

Introduction

The aim of this paper is to discuss the methodology used for the 2001 Output Area level classification. The classification places each output area in a group with those other output areas that are most similar in terms of census variables. This enables similar areas to be classified according to their particular combination of characteristics.

Choice of variables

The analysis was carried out using the Key Statistics tables produced from the census. The variables are socio-economic and demographic. The chosen variables cover the main dimensions of the census and for presentation purposes these have been defined as demographic structure; household composition; housing; socio-economic character; employment and industry sector. Strongly correlated variables were removed to avoid the duplication of particular factors. This allowed the minimum number of variables to be included so that the six main census dimensions were represented using the available data. For further details please see the variable selection paper.

Transformation of the data

Before any analysis was carried out the data was transformed to a log scale. The decision to do this was taken because of the effect of a large number of outliers at the high end of the value scale.

Log transformation

By transforming the data to a log (logarithmic) scale the problem of very high value outliers was greatly reduced as the difference between values at the extremities of the data set are reduced by more than those more typical average values.

A log is the exponent of the power to which a base number must be raised to equal a given number. For example the log of 100 is 2 to the bases 10. This can be seen as $10 * 10 = 100$ which is the same as 10^2 , with 2 being the exponent previously referred to. Before the data was converted to a log scale, all the values had 1 added to them. This was because 0 (of which there are many in the data) has no logarithmic function and will produce an error and any value between 0 and 1 produces a negative value, which would have confused the dataset. By adding 1 to every data point this problem was resolved.

Standardisation of the data

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

Three methods of standardisation were considered:

Z-score standardisation

This is the most common form of standardisation. Z-score standardisation compares the value of the variable X_i to the mean X . This is then divided by the standard deviation. Z-

score standardisation works well when the data are normally distributed but this may not always be the case.

$$Z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

Range standardisation

This method of standardisation was implemented in the 1991 classification; see Wallace and Denham (1996). Range standardisation compares each value of a variable, X_i , to the minimum value X_{\min} . This is then divided by the distance between the minimum value X_{\min} and the maximum value X_{\max} of the variable. This method does not work well if the data contain extreme outliers.

$$Z_i = \frac{X_i - X_{\min}}{x_{\max} - x_{\min}}$$

After the data has been range standardised each variable has a range of 1 with the maximum value being 1 and minimum value being 0.

Inter-decile range standardisation

This method is a slight variation of the range standardisation method that overcomes the problems associated with outliers. It compares each value of a variable, x_i , to the median, X_{med} , which is then divided by the distance between the 90th percentile, $X_{90^{\text{th}}}$, and the 10th percentile, $X_{10^{\text{th}}}$.

$$\frac{x_i - X_{\text{med}}}{X_{90^{\text{th}}} - X_{10^{\text{th}}}}$$

Choice of standardisation method

Initial experiments showed that highly skewed variables were given too much weight by the Z-score and inter-decile range standardisation. This problem was resolved when the data were standardised using the range standardisation method. The range standardisation method was used to standardise the output area level data.

Defining the clustering technique

Although there are various methods of cluster analysis available Ward's method and K-means are the most widely used techniques and were used in the Local Authority and Ward level classifications. We will look at both of these methods.

Wards method

Wards method is the most commonly used agglomerative clustering technique. It produces spherical clusters that are roughly the same size. The aim is to join objects together in clusters, using a measure of similarity or distance. This is a bottom up approach starting with n groups each containing one case. Two of the cases then combine to form a single cluster. At the next stage, either a third case is added to the cluster or two other cases are merged into a new cluster. This process continues until all cases belong to one cluster. Once a cluster is formed it cannot be split, it can only be combined with other clusters. In addition to choosing a similarity or distance measure to use when comparing two observations, there is also the choice of what should be compared when groups contain more than one observation. Wards method joins the two groups that minimise the Error Sum of Squares (within cluster sum of squares). Due to the agglomerative nature of Wards, the cluster centres change each time a new case is added. This might mean that by the end of the process some cases are no longer in the correct cluster. The solution given by Ward's method can be refined by k-means.

K-means refinement

This is a simple non-parametric clustering method that minimises the within cluster sum of squares whilst maximising the between cluster variability. The k-means method requires that the number of clusters are specified beforehand. It is an iterative relocation algorithm based the sum of squares. The algorithm repeatedly moves a case from one cluster to another to improve the sum of squares within each cluster. The case is assigned/removed from the cluster to which it brings the greatest improvement. When all cases have been processed then the algorithm moves to the next iteration. A stable solution is reached when there are no more moves in a complete iteration.

Classification of Output Areas

It soon became apparent that at the output area scale, Ward's method did not work as well as it did when working at the ward scale. When Ward's hierarchical clustering procedure was run on the 500 cluster centres as was done for wards, clusters were being produced which were several factors of scale in difference. The number of clusters that were produced ranged in size from 95,000 to 7 even at the top level of hierarchy where the target size was between five and ten groups. This dramatic difference in cluster sizes was caused by intricacies of the Ward's method. Ward's method works by grouping the nearest two output areas together and then repeating the process again at the next run but it treats the two output areas clustered on its first run as an unsplitable whole. This tends to increase the likelihood that unevenly sized groups are produced, even in a very large data set. An output area which is an outlier on several variables will be clustered last and left on a group on its own even though there maybe output areas clustered together which are further apart

