Introduction

The purpose of this paper is to explain the choice of variables for analysis in the 2001 Output Area level Area Classification. The selection process has followed that used for the Local Authority and Ward level classification and the initial chosen variables are similar with some minor changes due to the change in scale. The underlying objective in variable choice is to select the minimum number of variables that will adequately represent the main dimensions in the census. For presentation purposes these have been defined as demographic structure; household composition; housing; socio-economic character; employment and industry sector. The data are from the Key Statistics tables produced from the Census.

The steps involved in the selection of variables are summarised below:

- Variables from the Key Statistics tables were considered for use.
- Variables were merged to create composite variables; for example, the variable 'South Asian' represents people identifying as Indian, Pakistani or Bangladeshi.
- Strongly correlated variables were removed
- Variables with badly behaved distributions (e.g. a high proportion of zeros) were not included.

In all cases, the decision to include or exclude a variable also involved using the team's own judgement. Continuity with previous classifications was considered when deciding whether to include or exclude a variable. It was felt by the Project Board that it would aid the understanding of the user if the set of variables was the same as that chosen for the Ward level classification (allowing for some differences that were unavoidable due to the change in scale).

Initial set of variables considered

The Key Statistics have already been identified as being the most important variables so the initial data set included all variables from the Output Area Level Key Statistics Tables. The Key Statistics represent the most important variables within the published data from the census, and have a comparatively simple data structure that will aid data extraction.

Reducing the initial set of variables

As the aim was to represent the main dimensions in the census data with the minimum number of variables the initial data set was reduced. If a variable didn't add anything to the classification it was removed. For example, the variables "age group 65 - 74" and "age group 75+" showed very similar same population characteristics; "age 65+" was selected to represent those above the retirement age. In some cases a composite variable was used to reduce variables, for example Indian, Pakistani and Bangladeshi composite ethnicity has been used to represent respondents identifying from any of the individual countries. Thirdly variables that only identified very small sectors of the population were removed. It was not possible to include a migration indicator as these data were not yet available for England and Wales when the classification was carried out. Religion was also omitted because the question was optional and it experienced variable levels of response in different areas of the UK.

Further reducing the variable set

A further reduction was made, as it is likely that some of these variables represent the same population characteristic and will therefore not provide extra information. For example, the proportion of people at pensionable age might account for a high proportion of single pensioner households and a high rate of rooms per person. It was therefore necessary to further reduce the dataset by identifying and removing strongly correlated variables. The Pearson correlation coefficient was used to identify those pairs of variables where it was likely that a characteristic would be given too much weight if both were to be included in the classification. If a pair of variables had a correlation coefficient that was above 0.85, then one of the pair was considered for exclusion. Pairs of correlated variables were not just considered in isolation but alongside other pairs of correlated variables. Judgement on which variables to exclude were made following group discussion and in conjunction with the team who produced the Ward level classification. The distribution of each variable were also examined and any variables with problematic distributions (e.g. a high proportion of zeros were not included).

A note concerning the included and excluded variables

We consider that the chosen variables represent the main dimensions of the census. Since an Urban/Rural indicator was not available, population density was used as a proxy Urban/Rural indicator. A distinction was made between different age groups in the population. The variable age15-24 was not included as this is highly correlated with the variable "students" and "students" were thought to be a more distinct group than the whole of the 15-24 age group. The over 75s are highly correlated with pensioners so this variable is excluded. It was decided not to include the proportion of people who live in a communal establishment 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, which is something to be avoided. 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, university residences etc. These are very different types of people who should not be grouped together.

The Indian, Pakistani and Bangladeshi and Black variables are included to represent ethnicity. The Chinese were not included as there are comparatively few cases and they are fairly evenly distributed across the country. They would therefore add little to the grouping procedure. Consideration was given to including the proportion of people not born in the European Union and the proportion of people born in the European Union excluding the UK. As membership to the European Union changed in 2004 (between the 2001 census and the creation of the classification), it might have caused confusion if these variables had been used. Therefore it was considered to be better to use the proportion of people not born in the UK instead. People not born in the UK also adequately represented the other two variables with only a small loss of distinction.

There are four household composition variables that were identified useful groups in the population. Renters from both the private and public sector are included plus terraced and detached houses. Purpose built flats and converted flats were merged with purpose built flats to create an “all flats” variable because they were correlated. It was also considered that merged they would be more reliable as the flats variable has a bimodal distribution with many areas having either a very high value or an absence of flats. Semi-detached is included by proxy as it is most of the rest of the housing not covered by the other categories. It also does not represent such a distinct group as terraced or detached.

The average number of rooms (household size) was used as a measure of the size of each household. The average number of people per room is used to measure overcrowding.A "no central heating" variable and "people with routine/semi-routine occupations" are included to measure social exclusion.

Long-term unemployed could not be used as in general the numbers are low across the country and are particularly unreliable due to disclosure controls. For example there were several occurrences of Output Areas where one person was unemployed, 3 of who are long-term unemployed. "People without qualifications" were highly correlated with those "working in a routine occupations" so was not kept. "People with Higher Education (HE) qualifications" and "households with two or more cars" are included to represent the section of the population that is more prosperous. "People in professional or managerial occupations" were highly correlated with "people with HE qualifications" so this variable was not kept. "Average number of cars" and "two or more cars" are highly correlated. We decided to keep the variable "two or more cars" as a measure of affluence. "Households with no cars" are highly negatively correlated with "households that have two or more cars" so it was not necessary to also include this variable.

