Geodemographic segmentation systems for screening health data

Stan Openshaw, Marcus Blake

Abstract

***Aim* – To describe how geodemographic segmentation systems might be useful as a quick and easy way of exploring postcoded health databases for potential interesting patterns related to deprivation and other socioeconomic characteristics.**

***Design and setting* – This is demonstrated using *GB Profiles*, a freely available geodemographic classification system developed at Leeds University. It is used here to screen a database of colorectal cancer registrations as a first step in the analysis of that data.**

***Results and conclusion* – Conventional geodemographics is a fairly simple technology and a number of outstanding methodological problems are identified. A solution to some problems is illustrated by using neural net based classifiers and then by reference to a more sophisticated geodemographic approach via a data optimal segmentation technique.**

(*J Epidemiol Comm Health* 1955;49(Suppl 2):S34–S38)

Geodemographic classifications are widely used in many different application areas as a means of obtaining a useful descriptive summary of the principal types of residential areas that exist in the UK, based on a multivariate classification of census data.[1] Brown[2] writes, "Geodemographics has come into popular use as a shorthand label for both the development and application of area typologies that have proven to be powerful discriminators of consumer behaviour and as aids to market analysis". A typical geodemographic classification starts with 1991 census data for all the 145 716 census enumeration districts in Britain for which there are about 10 000 different census counts available (see ref [3] for a complete listing). These 10 000 variables are reduced by careful selection to about 50 to 100 composite indicator variables measuring a range of socioeconomic, demographic, and housing characteristics. This set of derived variables are then used to classify the census enumeration districts into clusters (or groups) of similar types of areas based on their multivariate data profiles. Current commercial geodemographic classifications contain between 10 and 161 residential area types; for example, the *ACORN91* system has 6 categories, 17 groups, and 54 types; the *MOSAIC91* system has 12 groups and 54 types; and the *SuperProfiles94* system 10 lifestyles, 40 target markets, and 161 clusters. Each of these clusters is given a simple descriptive label that offers an idealised and highly stylised portrait (or picture) of what the typical members are like. The resulting classification is then usually linked to postcodes via the *OPCS/GRO*(s) census to unit postcode directory.

In essence, geodemographics provides a means by which people can be characterised by the types of area in which they live, using postcodes as a simple indexing mechanism to a multivariate classification of small area census data. This is potentially relevant to health analysis because a geodemographic classification of this type might well be regarded as a more sophisticated approach to incorporating deprivation related effects than that provided by the more traditional ranking of index values for small areas such as wards (for example, see Jarman[4] and various other indexes of deprivation developed for the Department of Environment and Bradford *et al*[5]). Geodemographics might well be more useful because it offers a strongly multivariate view of the characteristics of areas and being census enumeration district based may well also provide a higher level of geographic discrimination. They also offer a proxy for lifestyle and prosperity, they act as a substitute for census data, and generally provide a quick, albeit crude, means of adding a census based socioeconomic, demographic, and housing context to virtually any postcoded health data. The results are also simple enough to be understood by nontechnical experts.

It is suggested that geodemographics are relevant to health database analysis because they offer a particularly quick and easy way of performing a broad brush screening of medical data in terms of different types of residential area. They offer answers to questions such as the following. Do disease rates vary by type of residential neighbourhood? Do poor housing areas with high levels of unemployment have significantly higher than expected mortality or morbidity rates? What sorts of residential area in what parts of the UK are associated with the highest incidence of a particular disease? An example of this type of analysis is that of Reading, Jarvis, and Openshaw.[6]

It is argued, however, that geodemographics does more than merely offer a convenient source of covariate information that statistical modellers can use to remove so called confounding socioeconomic effects. Health differences that vary in relation to the type of residential area, after allowances are made for age and sex effects, should perhaps become the focus of attention since (in theory) they may be treated or managed by social rather than purely medical means. They may also have

School of Geography, Leeds University, Leeds LS2 9JT
S Openshaw
M Blake

Correspondence to:
Professor S Openshaw.

political importance as a reflection of spatial inequalities in life expectancy and well being. It would seem, therefore, that a routine geodemographic style of screening of important health and death databases might be an extremely useful initial exploratory step in monitoring health information systems in both epidemiological research and for assessing the health needs of populations. It is perhaps surprising that much greater use is not as yet routinely being made of this technology. The reasons for this relative neglect probably reflect the cost of acquiring a commercial geodemographic system, unfamiliarity, and lingering concerns about the quality of some of the commercial products.

