Geodemographic Classifiability

Dr Paul Williamson, Dept. of Geography and Planning, University of Liverpool

Extended Abstract

1. Introduction

This paper has two goals. First, it seeks to demonstrate the various ways in which the precision of geodemographic classification is spatially variable. Second, it seeks to explore which particular aspects of the geodemographic classification process can be most effectively used to mitigate against this spatial variability of classification fit. Reflecting this two-fold goal, the paper is split into two main parts, the first dealing with the measurement of the geographic variability of geodemographic classification; the second dealing with refining the geodemographic classification process to minimise this spatial variability. Both sections use the Output Area Classification, devised by Dan Vickers for ONS, as a test-bed to explore the ideas being discussed. Hence the paper commences with a brief overview and critique of OAC.

2. OAC as an exemplar of Geodemographic Classification

Brief outline of geodem (rational for classifying areas; examples of classifications being used to inform policy).

Brief intro. to OAC. As many other authors have already noted, OAC provides an excellent test-bed for investigating the merits and failings of geodemographic classification. OAC, and the data from which it was derived, are all publicly available. Equally importantly, full details have been provided of the methodology adopted to create the classification. In addition, whilst there will inevitably always be room for improvement, the basic approach adopted for the creation of OAC is methodologically sound, and broadly representative of 'best practise' in use at the time OAC was produced.

Brief Critique of OAC. Whilst a number of criticisms of OAC have been raised by various authors, the main points of contention are that (i) OAC is a 'one sample' solution; (ii) OAC suffers from apparent variability in the quality of its spatial resolution with, for example, nearly all of London placed into only one Supergroup, whilst other regions of the UK generally contain areas from a far wider selection of Supergroups.

The next section of this paper investigates the main problem faced by OAC, and by geodemographic classifications in general – the geographical variability in classifiability of areas

3. Measuring geographic classifiability

(a) What do we already know about spatial variability of OAC?

Table or graph of % supergroups [n=7] /groups [n=21] /sub-groups [n=52] x geographic coverage [GOR?] showing lack of diversity of area classification in London etc. [Is criticism of London classification fair; or do sub-groups neatly break up the London supergroup?]

Geography of fuzziness (i.e. what other cluster(s) each area is most closely aligned with)

(b) Measures of geographic classifiability

In order to explore the spatial variability in the precision of geodemographic classification more closely, this paper explores the performance of OAC using a suite of three novel measures of spatial variability, focussing on the OAC 'Supergroup' classification of GORs.

Measure 1: Spatial variation in cluster compactness (GC_{cc})

- How close OAS in a given region are to the national centroid of the cluster into which they have been classified

Measure 2: Spatial variation in systematic deviation (GC_{sd})

Equivalent to Everitt et al 2001, p184 'relative tightness' [average inter-point distance within cluster? a.ka. 'homogeneity'?]

- How systematically OAs within a given region differ from other OAs in the same cluster

Measure 3: Spatial variation in intra-cluster variability (GC_{icv})

[Equivalent to Hatzichristos (2004) 'dispersion of areas' measure?]

- Average distance between OA pairs within region differ from average distance between all OAs in that cluster.

(d) Classifiability using LAs and ward geographies

Having established that some the output areas within some regions are better classified than others, the problem is revisited for the lower-level geographies of local authorities and wards. Consideration is also given to the change in precision that results from moving to the higher resolution area classifications of groups and sub-groups.

- Variation by supergroup
- Variation by group and sub-group

(e) The link between geographic classifiability and other metrics of cluster quality

The novel metrics presented above for the capture of spatial variability in the performance of geodemographic classification were specifically designed to tease out meaningful dimensions this performance, allowing the development of a better understanding of what factors this is attributable to. However, a wide range of other metrics for the evaluation of geodemographic classification already exist. In this section the novel measures presented above are compared to more conventional measures, including:

Plot of %SS; Davies-Bouldin index; Silhouette score against GC indices (Manhattan and Euclidean based)

(f) Conclusion

- What do these new measures tell us about the OAC that we didn't know already?

- To what extent do more finely resolved clustering solutions solve the problem (i.e. sub-groups v. supergroups)?
- To what extent to the findings influenced by spatial resolution at which 'classifiability' is analysed?
- How geographic classifiability is (not?) related to more conventional clustering metrics
- Lessons for future *area* classifications
    - Variables that better capture 'city' / 'remote rural' effects (e.g. values of neighbouring areas as attributes of each area being clustered)
    - More focus on variables that show strongest levels of spatial variability (via weighting or omission of non-varying variables)
- Space is 'just a dimension'; so lessons for other more general clustering problems

4. Refining the OAC clustering process

(a) Possible improvements to the OAC approach

The following are possible improvements to the existing OAC clustering process:

- **Starting point** (1 random; multiple random starts; pp)
- **Variable transformation** (log(x=1); symmetrifying; normalizing inputs; dealing with 0s)
- **Distance metric** (Manhattan v. Euclidean) (aka k-means v. k-medians)
- **Neighbourhood effect** (add 'neighbouring OA average' [or, as crude proxy, ward value] of each variable to list of OA characteristics; plot variance in $d_{i,M}$ and use stepwise-selection to identify an 'optimal' variable sub-set comprising, at most, same number of variables as were input to OAC.
- **Clustering algorithm** (K-means; k-medioids (pam); SOM/Neural Gas)

(b) Identifying an 'optimal' classification approach

To exhaustively test all possible permuations of the above algorithmic variations would be overly time-consuming. Therefore, in this section of the paper a series of experiments are conducted, changing one aspect of the algorithm at a time, in a way designed to solicit, if not 'the' best algorithmic solution, then at least 'a best' solution.

The OAC was created by classifying OAs into supergroups; then further classifying the OAs within each supergroup separately. (i.e. a 'top-down' approach). Hence classifications broadly comparable to supergroups are immediately reproducible (set k the same). Note, however, that the 'optimal' number of supergroups adopted for OAC might be sub-optimal for an alternative clustering solution. It would be even more difficult to create comparable sub-clusters, as judgements need to be reached about the appropriate number of sub-clusters into which each 'supergroup' should be broken (which will vary by 'supergroup' AND by clustering solution).

For each of following experiments, 'Supergroups' are created, and their success considered in terms of: % within-cluster sum-of-squares; Davies-Bouldin index; Silhouette plots; GC indices [at ONE spatial scale only: GOR or district]

	Starting point	**Data transformation**	**Distance metric**	**Variable set**	**Algorithm**
Baseline	First random	Log(x+1)	Euclidean	All	k-means
Exp. 1	Best random	Log(x+1)	Euclidean	All	k-means
Exp. 2	*winner*	Symmetrify	Euclidean	All	k-means
Exp. 3	*winner*	Normalize	Euclidean	All	k-means
Exp. 4	*winner*	Winner + adjustment for 0s			
Exp. 4	*winner*	*winner*	Manhattan	All	k-means
Exp. 5	*winner*	*winner*	*winner*	'best' sub-set	k-means
Exp. 6	*winner*	*winner*	*winner*	*winner*	k-medoids
Exp. 7	*winner*	*winner*	*winner*	*winner*	SOM/Neural gas

(c) Conclusion

(1) To what extent do the above changes in clustering process help to improve geographic classifiability?
(2) Of the above changes, which are the most important?
(3) Using ALL of experimental results, plot DB index, Silhouette score and % SS against GC indices to see if there is a clear relationship (always using relevant experiment's distance metric) (i.e. can focus of clustering still remain on optimising current cluster validity measures; or do the GC indices add a new dimension that cannot be ignored, at least for AREA classifications
(4) Point out that, although paper is expressly spatial, space is simply one dimension of an n-dimensional problem, so findings from paper likely to generalise to other dimensions of the clustering problem; and to clustering applications in other (aspatial) disciplines

(d) For the future

Even if one clustering solution is shown to be clearly superior to OAC in terms of internal valuation metrics (DB index; %SS; Silhouette score; GC indices), this still begs the question of whether this makes any difference in terms of actual applications. Therefore need one or more worked examples.

Could knock out one or more of input variables; recreate classification using favoured method and see how well this explains spatial variability in omitted variable compared to solution using least favoured method. (Can't compare directly with OAC, because OAC based on all variables – but could compare to OAC 'method' by selecting one of multiple random starting point solutions with same rank as OAC turns out to have relative to range of solutions derived from Exp. 1). Omitted variable(s) should be chosen on basis of policy relevance (e.g. health; not %50-54 year olds) and challenge (e.g. v. skewed v. relatively normal; spatially concentrated v. spatially diffuse). Could (also) use migration. C.f. Vickers Chapter 7 p236- , where migration rate (turnover – i.e. % of population that moved in and out) is shown to vary significantly by supergroup; and Vickers states migration wasn't used in classification. [Not in list of variables considered in Chapter 5 re. OAC classification. In chapter 4 (LA classification) not considered because migration data for N Ireland hadn't been released at that

stage). Used population change 1991-2001 as a proxy. Population change 1991-2001 presumably not available as a variable for OAs, as the OAs didn't exist in 1991!

The above avoids the challenge of finding a 'best' problem-specific solution, which is problematic because:

If you don't have the small-area data in the first place, how do you establish what are the key variables that drive small-area differences? Best logistic model for person-level data = ecological fallacy. Survey data coded only with higher-level geographies (e.g. region or district) might deliver misleading ecological models. Primary sampling units (for which full 'enumeration' may be assumed are, conversely, probably too small...)