Towards a new digital data infrastructure for urban analysis and modelling

P A Longley, R J Harris
School of Geographical Sciences, University of Bristol, University Road, Bristol BS8 1SS, England; e-mail: paul.longley@bristol.ac.uk; r.j.harris@bristol.ac.uk
Received 18 February 1999; in revised form 12 May 1999

Abstract. Formal models of urban systems have the potential to reveal a lot about the form and functioning of urban settlements, yet much of this potential has still to be realised. In this paper we focus on the extent to which this has reflected the dearth of digital data that are rich, relevant, and disaggregate. Geodemographic classifications have made important and enduring contributions to small-area analysis. Yet, on the one hand, reliance upon census data makes them outdated and irrelevant and, on the other, fragmentation and diversification of social areas in cities has made the 'mosaic metaphor' of small-area analysis untenable. As part of the quest for a new perspective on data modelling, we investigate in this paper the potential of 'lifestyles' data sets for creating richer, more relevant digital models of human activity patterns in cities.

1 Introduction
Simple powerful theories and models are the most established and enduring in social science. Yet they provide only normative sketches which are increasingly irrelevant to the understanding of the messy irregularity that characterises the patterning of the real world. Empirical generalisation requires quantitative data collected according to rigorous research designs, yet most such data are too infrequently collected, too coarse, and/or insufficiently relevant to the functioning of fast-changing systems. In this paper we begin to gather a new perspective on the practical foundations to model building and to seek a rapprochement between social scientific orthodoxies and the practice of generalising individual and household activity patterns. We develop our arguments in the context of the analytical tradition of measuring and modelling the spatial patterning of social groupings in city regions, extended to encompass the so-called 'lifestyles' analysis of household consumption and activity patterns. We argue that our ability routinely to measure, share, and concatenate rich digital data sources creates opportunities to develop relevant and timely depictions of what is going on right across urban systems but that social science has so far held back from embracing such sources. The reasons for this are valid, yet if ultimately determinate they will stifle the creativity of model building in the digital age. As one of us has argued elsewhere (Goodchild and Longley, 1999; Longley, 1998), the 'new digital infrastructure' to GIS-based analysis can be fundamentally unsystematic in design but this need not necessarily preclude all systematic analysis. In our discussion we will make much of the roles of improved data models of spatial distributions in fostering the development of systematic and thematic urban models of social systems.

2 Models of social patterns in cities
Measurement of the patterning of urban social areas is a long and rich tradition in urban geography, from the work of the Chicago human ecologists through to the development of computer-based methods for social area analysis in the 1950s and 1960s. The classic exemplar is the Burgess model, which was devised to depict the way in which a clearly identifiable process (rapid immigration to the core of the city)
became manifest in a differentiated mosaic of residential land uses, as waves of 'invasion and succession' swept through successive inner-city neighbourhoods. Although generations of students have since seen measurement of pattern as a goal in itself, it is important to remember that the original rationale for such measurement lay in a simple theory of urban dynamics and that an early objective was to relate generating process to spatial pattern.

The dynamics of Chicago's rapid growth in the 1920s present an unusually straightforward context for analysis of the evolution of urban form—a fast-growing city, the change dynamics of which could be traced simply to in-migration. The application nevertheless spawned the much more general 'social area analysis' tradition to analysis of residential differentiation, based on a developing range of principal component and factor analysis techniques. Reviews of this research (see Clarke and Gleave, 1973; also Timms, 1969) illustrate the way in which inductive generalisation about the similarities between residential areas, harnessed to the 'mosaic metaphor' (Johnston, 1999), took place in innumerable case studies. Much of this research was avowedly technocentric and arose out of the development of computers to handle, by the standards of the time, large and complex data sets. Yet even as taxonomies of social areas and the groups resident within them became more detailed and sophisticated, the approach came to be seen as increasingly irrelevant to any understanding of the way in which towns and cities functioned (see Harvey, 1973). Over time, it became apparent just how dependent statistical classifications were upon the particular cocktail of variables that were used to generate them, and the constructs and labels that were appended to statistical 'dimensions' (such as 'stage in family life cycle') themselves came under greater scrutiny (see Stapleton, 1980). At the same time, within mainstream quantitative analysis, there was heightened awareness of ecological fallacy and modifiable areal unit effects in geographical analysis (see Openshaw, 1984). If the characteristics of areas could misleadingly be confounded with the characteristics of individuals in areas, then there were clear problems in overreliance upon publicly available data sets available only for (often very coarse) areal aggregations (see Cole, 1993). In short, from the mid-1970s onwards, data models of multivariate spatial distributions (and thence the thematic urban models that were built upon them) became viewed with increasing suspicion.

Data models of spatial distributions provided the foundations to thematic models of urban systems (Batty, 1981), and disillusionment with the measurement paradigm no doubt contributed significantly to the demise of urban modelling in the late 1970s. Not only were the units of analysis too coarse and spatial attributes too ill defined, but the state of computation was also too rudimentary and the scope of urban models too ambitious to capture the richness and diversity of urban systems (Birkin, 1995; Sayer, 1979). Such problems were compounded in any attempt to model the dynamics of increasingly rapid change. The growth of Burgess's Chicago was undoubtedly an anomaly in the history of urban dynamics, in that the single process which fuelled its short-term growth was clearly defined and spatially manifest. Yet even here the evolving physical layout and land-use configuration was not the regular and idealised 'city of pure geometry' (Batty and Longley, 1994) of textbook illustrations. The urban modelling tradition of the 1970s was able to come to terms neither with the myriad forms of human agency, nor with the jagged irregularity of urban morphology that arises out of urban growth dynamics in the real world. The 1960s and 1970s saw the morphology of urban land use affected by changes in affluence, increases in car ownership, and fragmentation of consumption patterns—in short, the lifestyles of those resident within the physical carcass of the city became increasingly diverse. As this took place, so the representation of urban dynamics by using crude surrogate data models, and crude spatial partitions, became increasingly irrelevant to the understanding of city systems.
In this context, the quest to relate form to function, patterning to social process, was largely abandoned (Batty and Longley, 1997). Since that time, urban geography has arguably been overwhelmed by the task of representing the statics and dynamics of spatial structure, to the point at which the discipline appears to have all but withdrawn from the task of generalisation. From the innovation of behavioural geography through to the depictions of individuals in cultural geography, the balance of intellectual activity has shifted from system-wide generalisation to richer yet haphazard depiction of disparate and fragmented subgroups within society. This has had knock-on effects in terms of the confidence of the discipline to suggest prescriptions and prognoses for urban change. Generalisation is a cornerstone to rational planning policy, and an urban geography which eschews system-wide generalisation is likely to become relegated to the sidelines of all but academic discourse. The demise of applied geography is particularly apparent in this context (see Pacione, 1999), as is the reduced esteem in which ‘predict and provide’ planning is presently held.

3 Geodemographics and the emergence of ‘lifestyles’
Academic qualms about the validity, scope, and applicability of social area analysis have had few implications for applied marketing geography. Within the United Kingdom, for example, classifications of residential areas have become an established marketing tool ever since digital census data first appeared following the 1971 Census (Beaumont, 1991). This applications field has become known as ‘geodemographics’, defined by Brown (1991, page 221) as “a shorthand label for both the development and the application of area typologies that have proved to be powerful discriminators of consumer behaviour and aids to ‘market analysis’”. In the United Kingdom and USA, data models of geodemographic distributions of entire populations have been built by retail consultancies, by applying techniques of cluster and principal components to census data. Qualitative analysis of the results of data reduction leads to the assigning of labels to the different groups (such as ‘affluent achievers’, ‘have nots’, ‘thriving greys’, etc) which marketeers have associated successfully with particular product and service niches in retailing (Goss, 1995). In an important review paper, Batey and Brown (1995) trace the transfer of the techniques of social area analysis to applications in marketing through a range of ‘near-market’ research activities. Geodemographics has no core theory beyond the notion that ‘birds of a feather flock together’ (Flowerdew and Leventhal, 1998) yet experience has shown that this provides no practical barrier to successful application. Various refinements have been carried out to the classification methodologies and their marketing to clients, including the use of supplementary noncensus data to label census classifications (Batey and Brown, 1995). By the mid-1990s geodemographics had become a successful and standard tool of the marketeer. Different proprietary systems have used different cocktails of census counts and classifying algorithms (with the SuperProfiles system having perhaps the best academic pedigree; Openshaw, 1996) yet successive geodemographic systems produced from 1971, 1981, and 1991 UK Census data represent applications of a core technology which have enjoyed repeat purchase by a range of business clients.

