

**A Proposed Quantitative Comparative
Analysis for Geodemographic
Classifications**

A. Ojo

**Department of Geography, University of Sheffield,
Sheffield, UK**

for

Yorkshire and Humber Public Health Observatory

Abstract

Social area classification focuses on segmenting geographical units into groups based on the socio-economic characteristics of their residents. ¹⁵

The benefits of geodemographic classification systems for informing health intelligence are increasingly being exploited. In this paper, we present a framework for evaluating the usefulness of geodemographic segmentations. We embark on a demonstration to compare six classification systems at our disposal.

Alongside Lorenz curves, we have also employed a statistic called the Gini-coefficient. We use this to measure discriminatory power for each of the different systems for four health conditions including admissions for Asthma, Coronary Heart Disease (CHD), Chronic Obstructive Pulmonary Disease (COPD) and Diabetes.

We also attempt to use some other factors to define usefulness of a system and conclude that each system has its pros and cons.

Key words: Geodemographics, Public health, Hospital Episodes Statistics,

Introduction

The rise of population and consumer segmentation systems in countries like Britain and America has been greatly influenced by the appreciation of these systems within the commercial industry. The transition from mass marketing to niche marketing contributed greatly to the growth^{8, 4} with numerous commercial companies opting for target-marketing and adopting intelligent mail-shot systems.

With quite a number of classifications now available, the choice of the system to adopt becomes difficult for many users. For obvious reasons, most vendors claim to have the best system for almost any purpose. This may not be true as there is a relationship between the purpose of a classification and the variables it encapsulates.⁷ The onus therefore rests on the user of the system to decide on what to use and what not to use.

In this paper, we present a comparative analysis of a number of classification systems. We have used data from the national Hospital Episodes Statistics (HES). Admissions for Asthma, Coronary Heart Disease (CHD), Chronic Obstructive Pulmonary Disease (COPD) and Diabetes were explored using each of the six classification systems at our disposal.

Methods

Area typologies have provided room for informed decision making within the public sector.⁴ Some arena's where area classifications continue to function include community safety and policing¹, education^{13, 16}, regional planning and development² and public health.^{12, 14} Most systems are created in hierarchical structures meaning that the number of clusters at various hierarchies differs.

Aside the cost of purchase, a number of other important issues can be considered when evaluating geodemographic systems. These issues include:

- The use for which the system is intended
- The input variables
- Knowledge of clustering algorithm used
- The discriminatory power of the system
- The predictive power of the system

Some of these issues have been examined briefly in this paper with particular focus on discriminatory power. Discriminatory power quantifies how well a system measures the within-population variation of the rates or concentration of a target characteristic/variable relative to a base population. The use of Gini-coefficient and Lorenz curves has been suggested as a useful measure for assessing this discrimination.^{9, 11}

Datasets

Hospital Episodes Statistics (HES) is a data warehouse containing information about patients treated by NHS providers in England. It is mainly populated from

routine data flows exchanged between providers and commissioners. The database is very detailed and dates back to 1989/1990.

From the HES database, the conditions were identified using:

Primary diagnosis codes: -

Asthma ICD10 : J45, J46

CHD ICD10 : I20-I25

COPD ICD10 : J40-J44

Diabetes ICD10 : E10-E14

A count of finished admissions was used for each patient. A finished admission episode is the first period of in-patient care under one consultant within one healthcare provider. Admissions do not represent the number of in-patients, as a person may have more than one admission within the year. In this paper we have explored admissions for three years; 2003/2004, 2004/2005 and 2005/2006. The admission conditions analysed include Asthma, Coronary Heart Disease (CHD), Chronic Obstructive Pulmonary Disease (COPD), and Diabetes.

