Integrating GIS and Spatial Statistical Tools for the Spatial Analysis of Health-related Data

Jingsheng Ma

Thesis Submitted for the
Degree of Doctor of Philosophy

November 2000

Department of Geography
University of Sheffield
Sheffield, United Kingdom

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my supervisor Professor R. P. Haining for his guidance, help and encouragement through the course of this seven- year part-time study. His comments on each thesis chapter have done much to improve the quality of the thesis on both its contents and its English. I hope that this thesis repays him in some part for his kindness he has shown.

I am grateful to S. M. Wise for his support and encouragement. His thoughtful comments on the manuscript of the thesis and his effort in improving its English are highly appreciated.

Peter Bragg is thanked for helping me get access to departmental computer facilities for this study.

I should like to thank my colleagues, Dr. Gillian Pickup in particular, in Department of Petroleum Engineering, Heriot-Watt University, for their perpetual encouragement, understanding and patience. The department is thanked for its every support that makes the completion of this possible.

No work presented in this thesis would have been possible without the ESRC funded project (grant No R000234470) on which I have worked full-time for three years.

I am most indebted to my wife, Ping Tian, my parents-in-law and my parents for their warm concern, encouragement and sincere understanding and support. To them I dedicate this thesis.

ABSTRACT

Spatial Statistical Analysis (SSA) and Geographical Information Systems (GIS) are instrumental in many areas of geographical study. However, their use tends to be separate one from another. This has prevented their potential in many application areas from being realised. This research is an attempt to bring the two technologies together for a specific application area - health research. There are two research objectives. The first and main objective is to construct a software package – SAGE – by integrating necessary SSA techniques with ARC/INFO, a GIS, to enable the user to undertake a coherent study of area-based health-related data. The second objective is to evaluate and demonstrate SAGE through a case study.

A range of SSA techniques was identified to be useful for addressing typical health questions. A three-tier client-server model was suggested and argued to be the most appropriate for integration as it takes advantages of both the loose-coupling and close-coupling approaches. Under this model, a SSA component forms the client, while ARC/INFO functions as the server. They are linked through the middle tier – the linking agent. The development of SAGE provided experiences useful for developing a generic SSA module in the future for any GIS that conforms to a set of well-defined standard application interfaces.

An empirical study of colorectal cancer (CRC) incidence for the city of Sheffield using SAGE is presented. It shows the usefulness of the SAGE regionalisation tool in constructing an appropriate regional framework for subsequent data analyses and of both exploratory and confirmatory spatial data analysis in exploring the characteristics of CRC incidence. Some weaknesses of SAGE are identified, while remedies for them are suggested. Future work is recommended.

The SAGE User Guide, related publications and the SAGE source and executable code as well as the data used in the case study are enclosed for reference.

CHAPTER 1. INTRODUCTION

1.1. Spatial statistical analysis and GIS

"A spatial data set consists of a collection of measurements or observations on one or more attributes recorded at specified locations" (Haining 1993, pp. 3). Spatial analysis contains three types of analysis: map-based analysis (sometimes called cartographic modelling), mathematical modelling, and spatial statistical analysis (SSA) (Haining *et al* 1996, Haining and Wise 1991). In map-based analysis, layers of spatial information are manipulated and overlaid with simple map algebra to produce new information (Tomlin 1990). In mathematical modelling, spatially distributed systems are analysed using deterministic methods, such as location-allocation analysis (Bailey and Gatrell 1995). In SSA, spatially distributed events are analysed using statistical techniques where the forms of analysis and the interpretation of results depend on the arrangement of the events in some defined space (Haining *et al* 1996). It is SSA that is of concern in this thesis.

SSA comprises four main types of analysis corresponding to four different types of spatial data at different geographical scales (Cressie 1993). These four types of spatial data are distinguished by the stochastic model that is responsible for them. In a general form, a spatial stochastic process is defined as

$$\{\mathbf{Z}(s);\ s \in D\} \qquad (1.1)$$

Where **Z** is a vector of random variables, s is a vector of co-ordinates of a location, and D is an index set of locations in d-dimensional Euclidean space $\Re^d$. The four data types are:

1. geo-statistical data where attribute values (such as soil pH values) are recorded at (sampled) locations on a continuous surface;

1. lattice data where attribute values (such as disease mortality counts) are

recorded for fixed areas that may be regular or irregular in shape or at fixed locations that may form a regular or irregular arrangement;

2. point data where the location of events is the outcome of some random process (and a distinction is drawn between processes where all locations on a continuous surface are available as in the case of seed dispersal across a field and processes where only specific locations are available for the event, such as diseased trees in an orchard);
3. objects where the location of the events is the outcome of a point process and attributes are random sets. Vegetation patches are one example of such events, as are 3D geological blocks.

Spatial statistical analysis for the second and third types of spatial data is referred to as lattice data analysis and point pattern analysis respectively.

The purposes in using spatial statistical methods in geography can be summarised as follows (Haining 1994):

1. to allow the quantitative description of events in geographical space;
2. to allow the systematic exploration of the patterns of events in geographical space and associations between events in space in order to gain a better understanding of the processes that might be responsible for the observed patterns;
3. based on the above, to facilitate the prediction of where events of a disease, for example, may occur; and
4. consequently where appropriate to facilitate intervention and other forms of management and/or monitoring.

Analysis of spatial data using SSA often involves two phases: exploratory spatial data analysis (ESDA) and confirmatory spatial data analysis (CSDA). ESDA emphasises dynamic visualisation as an effective way of identifying data properties at different geographical scales and generating hypotheses. It draws on methods that are resistant in the presence of a small number of aberrant or extreme values, and require few *prior* assumptions about the data. The dynamic visualisation is featured by the 'hot' links of maps and statistical plots (Haslett, *et al* 1990, Fotheringham 1999). In contrast, CSDA is concerned with statistical inference including hypothesis testing, model specification, estimation and criticism.

ESDA is of great value not only for data description but also for model specification and subsequent model criticism, especially in a data-driven approach as opposed to a model-driven approach to model building. The model-driven approach starts from theory, and posits a structure for spatial dependence to be incorporated in a formal model specification (Anselin 1988). On the other hand, the data-driven approach begins with looking at the data properties using ESDA techniques in particular, in order to propose an initial "soft" model for the data. Then the model is fitted to the data and assessed using ESDA techniques. From this process, a "harder" model may be proposed, and a new cycle of fitting and assessment starts. This cycle may go on until a "hard" model is achieved. This process usually draws on different ESDA and CSDA techniques and allows the analysts to introduce subject matter theory into the specification and development of the model (Haining 1993).

1.2. GIS and integration of SSA and GIS

A Geographical Information System (GIS) is a general-purpose computer system for handling spatial data. In terms of system functionality, one definition of a GIS is "a system for capturing, storing, checking, manipulating, analysing and displaying data which are spatially referenced to the Earth" (DoE 1987). Since GIS evolved from many other systems, including computer cartography, database management, CAD and remote sensing, it shares many features with these systems, but one unique feature of GIS is thought to be its ability to perform spatial analytical operations (Goodchild 1987).

GIS ought to be able to answer five types of questions: "what is at...?", "where is it..", "what has changed since ...?", "what patterns exists..?" and "what if ....?" (Heywood *et al* 1998). Questions of the first type are concerned with characteristics at particular locations, while questions of the second are concerned with the locations of areas which have certain characteristics. Questions of the third type seek the differences in characteristics at given locations over time, including questions of the first two types asked with respect to the change of time. Questions of the fourth type are concerned with the kinds of spatial patterns that exist in the data and the locations of spatial patterns such as clusters of large values. Questions of the last type are modelling questions and are concerned with possible outcomes resulting from changes in locations, attribute values or both. Currently GIS is, however, weak in spatial analytical capability to answer questions of the last two types. This has been widely acknowledged (Masser 1988, NCGIA 1989, Openshaw 1990, Haining and Wise 1991, Goodchild *et al* 1992).

This lack of development results partly for the following reasons (Goodchild *et al* 1992). First, the use of GIS has been mainly for the purpose of spatial data management where only rudimentary spatial analytical techniques are required. Second, techniques capable of answering the questions of the last two types are often quite complicated and require the user to have a sound statistical knowledge. Third, many of these techniques have been developed independently of GIS and are not known about by the GIS community. Many of them have to be modified in order to work with GIS.

It has been argued and demonstrated by numerous authors that the integration of both classes of technologies could benefit each other greatly (National Center for Geographical Information Analysis (NCGIA) 1989, Masser 1988, Goodchild 1987, Openshaw 1990, Haining and Wise 1991, Goodchild *et al* 1992). First, the fusion would help GIS to realise its potential as a general purpose tool for handling spatial data, and consequently to meet the GIS market demands for sophisticated spatial analytical capabilities. Second, GIS offers a variety of data models useful for many types of spatial analysis and supports data capturing, management, and display operations that are fundamental for spatial analytical techniques. Thus, re-inventing them would be wasteful.

Classes of SSA techniques have been identified to be in demand in the GIS community and developing intuitive SSA techniques based on both novel and traditional SSA methods for GIS has been called for (Goodchild *et al* 1992). Bailey (1994) provides a review of the integration of spatial analytical tools with GIS. Recent developments on linking spatial statistical techniques and GIS have been reviewed by Anselin and Bao (1996), Levine (1996) and Haining *et al* (1996).

1.3. Health research and the roles of SSA and GIS

One application area that may benefit from the integration of GIS and SSA is health research, which contains two distinct but related areas - epidemiology and health service research. Epidemiology is defined as "the study of the distribution and determinants of health-related states or events in specific populations and application of this study to control of health problems" (Last 1995 pp. 55). Health services research is concerned with health promotion and disease prevention and deals with the need for, provision and use of health services as well as the effect of changes in service provision (Majeed *et al* 1994, Wright *et al* 1998, Haining 1998).

In both areas, a spatial perspective is important because where people live and work determines many factors that influence their health and the treatment they receive from health services. Taking a spatial perspective, epidemiology focuses on describing the spatial variation of health events and elaborating relationships between the events and social and physical environmental aspects. Thus it may help generate the hypotheses of disease aetiology, identify possible causal factors for subsequent and more detailed studies and develop models for disease prediction, prevention and control (Thomas 1990, Haining 1993, Elliott *et al* 1992, Bailey and Gatrell 1995, Cliff *et al* 1985). On the other hand, health service research focuses on identifying the problems of geographically defined populations for health services arising from health inequalities. It tends to address these problems through the locality planning of health provision as well as monitoring and assessing equity in use and access to health services (Benzeval *et al* 1995, Wright *et al* 1998, Haining 1998).

Spatial epidemiology and health services research are closely related. Spatial epidemiology can help the analyst describe the health needs of geographically-defined populations in terms of the spatial distribution of diseases, and gain a better understanding of the links between mortality or morbidity and other aspects such as socio-economic deprivation (Gatrell 1997, Williams and Wright 1998). Knowledge gained can help derive accurate estimates of the needs of a local population and thus deliver appropriate health resources to target the problems of that population. Results of health services research are useful for spatial epidemiology. The study of inequalities in the provision or uptake of preventive health services, for example, may help explain why a disease occurs in a certain place, and help to anticipate where a disease may occur (Meade *et al* 1985).

Spatial statistical techniques, point pattern analysis and lattice data analysis in particular, comprise two sets of techniques particularly useful in health research. Point pattern analysis is used widely in spatial epidemiological studies where disease events are recorded where they occur. The primary interest about such disease events is whether they exhibit any clustering, that is, their locations represent a significant pattern, after allowing for the variation of the underlying population at risk. If they do, questions may be asked about the spatial extent of the clustering and clusters as well as the association of them with surrounding putative sources (Bailey and Gatrell 1995).

Lattice pattern analysis is useful for both spatial epidemiology and health services research (Haining 1993, Bailey and Gatrell 1995). In spatial epidemiology, although point pattern analysis is preferred to the lattice data analysis (Thomas 1990), lattice data analysis is still widely used because some important health-related data are available only at area levels. This is often the case for census data for reasons of confidentiality and for environmental data, such as levels of air pollution, because individual exposure to such a factor cannot be measured accurately (Quinn 1992, Haining 1998). In the context of health services research, the use of lattice data analysis is very common because health services are often organised in areas and need to be analysed at areal levels (Judge and May 1994, Majeed *et al* 1994, Benzeval *et al* 1995).

However, one must note that lattice data analysis requires the lattice of areas or the regional framework to be appropriate for the intended study, otherwise a number of problems can arise. These include the small number problem, and the heteroscedasticity problem (Kennedy 1989, Clayton and Kaldor 1978). In addition, the interpretation of analysis results is subject to the modifiable areal unit problem (MAUP) since the analysis is performed on, at best, only a few of many possible regional frameworks (Openshaw and Taylor 1978). Therefore, regionalisation - a process of constructing an areal framework – is often needed to construct an appropriate areal framework for a study and it may be desirable to consider alternatives to check the sensitivity of results to the area configuration (Haining 1993, Wise *et al* 1997).

GIS is a valuable aid for health research (de Lepper 1995, Gatrell and Löytönen 1998, Gatrell 1999). GIS provides an effective way to manage spatial data and to link them up for studies involving the use of large and multiple data sets that include disease incidences, socio-economic, demographic and environmental measures. It also enables spatial features and attributes to be overlaid, mapped and queried. GIS spatial manipulation capabilities are fundamental to defining study areas around pollutant sources for study (Elliott *et al* 1992, Gatrell and Dunn 1995), to creating appropriate area frameworks (Haining *et al* 1994, Wise *et al* 1997 and Wise *et al* 2000) or to demarcating service areas or communities for managing and delivering health resources (Kivell *et al* 1990, Bullen *et al* 1994).

With respect to health research, recent developments in linking GIS with SSA techniques have strengthened GIS capabilities to analyse spatial patterns of disease events. For example, the Geographical Analysis Machine (GAM) is able to detect patterns automatically from a set of point data although the underlying population is estimated using area-based census data (Openshaw *et al* 1987). Gatrell and Rowlingson (1994) report linking a K function-based point pattern analysis package in S-Plus with a GIS - ARC/INFO. A point pattern analysis technique, devised by Diggle (1990), has been linked with ARC/INFO (Diggle *et* *al* 1990). This technique enables the analyst to detect point patterns in relation to potential hazards, while it still allows for the variation of the underlying population at risk. By contrast, the integration of lattice data analysis techniques with GIS lags behind. Although some techniques have been linked with GIS or available as stand-alone packages, at the time when this project began[1] there was no such package that provides a suite of techniques for regionalisation, ESDA and CSDA in a GIS environment in order to make possible a coherent analysis of area-based health-related data.

1.4. Objectives

There are two objectives in this research. The first one is to construct a software package that integrates a set of regionalisation, ESDA and CSDA techniques with a GIS to enable the user to undertake a coherent analysis of area-based health-related data. In order to achieve this objective, two tasks have to be accomplished. The first task is to identify appropriate lattice data analysis techniques that are required in the analysis of area-based health-related data and are suitable for multidisciplinary users, such as epidemiologists, geographers, public health officers in charge of health planning and health service management, to use. The second task is to explore the ways in which these techniques can be tied together and integrated with a GIS seamlessly to realise fully their potential in this research area. It is also hoped that this will contribute to an understanding of how SSA techniques may be organised so that they can complement one another to support a coherent data analysis and make best use of GIS functionality. A GIS chosen for the integration will be ARC/INFO from ESRI (Environmental System Research Institute, Redland, California, USA). The integrated system is called SAGE (Spatial Analysis in a Geographical Environment).

The second objective is to evaluate and demonstrate the capabilities of SAGE in the analysis of area-based health-related data. A case study of colorectal cancer (CRC) incidence for the city of Sheffield, UK, using SAGE is given. The major concern of this study is to describe the spatial variation of the disease and to

[1] This project began in April 1997.

look at its relationship with socio-economic deprivation.

1.5. Thesis organisation

This thesis is organised into seven chapters, including this chapter. Chapter 2 discusses how to analyse area-based data and some problems arising in the analysis of area-based data. This is followed by a discussion of the strengths and weaknesses of GIS with a reference to SSA and the mutual benefits that may be gained by linking GIS and SSA techniques. Some issues concerning system integration are discussed and the previous examples of integrating GIS and SSA techniques are reviewed.

Chapter 3 aims to show the usefulness of SSA and GIS in the area of health research and is divided into three parts. In the first part, several main types of health studies and questions posed are considered. Examples are given to show how SSA may help answer these questions. Some problems likely to arise in these studies are discussed. The second part summarises some major data sources for health research and issues relating to data confidentiality and data accuracy as well as their implications for SSA. The third part examines the roles of GIS in the analysis of area-based health-related data and what forms of SSA techniques which should be integrated with a GIS to meet the needs of multidisciplinary users.

Based on the discussion in Chapter 3, Chapter 4 reviews some lattice data analysis techniques useful in the analysis of area-based health-related data. These techniques are divided into three categories: those for preparing data for analysis, including regionalisation techniques, and those for performing ESDA and CSDA.

Chapter 5 discusses the development of SAGE from both a system developer perspective and an end-user perspective. From a system developer perspective, it summarises some important aspects concerning system analysis, modelling, design, and implementation as well as testing. From a user perspective, it shows what the user needs to know to start using SAGE. Some drawbacks of SAGE are identified.

Chapter 6 presents a case study of CRC incidence and socio-economic data for the city of Sheffield, UK, using SAGE. It demonstrates how SAGE can be used to describe the spatial variation of the disease incidence and to quantify its relationship with socio-economic deprivation. Some weaknesses in SAGE SSA capabilities for the analysis of area-based data are identified and remedies for them are suggested.

Chapter 7 summarises each of the six chapters and concludes the thesis with suggestions for future work.

An appendix contains the SAGE User Guide, copies of publications relating to this project, a CDROM including the SAGE source and executable code. Due to the confidentiality reason, the CDROM contains all but CRC data for individuals used in the case study.

CHAPTER 2. SPATIAL STATISTICAL ANALYSIS AND GIS

This chapter reviews two sets of issues that are important to this research. The former is concerned with analysing area-based data, while the latter with integrating SSA techniques with GIS.

The chapter is organised into five sections. The first section discusses how to analyse area-based data, while the second section examines some problems likely to arise in the analysis. The third section considers the strengths and weaknesses of GIS for SSA and argues about the mutual benefits that may be gained if SSA techniques are integrated into GIS. The fourth section reviews some approaches to integrating SSA techniques with GIS, and examines their advantages and disadvantages.

2.1. Spatial statistical analysis of area-based data

An area-based data set consists of a collection of observations (typically counts or rates) on one or more attributes taken at a set of regular or irregular areas, called here the regional framework. In health research, attributes may be those relating to health status, socio-economic and environmental characteristics of a population.

One aspect about such data is that the observations are likely to be spatially dependent rather than spatially independent. Spatial dependence is a fundamental property of spatially-referenced data, according to Tobler's (1976) first law of geography, which states that everything is related to everything else and near things are more related than distant things. Another aspect is that observations may be spatially heterogeneous following from the intrinsic uniqueness of locations (Anselin 1988). Spatial dependence and heterogeneity may exhibit at different geographical scales and give rise to complex spatial structures in observations. A third aspect is that a data set is defined on a pre-selected area framework characterised by two factors: scale (the number of areas a region is divided into) and zoning (the way in which that region is partitioned). Areas may differ in terms of some characteristics, such as the size of the underlying population. Results drawn from an analysis are conditional on the framework. Besides these, it should be noted that data does not result from a carefully controlled experimental design undertaken by the investigator specifically for the purpose of analysis but from surveys, is often measured on inadequate and mixed measurement scales and contain spatially varying errors (Anselin 1988).

The remainder of this section considers how area-based data may be analysed given these features. The discussion is focused on conceptual frameworks of statistical analysis of such data. Statistical methods and techniques that fit into the frameworks and are particularly useful for health research will be reviewed in Chapter 4.

As with SSA of other types of spatial data, SSA of area-based data consists of two broad phases: exploratory and confirmatory spatial data analyses - ESDA and CSDA. They are the extensions of traditional exploratory data analysis (EDA) and confirmatory data analysis (CDA) respectively to handling geo-referenced data. EDA is concerned with summarising data and detecting patterns in data, generating hypotheses for data and assessing models. It employs a set of statistical techniques resistant to atypical values or outliers. The use of resistant techniques is essential given the inaccuracy of observations. EDA techniques often take graphical forms (Tukey 1977, Cleveland and McGill1988) and are essentially descriptive rather than inferential although they are also used to assist model specification and subsequent model assessment.

The aims of ESDA extend to detecting spatial patterns in data, to formulating hypotheses and to assessing models based on the geography of the data (Haining *et al* 1998). Clearly, the emphases shift to the spatial aspects of the data. Resistance is again the key feature of ESDA techniques, while cartographic mapping is a central feature of them. With the fast-growth in computer power, dynamic linkage of maps and statistical plots have become a distinct feature of current ESDA (Haslett, *et al* 1990).

For a single variable data set, one conceptual data model for EDA comprises two components - smooth (sometimes called fit) and rough (sometimes called residuals) (Tukey 1977):

$$\text{Data} = \text{smooth} + \text{rough} \tag{2.1}$$

The smooth may include the central tendency and dispersion of the distribution of data, while the rough is the difference between the data value and the value of the smooth component. For example, the median and the inter-quartile range may be used as measures of the central tendency and dispersion of the distribution and can be visually represented using a box plot (Tukey 1977).

