Papers

Designing geodemographic classifications to meet contemporary business needs

Richard Webber

Received (in revised form): 6 October 2003

Richard Webber
is generally recognised as the originator of geodemographic systems, having classified UK neighbourhoods using statistics from each of the past five Censuses. Formerly Managing Director of Experian's Micromarketeing Division, he is currently Visiting Professor at the Centre for Advanced Spatial Analysis, University College London.

Abstract

With statistics from the UK 2001 Census now available for small areas, the many marketers who make use of geodemographic classifications are having to familiarise themselves with new and updated segmentation systems. The release of these classifications is therefore an appropriate moment to review the current 'state of the art' in this particular form of consumer segmentation.

This review starts with a brief account of how the marketing applications to which these segmentation tools have been put have evolved since their introduction in 1979, and an assessment of their relevance and scope in relation to other sources of data that marketers can use to segment prospects and customers. The following section of the paper challenges the assumption that the finer the geographic granularity of the units being classified the better the predictive power of the classification, and questions the popular view that geodemographics is useful primarily in contexts where demographic data at the household level are unavailable. This paper presents evidence to suggest that neighbourhood often contributes incremental predictive power in behavioural models over and beyond individual-level characteristics. It is suggested, however, that there is no optimal scale for classifying neighbourhoods. Consumer behaviour within some product categories is better predicted using demographic data for areas more geographically extensive than Census output areas, while for others the appropriate granularity is as low as unit postcodes. The paper then advances a basis for judging which level of granularity, fine or small, is likely to be most valuable for predicting usage levels in different product categories.

The concluding sections explain how changes in the marketing applications to which geodemographic classifications are now put are affecting the manner in which these systems are now constructed, resulting for instance in the use of data from sources additional to the Census and in the use of new statistical methods for optimising their predictive performance.

Keywords: geodemographics, segmentation, 2001 Census, cluster analysis, Mosaic, neighbourhoods

Richard Webber
Centre for Advanced Spatial Analysis
University College London
16 Broadlands Road
Highgate
London N6 4AN, UK
Tel: +44 (0)20 8340 3034
E-mail:
richardwebber@blueyonder.co.uk

Introduction

It is now close to a quarter of a century since Ken Baker, at that time chief statistician at the British Market Research Bureau, introduced neighbourhood as a useful method of classifying consumers.[1] During this period the practice of classifying consumers according to the type of neighbourhood in which they live (or geodemographics as it is now popularly referred to) has become a trusted and familiar element within the marketer's armoury of segmentation tools. Indeed, it has established itself as part of the syllabus of those courses leading to academic qualifications in marketing.[2] But, notwithstanding the familiarity of proprietary systems such as Mosaic and Acorn, much more has been written on the commercial applications of geodemographics than on how such systems are built, whether and why they 'work' and, if they do, for whom and in what contexts. Most published books and articles on geodemographics still focus on how marketers use it.[3]

Design of geodemographic systems seldom written up for publication

In the 25 years since the Baker *et al.* paper little if anything has been written about how the various classifications are engineered, what options are available to those who design them and what effect these choices have on the effectiveness of their performance. Indeed, the paucity of academic literature results in many university geography departments, which one might otherwise expect to know about these things, being almost wholly ignorant of the empirical knowledge that has been built up by marketers on the relationship between purchasing patterns and the patterns of neighbourhood segregation which characterise modern societies.

The release of the 2001 Census at small area level and the consequent reconstruction of the well-known classification systems — as well as the 25th anniversary of the first use of geodemographics in marketing — are perhaps appropriate occasions for a wider review of the current state of the art of building geodemographic systems.

Background

MRS conference, Brighton, 1979

In the UK the seminal occasion at which marketers were first exposed to neighbourhood as a useful classifier of consumer research findings was the British Market Research Society (MRS)'s annual conference in Brighton in 1979, when Baker *et al.* presented analyses of Target Group Index data by neighbourhood type.[4]

The neighbourhood typology used by Baker *et al.* for this purpose was not originally developed for marketing applications. Its primary intention was to provide a better understanding of patterns of urban deprivation. It is ironic that a tool originally designed to distinguish between different categories of poor areas came to be used as a tool for differentiating between different categories of rich people.[5]

Likewise, the original purpose to which the classification was put by the British Market Research Bureau, to validate the reliability of their then current sample for the Target Group Index survey, was very different to the principal applications to which neighbourhood classifications were later put. Baker's original requirement was to establish that survey respondents were being recruited in correct proportions between different sorts of area, such as high-rise flats, ethnic ghettos, rural villages and suburban semis. Indeed, it was only as an afterthought that Baker used the classification to test differences in the profiles of users of particular products covered in his survey. The ability of neighbourhood to

Contrasting readership profiles of *The Guardian* and *The Daily Telegraph*

discriminate between the typical *Guardian* reader and the typical *Telegraph* reader, notwithstanding the very similar profiles of the papers' readers by social grade, was a key highlight of the Baker *et al.* MRS paper.[6]

Using geodemographics for customer acquisition

Unlike the majority of tools that are designed for specific applications, neighbourhood classifications have unwittingly enjoyed a stream of new and unanticipated marketing applications over the succeeding quarter-century without any conscious effort on the part of their designers to adapt their classification methods. After the initial applications in sample frame validation and readership research, the systems started to be used for prospect targeting. Door-to-door distributors adapted their distribution schedules to allow users to be selective about which types of neighbourhoods they wanted leaflets or samples dropped to.

In 1982 the Post Office, in collaboration with CACI and CCN (as Experian was then called), launched a system branded the 'Consumer Location System' which promoted the electoral register, enhanced with a geodemographic overlay, as a rentable source of names and addresses for prospect mailings.[7] Sales departments of local media were quick to recognise the use the neighbourhood profile of their circulation areas as a tool for promoting the suitability of their titles for prospective advertisers. From the start, and in particular in the USA, retailers found it cost-effective to purchase profiles of the trade areas around potential new sites before undertaking a site visit or engaging in expensive site assessment strategies.