Table 1
Comparison of leading geodemographic classification tools (adapted from Dedman et al.⁵ and taken from APHO Technical Briefing No. 5: *Geodemographic Segmentation* – due to be published in December 2008)

Supplier	CACI	CACI	Experian	ONS	ONS	Beacon Dodsworth	Acxiom		
Tool	ACORN	Health ACORN	Mosaic UK	Output Area ^a Classification (OAC)	2001 Area Classification (above OA level)	People & Places P ²	Personix Geo		
Categories	Category (5) ↓ Group (17) ↓ Type (56)	Group (4) ↓ Type (25) ↓ Sub Type (59*)	Group (11) ↓ Type (61) ↓ Segment (243*)	Super Group (7) ↓ Group (21) ↓ Sub Group (52*)	Super Group ↓ Group ↓ Sub Group*	Numbers depend on spatial level	"Tree" (14) ↓ "Branch" (41) ↓ "Leaf" (160*)	Category (5x4) ↓ Group (20) ↓ Type (60)	
	Some areas unclassified? ^b	Yes	No	No	No		No	Yes	No
	Sorting of categories	Affluence	Health Outcome	-	-		-	Affluence	Lifestage / Affluence
	Variables	Construction variables - UK 2001 Census - Survey data	- UK 2001 Census - Survey of food consumption - Survey of health & consumer lifestyles	- UK 2001 Census - Survey data	- UK 2001 Census		- UK 2001 Census	- UK 2001 Census - TGI Household Survey	- UK 2001 Census - Survey data
Other	Cost	Subscribe	Subscribe	Postcode: Subscribe. LSOA using aggregated data: free for academic use	Free	Free	OA: Subscribe. LSOA using a rebuild of data: free for NHS	Subscribe	
	Smallest geographical level	Down to Postcode	Down to Output Area	Down to Postcode / Household	Output Area	LA down to Super Output Area	Down to Output Area	Down to Postcode	
Contacts	Web address	www.caci.co.uk	www.statistics.gov.uk	www.experian.co.uk	www.area.classification.org.uk	http://tinyurl.com/6bxy42	www.beacon-dodsworth.co.uk	www.acxiom.co.uk	

LSOA, lower super output area level.
^a Output Areas are the smallest areas for which Census data is published.
^b Some tools leave areas 'unclassified' if they do not fit easily into a category. Examples might include areas with an unstable population, or communal establishments such as halls of residence, prisons and army barracks.

Table 1 presents a comparison of the features of the six geodemographic systems which were used to analyse the health statistics at our disposal.

Indexing

Indexes are widely used in geodemographic analysis to create an understanding of patterns. They make it possible to quantify how under-represented or over-represented a variable is within a base population. In other words, indexes are useful measures of demand.

In this exercise, we have derived the average number of admissions for each available postcode within the HES database across three years (2003/2004, 2004/2005 and 2005/2006). Each postcode was subsequently linked with its corresponding geodemographic type allowing us to aggregate admissions by geodemographic clusters. This was done for each health condition and at all available hierarchies of the different classification systems. Unclassified clusters were not included in the analyses.

The indices for each health condition were calculated by relating the propensity of admissions for such condition by geodemographic cluster with the base population. An index of 100 represents the U.K. national average based on aggregate national data. A value greater than 100 indicates that people living within such neighbourhood have a higher likelihood of been admitted for the health condition while an index lower than 100 indicates a lower likelihood.

Quantifying discriminatory power

The Gini coefficient is often used to measure the degree of concentration of a variable within a distribution of its elements. The coefficient allows a graphical comparison of inequality using a Lorenz curve. Values for the Gini coefficient range from 0, where there is perfect equality, and 1, where there is perfect inequality. The Gini represents an expression of the area located between the line of perfect equality and the Lorenz curve.

The use of this measure and the Lorenz curves has been questioned in certain quarters but Leventhal⁹ describes it as a method which can help

mitigate the challenges posed by numerical methods of comparison. Specifically he suggests numerical methods ‘cannot evaluate the usefulness of a discriminator’ and may not be able to control for the different number of clusters characteristic of different classifications.

In this exercise, we have derived our gains charts by calculating the index as described above. This index, an indicator of ‘need’ has been used to sort the percentage of each health condition and base populations in descending order (see Table 2).

