geographically or demographically; certain sections of society are likely to be overrepresented or underrepresented in the data. There will also be many other sources of uncertainty. It is impossible to know whether the data contain undocumented and unidentifiable errors.

The most obvious topic that was not covered by the 2001 census is information on income and wealth. It is believed that creators of commercial geodemographic classifications add data on income from sources such as credit and consumer card records (Harris *et al.*, 2005). Some of the data are obtainable or available on request and under licensed conditions.

It is important that any non-census data which are brought into the classification are at the same spatial scale and refer to the same geographical system. Although several methods of transferring data between spatial scales have been formulated and indeed some are used widely, no system has yet been formulated that can transfer data between overlapping areal units to a satisfactory level of accuracy (Vickers, 2003). There are census variables such as car ownership and type of housing which give a good proxy for income and wealth. These have the advantage over the data from other sources of being complete in coverage and open and reliable in terms of data collection methodology. It is hoped that a question on income will be added to the 2011 census, which would be very welcome and would enable an improved classification to be created.

By reviewing the census data that are available at OA level, five main domains have been identified: demographic structure, household composition, housing, socio-economic group and employment. The key statistics were identified by the ONS as being the most important census variables in the design of this data set and were the first to be published. The initial data set that was assembled for this study included all variables from the OA level key statistics tables. The key statistics represent the most important variables in the census and have a comparatively simple data structure that aids data extraction. An initial selection of 94 variables was made with the intention of representing the main domains of the census; this list was then reduced significantly following detailed assessment of each variable (Vickers *et al.*, 2005). A number of principles for variable selection were followed. They are outlined in Sections 4.1–4.9.

4.1. Highly correlated variables

The inclusion of pairs of variables with strong correlations within a data set is undesirable for cluster analysis, because they represent data redundancy. To look for redundancy within the data set a correlation matrix of all 94 variables was constructed. Including highly correlated variables makes it very difficult to assess the effect of any individual variable on the clustering process. Some strong correlations were found in the initial set of variables. Common sense suggests that one of each pair of highly correlated variables should be removed because much of the information is redundant; 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 for use in the classification (Voas and Williamson, 2001). It is likely that variables that can predict the value of other variables would enable the classification to predict other behaviours. Therefore, there is an advantage in retaining a high proportion of highly correlated variables as they can be seen as powerful predictors. Therefore no general rule can be implemented when considering the correlations between variables. In the main an attempt was made to remove the strongest correlations from the data set. However, each case needs to be considered on its individual merits rather than the implementation of a general rule.

4.2. Variables with badly behaved distributions

Methods of clustering and standardization work reliably with data that have a normal distribution. However, highly skewed distributions can create problems in both the standardization and the clustering procedures. The skew that is observed most often in census data and the one that causes the most problems when clustering is a positive skew, i.e. the majority of the data are found at the lower end of a 0–100% scale with only a few high values. The most common form is when a variable identifies only small sections of the population. Another way to look at this is that the majority of areas have an absence of a particular feature leading to a large number of zeros within a specific variable. These problems become more acute as the spatial scale reduces because the likelihood of extreme values increases.

4.3. Composite variables

Composite variables can be formed from two related variables which show similar patterns. These variables must share the same denominator (otherwise the proportion of people relating to that variable could exceed 100%). This method can be used to group 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 a single all-flats variable. This is especially important when working with OAs because the numbers can be small and affected by disclosure controls that are imposed on the data (Stillwell and Duke Williams, 2007). For an explanation of what disclosure control entails and the effect that it has on the data see Office for National Statistics (2003a).

4.4. *Constancy of variables across the whole UK*

Some questions were asked in all countries of the UK (England, Wales, Scotland and Northern Ireland) 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 Christian denominations whereas other religions were combined into one category. In England and Wales all types of Christians were reported in a single variable and other religions were reported separately (Buddhists, Hindus, Jews, Muslims and Sikhs). Some interesting patterns may be visible, but if data are not available for all parts of the UK it is not possible to include the variable in the classification.

4.5. *Vague or uncertain variables*

It seems sensible to assume that all census variables are produced in the same way, i.e. from the answers that were written on each census form. However, this is not so for all variables. For example, the 'household spaces with no residents' variables in census Table KS16 was coded as either 'Vacant' or 'Second residence/holiday accommodation'. Unlike other census variables there was no-one to fill in a form for these variables because the properties were empty on census day. The values of the variables were imputed by the census enumerator on the basis of their own judgment. It is widely accepted that 'Second residence/holiday accommodation' was underrecorded by using this method, especially in the more rural parts of the country.

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

4.6. *Uninteresting geographic distributions*

For variables to work in the classification they need to show variation over space. Not all ethnic groups show the same distribution over space. Some are distributed fairly evenly, but others show a more ghettoized population (Simpson, 2007). Peach (1996) explored this phenomenon by asking the question 'Does Britain have ghettos?' to investigate to what extent different ethnic groups are dispersed throughout Britain. For all groups apart from white and Chinese, over 60% of their population are found in four major urban centres. White and Chinese populations vary less significantly over space. Black Caribbean, black African, black other, Indian, Pakistani and Bangladeshi variables would add more to the classification than white or Chinese variables because their distributions vary more over space.

4.7. Consistency of the variable for the lifetime of the classification

The longevity of the classification must be considered as the classification is likely to remain the most current ONS area classification until after the release of the 2011 census results. Any variable whose understanding by the user may change over the life course of the classification should not be included. A variable that was considered for use in the classification was 'Born in other European Union (EU) country (excluding UK and Republic of Ireland)'. On census day April 29th, 2001, there were 15 members of the European Union; on May 1st, 2004, a further 10 countries joined. The consequence of this is that the 'Born in other European Union' variable in the census no longer reflects the current membership of the European Union. There are also applications to join from Balkan states and Turkey. If and when these countries join, the number of member countries of the European Union 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.

4.8. Standardization of variables affected by the population's age distribution

The percentage of the population suffering from limiting long-term illness as provided in the key statistics Table KS08 could have been used in its raw form, as it was to create the local authority classification (Office for National Statistics, 2005a). However, this was considered unsatisfactory as crude rates are greatly affected by the age structure of the population at fine geographic scales. An area which has a high proportion of older people (all other things being equal) tends to have a much higher rate of illness than an area with a younger population irrespective of their relative mortality rates.

