Transactions in GIS, 2010, 14(3): 283–297

Research Article

Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases

Muhammad Adnan
Department of Geography
University College London

Paul A Longley
Department of Geography
University College London

Alex D Singleton
Department of Geography
University College London

Chris Brunsdon
Department of Geography
University of Leicester

Abstract
Geodemographic classifications provide discrete indicators of the social, economic and demographic characteristics of people living within small geographic areas. They have hitherto been regarded as products, which are the final "best" outcome that can be achieved using available data and algorithms. However, reduction in computational cost, increased network bandwidths and increasingly accessible spatial data infrastructures have together created the potential for the creation of classifications in near real time within distributed online environments. Yet paramount to the creation of truly real time geodemographic classifications is the ability for software to process and efficiency cluster large multidimensional spatial databases within a timescale that is consistent with online user interaction. To this end, this article evaluates the computational efficiency of a number of clustering algorithms with a view to creating geodemographic classifications "on the fly" at a range of different geographic scales.

1 Introduction

Geodemographics are small area classifications that provide summary indicators of the social, economic and demographic characteristics of neighbourhoods. They continue a lineage of socio-spatial differentiation in human geography and urban sociology which extends over an 80 year period to formative work in urban ecology and social area

Address for correspondence: Muhammad Adnan, Department of Geography, University College London, Gower Street, London WC1E 6BT, UK. E-mail: m.adnan@ucl.ac.uk

doi: 10.1111/j.1467-9671.2010.01197.x

analysis (see Batey and Brown 1995, Harris et al. 2005 for reviews and interpretations). The term was first coined in the 1970s in work seeking to identify deprived inner city areas (Webber 1975, Webber and Craig 1978). The success of these public sector applications led to the commercialisation of geodemographics during the 1980s when the approach became widely used in the private sector as a method of target marketing (Birkin et al. 2002). More recently, public sector applications have moved somewhat to the fore (Longley 2005), with applications in health (Farr and Evans 2005, Shelton et al. 2006), policing (Ashby and Longley 2005), education (Singleton and Longley 2009) and local government (Longley and Singleton 2009).

The increasingly urbanized, complex and connected nature of human settlement is driving a demand, in developed countries at least, for better contextual information to inform decisions about the needs and preferences of people and the places in which they live and work. Decennial censuses of population have in the past been appropriate for this task, but the far-reaching and rapid changes that today characterise population change are making it increasingly necessary to supplement census sources with data that are more timely and relevant to particular applications. Better and more intelligent integration of a wider range of available data sources can open new horizons for depicting salient characteristics of populations and their behaviours. The art and science of creating geodemographic classifications has always been about much more than computational data reduction, and a key consideration in this quest is the availability of decision support tools to present areal data from a range of attributes in a format that is readily intelligible to the user. Thus, for example, in devising a local indicator of future risk of obesity, it might be appropriate to use data sources that variously measure demographic structure, school attainment, deprivation and existing health problems. In assembling such sources together, the analyst should be made aware of issues of data collection, normalisation, weighting and data reduction method.

Geodemographics typically classify neighbourhoods at two or more hierarchical levels: for example, the five Categories that make up the top level of the commercial ACORN classification (CACI Ltd, London) are divided into a total of 17 Groups at the secondary level, and these 17 Groups are divided into a total of 56 Types at the lowest, tertiary, level. The characteristics of each class within the typology are summarised by a label (e.g. "City Living"), a verbal "pen portrait" and other descriptive material such as photographs, montages and videos. These multimedia are used to give users of the classification a clearer understanding of the characteristics of the underlying population.

The data used to create geodemographic classifications may derive from any of a range of secondary data sources (Harris et al. 2005). Core to most classifications are census data and some classifications (such as the UK Output Area Classification: Vickers and Rees 2007) use no other source. Other data may be derived from behavioural or attitudinal surveys (e.g. "lifestyle" surveys from commercial sources), financial data (e.g. county court judgments, directorships) and property information (e.g. property tax bands). The process of creating geodemographic classification has been regarded as a computationally intensive activity, by which clustering algorithms are used to partition the records of large multidimensional datasets into classes that exhibit strong within-class homogeneity. The art and science of geodemographic classification is not without its critics. Voas and Williamson (2001), for example, suggest that the GB Profiles and Super Profiles classifications are based upon "clouds" rather than "clusters" of areas, and that between cluster variation may be no greater than some within cluster variation. Partly because they have been viewed as "high end" computational techniques, but partly also

for reasons of commerce, data ownership and control, the creation of geodemographic classifications has become the preserve of the expert. Most commercial systems use closed methods and provide limited documentation of the detailed data inputs, the precise clustering methods, or the weighting and normalisation procedures that are used.

In this article we present preliminary work that not only contributes to the creation of more responsive and open geodemographic information systems, but also seeks to question the authority implied by classifications that purport to present "best" solutions. There are a number of motivations for this work. First, current classifications are created from static data sources that do not necessarily reflect the dynamics of population change in modern cities. Data are increasingly available at high temporal resolution and offer the potential to be integrated with other traditional sources to create more timely systems. For example, travel data recording the flow of commuters across a city network could be used to estimate daytime population characteristics. A further example could be extracting frequently updated patient registrations with doctors' surgeries in order to provide a more up-to-date picture of the residential composition of neighbourhoods. Second, application specific classifications have been successfully demonstrated across a variety of domains, and there are many more sectors that could potentially benefit if the methods of construction and interpretation were more accessible and transparent. We argue here that there is a need for web-based applications that enable the creation of general purpose geodemographic classifications "on the fly". In these applications, the specification of (possibly real time) classification inputs should be guided to fulfil the objectives of the problem under investigation, with output from such analysis computed within a reasonable wait time. The exact specification of such a software tool lies outside of the remit of this article; however, we aim to investigate here a core component of a future system. A major problem is the speed at which the clustering algorithm can be used to create robust partitions of datasets into homogeneous groups. Thus, this article presents work on clustering efficiency, that ultimately is hoped will lead to the creation of online tools that manage the estimation and testing of user-specified segmentations tailored for specific purposes. As such, we believe that this is a first step towards a new method of creating geodemographic classifications that fundamentally challenges the prevailing view of segmentation produced by a limited number of expert producers.