However, fundamental problems remain with census-based classifications, which may be illustrated with respect to the UK case. The raw data of census counts provide at best imperfect indicators of likely consumption behaviour because crucial information is not collected (notably income data). In turn, composite indicators of consumer

(1) Supplementary descriptors include the use of the National Readership Survey and data from large-scale public-sector surveys such as the General Household Survey and Family Expenditure Survey. The ascription of labels from the coarser and varied geographies of such sources introduces the risk of invoking additional ecological fallacies in the classification process.
behaviour are thus also dependent upon crude surrogate data. Moreover, current
geodemographic systems are frequently reliant upon census data that are over a decade
old, and this represents an increasing handicap in fast-changing, consumer-led markets.\(^{(2)}\)
Census data are comprehensive in terms of coverage of the UK population—notwith­
standing problems of (illegal) nonresponse by the ‘missing millions’ in the 1991 UK Census
(Marsh, 1993)—yet this is rarely a key criterion for market area analysis, with its usual
focus on subgroup behaviour. Outside of the USA, many public-sector agencies have
found it necessary to introduce aggressive data-pricing regimes in order to recover
some of the costs of data creation, and this presents a further disincentive to the use
of census sources. Taken together (and contrary to the claims of some of those who sell
godemographic systems) it is clear that the data infrastructure provided by the census
does not present a panacea for analysis of consumption and activity patterns. It is
certainly well founded in survey research terms, but its content is increasingly marginal
to the understanding and prediction of what is going on in modern Britain.

At the same time, the capture of digital data by using a range of new technologies
has become commonplace. This has led to the advent of a wide range of so-called
‘lifestyles’ data, originating from such diverse sources as consumer product-guarantee
returns, store loyalty programmes, and recorded travel behaviour. A wide definition of
lifestyles data would emphasise their chief facets: they ‘capture’ (measure) some of the
varied consumption choices, shopping habits, and practices of identifiable individuals.
The wider definition thus includes a range of nonsurvey-based sources of lifestyles data,
such as guarantee-card returns, electronic point-of-sale (EPOS) from retail purchases,
loyalty card data, share ownership records, and country court judgments, many of
which have been available since the beginning of the 1990s. Companies such as ICD,
Claritas, Experian, and Psychnographics have built up huge ‘data warehouses’. CACI
Information Services claim that by using their ‘LifestylesUK’ product it is possible to
target over 44 million individuals by using 300 lifestyle variables (source: promotional
brochure). Claritas UK’s ‘Micromarketing’ offers information on 75% of UK house­
holds (source: promotional brochure).

In the empirical sections of this paper (and following Harris, 1998, page 2) we will
take a narrow definition of GB lifestyles data, as “data obtained and stored at the non­
aggregated level of a named individual and geo-referenced by their address; the data
[are] collated from the return of (consumer) questionnaires mailed directly to eligible
voters, as recorded upon the Electoral Register for Great Britain”. An example of one
such lifestyle questionnaire is shown in figure 1.

Lifestyles data sets elicit information on a far wider range of themes than the
census, from household structure and demographic characteristics (often including
income) to consumption habits and recreational pursuits. As such, today’s lifestyles
data provide a range of relevant (direct and/or indirect) indicators of individual and
household propensities to consumer particular goods and services, as well as detailed
information about individual and household activity patterns. Data pertain to the

\(^{(2)}\) This said, there is a counterargument that geodemographic classifications remain effective long
after the data from which they are derived have become outdated. This is because, the argument
goes, the classifications are ‘driven’ or influenced by variables that reflect directly the structure of
the property market. If we assume that typologies capture property attributes as well as household
characteristics then, as time goes by, although the residents will change, the property market acts
as an effective filter which tends to have the effect of making those who replace them likely
to share many essential characteristics of their predecessors. In marketing terms, this means that
targeting the same areas remains likely to reach the same kinds of consumers, or that mail shots
will continue to reach people who share similar tastes, aspirations, and patterns of consumer
behaviour—subject to changing small-area fashions, as the 1980s phenomenon of gentrification
illustrates.
Towards a new digital data infrastructure for urban analysis and modelling

Figure 1. An example of a ‘lifestyles’ questionnaire.
widest range of purchases (from cars to pet food), leisure pursuits (for example, theatre, fitness), activity patterns (such as propensity to take weekend breaks), health (for example, asthma, backaches), and travel opportunities (for example, details of transport, regular and occasional trip-generating behaviour). These vivid depictions of what is going on in modern Britain stand in increasingly stark contrast to conventional census sources, with their tired socioeconomic classifications (Snowdon, 1998) and all-but-irrelevant surrogate income measures. Lifestyles data are widely used in mail marketing activity and as a basis to 'one-to-one' marketing (Peppers and Rogers, 1997; such activities are arguably more ethnical than area-based targeting, insofar as lifestyles survey respondents often have the option to withhold their address from marketing activities).

In this fast-changing context, lifestyles data offer a number of potential advantages over conventional geodemographic indicators, not only in direct mailings to respondents but also as the basis to area-based generalisations. Yet the negatives associated with such sources are manifold. Some are basic to the design used to mail out questionnaires and are transparent in the analysis of lifestyles data. One such consideration is the common practice of excluding houses in multiple occupation, because residents of such properties tend to be low consumers and in any case are less likely to generate responses because of their high propensities to move. Other problems are more insidious and arise out of the voluntary basis to completion and reliance upon the postal questionnaire survey instrument (characterised by highly variable and unsystematic response rates, incomplete returns, misunderstanding and misinterpretation, etc; Dixon and Leach, 1977). These latter considerations make it doubtful whether lifestyles data are representative of the characteristics and habits of the survey respondents, never mind the broader populace of nonrespondents. Some of the characteristics of lifestyles data vis-à-vis conventional geodemographics are summarised in table 1.

Taking these considerations together, there is a number of respects in which lifestyles data may provide much better digital depictions of human activities and may allow GIS representations to move beyond the static mosaic metaphor of conventional social area analysis. Yet the principles and practice of lifestyles data collection are

| Table 1. Some characteristics of geodemographics and lifestyles data in the United Kingdom. |
|---------------------------------|-----------------|-----------------|
| **Unit of aggregation** | Geodemographics data | Lifestyles data |
| Population coverage | Census ED | Household or individual |
| | 100% or 10%, depending on variable | 10.8% |
| Sampling | 100% or 10% random sample | Self-selecting |
| Consumption or behaviour indicators | Indirect | More direct |
| Compatibility with postcode geography | 70–80% (using ED-to-postcode directories) | Perfect |
| Frequency of update | Decennial | Dependent upon data warehouse priorities, but like to be frequent |
| Bias | The ‘missing millions’ of the 1991 Census | Part quantifiable if we know where HMOs and non-ER registrants are; response bias, however, is likely to be multivariate and unquantifiable |

ED, enumeration district; HMOs, houses in multiple occupation; ER, Electoral Register.
Towards a new digital data infrastructure for urban analysis and modelling

transparently unscientific. Thus the emergent digital data infrastructure provided by lifestyles data may be contemporary and relevant but it is set on far shakier foundations than those of the census and other large-scale public-sector surveys. Lifestyles data sets are data rich, in terms of the number of variables they contain and their relevance to measuring a diversity of lifestyles from hobbies to holidays. But the standards used in the assembly of these data sets fall far short of those required by the ‘linear project design’ of conventional survey research (Goodchild and Longley, 1999).

If lifestyles data are profoundly unscientific in their collection, is there any contribution that they are likely to make to the geographical analysis of social conditions? And if so, might the richness and diversity of social characteristics that they depict make an important contribution towards new attempts at system-wide urban modelling? Any revitalised systematic analysis of the form and functioning of urban social areas must be founded upon appropriate data models of the spatial form of the city and the constellation of human activities that take place within it. Yet conventional public-sector sources are unlikely to keep pace with the fragmentation and diversity that characterises postindustrial societies—indeed governments sometimes seem to lack the will even to maintain the sedate pace of renewal of conventional public-sector data infrastructure.

We do not concur with those commentators (see Curry, 1995) who have suggested that digital depictions of geographical reality can never provide meaningful abstractions. Geodemographic ‘cocktails’ of census data retain enduring popularity in modelling numerous short-term aspects of consumer behaviour (Birkin, 1995), and repeat purchasing by industry provides ample evidence that they have an important niche role to play. Indeed they are today used in a wider range of business and service planning contexts than ever before (for example, monitoring access to university education; Batey et al, 1999; Utley and Thompson, 1999). But the domain of their application is ultimately limited by the constraints governing content, organisation, and dissemination of national censuses. In contrast, lifestyles databases are up-to-date, relevant, but of dubious scientific validity. Taken together, a best course might require us to identify the degree to which lifestyles data are representative of populations at large, in order that we might identify the applications domains within which data-rich models of spatial distributions might be developed as a precursor to further generalised analysis.

In the next section we begin to explore the characteristics of a major lifestyles data set to begin to identify whether lifestyles data are sufficiently robust to be considered part of the new digital infrastructure of urban analysis.

