An introductory review of geodemographic information systems

Received (in revised form): 1st March, 2003

Peter Sleight

is a leading authority in the use of geodemographics and lifestyle data. He spent his first ten years in FMCG marketing, with Cavenham, Unicliffe and RHM. He became professionally interested in geodemographics while marketing director at Allen, Brady and Marsh, joined CACI as head of consultancy, and later helped to set up Pinpoint Analysis. Since 1991, he has run Target Market Consultancy, providing expert advice on customer targeting and geodemographic analysis.

Abstract This paper is based on Chapter Three of the book 'Targeting customers — How to use geodemographic and lifestyle data in your business' by Peter Sleight. It provides a straightforward introduction to using geodemographic information systems. It reviews the rationale for neighbourhood classifications, how they are constructed, the arrival of the new census data and the implications for new products. Finally, it reviews the neighbourhood issues implicit in the assumption of such classifications. The third edition of 'Targeting customers' is published to coincide with the new system launches following the release of the 2001 Census data. The book (published by World Advertising Research Center, www.warc.com) was published in March 2004.

WHY NEIGHBOURHOOD CLASSIFICATIONS?

The broad definition of geodemographics (any use of demographics within units of area) implies that one may use 'raw' census variables within (say) postcode sectors; for example, age bands, type of housing, etc. Indeed, some applications use demographics in just this way. One of Pinpoint's first big projects, 20 years ago, identified areas with large concentrations of children by ranking census enumeration districts (EDs) on this one criterion.

Most commercial users of geodemographics, however, tend to use neighbourhood classifications (ACORN, MOSAIC, etc). Why should this be?

Why has this methodology become pre-eminent? In the author's view, the answer comes down to one thing — the ability to profile, which comes hand-in-hand with neighbourhood classifications. Customers, areas and market research surveys can all be profiled — this makes the technique easy to use and leads to all the applications which users take for granted. The use of 'raw' variables is fine if it is only areas that need to be profiled; but individuals may only be profiled by using a classification. Some of the issues involved are covered later in this paper, as are the linking mechanism of areas to addresses, and the methodology by which

Peter Sleight

Partner, Target Market Consultancy, Woodlands, Woodlands Close, Holmer Green, Bucks HP15 6QG, UK.

Tel: +44 (0)1494 712 371;
Fax: +44 (0)1494 714 203;
e-mail:
peter.sleight@btinternet.com

Table 1: Customer profile for product X in Great Britain

PiN group	Profile count	Profile (%)	Base (%)	Index
A	150	7.5	5.1	147
C	1,040	52.0	14.1	369
D	750	37.5	10.0	375
F	60	3.0	17.0	18
Totals	2,000	100	46.2	

Note: This is slightly unrealistic, in that all the customers are in four of a total of 12 PiN groups; nevertheless, it shows the principle of profiling clearly.

neighbourhood classifications are constructed.

The linkage of neighbourhood classifications to addresses enables individuals (or customers, or research respondents) to be profiled by those classifications. Profiling is simply a comparison exercise. It can be illustrated with a simplified example. For reasons of pure nostalgia, PinPoint Identified Neighbourhoods (PiN) is used at its 12-fold ‘group’ level, and for illustration, just four of the 12 PiN types are used:

- A — Rural
- C — Upwardly mobile young families
- D — Affluent households
- F — Suburban middle-aged or older.

Table 1 is fairly typical of the layout of profiles provided by geodemographics agencies. It is a customer profile for a fictional product.

To explain the columns in Table 1:

- (a) ‘profile count’ is a count of the customers in each of the PiN types. This will have been derived automatically by the profiling system, by analysing customer postcodes;
- (b) ‘profile %’ is the percentage represented by each PiN type of total customers; these numbers always add up to 100;
- (c) ‘base %’ is the percentage

represented by each PiN type of the base area, in this case Great Britain (GB). Most profiling systems allow the base to be specified;

(d) ‘index’ is simply an expression of ‘Profile %’ divided by ‘Base %’, multiplied by 100. If the index = 100, this would indicate that the percentage of customers in that PiN type was exactly equal to the percentage of population in that PiN type, in GB, ie on the national average.

So, in this example, there are nearly one-and-one-half times as many PiN type ‘A’ customers (147 on the index) as one would ‘expect’; there are three-and-three quarters as many PiN type ‘D’ customers (375 on the index) as there would be if the customers were quite typical of the population.

So that is the principle of geodemographic profiling — comparing the penetration of (say) customers within each PiN type, with the base area.