2 Real Time Data Modelling and Bespoke Indicators

The task of creating real time geodemographics by integrating diverse and possibly disparate spatial databases raises a number of computational challenges concerning data normalisation and optimisation for fast transactions. This preliminary work assesses three different clustering algorithms for their suitability for integration in an online environment where efficiency of processing is a key priority.

The finest level in most geodemographic classifications is created using the *k*-means algorithm which seeks to find the set of cluster centroids that minimises expression (1) below.

$$V = \sum_{j=1}^{k} \sum_{i=1}^{k} (x_j - \mu_i)^2 \tag{1}$$

where k is the number of clusters, and μ_i is the mean centroid of all the points x_i in cluster i.

The *k*-means algorithm begins by randomly allocating a set of *k* seeds within the data matrix and proceeds by allocating each data point to its nearest seed in multidimensional space. A cluster centroid is then calculated for each cluster, and a new partitioning of the data points is made around the new set of centroids. The centroids are then recalculated for the new clusters of points, and the algorithm repeats these steps until a convergence criterion is met (usually when switching of data points no longer takes place between the clusters). Singleton and Longley (2009) have illustrated how the resulting classifications are sensitive to the placement of the initial seeds, with consequences for the performance of the cluster model. They suggest that, in order to optimise a classification, a model is run multiple times in order to suggest an optimal convergence solution based upon Equation (1). This process is computationally very inefficient, firstly because each *k*-means takes time to compute, and secondly because the outcome of only a single cluster analysis is saved. For classifications created offline this inefficiency is acceptable, yet this is not necessarily the case where users are seeking an interactive solution. An online tool thus requires results to be returned more quickly. For example, if the input UK data for the Office for National Statistics Output Area Classification (OAC: $n = 41$ variables) are clustered on a high specification computer (Intel® Xeon® CPU 5150 @ 2.66 GHz, 3.00 GB of RAM) where $k = 52$, then the processing time for this to converge is 4.23 seconds. Thus for this classification to run 10,000 times, a user can expect to wait for 42,300 seconds (11.75 hours) to obtain results. Considerable work has been undertaken to improve the efficiency of *k*-means. For example, Reynolds et al. (2004) describe a new way of choosing the initial seeds, describing the algorithm as "*k*-means++". This method selects initial centres based on the density of data points and improves the overall processing time because initial seeds are selected more intelligently, thereby enabling data points to converge on clusters more quickly. A more radical method of improving classification efficiency is to supplement *k*-means with other modern classification procedures. Two such techniques include Partitioning Around Medoids (PAM; Kaufman and Rousseeuw 1990) and clustering using genetic algorithms (GA; Maulika and Bandyopadhyay 2000, Painho and Bação 2000).

The PAM procedure is represented in Equation 2. This algorithm attempts to assign points from within a multidimensional data matrix into clusters based in their "nearness" to a series of randomly selected representer points. Unlike k-means, representer points are actual data points from within the data matrix, rather than any point within the Euclidean space. In PAM "nearness" is calculated using a pre-computed dissimilarity matrix across all variables and data points within a data set. This offers improved efficiency over *k*-means because it reduces the on-the-fly distance calculations, and additionally is less sensitive to outliers – because positioning the centroid utilises a median rather than the mean in the optimisation procedures. Effectively, this minimises

$$V = \sum_{j=1}^{k} \sum_{i=1}^{k} |x_j - \mu_i| \tag{2}$$

where the variables are defined as in Equation (1). The pre-computation of a distance matrix is memory intensive and PAM struggles when applied on large data sets. Thus, Kaufman and Rousseeuw (1990) developed a sampling algorithm called Clustering Large Applications (Clara). Clara draws multiple samples of the dataset, applies PAM on each sample, and returns its best clustering as output. Because Clara applies PAM on samples rather than on the whole dataset, it can cope with larger data volumes.

Brunsdon (2006) and Fernández et al. (2005) demonstrated that GA can be combined with PAM to offer further improvements in classification efficiency by supplementing the initial random representor point selection with a genetic algorithm which generates multiple possible sets of representor points. Genetic algorithms run repeated analyses and works through a breeding procedure that preserves the characteristics of the best data points for the subsequent generation. The subsequent generation is created by mutation (change of a random position of the chromosome, i.e. change of a data point in a cluster) and crossover (change of slices of chromosomes between parents, i.e. change of slices of data points). After a number of generations, the genetic algorithm converges on the optimal solution of the clustering problem. For this paper, an R-based genetic algorithm was used varying values of number of generations to devise a clustering solution, retaining a number of data items from one generation to another, and invoking chance mutation.

3 A Comparison of Clustering Routines

This section develops a comparison between *k*-means, Clara, and genetic algorithms (GA). Three metrics are compared: computation efficiency (time), classification optimisation efficiency using average silhouette width, and computation efficiency using different variable standardisation techniques. The aim of this analysis is to examine which type of classification procedure would be most appropriate for computing real time geodemographic segmentations online. To compare *k*-means, Clara, and GA we used the input data for the National Statistics Output Area Classification (Vickers and Rees 2007) aggregated at three geographical levels (Output Area (OA), Lower Super Output Area (LSOA), and Ward for the UK. We used an Intel® Xeon® CPU 5150 @ 2.66 GHz with 3.00 GB of RAM for these comparisons.

3.1 Measuring Clustering Efficiency Based on Computational Time

In order to measure computing time for each of the algorithms we ran *k*-means, Clara, and GA for 1–100 cluster solutions at all of the three different geographic levels of the OAC classification for UK, and then compared the CPU clock time (in seconds) for each algorithm to converge on a specified number of clusters. Figures 1–3 show the relationship between CPU clock time (in seconds) and the number of clusters (1–100) by using *k*-means, Clara, and GA for the three different geographical levels. These algorithms were run a single time for each value of *k*.