4 Lifestyles in Bristol, United Kingdom

Despite the remarkable developments in data capture, warehousing, and application in recent years, there are very few documented examples of linkage of lifestyles data to ‘framework’ data (Rhind, 1997), such as the census, in any systematic manner (but see Birkin and Clarke, 1995, pages 372 – 384). We are unaware of any substantial scientific analysis of the content and coverage of lifestyles data—although there has been some informed speculation as to the relative merits of lifestyles and geodemographic analysis (Birkin, 1995; Cosijn and Brown, 1993a; 1993b). This is a glaring omission from the literature, and testimony to the gulf that presently divides data modelling research and practice. Any first attempt to begin to resolve this must necessarily be preliminary and here we will develop a four-point investigation of the feasibility of using lifestyles data

(3) One approach to these questions has been the suggestion that large and complex data sets constitute a fertile application area for a new range of data-mining technologies within the ‘geocomputation paradigm’ (Openshaw, 1998). However, it seems unlikely that machine intelligence can resolve the biases inherent in the collection of ‘nonscientific’ quantitative data.
in urban modelling. Given the vagaries inherent in the collection of lifestyles data, our approach will seek to anchor lifestyles analysis to the framework provided by conventional sources—that is, census data and their geodemographic derivatives. As has been suggested above, conventional geodemographics had developed considerably over the last twenty-five years and some of these developments have also entailed additional assumptions. Thus, in addition to establishing the credentials of lifestyles analysis, we will also begin to examine the relative strengths of the assumptions made in contemporary geodemographics versus lifestyles approaches. Thus we will adopt a four-stage preliminary assessment of the scope for urban modelling by using lifestyles data.

(1) We assess the degree to which a lifestyles database might be 'anchored' to 1991 Census data. The census framework provides a potential means of justifying (subject to caveats) the development of household classifications based on lifestyles data—which are likely to be richer and more up-to-date than their existing geodemographic counterparts. Our case study will also seek to establish a basis to generalisation across the wider British system.

(2) We carry out a cluster analysis of lifestyles data and compare the results with a more conventional (SuperProfiles) geodemographic classification.

(3) We investigate the degree of heterogeneity within small (census enumeration district) areas—in order to gain an idea of the fission of consumption activities within small areas and quantify the likely implications for conventional geodemographics.

(4) We compare the detail of the clustered lifestyles data with the 'pen portraits' devised for the 'freshened up' SuperProfiles geodemographic system—as a preliminary assessment of the validity and accuracy of descriptive labels that are not integral to the classification schema and which are vulnerable to scale and aggregation biases.

4.1 Lifestyles: a national and regional snapshot

The case-study data form a subset of a database collated from responses to a national postal questionnaire survey undertaken by a commercial data-warehousing company during September to October 1996. The questionnaire, which took about twenty minutes to complete in full, was mailed to addresses recorded on the February 1996 British Electoral Register (which was based on residence information as at 10 October 1995). One survey was mailed to each address, apart from: addresses with more than three different surnames per register entry (that is, those properties deemed to be houses in multiple occupation, HMOs); and those who opt out from receiving 'mail drops' through the Mail Preference Scheme. Wherever possible the questionnaires were addressed to females or to respondents to previous questionnaires. In total, 20 million questionnaires were mailed across Great Britain, of which approximately 2 million were returned, a response rate of approximately 10%.

The data warehouse anticipates that such surveys will disproportionately enumerate people who are 'mail responsive'. The characteristics of this population, which form the lifestyle database (henceforth, 'lifestyle population'), will therefore likely differ from the 1991 Census-enumerated population (henceforth, 'Census population'). Identification of response bias is difficult to disentangle from sample bias created through exclusion of HMOs and Mail Preference Scheme opt outs, as well as changes in population characteristics between 1991 and 1996. Nevertheless a crude comparison of the age profile of the lifestyles survey with that of the census (table 2) makes the underenumeration of young adults very apparent. Table 3 illustrates the relative underenumeration of individuals living in terraced properties or flats, in contrast to the relative overenumeration of those residing in semidetached property.

The case-study region comprises the 1658 1991 Census enumeration districts (EDs), which have 'BS' (Bristol) postcodes (and shaded grey in figure 3, see over). There were
Towards a new digital data infrastructure for urban analysis and modelling

Table 2. National age profile of lifestyle and census populations (source: authors’ calculations based on data supplied by the lifestyles data company).

| Age group | Lifestyles | Percentage | Census | Percentage | Difference |
|-----------|------------|------------|--------|------------|------------|
| 18–24     | 100,531    | 4          | 5,614,045 | 13         | -9         |
| 25–34     | 545,143    | 20         | 8,361,463 | 20         | 0          |
| 35–44     | 574,076    | 21         | 7,665,001 | 18         | 1           |
| 45–54     | 519,389    | 19         | 6,371,795 | 15         | -4         |
| 55–64     | 420,817    | 15         | 5,662,670 | 13         | -2         |
| Over 65   | 560,146    | 21         | 8,811,230 | 21         | 0          |
| Total     | 2,720,102  | 100        | 42,486,204 | 100       |            |

Table 3. Great Britain property types: lifestyles and 1991 Census analysis (source: authors’ own calculations based on data supplied by the lifestyles data company).

|          | Lifestyles | Percentage | Census | Percentage | Difference |
|----------|------------|------------|--------|------------|------------|
| Detached | 307,704    | 19         | 4,383,168 | 20         | -1         |
| Semidetached | 613,897  | 37         | 6,481,705 | 30         | 1         |
| Terraced | 449,203    | 27         | 6,343,156 | 29         | -2         |
| Flat or maisonette | 284,307 | 17         | 4,389,770 | 20         | -3         |
| Total    | 1,655,111  | 100        | 21,597,799 | 100       |            |

73,310 individual and adult respondents to the lifestyle survey in this region, from 51,882 households. The data were supplied in a form that allows the location of the households to be identified to the level of the unit postcode (for example, BS8 1SS). Unit postcode geography is not coincident with that of the 1991 Census in England and Wales and so it was necessary to use the 1991 and 1995 enumeration district-to-postcode directories to obtain an approximate match (see Martin, 1992). In practice, where a ‘postman’s walk’ (the basis of definition of UK unit postcodes) is identified as crossing the boundaries of one or more EDs, then the postcode was deemed to be located in the ‘pseudo-ED’ in which the directories deem the majority of the postcode population to lie. Approximately one quarter of unit BS postcodes cross ED boundaries, although this does not create any problems if the two or more EDs share common attributes (such as the same geodemographic category, which is assigned at the ED level). Nevertheless the mismatch between census and postal geographies creates ambiguity in assigning SuperProfile classes to addresses in 15% of all BS unit postcodes.

The adult census population of the study region comprises 623,132 individuals, implying (if all survey respondents were adults) that the survey was completed by 11.8% of the adult population. This is a large proportion—and larger than the absolute number of census returns used to compile 10% census returns at the ward scale (for example, for occupational data). However, ‘excluded’ members of HMOs (to which no surveys were mailed, amounting to 1.7% of all households in the study region) aside, the lifestyle respondents are entirely self-selecting. Hitherto there has been no research into the problems inherent in generalising from self-selecting lifestyles samples, and any attempt flies in the face of scientific approaches to statistical generalisation. Survey research practice is rightly and avowedly sceptical of the dangers of postsurvey stratification and differential grossing in light of subgroup response rates (see Moser and Kalton, 1993). Response biases are also likely to be compounded by partial completion of survey forms by respondents. This is most evident with respect to measurements of income—20% of survey respondents in the survey used here did not state their incomes,
Table 4. Regional and national enumeration district breakdowns according to SuperProfiles cluster [source: the SuperProfiles (SP) ten-cluster typology, after Brown and Batey (1994) and the 1991 UK Census].

| Cluster (economic rank)       | Percentage in study region | Percentage in Great Britain | Percentage in England and Wales | Percentage in England and Wales (not London) |
|-------------------------------|----------------------------|----------------------------|---------------------------------|-----------------------------------------------|
| Affluent achievers (SP1)      | 10                         | 10                         | 10                              | 10                                            |
| Thriving greys (SP2)          | 13                         | 11                         | 12                              | 13                                            |
| Settled suburbs (SP3)         | 12                         | 12                         | 12                              | 12                                            |
| Nest builders (SP4)           | 18                         | 16                         | 16                              | 18                                            |
| Urban venturers (SP5)         | 11                         | 10                         | 10                              | 6                                             |
| Country life (SP6)            | 2                          | 3                          | 3                               | 3                                             |
| Senior citizens (SP7)         | 8                          | 7                          | 6                               | 7                                             |
| Producers (SP8)               | 16                         | 12                         | 12                              | 13                                            |
| Hard-pressed families (SP9)   | 3                          | 8                          | 7                               | 8                                             |
| The ‘have nots’ (SP10)        | 5                          | 12                         | 11                              | 10                                            |
| Pearson correlation           | 0.79                       | 0.85                       | 0.82                            |                                                |

a figure that is broadly in line with the results of National Office of Survey trials for an income question in the 2001 Census.