Therefore the limiting long-term illness data were age standardized. Only when this is done will the relationship of illness with other variables become clear. The technique that was used to do this is the standardized illness ratio SIR, which compares the expected count of illness for an area with the observed count. The expected count is created by multiplying age-specific rates of illness for the whole UK population by the OA population by age. SIR for an area is defined as

$$\text{SIR}^i = 100I^i / \sum_a r_a^n P_a^i \quad (1)$$

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

SIR is a relative measure. The national illness ratio always has an average value of 100. A value of 150 shows that an OA experiences a rate of illness that is 50% greater than the age-specific rates for the whole population of the UK. A value of 50 means that the OA experiences a rate of illness that is 50% lower than in the UK population. There is substantial variation between the OAs with values ranging from 0 to 505.

4.9. Standardizing the variables

All clustering techniques are based on the similarity or dissimilarity of the cases to be clustered. This is measured by constructing a distance matrix, reporting the distance between pairs of cases (OAs) for each variable. It is clear that problems will occur if there are differing scales or magnitudes between the variables. In general, variables with larger values and greater variation will have more effect on the final similarity measure. It is therefore necessary to make each variable equally represented in the distance measure by standardizing the data. The preferred method of standardization for the OA classification is range standardization. This method was implemented in the ONS 1991 classification of local authorities (Wallace and Denham, 1996).

The data were standardized by the method which yields for each variable a value lying in the range 0–1. It defines values R_i for variable with value x_i in area i as

$$R_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (2)$$

where x_{\max} is the maximum value taken by x in the data and x_{\min} the minimum.

5. How were the variables selected?

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 merged. This section outlines the decisions that were taken. All variables were compared with all others to remove redundancy, although other factors were also considered in the choice of variables.

The choice of variables was made in collaboration with the team from the ONS who created the ward level classification. This joint effort was intended to match the variable selections at both geographies as closely as possible. This was done with the aim of making the classifications as simple and comparable as possible. The comparability across scales is an important part of the project. The area classification systems that were created were to be disseminated as a suite of systems to be used together. Within the process of variable selection some sacrifices were made at one scale to aid comparability with the other. How the various scales of classification can be used together and how they complement each other are explained in Vickers (2006).

5.1. Gender

Both the percentage of males and females variables were rejected as gender told us very little about an area. The majority of areas had very similar gender proportions.

5.2. Communal establishments

It was decided not to include the proportion of people who live in communal establishments as there are many areas with a zero value for this variable. Inclusion could lead to objects being clustered 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 term that covers residences for several different population groups, including care homes, hostels, prisons and university residences. These house four different types of people with little in common who would be grouped with the inclusion of this variable.

5.3. Urban–rural status

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

5.4. Age

There are 16 age categories in the 2001 census results. The age variables contain much data redundancy as they are highly correlated with each other. The age categories can be combined to create wider, more stable age groupings. The age variables which were selected for inclusion were the percentage of residents aged 0–4 and aged 5–14 years to pick up the difference between

younger and older children. The percentage of people who were aged 15–24 years was not used as it was highly correlated with students. The percentage of people aged 25–44, aged 45–64 and aged 65 years or older were all selected.

5.5. Marital status

The percentages of people aged 16 years and over who were married, cohabiting or 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. The category divorced was combined with separated as they were strongly correlated, which brought more stability to the variable.

5.6. Ethnic identity

The percentage of people who were born outside the UK was retained as a variable as it gave an indication of international migration. The percentage of the population with Indian, Pakistani and Bangladeshi ethnicity was kept as was the percentage of people who are black African, black Caribbean or other black as they showed an interesting geographic distribution and identified significant minority populations. The percentage of people who are Chinese was not included as their geographic distribution showed much less variation. All the religion variables were dropped owing to a high correlation with ethnicity and the voluntary nature of the question in England and Wales.

5.7. Health

Two of the health variables were included. Limiting long-term illness was included but it was standardized by age, creating the standardized illness ratio SIR. This enabled 100% of the population to be used, which is important as the OAs are small areas. As the population of some areas may be mainly outside the working age population, using the percentage of the working age population who report illness may not be reliable for some areas with a high elderly population. However, this was considered suitable for the ward level classification as the populations of wards are significantly larger. The percentage of people whose health is good, fairly good and not good were all found to be highly correlated with limiting long-term illness. The other health variable that was included was the 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 limiting long-term illness variable, would also give an indication of how well people are cared for.

5.8. Employment

People working part time and people who were unemployed were included; those working full time were not owing 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. Never worked and long-term unemployed were not included as they only identified small sections of the population and were highly correlated with unemployment. The percentage of unemployed who are long term unemployed was used in the ward classification but could not be used in the OA classification owing to the effect of disclosure control on the data. Several OAs reported values of over 100% when values for this variable were calculated. Because of the errors in the variable that was created we could have little confidence in this variable and it was therefore not used in the OA classification.

5.9. *Industry*

Of the 12 industry sector groups in the original list seven (agriculture, hunting, forestry and fishing; mining, quarrying and construction; manufacturing; hotel and catering; health and social work; financial intermediation; wholesale and retail) were included as they showed interesting geographic patterns. The other five (electricity, gas and water supply; transport, storage and communication; real estate, renting and business activities; public administration and defence; education) were rejected because of less distinctive geographic distributions and limited representation in terms of numbers.

5.10. *Occupation*

The nine occupation groups 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, percentage of people with qualification levels 4 and 5 (degree level and above) was included; those with 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 were highly correlated with other variables such as employment, qualifications, ethnicity and health. The only two variables from the original list that were included were semi-routine occupations and routine occupations, which were combined to give an extra variable indicating lower social standing.

5.11. *Commuting*

The percentage of people who 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 geographic patterns.

5.12. *Housing tenure*

Renters from both the private and the public sector are included as they indicate several things including transitory status (rent private) and lack of wealth (rent public). The second residence or 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 recognized that these data are unreliable, especially at such a small scale.

5.13. *Type of accommodation*

Detached and terraced housing variables were included; semi-detached housing was not included as it was highly correlated with other types of housing and was less descriptive. Also it 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 accounted for only a very small part of the population.

5.14. *Car availability*

The variable 2 or more cars was included in preference to no car households because the two variables are very highly correlated and to add additional information on affluence.

