Modifying the 2001 Census Output Area Classification (OAC) for applications in Higher Education.

A. Singleton, P. Longley.

Department of Geography - University College London - Pearson Building - Gower Street - London - WC1E 6BT
Telephone: +44 (0)20 7679 0500
Fax: (+44 (0) 20 7679 0565
a.singleton@ucl.ac.uk p.longley@geog.ucl.ac.uk , http://www.spatial-literacy.org/

1. Introduction

There are a variety of commercial Geodemographic classifications which can be adapted to inform spatial decision making in Higher Education. However none have been specifically designed for this purpose. Although contemporary Geodemographic classifications originated in the public sector through studies of deprivation in the 1970's (See Webber, 1975, Webber, 1977), the application of these techniques though the 1980s and 1990s remained predominantly within the private sector for market segmentation across a range of industries. Only recently have commercial vendors augmented these classifications with public sector data, however this has occurred superficially at the level of "pen portraits" to provide additional descriptive material of clusters (Longley and Goodchild, 2007). Although these classifications may be promoted as tailored solutions for the public sector, they do not address a number of key concerns including whether it is appropriate for a general purpose classification describing private sector consumption of goods and services to be applicable for public goods that are consumed collectively. Furthermore, for public sector use of geodemographic classifications, robustness in terms of social equality is acutely important as the misspecification of areas could have far reaching negative impacts. For example, in an advertising campaign targeting educational funding opportunities, residents of in an incorrectly prioritised area may lose real life chances by not receiving appropriate information, despite being stakeholders in the educational and taxation systems. Because of these issues greater transparency of classification procedures is required by public sector end users, including a greater level of methodological detail than has hitherto been provided by commercial vendors. The requirement for an open methodology is not easily achievable by commercial companies as the release of detailed information on how classifications are constructed is perceived as undermining competitive advantage. To date, details on classification methodology have only been made available at rather general levels, including detail of which clustering algorithms have been used (Harris et al, 2005) or the sources and broad mix of input data (Experian, 2006). It is in this context that this paper explores how a bespoke geodemographic classification can be created by combining public domain and sector-specific data, using explicitly specified techniques and tools required in classification. It then presents a pilot which could be refined and deployed though a centralised service to Higher Education (HE).

2. Variability of K-Means Clustering

Numerous clustering algorithms are available to create groupings from large multidimensional datasets and a thorough review can be found in Everitt (1980). These methods all aim to create homogenous groups from a multidimensional data matrix. Before running a cluster analysis data must be standardised to reduce the effect of outliers and measure the data on the same scale, in order to ensure that all variables have the same weighting. Romesburg (1984) discusses how standardisation prior to clustering prevents the units used for attribute measurement from affecting the similarities between objects, and therefore allowing more equal contribution by each of the variables. There are many ways in which the data could be standardised and these are evaluated by Vickers (2005), who found that a range standardisation method (Wallace et al, 1996) performs most effectively at reducing outlier effects in Output Area (OA) level classification. The Output Area Classification (OAC) classification is built using a *k-means* algorithm which partitions a multidimensional data matrix into *k* clusters or groups based on local a optimisation criteria. For a full evaluation of why this algorithm is most suited to clustering applications at an Output Area geography see Vickers (2005). After the first iteration of the model where initial cluster centroids (seeds) are randomly placed in the data matrix and all data points are temporarily assigned to their nearest seed, the *k-means* algorithm attempts to find a local optimum through an objective function that reallocates data points iteratively from their initial assignments. Each data point is considered for reallocation to other clusters, and after each test the model objective function is recalculated. If the outcome is larger, i.e. a less homogenous cluster, no further reallocation of data points takes place. Where reassignment of data points does occur, the cluster centroid values for the gaining and losing clusters are recalculated. The maximum number of iterations for this optimisation process can be specified by the user. However, with current computational power it is possible to leave the models running until the iteration process converges, i.e. further reassignments of data points does not improve the sum of squares statistic. Everitt (1974:26) astutely observes that “there is no way of knowing whether or not the maximum of the criterion has been reached”. This is because in a single *k-means* model there are multiple local optima, since the random placement of the initial cluster seed centroid means that there are multiple possible locally optimised models. This can be illustrated when two separate models are run to convergence where $k = 9$ on a two variable dataset extracted from the OAC input data (see Figure 1 and Figure 2).