Table 4 compares the incidence of the ten SuperProfiles geodemographic categories in the study region with some broader aggregations. Bristol and environs is broadly representative of the broader national picture, as indicated by the Pearson correlation coefficients, although the ‘lower status’ SuperProfiles (clusters SP9 and SP10) are underrepresented in the study region. The smallest category in the national classification (cluster SP6) is also underrepresented in and around Bristol. This finding is also substantiated by reference to other composite geodemographic sources (CACI Information Services, 1998; see also Harris, 1998). However, the characteristics of the lifestyle respondents differ from those recorded in the 1991 Census, and only part of the observed discrepancy can be attributed to the five-and-a-half year interregnum between the census and completion of the lifestyle survey. We have chosen not to try to estimate the extent to which this is a consequence of excluding households in multiple occupation from the lifestyles survey, as the necessary grossing would risk committing ecological fallacy. Figure 2 shows the percentage of the regional populations of each lifestage group according to the lifestyles data set and according to the census. It is apparent that, as with the national picture, it is the young who do not return lifestyle questionnaires. As Rae (1998, page 6) has commented, referring to the data of figure 2, “groups that are equally common in the population[-at-large] have radically different representation
in lifestyle data”—compare the 18–24 and 55–64 age bands, for example. Such differences could lead to misleading area profiles (see figure 3; and for further details see Harris, 1998).

4.2 Classification of a lifestyles data set

An iterative cluster analysis was developed by using the lifestyles data at the household level (for details, see Harris, 1998). The household ‘response rate’ for the survey was 16%. As with the creation of conventional geodemographic clustering, the outcome of clustering obviously depends upon the range and type of variables included in the analysis (Openshaw, 1996). We used 241 variables, chosen to represent a wide range of socioeconomic characteristics and behavioural information, as the basis for a clustering procedure. The cluster program performed best when the data were divided into sixteen clusters. The characteristics of the sixteen clusters are summarised in tables 5 and 6.

Table A1 in the appendix shows the principal defining characteristics of the clusters. They encompass a far broader range of household and individual characteristics than conventional census-based geodemographic indicators and it is interesting that table A1 reveals the importance of leisure, holiday, and consumption interests—and also other characteristics such as health. Indeed, in a number of instances, groups seem at least as much tied together by consumption as by conventional age, socioeconomic status, and family cycle considerations, if not more so. As with all conventional geodemographic classifications, and as noted above, the nature of the end classification is conditioned foremost by the nature and range of the input variables. The input variables used here are much more suggestive than conventional geodemographic indicators of whether people are sedentary, limited in physical mobility, participants in neighbourhood or city-wide activities, patronise ‘traditional’ or out-of-centre retailing,
Table 5. Variables used in the formation of the household typology.

| Type of variable                                      | Number of variables |
|------------------------------------------------------|---------------------|
| Age of household member                              | 6                   |
| Alcoholic beverages consumed                         | 9                   |
| Children: number in household and age                | 8                   |
| Consumer goods owned                                 | 7                   |
| Daily newspaper read                                 | 10                  |
| Household income                                     | 7                   |
| Financial investments and plans                      | 15                  |
| Gender                                               | 2                   |
| Have credit cards, store cards, etc                  | 7                   |
| Hobbies and pastimes                                 | 32                  |
| Holiday choices                                      | 22                  |
| Home improvements made                               | 11                  |
| Home type, tenure, and value                         | 18                  |
| Household size                                       | 4                   |
| Illnesses                                            | 9                   |
| Duration of residence                                | 6                   |
| Mail order purchases                                 | 7                   |
| Marital status                                       | 4                   |
| Charity support                                      | 14                  |
| Number of cars owned, make, and value                | 23                  |
| Smoking                                              | 2                   |
| Social-economic group                                | 5                   |
| Supermarkets regularly visited                       | 10                  |
| Other                                                | 3                   |

and so forth. On average, across the sixteen clusters, any one household value would share the same (yes or no) value as its cluster for 84% of the 241 variables. As with any cluster analysis, this does not necessarily represent the optimal solution but an optimised solution. It is also important that the results of the cluster analysis make sense in substantive terms.

4.3 Lifestyles and geodemographics: competing or complementary classifications?
The results of the lifestyles classification were aggregated into EDs in order to facilitate comparison with the ED-scale SuperProfiles for the study region. With regard to the lifestyles data, EDs were given the lifestyles descriptor which pertained to the largest absolute number of households in the ED. Table 7 (see over) shows how households in each SuperProfile category are spread across the sixteen lifestyle groups; and table 8 (see over) shows the spread of each lifestyle category across the SuperProfile categories. \( \chi^2 \) analysis confirms that lifestyle groups are not uniformly distributed across all the SuperProfile categories; Harris, 1998.) These tables suggest strong correspondence between the two classifications, with most SuperProfile categories being spread out between two or three lifestyle categories, and vice versa. This is an important finding, on at least two counts. First, the representation of all SuperProfile groups in some shape or form suggests that there are no gaping holes in the classification (arising particularly because of the underrepresentation of the young in the lifestyles survey). Second, it follows that classifications that are richer than conventional geodemographics can be built at disaggregate scales.

Differences between the two classifications are likely to have arisen from the following, alone or in combination.
Table 6. The sixteen consumer clusters (summarised).

| Cluster | Percentage | Summary |
|---------|------------|---------|
| A       | 5          | Wealthy older couples in upgraded homes and with diverse financial investments that afford many pastimes and regular holidays |
| B       | 4          | Computer-friendly couples and families with financial provision for retirement. Church attendees with interest in the arts |
| C       | 3          | Comfortable couples living in improved homes with PCs and satellite TV. Suffer from aches and pain and holiday in 'the Med' |
| D       | 5          | Comfortable city-dwelling (older) couples. Long-time residents of improved homes, within which their interests are pursued |
| E       | 6          | Younger couples and families living in improved semidetached properties. Holiday in the United Kingdom on camping and caravan trips |
| F       | 7          | Other couples in owner-occupied properties and with no children |
| G       | 8          | Affluent retired couples living in upgraded (city) homes. Diverse financial investments afford overseas and UK holidays |
| H       | 7          | Younger and middle-aged couples in improved homes. Financially comfortable with provision for retirement. Have home PCs and gamble upon the Pools or National Lottery |
| I       | 5          | Young outgoing singles. Have active social lives and often frequent holidays |
| J       | 6          | Couples and families taking few holidays and gambling upon the Pools or National Lottery. Do not smoke |
| K       | 5          | Mail-order responsive, lower-income couples residing in the city. Suffer from stress or other aches and pains. Smoke |
| L       | 4          | Female residents living alone and with home-based interests |
| M       | 11         | Low-income retired couples |
| N       | 5          | Elderly female widows |
| O       | 11         | Other lone-female households |
| P       | 6          | Low-income, single females and single mothers, living in housing association or local authority properties |

* Percentage of Bristol population.

(a) Inherent differences in the data—that is, differences arising out of the different constructs measured in the data sets, temporal changes between the two surveys, and the effects of response and sampling bias in the lifestyles data set.

(b) The effects of aggregation—as Birkin (1995) has pointed out, geodemographic classifiers may be misleading if used to suggest that ED labels pertain to every household within each classified ED, because this is patently almost invariably not the case in reality. This is an inherent problem in geographical classification and analysis, which is only ultimately resolvable through recourse to individual or household units of analysis (Openshaw, 1984). Information has been discarded in the compilation of tables 7 and 8, in that a simple ‘highest count’ rule has been used to label every ED with a lifestyle category.

It is difficult to disentangle these different considerations, although what is evident from figure 4 (see over) is a quite staggering diversity of lifestyles within EDs. This kind of small-area heterogeneity is hidden in conventional geodemographic analysis, yet the clear implication is that the mosaic of small areas used in conventional geodemographic analysis conceals considerable diversity. Moreover, table 9 (see over) shows that the EDs assigned to different SuperProfile categories are characterised by different degrees of diversity—with the most affluent EDs (in SuperProfile terms) characterised by the greatest degree of diversity. This table suggests that almost all census areas are neither ghettos of ‘have nots’ nor islands of ‘affluent achievers’—and that prescriptive urban
### Table 7. The spread of SuperProfiles across the lifestyle groups (household analysis).