Table 2. Full list of the 41 variables that were selected for input to the classification

<i>Variable</i>	<i>Definition</i>
<i>Demographic</i>	
v1	<i>Age 0–4</i> : percentage of resident population aged 0–4 years
v2	<i>Age 5–14</i> : percentage of resident population aged 5–14 years
v3	<i>Age 25–44</i> : percentage of resident population aged 25–44 years
v4	<i>Age 45–64</i> : percentage of resident population aged 45–64 years
v5	<i>Age 65+</i> : percentage of resident population aged 65 or more years
v6	<i>Indian, Pakistani or Bangladeshi</i> : percentage of people identifying as Indian, Pakistani or Bangladeshi
v7	<i>Black African, Black Caribbean or Other Black</i> : percentage of people identifying as black African, black Caribbean or other black
v8	<i>Born outside the UK</i> : percentage of people not born in the UK
v9	<i>Population density</i> : population density (the number of people per hectare)
<i>Household composition</i>	
v10	<i>Separated/divorced</i> : percentage of residents 16 years old or older who are not living in a couple and are separated or divorced
v11	<i>Single person household (not pensioner)</i> : percentage of households with one person who is not a pensioner
v12	<i>Single pensioner household</i> : percentage of households which are single-pensioner households
v13	<i>Lone parent household</i> : percentage of households which are lone parent households with dependent children
v14	<i>Two adults no children</i> : percentage of households which are cohabiting or married couple households with no children
v15	<i>Households with non-dependant children</i> : percentage of households comprising one family and no others with non-dependent children living with their parents
<i>Housing</i>	
v16	<i>Rent (public)</i> : percentage of households that are public sector rented accommodation
v17	<i>Rent (private)</i> : percentage of households that are private or other rented accommodation
v18	<i>Terraced housing</i> : percentage of all household spaces which are terraced
v19	<i>Detached housing</i> : percentage of all household spaces which are detached
v20	<i>All flats</i> : percentage of households which are flats
v21	<i>No central heating</i> : percentage of occupied household spaces without central heating
v22	<i>Average house size</i> : average house size (rooms per household)
v23	<i>People per room</i> : average number of people per room
<i>Socio-economic</i>	
v24	<i>HE qualification</i> : percentage of people aged between 16 and 74 years with a higher education qualification
v25	<i>Routine/semi-routine occupation</i> : percentage of people aged 16–74 years in employment working in routine or semiroutine occupations
v26	<i>2+ car household</i> : percentage of households with 2 or more cars
v27	<i>Public transport to work</i> : percentage of people aged 16–74 years in employment who usually travel to work by public transport
v28	<i>Work from home</i> : percentage of people aged 16–74 years in employment who work mainly from home
v29	<i>LLTI (SIR)</i> : percentage of people who reported suffering from a limiting long-term illness (standardized illness ratio, standardized by age)
v30	<i>Provide unpaid care</i> : percentage of people who provide unpaid care
<i>Employment</i>	
v31	<i>Students (full-time)</i> : percentage of people aged 16–74 years who are students
v32	<i>Unemployed</i> : percentage of economically active people aged 16–74 years who are unemployed
v33	<i>Working part-time</i> : percentage of economically active people aged 16–74 years who work part time

(continued)

Table 2 (continued)

<i>Variable</i>	<i>Definition</i>
v34	<i>Economically inactive looking after family</i> : percentage of economically inactive people aged 16–74 years who are looking after the home
v35	<i>Agriculture/fishing employment</i> : percentage of all people aged 16–74 years in employment working in agriculture and fishing
v36	<i>Mining/quarrying/construction employment</i> : percentage of all people aged 16–74 years in employment working in mining, quarrying and construction
v37	<i>Manufacturing employment</i> : percentage of all people aged 16–74 years in employment working in manufacturing
v38	<i>Hotel and catering employment</i> : percentage of all people aged 16–74 years in employment working in hotel and catering
v39	<i>Health and social work employment</i> : percentage of all people aged 16–74 years in employment working in health and social work
v40	<i>Financial intermediation employment</i> : percentage of all people aged 16–74 years in employment working in financial intermediation
v41	<i>Wholesale/retail trade employment</i> : percentage of all people aged 16–74 years in employment working in the wholesale or retail trade

5.15. Household size, density of occupation and amenities

Average household size was rejected as it did not reveal information about distinct types of household. The average number of rooms per household was included as it gave a good indication of affluence. The variable ‘occupancy rating of –1 or less’ was rejected in favour of a new variable, people per room. The occupancy rating provides a measure of underoccupancy and overcrowding. A value of –1 implies that there is one room too few and there is overcrowding in the household. The occupancy rating assumes that every person in the household requires a minimum of two common rooms excluding the bathroom (Office for National Statistics, 2003b). The percentage of households with no central heating was included as it is a good indicator of poor living conditions, but the percentage of households with no bath or shower was rejected as the numbers are very small. Lowest floor above ground level was not included as it was highly correlated with flats.

5.16. Household composition

Single-pensioner households and single-person non-pensioner households were both included as they identify a housing situation which is increasing in prevalence. All pensioner households (family) was rejected as this variable was highly correlated with single-pensioner households and age 65 years or more. The percentage of households containing two adults and no children and lone parent households were both included as they show interesting opposing residential situations. All-student households was rejected as it is highly correlated with all students. All-pensioner (other) households was rejected owing to correlation with similar variables. The percentage of households containing no adult in employment with dependent children was not included as it was highly correlated with lone parent households. 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 that they experience in trying to get onto the housing ladder.

5.17. The variables chosen

The decisions on the variables to be included in the classification were made by comparing statistical information and working within a theoretical framework considering such things as

redundancy and frequency of the phenomena. However, as with any choice of variables for a classification, a different group of people may have made different decisions, resulting in a different list of variables.

Table 2 lists the 41 variables that were selected for input to the classification and gives them a short definition and a longer verbal description. This final list of variables results from the implementation of the decisions that were made.

6. Assembly and checking of the database

To cluster the OAs into groups the data all need to be in one database. This sounds sensible and quite simple. However, for each key statistics table from the 2001 census there are 12 separate tables that need to be joined: nine representing the English Government office regions (GORs), one for Wales, one for Scotland and one for Northern Ireland. The data were published in this way because to put the data into one file would have made it too big to be opened in one of the most commonly used packages for handling statistics 'Microsoft EXCEL'. The tables could not simply be joined in sequence because some table formats were different in each UK country. Therefore, a computer program was written to extract the data from each table in the relevant format and to write the variables to a single file in a common format. For a detailed description of the data sources and extraction see Vickers *et al.* (2005).