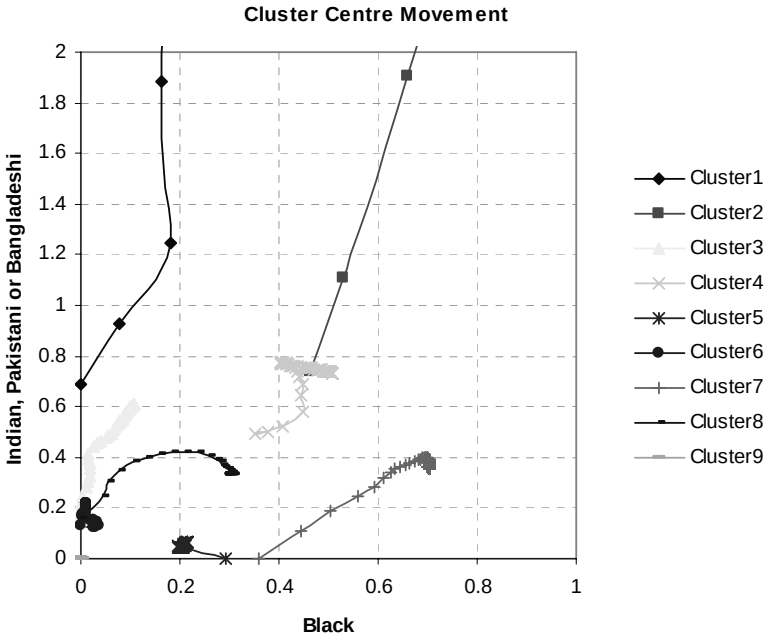

Figure 1: Cluster mean paths

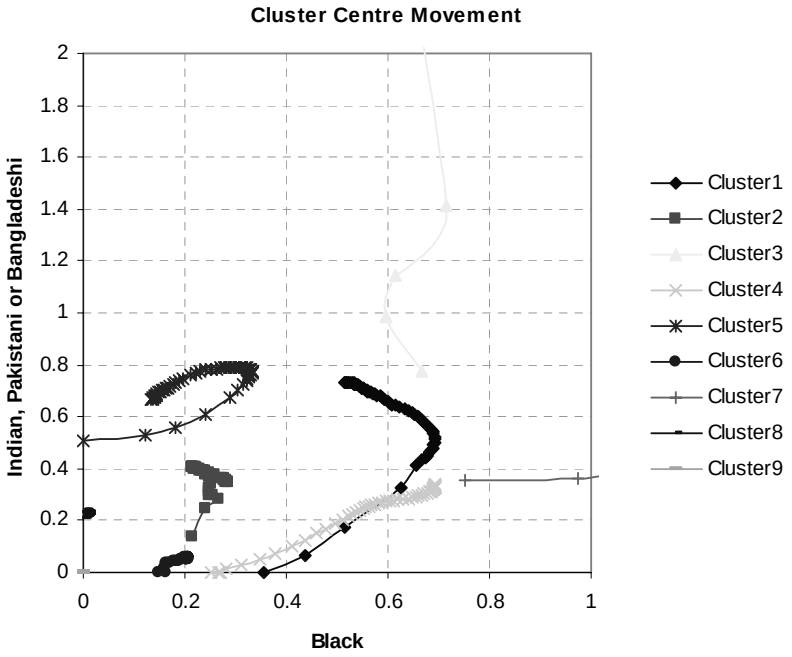

Figure 2: Cluster mean paths

Ignoring the arbitrarily assigned cluster names, these graphs show how the path of the cluster centroid can converge upon entirely different locations, depending upon the random initial seed location. Furthermore, as each iteration of the model reallocates data points to cluster centroids, "making the 'best' decision at each particular step does not necessarily lead to an optimal solution overall" (Harris *et al*, 2005:162). The most effective partitioning of the input data in a cluster model is globally optimised, although in reality this is not obtainable as there is no benchmark of global model performance for an individual data set. However, with sufficient computational power a globally optimised local model can be obtained by running *k-means* multiple times to convergence, comparing the results from each cluster analysis and saving the best performing classification. Figure 3 shows the results from a k=9 model which was run with a random seed allocation 150 times, and for each model an R-squared statistic was generated to estimate the quality of the model discrimination. This graph highlights the variance in overall model performance through selection of different initial seeds. Therefore, for an HE specific geodemographic classification it is essential that the model be run to convergence multiple times, with the explanatory power of each model compared.

