INNOVATIONS IN GIS

GIS and Evidence-Based Policy Making

Edited by
Stephen Wise
Max Craglia

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-0-8493-8583-4 (Hardcover)

Library of Congress Cataloging-in-Publication Data

GIS and evidence-based policy making / editors Stephen Wise and Max Craglia.
p. cm. -- (Innovation in GIS)
Includes bibliographical references and index.
ISBN 978-0-8493-8583-4 (alk. paper)
1. Geographic information systems--Government policy. 2. Geospatial data. 3. Information storage and retrieval systems--Geography--Government policy. I. Wise, Stephen. II. Craglia, Massimo. III. Title. IV. Series.

G70.212.W57 2008
910.285--dc22 2007029937

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

3

Geodemographics

Richard Webber

CONTENTS

3.1 Context

Geodemographics is a term used to define an increasingly important field of research that involves the classification of consumers according to the type of residential area in which they live. The practice was pioneered in the early 1970s to assist governments with the identification of inner-city communities for which different policy interventions were appropriate (Webber, 1975; Webber and Craig, 1978). Since the early 1980s, the application has subsequently spread to commercial organizations who have sought to tailor their investments in facilities and in communications to the specific interests of the local communities that they service (Weiss, 1988; Sleight, 2004). Today most of the large consumer-facing international brands use geodemographic classification to improve their business

performance in applications such as retail-site location, the setting of local sales targets, the distribution of promotional material, customer relationship management, and risk management. As governments seek to adopt proven techniques from the private sector, recent years have witnessed a renewed interest in the application of geodemographic classifications in sectors such as policing, health and education, and areas of public sector service provision which absorb high levels of funding but for which responsibility is devolved to local delivery units because of the wide variations in service need at a local level. During recent years the use of geodemographics has extended beyond the United States and the United Kingdom to cover most of continental Europe and much of East Asia.

Because of the geographical nature of the application, most users of geodemographics recognize the need for the investment in some form of information system for manipulating the geographical information they hold regarding the home locations of their customers, the postal, administrative, media, and sales geographies used in their business, and the locational information they hold about their outlets and those of their competitors. However, many geographers have found it more difficult to recognize the differences between conventional GIS and geodemographic information systems than their similarities. This has often led to a failure to recognize the bespoke investments that are needed in software solutions as well as in data and visualization tools in order to sustain effective returns from this form of analysis.

Many successful commercial applications of GIS to the analysis of human behavior involve common elements structured in familiar ways but in a bespoke development. The developer is likely to work to a brief which will list the most critical applications to which the system will be put. An assessment is made from various datasets needed to support the application. These will be referenced to each other and configured within an established set of software tools. Query opportunities will be made available to users via some form of network. Operators will then be trained in the use of the system to support the set of applications agreed at the outset of the project.

Typically the system will then be capable of supporting additional queries. However, in practice, the modifications needed to support extra functions will often need to be handled by information specialists. Such a model, to which real-life applications only approximate, typically proves highly effective in applications which are predictable, involve use by operators on a routine rather than an occasional basis, support operational rather than strategic queries, and where operational savings are easy to quantify and demonstrate. Elsewhere, and commonly in academic and research environments, users make use of powerful GIS packages to undertake a series of bespoke analyses.

The key difference between geodemographic information systems and mainstream GIS is that whereas conventional GIS tools and datasets are application independent, geodemographics involves the structuring of GIS

software and geographical databases in a generic form which is designed to support a general class of users thought to have similar application requirements. The customer of a geodemographic system therefore purchases, or more often leases, an application which is largely prebuilt, and in which different types of data are preconfigured both in relation to each other and standard GIS tools. Such systems are then supported with an ongoing training, consultancy, and updating service which is of a standard level of service and supplied at prices based on a standard rate-card.

Such an approach necessarily reduces the specificity of each application because the product itself is generic. However, the approach does assure users of access to standard industry methods of tackling particular applications. The other principal benefit is the lower cost of access to these applications and, in a commercial environment, the security of knowing that one is no longer at a competitive disadvantage to rivals who may have the resources to design and commission their own systems.

3.2 Origins of Geodemographics: The Classification of Residential Neighborhoods

Geodemographics originated as a distinct concept in 1974. During that year geographers in the United States and in the United Kingdom independently experimented with the concept of a nationwide classification of residential neighborhoods using the finest level of geography for which census statistics were published in these two countries. These were the "block group" in the United States and the "census enumeration district" in the United Kingdom. Using cluster analysis techniques, researchers identified that whereas every census output area was unique, there were nevertheless significant numbers of census output areas whose demographic patterns were broadly similar. By using the computer to search census output areas whose demographics were broadly similar across all the different topics covered by the census, it was possible to identify a limited number of neighborhood types to which every census output area to a varying degree approximated. By examining the key features which differentiated each of these clusters from their respective national averages, it was possible to create statistical profiles to help researchers understand the function that each type of neighborhood played in a complex urban residential system.

What transformed a basic urban research tool into a concept of relevance to a much wider audience was the emergence of tools which could relate residential addresses to these neighborhood clusters on a national basis. This was made possible in the United States by the development of geocoding systems. These allowed researchers to take a list of names and addresses and append block group identifiers to them. Using the correspondence table listing the classification assigned to each block group, this made it possible to code each individual address by a type of neighborhood. Finally,

TABLE 3.1

Variations in Victimization Rates for Different Types of Crimes in North and East Devon [Rates as a Percentage of the Average Rate for the Study Area]

Mosaic Groups	Incidents per 1000 Households	Same Postcode as Offender	Offender Detected
High-income families	69	58	60
Suburban semis	70	58	64
Blue-collar owners	112	95	143
Low rise council	146	145	193
Council flats	318	414	383
Victorian low status	193	216	227
Town houses and flats	118	117	112
Stylish singles	198	225	177
Independent elders	56	58	43
Mortgaged families	98	96	115
Country dwellers	72	72	60

by comparing the proportions of these names and addresses falling within each class of neighborhood with the corresponding proportions for the country as a whole, it became possible to profile an address file, in other words to identify whether the persons' addresses one was analyzing were predominantly from high- or low-income areas, from areas of young people or old, from urban or rural, and from ethnic or white neighborhoods. In the United Kingdom, exactly the same method of analysis could be used provided that the address files contained postcodes, using a correspondence table between postcodes and census output areas.