This paper briefly describes the development and application of a particular geodemographic segmentation system which was produced as a part of an ESRC funded project and is available free for academic researchers. This paper illustrates the use of this system and then outlines how it can be further developed to generate robust, safe, data optimal segmentations of virtually any postcoded medical data.

Spatial classification of 1991 census data

The statistical technology needed to create a crude national classification of small area census data is now widely diffused via popular statistical packages; for example, *SAS* or *SPSS*. However, it is not often recognised that many of the available classification methods date from the 1960s and are not well suited to handling the special nature of spatial (as distinct from non-spatial survey) data; in particular, problems of non-normal distributions, non-linearity, and spatial dependency are endemic. Census data also introduce a number of additional difficulties; especially those related to small number effects and spatially varying levels of data precision. Whether these matter depends on the nature of the application, on the level of skill used to develop a classification, and the context in which it is used. Better still would be the use of classifiers that at least attempt to handle some of the problems.

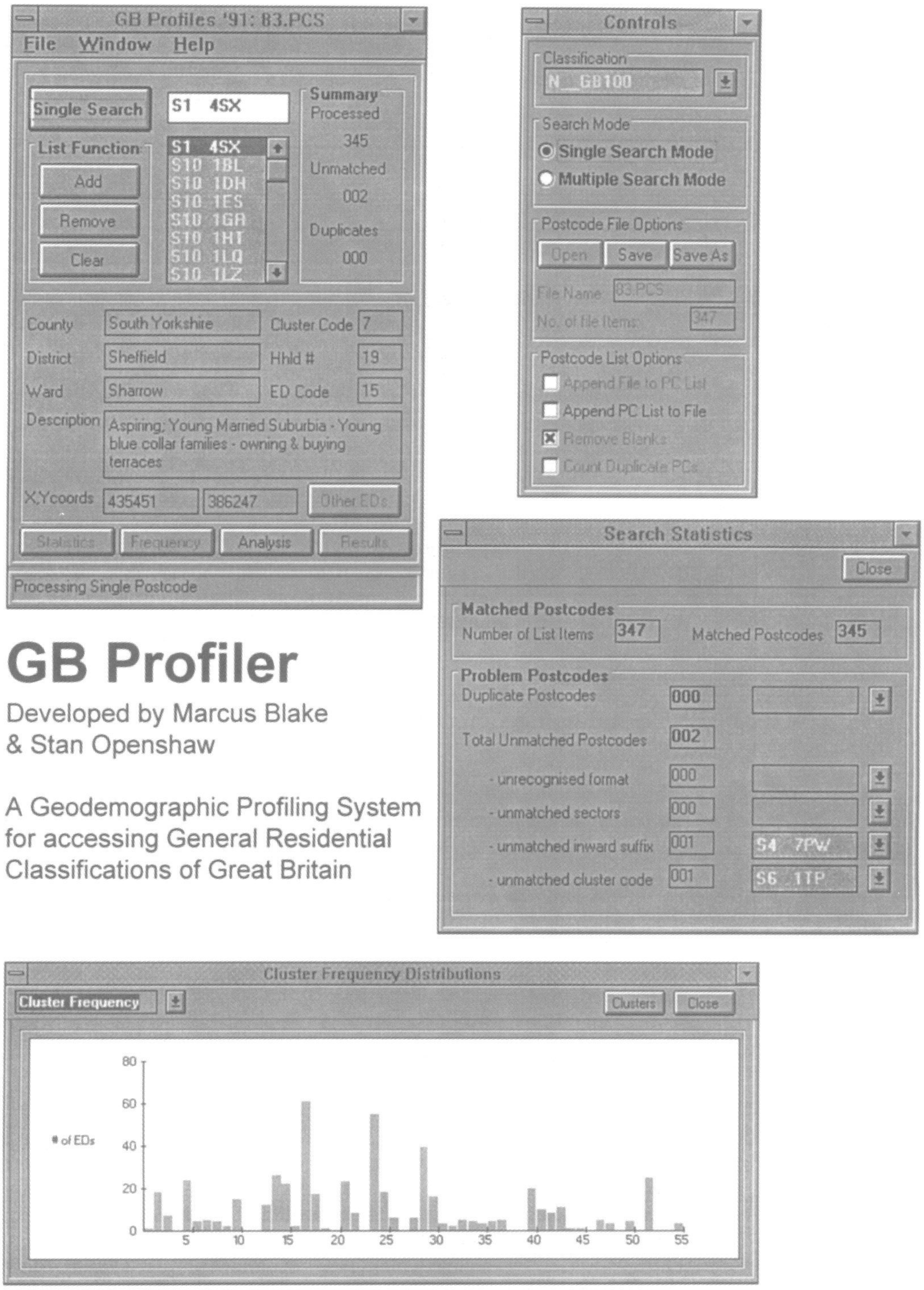
Accordingly, the classification method used here is based on a particular type of unsupervised neural net known as Kohonen's self organising map.[7] This has been modified to handle the data uncertainty present in census data.[1 8 9] The attractions of this method include its simplicity and flexibility. It can handle noisy census data and size related data precision issues, and there is a minimum amount of data preprocessing. However, as with all neurocomputing approaches, it is always useful to compare the results that are obtained with more conventionally produced classifications in order to provide performance benchmarks, against which any improvements can be assessed. This naturally leads to having not one but multiple classifications based on different methods, perhaps also using different sets of variables, and offering different levels of data generalisation obtained by varying the numbers of clusters present. The user is expected to choose what is best in the context of a particular data analysis application.