The England and Wales data were count data which were converted into percentages, whereas the Scotland and Northern Ireland data were available as percentages. Some variables were more problematic than others: working part time presented a particular challenge. The variable is in the second column of key statistics Table KS09 for England and Wales, and Scotland, but is in the third column for Northern Ireland. In the Northern Ireland table working part time and working full time are in the opposite order from that for England and Wales, and Scotland. Relatively few variables have the same table reference so constant checking and rechecking were needed to ensure that the correct data had been selected. Two forms of data checking were conducted on the database to ensure that correct values had been entered.

The first form of data checking was to extract a sample of variable values for individual OAs manually, to establish whether the data in the database matched the data in the original census tables. This check essentially tested the reliability of the data extraction program. The database is assembled from 12 tables (for GORs and countries) and 41 variables. Therefore, to test that each table was extracted and reassembled correctly, a check on the data for each GOR for each variable was performed: a minimum of 492 (12×41) separate checks were needed. As the data were extracted automatically it can be assumed that if one item is wrong then everything extracted from that table is wrong. However, to add more rigour to the test the same OA was not selected for each table. Every 2000th 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 that was done by the extraction program was repeated manually. The value of the relevant OA and variable from the original census tables was compared with the corresponding value in the database.

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 with 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 then 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. the total population, number of households

and people of working age). All the OAs in each GOR or country were then summed and the result was checked against the value given for the GOR or country in the census table. Some small differences are unavoidable due to rounding when multiplying out the data and because of the effects of disclosure control.

Three of the GORs (Eastern, South East and London) showed very large differences for the variable 'percentage of people who provide unpaid care'. Each of these GORs was found to have approximately 500 000 people fewer in the OA data than in the GOR or country data. 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 right? Which was wrong? This was fairly simple to deduce as the GOR tables showed a similar level for the variable across all GORs. In the OA data the percentage of people who provide unpaid care was significantly lower in the three GORs in which the discrepancies were found. 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 a result of errors in the original census data.

These data checks are not 100% foolproof: for such certainty all 9 million data points in the clustering database would need to be individually checked. However, the data checks do provide a strong indication that it is unlikely that any errors remain in the data set. The checks were designed to find errors both by checking back to the original OA data and against data at another geographic scale to see whether the values were consistent; the error that was picked up shows that the data checking procedures worked. Input data validation is a time-consuming process, but essential for quality assurance of the inputs.

7. Clustering method

After experiments with various clustering methods, the k -means algorithm in the SPSS statistical package was chosen, although several other statistical packages are available which could also have done the job. The use of a hierarchical method (Ward's algorithm) proved problematic as Ward's method struggled with the large number of objects to be clustered, in terms of both the computing power and the time that is required to run a hierarchical method on a large data set. Unequal clusters are also produced because of the method's inability to separate objects again once joined, an issue which was recognized by Everitt *et al.* (2001). Experiments, using a combination of a k -means algorithm to create 'starter clusters' followed by a hierarchical clustering, resulted in an unsatisfactory solution. The process produced a few large clusters and many clusters with few members and was rejected (Vickers *et al.*, 2005). Instead a method of generating a hierarchical classification by successive applications of the k -means algorithm was devised. The objective of the k -means algorithm is to minimize the within-cluster variability. If the number of clusters within the data set has already been prespecified, a k -means classifier can be used, for example, to form five clusters that are as distinct from each other as possible (Everitt *et al.*, 2001). The k -means method is one of the most commonly used methods in the geodemographics industry (Harris *et al.*, 2005). The k -means method uses an iterative relocation algorithm based on an error sum of squares measure. The basic operation of the algorithm is to move a case from one cluster to another to see whether the move would improve the sum of squared deviations within each cluster (Aldenderfer and Blashfield, 1984). The case will then be assigned or reallocated 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) to assess the distinctiveness of the clusters (Everitt *et al.*, 2001).

The way in which clusters are identified is by a measurement of how close objects are in multidimensional space. This can be calculated by either a similarity or a dissimilarity measure. A similarity measure (proximity) will report the largest value for the two objects that are closest together and the smallest value for the two objects that are furthest apart. Conversely a dissimilarity measure (distance) will report the smallest value for the two objects that are closest together and the largest value for the two objects that are furthest apart (Everitt *et al.*, 2001). Many different measures of both similarity and dissimilarity can be used within cluster analysis. Squared Euclidean distance is the measure that was used here as the basis for the OA classification.

As a consequence of the use of squared Euclidean distance based on m variables a measure of the overall similarity between the n_c entities forming a cluster indexed by c is

$$E_c = \sum_{i=1}^{n_c} \sum_{j=1}^m (Z_{ij} - \bar{Z}_{cj})^2. \quad (3)$$

Here Z_{ij} is the value for the i th member of the cluster on variable j and \bar{Z}_{cj} is its cluster mean. When summed across the clusters forming a classification this is the within-cluster variation that forms the clustering criterion that is used in the algorithm above.

8. Choosing the cluster levels and numbers

The creation of the classification works as follows: the k -means algorithm is run on the data set and (for example) five clusters are produced. The original data set is then split into five separate data sets (representing the highest level of the hierarchy, supergroups). Each of the five new data sets is then separately input into the k -means algorithm to create the second level of the hierarchy (groups). The second level of the hierarchy is then split into separate data sets (not necessarily containing the same numbers of clusters) and each is input into the k -means algorithm to create the lowest level of the hierarchy (subgroups).

The principles that were used to make the choice of the number of clusters and the choices themselves are now explained. The decisions were based on a plethora of information that can be output from the clustering process. The first principle that we adopted was to aim for a nested hierarchy of area classes. GB Profiles, the publicly available small area classification based on 1991 census enumeration districts (Rees *et al.*, 2002a) offered three cluster tiers but the clusters at each tier were independently generated. Most commercial classifications link two or three tiers in the classification together (Harris *et al.*, 2005). Why should hierarchical classifications be more popular? Hierarchies exist in all societies within and between organizations. The logical realm is described through hierarchical taxonomies, linked as a result of evolution. It is thus 'natural' to see smaller groups as parts of larger groups rather than small and large groups being separate and unlinked.