Where spatial data is concerned, both absolute locations of areas and relative locations between areas are taken into account in defining the smooth and the rough components, each of which may comprise elements relating to different geographical scales. One possible model for ESDA may be defined as (after Haining *et al* 1998):

$$\begin{aligned}&\text{Data} = \text{smooth (trend, global patterns)} + \\ &\text{rough (local patterns, spatial outliers)}\end{aligned} \tag{2.2}$$

A spatial trend is a large scale gradient (the first order effect), while superimposed on a trend are patterns (the second order effect) (Bailey and Gatrell 1995, Wise *et al* 2000).

Patterns may be global or local, that refers to the whole or part of a study space because different processes operate on the space at different scales (Besag and Newell 1991). Global patterns are considered as part of the smooth component in model 2.2 whereas local patterns as part of the rough component. Patterns are often expressed in terms of spatial autocorrelation or concentration, which measures the propensity of similar data values, either large or small values, to be close to one another across the space (Cliff and Ord 1981, Anselin 1995, Getis and Ord 1992 and 1995, Haining *et al* 1996). Local patterns are often expressed in terms of localised spatial clusters. A cluster refers to a set of adjacent areas that are similar to one another in terms of the values of the variable of interest. A spatial outlier refers to a single area in which a single value differs markedly from the values found in the adjacent areas. Depending on the geographical scale of concern, spatial clusters of large or small values and outliers may be referred to as 'hot' spots or 'cold' spots (Haining 1993, Anselin 1995).

Model 2.2 is just one of many possible models for spatial variation. For example, a model may decompose spatial variation into global and localised patterns. With such a model, the focus of the data analysis is to identify global and local spatial dependence in the data (Cliff and Ord 1981, Getis and Ord 1992 and 1995, Anselin 1995). Note that the spatial dependence identified in these ways may reflect a mixture of first and second order effects with respect to model 2.2.

On the other hand, CDA is concerned with testing hypotheses about data and modelling the variations in the data. CDA involves model specification, fitting, validation and diagnostics (Haining *et al* 1999). CSDA extends CDA to testing hypotheses about spatial aspects of the data, and to modelling spatial variations by taking spatial relationships between the areas into account explicitly (Anselin 1988, Haining 1993, Bailey and Gatrell 1995). In the context of regression modelling, for example, spatial relationships may be specified for the response variable and exploratory variables as well as error terms to take into account different spatial variations and co-variations (Haining 1993, see Section 4.5). They are also required in testing for spatial dependence in model residuals for the purposes of model validation and model diagnostics (Martin 1992, Haining 1994).

It has been recommended that CSDA should be performed with the assistance of ESDA. This is important when the analyst has little knowledge about the spatial data and the underlying processes that yield the data (Haining 1993). Martin (1987) describes a framework for data adaptive modelling. The framework considers SSA as an iterative process of multiple stages and emphasises the alternating use of ESDA and CSDA. ESDA techniques are called upon not only for the preliminary analysis leading to a "soft" model, but also for the analysis at later stages to evaluate the fit of the model. This leads to a refined model and eventually to what might be called a "hard" model.

This framework may be illustrated by regression modelling on a multivariate spatial data set. The process starts with the preliminary examination of spatial data where ESDA techniques are employed to identify the distribution and spatial properties of the data. Bivariate relationships may be explored graphically and statistically. If a linear relationship is suspected between a response variable and some explanatory variables, one may fit an ordinary regression model by least squares and assess model residuals using both ESDA and CSDA methods. If the model residuals are spatially dependent and there is no reason to add new or drop existing explanatory variables into and from the model, a model allowing for spatially correlated error terms may be chosen and fitted by maximum likelihood estimation. Again, ESDA and CSDA methods are called upon to check the model residuals. But this time, ESDA methods may play a more important role than in the previous stage since many formal hypothesis tests are no longer valid (Anselin 1988, see Section 4.5). This cycle may go on until the fit of the model is acceptable.

2.2. Problems in the analysis of area-based data

The analysis of area-based data using statistical methods poses a number of problems. These problems are both methodological and interpretative, stemming from the nature of area-based data briefly mentioned in the previous section. This section will consider how these problems may arise and their implications to the analyst.

2.2.1. Methodological problems

We consider here the methodological problems that stem from the process of spatial dependence, heterogeneity and the area configuration.

2.2.1.1. Spatial dependence

Spatial dependence may arise in the following ways. When an attribute varies continuously across several areas of a regional framework, its measurement for those areas tends to be spatially correlated (Anselin 1988, Haining 1993). This is usually the case for socio-economic data such as census data. Socio-economic attributes often vary continuously at certain scales in space, possibly across several census tracts because they are small and delineated not to reflect the continuity of those attributes but to minimise the efforts of collecting the data (Rhind 1983). Spatial dependence may also be an outcome of the operation of some processes where the way the processes behave is conditional on spatial relationships. The diffusion of an infectious disease involves the close contacts of individuals. The contacts may take place locally or at other places to which people carrying disease agents travel. Other types of competitive processes may also operate in space and give rise to spatial dependence (Haining 1993).

Spatial dependence raises distinctive problems in the analysis of spatial data. When observations are spatially correlated, 'the information content carried by them is less than would have been obtained from independent observations'(Haining 1993, pp.41). In this situation, classical inference procedures based on the assumption that the observations carry equal and independent amounts of information are likely to be misleading. Haining (1993) discusses the problem in estimating the constant mean of a variable. He demonstrates that the sampling variance of the mean estimator under the assumption of non-autocorrelation, underestimates or overestimates the variance of the mean estimator if there is positive or negative spatial autocorrelation in the data.

Classical correlation tests may be affected by the presence of spatial correlation in variables (Lazar 1981). When two variables themselves are positively spatially autocorrelated, the sampling variance of the Pearson correlation coefficient is underestimated (Clifford and Richardson 1985).

In regression modelling, when model residuals are spatially correlated, the usual likelihood ratio tests and F tests are no longer valid for hypothesis testing (Anselin 1988) and the estimates of regression coefficients may be biased. Bailey and Gatrell (1995, pp. 287) show how misleading the fit of a regression model could be without taking spatial autocorrelation in model residuals into account.

Classical model diagnostic techniques, like the Cook-distance that assumes independent errors (Cook and Weisberg 1982), cannot be used without modification (Martin 1992, Haining 1994).

2.2.1.2. Spatial heterogeneity

Spatial heterogeneity refers to spatial differentiation that follows from the intrinsic uniqueness of each location (Anselin 1988). Different areas may respond differently to the same conditions. The same house component (e.g. the number of bedrooms) may have less influence on the house prices in one area than in another (Brunsdon *et al* 1996). People living in different areas may respond to identical health preventive programmes, perhaps because they perceive the programmes differently or different conditions particular to those areas may limit people in one area but not another in accessing services (Holland and Steward 1990).

Spatial heterogeneity may be evidenced in spatial regimes (Anselin 1988, Haining 1990, Fotheringham *et al* 1996 and Brunsdon *et al* 1996). A global model that assumes the parameters are constant over the space may not explain the relationships that exist between variables. In modelling the relationship of house price to the number of bedrooms, Brunsdon *et al* (1996) argue that the marginal price increase associated with an additional bedroom is unlikely to be a fixed rate determined by a global utility but a rate determined by local culture or local knowledge. 'T(t)he value added for an additional bedroom might be greater in a neighbourhood populated by families with children where extra space is likely to be viewed as highly beneficial than in a neighbourhood populated by singles or elderly couples, for whom extra space might be a negative feature'(pp. 283).

It should be noted that the term heterogeneity or non-stationarity could refer to the mean, variance, or covariance of a variable. The example given above is concerned with the heterogeneity in the mean. In later sections, the heterogeneity in the variance will be considered and a term for this is heteroscedasticity.

2.2.1.3. Area configuration

An area framework can be characterised by two factors: scale - the number of areas that a region is partitioned into and zoning - the way that a region is partitioned at a given scale. Since a data set is defined on a regional framework, one has to ask whether the data can be analysed properly especially when a selected regional framework is designed originally for other than the intended study. For example, census tracts have been widely used as a regional framework in epidemiological studies on the small area scale (Elliott *et al* 1992). This is because the socio-economic and demographic data is readily available for census tracts although they were designed for a different purpose. For example, the UK Enumeration Districts (EDs) are designed for optimising the census workloads.

An inappropriate area framework could raise problems in SSA. If the study region is divided into many small areas, each area contains too few samples to derive reliable estimates for the population of concern. The UK EDs, for example, would be too small for a study of the spatial variations of a rare disease such as cancer occurring at a rate less than 5/10000, because each ED contains only a few hundred people and would attract very few, if any, disease cases. Thus, the estimates of the incidence rates are likely to be affected by very small changes to the number of cases. Indeed one extra case in an area may alter dramatically the estimate for that area. This problem is sometimes referred to as the Small Number Problem (SNP) (Kennedy 1989). In addition, if there are too many areas in a data set many spatial statistical techniques may become computationally intractable. In this case, if spatial data is available on a fine geographical scale, one may choose to construct an appropriate area framework on the fine framework through regionalisation (Carstairs 1981, Haining *et al* 1994). Techniques for doing this are reviewed in the next chapter.

Problems may also arise if the study region contains too few areas. In this case, the intra-area variability of a variable is likely to be great, as is the loss of information resulting from data aggregation. Observations for areas (e.g. the mean) become a poor measure for that area. If a study is to investigate possible environmental causes of a disease, the analyst would certainly like to control for socio-economic confounding. The bigger the area is the more diverse the socio-economic circumstances and the more information is lost. Consequently, the results of the analysis could be quite misleading.

The zoning of a region may raise problems for analysis. One problem is heteroscedasticity - a situation where the variance of the random variable is non-constant across areas. This is likely to arise if the population varies across areas (Anselin and Griffith 1988). Estimates of incidence or mortality rates, for example, are less reliable for less-populated areas than for more populated areas. A map of these estimates may be very misleading (Clayton and Kaldor 1987). Statistics, like the Moran's I test under the independent and identically distributed assumption, will be invalid (Cliff and Ord 1981, Walter 1992a and 1992b, Waldhor 1996, Oden 1997). In regression modelling, the presence of heteroscedasticity in the dependent variable may lead to a non-constant variance in the model errors. In this case, standard least squares is no longer valid as a fitting and inference procedure (Anselin and Can 1986, Haining 1993).

The extent of the loss of information on an attribute also depends on whether the zoning reflects the spatial variation of that attribute. Clearly, the better a partition reflects the 'real' variation of a variable, the less information of that variable will be lost.

2.2.2. Interpretative problems

Areal data raises not only methodological problems for statistical analysis but also problems of interpreting analysis results. One problem arises when area-based data is the only source available for a study but the object of the study is to inference individual-level characteristics and relationships (Wrigley *et al* 1996). Often it is impossible to ascribe even the very dominant characteristics of areal data to individuals. An inappropriate inference of individual level relationships from area-level results is termed the ecological fallacy (Yule and Kendall 1950). Openshaw (1984) demonstrates the extent of this problem by comparing the correlation of the same set of variables at individual level and census district level for the city of Florence, Italy. He found that the correlation at census district level is stronger, in an absolute sense, than at individual level. In other words, the correlation coefficients at the individual level and the areal level show a "S" relationship. Using a set of the same variables from the UK 1991 census data and the 2 percent samples of individuals, other authors also confirm this typical "S" relationship between the correlation coefficients at the individual level and the ED level (see Wrigley *et al* 1996). The ecological fallacy problem is related closely to the following problem.

The interpretation is also complicated by the fact that an areal framework employed in a study is only one of many possible partitions of the study region, and conclusions drawn from the study are conditional on that framework – the Modifiable Areal Unit Problem (MAUP) (Kendall 1939, Openshaw and Taylor 1978). MAUP has been studied with respect to both scale and zoning factors. Openshaw (1984) found that the correlation between the percentage of the Republican votes and the percentage of the population aged 60 or over in Iowa could vary between -0.92 to +0.92 over a range of areal frameworks containing 24 units constructed from 99 original counties. Fotheringham and Wong (1991) reported the sensitivity of the results of calibrating a multiple linear regression model to changes of the scale and the partitioning. They concluded that the behaviour of parameter estimates was complex and unpredictable. Sui (1999) reports similar results in a study of the relationships between minorities, income levels, and environmental hazards in USA.

These interpretation problems are found to relate to the spatial dependence. In the case of density data, Arbia (1986) found that the correlation of two variables tends to increase with increasing scale. The strength of the correlation is associated inversely with the level of spatial autocorrelation in the variables. In the context of regression, Green and Flowerdew (1996) found that the relationship between the response variable and explanatory variables varies with the scale, but the regression coefficients do not if the cross-correlation between the response variable and the explanatory variables is absent. This indicates that Fotheringham and Wong's findings (1991) may be the result of the cross-correlation.

Besides these two problems, there are other problems that may affect the interpretation of results. A spatial process may well extend beyond the boundary of a study region. If the study does not have data for the outside regions, any results derived are likely to be influenced by this “edge effect” (Haining 1993, Fotheringham and Rogerson 1993). The quality of spatial data may cause interpretative problems too since some statistical techniques may be affected greatly by errors (Goodchild and Gopal 1989). Since the relative locations between areas are often specified (see Section 4.1.2) by the analyst to reflect his/her belief of inter-area relationships, analytical results would be conditional on the specification of inter-area relationships. Because of the complexity of reality, the analyst might at best be able to describe the major characteristics of the data and probably to build over-simplified models for the data. Some confounding variables are bound to be left out in a study and others might not be controlled sufficiently (Greenland and Morgenstern 1989).

This section has considered issues arising in the analysis of area-based data. It has discussed problems arising from the very nature of such data that must be recognised and dealt with where possible as part of SSA.

2.3. GIS for SSA

This section discusses the strengths and weaknesses of GIS for SSA and the mutual benefits that could be brought for SSA and GIS through their integration as well as types of SSA techniques currently in demand in the GIS community.

2.3.1. GIS - Strengths and weaknesses

A GIS is defined as “a system for capturing, storing, checking, manipulating, analysing and displaying data which are spatially referenced to the Earth" (DoE 1987). GIS shares many features with systems, including auto-cartographic systems, CAD systems and database management systems, and may mean different things to different people (Maguire *et al* 1991). Some people view it as a map processing or display system where spatial data is represented as map layers. These layers may be manipulated by arithmetic in order to search for patterns in the data (Tomlin 1990). Some people view it as a database management system, an integral part of GIS where the data of spatial features and related attributes are stored and managed so that the spatial features, their relationships and the attributes associated with them can be queried and retrieved efficiently (Frank 1988). Another view that dominates in the GIS field is that GIS is a spatial analysis system with powerful spatial analytical capabilities. Spatial analysis refers to a set of analytical and modelling methods that require access to both the attributes and their locations in such primitive classes of spatial features as points, lines, areas and surfaces (Goodchild 1987, Maguire and Dangermond 1991). This view stresses spatial analysis as being the feature of the GIS most distinctive from other systems.

The different views above reflect the strengths of GIS from different angles. What underpins these is the variety of data models and operations associated with the models GIS provides and supports. A data model is an abstraction of reality with a level of completeness and may be viewed at three levels: conceptual, operational or logical, and storage or physical levels (Peuquet 1984, p 252). Data models, at the conceptual level, may take two broad forms (Goodchild *et al* 1992). Either reality is perceived as an empty space populated by objects or a set of fields, each defining the spatial variation of one variable. Objects are normally modelled as points, lines or areas, each having attached values of a set of attributes. Fields may be modelled in at least six ways (Goodchild *et al* 1992): 1) a raster of cells, each defining the average value of the field within each cell; 2) a raster of regularly-spaced point samples; 3) a set of non-overlapping, space-exhausting polygons, each defining a class; 4) a raster of irregularly-spaced point samples; 5) a set of digitised isolines; and 6) a set of non-overlapping space-exhausting triangulates, each assumed to approximate evaluations within the triangle with a simple plane. In either form, attribute values may be at the nominal, ordinal, interval, or ratio scales of measurement.

At the operational level, a data model is represented by data structures. Two common types of structures are vector and raster (Peuquet 1984). In a vector-based model, spatial features are represented as a series of 2D or 3D points, while in a raster-based model spatial features are described as "polygonal units of space in a matrix" (Maguria and Dangermond 1991, pp. 320). Vector-based models are divided into two classes - topological and non-topological models, depending on whether the topological relations (adjacency, containment, or distance-away) between spatial features are defined explicitly. At the storage level, a data model is mapped to one or more databases or related files containing spatial features and attributes managed by database management systems (Peuquet 1984, Heywood *et al* 1998). Different types of spatial features are often stored in separate data sets.

Most GIS support a set of basic data manipulations and analysis functions, including generalisation, proximate analysis, map overlay, measurement and data query (Dangermond 1984, Maguire and Dangermond 1991, Heywood *et al* 1998). Generalisation refers to operations that smooth and aggregate spatial features and attributes. Polygon dissolving is one such operation. Proximate analysis is another important operation that identifies the proximate of spatial features in either a metric space or a non-metric space such as the topological space. For example, with this operation one can define a buffer around a given point, a line or a polygon, possibly representing a pollution source. A unit-wide buffer of a polygon is a unit-wide zone around that polygon in the Euclidean space. But it would contain all polygons adjacent to that polygon in the topological space. Map overlay is "the process of comparing spatial features in two or more map layers" (Maguire and Dangermond 1991 pp. 329). The output of such a process is a single map. Map overlay is a key GIS function that enables the integration of different map layers or different data sets for the same geographic region. Vector-based map overlay relies on the geometry and topology of spatial features. Three types of vector overlay are point-in-polygon, line-in-polygon, and polygon-in-polygon. Taking measurements about spatial features is an important spatial operation. Common measurements include the straight or curved distance between two or a series of points, the length of line or its segment falling inside an polygon, the area of the intersection of two polygons, the perimeter and area of a polygon and so forth. Data query is a process of searching a single or multiple data sets for data that meet specific conditions expressed in terms of spatial features and/or the attributes. It may call for other spatial manipulation operations, including the proximate analysis and the map overlay. For example, one can make a query to find all sampling stations falling in the 50-mile radius of a specific point or in the northern region recording a temperature of below 15° C.

Goodchild (1987, pp. 332) classifies spatial analysis into six basic classes according to the relationships of spatial features involved in operations:

1. Operations requiring access only to the attributes of one class of spatial features (i.e. normal non-spatial analysis);
2. Operations requiring access to both attributes and location information for a single class of spatial features (e.g. calculating location of mean centre);
3. Operations which create spatial feature-pairs from one or more classes of spatial features (e.g. nearest-neighbour based point pattern analysis (see Cressie 1993));
4. Operations which analyse attributes of spatial feature-pairs (e.g. the Moran's I test of spatial autocorrelation (Cliff and Ord 1981) and spatial regression modelling (see Haining 1993));
5. Operations requiring access to attributes and location information for more than one class of spatial features or spatial feature-pairs (e.g. spatial interaction modelling);
6. Operations involving the creation of a new class of objects from an existing class; e.g. generating polygons from points or buffer polygons around points or line segments.

Haining (1994) discusses and illustrates the first four types of operations in the context of spatial statistical analysis. The author identified another class of spatial operations that requires modifying the relationships between pairs of spatial features.

The GIS has been widely recognised as an important tool for handling spatial data. As the Chorley Report (Department of Environment (DoE) 1987:8) stated, "...the Geographic Information System... is as significant to spatial analysis as the invention of the microscope and telescope were to science, the computer to economics, and the printing press to information dissemination". However, it has been recognised that this potential has not yet been realised owing partially to the weakness of GIS in spatial analysis. Openshaw (1987, p431) commented on GIS and said that "such a system is basically concerned with describing the Earth' surface rather than analysing it. Or if you prefer, traditional 19th-century geography reinvented and clothed in 20th-century digital technology". In reality, GIS today might be described as: "A database containing a discrete representation of geographical reality in the form of static, two-dimensional geometrical objects and associated attributes, with a functionality largely limited to primitive geometrical operations to create new objects or to compute relationships between objects, and to simple query and summary descriptions"(Goodchild *et al* 1992, pp. 408).

What may be responsible for the current situation is that the demands for spatial data analysis have not been strong for past decades. The GIS marketplace has been primarily dominated by applications in areas of resource management, infrastructure and facilities management, and land use management where GIS tends to be used for simple record-keeping and query (Goodchild *et al* 1992). The need to undertake spatial data analysis arose mainly from academia rather than industry.

Another reason for this situation is that spatial analysis remains a comparatively obscure field, and many spatial techniques in the field have been developed parallel to the GIS techniques. Taking SSA for example, there is only a handful of books about SSA, including Unwin (1981), Upton and Fingleton (1985), Cressie (1991) Haining (1990), and Bailey and Gatrell (1995). Although many sophisticated and complicated spatial statistical techniques have been demonstrated to be useful, they have not been fully appreciated in the GIS community. Implementing them in GIS often requires GIS to provide special support. For example, with reference to Cressie's lattice data (see Chapter 1), a vector-based topological model may be sufficient for representing the partitions of a region of interest, but insufficient for such important properties of SSA as inter-area relationships (Anselin 1988, Haining 1994).