The figures show that for large numbers of data points (OA and LSOA levels) Clara runs faster than *k*-means and GA given the number of clusters is a small number (<55 for OA level and <30 for LSOA). However, when the number of clusters is a large number, *k*-means runs faster than Clara. For a small number of data points (Ward), *k*-means runs faster than Clara and GA. As noted earlier, the *k* means algorithm performance is sensitive to the selection of initial seed points, and as such the actual computation time for *k* means may have to be multiplied numerous times if a global optimal solution is sought. Future work is required to examine how heuristics could be used to measure the optimal number of clustering runs required to approximate an optimal solution.

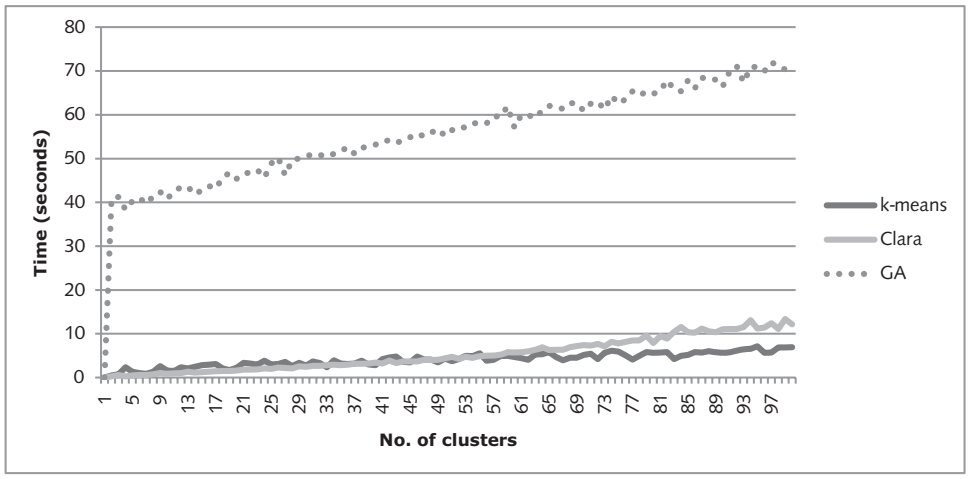

Figure 1 Comparing computational efficiency in an Output Area (OA) level data set covering the UK

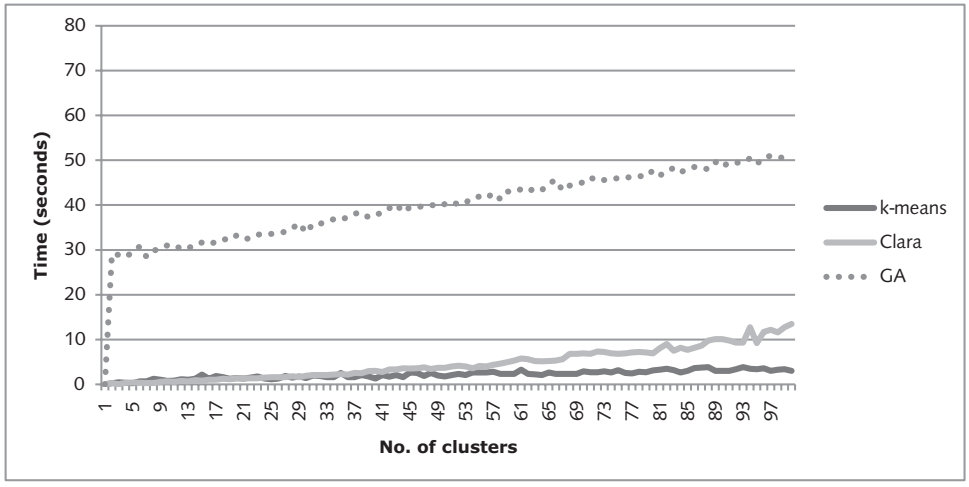

Figure 2 Comparing computational efficiency in a Lower Super Output Area (LSOA) level data set covering the UK

3.2 Measuring Clustering Efficiency Based on Average Silhouette Width

Kaufman and Rousseeuw (1990) introduced silhouette width as a plot showing which data points lie within a cluster and which ones are between clusters. Width of a cluster can be considered as a measure of a good or bad clustering outcome, and enables multiple algorithms to be compared. A large silhouette width indicates a good clustering solution and a small silhouette width indicates an average or bad clustering solution. An average silhouette width is the average of all the silhouette width of different clusters in a clustering problem. Reynolds et al. (2006) demonstrated how average silhouette width

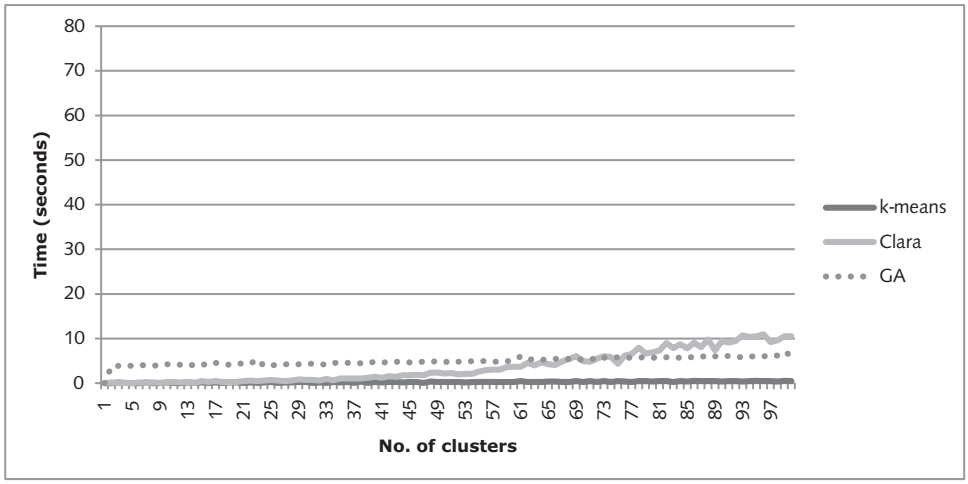

Figure 3 Comparing computational efficiency for a Ward level data set covering the UK

(Kaufman and Rousseeuw 1990) could be implemented as a method of comparing clustering efficiency. They present the following equation for silhouette width:

$$S(k) = \frac{x(k) - y(k)}{max(x(k), y(k))} \tag{3}$$

where $y(k)$ is the average distance of k from all other objects in the cluster C_k. For each $C \neq C_k$ average distance of k from the object C is given by $d(k, C)$. $x(k)$ is the smallest result after computing $d(k, C)$ for all clusters $C \neq C_k$. The mean of $S(k)$ for all objects k, is said to be the "average silhouette width" of that cluster solution. $S(k)$ ranges between 1 for a good clustering solution and –1 which would be a bad clustering solution (Reynolds et al. 2006).