Using geodemographics for customer management

During the past ten years, while it is undeniable that these early applications have matured, the focus of geodemographic applications has increasingly shifted beyond the recruitment of new customers towards the development of differentiated treatment strategies for existing customers and branches. For example, in the credit and insurance industries, where customer profitability is so dependent on bad debt levels and claims rates, the data are used to forecast risk at the consumer level and to set credit limits, insurance premium levels and even annuity rates accordingly.[8] It is this analysis of risk as much as of responsiveness which now informs what messages are communicated to existing cardholders and insurance policyholders. Retailers and leisure operators use the data to consider what format a store or pub should take, what decor would be appropriate, even what style of service their cashiers or bar staff should adopt. Call centre operators and internet site designers use the data to consider how to vary the script, site design and offers according to known information about the caller/visitor, not just as a tool for targeting outbound calls. Financial services organisations structure their systems to use the data as an input into their customer relationship management systems, with the intention of making communications more relevant to existing customers, not just new ones.

As channels proliferate and businesses become more aware of the differential costs of servicing customers according to the channel used, it becomes increasingly important that a geodemographic classification — or indeed any strategic tool for customer segmentation — should deliver significant discrimination in terms of channel usage.

Likewise, as more and more public domain data become available at the person or household level, there are opportunities for neighbourhood segmentations to be used in combination with person- and household-level data in order to optimise the effectiveness of consumer segmentation systems that operate at the person or household level.[9] This practice is particularly developed in the USA, whose business culture is more favourable than Europe's towards the trading and pooling of consumer data.

Compound revenue growth by around 10–15 per cent per annum

Across the 15 markets (other than the USA) covered by the largest of the international geodemographic networks, Mosaic, revenue growth from the supply of geodemographic services currently runs at around 10–15 per cent per annum. Within this overall average, however, revenue growth is faster in complex and hybrid applications and in bespoke systems applications than it is from the delivery of standard data products and standard applications systems.

In the USA, where there is greater separation of geodemographics suppliers and applications developers, virtually all market growth is in bespoke applications development and analysis, as distinct from data development and sales of standard applications.

Combining neighbourhood with other segmentation systems

The role of geodemographics in relation to other sources of segmentation data

In many of the early applications of geodemographics which involved the targeting of new customers, the segmentation would often be the sole or primary basis for targeting. Electoral roll selections and door-to-door distributions may have involved a regional filter, but primarily they were driven by the set of neighbourhoods deemed to be worth targeting. But as the focus of attention shifted from recruitment of new customers to the management of existing ones, neighbourhood data become just one of many possible sources of information that may be known about that customer. As a result interesting issues arise as to the status or relevance of geodemographic data in relation to other sources of segmentation data. These issues are poorly discussed, if at all, in the literature, and yet are very relevant to the effective implemention of customer management strategies and deserve wider debate.

Event data

The content of the CRM database that many businesses would aspire to build would include a number of qualitatively different types of data. These, for convenience, can be organised into three principal groups. There is likely to be a substantial amount of 'event'-related data, for example dates of key purchases, enquiries, disputes and records of poor service, previous mailing history, returned goods, insurance claims, missed payments etc. Some of these events will be chained, for example the events leading to the renewal of an annual contract for insurance, breakdown recovery or membership subscription.

Product usage and ownership

A second major group of data items is likely to describe the products held by each customer as well as summary data, derived from event records, showing the levels of usage, profitability or value at the customer level over a number of time periods, these often corresponding to accounting cycles. So a retailer with a loyalty card might register total amount spent by month or a credit card issuer the number of transactions and the average balance outstanding.

Demographic data

Geodemographic data belong in a third group together with other sources of information about the consumer as a person, as distinct from as a customer. Such information might typically also include date of birth, gender, perhaps other data on an application form and other externally sourced data, whether from lifestyle surveys or from comprehensive datasets built from public data.

So the first question is what value should be given to geodemographic data in relation to other demographic data known at the individual level? A helpful way of assessing this is to use market research survey data to undertake a series of tabulations of a product or a set of products against both neighbourhood and personal-level classifications. By measuring the variability of product usage across different age groups, the two genders, various categorisations of household composition, income bands, levels of educational attainment or daily newspaper read, it is possible mathematically to rank order each of these different discriminators according to how well they distinguish between heavy users/all users/non-users of a particular brand or product category. The same statistics can be created from the same source to evaluate the discriminatory power of neighbourhood.

Comparative discrimination of demographics and geodemographics

Unpublished research, undertaken at the time of the launch by the Post Office of the Consumer Location Service, showed neighbourhood classification, on average, to be as good a discriminator as other discriminators, such as age, income, educational attainment or household type, that operate at the personal or household level. These analyses suggested that there were very few product categories for which neighbourhood was the most powerful discriminator, yet also very few for which it was the worst. If this is the case two conclusions follow. The first is that, in terms of their behaviour as consumers, the set of people who live in the same street are no more nor less similar to each other than a group of people who fall within a similar age band, or who fall within the same income bracket or who have similar levels of educational attainment.

Although neighbourhood may be a more actionable discriminator because it is a piece of information that is known about all consumers, not just those who are customers or who fill in lifestyle surveys, it is not the case that it is less useful than personal or household-level data merely as a result of being a statistical aggregate. Streets may contain households in many different income groups — but income groups themselves contain households which are equally diverse in terms of how much disposable income they have, how much of that income they save or spend and what they spend it on. Perhaps the best way, therefore, to view a neighbourhood classification field on a CRM database is as just another demographic attribute known about that individual.

The second conclusion is that the more multidimensional the nature of an organisation's relationship with its customers the more relevant to its targeting geodemographic classifications are likely to be as compared with other demographic discriminators. Geodemographics is likely to be of greater use in contexts requiring consumers to be differentiated at one time on the basis of their propensities to buy many different products. Likewise, many segmentation schemes need to address consumers on multiple dimensions, such as, for example, responsiveness, channel preference, payment method and reason for purchasing the product as well as likelihood of purchase.

In which contexts should different types of data be used?