The hierarchy was implemented by first clustering the whole set of OAs to create the supergroups. Then the data for each supergroup were stored in a separate file. Each data file was then reclustered separately. This was then done again for the groups (the middle tier) to create the subgroups (the lowest tier).

The process of selection of the number of clusters involved consultation with potential users and members of the ONS area classification advisory board. Callingham (2003) supplied advice about which would be the most suitable number of clusters for users, based on many years of experience in using classification systems, which it is useful to quote in detail (Callingham (2003) (highlighting in italics by the authors)):

‘At the highest level of aggregation, the cluster groups should be about 6 in number to enable good visualisation and these clusters should also be given descriptive names. At the next level of aggregation, the number of groups should be about 20. This would be good for conceptual customer profiling (that is, when one wants to gain some conceptual understanding of one’s customer base) and would also allow market propensity measures to be established with comparatively small surveys (for example, two waves of an omnibus). This level could also be used for setting up sampling points for some market research surveys and would ideally also have descriptive names. At the next level of aggregation, the number of groups should be about 50. This can be used for market propensity measures from the larger commercial surveys such as TGI [target group index] 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’s comments give good guidance on the suitability for use of different numbers of clusters in the solution. Each level has a different purpose. The classification needs to be fit for purpose so much attention needs to be paid to the number of clusters that are created for each hierarchical level.

Callingham (2003) suggested that the most useful number of clusters in the first level would be around 6. Taking this as a starting-point, cluster solutions from 4 to 8 were examined to see how the average within-cluster distance changed. Fig. 1 shows how the average distance to the cluster centre is reduced as the number of clusters increases. The smaller the average distance to the cluster centre the more compact the cluster solution is. However, the reduction in the average distance to the cluster centre with the increase in the number of clusters does not happen at a constant rate. It is therefore necessary to consider not which solution produces the smallest average distance to the cluster centre (as this would simply be the solution with the largest number of clusters), but which solution is the most compact relative to the number of clusters within the solution. Fig. 1 enables the identification of most compact cluster solutions relative to the number of clusters. These solutions are those which display the steepest increase

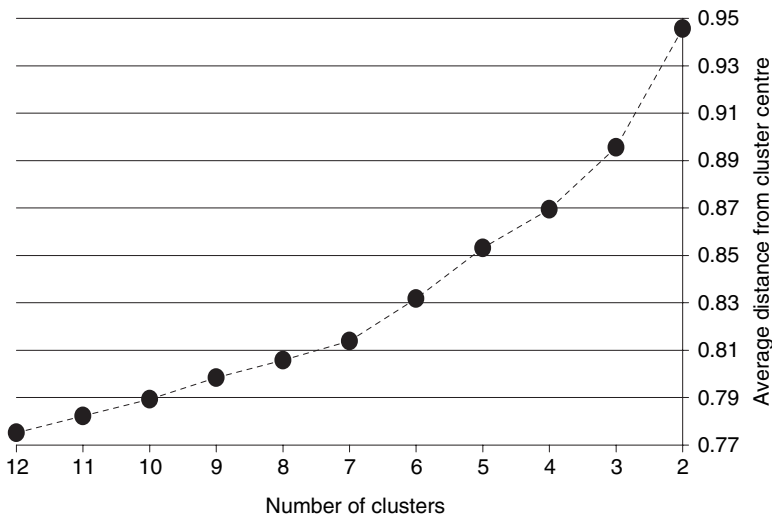


Fig. 1. Average distance from the cluster centre by number of clusters

in average within-cluster distance between themselves and the solution which creates one fewer cluster. Within this range it was not evident that there is any significant difference in the increase in the average distance from the cluster centre, although there appears to be a peak at 5 (which is less compact than would be expected relative to the number of clusters) and a slight trough at 7 (which is more compact than would be expected relative to the number of clusters).

Another factor that must 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 the number of members). It is preferable to have the clusters as closely sized to each other as possible. If creating two clusters from 10 objects, two clusters both containing five members would be the optimal solution in terms of usability of the resulting classification. In contrast, a solution of one cluster with nine members and another with only one member would be the worst solution. This would not have actually created two clusters, but only removed an outlier from the original data set. An explanation using 10 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 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 size of membership of the clusters becomes smaller. If the cluster was small to start with, this greatly increases the chances of creating even smaller clusters at a lower level.

To make sure that the classification did not produce unevenly sized clusters in terms of membership, 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 most preferable solution in terms of the number of members in each cluster. The simple example in Table 3 shows three possible solutions for clustering 12 objects into two, three or four clusters. The two-cluster solution has an average difference from the mean (in this case 6) of 2. The three-cluster solution has a smaller distance from the mean (in this case 4) at just 1.33. The four-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 homogeneous they are in terms of the number of members, the three-cluster solution would be selected as the optimal solution.

Table 3. Example of the method for calculating which solution is most homogeneous in terms of the number of members in each cluster†

	<i>Number of members in each cluster for the following solutions:</i>		
	<i>2 cluster</i>	<i>3 cluster</i>	<i>4 cluster</i>
	8	4	2
	4	2	4
	NA	6	1
	NA	NA	5
Average distance from the mean	2	1.33	1.5

†NA, not applicable.

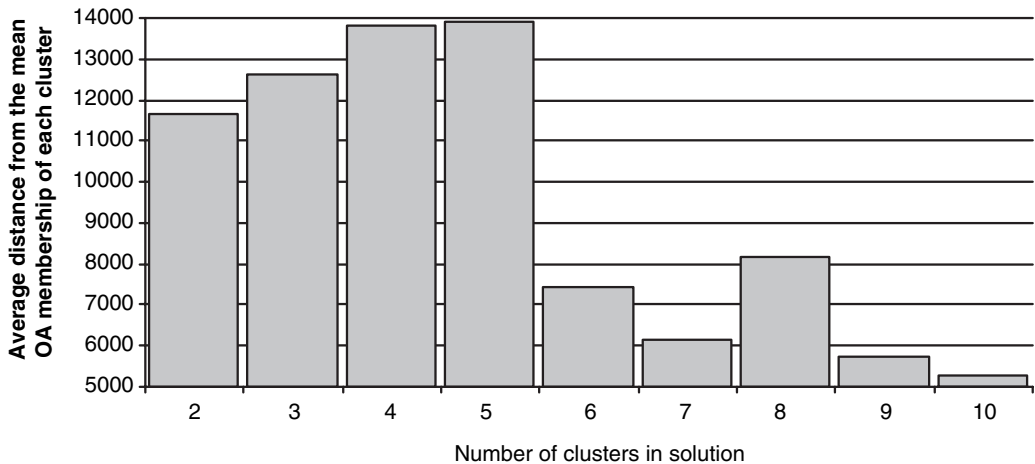


Fig. 2. Range in the size of clusters by the number of clusters

Table 3 shows how the method works on a small data set, but what results were produced by using this method on the possible solutions for the OA classification? Fig. 2 shows the average distance from the mean cluster membership for solutions of cluster numbers 2–10. The best solution based on this criterion is 10 clusters, followed by nine, then seven clusters. The worst solution is a virtual tie between the solutions containing four and five clusters.