In order to measure relative efficiency of the algorithm optimisation procedures, the average silhouette widths were calculated for *k*-means, Clara, and GA for 1–100 cluster solutions on the three different levels of geographies for UK. Figures 4–6 show the relationship between average silhouette width and cluster frequency for *k*-means, Clara, and GA on the three levels of geographic data for UK. These algorithms were each run a single time for each value of *k*.

These charts show that for large numbers of data points, GA works better than *k*-means and Clara. However, for small numbers of data points, Clara still gives better results than *k*-means and GA.

3.3 Measuring Clustering Efficiency Based on Variable Standardisation Techniques

An important aspect in the creation of every geodemographic classification is the selection of a standardisation technique that converts multiple input variables into measures related on a single scale. In this section we measure the efficiency of clustering algorithms when using three different variable standardisation techniques of: *z*-scores, range standardisation, and principal components analysis.

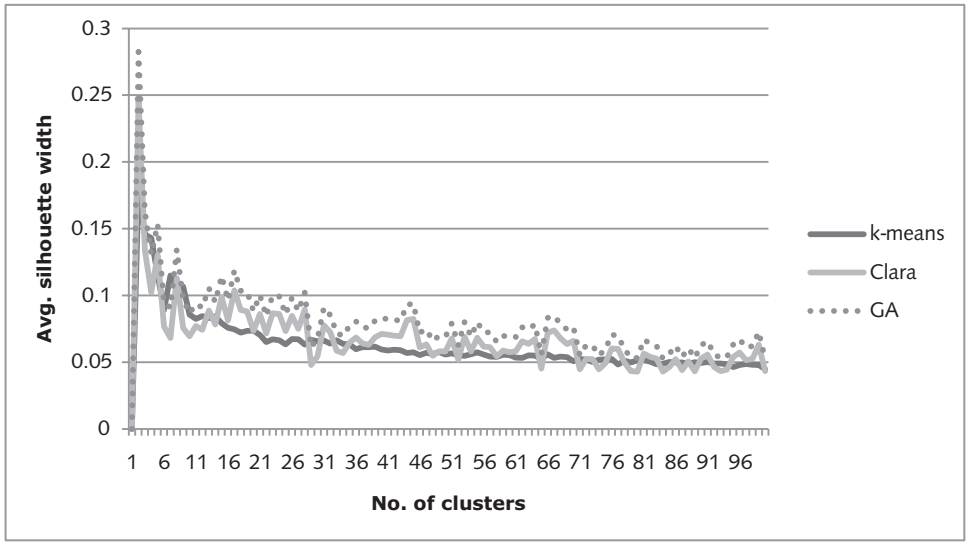

Figure 4 Comparing classification optimisation efficiency in an Output Area (OA) level data set covering the UK

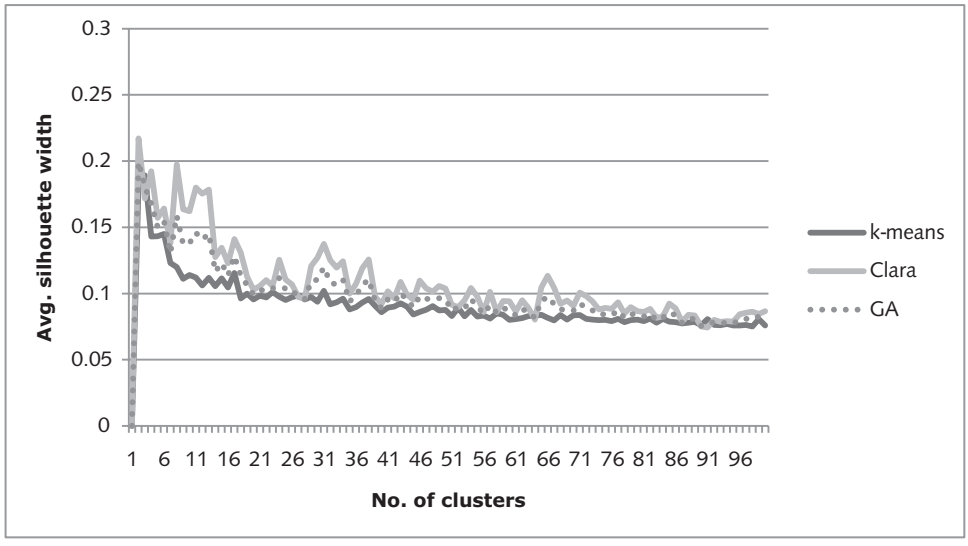

Figure 5 Comparing classification optimisation efficiency in a Lower Super Output Area (LSOA) level data set covering the UK

Z-scores provide one of the most widely used methods of variable standardisation. If x_i is the value of a variable for area i and x_{mean} is the average value of the variable across all n areas, then the z-score is defined as:

$$Z_i = \frac{x_i - x_{mean}}{\sigma_x} \tag{4}$$

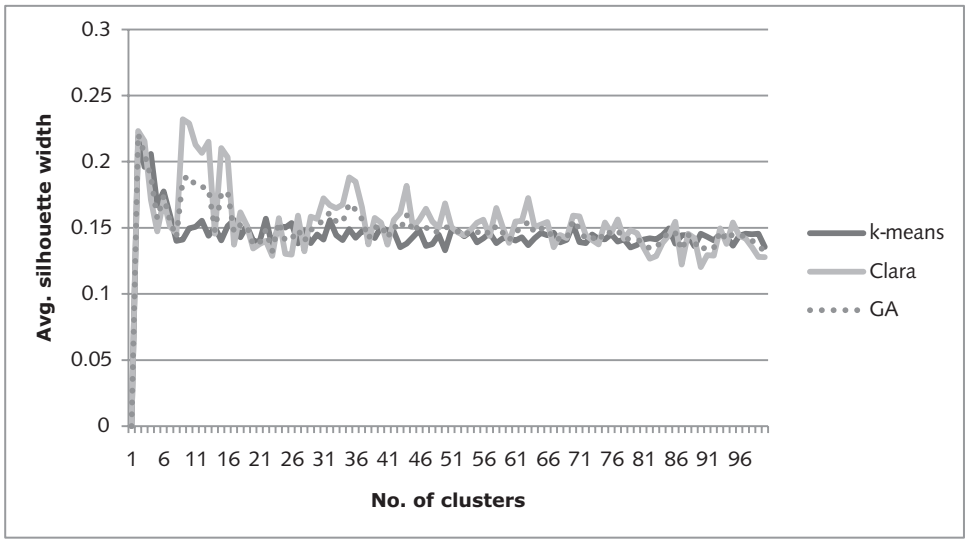

Figure 6 Comparing classification optimisation efficiency in a Ward level data set covering the UK

However, *z*-scores are vulnerable to outlier values within the dataset, and thus it has been argued elsewhere (Vickers and Rees 2007) that transformed variables are not suitable for analysis using *k*-means clustering algorithms. Given that *k*-means is also vulnerable to outliers, a robust solution is likely only to be obtained after numerous re-runs of the algorithm with different initial seed assignments. Additionally, seeds selected with initial positioning influenced by outlier values will likely lead to increased run processing time before meeting convergence criteria. As such, in an online environment where computational time is critical, *z*-scores may add to this computational burden. Vickers and Rees (2007) uses a different variable standardisation technique called the range standardisation method. This method standardises the values of each variable within a 0 and 1 interval and as such limits potential for outlier values. For a variable (x_i) the range standardisation index for area i is given by:

$$R_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{5}$$

A final method that can be used to limit the potential effects of outlier values is principal components analysis.

Each principal component represents a weighted combination of the original variables of the data set. The first component defines the dimension that accounts for the most variation, and each subsequent component accounts for successively smaller amounts of the remaining variation. However, it is important to remember that no method offers a panacea for standardisation. Range standardisation effectively compresses the data into the range of 0–1, while principal components analysis arguably places emphasis on those parts of the dataset that account for maximum variance. As such, both of these variable standardisation techniques may omit some interesting patterns within the dataset.

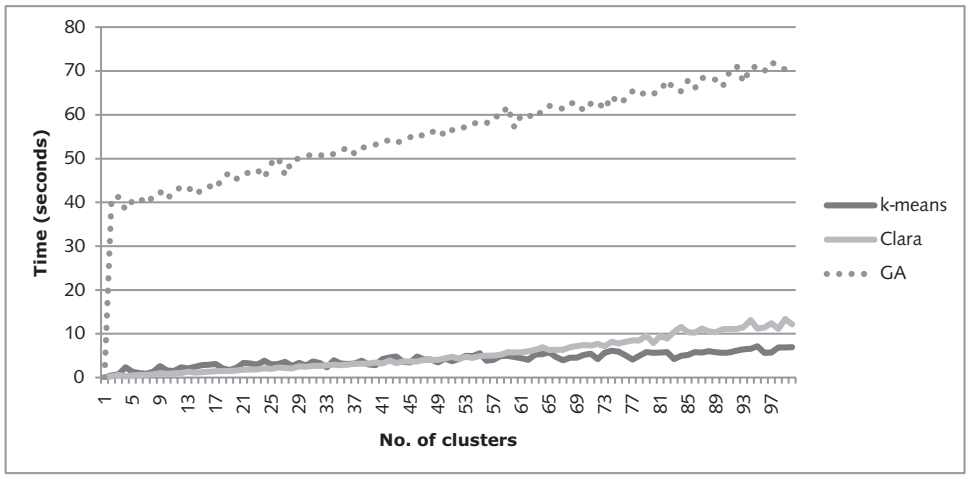

Figure 7 Comparing computational efficiency using z-scores as the standardisation technique for a UK Output Area (OA) level data set

The debate over which standardisation procedure is most effective lies outside the remit of this article, and to some extent is down to personal choice of the classification builder. Thus, in the remainder of this section we aim to develop a comparison of the computational run times of the three clustering algorithms against the number of standardisation methods.

In order to measure the computation time efficiency of the algorithms using *z*-scores as a variable standardisation technique, we ran *k*-means, Clara, and GA for 1–100 cluster solutions again on the three different geographic levels for the UK, and then compared the CPU clock time (in seconds) for each algorithm to converge on a specified frequency of clusters (see Figures 7–9). As with the previous analysis, the algorithms were run only a single time for each value of *k*.

These charts indicate that for a data set with high dimensionality (i.e. the OA and LSOA levels) Clara runs faster than *k*-means and GA when the number of clusters is small (<55 for OA level and <30 for LSOA). However, when the number of clusters is large, *k*-means runs faster than Clara. For small numbers of data points (Ward level), *k*-means runs faster than Clara and GA.

A repeat of the analysis was conducted with z-scores being supplemented for range standardisation (see Figure 10–12).

The results are similar *to z*-scores and broadly show that again for large numbers of data points (OA and LSOA level) Clara runs faster than *k*-means and GA when the number of clusters is small (<44 for OA level and <40 for LSOA). Again, when the number of clusters is increased, *k*-means runs faster than Clara. For smaller datasets with lower dimensionality (e.g. Ward level), *k*-means runs faster than both Clara and GA.

A final comparison was made using principal components as the standardisation method. For the OA level data, principal components analysis was applied on the dataset and 21 principal components were identified. These accounted for 90.69% of the variance in the dataset. The three clustering algorithms were run on these components and results are shown in Figure 13.