Neighbourhood classifications are the result of multivariate analysis techniques (many separate ‘variables’ go into their solution), so the neighbourhoods thus classified can be described in terms of all the variables that were used to construct them. Typically, they can be described in terms of ‘housing’ variables (number of rooms, type of tenure, amenities, etc); demographics of the people living there

(age profile, country of origin, family or household composition, qualifications) and socioeconomics of the inhabitants (social class or socioeconomic group, industry of occupation, etc). Thus a very rich description of the type of neighbourhood concerned, for each neighbourhood type, is held. This 'multifaceted' picture of the classifications is enriched still further by relating them to industry research, to flesh out the behavioural characteristics of the people who live in each neighbourhood type.

HOW NEIGHBOURHOOD CLASSIFICATIONS ARE CONSTRUCTED

As the geodemographics industry has grown, the methodologies have evolved, until now the details of each proprietary system are quite different. They all, however, share a basic clustering methodology. For the purposes of this discussion, the 'classic' methodology which uses census data alone is described.

All the neighbourhood classifications developed for the UK market to date are the result of performing cluster analysis. Purely census-based classifications involve clustering census 'neighbourhoods'; in the past, this has involved census EDs, but for developments in late 2003 (using 2001 Census data), the neighbourhoods in question will be census Output Areas (OAs). These terms will be explained shortly.

Readers will probably be familiar with cluster analysis — a means of grouping together similar things and thus differentiating them from things that are less similar. For a census-based neighbourhood classification, census neighbourhoods are subjected to cluster analysis. The technicalities of this are beyond the scope of this paper; readers wishing for more detail should consult the references at the end.

THE NATURE OF CENSUS DATA

The census neighbourhoods which will form the 'raw materials' for 2001-based neighbourhood classifications are OAs. Previously, enumeration districts (EDs) have served this purpose. While EDs averaged about 170 households, OAs are considerably smaller at 100–125 households each (Scotland is different — OAs there average 50 households each). The Office for National Statistics (ONS) has used geographic information system (GIS) techniques extensively in planning the 2001 Census. In outline, first the Ordnance Survey Addresspoint locations of each address were used, in combination with their postcode affinity. Then, postcode boundaries were constructed and these unit postcodes were used as the building blocks for OA design. ONS decided to separate out census collection geography from output geography — EDs were used for collection, much as before, but output geography was actually designed after the Census forms had been collected and processed. This fact — and the extensive use of GIS techniques — meant that OAs could be designed to be more 'fit for purpose' than EDs were. For example, OAs have been 'controlled' for maximum and minimum household and population numbers, shape and homogeneity (the latter only to a degree, based on tenure and dwelling type — but still a desirable consideration, particularly for producing neighbourhood classifications).

An automatic system for aggregating unit postcodes was developed by Professor Dave Martin of Southampton University; the system took account of the design criteria mentioned above, and went through a procedure which started with a random allocation of postcodes to OAs, checked the result against the design criteria, and swapped postcodes between adjacent OAs until an optimal

solution emerged. This process is illustrated in Figure 1, which is taken from Dave Martin's paper 'Output Area Construction'.¹

The most useful census data for producing neighbourhood classifications are Census Area Statistics, or CAS (post-2001; prior to that, Small Area Statistics, or SAS, were roughly equivalent). CAS are available at OA level, and constitute a vast data resource. There are 188 tables; 118 of these are the familiar cross-tabulations (see Table 2), the remaining 70 are univariate tables, which as their name implies, each contain information about a single variable (such as country of birth) and tend to be more detailed than cross-tabulations can be. There are some 15,000 counts, and over 3.2 billion statistics, in the CAS database.

The counts in the table — which are statistical counts, relating to each individual OA — are represented by the cell numbers, so this one simple table generates 225 counts. Some of them will be zero — and in total, many of the counts in the total national CAS data matrix will be zeros.

The reason for the huge data volume within CAS is this cross-tabulation within the tables. The census form itself asks a relatively short list of basic questions — but from these questions, this voluminous output is derived. In addition to cross-tabulating variables, the census output includes many derived variables, for example, socioeconomic classifications. For the first time, the 2001 Census includes social grade (ABC1C2DE), much used in marketing and market research, of course. This is a valuable innovation (the relevant people within the Market Research Society (MRS)'s Census and Geodemographics Group must be thanked for initiating this, devising the algorithm and championing the cause).