Table 2 A profile report re-ordered for OAC

Index	OAC Super Group	Profile (%)	Base (%)
139	5	13.82	9.95
105	1	17.68	16.83
99	2	4.26	4.29
98	7	12.34	12.65
94	6	17.89	18.99
92	4	22.16	23.96
89	3	11.85	13.33

The profile and base percentage values were subsequently accumulated and used to derive the Gini co-efficient and gains charts. For the Gini, we have adapted the following formula suggested by Brown³.

$$\mathbf{G} = 1 - \sum_{i=0}^{K-1} (\mathbf{y}_{i+1} + \mathbf{y}_i)(\mathbf{x}_{i+1} - \mathbf{x}_i)$$

Where

G is the value of the Gini coefficient

k is the number of data points for the profile and base populations

y is the profile population for a selected geodemographic cluster

x is the base population for the selected geodemographic cluster

Figure 1 Lorenz curve for Mosaic at tier 2



We have also tried to adapt the suggestion of Brown³ about the area between the Lorenz curve and the line of perfect equality. He wrote as follows:

'Defined graphically, the Gini coefficient formally is measured as the area between the equality curve and the Lorenz curve, divided by the area under the equality curve.'

From this suggestion we deduce that the area between the Lorenz curve and the line of perfect equality is half the value of the Gini coefficient.

It has been suggested by Callingham⁵ that the value of this area can be plotted against the number of clusters for each classification system to provide a graphical comparison of discrimination. We have multiplied the value of each area (**A**) by 100 to convert it to a percentage.

Results

One of the major challenges of this exercise is the fact that we are comparing discriminators which do not have equal number of clusters (see Table 1). It may be argued that classification systems with more clusters are likely to provide greater detail of discrimination.

To address this problem we chose to explore the classifications by looking at the mutual interdependence of the different cluster levels. Figures 2, 3, 4 and 5 provide charts which compare the area (**A**) derived from the Gini and Lorenz curves with the number of clusters. For each system, the value of **A** appears to increase with increasing number of clusters.