What, then, is the relevance of geodemographic (and other demographic) data in relation to product data? Where a business holds both types of data about its customers, as most do, it is likely that both types of data have a role to play but in different degrees.

It may be instructive for a business, when comparing its segmentation practices with 'exemplars' in its own or other industries, to consider where it falls along a continuum between data richness and data poverty. This might be done by estimating the total number of useful items of information captured from customer transactions and communications in the course of a year. Businesses offering credit accounts, phone companies and retailers with loyalty cards are examples of businesses which may on average generate between 500 and 1,000 useful items of information about each customer in any year from their operational systems.

Some organisations are richer in data than others

A provider of a personal pension, by contrast, may generate no more than one useful item a year. Car manufacturers, publishers and insurance companies generate very few useful items, energy companies little more. Credit card operators, airlines and savings account providers generate many more useful items, but their averages often conceal wide variations between their 'better' customers and their 'worse' ones. Self-evidently the more data rich a business is the less likely it is to have a big requirement for neighbourhood classification data when segmenting customers.

Some customer relationships are more mature than others

A second dimension relevant to this issue is level of engagement of customers. In most businesses fewer useful items of information are known about new customers than about long standing customers. Less is known about applicants than about accepted customers. On mail-order companies' databases there will always be a large number of dormant or scarcely active customers. This is true of many financial services providers also. Whether a business is data rich or data poor, external data such as geodemographics will be relatively more useful in the segmentation of new or infrequent customers than in long standing and highly active ones.

Joining these two dimensions, as is done in Figure 1, it is possible to see the relative value of these different sources of data in relation to each other for targeting different types of customer in different types of business environment.

Some organisations have more products to cross-sell

One respect in which this general schema needs to be qualified is according to the width of the range of categories that a business offers to its customers. As a general use we can conclude that 'external' data, such as geodemographic classifications, are more useful to businesses that have a wider variety of products that they cross-sell to their customers than to those whose communications programmes focus on the upselling of an existing product range to the same customer. For example, were Tesco to

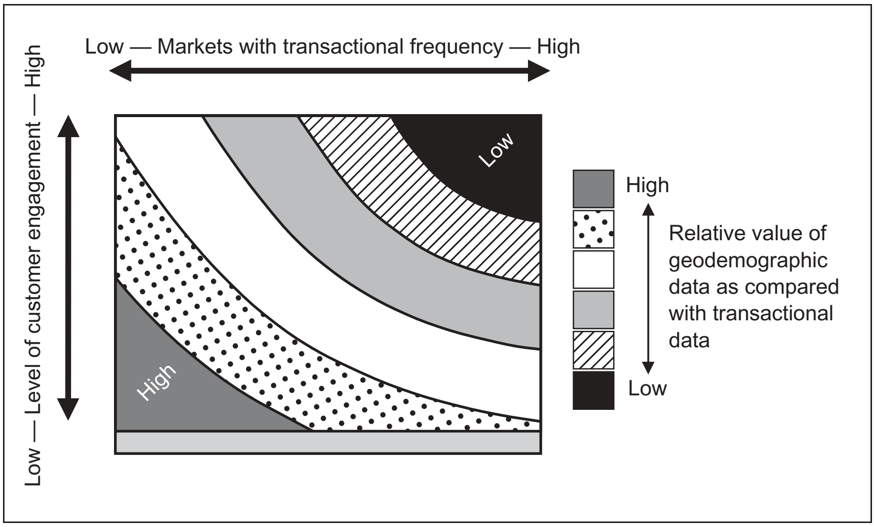
Low — Markets with transactional frequency — High
Low — Level of customer engagement — High
Low
High
High
Low
Relative value of geodemographic data as compared with transactional data

Figure 1: Relative value of transactional and geodemographic data

enter the car market, it is unlikely, though not impossible, that the vehicular demand of its customers could be predicted from even the most detailed inspection of their grocery purchases. Geodemographic data would have a more valuable role, unrivalled though Tesco is in terms of the richness of its transactional data. Conversely a phone business attempting to identify which tariffs or service add-ons to promote to its existing customers would undoubtedly do best to base their targeting on a very detailed analysis of existing customers' calls.

Some organisations share customers with competitors

A second qualification to these general rules may prove useful. In principle it is possible to differentiate between most businesses by considering the extent to which they share their customers with their competitors. Most consumers would, for example, visit more than one retail chain, would hold financial products with more than one financial services group, would travel using more than one airline and would eat at more than one restaurant chain. This is by contrast to the situation in sectors such as energy, home loans, mobile phones and insurance where, at any one time, most consumers have monogamous rather than promiscuous relationships.

Business operators in the first of these two groups experience a recurring problem. They know which customers are highly engaged with the brand and are responsive to offers. They can target these customers easily using transaction data. But among the residue of occasional users it is invariably difficult for them to identify the customers for whom their product offering is inappropriate — and to avoid wasting money attempting to reactivate these customers — from those who are heavy users of competitors' products and services, customers at whom they might logically wish to target their heaviest communications spend.

For organisations in this 'share-of-wallet' category it is likely that the most effective segmentation strategy will be one which matrixes for each individual customer their current level of spend with the business against likely level of spend on the product category across all competing suppliers (see Figure 2).[10] Using such a strategy, infrequent users will fall into separate segments according to whether or not they justify investment effort. Likewise, among the more profitable customers it will be possible to differentiate those who could contribute even more business, and who therefore are ones most likely to be targeted by competitors, from those for whom a policy of thanks and effective stewardship would be more appropriate than attempts to sell additional products.

Segmentation on the basis of events

How, therefore, do geodemographic data relate to 'events' as a source of information with which to segment consumers?

Events vary according to frequency and significance

Turner[11] has shown that events can take many different forms. Some events, such as the phone calls a consumer makes, are of high frequency and low significance to a telephony operator, in which case they may be best used in the form of summary data. By contrast others, such as non-payment, are potentially very significant. Some, such as mailings, are initiated by the supplier while others, such as a change of address, are notified to the business by the consumer. Some, such as the maturity of a savings plan, can be foreseen before they happen while others, such as a missed payment, can only be recognised after they occur (or fail to occur).

