The Role of Geographical Context in Building Geodemographic Classifications

Alexandros Alexiou[*1], Alexander Singleton[♦1]

[1] Department of Geography and Planning, University of Liverpool

November 7, 2014

Summary

Geodemographic analysis is a methodology that simplifies differentiated patterns of socio-economic and built environment structure for sets of small area geography. A particular issue with many current geodemographic classifications is that these lack any explicit specification of geographic context within the clustering process. Within the broad range of geodemographic applications, current techniques arguably smooth away geographic differences between proximal zones, thus limiting classification sensitivity within local contexts. This research begins to address the issue of geographic context by analyzing and evaluating various local, regional and national extents that can be used as attribute contextual weights.

KEYWORDS: Geodemographics, Geographic sensitivity, K-means

1. Introduction

Geodemographic analysis is an established methodology that can provide a simplified measure of socio-spatial structure of small area geography. Such classifications have demonstrated utility over a range of public and private sector applications (Longley, 2005; Singleton and Spielman, 2013). Geodemographic analysis typically uses the K-means clustering algorithm of multidimensional socio-economic variables. This methodological framework can capture a wide set of input attributes, taking advantage of the plethora of census variables and other geographically referenced data to generate aggregate multidimensional profiles (Harris et al., 2005).

A particular issue when constructing such classification is the way attributes are used in the clustering process. Due to the aspatial nature of the K-means clustering algorithm, geodemographic classifications account only for similarities in the clustering process and not the geographical context of each area; areas are essentially treated as independent from one another. Arguably, national aggregations could sweep away contextual differences between proximal zones, reducing the local sensitivity of classifications and thus obscuring potentially important patterns. This type of ecological fallacy raises methodological questions regarding the accuracy of geo-classifications, given the inherent loss of within-cluster variation (Voas and Williamson, 2001).

Proposed methodologies use a number of techniques to address these limitations, typically through the implementation of radial buffers for zones, and selecting attribute locational contextual measures. Although there are many national and proprietary classifications available (i.e. the OAC National Classification by the ONS, MOSAIC by Experian and ACORN by CACI), these classifications may not be suitable when assessing local patterns for policy applications. There are indicators that private classifications incorporate locational attribute sensitivity, however, underlying techniques are typically

[*] a.alexiou@liverpool.ac.uk

[♦] alex.singleton@liverpool.ac.uk

obscured and impeding thus impede reproduction, and as such, there are no established tests to their validity (Harris et al., 2005; Longley, 2007). Counter to this argument is that classifications constructed at the national, regional and local extent are effectively built for different purposes, and as such undermines comparison. This is a longstanding debate originating in the earliest of UK classifications (see Openshaw, Cullingford and Gillard, 1980 and Webber, 1980).

2. Methodology

This research uses a set of fixed input attributes for Output Area zonal geography to build classifications with different geographic extents. For this purpose, a number of scales are considered (local, regional, national) to demonstrate the impact on final classification outcome when input variables are kept constant.

Following the methodology of Harris, Sleight and Webber (2005) and Vickers and Rees (2007), a data set was assembled (Table 1) that includes demographic, economic and housing attributes of England and Wales. The dataset is assembled in its entirety with 2011 census variables, provided by the Office for National Statistics and aggregated at the Output Area (OA) level. Values where converted into percentages in accordance to their respective denominator (with the exception of *V10: Population Density*). In order to minimize the influence of certain attributes in the clustering process, highly correlated variables were later discarded at a cut-off point of 70% and above. The final remaining dataset was then normalized using a Box-Cox transformation and converted into z-scores for standardization:

$$z_{i,\alpha} = \frac{x_{i,a} - \mu_S}{\sigma_S} \tag{1}$$

where $x_{a,i}$ is the attribute value i of area a and μ_S is the mean and σ_S is the standard deviation of the observations in the dataset S. In order to measure the contextual differences between the three geographical levels, the mean and standard deviation of the OA observations for the Local, Regional and National datasets S_L, S_R, S_N where calculated, and z-scores where adjusted accordingly in equation (1). Each of the three final datasets produced where used for the clustering process in order to measure differences in classification performance.

Table 1 Initial attribute dataset used. Attributes are aggregated per OA code.

