

Methods for National Statistics 2001 area classification for local authorities

Introduction

This document outlines the methodology used in the 2001 area classification at Local Authority District (LAD) level. This includes the process undertaken to select a variable set which represents population characteristics captured in the Census and the clustering techniques used to create the classification.

Before the analysis was carried out, the Isles of Scilly was combined with Penwith, and City of London was combined with City of Westminster. This was because of the small populations found in these two local authorities.

Variable selection

The procedures adopted for the selection of variables were conducted via a series of team meetings using a rigorous and logical approach, designed to gain an efficient representation of the census data. For any method of classification, the results will depend on the variables used for the analysis. The underlying objective in variable choice was thus to select the minimum number of variables that would adequately represent the main dimensions in the census data. These have been defined as demographic structure; household composition; housing; socioeconomic character, employment and industry sector.

Selection procedures

Variables were selected from the 2001 census key statistic tables, as they contain important census data and they are accessible to the public. The starting point was to consider all possible variables that were available at the appropriate time, then to reduce the data set by a process of elimination. Where possible all variables that were included in the 1991 classification were considered for the new classification as well as new variables that were available for the first time.

The initial data set was reduced by three methods. Firstly, if a variable didn't add anything to the classification or was considered unreliable it was removed. For example, answering the religion question in the 2001 census was optional and may contain missing values. Secondly, in some cases a composite variable was used to represent similar variables, for example the variable 'Flats' was used to represent respondents who live in either purpose built or converted flats. Thirdly, variables that only identified very small sectors of the population were removed. It should also be noted that it has not been possible to include some variables which were not available at the time of producing the classification.

A further reduction was made based on the matrix of correlations between variables. If the Pearson Correlation Coefficient between two variables was greater than 0.85, one was removed. It is likely that highly correlated variables represent the same population characteristic, inclusion of both would result in overrepresentation of certain population characteristics. The final data set contained forty-two variables.

Clustering methods

The cluster analysis method places each area in a group with the other areas to which it is most similar in terms of the forty-two census variables selected. This enables similar areas to be classified according to their particular combination of characteristics. The classification consists of two parts: a hierarchical classification of supergroups, groups and subgroups, and an overlapping classification of 'corresponding areas'. This second part lists the authorities most similar to each authority.

The data were first standardised in order to ensure that the scale of the variables were comparable. In order to create the hierarchical classification, the Ward's clustering method was adopted followed by the *k*-means method, which optimises the solution attained. Ward's clustering method uses the Squared Euclidean Distance as a similarity measure. This was also used to attain the classification of corresponding authorities. Ward's method is well established and has given valid and reliable results in previous ONS classifications. It finds groups which are as homogenous as possible at each level and uses every variable in each LAD.

Standardising data

The data were standardised using an *Inter-decile Range* method. This compares each local authority's value, X_i , for each variable to the UK median, X_{med} , and is then divided by the difference between the 90th percentile, X_{90th} , and the 10th percentile, X_{10th} :

$$\frac{X_i - Y_{med}}{X_{90th} - X_{10th}}$$

This was considered more appropriate for the distribution of the data than the more frequently used 'z-scores' standardisation and is more robust to outliers than the range standardisation that was used in the 1991 classification.

Distance measure

The Squared Euclidean distance was used to measure similarity between clusters. Two local authorities X and Y, are said to be similar if the 'distance' between them, based on census characteristics is small. It uses the following formula:

$$\sum_i (X_i - Y_i)^2 \quad \text{where } X_i - \text{value of variable } i \text{ for LAD X and } Y_i - \text{value of variable } i \text{ for LAD Y}$$