A minimum cluster membership target of 50% of the average size of membership for each cluster levels was set. Therefore, if the first level contains six clusters, the minimum size would be $223060/6 \times 0.5 = 18588$. If the middle layer consisted of 25 clusters the minimum target would be $223060/25 \times 0.5 = 4461$. This target was put in place to achieve groups of fairly even sizes. However, this criterion was viewed flexibly and if a sensible group formed that was within about 10% of the target it would be acceptable. Also smaller groups were allowed if it meant that their non-formation would have prevented the splitting of a cluster into a lower level.

These two forms of analysis (change in intracluster distance and departure from average size of cluster) were run on the clusters to establish which cluster solution was most suitable to represent the first level of the hierarchy. The choice was based on the solution which performed well on both tests.

The four-cluster solution performs well in Fig. 1 but poorly in Fig. 2. The five-cluster solution performs poorly in both tests. The six-cluster solution performs reasonably in both tests. The seven-cluster solution performs reasonably in Fig. 1 and well in Fig. 2. The eight-cluster solution performs reasonably in both tests. Therefore four- and five-cluster solutions can be rejected for performing badly in one or both of the tests. This leaves six-, seven- and eight-cluster solutions which all performed equally well in Fig. 2, but in Fig. 3 the seven-cluster solution outperforms those of six and eight, suggesting that it is the best solution. Therefore the seven-cluster solution has been selected as the solution for the first level of the hierarchy. Table 4 shows that there is a fairly even distribution of OAs across the seven clusters.

Once the first level of the classification (supergroups) had been decided on as containing seven clusters, each cluster needed to be broken down to create the second level (groups) of the hierarchy. This was done in the same way as for the first level, by examining the average distance to the cluster centre, and identifying the solution which is the most compact relative to the number of clusters in the solution, as described earlier in this section. However, at this level only two, three or four clusters were considered to ensure that the number of clusters reflected

Table 4. Number of OAs in each cluster

	<i>Cluster 1</i>	<i>Cluster 2</i>	<i>Cluster 3</i>	<i>Cluster 4</i>	<i>Cluster 5</i>	<i>Cluster 6</i>	<i>Cluster 7</i>	<i>Range</i>
OAs	35837	16638	27743	47251	33166	40769	21721	30613
OA (%)	16.0	7.5	12.4	21.2	14.8	18.3	9.7	13.7

Table 5. Cluster names

<i>Supergroup</i>	<i>Subgroup</i>
1. Blue collar communities	1a. Terraced blue collar 1b. Younger blue collar 1c. Older blue collar
2. City living	2a. Transient communities 2b. Settled in the city
3. Countryside	3a. Village life 3b. Agricultural 3c. Accessible countryside
4. Prospering suburbs	4a. Prospering younger families 4b. Prospering older families 4c. Prospering semis 4d. Thriving suburbs
5. Constrained by circumstances	5a. Senior communities 5b. Older workers 5c. Public housing
6. Typical traits	6a. Settled households 6b. Least divergent 6c. Young families in terraced homes 6d. Aspiring households
7. Multicultural	7a. Asian communities 7b. Afro-Caribbean communities

as closely as possible the target number of clusters of around 20 and that the supergroups 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. The second level (groups) then needed to be split down again to create the third level (subgroups) of the hierarchy with a target size of around 50 clusters. To create the subgroups the clusters in the groups were spilt into two, three or four clusters, again considering the within-cluster difference and the number of OAs in each cluster. The subgroups consist of 52 clusters. Table 5 shows the structure of the classification, indicating into how many groups each cluster was split.

9. Profiling and naming the clusters

The final two steps in creating the classification are to profile each cluster and to name it on the basis of the profile features which distinguish the cluster from its fellow clusters. The idea behind cluster profiles is to create a short description, using text and visuals, which only takes a few seconds to read but significantly expands the user’s understanding of the group. The cluster profiles include graphs, photographs of typical homes or neighbourhoods and some statistical information along with an extended description of the clusters.

The cluster profiles were not easy to produce, especially for the subgroup level where the clusters are more numerous and in some cases not easy to distinguish from each other. However, at the subgroup level there are more extreme values. Therefore for many subgroups it is easier to see which variables are distinguishing that cluster from other subgroups. Clusters that show extreme values for one or more variables are easier to describe than clusters which have average values for all variables. This is perhaps not surprising as researchers tend to focus on exploring extremes, whether it is poverty or affluence; social and economic averageness is not studied as intensely. The descriptions also, where appropriate, contain information about the geographical distribution of the groups whether the group is found in a particular geographical *milieu*, in particular parts of towns and cities or only in rural areas. Specific place names are avoided, because these have resulted in geographical mislabelling in past classifications.

Fig. 3 shows the cluster summary of supergroup 3 'Countryside'. Portraits displaying radial plots for every cluster can be found in Office for National Statistics (2005a). Each summary has a radial plot which represents the standardized values for each variable. The numbers on the scale represent the difference from the mean value for that variable; therefore the mean for all variables is 0. The mean is represented by the middle ring at 0; the value of each variable for that supergroup can then be seen by the amount that each point (showing the variation from the mean for each value) is above or below the inner ring. The outer ring represents a value of 0.5 and the inner ring a value of -0.5 .