This may be contrasted with a more traditional, commercial geodemographic approach, based on a single all purpose system which is applied to virtually everything. While of general descriptive utility and of considerable simplicity, it should be appreciated that particular applications may require either their own bespoke classifications, tuned to specific requirements (although this tuning is a very imprecise process), or a means provided of choosing the best from among many alternative classifications. A customised or tailored segmentation system is one in which a purpose specific classification is developed to meet the specialist need of a particular application.[10] There is no reason why such systems can not be developed and optimised for use in a health context. The customisation process may involve the selection of a specific set of variables to be used and the careful choice of the best number of clusters relevant to a specific context. This can be a lengthy and costly process. Fortunately, it seems that the principle unknown, but critical, variable in the classification process is often the best number of clusters to use. Research suggests that it may well even be more important than the choice of classifier as a sensitivity optimising device.[11] There has been a tendency in a marketing context to only seek highly parsimonious geodemographic segmentations with few clusters, whereas some health applications may well require 5 to 10 times as many in order to allow place or region specific variations of mortality and morbidity rates to appear.

The *GB Profiles* geodemographic system

Research performed in the School of Geography at Leeds has produced a series of over a 100 different 1991 census data based geodemographic systems designed solely for academic research usage. The licence that makes the 1991 census data available to academics prohibits commercial or non-university based applications. This restriction made it feasible to develop census classification systems without being hindered by any market factors; for example, similarly to previous products or constrained by conventional geodemographic practices. The resulting *GB Profiles* system,[12] based on the best possible available technology, is designed to use broadly representative census data, to offer multiple classifications at varying levels of resolution, packaged so that it is easy to use, and is freely available for academic research and teaching purposes.

The *GB Profiles* system runs under both PC and UNIX environments. The Microsoft *Windows* PC-based system provides easy access to a whole series of census classifications with variable levels of resolution; ranging from 2 to 5000 clusters. However, to keep matters simple, the system currently restricts external users to four specific classifications (with 10, 49, 64, and 100 clusters) derived by both the neural net classifier and a more conventional K-means method (see [1] for details). These are



Single Search: S1 4SX

List Function

S1 4SX
S10 1BL
S10 1DH
S10 1ES
S10 1GA
S10 1HT
S10 1LQ
S10 1LZ

Summary
Processed: 345
Unmatched: 002
Duplicates: 000

County: South Yorkshire
Cluster Code: 7
District: Sheffield
Hhld #: 19
Ward: Sharrow
ED Code: 15
Description: Aspiring; Young Married Suburbia - Young blue collar families - owning & buying terraces
X,Ycoords: 435451 386247

Processing Single Postcode

GB Profiler

Developed by Marcus Blake & Stan Openshaw

A Geodemographic Profiling System for accessing General Residential Classifications of Great Britain

Classification: N_GB100

Search Mode
● Single Search Mode
○ Multiple Search Mode

Postcode File Options
File Name: 83.PCS
No. of file items: 347

Postcode List Options
□ Append File to PC List
□ Append PC List to File
☒ Remove Blanks
□ Count Duplicate PCs