Table 3.1 provides a good example of how police use information from operational databases to identify variations in the level of victimization experienced in different types of neighborhood within a force area, variations in the success of the police in clearing up the crime, and variations in the concentrations of offenders between neighborhood types. Although the areas of "Independent Elders" generate very few offenders and victimization rates are low, these prosperous retirees have legitimate complaints that the police are relatively ineffectual at apprehending offenders in these areas compared, for example, with areas of low-rise local authority housing.

3.3 Applications of Neighborhood Classification Systems

Although both in the United States and the United Kingdom the principal intended use of neighborhood classifications was for public policy applications, the tool rapidly "escaped" into the private sector. In 1978, by a curious accident of history, Ken Baker, who was the head of statistics at the United Kingdom's largest consumer research company, the British Market

Research Bureau (BMRB), and who was troubled by the possible bias in the location of the respondents to the Target Group Index survey for which he was responsible, attended a seminar on social deprivation in order to evaluate the possible role that a neighborhood classification could play in analyzing respondent bias. Taking away from the seminar a copy of the classification, Ken coded up a 12-month sample of survey respondents with the classification in order to check its representativeness.

As an afterthought Ken decided it might be interesting to examine various consumer behaviors by type of neighborhood and began to realize that neighborhood classifiers linked to market research data provided very interesting and highly actionable insights to consumer marketers (Baker et al., 1979).

Unlike their public sector counterparts, consumer marketers are unable to have questions of interest to them included as questions on a decennial census. Whereas educational administrators can and do require the census to carry a question on educational qualifications and housing policy experts can successfully argue for the inclusion of questions on tenure, number of rooms, accommodation, and in some countries, age of dwelling, consumer marketers have to make do with asking questions on market research surveys whose coverage is typically restricted to a set of 40,000 respondents in any 1 year.

The significance of this is that the data relevant to most public policy issues are available for geographical areas of very great detail, whereas the data relevant to consumer marketers are unlikely to be statistically reliable below the level of the standard region or regional media area. In order to develop advertising campaigns at a local area level, consumer marketers need some method which a geodemographic classification can provide for interpolating reliable estimates of product and service needs for individual streets and communities at local level from random sample data collected at national level. In contrast, public sector professionals do not.

The applications that neighborhood classifications were first used to support were the recruitment of new customers. Businesses, through their advertising agencies, were continuously on the lookout for media which were particularly cost effective in reaching specific target audiences. High levels of sophistication were applied by agencies to the selection and purchase of TV spots. On the contrary, owners of more localized media channels, such as radio, door-to-door distribution, poster sites, and direct mail, were unable to provide the same level of detail about the audiences they could reach, which was ironic since by being more local in their coverage they were potentially much more attractive to advertisers who were interested in reaching tightly defined consumer groups. Table 3.2, which was derived from Testologen, a market research survey in Sweden, would be helpful to businesses in the leisure market with their media targeting as well as with distribution.

In 1979, the U.S. market research organizations Simmons and the United Kingdom's BMRB initiated a service whereby clients could access tabulations of consumer behavior analyzed by a residential neighborhood

TABLE 3.2

Variations in Leisure Activities by Type of Neighborhood [Index Values: 100 Represents the National Average, Sweden]

Mosaic Groups	Very Interested in Hunting	Very Interested in Playing Golf	Own a Caravan	Own a Summer House
A. Well-educated metropolitans	40	149	18	112
B. Low–middle income earners	53	121	35	99
C. Pensioner areas	70	72	66	107
D. Low educated in villages	92	81	84	87
E. Younger low income	58	60	51	51
F. High-income villas	58	175	65	148
G. Terraced houses and villas	74	124	136	116
H. Middle incomes, detached houses	109	92	144	95
I. Countryside	222	50	161	81

Source: Testologen.

(Weiss, 1988). This enabled media owners to link national research profiles to local demographic data so as to create credible statistics showing the goodness-of-fit between the target audiences that were reached by their titles and the target audiences of the most heavily advertised brands. Local newspapers began to equip their sales forces with neighborhood profile of their circulation areas as the centerpiece of their sales proposition. Door-to-door distributors enabled do-it-yourself (DIY) retailers to avoid the waste of having their catalogs dropped through letter boxes in high-rise flats. The post office vigorously promoted the rental of names and addresses selected geodemographically. Poster companies began to break up their stock by the geodemographics of passing drivers and pedestrians.

In parallel the people responsible for the siting of new stores began to recognize the merit of using statistical information as well as experience and judgment in the evaluation of new sites. In businesses whose boards of directors did not subscribe to "seat of the pants" methods, it was often a requirement that a geodemographic profile of a proposed new store's trade area should be included in any investment appraisal.

While the link between neighborhood classifications and market research allowed the targeting of new customers to be undertaken in a more scientific manner, the facility for geocoding (or postcoding) excited the interest of those responsible for the targeting of communications to existing customers. The link allowed a bank or a mail-order company for the first time to examine whether there were differences between the types of neighborhood in which their profitable and unprofitable customers lived; whether their new customers were similar or different to their old customers, in which sorts of neighborhoods they had good payers, and in which bad payers. Likewise the link allowed a mail-order company to compare the demographics of customers between who bought shoes and who bought washing machines.

Using geodemographic classifications, they were able to profile customers at a much more detailed level than was possible using national market research surveys where the list of questions that could be asked of their own customers was limited and the number of respondents too small to support answers other than to very broad-brush questions (Sleight, 2004).

3.4 Methods of Accessing Geodemographic Information

When confronted with these possibilities most marketers realized that it would probably be inappropriate to use traditional GIS to undertake the particular forms of analysis that they required. The systems were too complex in their functionality, too expensive to install, and often unsuited for use other than by specialists. Marketers in any case found it difficult to articulate what they needed, since their needs followed necessarily unpredictable changes in the competitive landscape. For an application which was not operational and therefore had no definable cost savings, it was difficult to prepare a financial justification.

