6 Modern geographic information systems and model linking

P A Longley and M F Goodchild

6.1 Introduction

Geographical analysis involves the use of a wide range of techniques and models to apply formal, usually quantitative structures to systems in which the prime variables of interest vary significantly across space (Longley and Batty 1996). Such techniques of spatial analysis form a well-defined subset of the larger set of analytic methods, defined by an invariance property: the results of spatial analysis are not invariant under changes in the locations of the objects being analyzed. The tradition has its historical roots in locational analysis (e.g. Haggett 1965), geostatistics and the theory of regionalized variables (Matheron 1971; Cressie 1991) and spatial statistics (e.g. Cliff and Ord 1973), yet there are many reasons why modern geographic information systems (GIS) are causing the field to undergo profound change. Today's GIS (broadly defined) provides a seamless computing environment for every stage of spatial analysis. In this chapter we assess the implications of this state-of-affairs, and its possible implications for the future development of spatial analysis in a GIS environment.

GIS is a fast-developing and multi-faceted technology for producing simplified digital depictions, or models, of the real world. Our ability to apply this technology to create rich depictions and images of geographical reality has been spurred by the innovation of new data capture technologies and the proliferation of secondary digital data sources. Any GIS-based model of the real world requires input of information, and the foundations to GIS-based analysis are provided by an emergent digital data infrastructure, assembled from diverse sources and based upon different data structures (Goodchild 1998). Data models of spatial distributions provide a link between raw data taken from the real world to the spatial analytical models that are used to create higher order generalizations. Much of spatial analysis as conventionally understood involves building 'models of models', and as such the outcome can be no better than the data and data models upon which it is founded. In this chapter we will argue that, in practice, no spatial analysis system is founded on solid rock: rather, the practice of spatial analysis requires us to sink piles into the shifting sands of empirical reality until they are secure enough to support the structure that we wish to erect. The visual medium of today's GIS is slicker and more sophisticated than ever before, yet this medium can in practice obscure a more ambivalent message: that the foundations to analysis are insufficiently robust and sensitive to context. The objective of this chapter is to describe how and why model linkage, in physical and social science alike, requires understanding of the digital data that provide the infrastructural framework to analysis.

In a general sense, infrastructure may be defined as 'the system of services (public and private) which provide a systematic framework for human living arrangements' (Longley 1998). We think of traditional physical infrastructure as being designed with function and foresight in mind and as being created in a rational and orderly manner. It is also fixed to a unique location on the Earth's surface, and as such the immediate spatial context to its usage is obvious and apparent. Yet there are a number of important differences between physical infrastructure as commonly conceived and the new emergent digital infrastructure to GIS-based analysis. The large and complex data sets available today may pertain to unique locations, but they have usually been captured using a range of generic technologies, for a multitude of purposes. They may also be shared between users who have no particular geographical affinities or understanding of sensitivity to context (Goodchild 1997). As such the emergent digital infrastructure to spatial analysis may be less rational and ordered than is desirable for particular applications. Yet GIS also provides us with means of tailoring multi-purpose generic data to the context of particular applications. Later in this chapter we will illustrate this point through case studies in the diverse areas of remote sensing (see Curran et al 1998 for more general arguments) and geodemographics (Harris 1999). If the analytical elegance and theoretical transparency of the spatial analytic tradition is to be reconciled with our messy empirical world, the digital infrastructure to GIS must be designed with the same considerations in mind as its physical counterpart. That is to say, it should be designed with function and foresight in mind and be built on clearly recognizable (though not necessarily conventional) foundations. In many respects the new digital infrastructure to GIS is less rational and ordered than we would like!

GIS provides us with means of tailoring multi-purpose generic data to the context of particular applications

6.2 Spatial analysis in context: past and present

We have described the traditional practice of spatial analysis elsewhere (Goodchild and Longley 1999), and this section borrows from our previous review and interpretation.

6.2.1 The linear project design

Generations of students have been introduced to geographical research through being encouraged to hone their general interest in 'finding out' about something into tightly prescribed and clearly defined hypotheses. The myriad detail of geographical reality is of consummate interest, yet only a tiny part of real world complexity is 'researchable' within the resource constraints of a student dissertation or project. And in practice, the same kinds of resource constraints apply to all surveys and investigations of the real world. In the idealized linear schema shown in Fig. 6.1, research begins with the formulation of hypotheses, and culminates in their formal testing in the spirit of classical statistical inference. In between lies a clear and strict sequence of procedures – choosing a data collection method (and designing a survey schedule, as appropriate), identifying a sample design, piloting, field collection of data (with verification and resampling), collation of results, analysis, and report-writing. This sequence is strictly linear and independent of external events. Hypothesis formulation precedes the