Matched Postcodes
Number of List Items: 347
Matched Postcodes: 345

Problem Postcodes
Duplicate Postcodes: 000
Total Unmatched Postcodes: 002
- unrecognised format: 000
- unmatched sectors: 000
- unmatched inward suffix: 001 S4 7PW
- unmatched cluster code: 001 S6 1TP

Cluster Frequency

Figure 1 A mosaic of the GB Profiles *system.*

provided with a full set of cluster labels and are designed to provide a broadly based description based on the available range of 1991 census data. It was thought that this would appeal most to the majority of potential social science and other research users. An easy to use Microsoft *Windows* front-end offers the user the ability to select a classification, attach cluster codes to a postcoded data file, export the results in one of several different formats and perform some preliminary analysis. It is also possible to investigate the cluster labels that are used to represent each area type, and, if appropriate, to change them. Figure 1 presents a mosaic of some of the *GB Profiler* screens.

A file of postcoded cancer registrations was run through *GB Profiler* to illustrate some of the benefits and problems of a geodemographic approach. Postcoded data were available for the 1622 colorectal cancer registrations that occurred in Sheffield between 1979 and 1983. The 100 cluster neural net based classification was applied to these data to identify which residential area types were associated with the highest disease incidence, see table 1. Here attention is restricted to areas with a 40% higher than expected cancer incidence. These areas are mapped in figure 2. The geodemographic labels associated with the cluster numbers suggest that most of these high in-

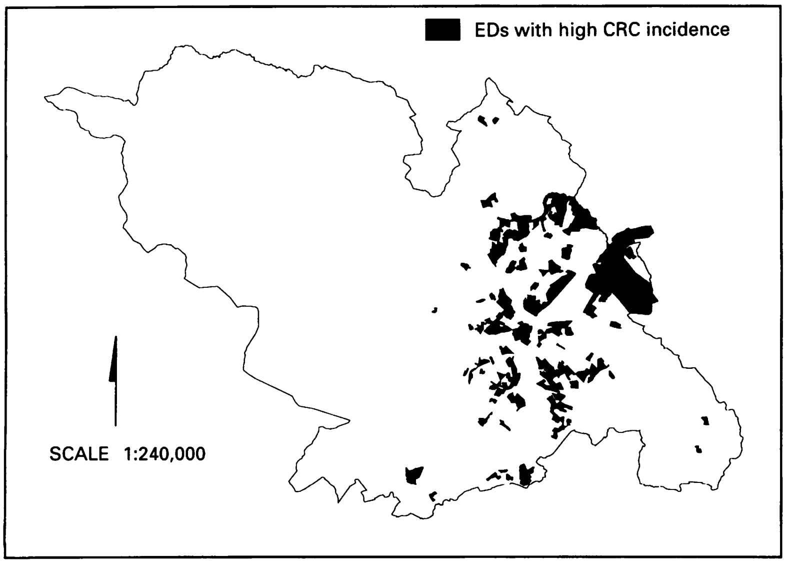


Figure 2 Sheffield District Health Authority showing areas of high colorectal cancer (CRC) incidence identified by the 100 cluster neural net classification.

Table 1 Residential cluster types with the highest disease rates

Cluster no	*Observed cases*	*Expected cases**	*Index value†*
34	2	1	277
10	5	2	261
94	11	4	225
71	7	3	207
97	7	3	205
62	4	2	200
80	13	6	180
64	5	2	175
15	45	22	173
49	44	23	163
27	20	11	156
87	13	7	152
22	20	12	145
28	10	6	144
79	6	4	144
16	66	40	142

* This is an age-sex estimate of the expected numbers of cases.
† 100 is the average for Sheffield.

cidence areas are relatively poor. These are characterised as either struggling, unemployed families and single parents living in council housing. This is a quick and simple way of identifying the type of person and the kind of residential areas that have high numbers of, in this case colorectal cancer registrations. This descriptive information would be of use as part of the larger picture of health needs assessment and disease monitoring. However, this is a very basic approach that may well be too simple to provide completely reliable results.

A data optimal segmentation system

The example presented here is a very brief but not untypical illustration of how a geodemographic approach would be used in health analysis. There are a number of potential problems: (1) the choice of classification, (2) the best number of clusters to use, (3) problems of an ecological fallacy nature, and (4) possible small number problems that render the results uncertain. With *GB Profiles*, the greatest immediate source of uncertainty is which classification and how many clusters to use. Select too few clusters and the results might well be over generalised and important associations lost; select too many and the results might be spurious due to small number effects. This dilemma between "too few" and "too many" clusters is problem dependent and thus data specific. In a highly descriptive preliminary data screening exercise this may not matter. However, there is a world of difference between geodemographics as applied to "junk mail" in the commercial sector and its use with health database analysis of a more critical nature where higher standards should apply and the problems deserve more explicit consideration.