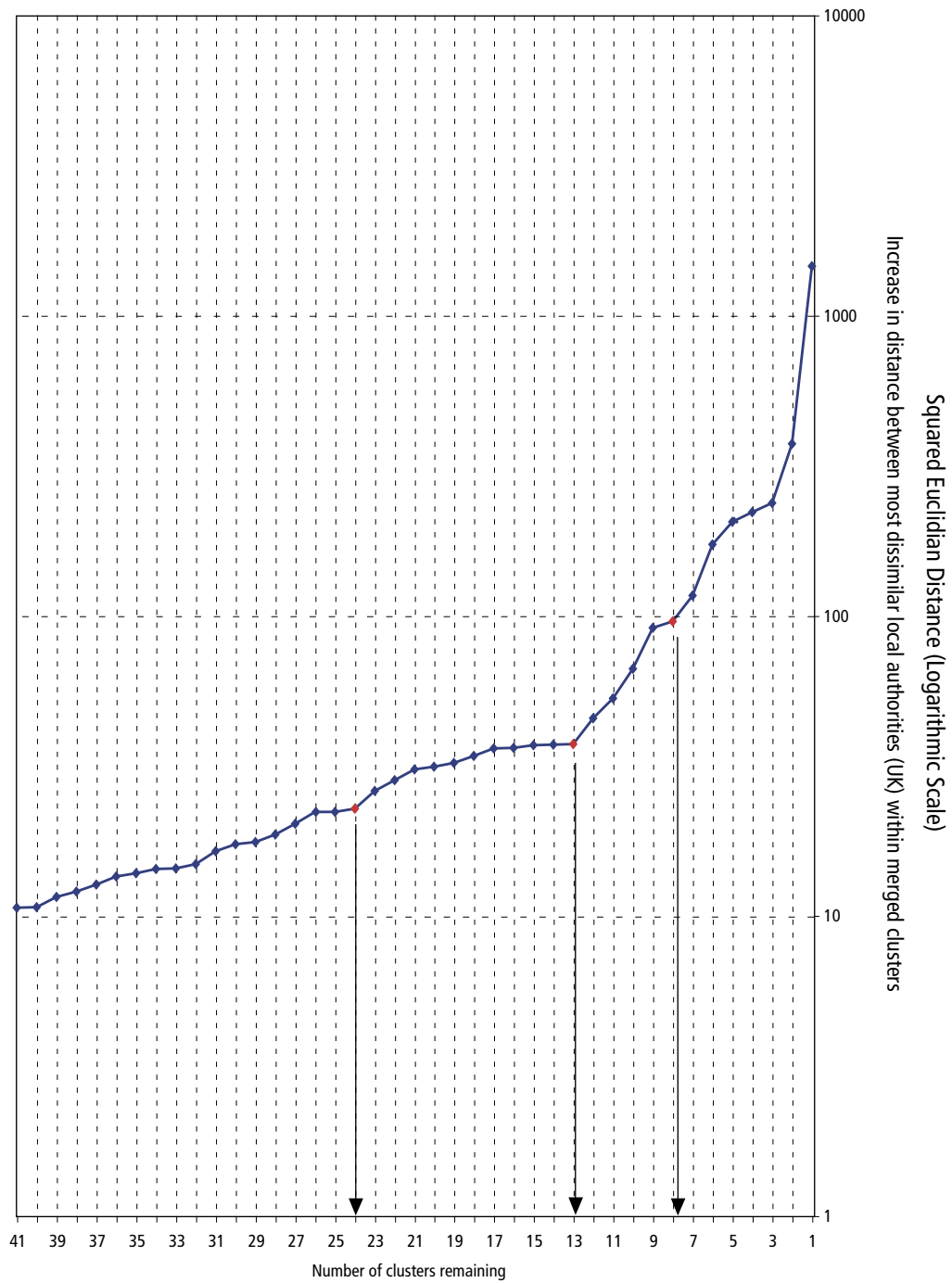
so that the distance between the two local authorities is the sum of the squared differences between their values for each and every variable (i=1 to 42).

Ward's clustering method

The method initially treats all 432 local authorities as separate clusters, and then combines clusters by maximising the within-cluster homogeneity. To measure homogeneity a within-cluster sum of squares is used. At each step, all possible solutions are considered and clusters are formed using the solution that gives the smallest within-cluster sums of squares. The within-cluster sum of squares that is minimised is also known as the 'Error Sum of Squares' (ESS). For each case the squared Euclidean distance to the cluster means is calculated. As each case initially starts off as a cluster, the ESS is zero. The next step would then form 431 clusters, one cluster of size two and the others all of size one. The ESS is calculated for all possible solutions and the cluster solution that produces the least ESS is chosen and the process is repeated. This continues until there is just one cluster containing all local authorities. An agglomeration schedule is produced (see figure) which shows the difference in ESS at each stage. This was used when determining the cut-off points of 8, 13 and 24.

Once the clusters have been formed, a check must be carried out to ensure that each local authority is assigned to its correct cluster. Due to the agglomerative nature of the technique the cluster centroid will have changed at each step, as new districts are added. This might mean that by the end of the process some districts are more similar to districts in other clusters than they are to districts in their own cluster. To ensure that all authorities are in the right cluster, a *k*-means analysis was carried out. This technique reassigns districts to the cluster with the smallest distance between the district and the cluster centroid. This was carried out at the 24 cluster level and the higher levels of the classification were then created by reassigning LADs using the Ward's solution, starting from the centroids of the 24 *k*-means clusters.

Agglomeration schedule for local authority classification



Optimal levels can be identified on the basis of where there is a natural levelling off in the slope of the line.