Figure 2 Asthma admissions: relationship between cluster levels

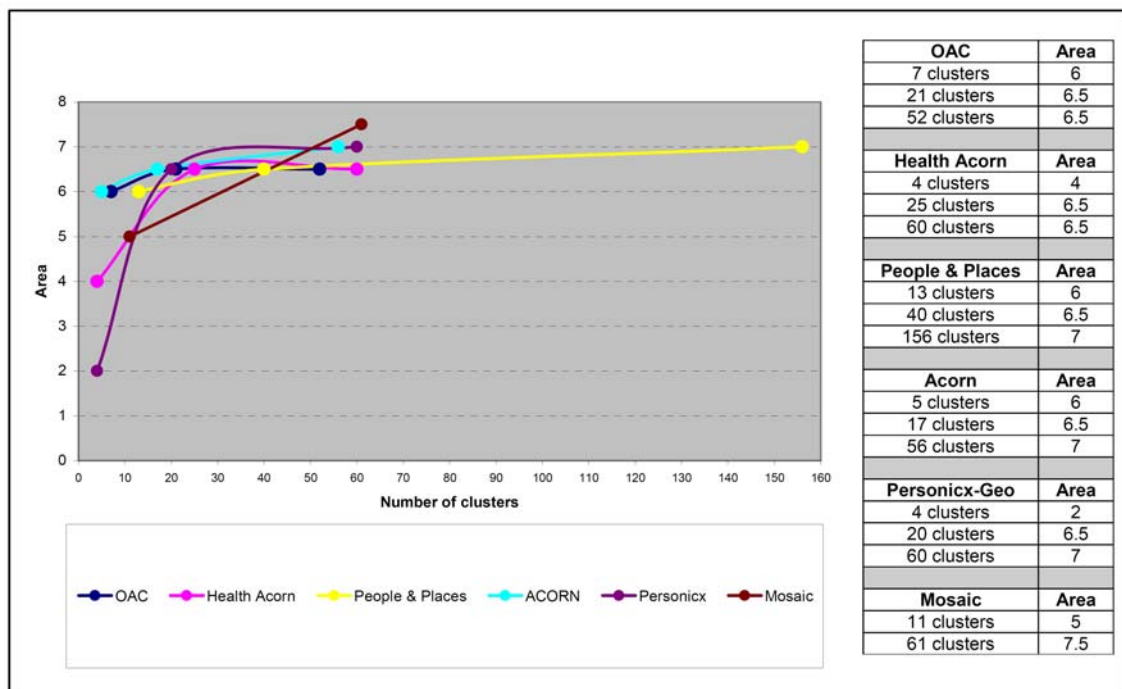


Figure 3 CHD admissions: relationship between cluster levels