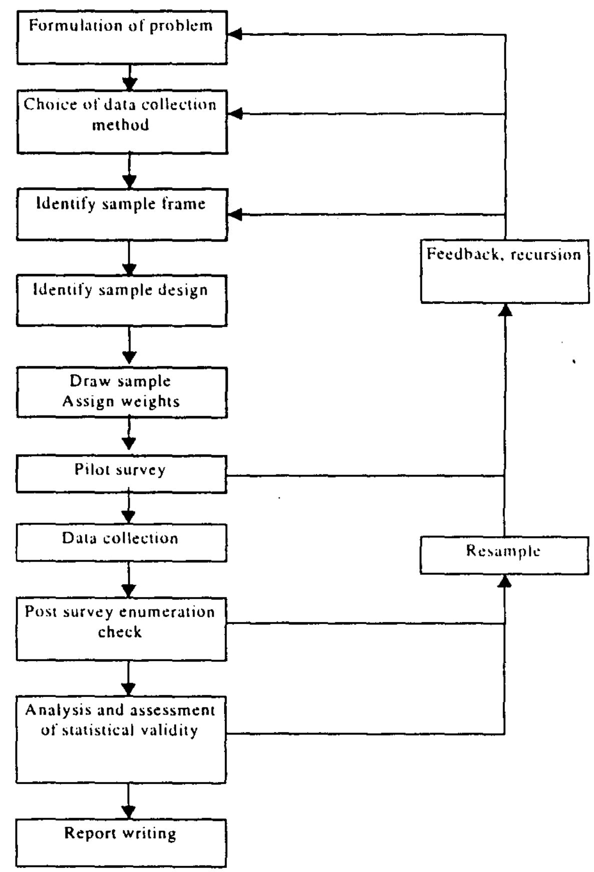

Figure 6.1. The traditional linear project design, with recursion. After Goodchild and Longley 1999.

collection of data and the performance of analysis. Availability of data has no influence on problem definition, and method of analysis is selected with the research hypotheses and characteristics of the new data set in mind (e.g. Levine 1981: Ch 17; Marascuilo and Levin 1983: inside cover; Siegel 1956: inside cover).

This 'linear project design' is to some extent recursive (e.g. major failures identified in the piloting phase may be fed into changes in the way in which the main data collection is carried out), and the search for hypothesis confirmation is reinforced by data exploration. Yet in no prevailing sense can it be described as data- or technique-**driven**—in stark contrast to 'data mining' approaches in which 'the data are allowed to speak for themselves', unconstrained by the conduits of hypothesis formulation. Traditional approaches to science have eschewed 'data-led thinking', instead taking a path towards understanding through cumulative, deductive reasoning.

6.2.2 *Spatial analysis: retrospect*

Implicit in our statement of the remit of spatial analysis at the start of this chapter is the view that the role of techniques and models is to help solve scientific or decision-making problems. The roots to the spatial analysis approach lie in the so-called 'Quantitative Revolution' in Anglo-American geography that was pioneered, *inter alia*, by Berry, Haggett, and Marble in the 1960s (e.g. Berry and Marble 1968; Haggett 1965).

The context to spatial analysis has been transformed over the last 30 years. The reasons for this are well known, and most obviously relate to the precipitous falls in the absolute cost of computer processing power. Looking back from the end of the twentieth century, the tools available to the first generation of quantitative geographers appear no more than rudimentary calculating devices capable of testing only the most restrictive of research hypotheses. The cost of computing in practice limited analysis to a single pass through the data in order to confirm (or refute) a research hypothesis. The rudimentary state of computing often required geographers to write their own (FORTRAN) programs (or adapt those of others), and thus researchers tended to have a specialist understanding of particular techniques rather than a general understanding of a wider range. And finally, the data that were generally available were restricted in quantity and quality by the budgets of particular projects and the lack of secondary georeferenced sources in digital form.

By the late 1980s, general technological changes in computing and the particular innovation of GIS had changed much of this. Increases in computing power led to routine exploration of data and multiple passes through data sets, while computer graphics made the spatial dimension to analysis much more explicit. The cumulative development of algorithms and subroutines led to the development of all-encompassing monolithic GIS packages with the widest range of analytical functionality. Together these developments encouraged spatial analysis in which the sensitivity of model forecasts could be measured against changes in control and predictor variables. These developments fundamentally separated the creation of spatial analytical functions from control of their usage.

As the 1990s progressed, two related processes characterized the subsequent development of GIS. First, the development of the Internet accelerated the break up of early GI systems into reusable (often application-specific) software modules which could be distributed across the Internet and be made available (through software vendors or other interested parties) as 'plug ins' to broader analytical functions. In some sense this development has brought spatial analysis back to the early years, in that responsibility for the design and precise configuration of particular applications has once again devolved to the analyst. It is not just software that has become distributed across the Internet, but also text (at individual and institutional Web-sites) and data: the context to spatial analysis has become digital. The second trend has been for software vendors to package 'fixed menu' offerings of analytical functions (and,

increasingly, data) as low cost 'desktop GIS', permitting GIS-based models to be linked in a quite tightly prescribed range of ways.