These problems have been widely acknowledged - the importance of identification of appropriate spatial analysis technologies and their link to GIS was mentioned (DoE 1987), and subsequently appears as a key issue in the agenda of both NCGIA and Economic and Social Research Council/Natural and Environmental Research Council (ESRC/NERC) initiative on GIS (NCGIA 1989 and Masser 1988). Many authors have expressed similar views and called for a greater range of spatial analytical tools to be linked to GIS (Goodchild 1987, Openshaw 1990, Haining and Wise 1991).

2.3.2. Benefits of integrating GIS and SSA

It has been argued and demonstrated by many researchers that the integration of both classes of techniques could benefit each other greatly (Bailey 1994, Levine 1996, Anselin and Bao 1996, Haining *et al* 1996). GIS offers a useful platform for developing SSA techniques cost-effectively. Its data models provide bases for organising spatial data, while its operations associated with the data models make it possible to support a wide range of data manipulations required by SSA. For example, a vector-based GIS is appropriate for area-based data analysis. Its data query and manipulation operations could assist the construction of appropriate area frameworks and the specifications of inter-area relationships. Its cartographic operations are particularly valuable aids for SSA since they support the human eye's extraordinary power to digest two-dimensional information. Many SSA techniques can easily be made available in a GIS environment. Also GIS encourages the development of SSA techniques by taking advantage of its facilities. The best use of GIS facilities would help reduce technical complexity commonly observed in many SSA techniques and help their users understand the techniques and the results they produce (Goodchild et al 1992).

GIS could also benefit from linking with SSA techniques. SSA techniques offer an opportunity for GIS to realise its potential as a general tool for handling spatial data to meet potentially huge demands for GIS packages to have strong capabilities in spatial data analysis given the number of organisations who have datasets in GIS databases. In addition, SSA techniques would surely make their contribution to the evolution of GIS by raising new requirements and proposing new ways to perform analyses in GIS environment. For example, lattice data analysis requires GIS to facilitate the specification of the inter-area relationships and manage them effectively (Cliff and Ord 1975, Anselin 1988, Haining 1993). Dynamic ESDA techniques, such as "linked windows", may also change the way in which analyses are performed in current GIS (Haslett *et al* 1990, Haslett 1992, Goodchild *et al* 1992, Bailey 1994, Haining *et al* 1996).

2.3.3. Development of SSA techniques for GIS

Goodchild *et al* (1992) have suggested that there are four areas where incorporating statistical tools within GIS might strengthen current GIS practice: 1) data rectification; 2) data assessment; 3) data sampling; and 4) data exploration and confirmatory analysis.

2.3.3.1. Data rectification

A common problem in GIS occurs when spatial data is collected from different and incompatible regional frameworks. Studies that require censuses taken from two or more census years would be likely to face this problem because census tracts are often not the same across years. Goodchild and Lam (1980) reviewed methods dealing with this problem of incompatible areas. Flowerdew and Green (1991) report a modern statistical method based on the E-M algorithm for assigning variables to intersected areas. This method has already been implemented using a statistical package GLIM linked to a GIS.

It is not unusual that some variables do not have values at some required spatial locations. For example, the level of air pollution is often measured at specific sampling locations. If a study is intended to investigate the relationship between the pollution level and the respiratory disease incidence at the level of areas, the analyst has to estimate the pollution level for each area with measurements at the locations (Collins 1998). The methods for doing this can range from fairly simple ones, depending largely on the geometric relationships between sites, to much more complex ones, such as regression and kriging analysis which depend on carefully modelling the mean and covariance properties of the data.

2.3.3.2. Data assessment

The issue of the inaccuracy in spatial databases has been paid much attention recently. It has been suggested that statistical techniques may be of use to the GIS community, including those for characterising sampling errors, those for detecting outliers in databases and those for estimating the effects of error propagation (Goodchild and Gopal 1989, Heuvelink and Burrough 1989). These techniques are particularly useful for small-area health studies of rare diseases because any data errors are likely to have strong effects on the analysis results.

2.3.3.3. Data sampling

Since GIS databases are often very large, methods for data sampling are required to extract information from the databases. Arbia (1991) described a sequential GIS-based procedure DUST (Dependent Areal Units Sequential Technique) that can be used to identify areas to be included in a sample in order to maximise the information content of the samples. Data sampling techniques would be very useful for health research because there is a large amount of detailed health-related data waiting to be analysed (Elliott *et al* 1992).

2.3.3.4. Exploratory and confirmatory analysis

There is a widespread agreement that general data exploratory methods would be of great value within GIS, especially in those situations where data is of poor quality and there is a lack of genuine prior hypotheses (Goodchild *et al* 1992). However, the opinion on what methods are appropriate to GIS is divided. Openshaw advocates the development of new, generic spatial analysis methods, customised for a data-rich GIS environment, automatically looking for potentially interesting patterns within large data sets (Openshaw *et al* 1987). To others, there is much to be learned from using traditional exploratory analysis methods for identifying properties, perhaps as part of a process of data modelling (Upton and Fingleton 1985, Haining 1990 and Cressie 1991). Since the traditional exploratory analysis techniques have been developed for a long time and are widely applicable in many areas, they are ready to be linked with GIS. Nevertheless, novel techniques, like automated pattern spotters, would complement traditional methods (Bailey 1994). Goodchild *et al* (1992) identify a list of ESDA methods that are thought appropriate to be integrated with GIS. Among them, the most favoured are exploratory data analysis methods that utilise the dynamic links between statistical graphics and maps.

Methods for confirmatory data analysis are also important for GIS applications. There is a large body of theoretical work on statistical modelling. A number of CSDA techniques for analysing point patterns have been linked into GIS. Examples include Diggle's model (1990) for the analysis of the pattern of points (e.g. disease incidences) in relation to a point source (e.g. potential hazard), and those based on the K-function (Diggle *et al* 1990, Gatrell *et al* 1994). Ordinary and generalised regression models, including those catered specifically for spatially referenced data are well-established (Anselin 1988, Haining 1993). Anselin (1990) reported the development of a spatial statistical package, SpaceStat, enabling the analyst to fit a set of regression models and perform a range of model diagnostics.

For the analysis of area-based data, regionalisation methods are of particular importance for building areal frameworks satisfying criteria for a specific study and for assessing the sensitivity of results to different but equal "plausible" areal frameworks. Openshaw and Rao (1994) introduced a set of region-building methods implemented in a zone design system, ZDES, which allows the user to define criteria in a number of ways. Wise *et al* (1997) discuss another method which could take into account three types of criteria for constructing areal frameworks. This method will be reviewed in some detail in Chapter 4.

2.4. The integration of SSA techniques with GIS

This section considers some issues concerning system integration, discusses some approaches to linking SSA techniques with GIS and summarises some previous work on integrating SSA techniques and GIS.

2.4.1. System integration issues

System integration is concerned with the integration of one or more GIS and SSA software packages or components, to form a co-operative system to provide additional SSA functions for applications. Such a system must be able to perform the required spatial analysis, must have an expected system performance and a friendly user interface, and require the least amount of efforts to integrate.

The system integration faces a number of issues and they have been discussed by many authors, including Nyerges (1992), Chou and Ding (1992), Goodchild et al (1992), Abel and Kilby (1994), Abel et al (1992), and Raper and Bundock (1993). The following summary is based on Abel and Kilby (1994) where the authors discuss the issues through comparing differences between components at their external, conceptual and internal schemas. An external schema describes services provided by a component to another component. Typical services may include such as a scripting language, protocols or application interfaces (API) and data structures for data transfer. A conceptual schema defines the conceptual structures of the data objects stored, related to each other and manipulated, or in short models for physical systems of concern. An internal schema describes the implementation of the corresponding conceptual schema within a particular computing environment.

Differences at component external schemas may exist because each component may support its unique scripting language, protocols or API, data structures for the data transfer. In order to make each component to be able to communicate with other components purposed-built transfer services must be provided to bridge them. Differences at component conceptual schemas are likely to exist among components developed in different disciplines. This is because different disciplines often employ their unique conceptual schemas, each specifies a certain structure and behaviour of a physical system as well as its associated schematics together with simplifications and abstractions in formalising that structure and behaviour. The differences may have their profound implications on system integration because they are likely to determine how the components should be integrated in order for them to work in a logically correct manner.

Differences at component internal schemas often exhibit as different components may have their unique specifications for implementation environments including programming languages and software and hardware facilities. The implementation environments are likely to influence the structure and behaviour of data objects specified by the conceptual schemas because different languages have different powers for the developers to implement certain data structures. For two components, the first one is developed in C and the second in FORTRAN 77. Then, the latter almost certainly would employ simpler data structures than the former would because FORTRAN 77 could not express complex data structures that C can. The implementation environments may influence the flexibility for the developer to carry out system integration because different environments may have different degree of openness and support for interoperability in hardware and software.

The discussion above indicates the importance of the differences between components at the three schemas for system integration. Upon the identification of differences, approaches to constructing purpose-built transfer services, then, become important for system integration. The following section will consider integration approaches that may be taken for system integration.

2.4.2. Integration approaches

Goodchild *et al* (1992) classified integration approaches as: stand-alone SSA software; loose coupling of, or close coupling of GIS components with SSA components; and full integration of spatial analysis tools within GIS. The first approach is an extreme case because there is no GIS component involved in system integration. Many early practices were implemented in this way and made their contribution to GIS evolution. An approach is regarded as loose-coupling if GIS components and SSA components exchange data using ASCII or binary files. Data exchanges are usually carried out manually. A close coupling approach implies that the GIS operations are modified and extended through those routines which might be implemented either inside or outside the GIS. The data exchanges between components are implicit rather than explicit to the user. The last approach is again an extreme case where SSA techniques are implemented entirely inside a GIS environment.

Nyerges (1992) categorised the approaches in a similar manner as isolated, loose-coupled, tight-coupled and integrated. With an isolated approach, different components do not share the same data model and data need to be converted manually between them and transferred off-line. The components may run on different computing environments, while the user interfaces are separate. With a loose-coupled approach, differences in data models need to be solved manually. This process results in a set of mappings serving as on-line cross-referenced indices for the subsequent operations. The data is transferred automatically. The user interfaces are still separate. With a tight-coupled approach, data models may or may not be the same, but the data is shared and accessed through a set of application interfaces (APIs). The user interfaces may or may not be separate. With an integrated approach, the same data model is shared by all components.

Chou and Ding (1992) proposed another classification scheme from three perspectives: 1) data sharing; 2) structures of implementations; and 3) the user interfaces. In terms of data sharing, the approaches are divided into two categories: direct data sharing and data transferring. With a direct data sharing approach, spatial data is directly accessed by the components without any transferring. With a data transferring approach, the spatial data is transferred between the components through intermediate files. From the structures of the implementation perspective, the approaches are divided in terms of internal and external modelling - a component is implemented in the environment of another component or not. By considering the user interfaces, the approaches are classified depending on whether the user interfaces are separated among the components, called shifting user interfaces, or are integrated as a single user interface. The recognition of this last aspect is useful and necessary because a package may have its own user interface not compatible with those of other components. Although the user interface of a component may be re-implemented in another component or a single user interface system can be constructed for all components, it is often not economical to do this. Therefore, the integrated system might have more than one user interface systems and system integration must consider how to ease, if not eliminate, the possible confusions arising from the multiple user interfaces, and control the user interactions with the system.

Although the three schemes outlined above are useful for classifying the existing integrated systems, they all fail to provide a framework to make a detailed study into different aspects that ought to be considered for a specific system integration. Abel and Kirly (1994) describe a model that appears to overcome this problem. The model is based on Sheth and Larson's (1990) inter-operable or federated database model dealing with coupling multiple databases maintained by autonomous heterogeneous database management systems, but has been modified to place emphasis on modelling the data exchange between different components. In this model, the data exchange activities are expressed using four types of linking operations - transformation, constructor, accessory and filtering operations. Transformation operations map data under one component's external schema to equivalents under another external schema. A transformation operation may convert a set of vector data for a component to raster data for another component. Constructor operations formulate a batch of commands, and then map the commands into a series of operations onto corresponding components. Accessory operations dispatch a sequence of operations defined by commands to corresponding components and execute them accordingly. They may also retrieve the data from different components and assemble them. Filtering operations validate the commands and the data, and may also translate the commands from one format to another. These operations are named as T, C, A and F operations for short respectively.

Abel and Kirly's model (1994) can be used to classify different integration approaches by considering the combination of these four types of linking operations and the configuration of components. The authors discussed several typical configurations of components and some of these will be examined in the next section. This model is important because it provides a framework for the designers to identify appropriate configurations for the components and to make detailed specifications on linking operations between these components.

Section 2.4.1 and 2.4.2 consider some system integration issues that are important to linking SSA with GIS. Chapter 5 will show how these issues are addressed in the implementation of SAGE.

2.4.3. Previous practices in the system integration

In this section, a review of some previous practices in system integration is given in line with stand-alone, coupling and fully-integrated approaches. For each integrated system, its corresponding configuration in Abel and Kirly's (1994) model will be discussed briefly.

2.4.3.1. Stand-alone approach

There are many working systems of the stand-alone approach. One of the earliest examples is INFOMAP which offers the features of a thematic mapping package and enables the user to perform a range of statistical analyses on different types of spatial data (Bailey 1990, Bailey and Gatrell 1995).

Another best-known example is REGARD, formerly called SPIDER by Haslett and co-workers (Haslett *et al* 1990). The package has been developed on a Macintosh system and takes advantage of the excellent graphic capabilities built into the system. REGARD is capable of producing multiple linked views of the same spatial data which may contain a map view and graphic views such as a histogram and a scatter plot. The graphic views are associated with the map view so that highlighting areas on the map will lead to highlighting the corresponding parts in the histogram and points in the scatter plot. MANET is another package developed for Macintosh by Unwin and his colleagues (Unwin 1996, Unwin *et al* 1996b). Besides inheriting many features from REGARD, it extends its capabilities further to enable analysis of categorical data and superficial assignment of physical conditions, such as rivers and highways. Dykes (1996) described an exploratory package, cdv (Cartographic Data Visualiser), which provides functions similar to REGARD but with enhanced choropleth mapping. cdv can run on all platforms with Tcl/Tk (Ousterhout 1994). Anselin (1990) has developed a system, SpaceStat. This package provides a set of comprehensive facilities mainly for the confirmatory statistical analysis of spatial data with the ability to fit different types of spatial regression model and to discriminate between different types of spatial models.

Openshaw and colleagues (Openshaw *et al* 1987) provide another example, based on a different approach to the problem of spatial data analysis. They argued that GIS users with large amounts of data, often of poor quality, will want to 'explore data' without knowing where to look or what to look for or when to look. They propose a suite of automatic 'pattern spotters' which can browse through GIS databases looking for interesting patterns and relationships automatically. The earliest of these was the Geographical Analysis Machine (GAM) which has been used to detect clusters of leukaemia cases.

It is obvious that if a stand-alone package needs to handle all spatial data by itself the focus of its implementation would inevitably be shifted to developing functions that have been already well supported in GIS. However, there are at least two reasons for writing stand-alone packages. First, the SSA techniques themselves may not really need the GIS facilities. This is probably true when the SSA techniques represent a totally different approach to the analysis of spatial data. GAM, for example, uses computer power to repeat automatically the same calculation many times for different positions in space. In contrast, most current GIS packages are optimised for interactive operations through querying and displaying facilities. In addition, owing to the need for rapid and repeated access to the geographical database in order to perform the statistical calculations many times, GAM requires a unique way to achieve very fast access to the spatial data. In GAM a unique data structure, the KDB tree, was designed for this purpose. As another example, REGARD represents a new way to explore spatial data where dynamic rather than static links of multiple views are essential. In order to achieve highly dynamic links between maps and statistical plots, data must be handled in some unique ways. Although many GIS mapping and graphing functions are written with flexibility in mind, they are not designed or optimised for such linking operations.

Second, the GIS facilities may not add much to the development of techniques. SpaceStat focuses on parameter estimation and hypothesis testing on attributes for each spatial unit and the relationships between spatial units (Anselin 1988). Although GIS would be very useful for handling attribute and feature data, it does not offer any explicit support for managing connectivity matrices and their computation. Therefore, the GIS functions are unlikely to have any significant impact on such a system unless there is a need to handle complex attributes and features from multiple data sets and to visualise the data and results.

2.4.3.2. Coupling of GIS with SSA techniques

One of the earliest examples of the close or tight coupling approach is SAM (Spatial Autocorrelation Machine) by Ding and Fotheringham (1992). SAM consists of two components, ARC/INFO, a GIS package from ESRI (ESRI 1995), and a set of functions written in the C programming language for calculating spatial autocorrelation statistics for the spatial data held in ARC/INFO. These functions are called by ARC/INFO directly using the AML (ARC/INFO Macro Language) (ESRI 1995) and their results are mapped using ARC/INO. Batty and Xie (1994) described another system, similar to SAM but more complex, for modelling the density of urban population. Again a set of functions were written in FORTRAN and C to perform time-intensive computation while other functions in the AML are written to handle the data display and linked-windows. Besides these research systems, there are also some commercial close-coupled systems. An example of such a system is the link of S-Plus and ARC/INFO by StatSci and ESRI where an object-oriented statistical language, S-Plus, provides statistical analysis capabilities inside ARC/INFO (MathSoft).

The systems discussed above are two component systems sharing the same features. The GIS component functions as a master whereas the spatial analysis component is a slave of the master. The GIS provides C, A and T operations for constructing a sequence of commands according to a specific analysis task, executing the commands sequentially, and transforming spatial data in correct forms forwards and backwards between the two components. With respect of Abel and Kirly definition (Abel and Kirly 1994), these systems follow an "embedded" configuration.

The most typical example of loose-coupled systems is linked systems through Geolink (Waugh 1986). The centre of Geolink is a programmable 'editor' to define data-converting programs to link different components into a pipeline. It also provides facilities to produce complex end-use systems. With reference to the Abel and Kirly's model, Geolink functions as a master and provides C, A and T operations where other components are the slaves of it. Such a system configuration is regarded as a common user-interface configuration, a direct extension of the embedded configuration. Geolink has been used to link ARC/INFO with other standard statistical packages.

Cook *et al* (1996) discussed a system which might also be regarded as a loose-coupled system. In this system, ArcView (ESRI 1995) and Xgobi, a visualisation package, are linked for ESDA. The former is used to display spatial reference maps and concomitant variables, while the latter is used to explore multiple variables. They are linked by RPCs (remote procedure calls, SunSoft 1994) and the different views of the spatial data are "hot" linked. Each component has its own user-interface and functions as either a master or a slave at any one time. Each component provides a set of C, A and T operations. Since the RPCs are implemented as a middleware to be shared by the components, this system may be regarded as a common-shared component configuration in Abel and Kilby's (1994) model. Anselin and Bao (1996) report development on linking SpaceStat with ArcView in a similar fashion although the linkage between the two packages is fulfilled by directly manipulating ArcView data files. This has largely enhanced the capabilities of SpaceStat in exploratory data analysis although statistical plots and maps are not 'hot' linked.

The choice of an appropriate integration approach or configuration for linking the components is a major step in system integration if the integrated system is to achieve the expected system performance, to have a friendly user interface system, and to require the least efforts to implement. In terms of coupling, the close coupling approach is suitable when components share the same or very similar data models. The use of this approach could result in a seamlessly integrated system. However, in practice, this approach is often limited by what GIS and SSA offer, and a 'real' seamless system is often hard to achieve.

Unlike the close coupling approach, the loose coupling approach represents a more flexible way to integration. GIS and SSA components do not necessarily share the same or even similar data models, and the differences are resolved through purpose-built transforming facilities. Typically, the user of such an integrated system may need to shift from one component to another, probably through different user interfaces in order to fulfil certain operations.

In terms of the configuration of components, the embedded configuration is simple and easy to implement, and probably results in a user-friendly system. Unfortunately, few packages provide sufficient functions to allow components to be linked in such a simple manner. Therefore, the shared-common component configuration and the common user-interface configuration may be more suitable.

2.4.3.3. Full integration

SSA techniques are implemented using functions provided by both GIS and the other components in such a way that they become genuine functions of the GIS. The full integration approach is an ideal way to make SSA techniques available to the GIS community. However, since most GIS packages are proprietary, the full integration approach may be an option for GIS developers only rather than analysts (Goodchild *et al* 1992). This situation is changing and more and more GIS packages have begun to provide facilities to allow users or developers to make extensions more easily. Packages, such as ArcView and MAPINFO, provide powerful programming languages to allow users to implement new functions that could access spatial data internally. Moreover, as software development is driven towards componentisation conforming to standard application interfaces (McKee and Kottman 1999), the full integration of different components will become less of a problem and the distinction between close, loose and full integration will become less clear.

As discussed above, the stand-alone approach is limited in the sense that an integrated system normally lacks many of the required GIS functions. On the other hand, the full integration approach is most likely to be employed by GIS developers rather than analysts. As a consequence, the coupling approach remains as the most plausible way for the integration of GIS with SSA techniques. In practice, the choice of an approach to or a configuration for the system integration depends critically on the facilities of the chosen GIS, SSA packages, other accessory packages and computing environments as well as the intended application areas where constraints on the integrated system are either explicitly or implicitly imposed. The data models and functions supported by GIS and SSA components are usually a decisive factor in choosing an integration approach. It is by no means trivial to make the right judgements on all of these factors and to bring out a technically feasible system design and implementation plan.