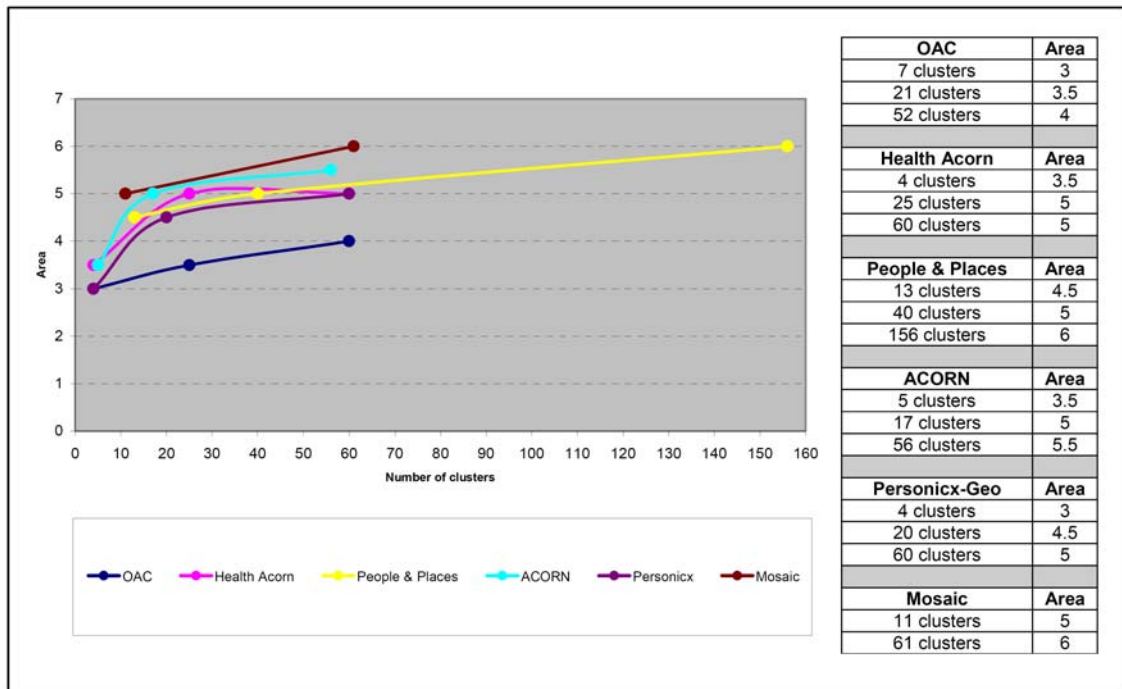


Figure 4 COPD admissions: relationship between cluster levels

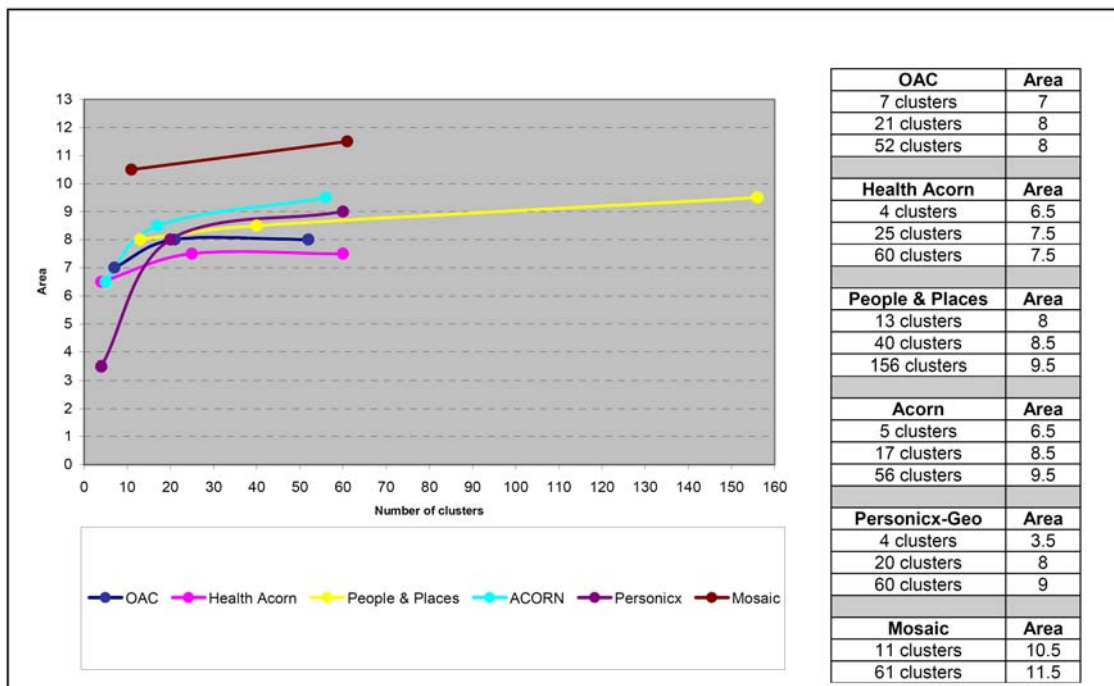
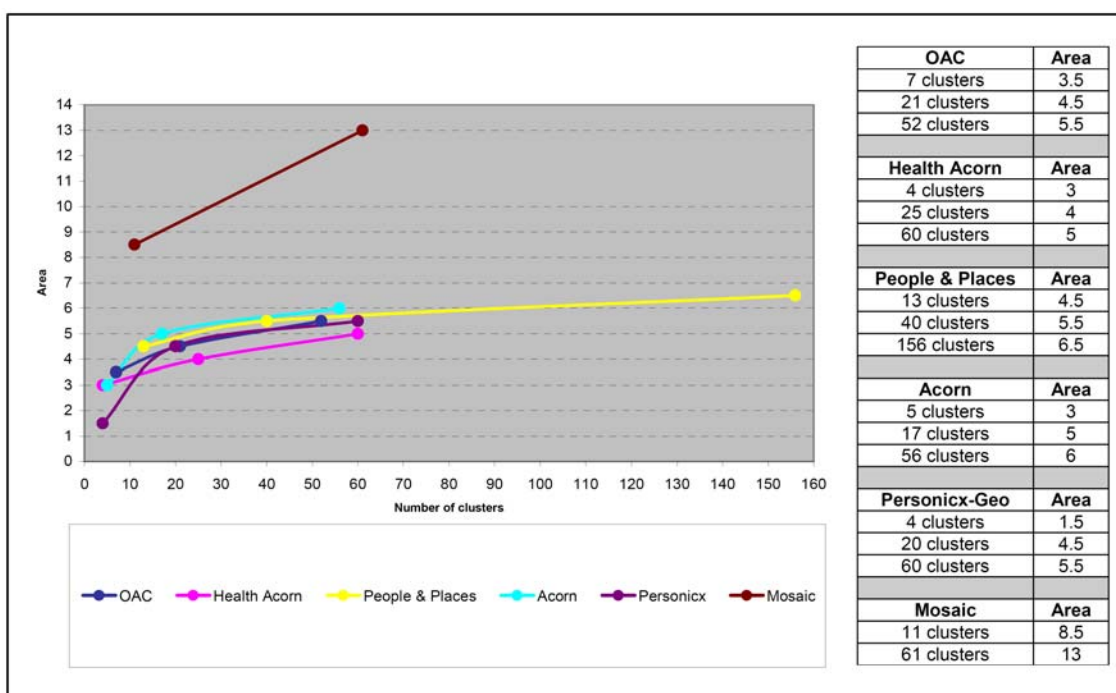