Finally, it was evident that some of business applications, such as the profiling of customers, were not standard features in any GIS software and, while extra functions could be added, this would involve further expense, delay, and uncertainty.

In the United States and the United Kingdom, it therefore became evident to vendors of the classification systems that the majority of client needs could seldom be cost effectively met without the development of a set of integrated software tools for delivery with the classification system itself (Longley and Batty, 2003). While obviously there is an overlap between their functions and those of standard GIS tools, there are a number of notable differences based on a need for the following requirements:

- The requirement to geocode a file, to display the distribution of records by type of neighborhood, and to index this against an external distribution which might, for example, be the national distribution or the distribution of an entire customer file.
- The requirement to accumulate statistics for ad hoc areas and to compare their distribution with that of, for example, the country as a whole or a set of comparator store catchments.
- The requirement to accumulate from a set of base zones a reusable set of user-defined zones, such as store catchments or media areas, to which other data may be accumulated for analysis at a later time. In many instances these areas will be overlapping and not mutually exhaustive in coverage.

- The requirement to rank order these user-defined areas by concentrations of particular demographic segments and to rank order the zones within them.
- The ability to use the neighborhood profiles of a consumer behavior and of a user-defined zone to interpolate relative levels of consumer spend by zone, and to report and rank either behaviors (according to their relative frequency by zone) or zones (by frequency of behavior).

In addition to specific functional features these systems also needed to meet user needs by incorporating a number of databases other than the census, as follows:

- The requirement that tabulations, rankings, and comparisons be available for standard media area definitions
- The requirement to define isochrones around points of interest
- The requirement to define the location and characteristics of key shopping centers
- The requirement that census-based population and household counts be updated on an annual basis with best current estimates

Table 3.3 illustrates how developers and retailers would typically describe differences in the population make up of two different regional shopping centers on the south coast of England, their catchment areas having been defined in terms of 45 min drive times.

TABLE 3.3

Population Characteristics of Two Shopping Catchment Areas (Defined as 45 min Drive Times) in Plymouth and Bournemouth

Mosaic U.K. Groups	Plymouth %	Bournemouth %	Plymouth as % of Bournemouth %
A. Symbols of success	3.75	10.19	37
B. Happy families	15.16	11.55	131
C. Suburban comfort	15.78	16.93	93
D. Ties of community	18.23	11.24	162
E. Urban intelligence	6.91	5.76	120
F. Welfare borderline	5.38	2.82	191
G. Municipal dependency	7.97	1.64	485
H. Blue-collar enterprise	13.73	9.57	144
I. Twilight subsistence	2.29	2.59	89
J. Grey perspectives	7.12	24.15	29
K. Rural isolation	3.63	3.54	103

These analysis tools, as is so often the case, not only incorporated popular methods of analyzing consumer data, with their standard report formats and specialized terminology, but, in due course, met the industry's needs for standards for negotiation. For example, when advertisers buy from media owners or retailers negotiate with developers, both parties want to use a common currency, whether in terms of mutually understood neighborhood classifications, industry agreed methods of updating population estimates, and commonly understood formats for displaying area statistics.

The synergy from the use of common standards is enhanced to the extent that third parties begin to make use of them. For example in some national markets, such as Japan, the adoption of geodemographic classifications is often inhibited until the most widely used market research surveys are coded up or until geocoding companies offer geodemographic coding as part of their address recognition systems. Only as a result of user demand will vendors of mailing lists set up arrangements whereby the names they rent can be selected on the basis of geodemographic category. If is often only when key suppliers to a particular vertical market support these classification systems that clients within it adopt the classifications as a standard.

3.5 Relation between Suppliers and Users

Insofar as the tools incorporate forms of analysis that are standard within particular industries there has been a recognition on the part of vendors and users that standard fixed annual fee licensing is a more appropriate charging mechanism than a once off fee linked to a smaller annual support charge.

The annual fee arrangement enables the user who is unsure of the financial payback of the system to enter into an agreement without long-term commitment. The user is assured not just of advice on how to use the system but some measure of strategic consultancy. The annual fee will typically incorporate updates of all the input datasets and automatic reconstruction of their linkages, and automatic upgrades to the software. Such an arrangement is, therefore, not unlike that between a client and a professional services organization.

Large organizations such as Marks & Spencer and McDonalds operate a culture that requires standard solutions to standard requirements throughout their international operations. It is for this reason that the strength of the relationship between a client and a supplier causes a supplier to invest in the development of geodemographic services in new international markets. For example, while evaluating the markets of Hong Kong and Japan for international expansion, Marks & Spencer made it a requirement of its relationship with Experian that the company should use its best endeavors to build geodemographic classifications.

3.6 Internationalization of Geodemographics

Since the first geodemographic systems were launched in the United States and United Kingdom in 1974 equivalent classifications have been introduced to 19 markets around the world as shown in Table 3.4, a process which has been extensively documented by Weiss (2000).

The take up of geodemographics in different countries has depended on a number of factors. In general, it is easy to introduce geodemographics in countries which make available census statistics at a fine level of geographic detail. It is for this reason that systems work more effectively in countries such as Canada, Italy, Peru, Finland, and Sweden (as well as in the United States and the United Kingdom) than in Spain and Germany. On the other hand, in markets where the cost of accessing the data is very high or where there is a virtual monopoly on the linkage between address and census geography, such as Italy, it is difficult to commercialize the service.

In a number of countries where census statistics are unavailable (the Netherlands) or available only at a coarse level of geography (Spain and Germany), geodemographic classifications have been built for finer levels of geography using statistics from sources other than the census. For example, the electoral roll is used extensively as a data input in both Spain and the United Kingdom, whereas in the Netherlands market research interviews and mail-order data are key alternative input sources. In New Zealand, public data on buildings allow geodemographic classifications to be taken right down to the building level.

There is no doubt that in some countries neighborhood differences operate at a much finer scale than in other countries. The United States, China, and Hong Kong are examples of countries where differences tend to be more evident at a coarse than a fine level. In Hong Kong, most blocks of flats are more populous than a U.S. census block. In Italy, France, and the United Kingdom, by contrast, the mesh of social differentiation operates at a much finer scale—in other words to identify the type of neighborhood a consumer lives in, you need to use the demographics for quite a small geographical area around which that consumer lives. In general, it seems that geodemographic differentiation operates at a coarser scale in those countries where population growth is faster. Indeed, within the United Kingdom, it is evident that in cities that grew very rapidly over a limited period of time, such as Glasgow, Liverpool, and Middlesbrough, one can see particular types of neighborhood extending over much larger contiguous residential areas than in those cities that have grown at a slower or more consistent rate, such as London and Bristol.