| Percentage | A   | B    | C    | D    | E    | F    | G    | H    |
|------------|-----|------|------|------|------|------|------|------|
| SP1        | 25.0| 32.7 | 13.0 | 20.0 | 8.5  | 21.8 | 21.5 | 6.3  |
| SP2        | 15.4| 21.2 | 13.0 | 11.1 | 12.2 | 5.9  | 34.5 | 1.0  |
| SP3        | 23.1| 1.9  | 13.0 | 28.9 | 36.6 | 6.9  | 14.1 | 17.7 |
| SP4        | 15.4| 19.2 | 39.1 | 24.4 | 32.9 | 23.8 | 11.9 | 55.2 |
| SP5        | 3.8 | 19.2 | 4.3  | 2.2  | 0.0  | 21.8 | 0.6  | 4.2  |
| SP6        | 5.8 | 1.9  | 0.0  | 0.0  | 0.0  | 5.9  | 0.0  | 0.0  |
| SP7        | 5.8 | 3.8  | 0.0  | 0.0  | 0.0  | 4.0  | 5.6  | 2.1  |
| SP8        | 3.8 | 0.0  | 13.0 | 11.1 | 7.3  | 8.9  | 11.3 | 10.4 |
| SP9        | 0.0 | 0.0  | 4.3  | 2.2  | 2.4  | 0.0  | 6.0  | 2.1  |
| SP10       | 1.9 | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 1.0  |
| **Sum**    | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|

| I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|
| 3.1 | 0.0 | 0.0 | 0.0 | 9.3 | 0.0 | 3.7 | 0.0 |
| 12.2 | 13.3 | 13.3 | 13.3 | 14.6 | 17.1 | 7.9 | 1.9 |
| 5.1 | 6.7 | 6.7 | 0.0 | 15.3 | 11.4 | 7.5 | 1.9 |
| 19.4 | 30.0 | 30.0 | 20.0 | 15.3 | 11.4 | 13.3 | 4.8 |
| 35.7 | 26.7 | 26.7 | 40.0 | 0.7 | 5.7 | 7.9 | 8.7 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.8 | 0.0 |
| 16.3 | 3.3 | 3.3 | 20.0 | 7.8 | 42.9 | 14.5 | 8.7 |
| 6.1 | 16.7 | 16.7 | 6.7 | 32.0 | 8.6 | 27.4 | 23.1 |
| 2.0 | 0.0 | 0.0 | 0.0 | 2.8 | 0.0 | 7.9 | 14.4 |
| 0.0 | 3.3 | 3.3 | 0.0 | 2.1 | 2.9 | 9.1 | 36.5 |
| **Sum** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

### Table 8. The spread of lifestyle groups across SuperProfiles (household analysis).

| Percentage | A   | B    | C    | D    | E    | F    | G    | H    |
|------------|-----|------|------|------|------|------|------|------|
| SP1        | 8.5 | 11.1 | 2.0  | 5.9  | 4.6  | 14.4 | 24.8 | 3.9  |
| SP2        | 4.2 | 5.8  | 1.6  | 2.6  | 5.2  | 3.1  | 31.9 | 0.5  |
| SP3        | 6.4 | 0.5  | 1.6  | 7.0  | 16.0 | 3.7  | 13.4 | 9.1  |
| SP4        | 2.8 | 3.5  | 3.1  | 3.8  | 9.3  | 8.1  | 7.3  | 18.3 |
| SP5        | 1.6 | 8.1  | 0.8  | 0.8  | 0.0  | 17.9 | 0.8  | 3.3  |
| SP6        | 25.0| 8.3  | 0.0  | 0.0  | 0.0  | 50.0 | 0.0  | 0.0  |
| SP7        | 2.4 | 1.6  | 0.0  | 0.0  | 0.0  | 3.2  | 8.0  | 1.6  |
| SP8        | 0.8 | 0.0  | 1.2  | 2.0  | 2.4  | 3.5  | 7.8  | 3.9  |
| SP9        | 0.0 | 0.0  | 1.8  | 1.8  | 3.6  | 1.8  | 1.8  | 3.6  |
| SP10       | 1.4 | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 1.4  |
| **I**      | 2.0 | 0.0  | 0.0  | 0.0  | 17.0 | 0.0  | 5.9  | 0.0  |
| **J**      | 6.3 | 2.1  | 2.1  | 1.0  | 21.5 | 3.1  | 9.9  | 1.0  |
| **K**      | 2.7 | 1.1  | 1.1  | 0.0  | 23.0 | 2.1  | 9.6  | 1.1  |
| **L**      | 6.6 | 3.1  | 3.1  | 1.0  | 14.9 | 1.4  | 11.1 | 1.7  |
| **M**      | 28.5| 6.5  | 6.5  | 4.9  | 1.6  | 1.6  | 15.4 | 7.3  |
| **N**      | 0.0 | 0.0  | 0.0  | 0.0  | 0.0  | 16.7 | 0.0  | 100  |
| **O**      | 12.8| 0.8  | 0.8  | 2.4  | 17.6 | 12.0 | 28.0 | 7.2  |
| **P**      | 2.4 | 2.0  | 2.0  | 0.4  | 35.3 | 1.2  | 25.9 | 9.4  |
| **Sum**    | 3.6 | 0.0  | 0.0  | 0.0  | 14.5 | 0.0  | 34.5 | 27.3 |
| **SP10**   | 0.0 | 1.4  | 1.4  | 0.0  | 8.5  | 1.4  | 31.0 | 53.5 |

P A Longley, R J Harris
Towards a new digital data infrastructure for urban analysis and modelling

Figure 4. The number of household consumer types per enumeration districts (EDs).

| Number of EDs | 1   | 2    | 3    | 4   |
|---------------|-----|------|------|-----|
| Affluent achievers | 153 | 1.3  | 22.2 | 68.0 | 8.5 |
| Thriving greys   | 191 | 4.2  | 29.8 | 54.5 | 11.5|
| Settled suburbs  | 187 | 1.1  | 17.1 | 62.0 | 19.8|
| Nest builders    | 289 | 0.3  | 19.0 | 56.4 | 24.2|
| Urban venturers  | 123 | 13.0 | 41.5 | 43.1 | 2.4 |
| Country life     | 12  | 8.3  | 66.7 | 16.7 | 8.3 |
| Senior citizens  | 125 | 9.6  | 35.2 | 45.6 | 9.6 |
| Producers        | 255 | 2.0  | 29.4 | 58.0 | 10.6|
| Hard-pressed families | 55  | 3.6  | 47.3 | 43.6 | 5.5 |
| The ‘have-nots’  | 71  | 25.4 | 54.9 | 18.3 | 1.4 |

Note:
1. least heterogeneous (8 or less household types per ED);
2. (9 to 11 household types);
3. (12 to 14 household types);
4. most heterogeneous (15 or more household types per ED);
5. table includes only the 1461 (of 1568) EDs in which 17 or more households were enumerated by the lifestyles survey.

modelling should move away from such crude conceptions of social patterning and neighbourhood function. The analysis does not present a direct comparison in that we have compared a geodemographic classification based on data that have been aggregated to the ED level with a lifestyles classification that has been based on individual observations. The lifestyles classification has subsequently been aggregated to the ED scale as a convenience to facilitate comparison, and this comparison has the advantage of making use of the most available detail from each of the two data sets. More direct comparisons could be made in either of two ways. First, we might compare the classification of individual lifestyles data with the results of a cluster analysis of individual-level census data (the Sample of Anonymised Records) at the district scale. Or, second, we might aggregate the lifestyles data to the ED level prior to clustering and then compare the results with the cluster analysis of census data. These should each be the focus of further research.
4.4 ‘Pen portraits’: an outline appraisal

The classification typology developed above is grounded in data and, the vagaries of
the conduct and response to the lifestyle survey aside, in an analytical rigorous way.
The more recent geodemographic systems have sought to ‘freshen up’ census-based
classifications with reference to ancillary data sources that are more recent, relevant,
and detailed with respect to consumer behaviour. Such sources have variously included
the General Household Survey, Family Expenditure Survey, National Readership
Survey, and a range of service industry sources. This potentially brings a wealth of
detail to geodemographic classification, although not at the fine spatial scales for
which classifiers are used to discriminate behavioural types—indeed they are rarely
capable of statistically valid comparison at finer spatial scales than the district level.
Thus, in practice, ancillary sources remain external to the classification procedure but
are used to provide ‘thick descriptors’ of the classification ex post facto. Of course this
procedure is inherently unscientific and potentially introduces a number of scale and
aggregation-induced effects into the interpretation of classifications.

Yet words are more seductive than numbers and the resulting ‘pen portraits’ add
intuitive plausibility to classifications. We can see this, for example, in the SuperProfiles
category of ‘Affluent Achievers’:

“High income families with a lifestyle to match. Detached houses predominate, reflect­
ing the professional status of their owners. Typically living in the stockbroker belts
of the major cities, the Affluent Achiever is likely to own two or more cars, which
are the top of the range, recent purchases, and are needed to pursue an active social
and family life. Affluent Achievers have sophisticated tastes and aspirations. They eat out regularly, go to the theatre and opera and take an active interest in sports (such as cricket, rugby union, and golf). They are able to afford several expensive holidays every year. Financially aware, with a high disposable income, this group invests in both quoted and privatised companies. They are likely to use credit and charge cards and are likely to have private health insurance. Investments are followed closely in broadsheets, such as The Financial Times, The Times, and The Telegraph. For more leisurely reading, Hello, Harpers & Queen, and Vogue are likely to be found in the home of the Affluent Achiever” [source: promotional literature (cited by Brown and Batey (1994). Italics added to highlight similarities with table A1].