Figure 5 Diabetes admissions: relationship between cluster levels



The other important issue to keep in mind about the number of clusters relates to visualisation. Too many subdivisions make mapping less aesthetically pleasing and difficult to interpret. Martin Callingham, a visiting professor at Birkbeck, University of London and geodemographic expert suggests that an ideal number of subdivisions for hierarchical classifications would be about 6 clusters at the lowest level, about 20 clusters at the next level and about 50 clusters at the highest level of aggregation.

Discussion

As depicted in Table 1, the sources of data which go into these classification systems are varied. Apart from the 2001 census, practically all other sources are life-style related. For obvious competitive reasons, these databases are not made publicly/freely available by the marketing outfits which conduct the

surveys. It is difficult to evaluate the quality of the data sources. The addition of two variables suggesting similar information only results in redundancy.

The scale at which some of the life-style surveys are conducted is also not very clear. Indeed, combining data harnessed at different scales can pose its own problems giving rise to need for generalisations. Such generalisations can result in loss of potentially useful information.

We do not suggest that data from multiple sources reduce the quality of a classification system. Indeed they can provide useful information not covered by questions asked in the Census. However variables should only be included if there is a good reason for their presence and they do not mask important patterns of other variables.^{10, 6} Since we know that the variables which are used to build a system can suggest its primary purpose, we believe that knowledge of these variables can also refine the interpretations of results of analysis conducted with the systems.

We may not be able to fully evaluate the methodology used to arrive at the clusters of many of the commercially available segmentations. However one of the clues that may prove useful in identifying the clustering algorithm is the presence or absence of unclassified groups.

Hierarchical agglomerative methods such as Ward's method, centroid clustering method, within and between group linkage methods have the ability to create multiple solutions.⁶ These methods do not require the specification of the number of clusters from the start of implementing the clustering algorithm.

This means when using these methods, there is a probability that a case (e.g. output area or postcode) may not be allocated to a particular cluster if its behaviour is deemed too different to all the clusters. A short-coming of these methods is their inability to cope with very large datasets.

In the case of iterative relocation methods such as K-means, there is usually the need to specify a particular number of clusters from the beginning. It has the characteristic of allocating every case to a cluster. K-means is about the most popular method employed in the creation of geodemographic systems because of its efficiency and capability to handle large datasets.

The likelihood therefore is that the classifications with unclassified zones (People and Places, Acorn and Mosaic) may have employed a hierarchical agglomerative method while those which have all zones classified (OAC, Health Acorn and Personix-Geo) may have adopted K-means clustering. However, it is possible to use K-means and still have some zones unclassified. If there are a total of n cases and a few zones (x) have been identified as too different or can not be seen to align themselves to any clusters, they can be exempted and the algorithm can be applied to $n - x$ cases.

We have used the Lorenz curves to measure how well the systems discriminate for differences within the population. In Figure 1, the horizontal axis represents the accumulated percentage of the population. These have been ranked in accordance with the percentage of the admissions population for each condition as represented by the curves. Since admissions are unevenly distributed (i.e. inequality exists), the curve shifts away from the diagonal line of

perfect equality. The larger the area between these two lines, the better a system can uncover the differences within the population.

We observe that for each system, discrimination improves with increasing number of clusters. For instance, OAC records values of 3.5, 4.5 and 5.5 for cluster levels one, two and three respectively for diabetes admissions. For the same condition, People and Places records values of 4.5, 5.5 and 6.5 for each of the three cluster levels respectively.