These developments have each empowered spatial analysis but in rather different ways: the first has enabled specialist spatial analysts to assemble the widest range of subroutines and programs for everything from designing samples, through grossing of sample estimates, to analysis and report writing with text and graphics; while the second has introduced the novice to a guided and prespecified range of spatial analysis options which should be safe to use and tailored to specific research problems. Although different in conception and packaging, these different developments each represent an extension of the ranges of uses to which computers are put in spatial analysis.

From a historical perspective, the computing input to the linear research design of Fig. 6.1 initially involved limited batch processing of data. It developed into wider interactive data analysis with recursive feedback into hypothesis reformulation, and has today culminated in the use of computer software throughout every stage of research design and analysis—that is, everything from literature searches through to report writing and dissemination. Thus the computer has evolved from an equation-solving and data-handling device to a beginning-to-end environment for the modeling of complex systems and processes. Couclelis (1998) sees this not just as an incremental drift but rather as a sea change in which the IT environment in general has evolved from one in which computing entailed specialized numerical analysis at particular clearly-defined stages in research to one in which computation pervades almost all stages of research from hypothesis formulation to report-writing.

The computer has evolved from an equation-solving and data-handling device to a beginning-to-end environment for the modeling of complex systems and processes

6.2.3 *Spatial analysis today: technique and application, induction and deduction*

In spatial analysis, as elsewhere in science, those that have developed new techniques of analysis have historically devised appropriate means of testing and describing them. (Cliff and Ord 1973; Getis and Ord 1992) Yet even in the early days of computing, this world was changing, as the sharing of specialized subroutines between researchers and packaging of techniques into broader methods (informally, or through the development of subroutine libraries) obviated the need for every application to 'reinvent the methodological wheel'. Thus, for example, urban geographers seeking to classify urban social areas in the 1950s and 1960s shared principal components and factor analysis programs (and agreed upon appropriate distance and rotation methods), and it was not long before applications were spawned in almost every major city in the world.

In point of fact, this example illustrates three facets to the change that was taking place in computing. First, the techniques of classification became 'black box' tools which would generate outcomes ('results') for expert and non-expert users alike. Second, repeated use of an algorithm could gauge the sensitivity of classifications to particular combinations of input variables. Perhaps one of the more enduring legacies of the social area analysis approach, which otherwise amounted to a routine and unimaginative quest to identify communality of social structure between seemingly diverse case studies (see Berry 1971; Clark and Gleave 1973; Timms 1969), was that it suggested to human geographers that the measurement and choice of particular quantitative variables tightly prescribes the outcome of spatial analysis. And third, successive applications were associated with refinement of the basic clustering techniques, thereby leading to the cumulative development of computer algorithms and associated subroutines.

Together, then, cumulative developments in computing and technology led to increasing separation of technical development and application, and there was a redeployment of research effort away from development and programming of spatial analysis procedures for particular applications towards study of the sensitivity of packaged techniques to particular combinations of data and technique. In this way the deductive linear project design of Fig. 6.1 was gradually supplemented and to some extent replaced by inductive procedures designed to suggest empirical generalizations.

This repeated 'stimulus–response' approach to computing was not restricted to self-contained, stand-alone procedures such as principal components and factor analysis. The development of computer graphics was also associated with the development of procedures for exploring data (e.g. plots of residuals and potential leverage points), which strengthened the feedbacks shown in Fig. 6.1, and in turn led to the popularization of further new confirmatory analysis tools (such as resistant fit procedures). During the 1990s, a further wave of inductivism has run through spatial analysis with the development of WIMP (windows, icons, mice, pointers) computer interfaces, to the point at which today the medium of computation allows us to evaluate the messages of spatial analysis in (and through) the widest possible range of senses.

6.2.4 *GIS and geocomputation*

Although much of the preceding discussion has been framed in terms of the implications of general changes in the information technology environment for spatial modeling, they are clearly pertinent to any discussion of GIS. For example, the WIMP environment is particularly conducive to the conception and formulation of spatial problems, while today's network servers bring geographically enabled analysis to orders of magnitudes more users than ever before. As with mainstream computing, GIS today provides an open environment for every stage of any spatial analysis project. And the development of better user interfaces has led to the use of GIS in an ever wider range of applications contexts (Longley et al 1999 report estimates of the value of the digital geographical information industry between $10 billion and $14 billion by the mid 1990s). Popularization and diversification has also led to the blending of GIS with other background technologies: today many of the routine functions of spatial analysis are as accessible and easy to use as word-processing, particularly through standard software such as geographically-enabled spreadsheets. 'True GIS' (Elshaw Thrall and Thrall 1999) nevertheless remains a distinctive software environment for the conduct of spatial analysis, by virtue of its data

management, transformation, and linkage functions, the range of analytical techniques which it supports, and its spatial representation and display functions. Much has been written about the linkage of specialist spatial analysis functions to GIS in ways which are consistent with the linear project design of Fig. 6.1 (Anselin and Getis 1992).