In each country the classifications are unconstrained. That is to say there is no a priori determination of what the clusters should be. This is left to the computer algorithm which was used to construct these classifications to determine, which it does according to its own optimization criteria

TABLE 3.4
Countries Covered by the Mosaic International Network

Country	Level of Geography	No. of Units	No. of Types	Households per Geographic Unit	Total Household Estimate
Australia	Mosaic microsegment	314,078	41	22	7,002,346
Belgium	Street segment	160,000	30	25	4,000,000
China (part)	Street	17,366	34	2,265	39,338,000
Denmark	Mosaic area	32,594	34	75	2,444,550
Finland	Mosaic unit	192,000	30	20	3,840,000
France	Ilot	246,109	52	50	22,300,000
Germany	Building	15,966,793	38	2	31,933,586
Great Britain	Postcode	1,632,261	52	14	22,851,654
Greece	Postcode/census block	49,143	33	85	4,100,000
Hong Kong	Street block group	2,757	30	570	1,571,490
Japan	Grid cell	184,684	39	400	73,873,600
Netherlands	Six-digit postcode	389,756	41	17	6,625,852
New Zealand	Mesh block	38,365	38	50	1,918,250
Northern Ireland	Postcode	44,219	36	13	574,847
Norway	Electoral unit	13,648	30	180	2,200,000
Republic of Ireland	Mosaic area	17,016	32	40	1,800,000
Spain	Mosaic area	621,408	48	27	16,662,339
Sweden	Mosaic Områden	68,461	34	56	3,833,816
United States	ZIP + 4	20,000,000	60	8	137,750,000

(Webber, 2004). For this reason, the types of neighborhood that are created do differ significantly from one country to another.

In general, in countries with large populations and fine geographic detail more residential differentiation will be apparent. The geodemographic systems in these countries will therefore have more categories. This is why the U.S. systems have more different clusters than the ones in Hong Kong and Ireland. Factors that cause the number of categories to vary include the extent of social housing (none in Brazil or Peru), young singles live away from their parents (which they do more in Canada and Australia than in Spain and Italy), the level of ethnic diversity (high in the United States), and the propensity of old people to retire to geriatric neighborhoods, often by the coast (high in the United Kingdom and Australia).

Notwithstanding the differences in the sources of input data, in the questions covered by the census, in the level of granularity of the census output areas, and the size of each country, there are significant similarities as well as differences among the geodemographic classifications that are created in different countries. Almost every country (except Hong Kong) has a set of clusters characterized by high levels of education, late marriage, employment in service occupations, young age profile, high mobility, and location close to the centers of very large cities. Such clusters, such as the "Elite Urbanas" segment in the Spanish Mosaic (Figure 3.1), almost invariably have high proportions of people working for international companies,

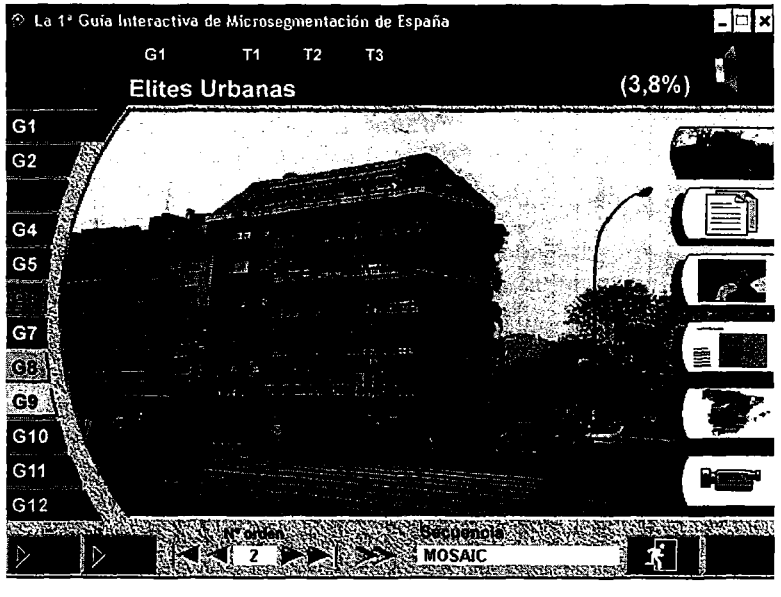

FIGURE 3.1
"Elitas Urbanas" classification of Spanish Mosaic. This is typical of similar categories found in classifications around the world.

TABLE 3.5
"Global Mosaic" Categories

Code	Label
A	Agrarian heartlands
B	Blue-collar self-sufficiency
C	Career-focused materialists
D	Deindustrial legacy
E	Educated cosmopolitans
F	Farming town communities
G	Grays, blue sea, and mountain
H	Hardened dependency
K	Inner-city melting pot
L	Low-income elderly
M	Midscale metro office workers
O	Old wealth
S	Shack and shanty

high levels of consumption of print media, a predilection for eating in quality restaurants, and good knowledge of international trends. This is the most cosmopolitan type of neighborhood that occurs within each country and many of its members would be more at home with their counterparts in similar types of neighborhood in other countries than they would with residents from blue-collar neighborhoods in their own homeland.

In contrast, most geodemographic classifications contain one or more types characterized by blue-collar employees, with low levels of education, high levels of home ownership, and elderly age profiles. These tend to exhibit high levels of community involvement, low levels of crime, and as good places to experience what is quintessential national about a country's cuisine. People from such neighborhoods, if they do travel abroad, want to meet others of a similar background and are least interested in absorbing alien cultures and other countries' food in particular.

Table 3.5 shows an attempt by Experian International to organize the different geodemographic clusters into common global categories. In each country, each type is classified into a single one of 13 different groups from "Old wealth" to "Dependency hardened." Not every one of these 13 categories is found in every market.