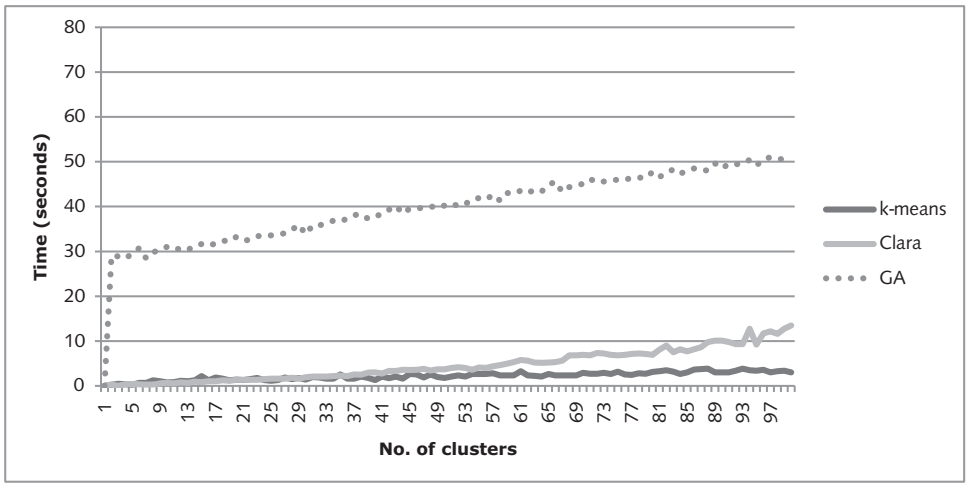

Figure 8 Comparing computational efficiency using z-scores as the standardisation technique for a UK Lower Super Output Area (LSOA) level data set for the UK

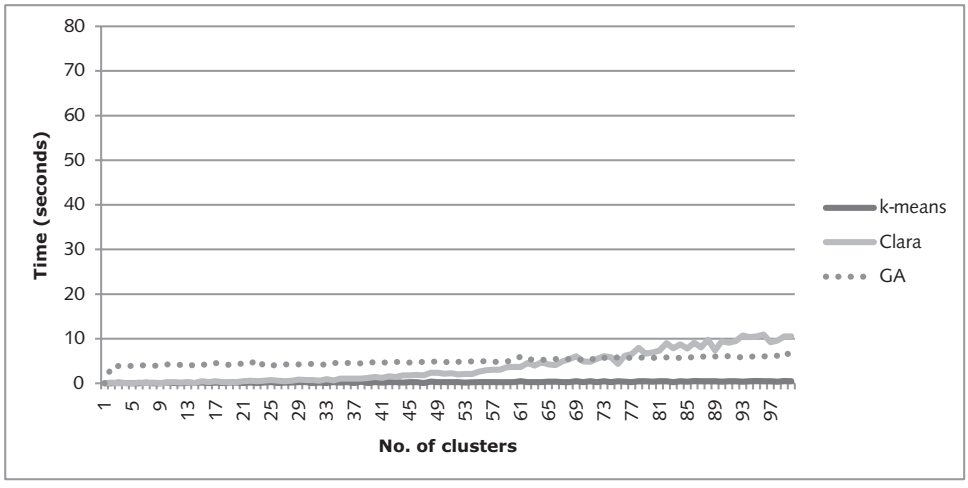

Figure 9 Comparing the computational efficiency using z-scores as the standardisation technique for a UK Ward level data set

For LSOA level data, 16 principal components were created that accounted for 90.8% of variance within the dataset. The three clustering algorithms were then computed on these components, the results of which are shown in Figure 14.

For the Ward level data, principal components analysis was used to identify 10 principal components that accounted for 91.11% of the variance in the dataset. The three clustering algorithms were again run on these principal components, with the results graphed in Figure 15.

Another similar result is returned by these analyses, again showing that for large numbers of data points (OA and LSOA level) Clara runs faster than *k*-means and GA

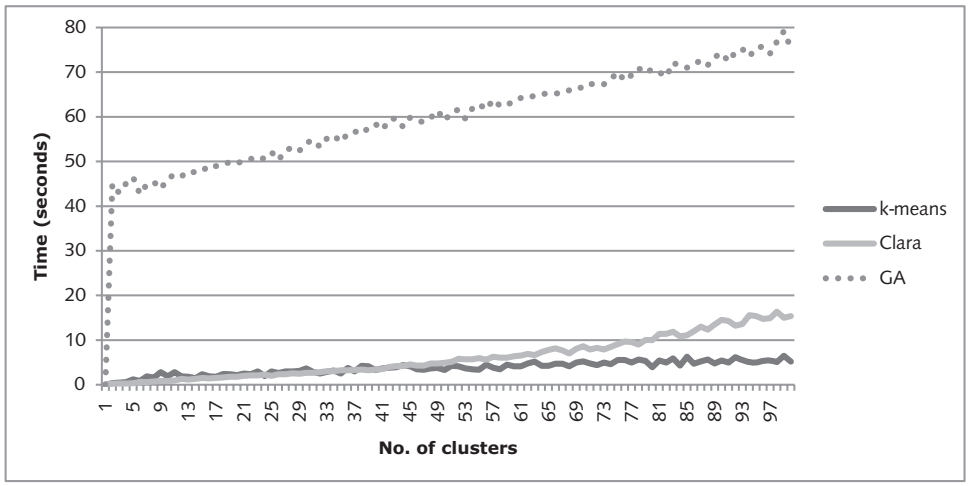

Figure 10 Comparing the computational efficiency when using range standardisation as the standardisation technique for UK Output Area (OA) level data set

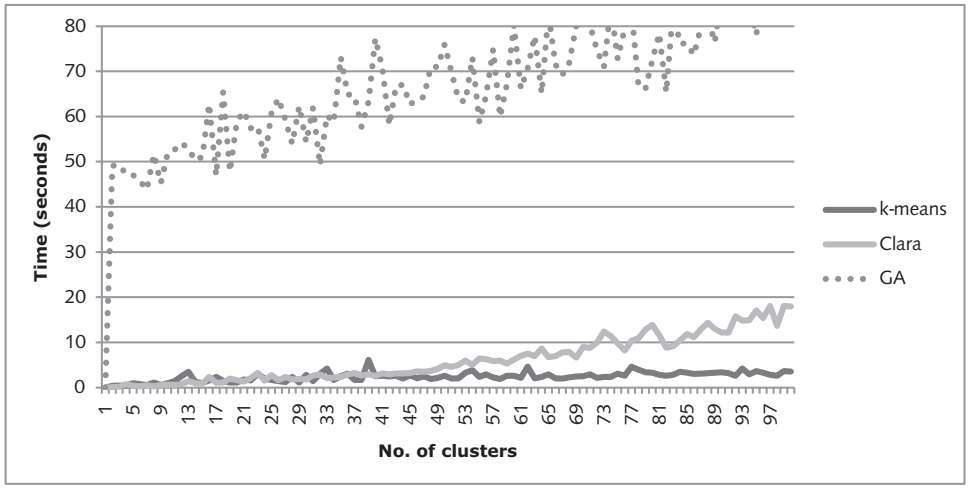

Figure 11 Comparing the computational efficiency when using range standardisation as the standardisation technique for a UK Lower Super Output Area (LSOA) level data set for UK

when the number of clusters is small (<45 for OA level and <30 for LSOA). However, when there are a larger number of clusters, *k*-means runs faster than Clara. For small numbers of data points (Ward level), *k*-means runs faster than Clara and GA.