Yet emergent computer technologies are also being harnessed to new computation-intensive techniques (see Openshaw 1998; Openshaw and Openshaw 1997). The emergent field of *geocomputation* (Longley et al 1998) is associated with new and novel ways of using computers to address spatial problems, and may engender new generic approaches to spatial problem solving. To date, as Couclelis (1998) observes, geocomputation has encompassed more or less the full 'grab bag' of spatial analysis techniques set within computationally-intensive environments: yet a common thread to most research applications has been their heavy emphasis upon inductive generalization.

As such, the geocomputational approach would seem set on a collision course with the principles of scientific research epitomized by the linear project design. The development of the approach appears to some to be more a reaction to developments in technology rather than science, although others see it as a deeper quest to develop new science using new tools. This is much more than just an academic or moot point: for some, the new automated 'artistic science of statistical analysis' epitomized by the work of Openshaw (1998) and colleagues is an over-egging of a more mundane reality—namely that geocomputation is essentially reactive, a 'black box' technological response, which is diverting attention from more truly active and rigorous use of GIS tools to develop GI Science (Couclelis 1998; Goodchild 1992). Technology empowers us with tools, so this argument goes, yet conventional wisdom asserts that we need consciously and actively to use them in developing science without surrendering control to the machine. The implication of this view is that better models invariably require better theories, and that these should be transparent and couched within more conventional scientific paradigms (e.g. Burrough 1998). An opposing view is that for too long the esteem in which academics hold traditional deductive science has diverted us from the other equally valid quest for predictive success, and that computationally-intensive inductive generalization is the best vehicle to achieve this. A middle path is suggested by Mandelbrot's (1986) observation that '[d]escription coming before theory is the normal pattern in science', and the view that visualization of pattern creates insight into process (Batty and Longley 1994). Certainly visualization and the creation of virtual realities (VRs) enhances our understanding of scientific problems, and the work of Shiffer (1998) has illustrated the ways in which networked VRs might be used by different users who can each bring their own expertise and insights to collective problem-solving.

From ontology to prediction and practice, debate is likely to characterize the evolution of geocomputation although, as Couclelis (1998) observes, there is no obvious single meta-theoretical framework for activity in prospect. While this state-of-affairs is resonant with recent postmodern fashions, it nevertheless leaves geocomputation with rather looser anchorings than the conventional spatial analysis paradigm. The jury is also still out as to whether the substantive conclusions of such approaches can be sustained under challenge, and whether rich digital descriptions will ever be more than an exploratory prelude to analysis within the more conventional spatial analysis paradigm.

6.3 The changing data infrastructure

Developments in information technology and computing have strong steering influences on the direction and development of spatial analysis research, yet this is only part of the story. However computer models are linked together, all empirical models are ultimately grounded in data. In this section we identify some of the changes to the data environment that are taking place, and use two case studies to illustrate some of the hybrid data modeling methods and techniques that are presently under development. We also speculate on the kinds of spatial analysis that we expect to dominate in the future.

6.3.1 Changes in supply, pricing and access

In physical and social science alike, the costs of data have generally been a (sometimes **the**) major component of the costs of GIS creation. The order of magnitude of data costs reflects a number of technological and secular imperatives which govern the supply, pricing, and access aspects to data availability.

In the early days of GIS, the 'data bottleneck' of (manual or semi-automated) digitizing presented a major impediment to the creation of spatially-referenced databases, particularly if the hard copy source documents were complicated or ambiguous. Early software systems provided (by present day standards) fairly unsophisticated procedures for detecting and correcting the results of error-prone digitizing. Moreover, 'framework' spatial data, such as those created and maintained by national mapping agencies (Rhind 1997) were available only in hard-copy printed form, and in the early days of GIS there was resistance to initiating the task of converting 'legacy' hard copy maps to digital form.

A wealth of digital data has since come into existence. First, and as with computer hardware, new technology is playing an important role. In particular, the wide (selective) availability of global positioning systems (GPS; Lange and Gilbert 1999) makes creation of new digital data sets much more straightforward than hitherto. Second, most national mapping agencies have gradually overcome their initial reluctance to create digital versions of their paper records, while at smaller scales private providers have created a range of digital atlas products. And third, computerized logging of the physical and social environment takes place with ever-increasing frequency, and to ever-greater levels of detail—for example through high-resolution remote sensing of the physical and built environments and the digital encoding of consumer purchasing behavior (through loyalty programs and the development of 'relationship marketing') in the socioeconomic realm.