However when compared against each other we observe that the notion that more clusters provide greater discrimination may not always be true. In the case of asthma admissions five of the six systems (excluding Mosaic) record values of 6.5 at the second level of aggregation even though the number of clusters for each system at this level of aggregation varies. People and Places has 40 clusters and Acorn has 17 clusters. If these two systems can discriminate for the same health condition at equal magnitude and the number of clusters contained in one is more than twice the other, then it makes sense to opt for the system with fewer clusters (for that particular health condition) to reduce the complexity of mapping and interpretation.

One of the systems (Mosaic) seemed to be consistent in its high discriminatory power relative to the others for conditions which appear to be prevalent amongst older populations. This was however not the case for asthma admissions. For instance at the first level of aggregation, OAC comprises 7 clusters resulting in a discriminatory value of 6 while Mosaic's 11 clusters account for a value of 5. This raises a question – 'Is Mosaic discriminating for

age'? To address this question, the entire exercise may have to be conducted using age standardised populations.

Conclusions

We have provided a review of the potential of geodemographic classification systems for health. We have also tried to evaluate the usefulness of these systems by examining their discriminatory power.

Our findings suggest that no one system supersedes the other as different systems have their benefits and/or shortcomings. In addition to how well systems uncover inequality, the use to which they are to be put, knowledge of input variables and other embedded analytics can be considered when deciding which system to use. Future directions to this work will consider using age standardised and possibly income (deprivation) populations to investigate the patterns of hospital admissions, discriminatory and predictive power.

References

1. Ashby DI, Longley PA. Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS* 2005; 9: 53-72.
2. Batey PWJ, Brown PJB. The spatial targeting of urban policy initiatives: a geodemographic assessment tool. *Environment and Planning A* 2007; 39: 2774-2793.
3. Brown MC. Using gini-style indices to evaluate the spatial patterns of health practitioners: theoretical considerations and an application based on Alberta data. *Social Science Medicine* 1994; 38: 1243-1256.
4. Brown PJB, Hirschfield AFG, Batey PWJ. Adding value to census data: public sector applications of super profiles geodemographic typology. *Journal of Cities and Regions* 2000; 10: 19-32.
5. Callingham M. The application and further development of OAC; 2006. Available at: (http://www.rss.org.uk/rssadmin/uploads/577303_Callingham%20Descriptions%203%20Oct%202006.pdf). (Last accessed 5/7/2007).
6. Everitt BS, Landau S, Leese M. *Cluster analysis*. 4th ed. London: Arnold; 2001.
7. Everitt BS. *Cluster analysis*. 3rd ed. London: Arnold; 1993.
8. Harris R, Sleight P, Webber R. *Geodemographics, GIS and neighbourhood targeting*. London: Wiley; 2005.
9. Leventhal B. Evaluation of geodemographic classifications. *Journal of Targeting, Measurement and Analysis for Marketing* 1995; 4: 173-183.
10. Milligan GW. Clustering validation: results and implications for applied analysis, In: Arabie P, Hubert LJ, De Soete G, (eds). *Clustering and classification*. Singapore: World Scientific; 1996.
11. Novak TP, Leeuw J, Macevoy B. Richness curves for evaluating market segmentation. *Journal of Marketing Research* 1992; 19: 254-267.
12. Shelton N, Birkin M, Dorling D. Where not to live: a geo-demographic classification of mortality for England and Wales, 1981-200. *Health and Place* 2006; 12: 557-569.
13. Singleton A, Davidson-Burnett G, Longley P. *University market area analysis for widening participation*. Cardiff. CEBE; 2007.
14. Todd P, Bundred P, Clarke JRE, Brown PJB, Forbes H. *GIS in health care planning: locating cancer treatment centres*. URPERRL working paper 41. Liverpool; 1994.

15. Vickers D, Rees P. Introducing the area classification of output areas. *Population Trends* 2006; 125: 15-29.
16. Webber R, Butler T. Classifying pupils by where they live: how well does this predict variations in their GCSE results? *Urban Studies* 2007; 44: 1229–1254.