THE 'NEIGHBOURHOOD' ISSUE

There is one crucial point which must be grasped when using geodemographics — these data are published about census neighbourhoods, and not about individuals. The reason for this is to protect personal privacy, and indeed, the people who work for the Census Offices (ONS for England and Wales, GROS for Scotland, and NISRA for Northern Ireland), must sign the Official Secrets Act and exercise great diligence in protecting privacy. When publishing the CAS output, the Offices examine the data for each OA and either suppress or disguise any fact which might lead to the identification of any individual person or household. So privacy is jealously safeguarded; the other side of the coin is that the CAS data can only be used at a neighbourhood level, and targeting of individuals is not possible. Thus, to put the case slightly differently, a vast amount can be known about the neighbourhood within which an individual resides, but nothing specific can be known about the individual.

This is either a problem or an opportunity, depending upon what the marketer is trying to do. If the marketer is planning to send a mailshot to an individual, geodemographics will give information about the characteristics of the neighbourhood (and via techniques that are explained in the author's book, a great deal can be adduced about the propensity of that type of neighbourhood to be 'in the market for' the product in question). A marketer cannot, however, be certain that the individual shares the characteristics of the neighbourhood. Taking housing tenure as an example — an OA of great potential for a product might have been identified, which for the sake of illustration, could be double glazing. Research indicates that householders who own, or are buying, their own home are better prospects than

Figure 1 Construction of OAs

These diagrams illustrate the construction of OAs. Underlying data are Crown copyright, Ordnance Survey and Office for National Statistics.

Table 2: Example of a CAS table

Table CAS001 Age by sex and type of resident Table population: All people Geographical level:					
All people	All people 0001	Household residents		Communal establishment residents	
		Males 0002	Females 0003	Males 0004	Females 0005
0–4	0006	0007	0008	0009	0010
0	0011	0012	0013	0014	0015
1	0016	0017	0018	0019	0020
2	0021	0022	0023	0024	0025
3	0026	0027	0028	0029	0030
4	0031	0032	0033	0034	0035
5–9	0036	0037	0038	0039	0040
5	0041	0042	0043	0044	0045
6	0046	0047	0048	0049	0050
7	0051	0052	0053	0054	0055
8	0056	0057	0058	0059	0060
9	0061	0062	0063	0064	0065
10–14	0066	0067	0068	0069	0070
10	0071	0072	0073	0074	0075
11	0076	0077	0078	0079	0080
12	0081	0082	0083	0084	0085
13	0086	0087	0088	0089	0090
14	0091	0092	0093	0094	0095
15–19	0096	0097	0098	0099	0100
15	0101	0102	0103	0104	0105
16	0106	0107	0108	0109	0110
17	0111	0112	0113	0114	0115
18	0116	0117	0118	0119	0120
19	0121	0122	0123	0124	0125
20–24	0126	0127	0128	0129	0130
20	0131	0132	0133	0134	0135
21	0136	0137	0138	0139	0140
22	0141	0142	0143	0144	0145
23	0146	0147	0148	0149	0150
24	0151	0152	0153	0154	0155
25–29	0156	0157	0158	0159	0160
30–34	0161	0162	0163	0164	0165
35–39	0166	0167	0168	0169	0170
40–44	0171	0172	0173	0174	0175
45–49	0176	0177	0178	0179	0180
50–54	0181	0182	0183	0184	0185
55–59	0186	0187	0188	0189	0190
60–64	0191	0192	0193	0194	0195
65–69	0196	0197	0198	0199	0200
70–74	0201	0202	0203	0204	0205
75–79	0206	0207	0208	0209	0210
80–84	0211	0212	0213	0214	0215
85–89	0216	0217	0218	0219	0220
90 and over	0221	0222	0223	0224	0225

Notes: Cells in this table have been randomly adjusted to avoid the release of confidential data.
Communal Establishment residents includes staff and their families, other residents and people sleeping rough.
Crown copyright, Office for National Statistics.

householders who are renting their home. The CAS statistics may indicate that 80 per cent of the households within the chosen OA are owner-occupied and 20 per cent are rented. If that OA is mailed with the double-glazing proposition, 80 per cent of the mailings will drop through the letter boxes of owner-occupied properties, the other 20 per cent will be wasted. This is not necessarily the end of the world, as all readers with experience of direct mail will realise, because a 20 per cent wastage on this one dimension is not too serious when a 1 per cent positive response leading to a sale is probably perfectly viable. But the point is this: geodemographics on their own cannot identify which households within the ED are in the core target market, they can merely give the overall 'odds' in favour. If, instead of mailing, the marketer used the analysis to direct door-to-door salespeople to areas of best potential, the marketer would possibly save a deal of wasted effort.