Yet this has not created a panacea for data modeling. In practice, accurate field recording of data remains an expert task and sound geographical analysis presumes sound data standards (Salgé 1999). Many national mapping agencies (such as Great Britain's Ordnance Survey: Rhind 1997, 1999) have only succeeded in 'going digital' in the face of increasingly stringent public expenditure constraints by recovering vastly

increased proportions of their creation and maintenance costs through user charges: the inevitable consequence is a rationing of framework data on an 'ability to pay' basis. Similarly hawkish data pricing regimes may apply to the data products from the new generation of high-resolution satellite sensors (see Barnsley 1999), while high royalty charges dissuade many business users from census data and census data products in some countries (such as the UK). At the same time, governments are reluctant to fund even their traditional linear project design-driven surveys, in view of the apparent tide of information created using new data capture technologies. With respect to the academic realm, the rise of interdisciplinary science is leading to a higher incidence of jointly-funded projects, and the commonplace situation in which the creators of spatial data may be widely separated from some of the communities of end users. As creators and users of data become more and more separated, in space, time, and intellectual tradition, the ability to describe data becomes increasingly critical. The creator must be able to tell the user about methods, accuracy, formats, and all of the details needed to transfer, open, and make effective use of the data. Moreover the user must be able to determine whether a given data sets meets or falls short of requirements. The term 'metadata' is increasingly used for data description; metadata combine the functions of a library's card catalogue, handling instructions on the outside of the data package, and the quality 'seal of approval'. Metadata have become the key to data sharing through clearinghouses (e.g. the National Geospatial Data Clearinghouse, *www.fgdc.gov*; and the Alexandria Digital Library, *alexandria.ucsb.edu*).

The rise of interdisciplinary science is leading to the situation in which the creators of spatial data may be widely separated from some of the communities of end users.

6.3.2 The changing remit and requirements of modeling

The early years of the spatial analysis paradigm were associated with the development of wide-ranging models of physical and social systems (e.g. Batty 1981). The remit of such models was avowedly ambitious, yet on reflection the data infrastructure was not commensurate with the tasks in hand. A number of commentators (Batty 1976; Sayer 1979) have identified reasons for the subsequent demise of large scale socioeconomic modeling activity, although the innovation of GIS has brought with it a renaissance in model-building activity. Moreover, any decline in large-scale modeling of socio-economic systems has been matched by the rapid growth of environmental modeling, much of it coupled with or otherwise making use of GIS (Goodchild et al 1993; Goodchild et al 1996).

The new is quite different from the old, however. Within the socioeconomic realm, Birkin (1996) has described how the current generation of spatial interaction models, for example, seeks only to model limited (in terms of spatial extent, time frame and attribute range) aspects of urban sub-systems. This in part reflects secular trends in all developed societies away from system-wide planning, yet it also reflects a profound reappraisal of what we now consider to be the appropriate domain and capability of analytical models. Today's urban models are much more data-rich in two respects. First, the revolution in the supply and availability of geographical information means that data no longer represent coarse zonal aggregations, and thus that the data model of spatial distributions bears a closer correspondence with reality. Second, the first generation of urban models used data derived exclusively from public sector sources and which were thus restricted to the limited range of variables of interest to officialdom. Whilst such data can be used, singly or in combination, to create crude **indicators** of human behavior and activity patterns, such indicators bear at best a very imperfect correspondence with reality.

Within the socioeconomic realm, the present status of modeling is rather ambiguous. Within academia, disenchantment with urban modeling leaves it as an area of activity with a significantly reduced real share of intellectual activity compared to, say, twenty years ago. Business applications of data-rich partial models of components of urban systems are buoyant, and today client repeat purchases provide vindication of the validity of spatial interaction and other modeling approaches. Within planning, there has never been a greater need for accurate data and analytical models of urban systems, because the rate, scale, and pace of change has never been greater. Yet, in the UK at least, there is disquiet about the 'predict and provide' approach to planning which has hitherto been based upon aggregate modeling approaches.

6.3.3 Model linkage: towards a new perspective?

The linear project design that presented in Section 6.2.1 presumed that resources were available for a linear, vertically integrated sequence of events. Today's research environment is much less straightforward. The strictures of public expenditure make it less likely that large-scale purpose-specific research will be funded, while information commerce makes it less than unequivocal that the best secondary data will be available. Yet data warehouses are bursting with data that might be combined to create richer profiles of landscapes, morphologies, households, and activity patterns than have ever been created before. While the developing geocomputation paradigm presents us with some 'brute force' mechanisms for searching out generalizations from large and complex data sets, we may have no way of knowing whether such generalizations hold any scientific validity.

A negative view of this research environment would suggest that a price has been put on scientific truth that lies beyond the budget of many researchers. There is some truth in this, yet economic imperatives need also to be viewed in their technological context. In truth, as our retrospective of urban modeling above has illustrated, data collected through the linear project design did not provide a panacea in practice. Today's digital data infrastructure is more detailed, relevant, and up-to-date than ever before. The problem is that this infrastructure is also more piecemeal, and hence possibly ill-founded and unsafe.