3.7 Limitations of Geodemographic Analysis

How well do these classifications work? To answer this question satisfactorily we need to consider the expectations that users may have of the systems. Clearly not everyone who lives in a single census output area will share the same age, housing characteristics, and socioeconomic profile. To the degree that census output areas are themselves heterogeneous,

no geodemographic classification can be as good as a classification built at the person level, at least in relation to predicting the demographics of its members. To this degree a neighborhood classification will always be a second best to person-level demographics in instances for which this is the required basis for targeting.

A different way of looking at their effectiveness is to consider how similar to each other are the different census output areas (or other zones) that are grouped together into a common geodemographic cluster. On this criterion, the systems can be said to be very efficient. Typically the loss of variability of the original dataset that would be lost by grouping output areas into clusters is around 50%. For many key variables, the loss of variance is very much less. Geodemographic clusters are particularly uniform in relation to housing type, tenure, dwelling age, and size. They are also particularly uniform in relation to car ownership, income, and travel to work, in relation to population density and to employment in agriculture. The proportions of the population married or single and the proportions with children are typically very similar across zones in any given category. On the other hand, zones tend to be less uniform in relation to the industrial sector in which employees work, in terms of age structure and, in most countries, ethnic origin. They are least uniform in relation to the proportion of women who work. In general, therefore, they are more uniform and predictive of behaviors that are related to building type and to status than to life stage and industrial structure.

An interesting measure of the efficiency of the U.K. postcode-level classification is that less variability in the input data is lost by grouping residents into 61 geodemographic categories than into 9000 postcode sectors. Knowing that a person lives in a cluster such as "New urban colonists" is more predictive of his or her demographics than knowing that he or she lives in postcode sector N6 4.

Returning to the issue of within-cluster uniformity it is not uncommon for ethnic minorities to be in the minority even in a cluster labeled "Asian enterprise" or for pensioners to be in a minority in a cluster "Sepia memories." One has to be careful not to suppose that labels necessarily apply to everyone in the cluster. However, if one's objective is to target specific groups, whether these be relevant to a public service, such as people burgled in the last year, on to a private organization, such as people who have purchased a new car, the key issue is whether a geodemographic classification is more effective in locating such groups as a targeting system based on any other single criterion, such as age, educational qualification, income, or whatever. Tests have established that over a wider range of consumer behaviors multivariate geodemographic classifications are typically as good as but not better than univariate demographic classifications such as age, gender, income, and so on. These tests have shown that geodemographics is seldom the best discriminator and seldom the worst among comparator demographics.

Whether the systems work or not may also be considered in terms of the nature of error. Some users of geodemographics, such as mail-order companies, use geodemographics as one of a set of criteria for selecting or deselecting customers for communications. In such instances whether or not there is a systematic pattern of bias at a regional or local level is unimportant. On the other hand, for a retailer using the system as an input to an estimate of market potential on a local basis, it is important that neighborhoods of a similar category should behave in a uniform manner throughout the country. In this context such systems would not work very well for predicting the local demand for snow chains in a country with such diverse weather as the United States. Nor would it work well for predicting the level of demand for whiskey or porridge in various supermarket catchment areas in the United Kingdom. These are examples of food products whose consumption varies on a regional level with Scots having particular predilections for both. On the other hand, for the estimation of the proportion of local children likely to apply to university in a particular town, geodemographics should work well because there is unlikely to be systematic error at the cluster level for such behavior. We say unlikely because such an assumption may well not have been tested. In summary, for interpolation from national to local levels, the systems probably work well in the majority of cases but more research would be useful to improve our understanding of the contexts in which the method is likely to be least effective.

3.8 Geodemographics and Government

As mentioned earlier, geodemographic classifications were developed initially to help with the spatial allocation of government programs for inner-city regeneration. Prior to this application, and much of the time since, deprivation has been viewed as a one-dimensional characteristic, much like temperature, and the objective of much statistical analysis has been to identify the associated measures which, once combined, could position each census output area on an appropriate ordinal scale.

One of the uses of geodemographics in this regard was to provide a formal evaluation of the input variables used to measure deprivation. Two examples will suffice. In the 1991 U.K. census, there are two alternative measures of overcrowding available from the census. One measure is the proportion of households living at over one person per room, and the other measure is the proportion living at over 1.5 persons per room. Left to decide which indicator is the more appropriate, housing specialists will debate which level of overcrowding is today considered acceptable. Left to decide which indicator is more appropriate, geodemographers will note that the sorts of neighborhood with high levels of persons living at over 1.5 persons per room are areas of large old divided houses (with very spacious rooms) in very wealthy areas of inner London. Here there are significant numbers of dwellings with two persons in one room.

This behavior is associated with foreign visitors and young singles adapting to very high rents. The alternative measure, in contrast, is associated with very poor peripheral council estates where councils tend to place families with four or more children in three-bedroom houses. The two measures are not measuring differences in level of overcrowding as much as the reason why households may be overcrowded.

A second nice example is taken from a Shanghai regional classification. Table 3.6 illustrates the contrasting demographics of the best and worst educated clusters in the city region. The cluster with the lowest level of unemployment (intensive farming) is also the cluster which seems to have the lowest income, namely the cluster with the most own account farmers. This cluster, despite (or because of) its low incomes has the highest incidence of residents who work 7 days a week, who are not retired, and are not unemployed. Unemployment in this context is not a meaningful indicator of deprivation.

These two examples show the potential use of classifications as methods of evaluating the appropriateness of potential social indicators. Such analysis could be undertaken using regression. However, the use of geodemographics is quicker and simpler. Often patterns which one was not looking

TABLE 3.6

Contrasting Demographies of Top Areas and Intensive Farming in Shanghai [Values Show Proportions of Different Demographic Groups Expressed as a Percentage of the Regional Average]

	Mosaic A1	Mosaic G32
	Comfortable Living/Top Areas	Rural/Intensive Farming
Aged 20–24	79	53
Aged 55–59	144	148
Two children	33	206
Percentage females 25–44 who are single	219	15
No education	38	252
University general and postgraduate	242	5
Farming	1	977
Manufacturing	74	43
Finance and real estate	193	3
Government occupations	199	21
Housewife	46	133
Retired	134	26
Unemployed	86	10
Work 7 days a week	37	199
Self-built house	2	428
House bought from public sector	292	1
Space 40 plus square meters	145	121
Use firewood for heating	0	1529

for emerge unsought—in the manner of true data mining. More often the contextual background provides an explanation of the relationship as well as evidence of it.