The working from home variable may identify interesting sub-groups, as would those using public transport. People with limiting long term illness represent poor health, this variable is to be standardised by age to adjust for the effect that the age distribution has on this variable. The percentage of residents who provide unpaid care identifies an important group of the population. The separate unemployment rates for men and women are highly correlated with total unemployment. There is also a high correlation between the unemployment rate for men and the unemployment rate for women so total unemployment was used instead of the two separate unemployment rates. It was also decided not to split variables by gender as it would halve the already small population within the Output Areas (especially in Scotland where they are smaller) increasing the chance of getting an unreliable value. A composite (men and women) variable that work part-time was used as the decision was made not to split by gender at the Output Area level. "People who look after the home" was used instead of women as no variables were split by gender at the Output Area level. Students and the percentage of the population that work in some industrial subgroups are included.

Standardisation of LLTI (SIR)

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 Output Areas than for higher level geographies, because of the relatively small size 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 it's age structure as the presence of a large amount of ill health.

It was therefore decided 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 Standardised Illness Ratio (NOT Rate) 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 i is defined as follows:

SIRi = 100 (Ii/a ran Pai) (1)

where

Ii = observed count of ill people in area i
ran = rate of illness for age group a in the national population
Pai = 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 it local age distribution. There is substantial variation between the OAs with values ranging from 0 to 505 the healthiest areas are Output Areas with SIRs below 70 and the least healthy are OAs with a SIRs exceeding 130

List of final set of variables

The variables are listed under the six domains

Table 1: Full list of 41 variables

Variable Number	Short Name
Demographic	
Age	
percentage of resident population aged 0-4	*Age 0-4*
percentage of resident population aged 5-14	*Age 5-14*
percentage of resident population aged 25-44	*Age 25-44*
percentage of resident population aged 45-64	*Age 45-64*
percentage of resident population aged over 65	*Age 65+*
Ethnicity	
percentage of people identifying as Indian, Pakistani or Bangladeshi	*Indian, Pakistani or Bangladeshi*
percentage of people identifying as Black African, Black Caribbean or Other Black (1)	*Black African, Black Caribbean or Other Black*
Country of Birth	
percentage of people not born in the UK	*Born Outside UK*
Population	
Population Density (number of people per hectare)	*Population Density*
Household Composition	
Living Arrangements	
percent of residents over 16 who are not living in a couple and are separated or divorced (2)	*Separated/Divorced*
Size/Family	
percentage of households with one person who is not a pensioner	*Single person household (not pensioner)*
percentage of households which are single pensioner households	*Single pensioner household*

percentage of households which are lone parent households with dependent children	*Lone Parent household*
percentage of households which are cohabiting or married couple households with no children	*Two adults no children*
percentage of households comprising one family and no others with non-dependent children living with their parents	*Households with non-dependant children*

Housing

Tenure

percent of households that are public sector rented accommodation	*Rent (Public)*
percent of households that are private/other rented accommodation	*Rent (Private)*

Type and size

percent of all household spaces which are terraced	*Terraced Housing*
percent of all household spaces which are detached	*Detached Housing*
The percentage of households which are Flats	*All Flats*

Quality/crowding

percent of occupied household spaces without central heating	*No central heating*
average household size (rooms per household)	*Rooms Per Household*
The average number of People per room	*People per room*

Socio-Economic

Education

percent of people aged between 16 - 74 with a higher education qualification	*HE Qualification*

Socio-economic class

percent of people aged 16-74 in employment working in routine or semi-routine occupations	*Routine/Semi-Routine Occupation*

Ownership/commuting

percent of households with 2 or more cars	*2+ Car household*
percent of people aged 16-74 in employment usually travel to work by public transport (3)	*Public Transport to work*
percent of people aged 16-74 in employment who work mainly from home	*Work from home*

Health and Care

percentage of people who reported suffering from a Limiting Long Term Illness (Standardised Illness Ratio, standardised by age) (7)	*LLTI (SIR)*
percent of people who provide unpaid care (6)	*Provide unpaid care*

Employment

percent of people aged 16-74 who are students (4)	*Students (full-time)*
percent of economically active people aged 16-74 who are unemployed	*Unemployed*
percentage of economically active people aged 16-74 who work part time (5)	*Working part-time*
percentage of economically inactive people aged 16-74 who are looking after the home	*Economically inactive looking after family*

Industry Sector

percent of all people aged 16-74 in employment working in agriculture and fishing	*Agriculture/Fishing*

percent of all people aged 16-74 in employment working in mining, quarrying and construction	*Mining/Quarrying/Construction*
percent of all people aged 16-74 in employment working in manufacturing	*Manufacturing*
percent of all people aged 16-74 in employment working in hotel and catering	*Hotel & Catering*
percent of all people aged 16-74 in employment working in health and social work	*Health and Social work*
percent of all people aged 16-74 in employment working in financial intermediation	*Finance*
percent of all people aged 16-74 in employment working in wholesale/retail trade	*Wholesale/retail trade*

Footnotes

(1) includes Scottish Black for Scottish Unitary Authorities
(2) from KS03
(3) for Scottish Unitary Authorities this is percentage of residents usually travel to work or place of study by public transport
(4) This includes economically active full time students and economically inactive students
(5) Part-time is defined as working less than 30 hours a week
(6) Provides at least one hour a week of unpaid care
(7) working age is 16-64 for men and 16-59 for women