Variables	*Variable Definition*
Demographic	
V1: Age 0–4	Percentage of resident population aged 0–4 years
V2: Age 5–14	Percentage of resident population aged 5–14 years
V3: Age 15-24	Percentage of resident population aged 15-24 years
V4: Age 25–44	Percentage of resident population aged 25–44 years
V5: Age 45–64	Percentage of resident population aged 45–64 years
V6: Age 65+	Percentage of resident population aged 65 or more years
V7: Ethnic Group, White	Percentage of people identifying as white
V8: Ethnic Group, Black	Percentage of people identifying as black African, black Caribbean or other black
V9: Ethnic Group, Asian	Percentage of people identifying as Indian, Pakistani, Bangladeshi, Chinese or Other Asian
V10: Population Density	Number of people per hectare
Housing	
V11: Privately Owned	Percentages of households that are privately owned
V12: Rent (Private):	Percentage of households that are private sector rented accommodation
V13: Rent (Public):	Percentage of households that are public sector rented accommodation
V14: Detached	Percentage of all household spaces that are detached
V15: Semi-Detached	Percentage of all household spaces that are semi-detached
V16: Terraced	Percentage of all household spaces that are terraced
V17: Flats	Percentage of households which are flats
V18: Central heating	Percentage of occupied household spaces with central heating
V19: No central heating	Percentage of occupied household spaces without central heating

Economic Activity	
V20: Working full-time	Percentage of household representatives who are working full-time
V21: Working part-time	Percentage of household representatives who are working part-time
V22: Unemployed	Percentage of household representatives who are unemployed
V23: Retired	Percentage of household representatives who are retired
V24: Student	Percentage of household representatives who are full-time students
V25: No Qualifications	Percentage of people over 16 years without further education qualifications
V26: Low Qualifications	Percentage of people over 16 years with some qualifications but not a HE qualification
V27: Higher Education	Percentage of people over 16 years for which the highest level of qualification is level 4 qualifications and above
V28: No car household	Percentage of households with no cars
V29: 1 Car household	Percentage of households with 1 car
V30: 2 Car household	Percentage of households with 2 cars
V30: 3 Car household	Percentage of households with 3 cars
V31: 4+ Car household	Percentage of households with 4 or more cars
V32: NSeC - managerial	Percentage of households with an HRP with a managerial position
V33: NSeC - intermediate	Percentage of households with an HRP with an intermediate occupation
V34: NSeC - semi	Percentage of households with an HRP with a semi-routine occupation
V35: NSeC - none	Percentage of households with an HRP with no occupation

The classification methodology to produce clusters is the iterative allocation–reallocation algorithm, known as the *K-means clustering* detailed in Milligan (1996) and Everitt, Landau and Leese (2001). K-means clustering uses squared Euclidean distance as a dissimilarity function. Essentially, K-means clustering assigns n observations into K clusters in such a way that within each cluster, the average distance of the variable values from the cluster mean is minimized. For the aggregate of the total clusters there is a set of arguments that minimize the total within cluster variation of the multidimensional data points:

$$\mathrm{WCSS} = \min_{c} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2 \tag{2}$$

where WCSS is the within-cluster sum of squares for a cluster distribution C with K seeds, $x_i \in N$ is the data observations and $\bar{x}_k$ is the k cluster mean. Since the algorithm is dependent on the initial seeds, it must run multiple times in order to obtain optimal results (typically minimizing the WCSS).

Once the optimised sets of K cluster assignments are calculated for each scale of input, clusters within each set are matched in order to determine which cluster ID from one classification fits best to another. Besides the typical qualitative way, i.e. cross-tabulation of the within-cluster distribution, an algorithm was also developed that for a set of different classifications calculates the minimum absolute distance between cluster attribute means to test if this process could be used for a wider set of comparisons in the future.

If $k_i = \begin{pmatrix} \mu_1 \\ \cdots \\ \mu_n \end{pmatrix} \in K_i$ represents a vector with the average attribute values μ cluster k_i of the set K_i, then that cluster is more similar to another cluster $k_j \in K_j$, given they come from the same set of observations S, when:

$$k_i - k_j = \underset{\mu}{\operatorname{argmin}} \sum_{n} \left\| \mu_{k_i}^n - \mu_{k_j}^n \right\| \tag{3}$$

Finally, this research uses the R programming language in order to perform the analysis and map the output classifications.