One way forward is to develop a more sophisticated geodemographic approach. We have outlined what is termed a data optimal segmentation system that might be developed.[11] This so called "geodemographic targeting machine" (GTM/1) attempts to identify the best classification to use and within it the best set of clusters so as to maximise coverage of the data and minimise problems due to small number effects. The GTM/1 evaluates a number of different geodemographic classifications by using a mix of Monte Carlo significance testing and boot-strapping to delete both unreliable classifications (that is, those yielding results little better than a random classification would) and also to delete clusters within acceptable classifications for which the results appear to be either highly uncertain (due to small number effects) or not particularly interesting (in terms of predefined performance benchmarks). In essence, GTM/1 is an optimisation procedure that evaluates a set of different geodemographic classifications to find that which captures the largest number of cases in clusters that meet user defined constraints.

The GTM/1 approach is illustrated by reanalysing the colorectal cancer data. A selection of 33 different classifications are examined covering a range from 10 to 5000 clusters. The following segmentation constraints are set:

- A minimum cluster size of 10 cases,
- A minimum cancer incidence 40% greater than expected taking into account age and gender factors, and
- Results significantly different from random.

Clearly these are arbitrary and can be readily changed as total run times are less than five minutes on a UNIX workstation.

The results are reported in table 2 and mapped in figure 3. It is very interesting that five of the classifications were dropped because they produced results that were not significantly different from random. The previous 100 cluster classification looked good in table 1 but did in fact only capture 278 cases compared with the 337 in the 25 cluster classification which produced the best results in table 2. Moreover, if the small and unreliable clusters are removed, then the 100 cluster classification only captures 219 cases. The labelling of the clusters in this 25 cluster classification would be performed using automatic labelling software and is the subject of continuing research. In general, the results reported here again identify poor housing areas, but with more precision than previously.

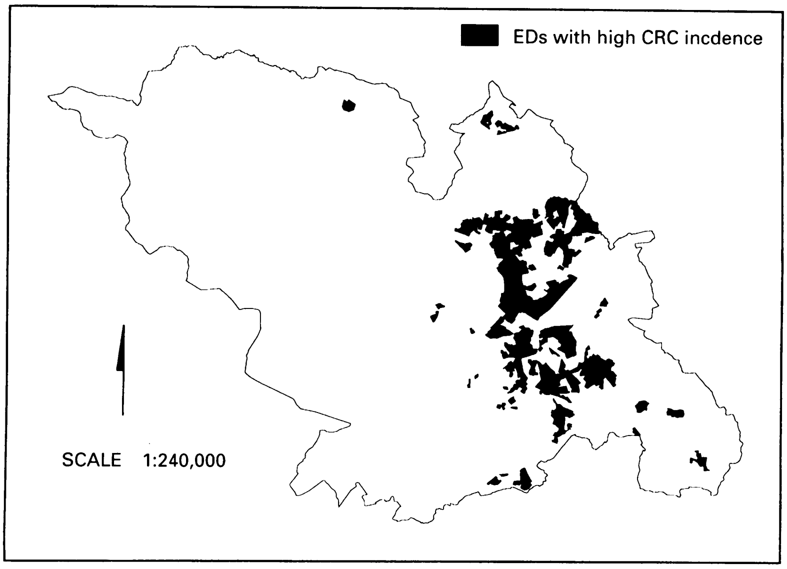


Figure 3 Sheffield District Health Authority showing areas of high colorectal cancer (CRC) incidence identified by the 25 cluster neural net classification.

Table 2 Data optimal segmentation results

*Clusters in classification**	*Clusters in segmentation†*	*No of observed cases*	*Expected no*
25	6	337	200
20	4	324	200
10	2	283	179
35	5	269	153
50	4	268	158
45	5	255	144
60	5	245	132
40	4	235	130
100	7	219	118
15	3	205	113
150	8	204	113
55	4	198	113
130	6	195	102
400	8	192	101
300	9	189	92
90	6	179	92
120	7	172	91
30	5	159	83
500	10	156	78
750	9	155	71
10	2	153	80
140	6	142	68
70	5	139	64
2000	9	137	62
49	4	113	61
4000	5	90	44
3000	7	82	41
5000	4	50	21

* The classification are ranked in descending order of the number of cases they represent.
† Clusters are deleted because they do not occur in the study region or they are too small or yield unreliable results.