4 Conclusions and Directions for Future Research

A requirement for distributed and simple-to-use online classification tools arises from changes in the supply of socio-economic data and the potential that this creates for end

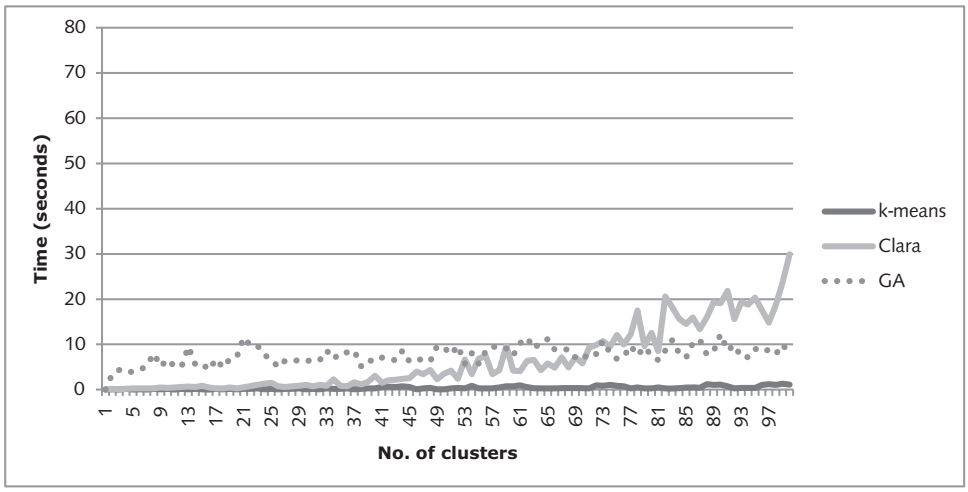

Figure 12 Comparing the computational efficiency when using range standardisation as the standardisation technique for a UK Ward level data set

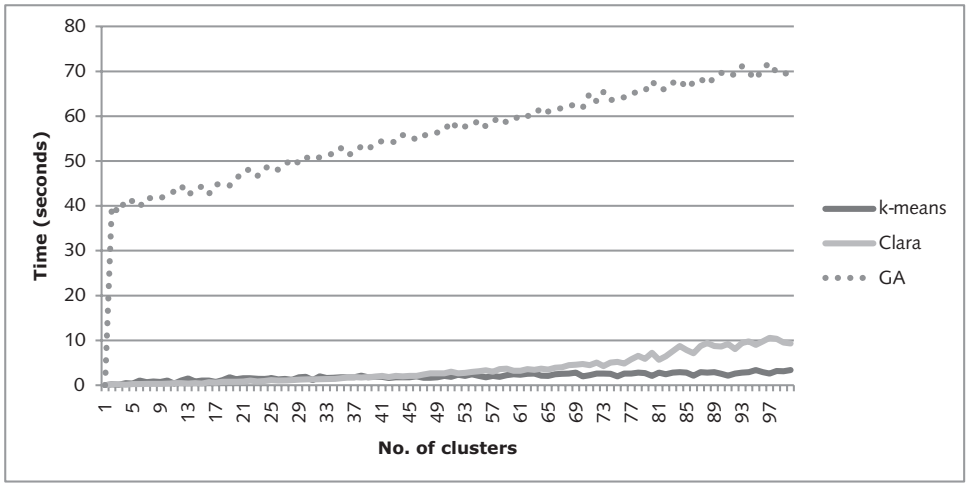

Figure 13 Comparing the computational efficiency using principal component analysis as the standardisation technique for a UK Output Area (OA) level data set

users to create new intelligence on socio-spatial structures. In addition to Census data that are collected every 10 years in the United Kingdom, numerous supplementary data sources are becoming available, some of which are already updated in near real time. The availability of such resources will increase the potential to create more responsive and application-specific geodemographic classifications that will make it less acceptable to uncritically accept the outputs of general-purpose classifications as received wisdom. As compared to the computational challenges posed in the past, real time computational solutions in an online environment are becoming possible and we anticipate that this will provide a stimulus to the development of real time application-specific geodemographic classifications.

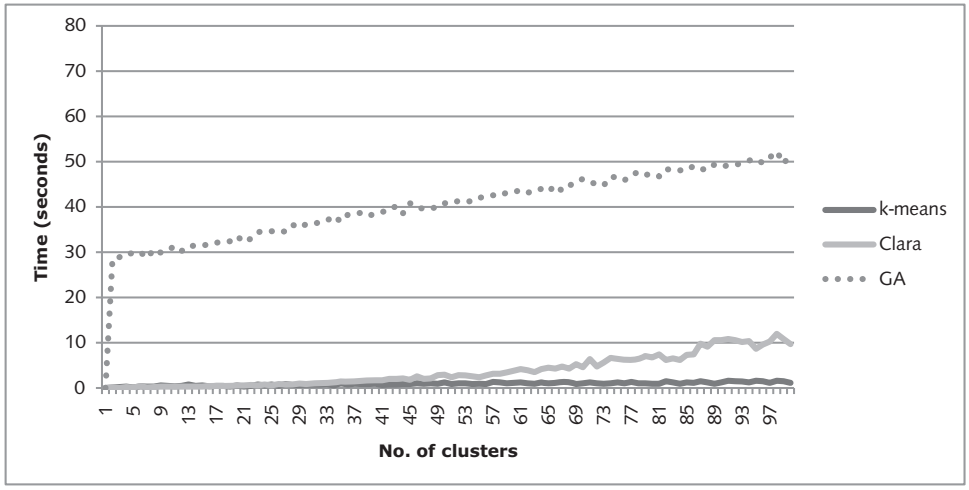

Figure 14 Comparing the computational efficiency using principal component analysis as the standardisation technique for a UK Lower Super Output Area (LSOA) level data set