The other intended use of geodemographics in deprivation studies was to provide a clearer context for the differentiation of programs according to the type not just the level of deprivation. At the rather crude level of electoral ward that is used to assemble neighborhood statistics in the United Kingdom, this distinction is less evident than it is at postcode level. The original Liverpool Inner Area Study, which spawned the first U.K. neighborhood classification, was able to differentiate three quite distinct types of environment for which quite different policy interventions were appropriate. One of these types consists mostly of dockside council estates with very high levels of unemployment and low levels of skill. Despite their hardship these were the areas with strong community spirit which exhibited some social controls over the most blatant examples of deviant behavior. The problems of these areas was primarily rooted in the decline of the local source of employment, the docks, and the poor level of skills needed for winning jobs elsewhere.

A second type for deprivation occurred among areas of older terraced housing. In days when lack of an inside toilet and incidence of damp and rodents was a cause of major concern, these neighborhoods would score high on a complex deprivation measure. In practice these were the thriving communities, where unemployment was lower than might be expected given the social class profile, but ones where outworn infrastructure and a dilapidated environment were the principal source of stress.

The third type of neighborhood was characterized by large old houses split up into tiny flats occupied by single parents, students, young professionals, dropouts, drug addicts, and prostitutes. Lack of social cohesion and support was the key source of stress in these areas which, in terms of income and educational attainment, fared no worse than the average for the city.

Figure 3.2 illustrates the architectural and environmental characteristics of these three different sorts of neighborhood.

The United Kingdom has, in recent years, seen a renewed interest in public sector applications of geodemographics, particularly in education, health, and crime. In a more evidence-based policy environment more interest is shown in what works where, and it is felt that public services could be more effectively delivered if they were supported by the targeting systems that are routinely used in the commercial sector. The speed at which such innovations can be made depends upon the linking of geodemographic classifications to datasets managed by the public sector. Examples of such databases are the Hospital Episode Statistics (HES), the Driver and Vehicle Licensing Office database, the Land Registry database, the Pupil Level Annual School Census (PLASC), and the British Crime Survey. While some of these generate revenue from an enhanced level of information service to their commercial clients, the last two have provided

(a)

(b)

(c)

FIGURE 3.2
Different types of disadvantaged neighborhoods in Liverpool. (a) Large Victorian villas, young, mobile, rootless singles, (b) Older terraces, strong community network, and (c) Peripheral council estates, high levels of antisocial behavior.

evidence which has been of practical value to universities and to police forces wanting to apply greater selectivity to the manner in which they handle individual clients.

Table 3.7 provide a generic overview of the key different neighborhoods in Great Britain, relating the level of social capital and the degree of trust to the types of crime these neighborhoods are particularly prone to suffer, and the crime prevention and detection strategies that are most likely to work within them.

3.9 Neighborhood Classification Systems in China

Mosaic China was constructed using statistics from the 2000 census for three provinces: Beijing, Guangzhou, and Shanghai. This can be attributed in part to the difficulty of obtaining statistics from other regions, there being no single source of statistics for the whole country, in part to the interest of overseas marketers which at present is concentrated in three large city regions. The total enumerated population living in these three provinces in 2001 was 39,405,000 divided between 17,366 enumeration areas.

The statistics available for these three regions are uniform in definition. For each of their enumeration districts there are some 396 published counts which describe the population in terms of their demographics, housing, and employment characteristics. For example, it is possible to access counts of people by age, marital status, ethnic origin, and number of years' residence within the locality. Features of the housing that are covered in the statistics are the decade in which it was built, the number of rooms in each dwelling, and their floor space. Information is provided on the ownership of the dwellings, on toilet and sanitation facilities, and means of heating. The Chinese authorities also provide information on whether a building is used solely for residential or for mixed residential and commercial purposes. Published information on employment include whether the population is employed, sick, unemployed or studying, the occupations and the industries that people work in, and their level of schooling and/or university qualifications. Information is also provided on number of days worked per week.

From this set of 396 data items, a set of 76 statistically reliable variables were created for use in the classification. Among these are a number of complex variables, such as for example the average number of rooms per person and the average number of square meters per dwelling.

To build the classification, a positive weight was given to only 66 of these variables, the other 10 being used only for analyzing differences between the clusters and not for creating them. Each of these variables was given a weight, reflecting to the level of influence it was thought appropriate to assign to that variable in determining the cluster each enumeration area should be allocated to. This weight was set for each variable on a scale from

TABLE 3.7

Social Capital, Crime, and Policing Methods in Different Neighborhood Types

	Social Capital				Crime Profile					
Neighborhood Type (Mosaic)	Level of Trust	Informal Contacts	Formal Association	Social Capital	Fear of Crime	Crime Level	Clear-Up Rate	Source of Offenders	Type of Crime	Appropriate Options
A. High-income families	High	Low	Medium	Weak community involvement	Fairly low	Low	Low	Imported	High-value burglaries/fraud	Alarms; high-level home security systems; private patrols; car trackers
B. Suburban semis	High	High	High	Strong support for community action	Medium	Low	Medium	Imported	Burglary	Neighborhood watch; geographic policing; target hardening
C. Blue-collar owners		Medium	Medium	Low	High levels of self-reliance	Medium	Medium	Medium	Vehicles/domestic violence/drinking	Neighborhood watch
D. Low rise council	Low	Medium	Low	Strong local knowledge	High	High	High		Petty theft/domestic violence/teenage vandalism	Community development; victim support
E. Council flats	Low	Low	Low	Self-policing gangs	Very high	Very high	High	Indigenous	Gangs/domestic violence/drugs	Community development; victim support; zero tolerance; high-intensity assurance
F. Victorian low income	High	High	Medium	Informal community networks	Moderate	High	Moderate	Indigenous	Burglary/attack/racial harassment	Development of community links; victim support groups; youth initiatives
G. Stylish singles	Low	Low	Low	Low levels of community involvement/students	Moderate	High	Moderate	Indigenous	Equipment theft	Engagement of ethnic leaders; campaign for reporting
H. Town houses and flats	Low	Medium	Low	Centered around local shops	Medium	Moderate	Medium		Burglary	Development of community links; postcode marking
I. Independent elders	High	High	High	Strong networks/people at home during day	Low	Low	Low	Imported	Fraudsters/identity issues	Physical evidence of police, reassurance; personal alarms; cameras/CCTV
J. Mortgaged families		Medium	Medium	High	Moderate networks/ people at work during day	Medium	Medium	Medium	Burglary/vehicles	Child security; crocodile
K. Country dwellers	High	High	High	Strong networks/few perpetrators	Low	Low	Low	Imported	Theft of antiques	Reassurance on response times; dogs; alarms; cameras; postcode marking