2.5. Summary

Section 2.1 considered a statistical methodology for the analysis of areal data underpinned by ESDA and CSDA. It emphasised the role of ESDA not only in identifying spatial properties but also in assessing CSDA for the purposes of model evaluation and diagnostics. Section 2.2 summarised some methodological and interpretative problems, stemming from the nature of such data. This section also highlighted the consequences of failing to recognise these problems.

Section 2.3 considered the strengths and the weaknesses of GIS in facilitating the analysis of spatially referenced data. It argued that GIS is an appropriate vehicle for integrating many useful SSA techniques. Section 2.4 summarised some issues concerning system integration and common approaches to integrating SSA and GIS. It reviewed some recent work in this area and examined the advantages and disadvantages of different approaches.

The main theme of this chapter has been the discussion of fundamental issues concerning the analysis of area-based data and the integration of SSA techniques and GIS. These issues will be elaborated in the following chapters in the context of the geographical analysis of health-related data. The next chapter will consider some common questions asked in health research and the roles that SSA and GIS might play in answering them.

CHAPTER 3. SPATIAL ANALYSIS AND GIS IN HEALTH RESEARCH

The previous chapter discussed statistical methodologies for analysing area-based data and some of the important problems that arise in spatial data analysis. It also considered the roles of GIS for SSA and the benefits that may be gained when two technologies are integrated. This chapter will consider the roles of SSA and GIS in health research and the types of SSA techniques that ought to be integrated with GIS in order to meet the needs of different users to carry out a wide range of health studies.

This chapter is divided into four sections. Section 3.1 considers some major types of studies in health research and questions with which these studies are concerned. It shows how SSA may help answer these questions and highlights some common problems likely to arise in answering these questions. Section 3.2 summarises the sources of health-related data commonly available to health research and two relevant issues: data confidentiality and accuracy and their implications for the analysis of health-related data. Section 3.3 examines the roles of GIS in health research. It argues that a GIS for health research needs to provide SSA tools capable of answering each of these questions and, for each question, there should be a range of SSA techniques varying in statistical sophistication to meet the needs of diverse users. A summary of this chapter is given at the end.

It should be noted that the main theme of this chapter is to show the roles which SSA and GIS may play in health research without dealing with all related health issues in depth. Health examples to be given in this chapter are therefore mainly for making the points. Also this chapter will lean toward an emphasis of the role of SSA and by no means try to give a comprehensive coverage of GIS applications in health research but rather focus narrowly on useful GIS functionality for supporting SSA.

3.1. Types of health study

As introduced in Section 1.2, health research falls into two areas: spatial epidemiology and health services research. This section will consider five main types of studies in spatial epidemiology, and two main types of studies in health services research.

3.1.1. Spatial epidemiology

Spatial epidemiological studies have been carried out on various geographical scales ranging from international, national, regional to sub-regional scales. Studies on coarse scales, in particular at the international level, have been successful. Some studies take advantage of the large differences in such attributes as genetic makeup and population lifestyles in different countries to identify factors that may give some clues as to disease causation. Stomach cancer, for example, has been found to vary dramatically around the world. Japan and Iceland are among the highest in mortality rates several times higher than that in other countries (Meade *et al* 1995). This has led to intensified search for the common factors shared by these populations. A number of hypotheses on the causal effects of factors, including diet, have been suggested (Meade *et al* 1995).

Given the very scale on which geographical surveys take place, a spatial epidemiological study at a coarse scale is unlikely to provide useful insights into phenomena that vary only on a sub-regional scale. Child leukaemia in young people in Western Europe, for example, shows no obvious spatial variation on this scale (English 1992). However, studies on the county and regional scales in the UK show some evidence of the localised spatial variation of this cancer associated with some localised socio-economic and environmental factors (Alexander 1991, Openshaw and Craft 1991, Draper *et al* 1991). Owing to the increasing awareness in public health and the availability of high-resolution health-related data for rare diseases and the population at risk in the economically-advanced counties, fine scale spatial epidemiological studies are often called on to address public concerns on possible localised health problems (Cuzick and Elliott 1992, Kulldorff 1998). It is the studies of rare diseases on the fine scale that is of particular interest in this section.

Of the many types of epidemiological studies (Elliott *et al* 1992), five common ones are:

1. Spatial descriptive study of disease events;
2. Investigation of the clustering of disease events as a general phenomenon;
3. Investigation of the localised clusters of disease events without reference to any specific putative sources;
4. Investigation of the clustering of disease events in the vicinity of putative sources; and
5. Investigation of the ecological association between disease events and risk factors.

3.1.1.1. Spatial descriptive studies

Spatial epidemiological studies require a detailed description of where disease events arise and what factors may relate to them in order to understand why diseases arise where they do. The spatial descriptive study is intended to address this by describing the geographical variation of disease events and factors of concern. This is often done with the assistance of statistical mapping techniques where the count of the disease events and the measurement of the factors may be mapped independently or together by overlaying. This allows the analyst to visualise the spatial distribution of the disease events and identify possible relationships between them and the factors (Muir 1987, Boyle et al 1989, Bailey and Gatrell 1995).

For the purpose of describing spatial variation of disease events across a set of areas, choropleth maps of relative risks for each area may be created. A relative risk for an area may be estimated by a standardised rate ratio - the observed number of disease events over the expected number of disease events for that area times 100. With the indirect standardisation, the expected number is estimated with respect to the population in a larger reference region adjusted for confounding variables such as sex and age (Meade *et al* 1985, Williams *et al*

1998). A ratio greater than 100 indicates a possible elevated risk in the corresponding area. For a non-contagious disease such as cancer, its occurrence in each area, assuming the absence of any variation across areas in particular disease determinants, would be expected to follow a Poisson distribution closely (Olsen *et al* 1996). So the extent of the observed number departing from the expected number in that area could be assessed with the probability, P value, under the null hypothesis that observed cases arise at the same rate in the area and the reference region. A P value of less than 0.05 is, conventionally, regarded to be statistically significant, leading to the rejection of the null hypothesis. A probability map of P values may be drawn along with the corresponding ratio map (Choynowski 1959, Muir 1989).

Caution should be taken in interpreting ratio and probability maps when the small number problem and heteroscedasticity are present. Less populated areas might stand out owing to a single disease case in a ratio map even if the "real" risks for areas are the same, whereas only areas with the largest populations would tend to stand out in a probability map (Clayton and Kaldor 1987, Kennedy 1989, Besag and Nevell 1991, Olsen *et al* 1996). One way to overcome these problems is to obtain 'smoothed' estimates of relative risks. Empirical Bayes procedures are widely used to obtain smoothed estimates by shrinking rates of less populated areas toward the mean rates of areas adjacent to each of them or with respect to all areas (Clayton and Kaldor 1987). Chapter 4 will review some of these techniques.

There are some technical issues concerning disease mapping. It is important to choose the number of intervals or categories for values to be mapped. The number of categories between 4 to 8 is recommended since the human eye cannot readily distinguish more than eight colour or greyscales (Olsen *et al* 1996). Choosing intervals is also important since different intervals may give rise to quite different visual impressions (Bailey and Gatrell 1995, Cromley 1996, Heywood *et al* 1998). No matter what data is to be displayed or how areas are coloured, physically large areas always attract the map readers more attention than small areas although the small areas may be more important in a study. For example, large census tracts are often less populated and least important in the mapping properties of the population.

3.1.1.2. Investigation of disease clustering

Studies of this type are concerned with the patterning of disease events over a region as a general phenomenon (Cuzick and Elliott 1992). As with the descriptive studies discussed above, even complicated patterns of non-random distributions may be made evident by visually assessing the distribution of disease events. However, they should be confirmed by statistical testing. Two null hypotheses often tested against are that the disease events arise from a Poisson distribution and that there is no general tendency of spatial dependence between those events occurring at one place and those at places nearby (Alexander and Cuzick 1992, Bailey and Gatrell 1995, Waldhor 1996). A statistical test under the first null hypothesis is often called a test for heterogeneity, $\chi 2$ test, or test for extra-Poisson variation. Knowing that there is an extra-Poisson variation in disease events may help formulate hypotheses. Evidence shows that there is extra-Poisson variation in childhood leukaemia incidence (Alexander 1991, Openshaw and Craft 1991). This has led to a hypothesis that a virus may be responsible for the disease although different interpretations are possible (Cuzick and Elliott 1992).

Alexander (1991) reports a study of an extra-Poisson variation of childhood leukaemia and non-Hodgkin lymphomas in England, Wales and Scotland where several statistical methods were employed, including the Cuzick-Edwards tests (Cuzick and Edwards 1990) and the Potthoff-Whittinghill test (Potthoff and Whittinghill 1966). With the whole data set for England, Wales and Scotland during 1966-83, the Potthoff-Whittinghill test showed some evidence of clustering of this disease, particular in age groups 0 to 4 and 0 to 14. Nevertheless, no tests provided evidence of clustering when a restricted data subset of the data set was used. Openshaw and Craft (1991) report a study of the same data where they applied four tests implemented under the GAM/2 framework. Using the whole data set, they found that there is some evidence of weak clustering of childhood leukaemia. Both studies demonstrate ways of testing for extra-Poisson variation and the difficulties in doing this for rare diseases (Draper 1991).

A test for extra-Poisson distribution, however, often tell little about the underlying spatial distribution of a disease, only that disease events arise at higher or lower rates than would be expected from a Poisson distribution (Olsen *et al* 1996). A test for clustering against the second null hypothesis may provide insights into the mechanisms of disease occurrence, which may relate to an infectious process, environmental ‘hot spots’, spatially-patterned underlying causal variables or various other factors (Oden 1995). Glick (1982) uses the spatial correlogram - a plot of spatial autocorrelation measures against spatial lags (see Bailey and Gatrell 1995) - to investigate the spatial organisation of cancer rates in USA. He shows that stomach cancer in counties across Pennsylvania has a steep correlogram for a few close lags, whereas bladder cancer has an almost horizontal correlogram for the first three lags. The former suggests isolated incidence possibly resulting from contagious disease occurring in large cities but not yet diffused across the counties, and the latter suggests an environmental carcinogen found over large areas, such as water catchment basins, rather than a diffusion process (Meade *et al* 1988). Two statistical tests useful for detecting spatial dependence of data over areas are Moran’s I test and the Getis-Ord test for spatial autocorrelation and concentration respectively. Cliff and Ord (1981), and Getis and Ord (1992) demonstrate the use of these statistics to test spatial dependence in the context of disease studies. These tests will be discussed in the next chapter.

3.1.1.3. Investigation of localised disease clusters

Studies of this type are often concerned with investigating alleged clusters of disease events and scanning data of disease events for clusters. Investigations of suspected or alleged disease clusters are often started in response to public concerns and may, consequently, help reduce the anxiety that often accompanies such allegations. However, such investigations often suffer from two problems that complicate the interpretation of statistical tests. One problem (pre-selection bias) is that the clusters of a disease reported somewhere might have well been selected from a set of diseases before a hypothesis is formed. Another relates to the definition of the population and the period for a study. The suspicion of a cluster often begins with the identification of a group of cases and only subsequently is the underlying population defined (Olsen *et al* 1996).

Scanning data of diseases for clusters is often done as part of a disease-monitoring programme. It is also important for the purpose of carrying out a more focused epidemiological study where researchers have to find appropriate sites for the study. Sites around clusters of disease events may be of most interest if there is no prior expectation of where the incidence of the disease may be much higher since the chance of finding clues to the disease's causation may be higher in the vicinity of a cluster than in other places (Openshaw 1990). Some diseases might not even show any sign of global clustering on a given scale if geographically defined populations react to different localised factors.

Probably the best known data scanning technique is GAM. Basically, GAM works by defining a fixed-radius circle on each point of a square grid. For each circle, it counts observed cases and computes the expected number of cases. Then it tests whether the observed number of cases departs significantly from the expected number based on the Poisson distribution. If the test is significant, it draws the circle. This procedure is repeated on circles with different radii. Consequently, if there are many circles with different radii overlaying at the same locations, those locations may indicate clusters of the disease. A variant of GAM, GAM-K employing Kernel estimation (Silverman 1986) removes the need for the user to interpret 'circle' maps (Openshaw 1990). One of the advantages of GAM and its variants is that the results drawn from them are insensitive to area configuration. The use of these methods for childhood leukaemia data in England and Wales provides evidence of clusters around Seascale and Gateshead as well as some other areas (Openshaw and Craft 1989 and 1991). Besag (1989) and Besag and Nevell (1991) suggest and explore a statistical method testing for significance of the distance to the kth nearest case. This method is considered to be an alternative test for the GAM system (Thomas 1991).

Several statistical tests have been devised to test localised clusters for area-based data, including the local Moran's I, Getis and Ord's Gi and Gi*, and the local Geary test (Getis and Ord 1992, Ord and Getis 1995, Anselin 1995, Getis and Ord 1996). These may be used to test for 'hot spots', to assess the stationarity of the spatial stochastic process (the second order effect) and to examine the scale of clusters. Getis and Ord (1992) use Gi* to investigate sudden infant death syndrome (SIDS) cases in North Carolina counties for the period 1979-1984 and identified a hot spot in the south central part of the state. Ord and Getis (1995) report the use of Gi to trace the spread of AIDS in USA.

Since these techniques involve performing statistical tests at many locations simultaneously, a conventional significance level such as 5% for all tests may no longer be appropriate. Suppose one chooses a significance level of 5%, an overall significance level, to assess the significance of 20 statistics derived for each of 20 areas at the same time. Even if the null hypothesis were true for every area, one would expect to find one out of 20 tests to be statistically significant at the 5% level. This problem is one form of the multiple selection (or multiple comparison) problem. In this case, a rather conservative individual significance level may be set to α/m based on Bonferroni inequalities where α is the overall significance level and m is the number of tests (Anselin 1995).

3.1.1.4. Investigation of clustering in the vicinity of putative sources

Studies of this type are concerned with the effects of one or more putative sources of pollution on the populations around it by seeking for evidence of any association between the sources and the occurrence of disease events. There have been many types of putative sources of pollution such as nuclear installations, waste incinerators and high voltage power lines which have been subject to this type of study (Cook-Mozaffari *et al* 1989, Diggle *et al* 1990, Elliott *et al* 1992, Gatrell and Dunn 1995).

In a study, the level of pollution from one or more sources may be measured directly or indirectly using surrogates. In small-area studies, distance from a source may be used as a surrogate for the level of pollution since direct measurements on it are often not available. In this case, a null hypothesis to be tested against is that the occurrence of the disease does not tend to increase with increasing proximity to the source.

A major problem in investigating disease clustering around putative sources is how to specify an appropriate area for the study (Bithell and Stone 1989, Olsen *et al* 1996). This problem is similar to the multiple selection problem mentioned in the previous section. Suppose that B, centred on the source P, is the area in which the increased risk is confined, and that A an area chosen for assessing the risk is also centred on P. If A is chosen to be much smaller than B in a study, although the chance to reject the null is high, the study may lack power when the number of cases is small. If A is chosen to be much larger than B, the study may fail to reject the null hypothesis (Bithel 1992, Hills 1992).

Stone (1988) and Bithell and Stone (1989) discuss methods for testing the increased risk around a source where a number of small areas that together cover more than any area of interest in a study can be specified. For each area, a relative risk for each area is estimated. A null hypothesis of homogeneous relative risks across areas is tested against an isotonic alternative corresponding to the order of areas with respect to the source (Bithell and Stone 1989, Hills 1992). Although these methods can deal with the pre-selection problem, results may be affected by the areal configuration. Diggle (1990) devises a method based on a point pattern analysis model. In this model, the local intensity of disease events is expressed by the multiplication of three components: the overall disease events, the background population and a parameterised distance-decay function to a putative source (Diggle *et al* 1990, Gatrell and Dunn 1995). The null hypothesis is then equivalent to finding that the coefficients of that distance-decay function do not depart significantly from zero. The background population is estimated from controls using the kernel estimation (Silverman 1986) within a pre-selected area.

Ideally, these methods above should be used without any prior knowledge about the disease incidence at a locality. When these methods have been used in situations where a suspected cluster of disease has been found and linked to a putative source, the replication of the investigation on other similar sources elsewhere is desired (Cuzick and Elliott 1992).

Diggle *et al* (1990) report a study of the incidence of cancer of the larynx around a closed-down industrial waste incinerator in Charnock Richard, Lancashire, England, using Diggle's method. The study shows some evidence of increased risks of the cancer around the incinerator. Elliott (1992) reports a study of the same problem using Bithell and Stone's method (Bithell and Stone 1989) and found no evidence that supports the previous finding. Gatrell and Dunn (1995) re-examined the same problem using Diggle's method. By estimating the underlying population using different controls and widening the study area, they found that their results are consistent with those of Diggle *et al* (1990).

3.1.1.5. Investigation of the ecological association between disease events and risk factors

When spatial variation in a disease is evidenced, it is of interest to search for evidence of any association at the aggregate level between aspects of the physical and socio-economic environment and the disease under study. If an association is found, the association between the response and explanatory variables needs to be explored by different methods (e.g. individual level studies). A confirmed relationship may give some insights into the disease aetiology and may be used to predict where and how disease cases might arise in future and in turn to direct possible prevention schemes to reduce them (Thomas 1990).

Law and Morris (1998) report a study that tries to explain "why is mortality higher in poorer areas and in more northern areas of England and Wales" by a multivariate Poisson regression analysis. One result from this study, using cause-specific mortality statistics for 403 local authority districts adjusted for socio-economic deprivation classified by latitude and urbanisation, age, sex, and proportion of ethnic minorities, shows that smoking accounts for about 45% of these excess deaths while the colder climate probably for more than 25%. Gatrell and Dunn (1995) report a study of the relationship between the incidence of cancer of the larynx and 44 hospital incinerators in Lancashire and Greater Manchester. They fit a generalised regression model to disease counts on a set of variables derived for a set of 2-km radius circles centred on each of 44 hospitals (cases) with and 44 hospitals (controls) without incinerators. The variables include the expected numbers of incidences, the estimate of social class, and a binary dummy value: 0 for each of the hospitals with incinerators and 1 otherwise. This study also shows how SSA could be used effectively with the assistance of GIS proximate analysis. In this study the result did not show evidence of a relationship between hospital incinerators and cancer of the larynx.

As the examples above and others (e.g. Eames *et al* 1993) show, classical correlation and regression analysis are two important techniques for associative studies in health research. However, they treat the data as if they are drawn independently from space and use locations mainly for linking up (point-to-point or case-to-case) the disease events and measurements on various factors. Two problems may arise. First, as mentioned in Chapter 2, if disease events are spatially correlated, classical inferences on correlation and regression coefficients may not be valid. Second, they cannot be used to examine cross-area but case-to-case relationships. For example, people may be in contract with a disease agent not necessarily at where they live but perhaps at their work place or other place nearby. In this situation, a meaningful association to be considered is to measure how the similar values of two variables are spatially close to one another. Tjøstheim (1978) proposes a test of spatial association of two variables although the inference framework does not allow for spatial structures. Haining (1990) reviews this technique and other regression techniques that allow for the spatial structures of the data in the analysis.

The results of ecological association studies need to be interpreted with caution. One particular reason for this is confounding effects. Age, sex, and socio-economic factors are known to confound the relationship between health outcomes and environmental aspects. Some diseases, like cancer, are more common in the population of one age group than in that of another and may occur frequently in one sex group but not in another. As mentioned previously, socio-economic factors are associated with many diseases. If these factors are not controlled, results from any study cannot be readily interpreted to provide useful insights into possible causal effects (Greenland and Morganstern 1989, Wakeford 1990). Indeed, one important aspect of the ecological study is to identify variables that need to be controlled in different health studies. Two examples given above demonstrate two approaches to controlling for demographic and socio-economic confounding by classification and regression respectively (Dolk *et al* 1995, Bithell *et al* 1995).

It is often the case that the analyst does not be able to know all the confounding variables beforehand and, even if he/she knows all for a specific study, data for them may not be available. In the case of regression analysis, if an important explanatory variable is omitted from the model, the model residuals by the least squares fitting may exhibit a spatial dependence that undermines statistical inference (Haining 1990, Bailey and Gatrell 1995). In this situation, a regression model allowing for spatially correlated error terms may be appropriate. Haining (1993) demonstrates a number of regression analysis techniques that cater for spatially dependent data in a study of the relationship between the mortality of some specific diseases and some socio-economic factors for the city of Glasgow.

3.1.2. Health services research

Health service research comprises many types of studies concerned with every aspect of the health services system. Wilkinson *et al* (1998 pp. 179) list some of them in the context of health research and the following lists some extracts relating to the health services research from that list.

1. Examination of disease rates and other health statistics by geographical areas to assess the health of a population;
2. Examination of spatial variations in health and the use of health services as a comparative approach to needs assessment and resource allocation;
3. Analysis of the spatial distributions of health care facilities and referral patterns to aid decisions on optimal locations of health services;
4. Studies of spatial variations in health treatments and outcomes for planning the development of health services;
5. Studies of health and health promotion interventions at community level.

This section is not intended to discuss all these types of study but to focus on two related areas: 1) assessing the needs of geographically defined populations for health care services; and 2) assessing the equality in access to and use of health care services. The intention here is to show how statistical methods may be used to provide the information needed to improve health services. In so doing, it inevitably excludes some types of studies. For a review of health services research in the spatial analysis and GIS context, see de Lepper *et al* (1995), Gatrell and Löytönen (1998), Wilkinson *et al* (1998), and Gatrell (1999).

