The Use of Consensus Clustering in Geodemographics

J. A. Cheshire[1], M. Adnan[1], P.A. Longley

[1]UCL Department of Geography,
Gower Street, London, WC1E 6BT.
james.cheshire@ucl.ac.uk

1. Introduction

Geodemographic classifications require clustering algorithms to partition the records of large multidimensional datasets into groups sharing similar characteristics. Many clustering algorithms have been developed but few have been as widely implemented as the "traditional" methods such as K-means or Ward's hierarchical clustering (Jain, 2010). No two methods create the same result, and multiple iterations of the same method may produce different clusters; it is left to the user to subjectively decide the best outcome. In addition most methods require an *a priori* impression of the number of groups in the data. This abstract outlines a new approach, known as consensus clustering, that utilises familiar clustering methods to produce more consistent results. The method offsets the weaknesses of one type of clustering with the strengths of another by establishing the consistent average outcome from multiple algorithms (Simpson et al. 2010). Consensus clustering has an additional advantage in that it provides a number of metrics that inform the researcher about the inherent groups within the data, and the robustness of the final cluster outcome. Still in its early stages of development, and largely applied in the fields of genetics and bioinformatics, the method has some performance issues when using large datasets but we are confident these can be overcome.

2. Consensus Clustering

Contemporary geodemographic classifications utilise clustering methods in isolation from one another; they do not combine their results in any way. Consensus clustering, proposed by Monti et al. (2003) and extended by Simpson et al. (2010), presents an alternative approach by representing the consensus across multiple runs of a clustering algorithm to determine the number of clusters in the data. This is especially useful when using methods that rely on random seeding to allocate the initial clusters (Monti et al., 2003). Confidence in the result will increase if the multiple clustering algorithms, or parameterisations of a single algorithm, produce comparable results. The output metrics from the Simpson et al. (2010) methodology inform the most appropriate clustering methodology in addition to indicating the optimum number of clusters.

Clustering was undertaken using the clusterCons package, developed by Simpson et al. (2010). A proportion of rows are sampled before clustering with the chosen algorithm and parameters. In this study we utilise the Ward's, K-Means and PAM algorithms. The sampling and clustering is repeated many times gauge the impact of feature removal. The results from each iteration, are stored in a consensus matrix which contains for each pair of items the proportion of the clustering runs in which they are clustered together. A merge matrix provides a way of combining the cluster

outcomes from multiple methods by weighted averaging of their respective consensus matrices. The weighting can be adjusted to increase/ decrease the influence of certain cluster methods. In this case all three are treated as equal. This process gives an indication of the cluster reliability because features consistently grouped together are more likely to be similar than those appearing in the same group less frequently. The merge matrix can then be clustered to yield the final outcome. The advantage of this approach is that it accounts for the different classification properties in each of the algorithms discussed above.

In addition to testing three algorithms, we group the data into a range of clusters. The optimal number in this case is defined by the criteria of Monti et al. (2003) who state that the true cluster number (*k*) can be estimated by finding the value of *k* at which there is the greatest change in cumulative density function (CDF) calculated from the consensus matrix across a range of possible values of *k*. By putting the unique elements into descending order it is possible to calculate a cumulative density function *CDF (c)* defined over the range c=[0,1] using the following equation.

$$CDF(c) = \frac{\sum_{i<j} 1\{M(i,j) \le c\}}{N(N-1)/2} \tag{3}$$

It is then possible to calculate the area under the curve, *AUC* as follows:

$$AUC = \sum \left[x_i - x_{i-1}\right] CDF\left(x_i\right) \tag{4}$$

where x_i it the current element of *CDF* and *m* is the number of elements. If every iteration from the consensus clustering identifies the same groups then the matrix elements will be either 0 or 1, thus producing an *AUC*= 1. This provides the benchmark against which to compare the different clustering results. By plotting the difference in *AUC* values it is possible to identify the appropriate cluster number as it exhibits the greatest reduction. Once the optimal number of clusters has been identified it is possible to re-cluster the merge matrix. The advantage of this approach is the stability in the results produced due to the removal of bias in the clustering structure unique to each clustering technique.