Turner argues that in general high frequency/low significance events should be incorporated into segmentation via their impact on summary data, which may then be used to modify the segment into which a customer falls, while low frequency/high significance events should be used not to alter a customer's segment but in the form of business rules which override the treatment which would otherwise have been accorded to that consumer on the basis of whatever segment he or she may have been in.

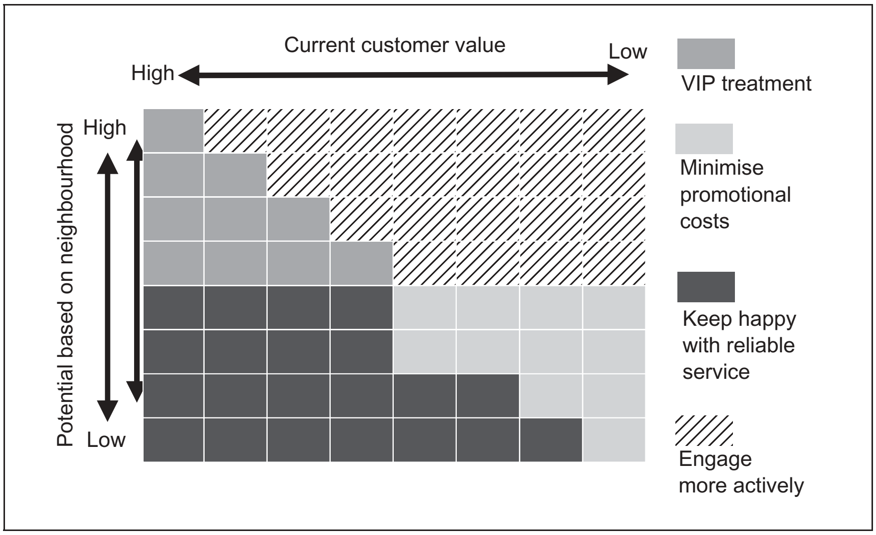
Current customer value
High
Low
Potential based on neighbourhood
High
Low
VIP treatment
Minimise promotional costs
Keep happy with reliable service
Engage more actively

Figure 2: Strategies in 'customer sharing' contexts

This argument is predicated on the not unreasonable assumption that in order to be effective a segmentation system needs to be constructed using stable rather than volatile data components. In this context it is possible to see demographic/geodemographic data, product data and event data falling in different places on a continuum from stable to volatile data. Thus a person's date of birth is fixed for life, as for most consumers is their gender. Data on product holdings hold half their value for perhaps two years, while the half-life of the value of volatile event data, such as an enquiry, a complaint or a bad service experience, may be less than two weeks.

Concept of a data item's 'half-life'

These are the grounds for believing that in customer relationship management applications the very stability of geodemographic classifications (and of demographic data generally) make them suited for use as or within strategic classifications that drive decisions on relationship objectives. Data sourced from product usage, being more volatile, are more effective when used in a tactical matter to decide which customers to include in or exclude from specific communications campaigns. There is, however, a secondary implication that, to maximise their effectiveness, geodemographic classifications should wherever possible be constructed from data variables whose local turnover is low, such as age of housing and type of building, rather than from data which themselves are volatile over time, such as the number of people who have purchased by mail order in the last 12 months or the frequency of house purchase transactions in the immediate postcode sector in the past three months. It is on the basis of this sort of reasoning that lifestyle data and other information on detailed purchase decisions may not be a very effective input into neighbourhood classifications, however interesting and relevant that particular information may appear to be as a predictor of local consumer behaviour.

Strategic segmentation should be based on data that changes infrequently

The influence of neighbourhood on consumer behaviour

The preceding section attempted to set out some simple frameworks for assessing the business circumstances in which neighbourhood (or indeed other external) data are likely to provide a worthwhile investment and those in which hybrid segmentation systems, based on a combination of demographic and product data sources, are likely to be more effective than either one source or the other source on its own.

While this may provide some theoretical explanation of where and how neighbourhood may prove useful, it does not necessarily help with the understanding of why neighbourhood may influence consumer behaviour and how the design of classifications can be tuned to deliver the highest possible level of discrimination in consumer markets.

Why should neighbourhood influence behaviour?

When people discuss the reasons why neighbourhood classifications usually 'work', the conventional answer with which they typically satisfy themselves is that 'birds of a feather flock together'. It may be useful to investigate in more detail the reasons why this may be so.

There are many circumstances in which a neighbourhood classification will prove useful because it can identify what might be called 'environmental' characteristics of a consumer: those attributes which

'Environmental' explanations

have a direct relationship to their needs. For example a consumer's postcode may be a single tower block. In this case living in a tower block makes the consumer a poor prospect for a lawn mower. Likewise, if a consumer's postcode is characterised as a military base then this consumer is likely to be a poor prospect for a mortgage. A consumer living in a postcode characterised as 'Rural Disadvantage' is likely to be a good prospect for Calor Gas. A consumer living in 'Asian Heartland' is likely to be a good prospect for Indian channels through a cable operator. In each of these cases the classification has introduced an dimension to the targeting armoury which is more relevant than the commonly used demographics, such as age and social class, conventionally used in research questionnaires or on customer application forms.

Clearly the effectiveness of geodemographics in incorporating this environmental variability depends upon the granularity at which environmental variation occurs on the ground and upon the fineness of local detail for which statistical information can be obtained.

The level of granularity needed to build a neighbourhood classification which enables marketers to target with acceptable levels of precision can vary by country. In Hong Kong, for example, where the majority live in very large developments of high-rise flats, granularity does not need to be finer than the apartment block. In New Zealand, by contrast, where many spacious older plots are being divided to accommodate infill housing, the release in 2002 of data at the level of the individual building has proved very important in improving the predictiveness of systems that had previously been built solely from statistics at mesh block (as Census output areas are called there) level, particularly in markets where local demand is driven by environmental characteristics.