3.1.2.1. Geographical health needs assessment

"Health needs assessment is the systematic approach to ensuring that the health service uses its resources to improve the health of the population in the most efficient way"(Wright *et al* 1998). It describes the health problems of a local population, identifies inequality in health and determines priorities for the most effective use of resources using methods including epidemiological methods (Williams and Wright 1998). Geographical health needs assessment focuses specifically on describing the spatial variations of the ill-health of a population, and identifying factors that may account for the variations to provide information possibly leading to a better understanding of inequality in health and therefore the health needs of the population.

The need for a geographical approach to health needs assessment is reflected by the fact that people who live in different places often have quite different health experiences. In the UK, spatial variation in health inequalities in health, in respect to both all health conditions and specific conditions, have been well documented (Black *et al* 1980, Benzeval *et al* 1995, Acheson 1998). Studies show that there is an increasing gradient in the level of ill health from the south to the north in England (Townsend *et al* 1988). Spatial patterning of ill-health is also evidenced on the regional scale (Eames *et al* 1993). A recent study shows that these still remain and the inequality gap in health in the north is widening (Phillimore *et al* 1994). In a comprehensive study of mortality for England and Wales, Law and Morris (1998) found that mortality for many diseases was higher in northern than in southern regions and in urban than in rural areas.

Although mortality and morbidity statistics are vitally important to obtain estimates of the health needs, the use of them alone is often not satisfactory. The use of mortality statistics for doing this has been criticised since dead people no longer pose any burden for health services (Jones and Moon 1987). Although morbidity statistics are preferred to mortality statistics, they are rarely available for diseases other than cancer. Even if the morbidity statistics of all causes were available, they would not be able to reflect aspects of behaviours, social and physical environments, genes and health care that together shape the health needs of a population (Wright *et al* 1998). Evidence suggests that socio-economic inequalities are powerful predictors of inequalities of health. "The persistent, consistent social gradients in health status (however measured) found in virtually every population are strongly and positively correlated with distributions of power,.... Social gradients in health also obtain for most diseases. The gradients cannot be adequately nor even largely explained by individual behaviours, by relative lack of access to effective health care, by exposure to hazards in the physical environment or by the artefact of measurement." (Hayes 1999, pp. 291). This summarises a body of work in past decades, including those taking a geographical perspective, which confirms that people who live in deprived areas tend to have worse health status than their counterparts living in affluent areas (Townsend *et al* 1989, Carstairs and Morris 1989, Eames *et al* 1993, Phillimore *et al* 1994, Benzeval *et al* 1995, Law and Morris 1998). Therefore, it has been a common practice to take the socio-economic gradient in health needs assessment at national, regional and community levels of health administrations (Carstairs and Morris 1989, Judge and May 1994, Benzeval *et al* 1995). Several indexes, often constructed using census variables that measure material deprivation, sometimes combined with other variables concerning health services, have been developed and used for estimating the health needs (Jarman 1983, Townsend 1989, Thunhurst 1985, Carstairs and Morris 1989). For example, the Jarman index (Jerman 1983) is well known and used for allocating primary care resources. For allocating NHS revenue to regional health authorities in the UK, the deprivation scores, although it is not the same for England and Wales, and Scotland, were calculated for each region. These became elements in the resource allocation formula to produce estimates of the needs for hospital services (Carstairs and Morris 1989, Judge and May 1994).

Evidence shows that relative socio-economic inequalities, after absolute material deprivation is controlled, plays a role in shaping inequality in health (Wilkinson 1996, Benzeval *et al* 1995, Gatrell 1997, Hayes 1999, Boyle *et al*

1999). In an empirical study for eighty-nine electoral areas in Morcambe Bay, England, Gatrell (1997) demonstrates that a geographical perspective can help shed light on links between health outcomes and relative deprivation. In that study, the author employed the following variables: aged-standardised mortality indicators for diseases of all causes, ischaemic heart disease, circulatory disease, Neoplasms and lung cancer; the Carstairs's deprivation score (measuring absolute deprivation); the mean deprivation score of adjacent areas (a local spatial average of that score); two categorical variables for areas below and above the mean deprivation score (measuring relative deprivation). Using a regression technique, he found that "For those areas where deprivation is less than the local mean there is no relationship at all between mortality and relative neighbouring deprivation. However, where deprivation in a ward exceeds that in surrounding areas mortality does vary significantly, though not for cancer" (Gatrell 1997, pp. 146). In an empirical study, Boyle *et al* (1999, pp. 791) show that "for small areas (wards) in England and Wales, morbidity is related to deprivation, variation in deprivation within and surrounding each area, and the proportion of the population that are migrants". An implication of these studies to health needs assessment may be that a relatively poorer area than its neighbouring areas may have relatively higher needs for health care than the neighbouring areas. Therefore, spatial analysis techniques for locating 'hot spots' of relatively deprived areas with respect to their neighbouring areas may be of assistance.

3.1.2.2. Geographical assessment of equality in access to health services

Geographical assessment of equality in access to health services aims to provide information on how well health services in a place meet the needs of geographically-defined populations. It raises questions such as "does where people live affect their chances of accessing appropriate health services?", and "where are the poorly performing areas?" and "what are the attributable factors?". Factors of interest may include those related to, for example, economic and demographic circumstances that may influence people's willingness of using health services, and those related to health policy and procedures taken by medical staff.

Many studies of the UK health system show that where people live does affect their chances when accessing health services. Jessop (1988) reports a study of access to some common elective operations among residents within the North East Essex region. Using standardised discharge ratios derived from the hospital discharge data for each elective operation, and the population estimate based on the 1981 Census data adjusted for sex and age, he found a marked variation in access rates across towns in the region. Harwich ranked the highest whereas Halstead ranked lowest for almost all operations. He also found that socio-economic factors could not explain the variation in full and other factors, such as referral behaviours of General Practitioners (GPs), were also important.

Recent studies show evidence of spatial variation in access to treatments of coronary heart disease (Ben-Shlomo and Chaturvedi 1994, Payne and Saul 1997, Ferris *et al* 1998). Payne and Saul (1997) compare coronary artery revascularision rates with the prevalence of angina and coronary mortality in the ward level in Sheffield. They found that some of the most deprived wards had only half the number of revascularisations per head of the population with angina compared with some more affluent wards. Ferris *et al* (1998) studied the carotid endarterectomy treatment use rates and the need rates estimated based on the number of patients eligible for the treatment using an age-sex standardisation method for six areas in the former Wessex regional health authority. They found that, among six districts, the use-to-need ratio varies substantially ranging from 0.28 to 0.47, and this variation may be partially explained by referral patterns and inducing people to use resources which are nevertheless still under supplies.

Pollock and Vickers (1998) studied variations in emergency admission for some cancer patients in the Thames region. Using 1991 ED-based census data and "finished consultant episodes" data, they found that people with cancers of the bowel, lung or breast in the most deprived 10% EDs, in terms of Townsend's scores, were more likely to be admitted as emergencies and ordinary inpatients than their counterparts from more affluent areas. They also found that patients with lung or breast cancers from the deprived areas were less likely to receive surgical treatment. The authors felt that there may be a link between socio-economic circumstances and delay in reporting symptoms. GP referral diagnostic procedures in deprived areas may at least partially explain this situation. They also found that patients with colorectal cancer from the most deprived area were less likely to be seen at hospitals with a large caseload, (i.e. with cancer treatment units) than patients from affluent areas. Their findings have profound implications for the ongoing re-organisation resulting from closing down cancer treatment units with low caseloads.

There is a growing interest in the geographical assessment of equality in the uptake of preventive health services. This is partly the result of the realisation that some preventive health services failed to reach those people who need them most (Holland and Steward 1990, Majeed *et al* 1994). Studies show that the uptake rates of preventive health services vary geographically and there is an association between them and socio-economic and demographic factors as well as the characteristics of medical practices (Johnson 1987, Majeed *et al* 1994, Gatrell *et al* 1998, Kreuger *et al* 1999). In a study of the effectiveness of cervical cytology screening in Sheffield, Johnson (1987) found that there is a marked spatial variation in routine cervical smear testing rates. Using both the Jarman's score and the proportion of residents in the last two social classes in each ward as measures of socio-economic circumstances, they found that "a highly significant inverse correlation between deprivation and smear uptake among women over the age of 35 years". Kreuger *et al* (1999) report a similar study carried out in Rotterdam in the Netherlands. They found that percentage attendance to cervical screening in 53 selected neighbourhoods varies from 36% to 58% and that, using regression analysis, marital status remains significant after putting in other relevant explanations.

Majeed *et al* (1994) and Gatrell *et al* (1998) carried out studies on the uptake of breast cancer screening in parts of London and South Lancashire respectively. Both studies were intended to explain variation in attendance by socio-economic factors and by characteristics of general practices. By fitting a multiple regression model to a set of demographic, socio-economic and practice variables on the uptake rate for 126 practices, Majeed *et al* found that four demographic and socio-economic factors and one practice factor - the presence/absence of a female partner were significant. The presence of a female partner increases the uptake rate by 12.5%. By fitting a logistic model and using the Carstairs score and the presence/absence of a female partner on the uptake rates, Gatrell *et al* found similar results for two rounds of screening (1989-1992 and 1992-1995). However, in the second round the presence of a female partner in a practice becomes less important than in the first round. One implication from both studies is that policy makers might need to take into account the practice characteristics for organising disease screening.

Spatial statistical analysis may be used to identify clusters of areas that have low uptake of immunisation of a contagious disease. If any of the clusters form a sufficiently large population at risk, such information may be used to predict where that disease may break out (Meade *et al* 1985, p238). In this case, an additional immunisation programme targeting that cluster may prevent the outbreak. Given the availability of treatments for breast and cervical cancers (Shapiro 1977, Quinn *et al* 1999), for example, directing additional resources to those areas likely to have low uptake rates for corresponding screening tests may help reduce substantially morbidity and mortality. Since areas with low uptake of some preventive health services are likely to be socio-economically deprived (Jonhson 1987), identifying 'cold spots' with respect to socio-economic deprivation may be useful in this respect. Information on covariation may be of use in controlling their confounding effects if a research question is asked about the role of other factors on the uptake of a preventive health service, such as distance to the health centres where the services are provided. Such information may be useful to 'predict' those areas that are likely to have low uptake of new preventive health services in the future and to design appropriate programmes.

The first part of this section discussed five types of health study that fall into main stream in the spatial epidemiology (Elliott *et al* 1992). It emphasised that they are closely related to and may complement one another in a specific study. Examples were given to show how statistical techniques might be used in these studies. Problems likely to arise in spatial epidemiological studies were highlighted. The second part considered two health services studies concerned with the needs of geographically defined populations for and their access to and use of health services. The review showed how a geographical perspective may help understand the variations, often observed in mortality and morbidity statistics as well as uptakes of preventive health services across areas, and relationships between them and other factors for providing useful information for planning and delivering health services.

3.2. Health-related Data

The previous two sections considered some major types of studies in spatial epidemiology and health service research. Some health-related data was mentioned but not discussed. The first part of this section considers the sources and availability of health-related data. The second part considers the issue of confidentiality of medical and census data and its implications in choosing SSA techniques. The last part is concerned with the accuracy of health-related data, including the ways that errors may be entered into the data sets and the effects on the analysis of health-related data.

3.2.1. Sources and availability of health-related data

Health-related data may be divided broadly into two groups - health data and data about the demographic, socio-economic and environmental circumstances of the population. Health data may include data on medical conditions of individuals or the population and on various aspects of health services activities.

Mortality and morbidity statistics are the two most important types of medical data. In many countries, mortality statistics are collected routinely through death registration (Lepoz 1992). In the UK, OPCS (Office of Population Census and Survey) maintains a digital file which records postcoded data for each person who has died in England and Wales. Each record contains such items as the date and place of birth, sex, marital status, occupation and cause of death coded according to the ICD (International Classification of Diseases) (WHO 1977). Other mortality statistics including perinatal and infant mortality "rates" (proportions) are also available (William *et al* 1998).

Some morbidity statistics are also routinely collected in some economically advanced countries. In the UK, the General Household Survey (GHS) is a continuous survey of non-institutional households. It records not only health data on self-reported illness but also data on housing, employment and education. The statistics are presented in tables dis-aggregated by age, sex, economic activity and socio-economic group (Jones and Moon 1987). In many countries, cancer incidences are collected routinely. The UK cancer registry records full postcoded data on all people who suffer from cancer. Each record contains items including age, sex, and the time of diagnosis, and the type of cancer (Swaldows 1992). Other sources of morbidity data may include registers for drug addiction, congenital abnormalities, specific diseases (such as diabetes and stroke) and communicable disease notification as well as morbidity data collected by the Royal College of General Practitioners from sample practices around Britain (William *et al* 1998).

Data on health service activities is often collected routinely. The UK hospital consultant episode statistics are collected and available from the Office for National Statistics and the Department of Health (Pollock and Vickers 1998). FHSAs (Family Health Service Authority) maintain records on the uptake of immunisations and disease screening, GPs and people registered with them. Such data are of use in understanding the uptake of preventive health services, assessing the requirements for GPs across regions, and deriving more accurate estimates of the population between census years (Majeed *et al* 1994, Gatrell *et al* 1998, Benzeval and Judge 1996, Lovett *et al* 1998).

Census data is a main source for demographic and socio-economic data and is collected routinely in many countries. In the UK, the census is carried out every ten years on households throughout the country. The census questionnaire forms cover two classes of questions: the first about the people at each household, such as age, sex, and marital, educational, occupational and employment statuses;

and the second about the household itself such as the nature of household (e.g. sharing entrance with other households), amenities and the number of rooms, people and cars (Rhind 1983).

Environmental data is often collected through specific programmes and may be available at a set of sampling locations or areas. For example, Co-ordinated Information on the European Environment programme, CORINE, resulted in a collection of many types of environmental data on different spatial scales (Maes and Cornaert 1995). The programme has established a data set of industrial emissions into air from 1985 to 1990. It also includes other data sets on soil types, water resources, and climates.

3.2.2. Confidentiality of health-related data

Although medical and census data is collected for individuals, for the reason of confidentiality they may be neither available to the public in their original forms nor allowed to be used in any publication where individuals or households being identified (Quinn 1992). For example, in the UK, there are many legal and other guidelines on confidentiality in place. For this reason, data that may lead to individuals or households being identified will not be released to the public. In medical data available to the public, items that can be used to identify individuals directly and indirectly are often excluded. For pure research purposes, although researchers may be allowed to access medical data such as cancer registry from which individuals can be identified, they can present only anonymous aggregated data in publications. Census data obtained at the household level are aggregated over some regional system before being made available to public. The UK census data is available in the form of Small Area Statistics (SAS) at a number of area levels, including enumeration districts (EDs) and district electoral wards in England and Wales, and postcode sectors in Scotland (Rhind 1983).

Because of confidentiality, the analyst may have to choose the most appropriate type of SSA techniques in a study according to the data available. For epidemiology, one particular task is to estimate the population at risk. As mentioned previously, the use of point pattern analysis techniques may need to estimate the underlying population using controls adjusted for socio-economic and demographic circumstances of the population since they are confounding. In this case, publicly available area-based census statistics cannot be used directly. In order to use these techniques, a study may have to draw controls and collect required data. Although controls may be easily drawn from other data sources such as health records of another disease as suggested by Diggle *et al* (1990), it is socio-economic data about the controls that may be expensive and time-consuming to collect since they usually do not exist in medical data sets. Moreover, the use of collected data is also subject to confidentiality constraints.

3.2.3. Data accuracy

The analysis of health-related data must recognise that the quality of health-related data may be very limited. For mortality statistics, misdiagnosis is a source of errors. Owing to the nature of the disease, exact cause of death is sometimes difficult to establish. Death caused by cancer of the pancreas and other internal organs cannot be determined without an autopsy (Meade *et al* 1985, p 12). Misdiagnosis is more likely to occur for people dying at age 75 or over since they are least likely to be autopsied (Boyle 1989, Lopez 1992).

Errors may arise resulting from inconsistency in disease classification. Cancer of the stomach may be coded according to its nature in different ICDs (Boyle 1989). Several studies have been undertaken examining this problem using methods including comparing clinical diagnosis against autopsy, reviewing complete case histories against death certificates, and duplicate coding of death certificates (Lopez 1992).

Morbidity statistics are also subject to errors (Boyle 1989). For example, GHS statistics are thought to be less accurate because of the lack of medical examinations and therefore may be limited for epidemiological purposes (Jones and Moon 1987) but may still be useful for health services research. For example, Benzeval and Judge (1996) used GHS data for 12,000 adults in England and census data to evaluate equality in the GPs provision across FHSAs. Although cancer registries are thought to provide the most reliable cancer data, they may still be subject to a number of errors (Swaldows 1992). For example, some cancer patients may never present to any doctor. Errors may also arise because of incomplete registration, occasional errors in registration, and crossing cancer registry registration. Duplications in recording may arise when the same person is diagnosed with cancer at more than one site. Data accuracy in uptakes may be subject to problems such as incomplete record and list inflation as letters of invitation may never reach those for whom they are intended (Holland and Stewart 1990).

In the collection of census data for households, errors may arise owing to the absence of people at census time and "double" counting of people absent from their places of residence (Rhind 1983). For confidential reasons, the census data for each ED is "distorted" by a quasi-random number +1, -1 or 0, and for those small areas that include fewer than 25 people, the data is suppressed. In addition, not all SAS are based on 100% populations. The SAS for the 1981 Census are presented in 53 tables where the last 10 tables contain statistics based on a 10% sample of the population.

Data accuracy may arise in other ways. If the health events, census and environmental data is available for individuals, census tracts and large areas respectively, and are to be used in a specific study, they may have to be projected onto the same areal framework (Pukkala 1992). So the health events may have to be aggregated, the environmental data dis-aggregated or discretised, and the census data on an incompatible area system interpolated (Gatrell 1989, Flowerdew and Green 1994). During these transformations, errors may enter into the resulting data set, depending on the nature of the data and the techniques used for the transformations. For example, if postcoded disease incidences are to be allocated to EDs in a point-to-area search approach using the Central Postcode Directory (CPD), which matches postcodes to EDs with an accuracy of 100 meters, many events may be assigned to the wrong EDs. The errors would be fewer if a more accurate address code such as PinPoint Address Code (PAC) is used with full-addressed disease cases (Gatrell 1989).

When environmental variables on continuous surfaces are discretised onto an area framework where a region is delineated into sharp areas, a certain number of errors may arise as a result of this misrepresentation (Goodchild 1989). The population for non-census years estimated using the census data might not be sufficiently particularly in urban areas where there has been, for example, housing redevelopment (Lovett *et al* 1998). Errors may exist in spatial features if area boundaries are digitised from a shrunk paper map.

It has been acknowledged that errors in spatial data may influence the analysis of the data (Goodchild and Gopal 1989). The extent of the influences may depend on the nature of a study and statistical techniques employed. If a study is to estimate the mean and standard deviation about the sampling mean for a variable, the estimates may be affected quite strongly. This is because these statistics are not "robust". However, a robust estimator such as the median would not be affected heavily by a few outliers. In the same study, if the values are the number of disease cases, a random error between -1, 0 and 1 in each area would have little impact on the estimates of the number of incidences if the population is roughly the same and sufficiently large in each area. If the population is small, such random errors may give rise to the small number problem discussed in Chapter 2. But the same errors will be less influential on more robust estimators (Claydon and Keldor 1987). Errors on spatial features may propagate during operations as overlaying and buffering (Heuvelink and Burrough 1989). It is clear that statistical analysis must provide statistical techniques that are resistant to data errors in order to help identify the errors and produce sound statistical results.

This section considered the sources and availability of some routinely collected data. It also considered issues relating to data confidentiality including the selection of appropriate SSA techniques. The last part discussed the quality of the health-related data and the ways in which errors enter the data as well as their influence on analysis results.

3.3. GIS for health research

The previous chapter discussed GIS and its strengths and weaknesses in SSA. This section will consider the potential of GIS for health research and kind of GIS that is the most appropriate for health research. As stated at the beginning of this chapter, this section is not intended to be a comprehensive review of what GIS has done and will be able to do for health research.

3.3.1. GIS for health research

Recently there has been a wide interest in examining and exploring GIS for health research (Twigg 1990, Hirschfield *et al* 1995, de Lepper 1995, Gatrell and Löytönen 1998, Haining *et al* 1996, Haining 1998,). One of the principal capabilities of GIS useful for health research is that it is able to manage large spatial data sets because contemporary health studies, in particular at small areas, often require analysing a large number of spatial features and attribute data. In a study of colorectal cancer in the city of Sheffield, UK (Haining *et al* 1994), a census data set contains small area statistics for over one thousand EDs in the city. Even though such a data set could not be regarded as large, it cannot be managed effectively for the purpose of spatial analysis without GIS. As powered by database management systems, GIS allows a large number of attributes managed in one or more separate data tables to be related to attribute tables of spatial features through primary keys. For example, FHSA postcoded GP register data can be linked with the census data by a postcode to ED lookup table (Raper *et al* 1992). This enables the analyst to estimate characteristics of the populations served by GPs and to analyse the uptake of preventative health services (Gatrell *et al* 1998). It also enables the analyst to derive more accurate estimates of the populations for areas between census years (Lovett *et al* 1998).