4. Data and Methods

For demonstration purposes we have taken a small dataset covering the London Borough of Southwark and the City of London. The boroughs represent a range of social characteristics. Their combined population is approximately 260,000 across 770 Output Areas (OAs). Each OA has the same 41 variables as the Output Area Classification (OAC) (see Vickers and Rees, 2007), standardised to z-scores. The data were consensus clustered over a range of *k* from 5 to 30. Figure 2 plots *k* against the change in AUC values. The greatest difference in AUC value occurs between 13 and 14 clusters, suggesting that 14 clusters will provide the optimal outcome. The resulting merge matrix was therefore clustered into 14 groups. In addition

conventional clustering without a final merge matrix was performed for comparison.

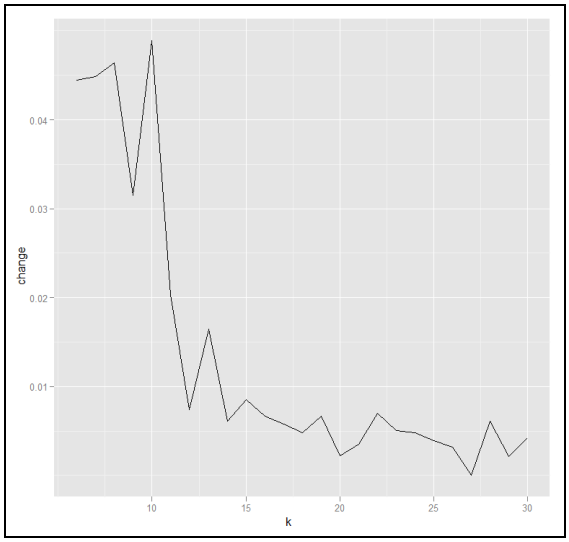

Figure 2: The change in AUC values at a range of *k* from 1 to 30.

5. Results and Discussion

Figure 3 demonstrates that the clustered merge matrix facilitates a more consistent outcome across all three methods with K-Means and PAM being almost identical. One of the most useful metrics from consensus clustering is the robustness measure mapped in Figure 4. The darker colours (signalling higher robustness values) are more prevalent when the merge matrix is clustered and there are significant improvements in the mean values when compared with the standard clustering approaches. In this case PAM produces the most robust cluster outcome that could be used as a basis for a final classification in this context.

Aside from the stability of its outcomes, one of the key advantages of the consensus clustering methodology is the metrics produced that can help inform the decision about the optimal number of clusters to use. In many contexts "optimal" can be defined quantitatively, but in geodemographics the outcomes are generally mapped, assigned group names and provide an important contextual basis for further research. For these reasons "optimal" in the quantitative sense, such as with the lowest within sum of squares value in the case of K-Means, may not be optimal in the practical sense. Consensus clustering does not circumvent these issues, but it does provide more information on which to base decisions. For example, in Figure 2 it is clear that a transition AUC values occurs between 13 and 14 clusters, partitioning the data further will clearly have less of an impact on the final classification (in terms of its robustness) than partitioning into fewer clusters.

A practical constraint to this methodology is it's computational intensity. A national-level classification could not be produced at OA level on a standard desktop workstation, for example. It is our intention to integrate the approach with ongoing research into the creation of geodemographic classifications using NVIDIA's Computer Unified Device Architecture (see Adnan et al. (2010) for more information). This process would enable the consensus clustering to be undertaken many times faster and facilitate fine-scale classifications on a national level.

In conclusion, this abstract has sought to outline consensus clustering in a geodemographics context. The method has demonstrated a strong potential for developing stable classifications and overcomes several of the limitations associated with the conventional implementation of well-known clustering techniques. More work is required to decrease its computation time and also investigate the practical relevance of the results when building a geodemographic classification.