GIS has always been an applications-led technology, and this is more true today than ever before. The sophistication of current applications requires a breadth and depth of data that could never have been sustained by established data collection methods. Today's open and desk-top GIS alike are geared towards the analysis of application-specific 'horses for courses' data sets. Such data sets are required to model real-world systems that are dynamic and fast-changing, and thus the time scale

between data collection and availability of secondary analysis needs also to be shortened. Our understanding of physical and social systems alike is now of such sophistication that infrequently collected, aggregate, and surrogate spatial data are simply not good enough. These are all crucial considerations, yet they all lie outside the remit of the linear project design. Are we therefore faced with a stark choice between scientific validity and 'making do' with inappropriate, overly-aggregate, out-of-date indicators? The rejection of Census-based geodemographics in favor of lifestyles (i.e. data warehouse) analysis in much of business geographics suggests that the road to scientific truth is no simple one-way street, and that proponents of inductive data-led thinking have their supporters in the world of application.

Table 6.1. Some characteristics of UK Census data and satellite imagery

Census	Satellite imagery
Vector data structure, e.g. UK SASPAC comprises attribute files compatible with vector boundary data at enumeration district (ED: c. 160 households) scale	Raster data structure: finer scale of spatial resolution (e.g. SPOT pixel size 30m x 30 m)
Infrequently collected – usually decennial	Frequent repeat passes
Residential building types only: averaged across EDs which are assumed uniform areas	All building types identified but not differentiated, plus spatial distribution of urban infrastructure and public open space
Very limited differentiation between non-residential land use functions	Greater differentiation of land use by inference from land cover
Information on socio-economic characteristics of households/individuals	-
Misrepresentation of fine scale spatial distributions—choropleth maps imply uniform within-area density of attributes	Better representation of detailed spatial form, but poor differentiation of function
Comprehensive coverage in theory—but evasion in practice, concentrated in inner city areas	Comprehensive coverage
Temporal inconsistency of (artificial) boundaries	Consistent, subject to minor variation in sensor characteristics
Long historical time frame for temporal analysis, but only in digital form post 1971	Landsat available post 1984

Framed in these terms, one of the big questions for GIS at the turn of the millennium must be: Can the new digital data infrastructure be assembled together in a sufficiently accurate, orderly and rational way to bridge relevance, richness and academic respectability? Goodchild and Longley (1999) use the term 'concatenation' to describe the integration of two or more different data sources, such that the contents of each are accessible in the product. The polygon overlay operation is one simple form of concatenation. They use the complementary term 'conflation' to describe the range of functions that attempt to overcome differences between data sets, or to merge their contents. Conflation thus attempts to replace two or more versions of the same information with a single version that reflects the pooling, or weighted averaging, of the sources.

Can the new digital data infrastructure be assembled together in a sufficiently accurate, orderly and rational way to bridge relevance, richness and academic respectability?

6.3.4 Model linkage in practice: RS–GIS concatenation

Census information and satellite imagery are diverse sources of information, and some of their different characteristics (taking the example of UK Census data) are shown in Table 6.1. Longley and Mesev (1997) use information from the 1991 UK small area census statistics as ancillary information to improve the classification accuracy of a contemporary (LANDSAT TM) image of Bristol. Information from the Census is used to assist in sample training and post-classification sorting. The full procedure is illustrated in Fig. 6.2. The resultant hybridized data set is designed with a specialized purpose in mind—to provide detailed data models of the distribution of population and domestic property. This is used to reappraise conventional analysis of the density at which urban space is occupied—and through comparisons Longley and Mesev (1997) develop density gradient profiles for different categories of urban space filling, such as 'built form', 'residential', 'households', and 'population'. They demonstrate that the differences between these apparently similar categories are more than semantic, and can heavily condition whether and to what extent we might consider density profiles characteristic of particular settlement types. The optimistic message of this work is that, once the differences between different conceptions of 'urbanity' have been clearly grasped, it is possible to develop a range of customized indicators of urban morphology. In this way, customized GIS-based data models are informing our thinking about the ways in which urban settlements fill space, as well as providing detailed information as to the morphology of particular settlement structures.

6.3.5 Model linkage in practice: conflating geodemographics and lifestyles

'Lifestyles' is a broad term that has been used to describe data pertaining to the consumption of a wide range of goods and services by identifiable individuals and households. Lifestyles data originate from a diverse range of sources, such as guarantee card returns, questionnaires attached to nationally circulated prize draw entries, and market research surveys. They are usually georeferenced through the postcode system (e.g. in the UK to the unit postcode, which typically comprises 15 or so addresses in urban areas). At least one UK 'data warehouse' estimates that it holds up-to-date information on 11 million UK households. Such data have evident use for direct marketing, for past consumption habits are key guides to future behavior. Harris (1999) has analyzed the anonymized individual/household records from one particular lifestyles questionnaire which was mailed out in October 1996. The number of