Overlay is another key feature of GIS allowing different data sets for a region to be integrated in a study. This is fundamental to health research because, as discussed above, health studies often need to link health-related data in different data sets possibly using different types of spatial features. For example, in a study of the relationships between air pollution and certain health conditions such as asthma, the major sources of air pollution, such as roads and industrial installations, may be overlaid one another over a layer of areas coloured according to its standardised asthma incidence ratio. This map may help the reader examine visually the geographical associations between the incidence of asthma and the level of air pollution. As discussed in Section 2.3.1, the use of overlay goes beyond generating an overlaid map to data integration and data query. For example, the point-in-polygon operation allows point data sets to be integrated with a polygon data set in a study. If postcoded disease incidences are contained in one or more point data sets where each incident is given the co-ordinates of the corresponding postcode (Heywood *et al* 1998, Gatrell 1989), one can obtain the number of disease incidences falling into each polygon.

GIS proximate analysis is also important in health research. As discussed in Section 3.1.1.4, in order to investigate the effects of putative sources on human health, the analyst has to define buffer zones around the sources which may be waste incinerators, nuclear installations, rivers, roads or high voltage electricity lines (Gatrell and Dunn 1995, Wilkinson 1998). The polygon dissolve operation is crucial when area-based data on a fine geographical scales are available and regionalisation is required to create an appropriate regional framework (Haining *et al* 1996, Wise *et al* 1997) and to delineate neighbouring communities for locality planning (Kivell *et al* 1990, Bullen *et al* 1996).

It is of great value for health research to be able to query spatial features, attributes and measurements. For example, the analyst may wish to identify areas with specific characteristics in their attributes, say, areas with SMRs greater than 200. In order to specify inter-area relationships, the analyst may query the adjacency between areas, the distance between area centroids and the length of the shared boundary between areas. The analyst may like to obtain the intersection of two overlapping areas for data interpolation (Goodchild and Lam 1980, Flowerdew and Green 1991). The analyst may wish to query the spatial data sets of road networks as well as attributes associated with them to derive measures of patient accessibility to health services (Gatrell 1999).

The discussions above clearly show the usefulness of GIS functions in health research. With these functions, the user can carry out the preliminary analysis on 'raw' data sets. Simple statistics on the data sets can be derived in conjunction with data querying, and displayed as maps for examination. This also helps identify data problems and suggest possible ways to overcome them. Spatial operations like the polygon dissolve make it possible to create an appropriate regional framework for analysis.

3.3.2. Which GIS is appropriate for health research

Common to the most GIS is the ability to answer questions: 'what is it at...?', 'where is it ....?'and 'what has changed since ....?'. As discussed in Chapter 2, GIS is often lacking sufficient spatial statistical tools to answer questions like: 'what spatial pattern exists in ...?', 'what explains the patterns in ...' and 'what if...?'. Given the discussion so far, clearly, it is these questions that are of particular interest in health research. Indeed, the first three questions may only be of interest in the light of information derived from the last set of questions. As discussed previously, for the purpose of health resources allocation it is important, for example, to identify areas that are outliers beyond what is being explained by variables such as deprivation. This usually requires the use of modelling tools with assistance of exploratory tools, like the XW plotting to be discussed in the next chapter.

This indicates clearly that, in order to enable a coherent data analysis, a GIS suitable for health research needs to provide SSA tools that can be used to answer every type of spatial questions. As stated in Chapter 1, the aim of this research project is to integrate state-of-the art spatial statistical techniques with a GIS so that they together meet this requirement.

Since many SSA techniques different in statistical complexity may be used to answer the same question, a GIS ought to support a range of them. This is because people who work in the area of health research may come from very different disciplines - epidemiology, geography, health planning, health service management and so forth. They are likely to have quite different levels of knowledge on statistical theory and methods. Clearly, different groups of researchers will prefer those SSA tools which they feel most confident. Haining (1994) considered the two extremes of this spectrum of SSA users as 'experts' and 'non-experts' and SSA tools that the experts and non-experts may prefer. For example, for the purpose of examining spatial trends of disease incidences across areas, two tools may be employed: mean or median smoothing and trend surface modelling. A non-expert analyst may prefer the first because of its simplicity, while an expert analyst may like to build a trend surface model to give a quantitative description of the data, though the model is likely to be built based on what he or she might have found with the first tool.

This section has examined the potential of GIS for health research underpinned by some fundamental GIS functions. It argued that a GIS suitable to health research must provide SSA tools to answer typical health questions, in particular those concerned with the spatial patterns of health-related events and modelling of them. It also argued that if a GIS is to serve health research well it ought to provide SSA tools with varying sophistication to meet the needs of researchers with varying knowledge of statistical theory and methods underpinning the SSA tools.

3.4. Summary and Conclusion

The main theme of this chapter has been to show how SSA and GIS may be of use in health research. Major types of studies involved in health research were discussed briefly. Examples were given to illustrate the use of some key SSA techniques in answering health questions. Problems likely to arise in health studies were also considered.

The possible sources of routinely collected health data were briefly summarised. The issues relating to data confidentiality and quality were considered. Data confidentiality may constrain the choice between point pattern analysis and lattice data analysis techniques in a study that relies on the use of censuses. The quality of health-related data may be affected as a result of the various types of errors entering into data sets and influences the data analysis in particular at small spatial areas. Consequently the analysis of such data calls for robust SSA techniques.

The role of GIS in health research was examined in line with types of GIS functions. Some common GIS functions for managing, manipulating and visualising spatial data were shown to be useful to help with some typical health questions that would otherwise not be answered easily, if at all. If GIS is to offer more for health research than it does at present, it needs to offer a range of SSA techniques in order to facilitate a coherent analysis of health-related data in order to answer the sorts of questions typically asked. In terms of complexity, GIS should provide the simplest exploratory techniques to the most complicated confirmatory and modelling techniques to meet the needs of users with varying levels of expertise in spatial statistics.

CHAPTER 4. SSA TECHNIQUES FOR THE ANALYSIS OF AREA-BASED HEALTH-RELATED DATA

The previous two chapters conveyed the followings. First, both SSA and GIS could play more important roles in health research than they do at the present if appropriate SSA techniques can be integrated with GIS. Second, SSA techniques to be integrated with a GIS should enable the analyst to undertake a coherent analysis of area-based health-related data. Third, SSA techniques with different complexity should be provided to meet the needs of diverse users. This chapter reviews some SSA techniques that could together satisfy the second and third requirements above. It also discusses the technical aspects of these techniques in some detail in order to show the requirements for implementing them. These techniques form the core SSA functionality in SAGE.

For the purpose of this discussion, these techniques are divided into three classes. The first class contains techniques for constructing appropriate regional frameworks and specifying inter-area relationships. These techniques are referred to as data preparation techniques and are considered in Section 4.1.

The second class contains techniques for analysing data of a single variable or map data, ranging from simple descriptive techniques to sophisticated statistical tests and modelling techniques. These techniques are divided into two groups according to whether they are used to identify non-spatial or spatial properties. The techniques in the latter group are sub-divided into mapping techniques, and techniques for identifying spatial trends, global spatial dependence or clustering, or localised spatial clusters and spatial outliers. Section 4.2 considers these techniques and highlights the benefits of 'hot' linking statistical plots with maps in the exploratory analysis.

The third class contains a set of exploratory and confirmatory techniques for analysing multivariate data sets. Section 4.3 reviews these techniques, giving a primary focus on regression analysis and modelling techniques catered specifically for spatially correlated data. Some related methods for fitting and validating the models and for diagnosing the model fit are also considered. Section 4.4 concludes this chapter.

It should be noted that the order in which the techniques are discussed does not imply that the analysis of area-based data should be carried out in this order. For example, the use of regionalisation is only justified if the 'raw' data would cause difficulties in using the necessary SSA techniques in a study. The analyst may be able to know whether one or more difficulties are present only when she/he analyses the 'raw' data. Indeed, the techniques should be used in a sensible order in line with an intended study.

4.1. Data preparation

4.1.1. Regionalisation

Regionalisation can be regarded as the following type of data process. Given an initial regional framework with N areas, regionalisation aggregates these N areas to form a new regional framework with K (K<N) areas according to a set of criteria. The attributes attached to the areas in the original framework are aggregated accordingly onto the new framework. Through such a process, a set of spatial data is transformed to another which is more appropriate for performing spatial analysis than the original.

An assumption made implicitly here is that all required raw data, which may be in different data sets in different forms such as point data and area-based data, have already been rectified so that all required attributes are correctly associated with areas in a regional framework. Data rectification, as briefly discussed in the previous chapters, consists of a body of spatial analysis methods and techniques in its own right and the discussion of them is beyond the scope of this research and will not be pursued further in this chapter.

4.1.1.1. Reasons for regionalisation

Sections 2.1 and 3.1 highlight two main problems that are likely to arise in studies of rare diseases in small areas: the small number problem and heteroscedasticity. As in many other studies, the problems have been highlighted in a study of colorectal cancer incidence for the city of Sheffield (Haining *et al* 1994). The number of colorectal cancer incidences per year for the city of Sheffield is about 300 but there are over one thousand EDs (1981 census) containing a population of roughly 300,000 aged between 30-85 (see Chapter 6). Even for a combined period of five years, the number of incidences is still too small for most areas to attract sufficient cases, if any. Moreover, the number of people in EDs varies more than ten-fold, ranging from just above 30 to 500. Thus, if EDs were taken as an analysis framework, both problems would certainly arise. As discussed in Section 3.2.3, data errors on this scale are likely to make these problems even worse.

In the UK, electoral wards form a much coarser area framework on which the censuses are available. In the City of Sheffield, there are just 29 wards in the 1981 census. Clearly, these problems are less serious at the ward level. However, at this level inter-area homogeneity in terms of socio-economic characteristics becomes a serious problem since the wards are demarcated for electoral purposes. For health studies, where relationships between disease events and socio-economic circumstances are sought, the electoral wards do not provide an appropriate framework.

Given that these problems are associated with the available areal framework, if data at fine scale is available for a study, regionalisation offers a way to build an appropriate areal framework on the right scale. Another reason for regionalisation is to reduce the size of a spatial data set in order to make statistical analysis tractable computationally. Many spatial statistical techniques involve intensive computation. For example, many modelling inferences may require computing the inverses of a NxN matrix where N is equal to the number of areas in an areal framework. A spatial data set with a thousand areas would make the computation "intractable" on many computers with modest capacity.

4.1.1.2. Criteria for regionalisation

Regionalisation must be able to take into account different criteria that specify how a regional framework is constructed. Three criteria important in health studies are homogeneity, equality and compactness (Wise *et al* 1997).

Homogeneity implies that only those areas similar in terms of the values of a set of selected attributes are allowed to be merged. This is essential if socio-economic and demographic attributes are to be used to explain the spatial variation of health events or to control for confounding effects.

Equality implies that a resulting regional framework should have a characteristic such that for selected attributes their values are similar across areas. As mentioned previously, the reliability of relative risk rates depends on the variability of the population across all areas. Therefore, equality in terms of the population could not only help overcome the small number problem but also make relative risk rates to be equally robust to possible data errors or perturbations.

Compactness imposes geometric constraints on a regional framework in order to eliminate areas with odd shapes. This is sometimes regarded as merely a cosmetic aim (e.g. Openshaw and Rao 1995) but Horn (1995) argues that it accords with our intuitive understanding that regions within cities form tight units of economic and social activity and hence are spatially compact. For the purposes of health service provision, compact areas are likely to be much more helpful than areas which have elongated shapes (Rossiter and Johnston 1981).

Besides the three criteria above, contiguity is another criterion required in regionalisation. Without it, regionalisation becomes classification. In this case, areas form a new area if they fall into the same group and happen to be adjacent to one another.

The selection of criteria for regionalisation depends primarily on the purposes of a study. If the study is intended to explore the spatial variations of disease rates by means of mapping, an equality criterion in the population is likely to be the most important whereas a homogeneity criterion on socio-economic status may not be a priority. The compactness and contiguity criteria may even be irrelevant. However, if the purpose of the study is to find the evidence of the association between the incidence of a disease and socio-economic factors, a regional framework ought to be created with a homogeneity criterion on these factors. The equality criterion on the size of population is also required. Again, the compactness and the contiguity criteria may not be relevant. If the results of a study are to be used for delivering health services, the compactness and contiguity criteria are likely to be required. It should be noted that the criteria discussed above would compete with one another during regionalisation and therefore they should be weighted properly. The table below summarises the above. The letter H, M, or L indicates high, medium or low priority respectively.

Objective	Homogeneity	Equality	Compactness	Contiguity
Mapping disease	M or H	H	L or M	L or M
Associative studies	H	H	L or M	L or M
Delivery health resources	H	H	H	H

Table 4.1. Priority in choosing different criteria.

4.1.1.3. Regionalisation methods

Regionalisation is a combinatorial problem. There are many ways to partition N area units into K (K < N) new area units. The number is often enormous with even a modest number of area units (Cliff *et al* 1975, Rossiter and Johnston 1981). Although it is feasible theoretically to enumerate all partitions in order to find the optimum with respect to a set of criteria, it hardly makes sense in practice to do so knowing no answer could be obtained in any reasonable length of time. Therefore, methods are employed to find a better but not necessarily optimal partition.

Many regionalisation methods are based on two types of classification or grouping approaches - hierarchical and heuristic approaches (Everitt 1979, Anderberg 1973, Spath 1980). A hierarchical classification process may take either an agglomerative or divisive procedure. Starting with N groups, each of which contains only one object, the agglomerative classification merges the two most similar groups in terms of some kind of similarity measure for every pair of groups. After a new group is formed, the similarity measures are updated. This process continues until there is only one group. Starting with a single group including the N individual objects, the divisive classification splits one group at a time into two until N groups are formed. Different strategies may be applied to decide which group, and how it should be divided up (Spath 1980). The product of a hierarchical classification is a hierarchical tree or dendrogram enabling the derivation of any specified number of groups. For both types of hierarchical classifications, merging or splitting strategies are usually designed optimally at each individual stage but not for the given number of groups (Spath 1980 p155). The hierarchical classification can be modified for regionalisation by allowing only those areas that do not violate the continuity constraint to be merged and split (Berry 1961).

Heuristic regionalisation employs iterative techniques to search for an 'optimal' partition. After each iteration, a better partition with respect to given criteria is found. Of the many heuristic methods, the K-means method is probably the best known and widely used (McQueen 1967, Anderberg 1973, Spath 1980). There are two basic elements in the K-means method: an initial partition with K groups on N objects (i.e. areas in regionalisation) and an objective function measuring whether one partition is better than another. Starting from an initial partition, the method aims to find a partition so that the objective function is 'optimal' through a procedure as follows. First, select any object from a group; second allocate the object experimentally to every other group in turn to find out an allocation that results in the greatest improvement to the objective function; third, assign that object to the group to form a new partition. Repeat the procedure for every object cyclically until no further allocation can improve the objective function. In this process, when a group contains only one object, that object is not allowed to be assigned to another group. Therefore, the K-means method always produces exactly K groups.

The K-means method has been widely used in regionalisation and is a central part in many regionalisation procedures. Openshaw (1978) designed an automatic zoning procedure (AZP) for the study of aggregated data. This procedure was adopted to solve a constituency-delimitation problem (Rossiter and Johnston 1980). For solving general zone partition problems, Openshaw and his colleague (Openshaw and Rao 1994) developed a zone design system (ZDES).

Horn (1995) developed a procedure for delineating electoral divisions in Australia.

Wise *et al* (1997) report a K-means-based regionalisation method capable of constructing an areal framework based on all criteria discussed above. In this method, homogeneity is expressed as the sums of squares of within-group variance for selected attributes. Equality can be stated as the sum of squares of within group differences between the sum of a given variable (e.g. the population) and the average of that variable for all groups. Compactness is measured as the sum of squares of within-group variance for the X and Y co-ordinates of area centroids. An objective function can then be defined by the combination of these three functions while contiguity can also be specified. These functions could be weighted appropriately to reflect the relative importance of them in regionalisation.

The K-means method is rapid and often converges on a solution in a few iterations (Spath 1980). However, it can only guarantee to converge to a local "optimum" rather than a global "optimum", conditional on the initial partition (Spath 1980). This problem may be more serious if the contiguity criterion is applied in a strict manner. In a recent development of the AZP approach, Openshaw and Rao (1995) studied techniques such as simulated annealing to prevent the search for a global optimal partition from being stopped at a local optimal partition. Macmillan and Pierce (1994) also used the simulated annealing technique to solve redistricting problems where equality in the population is considered. However all these methods result in much slower convergence.

Other procedures might be taken to deal with this problem. One procedure is to run the method many times with a different initial partition at each run. The best partition of all is then selected as the final partition. An initial partition can be generated using hierarchical classification. One can also select K seeds (i.e. areas) randomly and assign the remaining areas to the seeds to which they are closest.

Horn (1995), in his work on delineating electoral divisions in Australia, suggested a procedure. This procedure allows the number of contiguous sub-regions more or less than the expected to be formed temporarily, and therefore helps prevent regionalisation from converging to a local sub-optimal partition.

Wise *et al* (1997) suggested a procedure that deals with this problem as follows. First, a threshold is introduced and any allocation is allowed if the current value of the objective function relative to the previous value does not get worse than that threshold. Second, the following actions may follow. The worst group with respect to the defined criteria is broken into two to form a new partition, and the K-means method is run on it as the initial partition. Adjacent groups, the merge of which leads to the least change in the objective function, are merged and the K-means method is run on it. These may repeat many times where merging may take place before the breaking.

In practice, a trade-off between finding a better and finding an optimal partition has to be made. If very powerful computers are available, it may be best to use the K-means method in conjunction with simulated annealing technique in order to find a partition that is close to a real optimal partition. Otherwise, it would be better to use the procedures that find local optimal partitions. Since regionalisation may be used as a tool to explore the spatial distribution of multivariate data, it is arguable that a local optimal partition is sufficient at the ESDA stage (Wise *et al* 1997).

Figure 4.1 shows a flow chart of a five-stage regionalisation procedure incorporating a number of methods suggested in Wise *et al* (1997) and discussed above. Arrows in the figure indicate the possible iterations between stages. This procedure underpins the development of SAGE's regionalisation module.

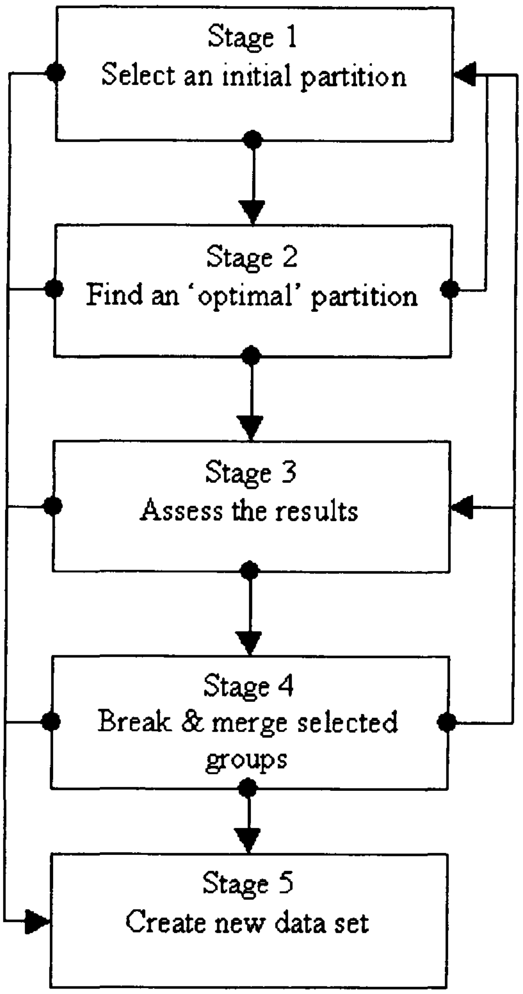

Stage 1
Select an initial partition
Stage 2
Find an 'optimal' partition
Stage 3
Assess the results
Stage 4
Break & merge selected groups
Stage 5
Create new data set

Figure 4.1. A regionalisation procedure for creating regional frameworks (after Wise et al 1997).

4.1.2. Specifying inter-area relationships

Many spatial statistical techniques require specification of the relationships between areas in the regional framework. In some studies, if there are well-developed theories available, the inter-area relationships may be specified based on them. For example, if a disease is known to spread mainly as a result of human contact, one may define the relationship between any two areas based on the measures of human contact taking place. When such direct measures are not obtainable, indirect measures may be developed. For example, the counts of commuters between two areas per unit time period may be an alternative (Haggett 1976, Cliff *et al* 1985 pp. 182-5). In assessing the needs for the health services, central place theory and Hart's inverse care law (Hart 1971) may together help specify inter-area relationships. Evidence indicates inverse care in declining industrial cites in the USA, particularly in the inner cites, and in inner cites in the UK as well as in rural areas and peripheral local authority housing estates (Jones and Moon 1987 p 239). These suggest a distance decay relationship relative to a fixed "central" location for those areas.

However, in many situations, as a result of the lack of understanding of underlying spatial mechanisms, the analyst may have to choose to specify inter-area relationships that he/she believes to be appropriate. This gives rise to a concern of the sensitivity of analytical results to the specification of the inter-area relationships (Haining 1993).