one to seven based on experience and judgment rather than formal rules. The variable given the highest weight in determining the cluster allocation was the variable "average square meters per dwelling." Its weight was seven times that of the variable with the lowest weight.

After various experiments with different selections of variable weights and different numbers of clusters, the preferred solution was the one which contained 33 separate clusters. The loss of variance resulting from the grouping of the 17,366 enumeration areas into 33 clusters was only 39.5%. As would be expected this loss of variance was much greater for some of the input variables than for others. For example, the loss of variance of the variable "% dwellings built prior to 1949" was only 10.4% while the loss of variance of the variable "% miscellaneous working status" was as high as 90.2%.

The 33 clusters were then grouped up into a coarser nine-level categorization to help with interpretation and mapping. This grouping contributed a further loss of only 16.3% of the original variance. In other words, by grouping 17,366 areas into only nine separate categories, as much as 44% of the initial variability in the data was still retained. This statistic is significantly higher than in other countries, reflecting the high level of within-cluster uniformity and high level of between-cluster variability in China

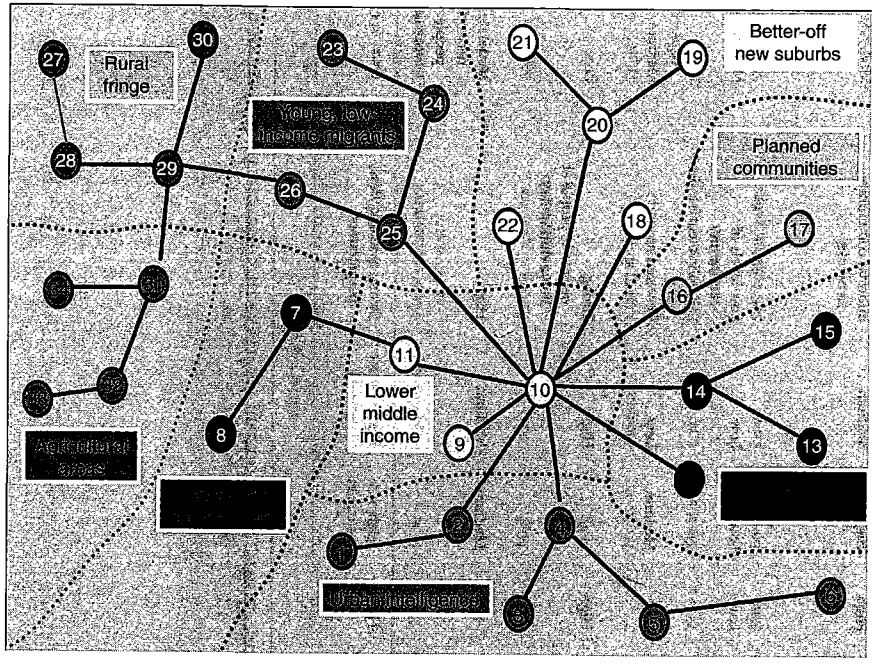

FIGURE 3.3
Minimum spanning tree of the clusters in the Mosaic China. This illustrates the different dimensions whereby Chinese residential neighborhoods are differentiated from each other.

compared with other countries. One reason why this statistic is so high in China is because of the comparatively large average population size of the enumeration areas, over 2200 people. On the other hand, it might be expected that a classification covering three cities in very different climatic zones and with very different industrial bases would not achieve such a high level of within-cluster uniformity.

In general, one wants to achieve a classification in which the population size of the clusters varies within as narrow bands as possible. In the case of China, this variation is larger than in other countries, with the smallest cluster, a set of rural enumeration districts containing concentrations of ethnic minorities, containing only 0.10% of the population. In the largest cluster, areas of older people renting collective flats accounts for 6.77% of the population. This variation in population size is justified by the fact that the smaller of these two clusters contains the largest amount of within-cluster variance of any of the 33 clusters, whereas the "collective flats" cluster, although very large in terms of population, consists of areas which are more similar to each other than is the case in any other cluster.

The minimum spanning tree of the clusters shown in Figure 3.3 provides clear evidence of the large number of different dimensions whereby Chinese residential neighborhoods are differentiated from each other.

3.10 Using Multilevel Geography to Improve Discrimination in the United Kingdom

For many years, U.K. users of neighborhood classification systems had supposed that the finer the resolution of the zoning system used to build a neighborhood classification, the higher would be its performance in discriminating between areas of different types of human behavior. It was this assumption that led developers of neighborhood classification to supplement the statistics that are published by the census offices with additional data items collected from other sources and summarized at the level of the unit postcode, a level of geography one-tenth the average size of a geographical census output area.

In recent years, this assumption has been questioned and it has been suggested that in certain domains, such as the level and type of crime, voting preferences, and children's performance in school, characteristics of the broader community in which a resident lives may have significant incremental predictive power over and behind the characteristics of the immediate microenvironment (Webber, 2004).

To investigate these possible effects when rebuilding its U.K. neighborhood classification based on the results of the 2001 census, Experian decided to create for each of the United Kingdom's 1.5 million full postcodes a series of 1, 2, and 5 km circles. For each of these 4.5 million zones, Experian then created average values on a number of variables derived from the census,

including the percentage of economically active persons engaged in agriculture, in manufacturing and mining, and in services, and the proportion in professional and managerial occupations. In addition, Experian created for each area an average level of population density and, for each postcode centroid, a series of metrics indicating relative accessibility to retail shopping centers, other populated areas, and coastal retirement resorts. Each of these variables were included in the set of variables used to construct the classification and were given weights considered appropriate.