Naming the clusters is a difficult and perilous process. 9145460 data values (223060 OAs by 41 variables) have been inputted into the clustering process. In the cluster profiles the number has been reduced to 3280 (80 clusters multiplied by 41 variables) standard cluster means. Naming the clusters involves reducing the data values to about 240 words (80 clusters multiplied by three words)! The names cannot capture more than a tiny part of the distinctiveness of the clusters and must be regarded only as convenient labels which provide signposts for the user but which must be used with great caution to avoid imputing to all households in an OA in a cluster the key attributes that are highlighted in the name. There was considerable discussion between the ONS and the authors about the names to be adopted and the authors have received many useful comments, mainly objections to candidate names rather than alternative suggestions. The outcome of the discussion was a decision not to name clusters in the official version of the OA classification, although the larger area classifications for local authorities, health authorities and wards using 2001 census information had been named. It was, however, agreed that the authors could proceed with naming of the supergroup and group tiers of the classification but that these names would not have official status. The main argument for naming is that discussion about the classification and its uses is facilitated. It is easier to talk about 'Prospering older families' than 'Group 4b' for example.

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

Some comments and suggestions on names were 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

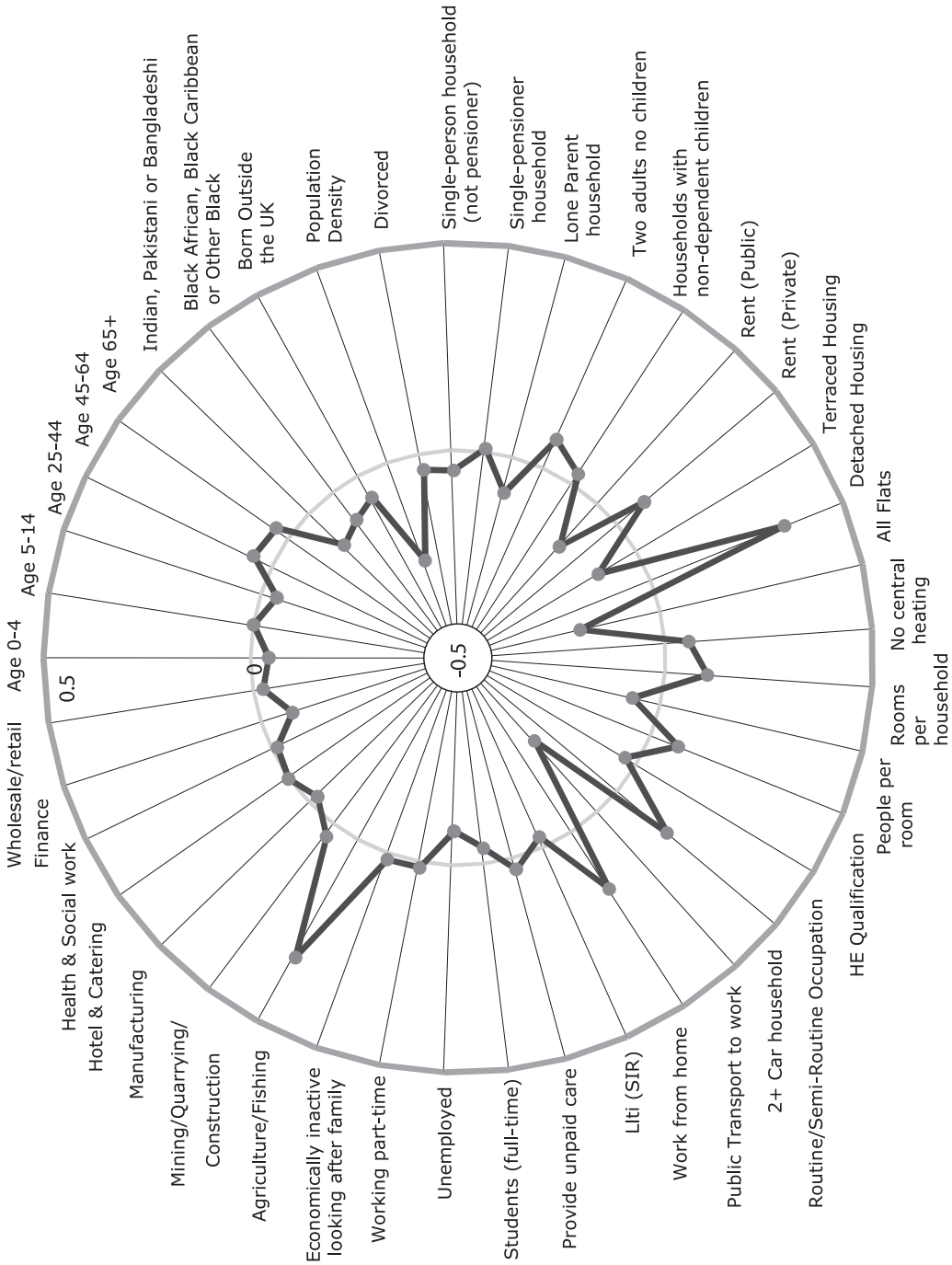


Fig. 3. Cluster summary of supergroup 3: 'Countryside'

revisions and names have moved from one group to another as it became apparent that a name that had already been given to a group was more suitable for a so far unnamed group.

The names (as displayed in Table 5) 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 that had been given to the previous classifications (local authority 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 the names which were selected had not 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 clusters had the same name would intuitively assume that the two clusters were intended to represent the same set of areas or people when this is not necessarily so. Armed with a dictionary and a thesaurus the task was then addressed with an open mind. The results are shown in Table 5.

10. Conclusions

This project has successfully created a unique picture of the geosocial make-up of the UK. This paper describes the creation of the OA classification, outlining the decisions that were taken and processes that were required in the production of the system. Area classifications and cluster analysis are introduced and outlined, providing the background to how the classification was created. The process of variable selection is carefully outlined, accounting for the decisions that were made in the inclusion or rejection of each variable. The clustering methodology that was used to assign the 223060 output areas into a hierarchy of seven, 21 and 52 clusters is outlined and explained. Names are given to each of the clusters to give an indication of their nature; examples of cluster summaries are given to show how the clusters can be interpreted with the use of explanatory material.

The classification has a range of potential applications, as well as being a data set in its own right showing socioresidential patterns. When mapped, it can be used to solve a variety of research questions. Geodemographic classifications are used in the planning stages of a project, usually at a relatively broad geographic level. They are also used in data profiling to code geographically sparse data to an area typology to enable further analysis or to simplify a complex data set. Consumer profiling by geographical area can help to establish how a product may sell in a certain area. Identification of product use across geographical areas can help to establish who is buying a certain product. Classifications are often used as a method of stratified sampling for opinion polls that are used to gauge the views of the nation most notably before a general election. Many academics regularly use classifications in their research; Rees *et al.* (1996) employed the ONS classification of districts to compare rates of migration across the UK. The classification is already being widely used by academics and both the private and the public sectors. As it is free it provides a bench-mark that everyone can access and refer to.