3. Preliminary results and future directions

In this particular example, the Local Authority of Liverpool is considered, which contains 1584 Output Areas, and is used as a basis to compare different classification outcomes. Figure 1 demonstrates how the local, national and regional classifications are mapped within this context. Between the classifications, there are differences in the emergent cluster patterns, with the local classification appearing to offer the greatest differentiation between areas. The cluster mapped with a red colour represents the most affluent residents (e.g. "*White Collar Families*").

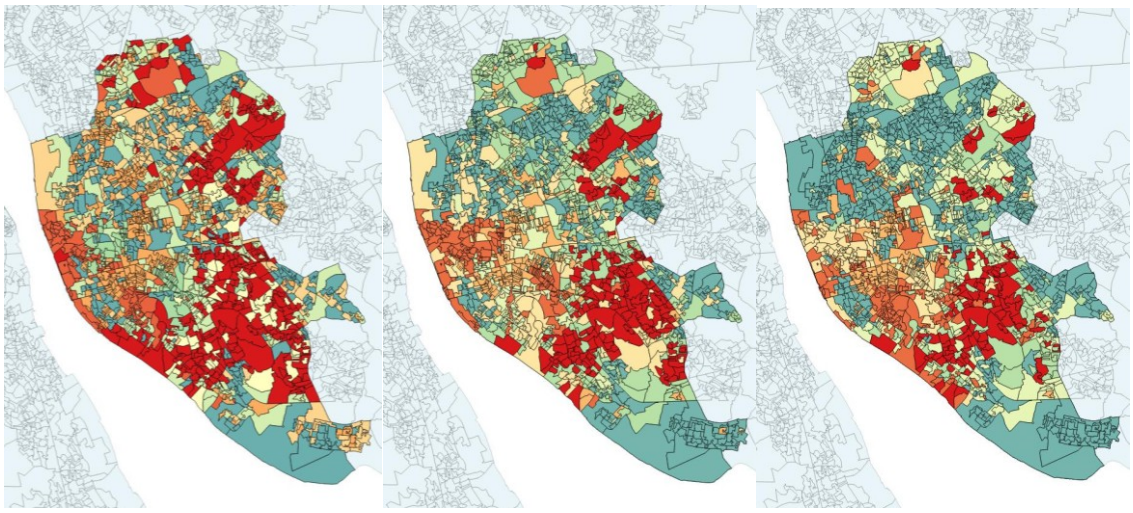

Figure 1 Differences in cluster patterns in Liverpool, UK. From left to right: Local, Regional and National geographical contexts used to calculate attribute extents.

Figure 2 shows a summary of the distribution of the values within this cluster portrayed as "*White Collar Families",* and gives an example how the developed cluster-fit algorithm works (in this case it fitted best the clusters 4, 5 and 7). It is also evident that the number of OAs in the cluster decreases as the attribute extents are scaled more globally in the case of Liverpool. For instance, an affluent family by local standards may not be as affluent by national ones. Since the Liverpool area is considered generally deprived, this number decreases from 234 OAs to 172 in the regional context and 118 in the national one.