Hyperbole aside, this pen picture accords with the characteristics of type A con­
sumer group identified in the lifestyles classification shown in table A1. This suggests
superficial correspondence between the lifestyles and geodemographic classifications,
yet conceals a more heterogeneous reality. Of the 231 high-concentration EDs for
consumer type A, 60 (26%) are found in SP1, another 48 (21%) are in SP2, and an
absolute majority (53%) of such households are scattered across the remaining eight
clusters! Further, a ‘high concentration’ need mean only that about 10% of the con­
sumer-classified households within the ED are of consumer type A. It would thus be
fallacious indeed to characterise the whole ED area as this consumer type when 90%
of the households are of a different consumer type. The magnitude of these differences
suggests that lifestyles analysis should be used to supplement, even replace, conven­
tional geodemographic typologies.

5 Discussion

Previous urban models have been deficient not because data were ‘unscientific’ in
collection but principally because the data models on which they were founded were
outdated, pertained only to coarse zonal aggregations and, perhaps most fundamen­tally, provided only very imperfect and indirect indicators of human decisions and
activity patterns. These deficiencies have become more apparent over time, as the scale, complexity, and diversity of society has increased, and have resulted in two dominant views of urban modelling in the research community. First, some formal urban analysis has 'carried on regardless', yet today undoubtedly accounts for a much reduced real share of intellectual activity in academic geography and planning. Second, some critics have suggested that digital data and urban models can never be up to the task of generating understanding of real-world problems and that the research effort should be channelled into other (by implication, more idiographic and/or small-scale) approaches. Thus academic discourse has become increasingly polarised between those who cling to lingering but increasingly marginal scientific certainties about data and those who refute the notion of any valid domain for quantitative analysis.

However, a third course is developing, as a result of the explosion in the extent and availability of digital data and the improvement in geographical data-handling technologies. This is based on the view that technology does not just cause consumption to fragment but it also empowers us to provide ever richer depictions of the diversity of population characteristics and behaviour within city systems. This is very much the view advocated in this paper and is also consistent with microsimulation approaches (Clarke, 1996). Indeed, following Johnston (1999), we suggest that data-rich GIS-based model building may be poised to move beyond the ‘mosaic metaphor’ which has governed almost all applications of GIS to date towards more convincing depictions of variety in space and time that are consistent with new theory. Description coming before theory is the normal pattern in science and the spirit of what we have described here is very much that data models can be made sensitive to context without sacrificing generality. Such an approach is also consistent with a reinvigorated contribution to rational planning policy.

A problem with this view is that the foundations to data-rich analysis, at least those explored here, are clearly unscientific. Yet our tentative empirical investigation has suggested that clear commonalities may be established between ‘framework’ data, such as the census and geodemographic systems, and new, relevant, and timely lifestyles data. We believe that application of concatenation and conflation procedures (Longley and Goodchild, 1999) to lifestyles data sets offers the prospect of creating vastly enhanced data models of the form and functioning of urban systems. Such models will need to manage error and bias and this may cause some unease in a modelling community that has been more focused on statistical formalism than messy empirical data problems. The social theory fraternity will doubtless be able to identify aspects of anecdotal historiographies that cannot (yet) be represented in digital form. But real-world business and service planning is already embracing the use of such data series and is using lifestyles data as successfully as their geodemographics forbear. The linear project design of conventional social scientific research was never a panacea in practice (Goodchild and Longley, 1999) and the subsequent flight to untested social theory has done much to marginalise academic contribution to rational planning policy. Today’s digital data infrastructure is not by any means perfect but it has much to offer a reinvigorated approach to urban modelling.

Acknowledgements. The authors are very grateful to a lifestyles data company for the supply of data, and to Peter Brown for the supply of information pertaining to the SuperProfiles geodemographic system. This paper was prepared for the CUPUM ’99 Conference, Venice, Italy.
References
Batey P W J, Brown P J B, 1995, “From human ecology to customer targeting: the evolution of geodemographics”, in GIS for Business and Service Planning Eds P Longley, G Clarke (GeoInformation International, Cambridge) pp 77 – 103
Batey P W J, Brown P J B, Corver M, 1999, “Expanding participation in higher education: a geodemographic analysis of recent experience in the English regions” Journal of Geographical Systems forthcoming
Batty M, 1981, “Urban models”, in Quantitative Geography: A British View Eds N Wrigley, R J Bennett (Routledge and Kegan Paul, London) pp 181 – 191
Batty M, Longley P, 1994 Fractal Cities: A Geometry of Form and Function (Academic Press, London)
Batty M, Longley P, 1997, “The fractal city” Architectural Design Profile 129 74 – 83
Beaumont J R, 1991 An Introduction to Market Analysis, in the Concepts and Techniques in Modern Geography series, number 53 (Environmental Publications, Norwich)
Birk M, 1995, “Customer targeting, geodemographics and lifestyles approaches”, in GIS for Business and Service Planning Eds P Longley, G Clarke (GeoInformation International, Cambridge) pp 104 – 149
Birk M, Clarke G P 1995, “Using microsimulation methods to synthesize census data”, in Census Users' Handbook Ed. S Openshaw (GeoInformation International, Cambridge) pp 363 – 387
Brown P J B, 1991, “Exploring geodemographics”, in Handling Geographical Information Eds I Masser, M J Blakemore (Longman, Harlow, Essex) pp 221 – 258
Brown P J B, Batey P W J, 1994, “Characteristics of Super Profile lifestyles and target markets: index tables, pen pictures and geographical distributions”, Super Profile Technical Note 2, Urban Research and Policy Evaluation Regional Research Laboratory, University of Liverpool, Liverpool
CACI Information Services, 1998 The Geodemographic Pocket Book (NTC Publications, Henley-on-Thames)
Clark B D, Gleave M B (Eds), 1973 Social Patterns in Cities Institute of British Geographers Special Publication 5 (Institute of British Geographers, London)
Clarke G P (Ed.), 1996 Microsimulation for Urban and Regional Policy Analysis (Pion, London)
Cole K, 1993, “The 1991 local base and small area statistics”, in The 1991 Census User's Guide Eds A Dale, C March (The Stationery Office, London) pp 201 – 247
Cosijn P N C, Brown P J B, 1993a, “Comparison of an individual level and an aggregate level market analysis system: performance of the lifestyles selector and Super Profiles classification. Part 1: Sources and classification of data” Journal of Targeting, Measurement and Analysis for Marketing 1 257 – 272
Cosijn P N C, Brown P J B, 1993b, “Comparison of an individual level and an aggregate level market analysis system: performance of the lifestyles selector and Super Profiles classification. Part 2: Discriminatory and penetrative power” Journal of Targeting, Measurement and Analysis for Marketing 1 355 – 365
Curry M R, 1995, “GIS and the inevitability of ethical inconsistency”, in Ground Truth: The Social Implications of Geographical Information Systems Ed. J Pickles (Guilford Press, New York) pp 68 – 87
Dixon C, Leach B, 1977 Sampling Methods for Geographical Research Concepts and Techniques in Modern Geography series, number 17 (Geo Books, Norwich)
Flowerdew R, Leventhal B, 1998, “Under the microscope” New Perspectives 18 36 – 38
Goodchild M F, Longley P A, 1999, “The future of GIS and spatial analysis”, in Geographical Information Systems: Principles, Techniques, Management and Applications Eds P A Longley, M F Goodchild, D J Maguire, D W Rhind (John Wiley, New York), volume 1, pp 567 – 580
Goss J, 1995, “Marketing the new marketing: the strategic discourse of geodemographic information systems”, in Ground Truth: The Social Implications of Geographic Information Systems Ed. J Pickles (Guilford Press, New York) pp 130 – 170
Harris R, 1998, “A comparative analysis of a lifestyle and geodemographic typology”, WP-2, School of Geographical Sciences, University of Bristol, Bristol; http://www.bitsmart.com/geodemographics
Harvey D, 1973 Social Justice and the City (Edward Arnold, Sevenoaks, Kent)
Johnston R J, 1999, “Geography and GIS”, in Geographical Information Systems: Principles, Techniques, Management and Applications Eds P A Longley, M F Goodchild, D J Maguire, D W Rhind (John Wiley, New York), volume 1, pp 39 – 47
Longley P, 1998, “GIS and the development of digital urban infrastructure” Environment and Planning B: Planning and Design 25 53 – 56
Towards a new digital data infrastructure for urban analysis and modelling

Longley PA, Goodchild M F, 1999, "Modern geographic information systems and model linking"; copy available from the author

Marsh C, 1993, "The validation of census data: general issues", in The 1991 Census User's Guide Eds A Dale, C March (The Stationery Office, London) pp 155 – 167

Martin D, 1992, "Postcodes and the 1991 Census of Population: issues, problems and prospects" Transactions of the Institute of British Geographers, New Series 17 350 – 357

Moser C A, Kalton G, 1993 Survey Methods in Social Investigation second edition (Heinemann, Oxford)

Openshaw S, 1984 The Modifiable Areal Unit Problem Concepts and Techniques in Modern Geography series, number 38 (Geo Books, Norwich)