'Social' explanations

By contrast to these 'environmental' characteristics there are others which we might call 'social'. In these cases the behaviours of individual consumers are not so much driven by their own objective needs or circumstances (as in the case of the need for lawn mowers in high-rise flats) as by the types of people with whom they come into contact during their daily lives. For example, unpublished research undertaken by Devon and Cornwall police shows that the reason why a person living in a council estate characterised by 'Peripheral Poverty' is likely to be a worse insurance prospect and receive a more expensive insurance quote than a person living in 'Rural Disadvantage' is not primarily as a result of their own personal characteristics but due to the proximity of their homes to the places where regular criminals tend to live. Likewise, the reason for a postman living in an inter-war council estate in Bexhill being more likely to vote Conservative than a postman living in an identical street (if one could be found) in Wath upon Dearne is almost certainly the result of the political culture of the community within which he or she may live.

Regional behaviours

Even coarser levels of geography may be the solution in certain markets. To take an extreme example, disparities in local demand in markets such as porridge, whisky and oatmeal biscuits are best understood in the UK at the level of standard region, while preferences for different types of alcoholic drink are significantly influenced by local traditions often related to the history of local employment. Today researchers at University College London are identifying some evidence, as yet unpublished, that variation in the local consumer demand for high-technology products is not unrelated to the presence of employment with IT vendors in the local economy. People living in regions rich in information technology industries, so it would seem, are more likely to be users of new information technologies during leisure hours than those living in regions historically dependent on heavy industry. In many product categories, therefore, consumer behaviours are as much determined by social influences, whether at the pub, the school gate, the church or with relatives, as they are by personal demographics or ‘environmental’ circumstances.

Impact of neighbourhood on unemployment

Perhaps the clearest indication of the incremental effect of neighbourhood over and beyond that of personal or household demographics can be seen from Table 1, which is based on 1971 Census statistics for Liverpool. In 1971, as today, unemployment levels in the lower socio-economic groups were many multiples the rate among professionals and managers. As a result local pockets of high unemployment could be explained as much by the social mix of an area as by the local availability of jobs. What Table 1 demonstrates is that, even after controlling for social class, residents in areas of big subdivided old houses in Liverpool were over twice as likely to be unemployed as residents of equivalent social class living in older terraces nearby. Knowing both social class and neighbourhood allows one to make a far more accurate prediction of a person’s vulnerability to unemployment than knowing just one or just the other piece of information.

What, then, is the relevance to the design of neighbourhood classifications of the distinction between the importance of environmental as against social explanations for differences in local consumer patterns? The answer is that in order to classify neighbourhoods optimally for

Table 1: Unemployment rates in Liverpool by social class and by type of neighbourhood (1971)

Great Britain average unemployment by SEG %	Liverpool average unemployment by SEG %	Socio-economic group (SEG)	Type of neighbourhood SEG specific unemployment rates expressed as % of the city average for that socio-economic group				
			A: Leafy Suburbs	B: Victorian Subdivision	C: Inner-City Council	D: Peripheral Council	E: Victorian Terraces
1.80	2.80	Professionals / managers	70	185	246	139	72
2.20	4.10	Non-manual	38	174	219	127	115
4.00	7.80	Skilled-manual	55	191	168	88	98
4.00	9.60	Semi-skilled	39	137	181	93	89
9.10	18.40	Unskilled	33	166	150	86	75

Notes

1. National unemployment rates are much higher among lower socio-economic groups than among higher ones.
2. In Liverpool all socio-economic groups are more prone to unemployment than in Great Britain.
3. In Liverpool unemployment among professionals and managers is only 70 per cent above the corresponding GB rate while among the unskilled the unemployment rate runs at 100 per cent above the national average.
4. Among professionals and managers, those who live in Inner City Council estates are 250 per cent more likely to be unemployed as those living in Leafy Suburbs.
5. Among unskilled workers, those who live in Victorian Subdivision are 400 per cent more likely to be unemployed as those living in Leafy Suburbs.
6. A professional or manager living in an Inner City Council estate is slightly more likely to be unemployed (6.8 per cent) than an unskilled worker living in a Leafy Suburb.
7. SEG specific unemployment rates are consistently much worse in inner-city neighbourhoods, whether Victorian Subdivision or Inner City Council estates, than in middle and outer areas, overspill estates and Victorian Terraces.

targeting in ‘environmental’ markets, such as lawn mowers, Calor Gas or mortgages, the finest possible granularity should be sought. On the other hand, in behaviours such as voting, crime reduction and alcohol, the most effective classification will be one which incorporates some information about a geographical area much broader than that at which the finest level of summary statistics is available.

Advances in the design of classifications

When Ken Baker first coded the Target Group Index with a neighbourhood classification, it was with a classification built solely using Census variables. It operated at the level of the ward/parish. Each variable used to build it was given equal influence in determining the cluster into which each of the 16,500 wards and parishes should fall.

Since 1979 the design of neighbourhood classifications has advanced in a number of respects. The most important of these have been in the range of data sources used, the levels of spatial aggregation at which they are used and in the flexibility with which different input data can be ‘weighted’.

Using non-Census data sources to build classifications

The Netherlands was the first market for which a neighbourhood typology was built for marketers using data sources other than the Census. The reason for this is that public concerns over data privacy in the Netherlands have prevented the government from releasing Census statistics at a small area level. When in 1984 Wehkamp, the largest Dutch mail-order company, commissioned a neighbourhood segmentation of Holland, it found that the most valuable sources of data for the project were already available from within its own files. The reason for this was that Wehkamp had had trading relationships at some time with more than half of Dutch consumers.

The data items which Wehkamp extracted from its files and accumulated to postcode level included information on household composition, use of credit and spend. Data were also collected by analysing surname patterns to provide a measure of ethnicity. This information was supplemented by market research data summarised by a coarser level of Census geography. Clever statistical algorithms managed to strike an appropriate balance between the need for a fine zoning mesh, to maximise detail, and the need for data to be used at levels of geography for which they were statistically reliable. This led to the same variables being used at more than one level of geography with appropriate weights.