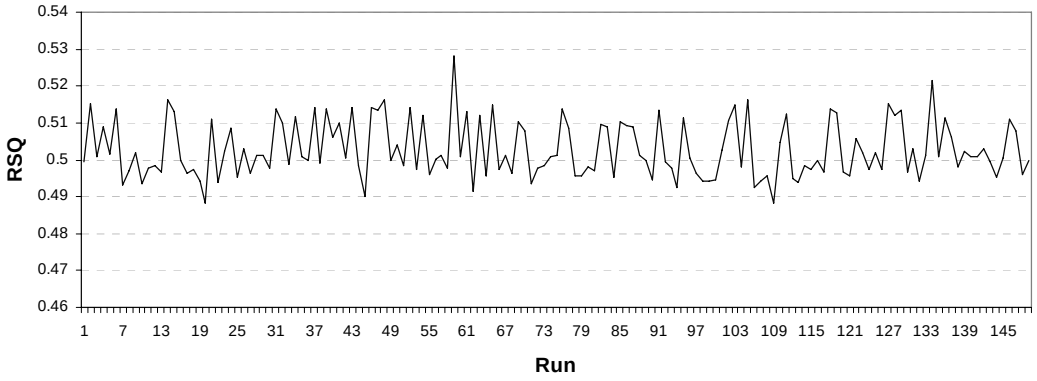

Figure 3: R-Squared Results from 9 cluster model

3. Creating an Educational OAC

An HE specific geodemographic classification could feasibly serve a number of different purposes including marketing, extending access, widening participation or subject specific targeting. When selecting input variables these purposes were considered with the aim of creating a classification useful for a variety of tasks demanded by HE decision makers. The data available for this study are from a subsection of the 2001 HESA database covering all students with English domicile studying within English institutions. This database contains of a variety of suitable variables for inclusion in the cluster analysis and the variables chosen are listed in Table 1.

Variable	**Numerator**	Denominator
Young participation rates	First year students aged 18-19.	Census 2001 18-19
Average distance from student's home to institution	N/A	N/A
Average A-Level Score of students	N/A	N/A
Proportion of students from low social class groups	Undergraduate degree students from the three lowest social classes (IIIM, IV, V)	All undergraduate degree students (Source: 2001 Census)
Proportion of students studying within degree course groupings	Those studying undergraduate degree courses within course groupings.	All undergraduate degree students (Source: 2001 Census)
Proportion of different ethnic minority Groupings compared to total HE population	Those undergraduate students from ethnicity minority groupings.	All undergraduate degree students (Source: 2001 Census)
Proportion of students previously educated in Independent Schools in Year	Those undergraduate students who previously attended independent schools.	All undergraduate degree students (Source: 2001 Census)

Table1: HE input variables for the cluster analysis

To investigate the most appropriate value of *k* in line with the observations from Figure 3, 10,000 separate cluster analysis were run from *k*=55 to *k*=65. The median, minimum and maximum R-Squared results are presented for each *k* value in Figure 4.