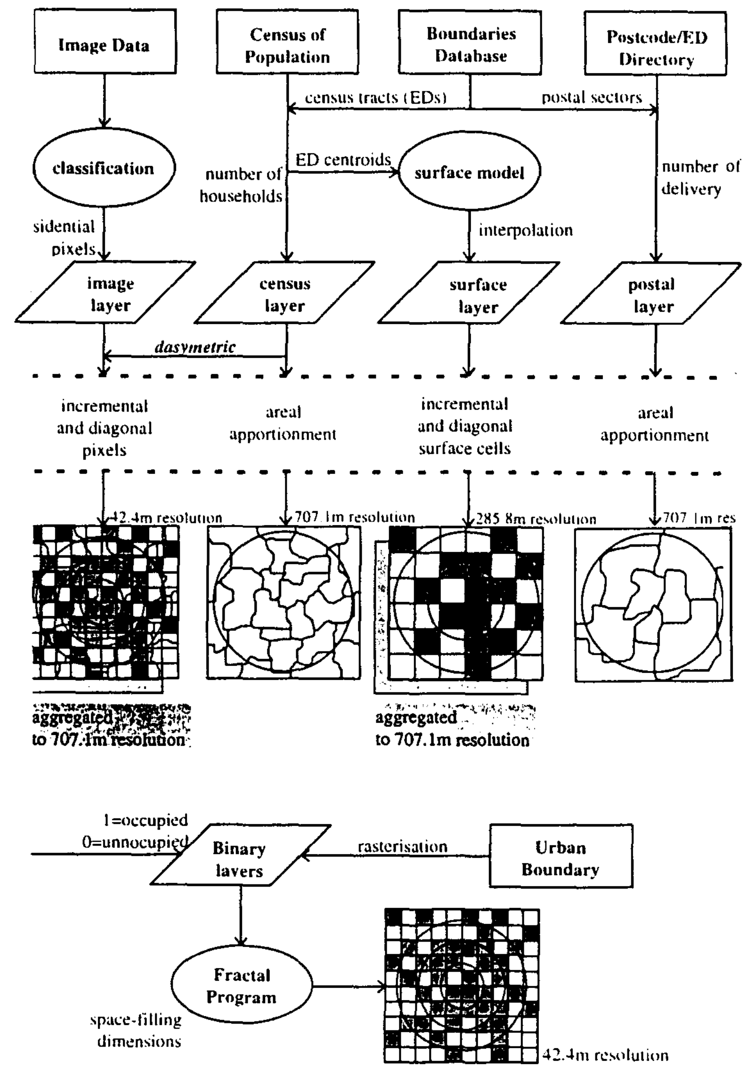

Fig. 6.2 Use of information from the UK Census to assist in sample training and post-classification sorting

respondents to this survey constitutes 10.8% of all households in Bristol, UK (population 636,000): this makes the survey larger in size than a mini census, yet the characteristics of non-respondents are likely to be very unrepresentative of respondents. In recent years, lifestyles approaches have gained some ground as tools for geomarketing at the expense of the use of census and composite geodemographic indicators, because the latter are increasingly out of date (the last UK Census was held in 1991), they are expensive to use because of UK royalty structures and, perhaps most damning of all, the census contains too few variables that bear an identifiable correspondence with consumer behavior (most notably in the UK, because of the absence of an income question in the Census). Some characteristics of lifestyles data vis à vis conventional census-based geodemographics are shown in Table 6.2.

The 'geodemographics–lifestyles' debate thus epitomizes the tensions described in Section 6.3.3 above. Geodemographics is based on tried and trusted techniques and derives from a data set (the Census) which has been designed and implemented using the most rigorous research design principles; and yet at the end of the day, it is out of date, and can supply at best only very imperfect indicators of real-world consumer behavior. Sampling theory tells us that reweighting of largely self-selecting samples on the basis of sub-group response rates is foolhardy; yet survey research practice tells us that quantitative indicators should be direct and transparent, and that survey results are only directly applicable to the population from which the respondents were drawn

Table 6.2. Some characteristics of geodemographics and lifestyles data

Geodemographics (UK)	Lifestyles (UK)
Population (Census) coverage	11 million of roughly 24 million households; little is known about response rates and response rate variability
1991 snapshot data	Continuous updates from waves of mailings
Individual variables of marginal relevance to consumption; no income question and conception of social class arguably outdated	Direct indicators of individual and household consumption
Available at ED level of aggregation	Analysis possible using anonymized individual records at postcode scale of analysis
Constrained to Census geography; integration with postcoded data requires further modeling assumptions	Potential direct application to all other postcoded data
Geodemographic indicators created through application of tried and tested cluster analysis routines	No obvious means of combining unweighted data in a scientifically valid way

(few of us would wholly identify with our digital past-selves who filled out a census form at the start of this decade).

A middle path between these two lies in Batey and Brown's (1995) assertion that lifestyle descriptors can be used as a wrapper to add depth to the labels assigned to different geodemographic groups. Thus, for example, the SuperProfiles category 'affluent achievers' has fairly distinctive Census characteristics in terms of house construction type, socio-economic status and car ownership, to which lifestyle labels about theatre and restaurant patronage, share registers, newspaper readership, and credit card usage are added. The data from which these labels are obtained are in many cases collected by unscientific means or strictly pertain only to coarser aggregations of households. Yet Harris's (1999) cluster analysis of (unweighted) lifestyle data finds some practical validity to this approach: it nevertheless runs rough-shod over conventional views about how scale and aggregation issues should be tackled.