As is often the case where necessity is the mother of invention, methods pioneered in the Netherlands were then deployed in the UK by CCN (now Experian) to construct a UK classification based on a mixture of Census and non-Census sources. This resulted in the use of data both at Census enumeration district level, of which there were then 130,000 units, and postcode level, of which there were then 1,300,000 units. Postcodes would therefore be allocated to market segments partly on the basis of information about the postcode and partly on the basis of information about the Census area to which that postcode belongs. Finer granularity should, at least in relation to ‘environmental’ behaviours, improve discrimination.

Using data sources from multiple levels of geography

The most valuable non-Census sources of data were the electoral roll and the postal address file. From the electoral roll it was possible to establish at full postcode level the proportion of addresses with one or more than one elector and with just one or more than one surname present. These measures provided good surrogates for the proportions of single adults, co-habitees, sharers and married couples in the postcode. The recording of the year each name first appeared on the electoral roll at the address allowed CCN also to produce an effective measure of mobility at the postcode level. From careful parsing of the postal address file it proved possible to assign addresses to different categories. In the author's London street, for example, there is a named house (Talbot House, where the owner of the local BMW dealership lives), numbered houses (such as 16, at which the author lives) and ones with a letter suffix, such as 22a, which typically consist of small rented flats created out of large old houses. Measuring the frequencies with which such addresses occur in the different postcodes within the Census output area helps significantly to provide greater precision and accuracy than would have been possible using the Census output area on its own.

Non-Census data sources facilitate inter-Censal updating

An obvious benefit of these non-Census data sources is that they can be updated more frequently than the Census. These updates help identify occasions where an area may have been subjected to rapid change, such as the London Docklands, and identify appropriate segmentation codes for postcodes added by the Post Office since the date of the last Census.

During the 1980s there was a celebrated case within Great Universal Stores in which it was discovered that one of the mail-order direct mail scorecards, which had used the original ward parish classification based on the 1971 Census, had never been updated with the more recent postcode-based classification. The reason was not oversight — tests showed somewhat counterintuitively that in a multivariate regression model the older ward/parish classification added more incremental predictiveness to a mailing response model that the more recent classification built at finer level of resolution. Was this a case of 'social' influences at work?

This led, in 1993, to experimentation whereby in addition to using information at postcode and Census enumeration district levels, CCN included information at postcode sector level on levels of employment in various industrial sectors among the set of data items used to build the classification system. The intellectual argument for this apparently counterintuitive approach is that variability across different categories of Census variable itself varies according to scale. Figure 3 sets out in theoretical form the degree to which area differences are likely to be affected by the geographical scale at which different demographic statistics can be presented. It suggests that housing characteristics vary very much more at a micro level than at a postcode sector or journey-to-work level, while differences in employment structure are relatively weak at postcode level but retain their strength even at a labour market area level. To achieve maximum performance from a neighbourhood classification system there could be a good argument for using variables relating to employment, and to a lesser degree status, for quite coarse levels of geography while using attributes relating to housing at the lowest possible scale.

Finest level of geography not necessarily the best

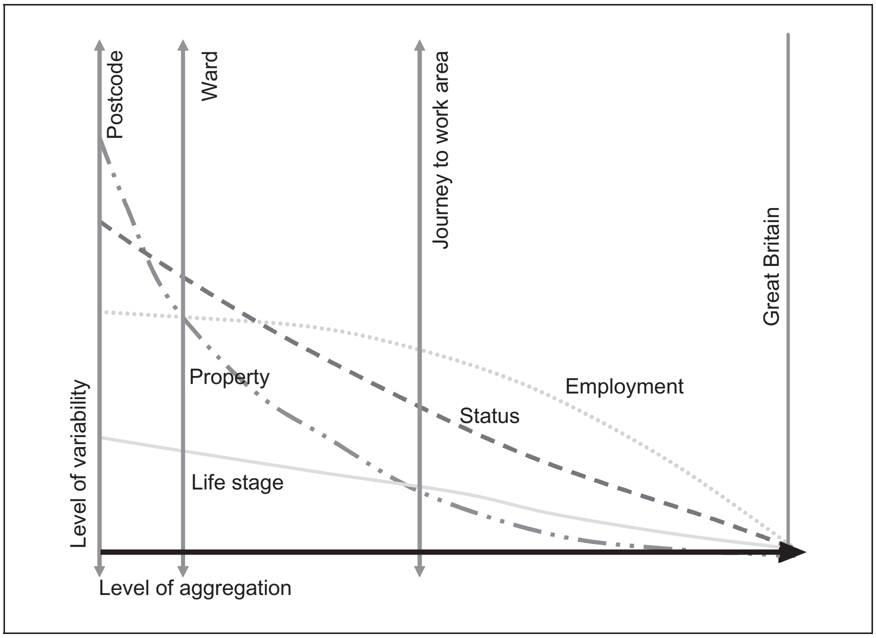

Figure 3: Loss of variance through aggregation

The specific relevance to marketers of this general tendency is the requirement to maximise discrimination on channel preference. In remote rural areas, where there are likely to be some small non-farming settlements, consumers are much more likely to use the telephone and direct mail for transactions. Without the use of data such as the proportion of workers in agriculture at postcode sector level, many rural postcodes in non-farming locations would be grouped into urban clusters with whom they had little in common.

Retail accessibility may impact on channel usage

To further the separation of these remote non-farm areas, of which the Western Isles and Shetlands were good examples, an additional non-demographic measure was introduced, namely a measure of the generalised accessibility of these postcodes to retail shops. It was hoped that the introduction of this dimension, which is not captured by the Census, would be relevant to channel preference, with people in remote areas being more likely than their suburban counterparts to transact over the telephone and to purchase by mail order.

Population density a useful input variable

In the UK rebuild based on the 2001 Census, the desire to identify non-agricultural rural locations has led to the use of moving point population densities as input variables. These density calculations are not based on the population density of the Census output areas themselves, which are sensitive to which output area non-residential land such as parks and railway sidings are assigned, but on the population densities of concentric circles around each output area. This has resulted in Scotland in the identification of a cluster of ‘miniature towns’, places such as Kyle of Lochalsh and Braemar, where few residents are employed in farming but which nonetheless lie at the centre of very sparsely populated areas.