The classification is housed on the National Statistics Web site: http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/oa/default.asp. This includes cluster membership and descriptive information, plus a detailed report on the creation of the classification. The classification can also be ordered on a compact disc by contacting info@statistics.gov.uk. The classification is now also available from additional sources with value-added information for the academic community. The classification is available to be added to any census table via CASWEB: <http://census.ac.uk/casweb/>. The classification can also be obtained with digital boundary data via the UK Borders

service: <http://edina.ac.uk/ukborders/>. Additional data sets and further information about the classification can be obtained by contacting the authors.

Acknowledgements

The funding provided for this project was by an Economic and Social Research Council and Co-operative Awards in Science and Engineering partner, the Office for National Statistics (Co-operative Awards in Science and Engineering award PTA-033-2002-00067) and an Economic and Social Research Council Postdoctoral Fellowship (award PTA-163-27-1006).

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. *Popln Trends*, **99**, 37–52.
- Brown, M. (2005) Second homes in Cornwall. *Personal Communication*.
- Callingham, M. (2003) Current commercial sector use of geodemographics and the implications for the ONS area classification systems. *Personal Communication*.
- Everitt, B. S., Landau, S. and Leese, M. (2001) *Cluster Analysis*, 4th edn. London: Arnold.
- Harrigan, K. R. (1985) An application of clustering for strategic group analysis. *Strat. Mangmnt J.*, **6**, 55–73.
- Harris, R., Sleight, P. and Webber, R. (2005) *Geodemographics, GIS and Neighbourhood Targeting*. Chichester: Wiley.
- Lorr, M. (1983) *Cluster Analysis for the Social Sciences*. San Francisco: Jossey-Bass.
- Martin, D. (2002a) Geography for the 2001 Census in England and Wales. *Popln Trends*, **108**, 7–15.
- Martin, D. (2002b) Output Areas for 2001. In *The Census Data System* (eds P. Rees, D. Martin and P. Williamson), pp. 37–46. Chichester: Wiley.
- Martin, D., Nolan, A. and Tranmer, M. (2001) The application of zone-design methodology in the 2001 UK Census. *Environ. Plannng A*, **33**, 1949–1962.
- Milligan, G. W. (1996) Clustering validation: results and implications for applied analyses. In *Clustering and Classification* (eds P. Arabie, L. J. Hubert and G. De Soete). Singapore: World Scientific Press.
- Milligan, G. W. and Cooper, M. C. (1988) A study of standardisation of variables in cluster analysis. *J. Classificn*, **5**, 181–204.
- Office for National Statistics (2003a) Edit and imputation—evaluation report. Office for National Statistics, London. (Available from <http://www.statistics.gov.uk/census2001/editimputevrep.asp>.)
- Office for National Statistics (2003b) *Census 2001: Key Statistics for Local Authorities in England and Wales*. London: Stationery Office.
- Office for National Statistics (2005a) Area classification for output areas. Office for National Statistics, London. (Available from http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/oa/default.asp.)
- Office for National Statistics (2005b) Beginners' guide to UK geography: census geography. Office for National Statistics, London. (Available from http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/.)
- Office for National Statistics (2005c) Rural and urban classification 2004. Office for National Statistics, London. (Available from <http://www.statistics.gov.uk/geography/nrudp.asp>.)
- Peach, C. (1996) Does Britain have ghettos? *Trans. Inst. Br. Geogr.*, **21**, 216–235.
- Rees, P., Denham, C., Charlton, J., Openshaw, S., Blake, M. and See, L. (2002a) ONS classifications and GB Profiles: census typologies for researchers. In *The Census Data System* (eds P. Rees, D. Martin and P. Williamson), pp. 149–170. Chichester: Wiley.
- Rees, P., Durham, H. and Kupiszewski, M. (1996) Internal migration and regional population dynamics in Europe: United Kingdom Case Study. *Working Paper 96/20*. School of Geography, University of Leeds, Leeds.
- Rees, P., Martin, D. and Williamson, P. (2002b) Census data resources in the United Kingdom. In *The Census Data System* (eds P. Rees, D. Martin and P. Williamson), pp. 28–36. Chichester: Wiley.
- Simpson, L. (2007) Ghettos of the mind: the empirical behaviour of indices of segregation and diversity. *J. R. Statist. Soc. A*, **170**, 405–424.
- Sleight, P. (2004) *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*. Henley-on-Thames: World Advertising Research Centre.
- Stillwell, J. and Duke-Williams, O. (2007) Understanding the 2001 UK census migration and commuting data: the effect of small cell adjustment and problems of comparison with 1991. *J. R. Statist. Soc. A*, **170**, 425–445.

- Vickers, D. (2003) The difficulty of linking two differently aggregated spatial data sets: using a look-up table to link postal sectors and 1991 Census enumeration districts. *Working Paper 03/2*. School of Geography, University of Leeds, Leeds. (Available from <http://www.geog.leeds.ac.uk/wpapers/03-2.pdf>.)
- Vickers, D. (2006) Multi-level integrated classifications based on the 2001 Census. *PhD Thesis*. University of Leeds, Leeds. Unpublished.
- Vickers, D. and Rees, P. (2006) Introducing the National Classification of Census Output Areas. *Popln Trends*, **125**, 15–29.
- Vickers, D., Rees, P. and Birkin, M. (2005) Creating the National Classification of Census Output Areas: data, methods and results. *Working Paper 05/2*. School of Geography, University of Leeds, Leeds. (Available from <http://www.geog.leeds.ac.uk/wpapers/05-2.pdf>.)
- Voas, D. and Williamson, P. (2001) The diversity of diversity: a critique of geodemographic classification. *Area*, **33**, 63–76.
- Wallace, M. and Denham, C. (1996) *The ONS Classification of Local and Health Authorities of Great Britain*. London: Stationery Office.
- Webber, R. and Craig, J. (1978) *Socio-economic Classifications of Local Authority Areas*. London: Office of Population Censuses and Surveys.
- Weiss, M. J. (2000) *The Clustered World*. New York: Little Brown.