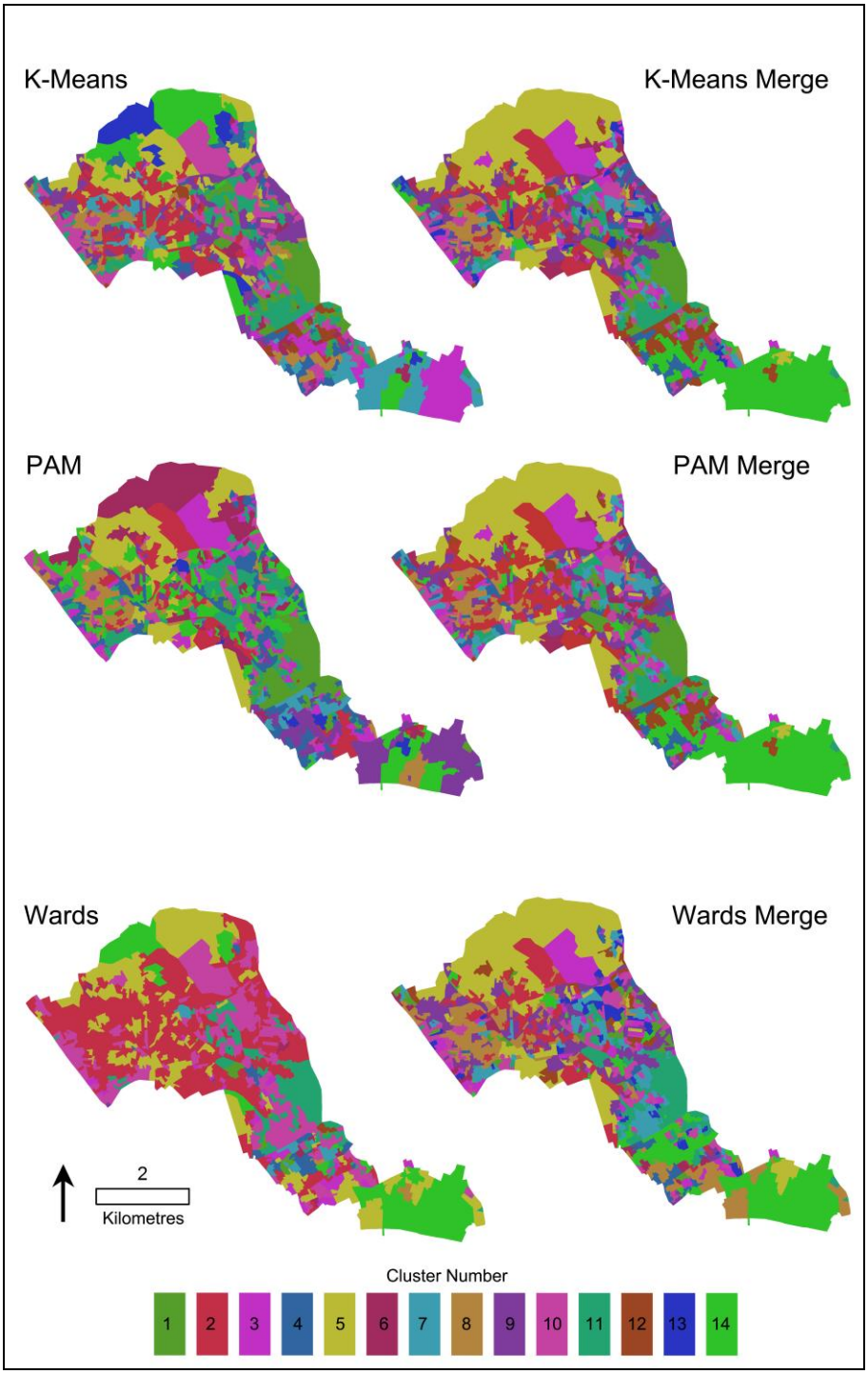

Figure 3: The mapped cluster outcomes from conventional clustering (on the right hand side) and merged consensus clustering (left hand side).