In Australia, in a similar initiative, Pacific Micromarketing thought that the effectiveness of its neighbourhood classification could be improved if it could differentiate coastal resort and retirement areas from outback tourist and vacation areas. To do this Pacific Micromarketing measured the distance of each Census area from the ocean. It is felt that by emphasising this aspect of neighbourhood variation it has been possible to achieve higher levels of discrimination, particularly in relation to leisure activities.

Measuring distance from the coast

Census topics focus on 'deprivation' indicators

How neighbourhood classifications were developed originally to assist in the identification of deprived areas has already been noted. In this context it is also worth noting that the questions which the government decides should be included in the Census tend to provide better coverage of low-income than high-income groups. Whereas the Census contains many measures of deprivation, such as sickness, unemployment, single-parent families, overcrowding and lack of exclusive use of a bathroom, there are relatively few measures of extreme affluence, other than access to three or more cars or enjoyment of homes with eight or more rooms.

To meet the requirements of marketers for finer delineation of upmarket areas there are clear arguments for supplementing Census statistics with newly available datasets such as council tax bands, average house price values and the proportions of residents with high levels of shareholding. The use of these non-Census data sources in the Scottish Mosaic results in a classification which more clearly differentiates areas of highest status and income from older up-market neighbourhoods. In the case of the latter it is high net worth rather than high income that is the most distinctive feature of the cluster.

Use of first names to infer age

The use of electors' names is another less obvious but nonetheless interesting innovation. For many years demographers have noted the tendency for parents at particular points in time to give particular first names to their children. Fashion appears to come and go with girls even more rapidly than with boys, so it is quite easy to make a good estimate of a woman's age from her name. Combining information on the first names of electors with those of their partners, where they appear to have them, and taking into account the number of years they have lived at their address — a statistic which is consistently lower for lower age groups — it has been possible to impute an estimated age for each UK elector with quite a high level of accuracy. It is only by averaging these age estimates at postcode level that it has been possible to identify postcodes which contain bungalows built to accommodate the very old from other postcodes within surrounding areas of apparently old people and low-rise council housing. Name-based age estimates have also been useful in evaluating the likely age distribution of postcodes of new housing added since the date of the last Census, only some of which accommodate young couples with children.

The 2001 classifications have extended the use of names to incorporate family as well as given names. For instance, in Scotland the percentage of electors with self-evidently Scottish names is significantly higher among consumers in highland and island communities than among consumers in student areas, defence establishments and areas of high-income singles and families in inner areas of Glasgow. Indeed, the percentage with

Scottish names has proved a more effective indicator than the Census indicator ‘speaking Gaelic’ in identifying areas with the most traditionally Scottish way of life.

Use of family names to infer ethnicity

The use of family names to improve the effectiveness of segmentation in England centres on the use of ethnic forenames, and more particularly the use of Indian names. Unlike British family names, which have no correlation with status, better-off Indians who have been successful in assimilating into British society tend to have very different family names from those who migrated from rural areas, often not speaking English, living in intensely Indian enclaves and working in manufacturing industry. The proportion of electors with high-status Asian surnames proved highly effective in the creation of a cluster ‘Asian Enterprise’, a type of neighbourhood now particularly common in the London boroughs of Harrow and Brent, which have one of the highest rates of readership of the *Financial Times* and *Time* magazine. If this variable can be updated on a regular basis it will be very useful for tracking the continuing suburbanisation of better-off Asian families which has been such a striking feature of the past ten years.

Balancing importance of different data items

Innovations such as these involve the widening of the set of statistical attributes beyond those covered by the Census and are intended to make classification more effective in meeting a number of very specific requirements that marketers now have. Other important innovations tackle the issue of how to establish the most appropriate level of influence accorded to the different measures used to build the classification, and how to build classifications which will generate the highest possible level of discrimination on customer files.

Neighbourhood classifications are created by assigning each postcode, and hence each consumer living within it, to a cluster whose ‘profile’ is most similar to the profile of the postcode. When calculating the overall measure of similarity to the profile it is necessary not just to decide which neighbourhood statistics should be included within the profile but to decide the relative importance (or ‘weight’) that should be given to each of the individual statistical items used to generate that overall measure of similarity.

If, for example, the analyst up-weights the relative importance allocated to housing variables, the cluster a postcode is assigned to will be determined to an increasing degree by its housing characteristics. The neighbourhood categories produced by such an approach will then be more effectively described in terms of their housing stock. If population densities are given higher weightings the resulting classification system will tend to partition areas in such a way as to discriminate older inner-city areas, inter-war suburbs, fringe and country areas. If setting the relative weights assigned to these different inputs can have a major influence on the shape of the solution, what can the analyst do to ensure he or she has appropriately balanced the relative weights that are accorded to housing versus income, age or household characteristics? What is the relative level of influence that should be accorded to Census versus non-Census data sources? What is the relative influence that should be accorded to data summarised at postcode level relative to data at output area level relative to higher levels of granularity? Until lately these decisions could only taken 'blind' on the basis of hunch, judgment, intuition or experience.

The solution to this problem is to train the classification system against 'external' data, that is to say data other than those used to build it. Such data might be market research data but, in practice, they are more likely to be lifestyle questionnaire data or data from client files, since training needs to be undertaken as the system is being built whereas market research data can only be used after the system is completed to compare its relative efficiency against previous or competing classifications.

Using training strategies to optimise discrimination

A sensible training strategy therefore may initially involve the aggregation of responses of up to 1 million lifestyle respondents to postcode level. These postcode-level statistics can then be added to the set of neighbourhood indicators used in the classification process. By allocating these input variables a zero weight in the process used to identify the cluster each postcode's profile is most similar to, it is possible to compare how well alternative classification solutions differentiate on these consumer behaviours without allowing the additional data to affect the classification result. By trialling different combinations of weights for different categories of input data, it is possible progressively to adjust the weights so that they optimise the discriminatory power of the classification at least in relation to the lifestyle variables chosen.