'BIRDS OF A FEATHER'

Geodemographics relies for its validity on what is often expressed as the 'birds of a feather' phenomenon; people tend to live in neighbourhoods among people who are, generally speaking, similar to themselves. Experience indicates that this is true as a generalisation, although there are exceptions which prove the rule. It is part of human nature to be more comfortable among people who share things in common, and this natural tendency is enforced by the likelihood of different types of housing stock to be clustered — from 'executive estates' at one extreme, to council estates at another; from inner-city 'bedsit land' to country cottages. Of course, one of the merits of geodemographics is that the base data themselves give the analyst a

means to examine this hypothesis; the very detailed CAS data will show just how homogeneous (or heterogeneous) the OAs within a particular 'neighbourhood type' actually are.

Before exploring how raw data are transformed into neighbourhood types, one more thought about the 'neighbourhood' issue. The assumption that one individual is necessarily typical of the neighbourhood in which they live is known by geographers and social scientists as the 'ecological fallacy'. The ecological fallacy arises when results from an area (or group) level analysis are incorrectly assumed to apply at the individual level. For example, an area may have high percentages for both unemployment and students, but it should not be the immediate conclusion that it is the students who are unemployed.

As with many aspects of geodemographics and associated techniques, the important thing is to keep a sense of proportion about what marketers are trying to do, and the pros and cons of the various ways they might do it. Geodemographics demonstrably works, there are many people who have used it effectively in their businesses. It is not perfect — a shortcoming it shares with many other aspects of marketing theory and practice — but if used properly, it is a considerable advance on what went before. The risk of a single representative of a neighbourhood (a research respondent say) not being typical of that neighbourhood can be guarded against, by using a sensibly-sized sample.

PRODUCING A NEIGHBOURHOOD CLASSIFICATION

The first requirement is the raw materials — the CAS data, as described above. The very first decision to make is which

variables to use in the solution. In practice, most 'authors' of neighbourhood classifications have chosen similar variables, although some have used more variables than others. It is not surprising that similar variables would be chosen, particularly if the intention is to build a 'general purpose' classification system, because the objective is to find differences between neighbourhoods, and everyday experience indicates that differences will manifest themselves in extremes of, for example, house size, size of family, age of head of household, number of cars in household, type of housing tenure. Other differences will occur in variables such as ethnic origin, socioeconomic classification (derived from occupation of head of household), industry of employment, means of travel to work, attainment of academic and/or professional qualifications, etc.

Having chosen the 'input' variables the next decision is whether to reduce them via principal components analysis (PCA) or not. Some of the systems have employed this technique, others have not. The reason for using PCA is simply expressed. In the type of 'sociological' variables dealt with here, there are usually a number of variables which express a similar underlying 'theme' in the data. The 'themes' that will surface in this type of analysis are such things as wealth versus poverty, rural versus urban, 'young' areas versus 'old' areas (in terms of the ages of the inhabitants), etc. If a set of 'raw' census variables are put straight into cluster analysis, there is the risk that the intercorrelation between variables (the extent to which certain variables are duplicating each other's effect) will cause 'noise' in the final result. The authors of the different systems each took a view on whether to go through this intermediate process or not; most did. Readers requiring more information could refer to Stan

Openshaw's chapter in the 'Census User's Handbook'² or, if requiring an excellent explanation of the multivariate techniques which they may encounter in marketing and market research, the splendid (if dauntingly titled) book by Mick Alt, 'Exploring Hyperspace'.³

Assuming that the marketer decides to use PCA, it is necessary to extract the input variables which have been chosen, for each OA. This will give a data matrix of n variables $\times$ 200,000 OAs and a PCA can now be run on this matrix. The analysis will identify a number of 'components', or dummy variables, each comprising 'scores' on some of the original variables. These latter scores will help to interpret what each dummy variable 'means'. For example, from Pinpoint's experience in using PCA to devise its PiN classification system, the most prominent component was labelled the 'wealth indicator'. While not relating directly to 'wealth' — neither income nor net worth are collected on the UK Census — in practice this component related to the 'trappings of wealth', such as large houses, owner-occupation, multiple car ownership, etc. One of the key aspects of PCA is that the dummy variables thus identified are uncorrelated with each other — they are each quite separate underlying themes, or dimensions, within the original data. These components are then put into cluster analysis.