Clearly there are all manner of interesting "trade-offs" that can occur between the choice of targetting constraints and the data being analysed. Nevertheless, it is quite clear that this optimal segmentation technology has considerable potential relevancy in a health context. A particularly nice feature is its ability to detect and reject random results.

Conclusion

Geodemographic classifications provide a useful descriptive tool for the analysis of health data. The data optimal segmentation system goes further and provides a simple and quick means of exploring health databases for potential interesting associations with residential area characteristics. The process is automated and designed to be intrinsically safe. With the large postcoded health databases that exist such as those kept by the cancer registries, the ability to screen quickly and easily these data for interesting patterns is an important need that the geodemographic segmentation systems described here could be used to meet.

The research was partly funded by an ESRC research grant (R000234436). The cancer data was provided by Sheffield District Health Authority as part of a MRC funded PhD project at Sheffield University. The analysis is based on census and digital boundary data provided with the support of the ESRC and JISC and are copyright of the Crown and/or Post Office. Information about how to obtain the *GB Profile* software and associated 1991 census classifications can be obtained by emailing; *marcus@geog.Leeds.ac.uk*, or *stan@geog.leeds,ac.uk*. See also the World Wide Web page: *http://www.geog.leeds.ac.uk/staff/m.blake/profile/hlp0-toc.htm*

1 Openshaw S, Wymer C. Classifying and regionalising census data. In: Openshaw S, ed. *Census users handbook*. London: GeoInformation International, 1995;239–70.

2 Brown PJB. Exploring geodemographics, in I. Masser and M Blakemore (eds) *Handling Geographical Information: methodology and potential applications*. Longmans: London, 1991;221–58.

3 Openshaw S. *Census users handbook*. London: GeoInformation International, 1995.

4 Jarman B. Identification of underprivileged areas, BMJ 1993; **286**:1705–9.

5 Bradford MG, Robson BT, Tye R. Constructing an urban deprivation index: a way of meeting the needs for flexibility. *Environment and Planning A*. 1995;**27**:519–33.

6 Reading R, Jarvis S, Openshaw S. Measurement of social inequalities in health and the use of health services among children in Northumberland, *Arch Dis Child* 1993;**68**: 626–31.

7 Kohonen T. *Self-organization and associative memory*. Berlin: Springer-Verlag, 1984.

8 Openshaw S. Neuroclassification of spatial data. In: Hewitson DC, Crane RG, eds. *Neural nets: applications in geography*. Boston: Kluwer, 1994;53–70.

9 Openshaw S, Blake M, Wymer C. Using neurocomputing methods to classify Britain's residential areas. In: Fisher P ed. *Innovations in GIS 2*. London: Taylor and Francis, 1994;97–112.

10 Openshaw S. Special classifications. In: Leventhal B, Moy C, Griffin J eds. *An introductory guide to the 1991 census*. Henley: NTC Publications 1993;69–82.

11 Openshaw S. Developing smart and intelligent target marketing system. *Journal of Target Marketing, Measurement and Analysis for Marketing*. 1994;**2**:289–301 and **3**:31–8.

12 Blake M, Openshaw S. *GB Profiler: A Windows front end for GB Profiles*. Leeds: School of Geography, University of Leeds. Working paper (in press).

Open discussion

BEN SHLOMO – Did you show post hoc classifications based on the best fit of the data, or who decided that that was the way you were going to classify those different groups?

OPENSHAW – Classifications were based on the census characteristics of the clusters. This was a labelling exercise. It was designed to reduce a set of 80 variables that would have different scores on all the enumeration districts in this particular cluster and come up with a label understandable to others as being broadly representative.

DOLK – Similar to an *ACORN* classification?

OPENSHAW – Yes. Most people identifying numbers of clusters do so by pretending they can detect break points in plots that show how within cluster variation diminishes as the number of clusters increases. Quite often moreover, no real break points can be found. There is usually a smooth graph. What I have done is to say "why not have multiple classifications?" The 33 classifications I looked at went from 10, 20, 30, 40, 50, 60, 70, 80, 100, 200, 300, and so on up to 5000 clusters and then let the data optimal segmentation system figure out what is best for your particular application. I think that neatly gets around the problem of having to identify an optimal break point on smooth graphs when none can be seen.