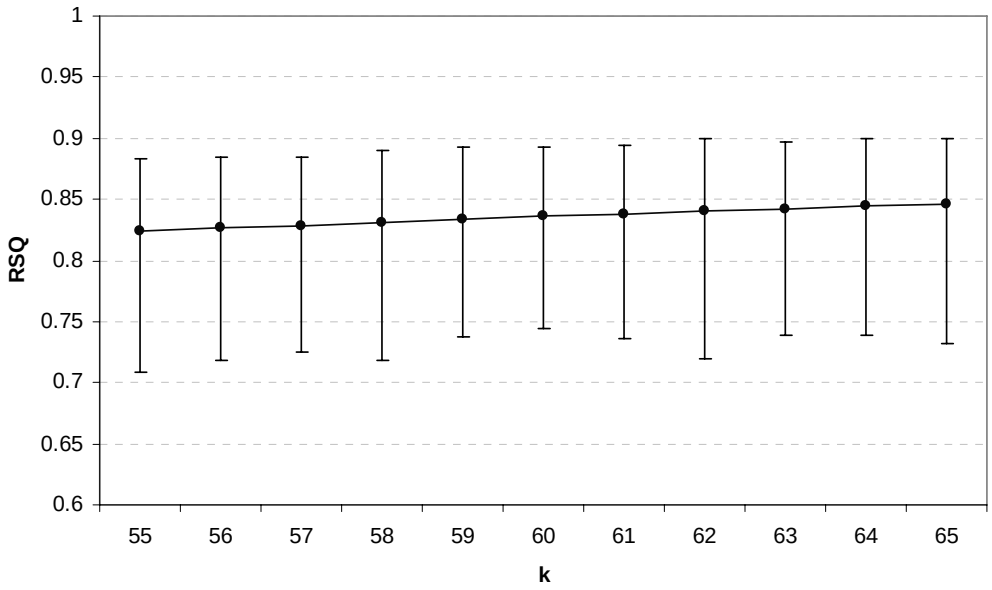

Figure 4: R-Squared Results from 10,000 Cluster Analysis

Each of these assignments of *k* appears to perform well at discriminating within the input data matrix and, as demonstrated in the earlier exploratory analysis, the minimum and maximum bars further illustrate the need to optimise each *k* allocation. Once the population distribution within each of these models were examined it was decided that *k*=64 was the most appropriate model as this had reduced outlier clusters. The Educational OAC Type level classification was therefore defined as comprising of 64 clusters, however, to create a more intelligible classification for end users a second hierarchy in classification was created which aggregates the finest level into larger groups. A second type of clustering algorithm was used to aggregate the 64 clusters (types) into the larger aggregation of groups. The Ward (1963) method assesses the loss of variance that would be associated by merging clusters together when those which are amalgamated minimise the "increase in information loss" (Everitt, 1993: 65). The result of which was an 8 group classification. Figure 5 demonstrates the variability in propensity to study particular subjects within one of these groups. Using an index score the average propensity is represented by the thin line at 100. Where Group D shows higher than average participation for a particular course, these scores rise above the line, and where participation is less than average these scores fall below the line. It is common practice in the geodemographic industry to consider an index score which is less than 80 and greater than 120 to be significant and these are represented by the two thicker lines. Therefore we could say that areas categorised as Group D shows an increased propensity to supply Veterinary Science and Language students.

This paper has demonstrated how a standard geodemographic classification can be adapted to serve a public sector market and an optimisation technique for the *k-means* algorithm has shown how great caution is required when building or interpreting classifications built using this method.

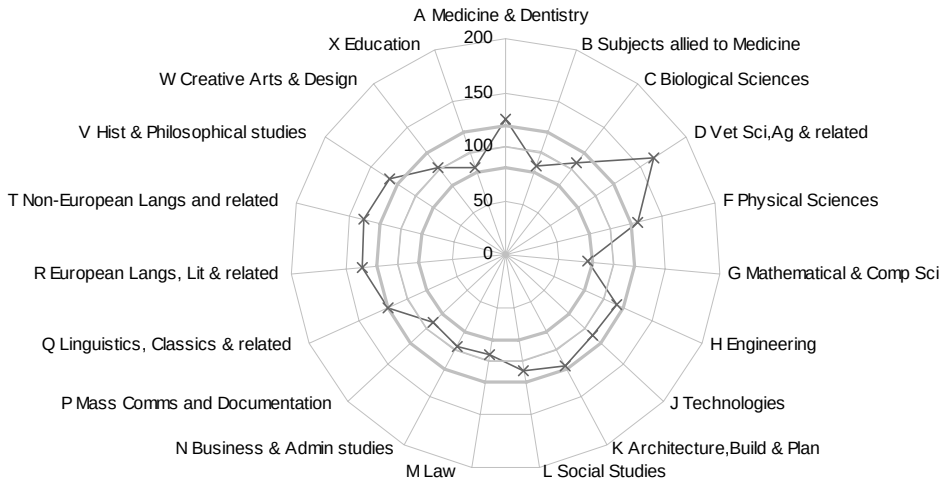

Figure 5: Group D – Course Propensity

4. References

Everitt, B. (1974) *Cluster Analysis*. London: Heinemann Educational Books.

Experian. (2006) *A Guide to Public Sector Mosaic*. London: Experian.

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

P.A. Longley, M.F. Goodchild (2007) The use of geodemographics to improve public service delivery. In J Hartley, C Skelcher, C Donaldson, G Boyne (eds) Managing Improvement in Public Service Delivery: Progress and Challenges. Cambridge, Cambridge University Press, in press.

Romesburg,H.C. (1984) *Cluster Analysis for Researchers*. Belmont, CA: Lifetime Learning Publications.

Vickers, D. (2005) *The National Classification of Census Output Areas*. Retrieved 21st August, 2006, from http://www.geog.leeds.ac.uk/people/d.vickers/OAclassinfo.html

Wallace, M., Charlton, J. and Denham, C. (1995) The new OPCS area classifications. *Population Trends*, 79, 15-30.

Ward, J.H. (1963) Hierarchical grouping to optimise an objective function. *Journal of the American Statistical Association*. 58, 236-234.

Webber, R. (1975) Liverpool Social Area Study, 1971 Data: PRAG Technical Paper 14. Centre for Environmental Studies. London.

Webber, R. (1977) Technical Paper 23: An Introduction to the National Classification of Wards and Parishes. Centre for Environmental Studies. London.