As mentioned previously, cluster analysis is a complex topic in its own right. In outline, the statistician conducting the analysis will take the principal component scores for each ED from the previous PCA stage, and will run their chosen method of cluster analysis on them. The statistician may try a number of runs, varying the total number of clusters (people who have

not been involved in cluster analysis may be surprised to learn that the number of clusters is imposed on the solution, rather than emerging from it) and testing the results, until an apparently 'acceptable' solution emerges. Knowing what represents an acceptable solution is very much a part of the skill involved in this process, because one of the fascinating properties of cluster analysis is that it will find patterns in any data thrown at, even if the data are garbage. (In which case, the solution will inevitably be garbage too.) Having arrived at an apparently sensible solution, a very useful check on its apparent validity is to apply it to an area with which its author is familiar (such as their home area) and map the results. Does it confirm the marketer's prejudices about the neighbourhoods that the marketer knows so well? If it does, it is probably viable! In most cases the number of clusters will be aggregated to a higher level, or several higher levels, as appropriate — another decision for the system author.

Having arrived at a cluster solution, the next task — and a very tricky one, at that — is to label the clusters. This involves adding those snappy, and necessarily concise, descriptions of the neighbourhood types that have emerged. The original census variables that went into the solutions will provide a clue here; what are the particular characteristics of the type in question that distinguishes it from the national average, and from its 'nearest neighbour' types? What are its predominant housing characteristics, tenure type, age profile, family composition? Is it urban or rural, lacking amenities and crowded, or car-rich and high status? Does a high proportion of its occupants commute, work part-time or remain unemployed? Is there a high child content, or is it full

of childless couples or retired people? A 'pen portrait' of each type can be built up in this way (supplemented by local knowledge, if this route has been followed). In the effort to find differences which can be exploited in the 'label', it is very easy to get over-excited by differences which, while striking in terms of indices against a national 'norm', are actually based on very small numbers within the neighbourhood type itself. Some of the system authors will provide a matrix of input variables by neighbourhood type, which shows both indices and GB counts for each variable used. This enables the user to establish the detailed census characteristics of each type.

LINKAGE OF ADDRESSES TO CENSUS AREAS

How census data are manipulated to produce a neighbourhood classification from 'raw' census variables at the OA level has been described. As it stands, the new system might have classified all the OAs in the country, but it still cannot classify addresses. That is because, historically, 'census geography' and 'postal geography' do not coincide. Back in the days of the 1981 Census, this was a real problem, because postcodes had to be linked to EDs by proximity analysis of their centroids. For the 1991 Census, the problem was partially resolved by the production of 'look-up tables', postcodes to EDs; but there were many postcodes that straddled EDs, which was still messy. Now, for the 2001 Census, the OAs have actually been built from unit postcodes, so the problem has largely disappeared. There are still (very few) split postcodes — which occur when OAs 'nest' into administrative geography, such as wards. So now there is an almost perfect link from an OA, to the addresses that make it up — which makes profiling of

individuals (customers, research respondents) simpler and more accurate.

References

- 1 See website www.geog.soton.ac.uk/research/oa2001. More detail is provided therein.
- 2 Openshaw, S. (1983) 'Multivariate analysis of census data: The classification of areas' in Rhind, D. (ed.) 'A Census User's Handbook', Methuen.
- 3 Alt, M. (1990) 'Exploring Hyperspace — A non-mathematical explanation of multi-variate analysis', McGraw-Hill.

Further Reading

Alt, M. (1990) 'Exploring hyperspace — A non-mathematical explanation of multi-variate

analysis', McGraw-Hill (for cluster analysis and principal components analysis).

Leventhal, B., Moy, C. and Griffin, J. (eds) (1993) 'An introductory guide to the 1991 census', NTC Publications Ltd (for ecological fallacy).

Longley, P. and Batty, M. (eds) (1996) 'Spatial analysis: Modelling in a GIS environment', GeoInformation International (for ecological fallacy).

Openshaw, S. (1983) 'Multivariate analysis of Census data: The classification of areas' in Rhind, D. (ed.) 'A census user's handbook', Methuen (for cluster analysis and principal components analysis).

Sleight, P. (1997) 'Targeting customers — How to use geodemographic and lifestyle data in your business', 3rd edn, WARC (for propensity modelling).

Census Offices' websites for much useful information:

ONS = www.statistics.gov.uk/census2001;

GROS = www.gro-scotland.gov.uk;

NISRA = www.nisra.gov.uk .