Openshaw S, 1996, “Developing GIS-relevant spatial analysis methods”, in Spatial Analysis: Modelling in a GIS Environment Eds P Longley, M Batty (GeoInformation International, Cambridge) pp 55 – 73

Openshaw S, 1998, “Building automated geographical analysis and explanation machines”, in GeoComputation: A Primer Eds P A Longley, S M Brooks, R McDonnell, W D Macmillan (John Wiley, Chichester, Sussex) pp 95 – 115

Pacione M, 1999 Applied Geography (Routledge, London)

Peppers D, Rogers M, 1997 The One-to-one Future: Building Business Relationships One Customer at a Time (Piatkus Books, London)

Rae J, 1998, “Why do you have to mix lifestyle and census data?” Marketing Systems Today 13 6 – 7

Rhind D, 1997 Framework for the World (GeoInformation International, Cambridge)

Sayer R A, 1979, “Understanding urban models versus understanding cities” Environment and Planning A 11 853 – 862

Snowdon R, 1998, “New system could spell end for ABCs” The Times 15 November, page 31

Stapleton C M, 1980, “Reformulation of the family life-cycle concept: implications for residential mobility” Environment and Planning A 12 1103 – 1118

Timms D, 1969 The Urban Mosaic: Towards a Theory of Residential Differentiation (Cambridge University Press, Cambridge)

Utley A, Thompson A, 1999, “Wraps of v-cs’ postcode database” Times Higher Education Supplement 29 January, page 3
Appendix

How to read the table
Taking the 18–24 age group as an example, a global mean of 2% of households per cluster are shown (by the lifestyles data) to have an adult aged 18–24 years resident. [Here 100 is an index value assigned to the global mean (GM) across all the clusters for a given variable.] These young adults are concentrated within cluster I: 'young outgoing singles'. This cluster has 6.5 times (650/100) the average proportion of young adult households, so 13% (650/100×2) of households in cluster I have a person aged 18–24 resident. By comparison, cluster N ('elderly female widows') has no young adult households. Instead, there is an above-average proportion of households with at least one member aged 65 or above.

Table A1

| GM | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 18–24 years | 2 | 0 | 0 | 50 | 50 | 50 | 100 | 0 | 100 | 650 | 150 | 150 | 100 | 0 | 0 | 150 | 450 |
| 25–34 years | 19 | 127 | 186 | 145 | 136 | 186 | 131 | 9 | 218 | 95 | 140 | 150 | 118 | 9 | 0 | 54 | 109 |
| 35–44 years | 21 | 180 | 171 | 209 | 176 | 142 | 128 | 42 | 109 | 61 | 123 | 142 | 95 | 33 | 9 | 61 | 80 |
| 45–54 years | 17 | 182 | 88 | 117 | 117 | 76 | 100 | 211 | 35 | 23 | 100 | 117 | 88 | 135 | 52 | 76 | 70 |
| 55–64 years | 24 | 58 | 16 | 8 | 16 | 4 | 12 | 283 | 0 | 4 | 16 | 29 | 37 | 312 | 350 | 100 | 29 |
| 65 or above | 80 | 108 | 107 | 110 | 106 | 105 | 90 | 106 | 107 | 75 | 101 | 112 | 111 | 98 | 98 | 92 | 95 |
| Gender | 66 | 131 | 119 | 113 | 113 | 122 | 112 | 122 | 128 | 80 | 127 | 122 | 21 | 121 | 31 | 62 | 54 |
| Marital status | 59 | 140 | 137 | 130 | 132 | 137 | 128 | 138 | 138 | 11 | 130 | 123 | 13 | 135 | 5 | 50 | 25 |
| Couple | 11 | 54 | 63 | 81 | 54 | 63 | 63 | 36 | 63 | 118 | 63 | 90 | 345 | 27 | 81 | 145 | 318 |
| Divorced | 16 | 43 | 50 | 50 | 75 | 43 | 93 | 18 | 43 | 468 | 75 | 81 | 312 | 25 | 75 | 131 | 237 |
| Single | 8 | 25 | 0 | 12 | 12 | 0 | 12 | 112 | 0 | 0 | 0 | 12 | 125 | 87 | 875 | 200 | 75 |
| Widowed | 82 | 117 | 69 | 108 | 108 | 66 | 101 | 148 | 46 | 130 | 88 | 95 | 108 | 140 | 145 | 54 | 83 |
| Household size | 12 | 108 | 208 | 125 | 133 | 216 | 116 | 0 | 291 | 25 | 141 | 116 | 91 | 8 | 58 | 116 |
| 1 adult | 10 | 140 | 110 | 160 | 150 | 130 | 110 | 90 | 80 | 120 | 100 | 140 | 60 | 70 | 20 | 80 | 70 |
| 2 adults | 4 | 175 | 100 | 225 | 200 | 125 | 150 | 50 | 75 | 200 | 100 | 125 | 50 | 50 | 0 | 75 | 75 |
| 3 adults | 5 | 77 | 188 | 122 | 100 | 188 | 100 | 11 | 222 | 66 | 133 | 133 | 122 | 11 | 0 | 88 | 177 |
| 4 adults | 4 | 108 | 208 | 125 | 133 | 216 | 116 | 0 | 291 | 25 | 141 | 116 | 91 | 8 | 58 | 116 |
| Children | 62 | 117 | 69 | 108 | 108 | 66 | 101 | 148 | 46 | 130 | 88 | 95 | 108 | 140 | 145 | 54 | 83 |
| No children | 9 | 77 | 188 | 122 | 100 | 188 | 100 | 11 | 222 | 66 | 133 | 133 | 122 | 11 | 0 | 88 | 177 |
| 1 child in household | 12 | 108 | 208 | 125 | 133 | 216 | 116 | 0 | 291 | 25 | 141 | 116 | 91 | 8 | 58 | 116 |
| 2 children | 4 | 75 | 200 | 75 | 125 | 200 | 100 | 0 | 250 | 25 | 150 | 75 | 75 | 0 | 50 | 175 |
| 3 children | 1 | 100 | 200 | 100 | 100 | 100 | 100 | 0 | 100 | 0 | 200 | 300 | 100 | 0 | 0 | 100 | 300 |
| 4 children | 2 | 76 | 207 | 92 | 123 | 184 | 115 | 7 | 261 | 30 | 138 | 153 | 107 | 7 | 0 | 76 | 184 |
| 5–10 years | 11 | 136 | 218 | 136 | 127 | 200 | 100 | 9 | 245 | 36 | 127 | 154 | 118 | 18 | 0 | 72 | 145 |
| 11–15 years | 10 | 138 | 159 | 163 | 161 | 163 | 155 | 23 | 172 | 157 | 153 | 140 | 80 | 17 | 4 | 38 | 55 |
| Economic status | 15 | 80 | 53 | 66 | 86 | 80 | 93 | 173 | 93 | 133 | 120 | 153 | 140 | 106 | 60 | 60 | 80 |
| Employed | 12 | 133 | 91 | 108 | 141 | 141 | 133 | 100 | 175 | 150 | 150 | 108 | 58 | 25 | 33 | 41 |
| Unemployed | 18 | 177 | 161 | 166 | 172 | 177 | 127 | 61 | 200 | 138 | 127 | 100 | 72 | 27 | 11 | 22 | 11 |
| Student | 6 | 266 | 316 | 233 | 200 | 216 | 166 | 33 | 150 | 133 | 83 | 66 | 50 | 0 | 0 | 16 | 0 |
| Retired | 4 | 350 | 350 | 325 | 250 | 175 | 175 | 25 | 100 | 125 | 75 | 25 | 25 | 0 | 0 | 25 | 0 |
### Car details

| Make         | 1 car in household | 2 cars in household | 3 cars in household |
|--------------|--------------------|---------------------|---------------------|
| Antiques     | 49                 | 24                  | 5                   |
| Cooking      |                    |                     |                     |
| Church       |                    |                     |                     |
| Self-catered holidays |            |                     |                     |
| Cruises      |                    |                     |                     |
| Volvo        |                    |                     |                     |
| Fax machine  |                    |                     |                     |
| Ford         |                    |                     |                     |
| Renault      |                    |                     |                     |
| Rover        |                    |                     |                     |
| Toyota       |                    |                     |                     |
| Vauxhall     |                    |                     |                     |
| Volkswagen   |                    |                     |                     |
| Volo         |                    |                     |                     |

### Car details (per annum)

- Under £300: 41, 72
- Over £300: 170, 228

### Make

| Make | 1 car in household | 2 cars in household | 3 cars in household |
|------|--------------------|---------------------|---------------------|
| BMW  | 200                 | 200                 | 200                 |
| Citroen | 300            | 300                 | 300                 |
| Fiat | 200                 | 200                 | 200                 |
| Ford | 23                 | 134                 | 100                 |
| Honda | 200               | 100                 | 100                 |
| Nissan | 160              | 120                 | 120                 |
| Peugeot | 175            | 150                 | 150                 |
| Renault | 166            | 233                 | 166                 |
| Rover | 12                 | 175                 | 150                 |
| Toyota | 200               | 200                 | 200                 |
| Vauxhall | 14              | 142                 | 135                 |
| Volkswagen | 3         | 166                 | 100                 |
| Volo | 250                 | 300                 | 150                 |
| Privately owned | 68    | 135                 | 119                 |
| Under 3 years old | 9    | 244                 | 188                 |