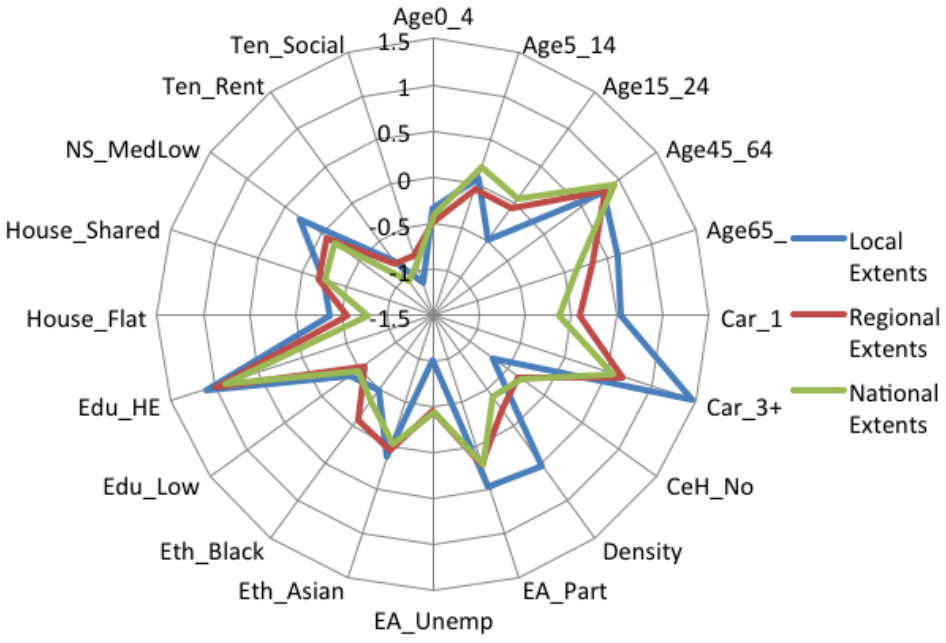

Figure 2 Distribution of average attribute values of Clusters 4, 5 and 7 (mapped red in Figure 1) for local, regional, and national extents respectively.

Although preliminary results show some degree of differentiation, a more extensive analysis is required to explore how these patterns may map between different geographic contexts, for example,

how might such patterns differ between Leeds, Liverpool, Manchester or Lancaster. Furthermore, research is needed to explore how classifications created at local or regional extents can be assembled in a way that national comparisons become possible. A challenge for future research is how these differences can be measured, and how between classifications created for different scales impacts upon the performance of the classifications when used for real world applications.

Finally, for simplicity, administrative definitions of context have been used for this study, however, we recognise that these may not represent true functional regionals or localities, and as such, further work is required about how local or regional extents might be defined, and what impact these geography will have on the final classification. In particular, at a local level, further work is also required to examine how built environment / transport infrastructure can be used to measure geographic extents and how this may impact up emergent patterns.

4. Acknowledgements

This research is part of a PhD project funded by the ESRC with an Advanced Quantitative Methods (AQM) award at the North West Doctoral Training Centre.

5. Biography

Alexandros Alexiou holds a diploma in Engineering with an MSc (Transport Planning) from the Aristotle University of Thessaloniki and an MPhil (Land Economy) from the University of Cambridge. Alexandros is currently a PhD candidate at the University of Liverpool, with research interests in the creation of new models of urban socio-spatial structure that better account for both geographic context and the dynamics of population.

Alex Singleton is a Reader in Geographic Information Science at the University of Liverpool. His research interests extend a geographic tradition of area classification and have developed a broad critique of the ways in which geodemographic methods can be refined through modern scientific approaches to data mining, geographic information science and quantitative human geography.

References

Everitt B S, Landau S and Leese M. (2001). Cluster Analysis. 4th edn. London: Arnold.

Harris R, Sleight P and Webber R (2005). Geodemographics, GIS, and Neighbourhood Targeting. Chichester: John Wiley & Sons.

Longley P A (2005). Geographical information systems: a renaissance of geodemographics for public service delivery. Progress in Human Geography 29(1): 57-63.

Longley, P. A. (2007). Some challenges to geodemographic analysis and their wider implications for the practice of GIScience. Computers, Environment and Urban Systems 31(6): 617–622.

Milligan G W (1996). Clustering validation: results and implications for applied analyses. In P. Arabie, L. J. Hubert & G. De Soete (Eds.), Clustering and Classification (Vol. 2, pp. 120-125). Singapore: World Scientific Press.

Openshaw S, Cullingford D and Gillard A (1980). A critique of the national classifications of OPCS/PRAG. Town Planning Review 51 (4): 421.

Singleton A D and Spielman S E (2013). The Past, Present and Future of Geodemographic Research in the United States and United Kingdom. The Professional Geographer.

Vickers D and Rees P (2007). Creating the UK national statistics 2001 output area classification. Journal of the Royal Statistical Society. Series A. Statistics in society 170(2): 379-403.

Voas D and Williamson P (2001). The diversity of diversity: a critique of geodemographic classification. Area 33(1): 63–76.

Webber R J (1980). A response to the critique of the national classifications of OPCS/PRAG. The Town Planning Review 51 (4): 440–450.