6.4 The future of spatial analysis

Box 6.1 (based upon Goodchild and Longley 1999) summarizes the implications of some of the arguments presented in the preceding discussion, in terms of the kinds of spatial analysis that are likely to become most common in the coming years.

Box 6.1 The implications in terms of the kinds of spatial analysis that are likely to become most common in the coming years.

- analysis of data whose meanings are clearly understood, making it easier for multidisciplinary teams to collaborate;
- analysis of data which are routinely collected in the day-to-day functioning of society and the everyday interactions between humans and computers;
- analysis of data with widespread use, generating demands that can justify the costs of creation and maintenance;
- analysis of data with commercial as well as scientific and problem-solving value, allowing costs to be shared across many sectors;
- methods of analysis with commercial applications, making it more likely that such methods will be implemented in widely available form;
- methods implemented using general standards, allowing them to be linked to other methods using common standards and protocols.

(source: after Goodchild and Longley 1999)

6.5 Concluding comments

Our discussion has highlighted the way in which the advanced information economy of the late 1990s has multiplied the number of potential sources of (rich) digital information, yet in ways which will be less standardized and project-specific than those implied by the linear project design. A major challenge to the GIS community is to devise methods to reconcile diverse data sets with different data structures or spatial referencing systems. Only in this way will GIS be able to tease out the complex relationships that exist between projects, data sets, and analytic techniques in modern science. The self-perception of rigor amongst spatial analysts has hitherto been misplaced because of the vagaries and inadequacies of data quality, resolution and richness: progress requires us to face up to the fact that the linear project design was never a panacea in practice. As Goodchild and Longley (1999) state 'Tomorrow's science will be increasingly driven by complex interactions, as data become increasingly commodified, technology increasingly indispensable to science, and conclusions increasingly consensual. New philosophies of science that reflect today's realities are already overdue.' In the short space of three decades we have moved from a world in which computers were seen as mechanical aids to calculation, to one in which digital information and computing are indispensable to interdisciplinary science. It is still far too early to foresee all of the profound changes that this will bring.

Exercises

1. What is the invariance property mentioned in the introduction? Does it provide a robust test of whether analysis is spatial or not?
2. Consider one of the global radiation models from Chapter 2. Under what circumstances would a GIS be suitable for modeling? What would be the benefits, and what the requirements? Consider all the elements of data, objects, research, and results.
3. Consider the problem of integration of data and suppose that a researcher wishes to apply GIS to the pea model (Chapter 1). Discuss the problems that might be encountered, and give some solutions.
4. Scientists are trained never to approach analysis as a 'black box', but always to understand all of the stages of manipulation of their data. Yet modern computing technology seems to leave the user more and more in the dark about the details of processing. Is this a problem, and how is it likely to be resolved?

7 Multi-scale approaches for geodata

M Molenaar

7.1 Introduction[1]

7.1.1 Four strategies

Topological object relationships in combination with object classification hierarchies appear to be fundamental in the definition of the aggregation rules for spatial objects. Such rules are essential building blocks for the construction of generalization procedures in spatial databases. A model for spatial database generalization can be formulated based on the syntax of the Formal Data Structure (FDS) as proposed in (Molenaar 1989). The syntax of the FDS will be formalized first, then database generalization procedures will be formulated with this syntax.
Four strategies will be explained for generalization:

- *geometry driven generalization*, the change of geometric resolution cells determines the transition from entities at a large scale to new entities at a smaller scale,
- *class driven generalization*, spatial objects at a large scale forming a region under one thematic class are merged for representation at a smaller scale,
- *functional generalization* links objects that are considered as response units in processes defined at different scale levels,
- *structural generalization* gives a stepwise simplification of a spatial process description in an area.

These different strategies will be explained and compared. In the discussion spatial data generalization will be presented as data transformation processes.

7.1.2 Spatial processes at multi-scale levels

Multi-scale approaches are at present a focal point of GIS research because awareness is rising that many processes on the earth surface can only be monitored and managed if they are understood in their geographical context. Part of this context is defined by the scale range at which these processes work. The development of land use in a district, for example is driven by actors at a lower aggregation level such as farmers, residents and companies. Their activities are constrained by socio-economic conditions and infrastructure at regional level and by the macro economic planning at national and supra national level. Another example is development of natural vegetation cover. Its actual state is defined by co-occurrence of species forming vegetation types, which are part of eco-systems. Its development is constrained by climatic conditions, geologic and soil condition and hydrology. Here too we find hierarchical levels of organization.

[1] The content of this chapter has been based on the concepts precribed in Molenaar (1998)