### Consumer goods owned

| Good         | Amount |
|--------------|--------|
| Camcorder    | 13     |
| Fax machine  | 5      |
| Hi-Fi        | 49     |
| Home PC      | 28     |
| Internet connection | 2 | 300 |
| Mobile phone | 16     |
| Satellite TV | 24     |

### Holidays

| Holiday Type | 1 overseas holiday per annum | 2 overseas holidays per annum |
|--------------|------------------------------|------------------------------|
| 1 overseas holiday | 42                 | 263                         |
| 2 overseas holidays | 22                 | 131                         |

### Destinations

| Destination       | Amount   |
|-------------------|----------|
| Australia/New Zealand | 5        |
| Canada            | 7        |
| Caribbean         | 4        |
| Europe            | 35       |
| Mediterranean      | 43       |
| rest of world     | 13       |
| United Kingdom    | 72       |
| USA               | 15       |
| Lakes/mountain    | 10       |

### Interests

| Interest    | Amount |
|-------------|--------|
| Antiques    | 19     |
| Betting     | 8      |
| Bingo       | 10     |
| Church      | 25     |
| Collecting  | 32     |
| Competitions| 31     |
| Cooking     | 45     |
Towards a new digital data infrastructure for urban analysis and modelling

| Interests (continued) | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|----------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Current affairs      | 22 | 250 | 209 | 104 | 122 | 77 | 77 | 145 | 86 | 113 | 68 | 95 | 104 | 81 | 104 | 31 | 45 |
| DIY                  | 39 | 182 | 141 | 74 | 143 | 141 | 51 | 148 | 158 | 64 | 141 | 161 | 71 | 76 | 30 | 30 | 66 |
| Eating-out frequently | 14 | 171 | 107 | 135 | 107 | 97 | 210 | 128 | 91 | 72 | 84 | 87 | 12 | 10 | 10 | 10 | 50 | 64 |
| occasionally         | 40 | 180 | 110 | 100 | 97 | 97 | 95 | 110 | 175 | 67 | 102 | 107 | 110 | 90 | 100 | 57 | 87 |
| rarely               | 31 | 77 | 103 | 90 | 109 | 87 | 103 | 80 | 83 | 45 | 106 | 135 | 93 | 112 | 100 | 103 | 148 |
| regularly            | 26 | 207 | 146 | 130 | 138 | 115 | 111 | 100 | 88 | 230 | 80 | 107 | 92 | 57 | 46 | 38 | 46 |
| Gardening            | 56 | 151 | 128 | 112 | 121 | 114 | 57 | 148 | 116 | 42 | 76 | 126 | 116 | 130 | 87 | 48 | 51 |

### Music

- classical/opera: 24
- easy-listening: 54
- folk: 28
- jazz: 14
- light classical: 40
- pop/rock: 54
- National Lottery: 64
- Photography: 24
- Playing the pools: 43
- Pubs: 38
- Reading: 63
- Sewing: 26
- Theatre: 27
- TV: 8

### Daily newspaper

- Express: 10
- Financial Times: 3
- Guardian: 5
- Independent: 5
- Mail: 21
- Mirror: 17
- Star: 3
- Sun: 27
- Telegraph: 13
- Times: 7

### Alcoholic beverages

#### Beer
- heavy drinker: 5
- light drinker: 23
- medium drinker: 11
- Brandy: 10
- Gin: 9
- Lager

#### Whiskey

#### Health

- Arthritis: 24
- Asthma: 24
- Backaches: 34
- Diabetes: 3
- Earaches: 13
- Headaches: 30
- Industrial accident: 9
- Stomach problems: 24
- Stress: 17
- Smoke?

#### GM

| 2A  | 2B  | 2C  | 2D  | 2E  | 2F  | 2G  | 2H  | 2I  | 2J  | 2K  | 2L  | 2M  | 2N  | 2O  | 2P  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 81  | 90  | 100 | 115 | 130 | 145 | 160 | 175 | 190 | 205 | 220 | 235 | 250 | 265 | 280 | 295 |
| 57  | 66  | 75  | 84  | 93  | 102 | 111 | 120 | 129 | 138 | 147 | 156 | 165 | 174 | 183 | 192 |
| 32  | 41  | 50  | 59  | 68  | 77  | 86  | 95  | 104 | 113 | 122 | 131 | 140 | 149 | 158 | 167 |
| 17  | 26  | 35  | 44  | 53  | 62  | 71  | 80  | 89  | 98  | 107 | 116 | 125 | 134 | 143 | 152 |
| 12  | 21  | 30  | 39  | 48  | 57  | 66  | 75  | 84  | 93  | 102 | 111 | 120 | 129 | 138 | 147 |
| 7   | 16  | 25  | 34  | 43  | 52  | 61  | 70  | 79  | 88  | 97  | 106 | 115 | 124 | 133 | 142 |
| 1   | 10  | 19  | 28  | 37  | 46  | 55  | 64  | 73  | 82  | 91  | 100 | 109 | 118 | 127 | 136 |
| 3   | 12  | 21  | 30  | 39  | 48  | 57  | 66  | 75  | 84  | 93  | 102 | 111 | 120 | 129 | 138 |
| 2   | 11  | 20  | 29  | 38  | 47  | 56  | 65  | 74  | 83  | 92  | 101 | 110 | 119 | 128 | 137 |
| 1   | 10  | 19  | 28  | 37  | 46  | 55  | 64  | 73  | 82  | 91  | 100 | 109 | 118 | 127 | 136 |
| 2   | 11  | 20  | 29  | 38  | 47  | 56  | 65  | 74  | 83  | 92  | 101 | 110 | 119 | 128 | 137 |
| 1   | 10  | 19  | 28  | 37  | 46  | 55  | 64  | 73  | 82  | 91  | 100 | 109 | 118 | 127 | 136 |
| 2   | 11  | 20  | 29  | 38  | 47  | 56  | 65  | 74  | 83  | 92  | 101 | 110 | 119 | 128 | 137 |
| Mail order purchases                | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|-----------------------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Have made                         | 78| 125| 119| 120| 120| 94| 98| 106| 115| 96| 94| 123| 114| 85| 94| 67| 93|
| Alcohol                           | 5 | 480| 300| 140| 160| 40| 80| 140| 80| 100| 40| 120| 80| 40| 60| 20| 20|
| Books                             | 38| 213| 186| 176| 184| 73| 78| 102| 110| 81| 65| 176| 121| 52| 76| 36| 68|
| Bulbs                             | 21| 290| 142| 128| 157| 66| 66| 171| 80| 33| 52| 147| 104| 104| 104| 38| 33|
| CDs/tapes                         | 28| 232| 150| 146| 142| 75| 82| 82| 121| 117| 75| 214| 114| 39| 57| 35| 85|
| Fashion                           | 48| 154| 131| 150| 150| 75| 81| 108| 143| 85| 70| 156| 147| 60| 87| 52| 85|
| Gifts                             | 14| 228| 114| 171| 100| 92| 28| 85| 164| 107| 71| 228| 135| 35| 71| 28| 114|

| Supermarkets regularly visited    | GM| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|-----------------------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Asda                              | 42| 138| 83| 147| 145| 102| 85| 97| 107| 145| 95| 114| 85| 90| 73| 69| 100|
| Co-op                             | 19| 147| 73| 89| 94| 78| 68| 147| 78| 52| 73| 136| 78| 131| 126| 89| 115|
| J Sainsbury                       | 44| 159| 163| 152| 97| 90| 88| 143| 88| 152| 79| 79| 143| 81| 120| 52| 45|
| Kwik Save                         | 23| 104| 56| 65| 78| 78| 65| 117| 100| 56| 86| 160| 82| 113| 100| 95| 173|
| Marks and Spencer                 | 19| 215| 136| 136| 121| 73| 89| 173| 73| 115| 63| 89| 115| 100| 131| 47| 47|
| Safeway                           | 24| 154| 120| 108| 120| 108| 104| 125| 125| 87| 95| 116| 104| 100| 83| 58| 62|
| Sommerfield                       | 17| 182| 105| 76| 117| 64| 88| 152| 70| 82| 64| 141| 94| 105| 117| 64| 88|
| Tesco                             | 61| 126| 116| 114| 113| 109| 109| 109| 116| 113| 111| 98| 113| 91| 70| 52| 88|
| Waitrose                          | 7 | 200| 314| 128| 114| 71| 100| 142| 42| 157| 57| 42| 171| 71| 128| 42| 28|
| Other                             | 42| 140| 83| 76| 73| 90| 59| 150| 95| 73| 80| 157| 90| 92| 111| 71| 159|