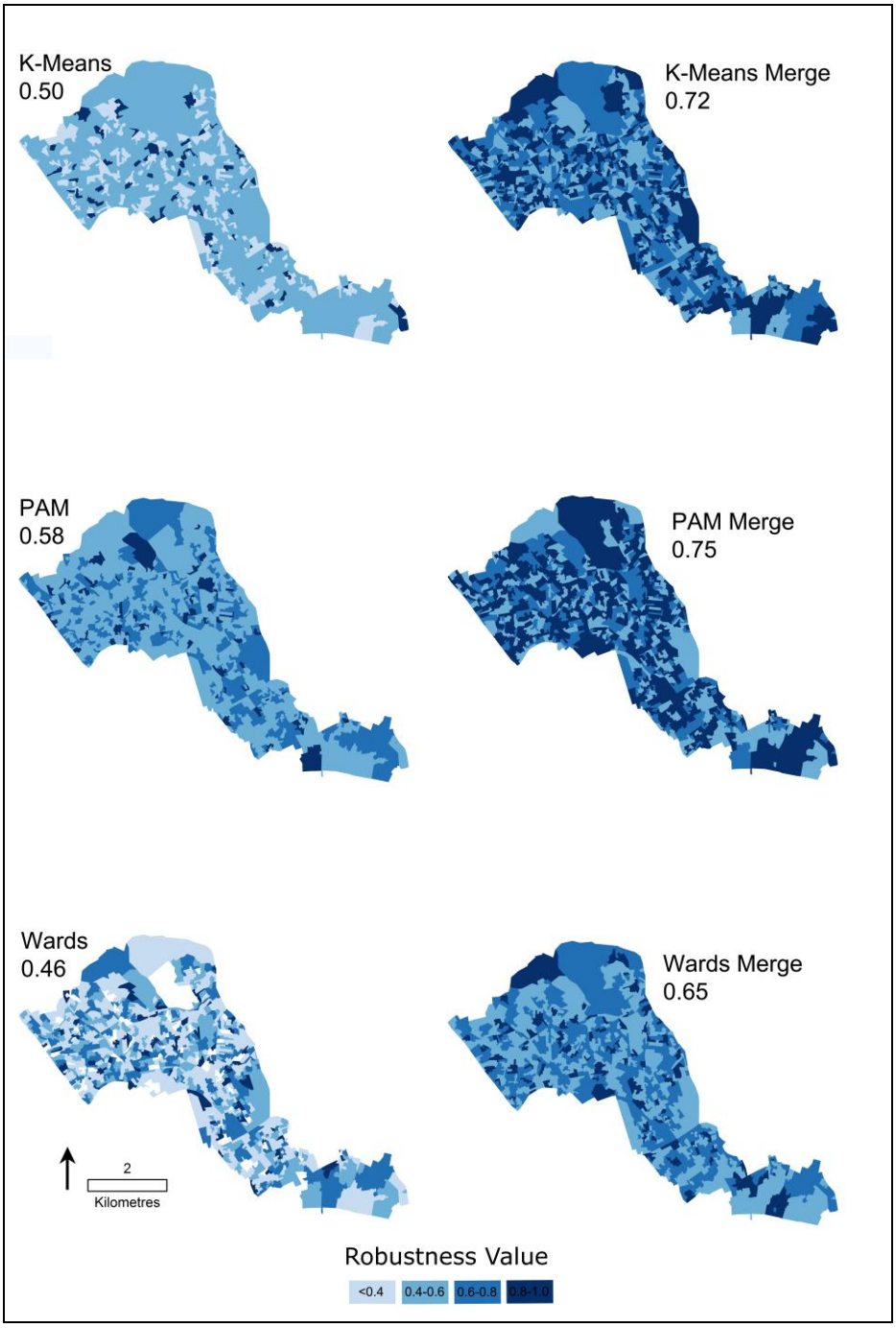

Figure 4: The mapped cluster robustness values outcomes from conventional clustering (on the right hand side) and merged consensus clustering (left hand side). Mean robustness values are also shown.

6. References

Adnan, M., Singleton, S. Longley, P. 2010. *Parallel K-Means Clustering Using Graphical Processing Units for the Geocomputation of Real-Time Geodemographics. Proceedings of the GIS Research UK 18th Annual Conference*. University College London.

Jain, A. 2010. Data Clustering: 50 years beyond K-Means. *Pattern Recognition Letters*. 31: 651-666.

Monti, S., Tamayo, P., Mesirov, J., Golub, T. 2003. Consensus Clustering: A Resampling Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*. 52: 91-118.

Simpson, I., Armstrong, D., Jarman, A. 2010. Merged Consensus Clustering to Assess and Improve Class Discovery with Microarray Data. *BMC Bioinformatics*, 11: 590.

West, M. 2002. Bayesian Factor Regression Models in the Large *p*, Small *n* Paradigm, *Bayesian Statistics*. 7.

Vickers, D.W., Rees, P.H. 2007. Creating the National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society, Series A*. 170 (2), 379-403.