On the completion of the classification, tests were undertaken to establish how effectively the new classification discriminated across a set of 120 customer files and lifestyle respondent databases. These databases were not used to build the classification but were used, in this way, to compare its discriminative power with that on the equivalent classification system, built largely based on 1991 census statistics, that it was designed to replace. Having made this comparison, the analysts responsible for the classification undertook further tests to establish whether the discriminatory performance of the classification could be improved either by up-weighting or by down-weighting the weights given to the "higher level area" variables relative to the weights given to the variables for micro areas.

Tests demonstrated that noticeable improvements could be achieved by significantly up-weighting the weights initially given to the higher level area variables until collectively they accounted for 1.4% of the total weight. By a process of "hill climbing," it was proved possible to generate an improvement in the discriminatory performance of the new classification of 1.8%.

During the process of interpreting the different clusters, it became evident that the weight given to employment data for higher level areas had been particularly effective in isolating council housing in what historically had been important mining towns. Subsequent analysis of information from the British Crime Survey showed that respondents in these clusters experienced very much lower levels of crime than did residents in otherwise similar housing estates in larger cities with an economy more dependent on services. Likewise the use of higher level data was influential in causing inner London neighborhoods to coalesce into a distinctive set of clusters clearly separate from those in outer London and in provincial cities. Accessibility data also seem to have the effect of differentiating villages from which people commuted to nearby towns from other upmarket villages in more remote areas of high landscape value. The use of the accessibility variables was also particularly effective also in causing neighborhoods of the Shetlands and Western Isles which are not dependent on agriculture to be assigned to clusters generally characterized by "rural isolation" rather than by "municipal dependency."

As a result of these tests, Experian have arrived at the opinion that the most effective neighborhood classification is built by using information on employment for multiple levels of geography within a single classification system, while using information on housing and household characteristics only at the very finest level of geography for which they can be obtained.

3.11 Conclusions

Geodemographic classifications have established themselves with consumer marketers as a very useful tool for segmenting consumers. They have been used as a research tool to better understand the profile of the market for specific products and services, media, and brands. They have demonstrated an ability to discriminate both the developed and developing markets. The development of geodemographics has been facilitated by an increased willingness of government organizations to release census statistics in aggregated form at the level of geographical resolution which meets the needs of commercial users.

However, it is evident that the extent to which these new tools are adopted does depend significantly on the level of local skills in the application of geographical information to marketing and communications as well as on the granularity of the postcode system.

By leaving the commercialization of these systems to a small number of international consulting organizations, problems exist for academic users in knowing how to access this information. In the United Kingdom, Experian have come to an arrangement with the Economic and Social Research Council whereby the Mosaic classification can be accessed free of charge by bona fide academic researchers. Notwithstanding this agreement, academic users often have difficulty in learning the applications of the classification systems as well as the meaning of the categories. Because of their confidential nature and the different priorities of commercial clients, it is often difficult to gain access to the results of demonstration projects. Outside the United States, the United Kingdom, and Australia, there is little evidence of access by the academic community to geodemographic classifications.

Nevertheless the information resource does represent a potentially very useful opportunity for academics to improve their understanding of the processes whereby particular groups segregate themselves within urban systems and hence of the contemporary patterns of migration and gentrification which cause neighborhoods to change (Butler and Robson, 2003). These resources should also facilitate a better understanding of the significance of area effects within multilevel modeling and improve the accuracy with which models can be used to interpolate local estimates from national surveys.

However, for these research opportunities to properly inform policy interventions, it is necessary that government organizations should recognize the value to researchers of the geodemographic coding of many of the large operational datasets which they maintain. Datasets in the United Kingdom, such as the HES, the Pupil Level Annual and School Census, and the British Crime Survey, are examples of files which have now been coded by type of residential neighborhood, making it possible to provide robust benchmarks for service delivery standards within highly localized areas. Where, how, and by whom such geodemographically coded datasets

are best analyzed is an important issue. Is this best undertaken within government, by commercial contractors, or by academic research institutes?

While academics and commercial organization can be relied upon to find productive applications for these research assets, there is little evidence of successful application in environments where operations are highly devolved such as policing, health, and education where, on account of the fragmented structure of delivery, there are few mechanisms for improving performance by reference to evidence from national datasets. On the other hand, one should not ignore the power of consumer-oriented Web sites such as "Up My Street" (www.upmystreet.com) to raise awareness of the value of these classifications among intelligent lay audiences or indeed even to school children. The intrinsic appeal of the characterization of streets and addresses to the man in the street may ultimately be an important reason why this approach to organizing geographical information achieves widespread recognition in the long term.

References

Baker, K., Bermingham, J., and McDonald, C., 1979, The utility to market research of the classification of residential neighbourhoods, Paper presented at the Market Research Society Conference, Brighton.

Butler, T. and Robson, G., 2003, *London Calling—The Middle Classes and the Remaking of Inner London* (Oxford: Berg).

Longley, P. and Batty, M. (editors), 2003, *Advanced Spatial Analysis: the CASA Book of GIS*, pp. 233–266 (Redlands, CA: ESRI Press).

Sleight, P., 2004, *Targeting Customers, Third Edition, How to Use Geodemographic and Lifestyle Data in Your Business* (Oxford: NTC Publication).

Webber, R., 1975, Liverpool Social Area Study, 1971 Data, *PRAG Technical Paper No 14* (London: Centre for Environmental Studies).

Webber, R., 2004, Designing geodemographic classifications to meet contemporary business needs, *Journal of Interactive Marketing*, 5(3), 219–237.

Webber, R. and Craig, J., 1978, Socioeconomic classifications of local authority areas, *OPCS Studies on Medical and Population Subjects*, No. 35 (London: HMSO).

Weiss, M., 1988, *The Clustering of America* (New York: Harper and Row).

Weiss, M., 2000, *The Clustered World* (Boston, MA: Little Brown & Co).