Given a regional framework, the inter-area relationships for all pairs of areas can be expressed using a n×n matrix, $W= \{w_{ij} \mid i, j = 1, 2,..., n\}$, called the connectivity matrix, where w_{ij} is a measure of the relationship from area i to area j. It should be noted that although a symmetric relationship for any two areas may often be appropriate, directional or asymmetric relationships may be more appropriate in a study of a diffusive disease along certain directions (Cliff *et al* 1981). w_{ij} is commonly specified using metric and/or non-metric information, including the distance between the centroids of any two areas, the length of the boundary shared by the areas, their adjacency or the combinations of these (Haining 1993, p. 69-74). A connectivity matrix based on the adjacency between areas only defines w_{ij} to be 1 if area i and area j are adjacent to each other or 0 otherwise. Since elements of a W matrix are often used as weighting factors by SSA techniques, they may need to be standardised. A common form of standardisation is that each row is standardised so that the sum of elements in that row is unity.

This section has considered the reasons and three important criteria for constructing areal frameworks and described in some detail a heuristic classification method, the K-means method, and a regionalisation procedure based on this method. It also considered ways in which appropriate inter-area relationships may be specified.

4.2. Analysing data on a single variable

As discussed in Section 3.1, health research needs to analyse the distribution of health events. This is often done by identifying the properties of data of a single variable, sometimes called map data. The properties can be classified as non-spatial properties and spatial properties defined in Equations (2.1) and (2.2) respectively.

4.2.1. Identifying the non-spatial properties

The centre of a distribution and the spread about the centre are two of the most important properties characterising the distribution of a variable. The centre may take the form of the mean or the median, and the spread the standard deviation or distance between the upper and lower quartile – the inter-quartile range. The mean and the standard deviation are the most frequently used classical statistics but these are likely to be influenced by extreme values. The median and the inter-quartile, on the other hand, are robust to extreme values and are often preferred in the analysis of health-related data when the data quality is low. If Fu and Fl denote the upper and lower quartiles, the samples which values are greater than (Fu + 1.5 (Fu-Fl)) or less than (Fu - 1.5 (Fu-Fl)) may be defined as outliers (Haining *et al* 1997). Since outliers can have an influence on analysis, the corresponding sample deserves a close look into whether its value is an accurate representation of it or a distorted representation due to the errors. In the latter case, samples might have to be excluded from any further analysis.

Although these statistics are usually summarised in numerical form, they could be displayed graphically for better intuition (Tukey 1978, Cleveland and McGill 1988). For example, one way to summarise the distribution features of a variable is to construct a histogram where bins are pivoted at the mean and the width of each bin is set equal to one standard deviation. Another way is to construct a box plot using Fl, Fu and the medium where outliers are represented as points (Tukey 1978).

Although these graphical techniques were devised mainly for analysing non-spatial data, they could be made to answer questions "where is it...?". When they are 'hot' linked with an area map, a selection of a part of a plot will cause all areas whose values fall into that part to be highlighted on the map (Haslett *et al* 1990). With a histogram plot of relative risks, for example, the analyst could select a bin at one of the ends to identify areas whose relative risks depart most from the mean. With a box plot of the same values, the analyst could easily locate those areas with extreme values. When the 'hot' linking is made bi-directional, an area could be selected to highlight its attribute values in one or more plots or an attribute table. In other words, a 'what is it at ...' question can be answered in a simple manner.

4.2.2. Identifying spatial properties

4.2.2.1. Mapping disease events

As discussed in Section 3.1, mapping disease events is a powerful tool in spatial epidemiology. Given an area framework, estimates of relative risk rates for each area are obtained and mapped. Relative risk rates may be formulated as O_i / E_i, where O_i and E_i are the observed and expected number of disease events respectively in a given period of time for the *i*th area. Events may refer to the number of deaths, new cases or prevalent cases. O_i / E_i has the following property. Let θ_i be the relative risk in the *i*th area to be estimated. Under the Poisson assumption on O_i with a mean $\theta_i E_i$, the maximum likelihood estimate (MLE) of θ_i is O_i / E_i. In what follows the specification of a time period is implicit.

E_i may be defined in many ways. Let P_i and O_i be the number of people and the number of observed cases in the *i*th area respectively, and E_i may be defined as $\left(\sum_{j=1}^{N} O_j / \sum_{j=1}^{N} P_j \right) \times P_i$. E_i defined in this way is sometimes referred to as a raw rate. One drawback of this definition is that everyone is assumed to have equal opportunity to catch, or die from a disease. However, this is not always the case. Some diseases may affect only certain age and sex groups of the population. For example, colorectal cancer may have a latency period of 20 years or more, and, therefore rarely occurs among people aged under 30. Cancer of prostate occurs only in males. Breast cancer occurs mostly in females, although it does occur very infrequently in males. Both types of cancer are unlikely to occur in younger people. A relative risk rate that can adjust for both age and sex may be defined as $E_i = \sum_{j=1}^{M}\sum_{k=1}^{2} d(j,k) \times P_i(j,k)$, where $d(j,k) = \sum_{i=1}^{N} O_i(j,k) / \sum_{i=1}^{N} P_i(j,k)$, and i, j and k are the indices for N areas, M age groups and two sex groups respectively. $P_i(j,k)$ and $O_i(j,k)$ are the number of people and the number of observed cases in the *i*th area respectively for the *j*th age group of either females or male. A relative risk computed in this way is regarded as an indirect age-sex standardised mortality or incidence ratio - SMR or SIR (Meade *et al* 1985).

As emphasised in Chapters 2 and 3, map interpretation may suffer from the small number problem and heteroscedasticity. Although the impact of these problems on the interpretation is likely to be less than would be after regionalisation, the problems may still be present when more than one criterion is involved in regionalisation since they compete with one another. Hence, techniques are required to produce reliable estimates. A chi-square statistic $(O_i - E_i)^2 / E_i$ is such an estimate (Jones and Moon 1987). The effect of this is that areas with small populations must be much more unusual before they appear at the tail end of this distribution. Suppose there are two areas with the observed number of cases and expected number of cases of a disease 20 and 15, and 200 and 150 respectively. Then the relative risks computed using the previous formula would give the same value of 1.333 but the chi-square statistic would give 1.66 and 16.6 respectively. Carstairs and Morris (1991) use another similar estimate, $(|O_i - E_i| - 0.5)^2 / E_i$, in their study.

Clayton and Kaldor (1987) discuss Bayesian estimation techniques that yield the estimates of relative risks, which, taken together, may be better estimates than those given above of $\{\theta_i\}$. In other words, these estimates are more comparable across the areas than the estimates yielded above. The main framework of the Bayesian estimation given by the authors may be summarised as follows (Clayton and Kaldor 1978, pp. 672). Suppose that $\theta = \{\theta_i$, i =1, 2, ... N$\}$

are unknown relative risk rate ratios for N areas and a parametric probability density function f(θ) is assumed for the distribution of them between areas. In addition, conditional on θ_i, O_i is a Poisson variable with expectation $\theta_i E_i$. If the parameters in f(θ) can be estimated through the marginal distribution of $\{O_i\}$, the posterior expectations of $\{\theta_i\}$ given $\{O_i\}$ may be estimated and then provide the empirical Bayes estimates of the relative risks. For a gamma model with ν and α as scale and shape parameters for θ_i, the Bayes estimates of it is $\frac{O_i + \upsilon}{E_i + \alpha}$. Since $\frac{O_i + \upsilon}{E_i + \alpha} = w\frac{O_i}{E_i} + (1-w)\frac{\upsilon}{\alpha}$, where $w = \frac{E_i}{E_i + \alpha}$, θ_i falls between O_i / E_i and ν/α - the estimate of the prior mean. When E_i is large the estimate is close to O_i / E_i, whereas when E_i is small it is adjusted or shrunk towards ν/α.

Unlike the gamma model where θ_i is treated independently, other models may allow for correlated θ_i. Clayton and Kaldor (1987) consider a log-normal model which allows for spatial autocorrelation in log relative risks. This has the effect of driving each θ_i toward the mean of θ_i for its adjacent areas. Marshall (1991) reviews applications of this framework and other approaches.

4.2.2.2. Detecting trends

Many ESDA techniques are available for detecting the global properties of a single variable. Basically, these techniques function as filters that filter out unnecessary detail and enhance major features of the data. The resulting data is then mapped.

One set of techniques is based on classifying values into a number of groups. Each group then is assigned with a single colour and mapped. Such a choropleth map would show less mosaic than an unclassified map and consequently help highlight the major spatial characteristics of the data. A simple method divides the range of values of a variable into a number of intervals and assigns the values falling into the same intervals into the same groups. Intervals may be defined to have the same length, or to contain the same number of data values (Heywood *et al* 1998). The hierarchical and heuristic classification methods discussed in Section 4.1.1.3 can also be applied to a single variable. Different classifications may have different powers of preserving information in the data. Cromley (1996) compared some commonly used classification schemes for choropleth mapping of area-based data.

Another set of techniques is based on a simple form of kernel estimation in two-dimensional space (Silverman 1986). It involves passing a filter with a fixed sized window over each area. Two useful filters are the "local mean" and "local median" filters (Haining *et al* 1997). When the former passes through the ith area it replaces the original value of that area with the weighted mean, $MA_i = \sum_{j=1}^{N} w_{ij} X_j / \sum_{j=1}^{N} w_{ij}$. w_{ij} is an element of a connectivity matrix W and X_j is the value of the variable for the *j*th area. When the median filter passes the *i*th area it replaces the original value with the median, MM_i, of the values of those areas adjacent to the *i*th area inclusively. The size of the selected window may be defined using the connectivity matrix to a certain order. The bigger the size of a window, the more details are filtered out and the less influential the errors. Consequently, only the larger scale variation remain. The smoothed component of the map can be extracted from the map by computing (X_i - MA_i) or (X_i - MM_i). A map of (X_i - MM_i) would make those areas with particularly high values to stand out more strongly than in a map of (X_i - MA_i). This technique may be used to map relative risks $\{O_i / E_i\}$in a slightly different way. If $\{O_i\}$ and $\{E_i\}$ are considered to be observed on two surfaces, O(x,y) and E(x,y), it may be appropriate to pass filters on $\{O_i\}$and $\{E_i\}$ respectively.

As discussed in Chapter 3, a spatial trend relative to a specific location may be of interest. For a putative source of certain pollutant emission, the analyst may want to know whether the incidence of a disease is higher in areas close to that source than in areas further away. For the purpose of assessing inequality in health and in access to health services, the analyst may be particularly interested in knowing whether the inequality tends to vary relative to the distance from a location, such as the city centre. In some declining industrial inner cities, there is evidence that inequalities in needs for health services tend to decline toward the city centres (Jones and Moon 1987 p 239). A simple visual technique for describing this is to draw a number of box plots. Each plot is constructed using data for all areas adjacent to the specific area at a certain lag order. All plots are arranged in either a descent or ascent "lag" order. This technique might work well if the areas are similar in terms of the population. Haining (1993) demonstrated the use of this technique in the analysis of mortality data with increasing distance from the city centre of Glasgow, UK.

The detection of the trend of a variable could be made formal by fitting a trend surface model to the variable. The general form of the trend surface model is as follows (see Haining 1993 pp. 251):

$$\begin{aligned} &y = A\theta + \varepsilon \\ &E[\varepsilon] = 0 \text{ and } E[\varepsilon\varepsilon'] = \sigma^2 I \end{aligned} \qquad (4.2.1)$$

where y is an nx1 vector; A is the matrix of locations of the n areas such as area centroids; and θ is the vector of trend surface parameters. A and θ depend on the order of the trend surface.

This model can be treated as a special case of an ordinary multiple regression model and fitted by least squares. Model (4.2.1) can be extended to allow for autocorrelation in the error terms. That is $E[\varepsilon\varepsilon'] = \sigma^2 V$ where V does not need to be diagonal and may be specified using a connectivity matrix. A discussion on parameter estimation and model diagnostics will be given in Section 4.3.2. A recent example of using this trend surface modelling technique is Haining's work (1990) where the author fits a second order trend surface to cancer SMR data for the city of Glasgow. A spatial trend peaking around the city centre is revealed.

4.2.2.3. Identifying global spatial patterns

As discussed in Section 3.1.1.2, detecting the spatial patterning of disease events at a global scale may give some insights into the disease's causal mechanism. As in Section 2.1.1, a global spatial pattern is expressed in terms of spatial association or dependence with respect to the whole data set. Spatial dependence takes two forms: spatial autocorrelation and spatial concentration.

Mapping spatial data may provide visual evidence of spatial patterning. However, the visual evidence often needs to be tested statistically. The Moran's I test (Cliff and Ord 1981) is one of several well-known statistical tests of spatial autocorrelation. The Getis-Ord statistic is well known for testing spatial concentration (Getis and Ord 1992). Moran's I is defined as follows.

$$I = \frac{n\sum_{i}^{n}\sum_{j}^{n} w_{ij} z_i z_j}{S_0 \sum_{i}^{n} z_i^2}, S_0 = \sum_{i}\sum_{j} w_{ij} \qquad (4.2.2)$$

where z_i is the standardised value of x_i by subtracting the mean of $\{x_i\}$ and dividing by its standard deviation. w_{ij} is the element of a connectivity matrix W defining the connectivity from area i to area j. n is the number of areas.

The distribution of Moran's I can be established using a random permutation of all z_i on n areas or approximated to normal if x_i is an observation of a normally distributed random variable X_i at every area and X_i is independent of X_j (Cliff and Ord 1981). Therefore, the significance of the Moran's I can be tested. A positive and significant z-value of Moran's I indicates positive spatial autocorrelation. Bailey and Gatrell (1995, pp. 282) explain the difference in hypothesis involved in the tests. The randomisation test assumes that $\{x_i\}$ forms a population rather than a realisation of a process as the approximate test does. The randomisation test is, therefore, appropriate for testing spatial autocorrelation of data such as election results. On the other hand, the approximate test is appropriate for testing spatial autocorrelation of, for example, disease rates.

Caution needs to be taken when applying the Moran's I test to disease rates or ratios where the size of the underlying population varies. This is because the original version of the Moran's I test under the approximate normal distribution assumes homoskedasticity of data (Walter 1992a and 1992b). Alternative tests that could overcome this problem have been suggested (Waldhor 1996, Oden 1997).

The Getis-Ord G statistic is defined by (4.2.3).

$$G(d) = \frac{\sum_i \sum_j w_{ij}(d) x_i x_j}{\sum_i \sum_j x_i x_j}, \ i{=}j \tag{4.2.3}$$

where x_i denotes the ith value of the variable, $w_{ij}(d)$ is 1 if the distance between area i and area j is not greater than d and 0 otherwise. The distance d is often a metric measurement but might be a non-metric measurement such as lag order (Getis and Ord 1996).

Ord and Getis (1995) derive the moments of G(d) statistics under a randomisation hypothesis. Let z be the z-score of G(d) with respect to its mean and variance, and then z approximates to normality. A positive or negative and significant z indicates the spatial concentration of large or small x_i.

4.2.2.4. Identifying spatial clusters and spatial outliers

Section 3.1.1 and Section 3.1.2 discussed the importance of identifying clusters and spatial outliers in both spatial epidemiological and health research studies. Techniques for identifying clusters in area-based data sets have been developed recently and are attracting much attention in the research community. Anselin (1995) devised a method to compute indicators, called the local Moran's I indicator, for testing for local spatial autocorrelation.

Local Moran's I is defined as:

$$I_i = (z_i / m_2) \sum_{j=1}^{N} w_{ij} z_j ; i = 1,2,...N. \tag{4.2.4}$$

where z_i is the standardised value of x_i by subtracting the mean and dividing by its standard deviation. w_{ij} is the element of a connectivity matrix W. m2 is a constant equal to 1 when W is row-standardised. The sum of I_i over all areas is equal to Moran's I.

Anselin (1995) derived the moments for I_i under a randomisation hypothesis and noted that the test may be affected if global spatial autocorrelation is present. A positive and significant I_i indicates a spatial cluster of similar values. A negative and significant I_i indicates a spatial cluster of dissimilar values. Since I_i is the *i*th component of global Moran's I, it also indicates the local instability contributing to global Moran's I.

Getis and Ord (1992) developed two local indicators of spatial concentration called G_i and G_i*. They can be used to detect clusters of either large or small values. Suppose that x_i is the value for the ith area and d is a measurement of distance, G_i and G^*_i are given as follows:

$$G_i(d) = \frac{\sum_{j=1}^{N} w_{ij}(d)x_j}{\sum_{j=1}^{N} x_j}, j \neq i; and\ G_i^*(d) = \frac{\sum_{j=1}^{N} w_{ij}(d)x_j}{\sum_{j=1}^{N} x_j} \qquad (4.2.5)$$

$$w_{ij}(d) = \begin{cases} 1, \text{if the distance between the } i\text{th and } j\text{th areas is less than } d; \\ 0, \text{if otherwise;} \end{cases}$$

Ord and Getis (1995) derived the moments of the two indicators. Let z_i be the z-score of either $G_i(d)$ or G^*_i (d) with respect to its mean and variance, then z_i may be interpreted as follows. A large positive and significant z_i implies a cluster of large values of x_i (above the mean x_i) within the distance d of area i. On the other hand, a large negative and significant z_i indicates a cluster of small values of x_i within the distance d of area i. As pointed out by the authors, these two tests should be used with caution when global spatial concentration is absent.

One important aspect in health research is to identify those areas that have exceptionally high rates of health events with respect to their neighbouring areas. A technique for this purpose is to compare each value of a variable with the weighted average of its neighbour values of the same variable. Suppose that x_i, (i = 1, 2,...N) is a value of a variable for the *i*th area, then the weighted average is:

$$\overline{x}_i = \sum_{j=1}^{N} w_{ij}x_j / \overline{w}_i; \overline{w}_i = \sum_{j=1}^{N} w_{ij}; \qquad (4.2.6)$$

where w_{ij} is an element in a connectivity matrix. Haining (1993) suggests fitting a bivariate linear regression of x_i (Y axis) on $\overline{x}_i$ (X axis) as a simple exploratory method. The cases, which have standardised residuals over 3.0, may be regarded as outliers. If ($\overline{x}_i$, xi) (i = 1, 2, ..., N) are drawn with the regression line, referred to as an XW plot, and linked with areas in a map, the outliers can be picked up from the plot and signalled on the map simultaneously.

4.3. Analysing multivariate data

4.3.1. Correlation analysis

As discussed in Section 3.1.1, ecological studies require techniques to examine the correlation between disease events (response variable) and socio-economic and/or environmental factors (explanatory variables). The analyst wishes to find explanatory variables strongly correlated with the response variable since a strong correlation suggests association. The analyst may be interested in correlation of events between one disease and another. In this case, a strong correlation may suggest that both diseases might be "caused" by the same set of factors.

ESDA and CSDA techniques are available for examining bivariate relationships. A simple ESDA technique is to draw a scatter plot of data of the two variables. A bivariate linear regression model can be fitted to two variables to quantify their relationships, and embedded into the scatter plot. This technique can be easily extended to examine the relationships for many pairs of variables simultaneously by arranging the plots in a matrix form (Tukey 1978, Cleveland and McGill 1988). A matrix plot may also be useful for identifying those explanatory variables that are collinear with each other.

The Pearson product moment correlation coefficient, $\hat{r}$ and the Spearman rank correlation coefficient, $\hat{r}_s$ are widely used in measuring bivariate correlation. These two coefficients were devised primarily for non-spatial data and do not take into account the locations of the observations. Although the coefficients themselves are not affected even when the observations are spatially correlated, the significance tests of them, which assume spatial independence, may be affected. For two variables, each of which is spatially positively correlated, the sampling variance of $\hat{r}$ is underestimated (Haining 1993, p314). Clifford *et al* (1985) suggest obtaining an adjusted value N for n, an equivalent number of independent observations, in order to assess the significance of correlation coefficients. The use of these techniques may require the data to be de-trended in advance since they apply only to stationary correlated data with constant variance. Haining (1991) calls for caution in using conventional procedures to test for the significance of correlation coefficients.

4.3.2. Regression modelling

Health research also requires techniques for quantifying the relationships between a response variable Y and a set of explanatory variables X_i (i = 1,2,...p-1). Regression modelling techniques are often used for this purpose. There are three major stages involved in regression modelling: model specification, model estimation and model criticism (Haining 1994). First, a model or a function is chosen for the data and assumptions on the model are made. Then an appropriate method for fitting the model is selected and the model parameters are estimated. Finally, the model is evaluated against the statistical assumptions, and is diagnosed on the extent to which the chosen model fits the data and the sensitivities or stability of results to perturbations of the data. The three stages may be iterated many times to fulfil the following objectives (Haining 1993 p 330): 1) identifying a model for the data; 2) obtaining good estimates of regression coefficients; and 3) providing an equation for prediction.

4.3.2.1. Ordinary linear regression model

One of the most frequently used models is an ordinary linear regression model defined as:

$$\begin{aligned} y &= X\beta + \varepsilon \\ E[\varepsilon] &= 0; E[\varepsilon\varepsilon'] = \sigma^2 I \end{aligned} \tag{4.3.1}$$

where y is the n×1 vector of the dependent or response variable, X is the n×p matrix with n cases (areas) on p-1 explanatory variables and a constant term. β is a vector of p regression coefficients. Least squares (LS) is often used to fit model (4.3.1) to data and make inferences on the goodness-of-fit of the model and model parameters.