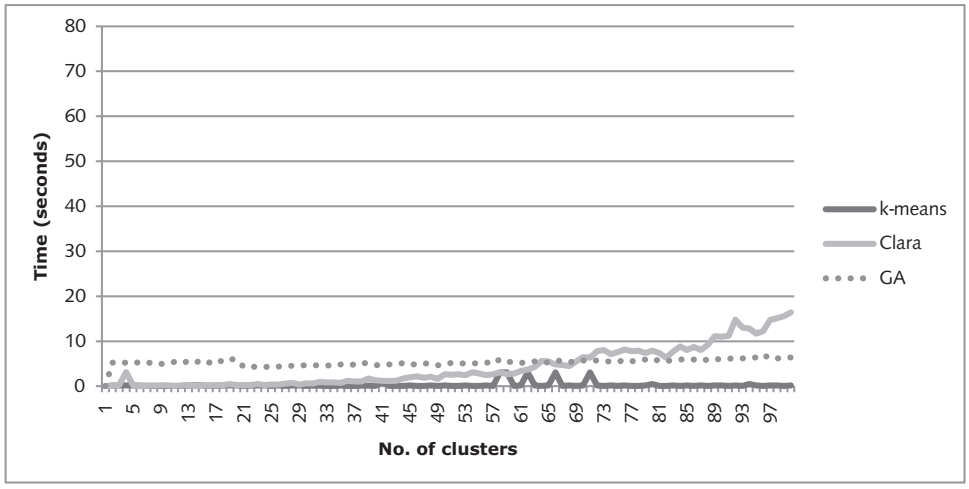

Figure 15 Comparing the computational efficiency of using principal component analysis as the standardisation technique for a UK Ward level data set

Based on the results of this article, we suggest that for the better partitioning of the data (as measured by average silhouette width) GA works better for larger data sets, but Clara is more appropriate for smaller data sets. For speed of computation, Clara works better than GA and *k*-means for larger datasets where the number of clusters is small. For larger data sets where a greater number of clusters is required, *k*-means runs faster than Clara and GA. When examining how these different algorithms are affected by a choice of different standardisation procedures there was little difference in computational times between methods. For an online geodemographic classification system, these algorithms and standardisation procedures could be chosen on the fly, based upon dataset size and user inputs.

Before any substantive conclusions can be drawn from this research further testing is required to compare other clustering techniques, and in particular those methods which could be used to make *k*-means run faster on larger datasets – such as those methods which optimise initialisation procedures and those methods which can use parallel processing architectures for achieving enhanced computational performance of *k*-means.

References

Ashby D I and Longley P A 2005 Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS* 9: 53–72

Batey P W J and Brown P J B 1995 From human ecology to customer targeting: The evolution of geodemographics. In Longley P and Clarke G (eds) *GIS for Business and Service Planning*. Cambridge, GeoInformation International: 77–103

Birkin M, Clarke G, and Clarke M 2002 *Retail Geography and Intelligent Network Planning*. Chichester, John Wiley and Sons

Brunsdon C 2006 A cluster based approach to the zoning problem using and extended genetic algorithms. In *Proceedings of the Fourteenth Annual GIS Research UK (GISRUK) Conference*, Nottingham, United Kingdom

Farr M and Evans A 2005 Identifying "Unknown Diabetics" using geodemographics and social marketing. *Interactive Marketing* 7: 47–58

Fernández V, García Martínez R, González R, and Rodriguez L 2005 *Genetic Algorithms Applied to Clustering*. Buenos Aires, School of Engineering, University of Buenos Aires

Harris R, Sleight P, and Webber R 2005 *Geodemographics, GIS and Neighbourhood Targeting*. London, John Wiley and Sons

Kaufman L and Rousseeuw P J 1990 *Finding Groups in Data*. New York, John Wiley and Sons

Longley P A 2005 Geographical information systems: A renaissance of geodemographics for public service delivery. *Progress in Human Geography* 29: 57–63

Longley P A and Singleton A D 2009 Linking social deprivation and digital exclusion in England. *Urban Studies* 46: 1275–98

Maulika U and Bandyopadhyay S 2000 Genetic algorithm-based clustering technique. *Pattern Recognition* 33: 1455–65

Painho M and Bação F 2000 Using genetic algorithms in clustering problems. In *Proceedings of Geocomputation 2000*, Greenwich, United Kingdom

Reynolds A P, Richards G, and Rayward-Smith V J 2004 The application of *K*-medoids and PAM to the clustering of rules. In Yang Z R, Everson R, and Yin H (eds) *Intelligent Data Engineering and Automated Learning (IDEAL 2004)*. Berlin, Springer Lecture Notes in Computer Science Vol. 3177: 173–8

Reynolds A P, Richards G, Iglesia B de la, and Rayward-Smith V J 2006 Clustering rules. Comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* 5: 475–504

Shelton N, Birkin M, and Dorling D 2006 Where not to live: A geo-demographic classification of mortality for England and Wales, 1981–2000. *Health and Place* 12: 557–69

Singleton A D and Longley P A 2009 Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Papers in Regional Science* 88: 643–66

Vickers D W and Rees P H 2007 Creating the national statistics 2001 output area classification. *Journal of the Royal Statistical Society, Series A* 170: 379–403

Voas D and Williamson P 2001 The diversity of diversity: A critique of geodemographic classification. *Royal Geographical Society with IBG* 33(1): 63–76

Webber R 1975 *Liverpool Social Area Study: 1971 Data*. London, PRAG Technical Paper No. 14

Webber R and Craig J 1978 *A Socio-Economic Classification of Local Authorities in Great Britain*. London, HMSO