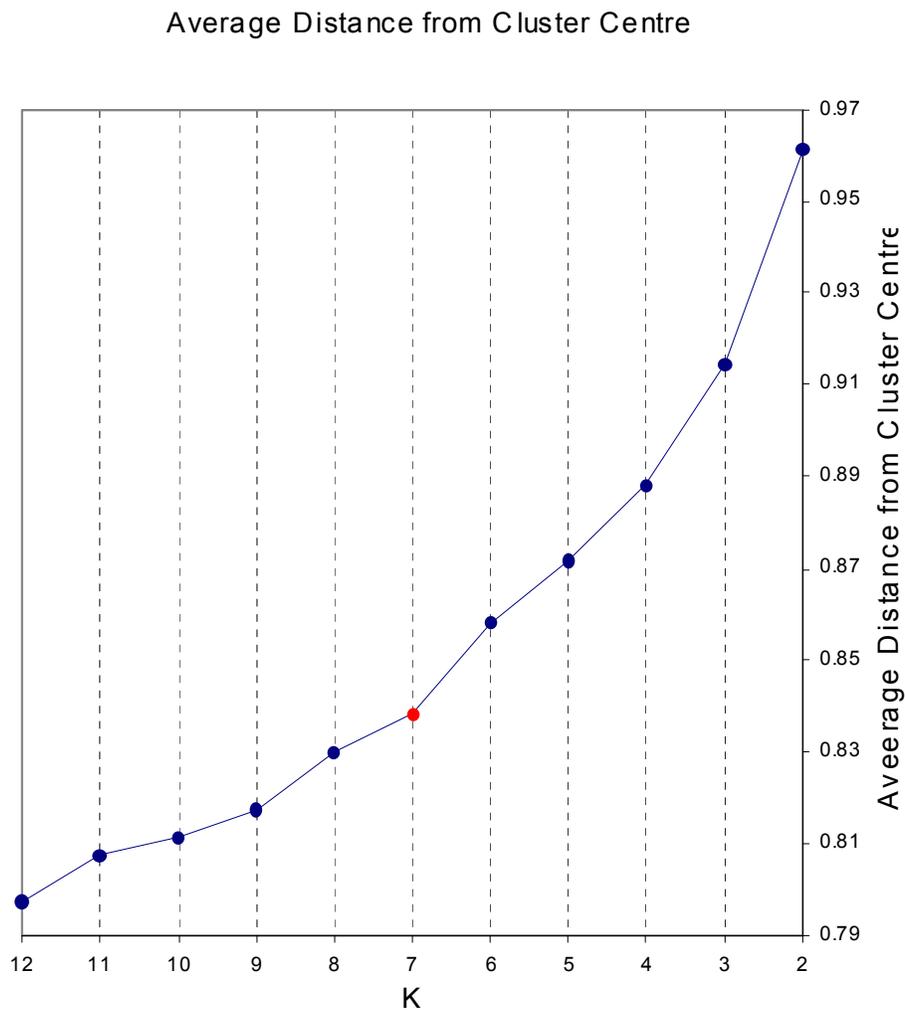
Creating a hierarchical classification using K-means

Since the experiments with Ward's algorithm were problematic an idea was formulated to adapt the k-means algorithm to produce a hierarchy. The K-means algorithm is run on the dataset and n clusters are produced, the original dataset is then split into n separate datasets (representing the highest level of the hierarchy). Each of the new datasets then has the K-means algorithm run on them separately to create the second level of the hierarchy. The second level of the hierarchy is then separated into m separate datasets and each one has the K-means algorithm run on them to create the lowest level of the hierarchy.

Using suggested numbers of cluster numbers, looking at how many output areas were in each cluster and how this effected the average distance from cluster centre, a hierarchy of 7, 21 & 52 was decided upon

It had been suggested that the most useful number of clusters In the first level would be around 6, taking this as a starting point clusters from 2 to 12 were examined to see how the average within cluster distance from centre changed. Fig. 1 shows how the average distance to cluster centre increases as the number of clusters is reduced. The target was a number of clusters around 6 this was then narrowed to an expectable range of 4 to 8, within this range it was evident that the largest increase in the average distance from cluster centre was when the number of clusters was reduced from 7 to 6 and the least increase in average distance from cluster centre is when number of clusters was reduced from 8 to 7. This suggests that the 7 cluster solution is the most suitable as its average distance to cluster centre is small than would be expected when you compare it to solutions 6 & 8.

Fig 1: Average distance from cluster centre for different values of K, using k-means clustering



Once the first level of the classification had been decided upon as containing seven clusters this then needed to be broken down to create the second level of the hierarchy. This was

done in a similar way to the first level, by examining the average within cluster distance, however at this level only 2, 3 or 4 clusters were considered to ensure that the number of clusters reflected as closely as possible the target number of clusters of around 20. Also taken into consideration was the number of output areas in each cluster, with the intention of keeping the clusters as similar in size as possible. A second level of 21 clusters was created splitting cluster 1 into 1a and 1b, cluster 2 in 2a and 2b etc. The second level then needed to be split down again to create the third level of the hierarchy with a target size of around 50 clusters. To create the third level the clusters in the second level were split into either two or three clusters, again considering the within cluster difference and the number of output areas in each cluster. The third level of the hierarchy numbers 52 clusters by splitting cluster 1a into 1a1, 1a2 and 1a3, cluster 1b into 1b1 and 1b2 etc.