Measuring the improvements in discrimination

How much of an improvement can training make? In the case of the new Scottish Mosaic classification released subsequent to the 2001 Census it is interesting to note that the initial set of weights chosen by the analysts produced an improvement in discrimination on the lifestyle behaviours of 7.3 per cent compared with the previous classification based on 1991 Census statistics and non-Census data sources for 2002. This comparison is between two systems with a similar number of categories. By adjusting the relative weights given to non-Census data for postcodes, Census data for output areas and Census data for higher-level data in such a way as to optimise discrimination across a wide range of lifestyles and customer files, the analysts were able to increase this improvement from 7.3 per cent to 8.9 per cent.

In the case of UK Mosaic, the equivalent overall performance improvement is somewhat greater, 13.0 per cent, but some element of this results from the number of classification codes being increased from 52 to 62. The impact of the training on performance improvement is broadly similar, however. The weights originally selected by the analysts produced an improvement compared with the previous classification of only 11.2 per cent.

These results confirm an informal working hypothesis that the predictive effectiveness of neighbourhood classifications deteriorates by an average of around 1 per cent per year, a rate incidentally not dissimilar from the rate at which new dwellings are constructed. What improvement in efficiency is gained from the use of data sources other than straightforward Census statistics for output areas but with no training is not known. The author's best estimate would be in the region of around 3 per cent, of which one half can be obtained by using subjectively arrived at weights, and the other by the use of the training techniques described above. Set in perspective this compares with an expected annual loss of approximately 1 per cent in predictive efficiency due to Census statistics becoming progressively more out of date.

New Census, new patterns

With the arrival of the new neighbourhood classifications, will marketers be better able to identify market segments that have existed for some time — from the use of more varied datasets and better methods — or will new segments come to light which are the product of social change?

The author's judgment is that as the result of updated Census statistics, the use of a wider range of data sources and the use of more sophisticated build methods, a number of key changes have become apparent. Eight trends of particular relevance to marketers are as follows.

Key changes since 1991 in where different types of people tend to live

Trend one is for the UK rural population to grow faster than the urban population, notwithstanding the continuing decline in the proportion of people employed in agriculture. Living in the countryside is increasing a lifestyle option rather than an employment necessity, and such consumers can only be located to the extent that measures other than employment in agriculture, such as population density, are given high weight in the classification process.

Trend two is the growth in what might be called 'dinky developments', the emergence of new housing developments suitable for singles and childless partners, often in less prestigious 'brownfield' sites. Information on the date of postcode introduction is a key indicator identifying residents of this new housing type and differentiating them from their longer-established neighbours.

Trend three, not unrelated to trend two, is the growth in the number of new flats for households without children, particularly in older suburbs and new docklands developments. Such areas have quite distinctive expenditure patterns, being good areas for eating out and entertainment, and ones which are significantly different from the older-established areas of divided flats in big old houses traditionally occupied by students and other younger singles.

Trend four results from the widening of access to higher education and the consequent growth in the number of students living away from home. This, combined with the entry of new private landlords into the housing market, is leading to quite large areas of older terraced housing, particularly in the inner areas of big provincial cities, shifting from family to singles' occupation and the marked differences in expenditure patterns that this entails.

Trend five is the emergence of middle-class Asian suburbs, of which North Brent and Harrow are good examples. These neighbourhoods cannot be identified without the use of ethnic origin and religion variables — religion being one of the few variables to have been added in the 2001 Census. The categorisation of all Asian areas into a single segment does not do justice to the variety of circumstances of different groups of immigrants from that continent.

Trend six results from attempts on the part of local authorities to address problems with high-rise blocks, either through demolition, refurbishment, new security systems or new letting policies. There is less of a stigma today than there was in the past in living in a tower. Differentiation between tower blocks may soon prove a necessity for marketers.

Trend seven results from the continuation of the process of selling council houses to their tenants. By 2001 it had become quite difficult on the basis of tenure variables alone to identify what once were built as council estates. This issue is particularly problematical in the post-war new towns — ‘white van territory’ — and in smaller towns and villages where the take-up of right to buy has been greatest.

Trend eight reflects the increasing polarisation of the elderly between those with significant capital and occupational pensions and those without. The dispersion of asylum seekers and the use of temporary accommodation by the DSS have reinforced the effect of cheap foreign holidays in undermining the status of a number of former seaside resorts such as Morecambe and Margate. This is at a time when other seaside resorts, such as Bournemouth and Eastbourne, perhaps with stronger retirement than tourist appeal, have managed to maintain their status. This increased polarisation within the growing grey market has significance for the targeting of many services relevant to old people.

References

1. Baker, K., Bermingham, J. and McDonald, C. (1979) ‘The utility to market research of the classification of residential neighbourhoods’, paper presented to MRS Conference, Brighton.
2. For example ‘The new IDM Practitioner’s Guide to Interactive and Direct Marketing’, IDM, Teddington.
3. Sleight, P. (1997) *Targeting Customers, Second Edition — How to Use Geodemographic and Lifestyle Data in Your Business*, NTC Publications, Oxford.
4. Baker *et al.*, ref. 1 above.
5. Webber, R. (1975) ‘Liverpool Social Area Study, 1971 Data’, PRAG Technical Paper No. 14, Centre for Environmental Studies, London.
6. Baker *et al.*, ref. 1 above.
7. The Consumer Location System, a new media database incorporating 1981 Acorn data, developed by the Post Office in conjunction with Billett and Company.
8. See the *Sunday Times*, 13 July 2003.
9. Webber, R. and Farr, M. (2001) ‘Mosaic: From an area classification system to individual classification’, *Journal of Targeting, Measurement and Analysis for Marketing*, Vol. 10, No. 1, pp. 55–65.
10. Webber, R., Pompa, N., Berry, J. and Reid, J. (2000) ‘Adopting share of wallet as a basis for communications and customer relationship management’, *Interactive Marketing*, Vol. 2, No. 1, pp. 29–40.
11. Turner, R. (2003) ‘The fuzzy art of decision science’, *Interactive Marketing*, Vol. 4, No. 3, pp. 243–256.