A number of problems may arise in using model (4.3.1) and can be grouped into two categories - those resulting from the nature of the data and those from the failure to satisfy the assumptions of least squares fitting (Weisberg 1985, Haining 1990 pp. 331-2). Problems of the first category may include incomplete and inaccurate data, and both spatial and non-spatial outliers in response variables and leverage effects associated with explanatory variables. Multicollinearity in explanatory variables is another problem. These problems may have influences not only on fitting the model but also on making inferences.

It is also of importance to identify problems of the second category. Model mis-specification is one of the main reasons why statistical assumptions may be violated. For a non-infectious rare disease, the number of observed events in area i, O_i, is approximately Poisson, so is O_i / E_i. If model (4.3.1) is fitted to O_i / E_i as the response variable and a set of explanatory variables which may act multiplicatively or exponentially rather than additively, the assumption of normally distributed errors may not hold. In this case, a log transformation may be applied to O_i / E_i prior to modelling (Pocock *et al* 1981, Haining 1994), while a generalised linear regression model with Poisson errors may be a good alternative (McCullagh and Nelder 1989, Bailey and Gatrell 1995).

In modelling area-based data using model (4.3.1), the assumption of homoskedasticity of errors may not hold if data is drawn from an unevenly distributed population. Anselin and Can (1986) noted this problem in the fitting of urban density functions. Pocock *et al* (1980) illustrated this problem in the context of modelling cardiovascular mortality rates across a set of areas on which the population varies. In this case, model (4.3.1) may be modified by allowing for non-constant variance, that is $E(\varepsilon) = \sigma^2 D$, where D is diagonal matrix with non-constant diagonal elements. The diagonal may be set inversely proportional to the size of population or the number of observed events (Pocock *et al* 1981, Haining 1993). This model can be fitted using weighted least squares or an intermediately weighted procedure (Pocock *et al* 1981). The authors considered the selection of these procedures. Haining (1991) reviewed these issues and compared regression results using three different fitting procedures in a study of intra-urban mortality data for the City of Glasgow, UK.

Model (4.3.1) assumes constant coefficients for all cases. This assumption might not be held if heterogeneity in explanatory variables were substantial. In this case, model residuals may also be spatially correlated. If an important explanatory variable spatially correlated with the response variable is not included in the model, the model residuals are likely to be spatially correlated. Different spatial mechanisms operating between areas may give rise to different forms of spatial interactions. Failure to take them into account may lead to spatially dependent residuals. Thus, model (4.3.1) may have to be modified.

4.3.2.2. Models for spatially heterogeneous data

Spatial heterogeneity in health data may arise as a result of people in different areas responding to the same socio-economic and environmental factors but to different degrees. The poverty in declining industrial city centres may be far more intensive than in rural areas (Jones and Moon 1987). On the other hand, some rural areas may be exposed to specific environmental problems not found in the cities. Facilities which are potentially dangerous to the public, such as nuclear power stations, are often installed in rural areas.

Model (4.3.1) can be modified for spatially heterogeneous data by allowing coefficients to vary across areas. If spatial heterogeneity is expected in two sub-regions of in a larger study region, a dummy variable, which takes a value of 0 and 1 for areas in the two different sub-regions respectively, can be included as an explanatory variable.

$$\begin{aligned} &y = X\beta + \alpha\delta + \varepsilon \\ &E[\varepsilon] = 0; E[\varepsilon\varepsilon'] = \sigma^2 I \end{aligned} \qquad (4.3.2)$$

where δ is a dummy variable and α is a coefficient. This model only induces the variation in the intercept but can be modified to allow for variations in slope coefficients (Haining 1993, pp. 339).

4.3.2.3. Model with spatially correlated errors

Residual correlation may be a consequence of excluding an important and autocorrelated explanatory variable. In a study of cancer, this situation is highly likely to occur. Analysts are unlikely to know all the factors which are important and should therefore be included in the model. Even though they might try out all known factors, there may still be other unknown factors. Moreover, data on some known factors may not be available to a study. Bailey and Gatrell (1995) show that model parameters can be altered after spatial autocorrelation in the error terms is modelled in the form $E[\varepsilon\varepsilon'] = \sigma^2 V$ where the variance-covariance matrix V is a non-diagonal matrix.

The form of V can be specified in different ways. In modelling heart disease and water hardness, Cook and Pocock (1983) specify the covariance of the error terms as a two-parameter exponential function of distance separating the observations where the functional form is derived by analysing the residuals of an ordinary regression model. More commonly, a form of V is specified indirectly by various interaction schemes. Model (4.3.3) is one of the interaction models:

$$\begin{aligned} &y = X\beta + u \\ &u = \rho Wu + \varepsilon \\ &E[\varepsilon] = 0 \text{ and } E[\varepsilon\varepsilon'] = \sigma^2 I \end{aligned} \tag{4.3.3}$$

where W is a connectivity matrix specifying the interaction term for the errors and λ is the autoregressive coefficient. Thus, V takes the form: $V = \sigma^2[(I - \rho W)'(I - \rho W)]^{-1}$.

Least squares can still be used to yield unbiased estimates for parameters. However, these estimates are inefficient because of the non-diagonal structure in the variance matrix and the autoregressive estimator is inconsistent (Anselin 1988). Ord (1975) discusses maximum likelihood estimation and inference for this model. Anselin (1989, pp181) describes an algorithm for parameter estimation.

4.3.2.4. Models for spatial interactions

Spatial interactions need to be considered in modelling health-related data. People may be exposed to risk factors in areas where they live and may also be exposed to risk factors in adjacent areas when they travel. For an infectious disease, the disease agent is passed from one person to another through contact. If the people who are close to each other geographically have more contacts than those who are far away, disease rates in one area are likely to be influenced by rates arising in adjacent areas (Cliff *et al* 1981). Model (4.3.1) may be adopted to take into account these two types of interactions.

A model for the first type of interaction may be specified as below:

$$\begin{aligned} &y = X\beta + WX_s\beta_s + \varepsilon \\ &E[\varepsilon] = 0 \text{ and } E[\varepsilon\varepsilon'] = \sigma^2 I \end{aligned} \qquad (4.3.4)$$

where y, X and β are as the same as in model (4.3.1). W is a matrix specifying a spatial interaction form for a subset of explanatory variables denoted as X_s, and β_s are the coefficients corresponding to X_s. This model may be appropriate for modelling the relationships between the incidence of a respiratory disease, like asthma, and the level of air pollution.

Model (4.3.4) can be fitted by least squares because the term WXs can be considered as another set of explanatory variables. However, care must be taken since X and WX may be co-linear. Model (4.3.4) is sometimes called a linear regression model with lagged explanatory variables.

A model for the second type of interaction may be specified as below:

$$\begin{aligned} &y = \rho Wy + X\beta + \varepsilon \\ &E[\varepsilon] = 0 \text{ and } E[\varepsilon\varepsilon'] = \sigma^2 I \end{aligned} \qquad (4.3.5)$$

where y is the dependent variable (nx1 vector), W is a matrix specifying an interaction form for the response variable, and ρ is the autoregressive coefficient. X and β are as the same as in model (4.3.1). This model may be appropriate for modelling contagious diseases such as flu, which is passed by human contact but also conditional on factors like the health status of the population.

Least squares is no longer valid for fitting this model since it yields biased and inconsistent parameter estimates (Whittle 1954, Mead 1967). Ord (1975) discusses a maximum likelihood procedure for fitting the model and making inference. Anselin (1989, pp.183) describes an algorithm for parameter estimation.

4.3.2.5. Model evaluation

After the data is fitted to a model, the model residuals need to be checked for the purpose of model evaluation. Both ESDA techniques and statistical tests have their roles to play. Residual normality may be checked using a rankit plot, which is actually a scatter of two variables - residuals and the normal scores (Cleveland and McGill 1988). The Shaprio W test can be applied to the residuals for the same purpose. For the models above, heteroscedasticity in the model residuals can be checked using plots. If the population varies across areas, a scatter plot of the model residual against the population for each area allows a visual assessment to be made (Haining 1993). Non-constant residual variance may be detected by plotting the dependent variable against each explanatory variable (Weisberg 1985). Anselin (1988) develops statistical tests for heteroscedasticity that are valid even when the spatial dependence in error terms is present, and these can be applied to all models discussed above.

Mapping residuals is a powerful tool for assessing residual spatial dependence. ESDA techniques discussed in Section 4.2 for identifying spatial properties of a single variable can be employed without change. Moran's I defined by equation (4.2.2) can still be used to test the spatial autocorrelation for the residuals of model (4.3.1). However, the mean and the variance under the null hypothesis of no spatial autocorrelation are no longer valid owing to a reduction of p degrees of freedom, where p is the number of the explanatory variables involved in the model. The asymptotic distribution of Moran's I for model residuals was developed by Cliff and Ord (1981).

Whether spatial interactions in either errors or the response variable should be included can be checked against model (4.3.3) or (4.3.5) respectively. Clearly, model (4.3.1) is a restricted form of either model (4.3.3) or (4.3.5) with λ or ρ set to zero respectively. The Likelihood Ratio (LR) test and the Lagrange Multiplier (LM) can be constructed to test the null hypothesis $\lambda = 0$ or $\rho=0$ to discriminate between models (4.3.1) and (4.3.3) or models (4.3.1) and (4.3.5). Both tests are asymptotically distributed as χ^2 with p degrees of freedom, where p is the number of parameters in the restricted model (Anselin 1988, pp. 67). The advantage of the LM test is that it does not require fitting the restricted model whereas the LR test does. A detailed discussion of these tests and others can be found in Anselin (1988 pp. 66-69). The author also discussed the LR and LM for a wider range of models.

The goodness-of-fit of different models to the same data may be assessed by comparing three values: the maximum of the log likelihood (ML), Akaike's information Criterion (AIC) and the Schwartz Criterion (SC) (Akaike 1981, Anselin 1988). An absolute increase in these values from one model to another indicates a better fit.

The aim of the model evaluation is not only to assess whether statistical assumptions are satisfied in order to make statistical inferences but also to seek a better model in the light of model evaluations. This point is particularly relevant in the analysis of health-related data given that the analyst usually has little knowledge about underlying processes that yield the data in advance.

4.3.2.6. Model diagnosis

Having fitted a model to a set of data, model diagnostics need to be performed to assess how well the chosen model fits the data with respect to the discrepancy between the data and the model. One aspect of this analysis is to identify outliers in model residuals. Outliers may be extreme residuals in both the non-spatial and spatial senses. The identification of outliers in model residuals can be useful in health research. If the purpose of a study is to search for possible causal factors, further and more detailed investigations may focus on the outliers to seek causal factors beyond what the model suggests. If the purpose of a study is to target health resources, these areas may be considered with high priority.

Two well-known diagnostics are standardised residuals and measures of leverage. The former is used to detect outliers which are extreme cases in the response variable, while the latter is used to detect extreme cases in one or more explanatory variables (Haining 1994, pp. 326). For model (4.3.1), the *i*th standardised residual is defined as (Cook and Weisberg 1982):

$$\hat{s}_i = (y_i - \hat{y}_i) / \hat{\sigma}^2 (1 - h_{ii})^{(1/2)}$$

where y_i and $\hat{y}_i$ are the *i*th observed value in the response variable and its predictor from model (4.3.1). $\hat{\sigma}$ is the estimate of σ by least squares; h_{ii} is the *i*th diagonal element in the matrix $X(X'X)^{-1}X'$. h_{ii} is the *i*th measure of leverage or potential, indicating the influence of x_i on $\hat{y}_i$.

The standardised residuals follow Student's t distribution with a common variance and n-p degree of freedom, while $h_{ii} \geq (1/n)$ and $\sum_{1,\dots,n} h_{ii} = p$. A standardised residual, greater than 3.0 or less than -3.0, may be considered large, while if $h_{ii} \geq 3(p/n)$ the leverage value is considered large. Martin (1992) derives the standardised residuals and measures of leverage for model (4.3.3) where parameter estimation is based on generalised least squares.

A second theme of model diagnosis is to assess the stability of the results to perturbations of the problem. One kind of perturbation is the exclusion of one or more cases. Analysis of this kind is called influence analysis (Cook and Weisberg 1982). An important aspect of this analysis is to find the cases or areas that are the most influential on the estimates of model parameters.

Cook's Distance statistic, D_i, is a measure of influence on the estimates of β coefficients after the *i*th case is excluded from model (4.3.1), and is defined below.

$$D_i = (\hat{\beta}_i - \hat{\beta})' X' X (\hat{\beta}_i - \hat{\beta}) / (p\sigma^2)$$

where $\hat{\beta}_i$ and $\hat{\beta}$ are the estimates of coefficients corresponding to explanatory variables with and without the *i*th case removed. Case i is considered to be influential if $D_i > F[0.5; p, (n\text{-}p)]$ (Cook and Weisberg 1982). Martin (1992) derives the Cook's Distance statistic for model (4.3.3) where parameter estimation is based on generalised least squares. It should be noted that if the covariance is specified using a connectivity matrix, excluding an area may require redefinition of the connectivity matrix so that those areas, which were previously connected with the deleted area, can be re-coupled. Haining (1994) reviewes the diagnostics discussed above and illustrates the use of them in highlighting those cases that might influence model estimation. A procedure is described for diagnosing many influential cases simultaneously.

4.4. Summary

The theme of this chapter has been to discuss some lattice data analysis techniques that together enable multidisciplinary users to perform a coherent analysis of area-based health-related data in order to answer those health questions considered in Chapter 3. The techniques shown in Section 4.1 are useful for the construction of areal frameworks and the specification of inter-area relationships. The regionalisation technique proposed by Wise *et al* (1997) and summarised in Section 4.1.1.3 can take three types of criteria into account in regionalisation and may be used for exploratory data analysis. Section 4.1.4 summarises some basic approaches to the specification of inter-area relationships required in the analysis of area-based health-related data.

Techniques in Section 4.2 are useful for the identification of both non-spatial and spatial properties of data of a single variable. The EDA techniques, such as the box plot, can be used not only to explore the distributional properties of the data but also to answer questions such as "where is it...?" and "what is at..." when they are 'hot' linked with a map. Techniques for estimating relative risk rates are fundamental for analysing spatial variations of disease events. Techniques, like the median smoother, the lagged box plot, the XW plot, the Moran's I and Getis-Ord G(d) tests and the local Moran's I and Gi* and Gi tests, can be used to identify spatial properties, while the trend surface model can be used to model spatial variation of the data.

Techniques in Section 4.3 are useful for assessing the correlation of pairs of variables and modelling the relationships between the response variable and explanatory variables. The main ESDA techniques are those based on the scatter plot of a pair of variables, while the main CSDA techniques are based on a range of regression analysis and modelling techniques. Techniques for assessing and diagnosing model fit are also important.

Throughout the discussion in this chapter, the technical aspects of these techniques have been discussed in some detail. This clearly shows what is required in order to implement them. The next chapter will discuss the implementation of these and other techniques in SAGE.

CHAPTER 5. SAGE - SPATIAL ANALYSIS IN A GIS ENVIRONMENT

The previous chapter discussed SSA techniques required for analysing area-based health-related data. This chapter is concerned with the integration of those SSA techniques with a GIS - ARC/INFO. This integrated system is called SAGE and enables the user to apply the techniques in studies in an interactive manner.

The discussion proceeds from both a system integrator perspective and a user perspective by considering how SAGE has been developed and how the user may use it respectively. There are seven sections in the chapter. The first four sections consider some important aspects involved in the development of SAGE at stages: system analysis, modelling, design, and implementation and testing. This is followed by a criticism of the integration. The sixth section shows how the user may use SAGE to analysis area-based data by briefly illustrating some of key SAGE functions, followed by conclusions.

As an integral part of this thesis, SAGE is available in the forms of source code and executable code (for Sun Solaris 2.5 or higher) from either the enclosed CDROM or the ftp site: ftp://ftp.shef.ac.uk/pub/uni/academic/D-H/g/sage/. Information on installation of SAGE and related documents can be found in the packages.

5.1. System Analysis

System analysis is concerned with the identification of functional and non-functional requirements for SAGE. First, we summarise those required system functions identified in Chapters 3 and 4. Second, we discuss why ARC/INFO was chosen for the integration and which functions ARC/INFO does and does not support for implementing SAGE. Finally, we consider some non-functional requirements for SAGE on its behaviour and deployment.

5.1.1. Functional requirements

Following the discussions in the previous two chapters, the system functions fall into three groups: data visualisation (DV), data analysis and modelling (DAM), and data management (DM). Functions in each group may be decomposed further into subgroups. Figure 5.1 shows a decomposition of the functions.

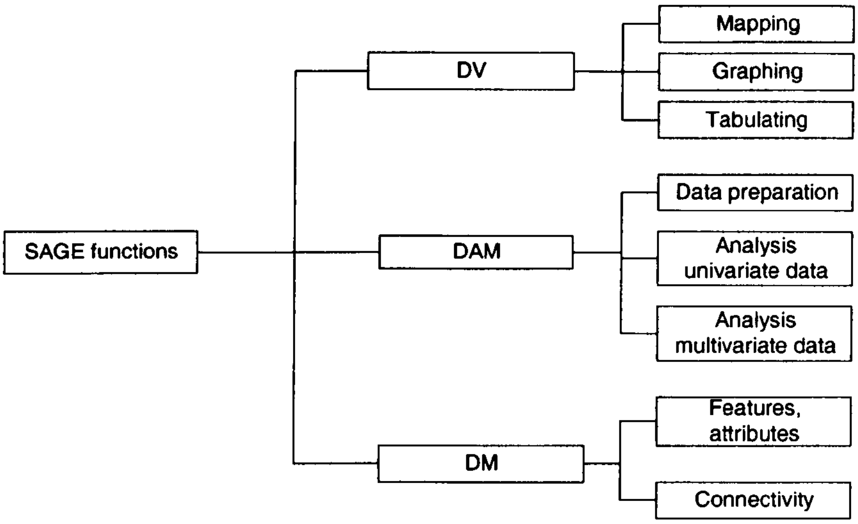


Figure 5.1. A decomposition of the system functions.

5.1.2. Data visualisation (DV)

To facilitate exploratory statistical data analysis, the system must support the visualisation of spatial data in three forms - area maps, attribute tables and statistical plots. A display in one of these is termed a view. Besides area maps and statistical plots, the roles of which in SSA have been discussed in the previous chapters, a table view is also essential for SSA given that attribute values are managed as tables. A table view contains N rows and M columns, corresponding to areas and attributes respectively. Objects that together make a view, such as areas in the map, points or lines in a plot, cells in the table, are referred to as view objects. Table 5.1 lists statistical plots to be supported by SAGE. Note that for the purpose of comparing statistical plots, plots of the same type may be drawn in the same view. The last column in the table indicates whether a single plot (s) or multiple plots (m) of each type can be drawn into the same view.

Type	Description	Purposes	s/m
Histogram	A set of bars representing the frequency of variable values.	Examine the distribution of a single variable.	s
Box plot	A graphical representation of resistant measures – medium, upper, lower quartiles and outliers.	Examine the distribution of a single variable and identify outliers.	m
Lagged box plot	A set of box plots, each of which is constructed using data from all areas at a certain lag away from a selected area.	Explore the trend of data in relation to a specific area.	s
XY plot	A plot of two variables in 2D.	Examine the relationships between two variables.	m
Rankit plot	A special type of XY plot.	Examine the normality of a variable by checking its values against the normal scores.	m
XW plot	A special plot of two variables; The X variable is constructed by, for each area, averaging the values of the X variable in the areas adjacent to that area defined by a W matrix.	Examine spatial auto-correlation and identify spatial outliers.	m
Matrix plot	A set of XY plots arranged as a matrix.	Examine bivariate relationships between pairs of many variables.	s

Table 5.1. A list of statistical plots to be supported.

The system should allow the user to perform some operations on each view, such as changing object properties like shading colours and patterns in a map or plot view. It needs to allow the user to query view objects by selecting them interactively. More important, the system must be able somehow to maintain "hot" links among views so that a selection of view objects in a view can be reflected in another view by highlighting the corresponding view objects. Each view should be automatically updated whenever the values used to construct it are changed. This dynamic feature is essential to explore different properties of spatial data simultaneously.

5.1.2.1. Data analysis and modelling (DAM)

As discussed in Chapter 4, the system needs to support data preparation, spatial statistical analysis and modelling of a single variable data set or multivariable data sets. These functions require accessing spatial feature attributes (e.g. distance, perimeters etc.) and attributes, connectivity data (W matrices), or all of these at the same time, and may generate results in the form of either an attribute (having a value for each area) or a set of numeric values. Some functions should be applicable to a selected subset of the whole data set. Table 5.2 summarises data analysis and modelling functions. The last column shows whether a function is expected to work only on the whole data set (w), or part of it (p) as well.

Category	functions	whole/part
Data preparation	1) Construct regional frameworks and perform classification; 2) Specify inter-area relationships, i.e. W matrices; 3) Create new attributes.	w w & p w
Analysis of univariate data	1) Describe non-spatial properties of the data – such as mean, standard deviation, median, the upper and lower quartiles, and outliers; 2) Describe spatial properties of the data –mapping, detecting trends, global and localised spatial associations and spatial outliers.	w & p w & p
Analysis of multivariate data	1) Perform bivariate correlation analysis; 2) Fit regression models to spatial data, and perform model evaluations and diagnoses.	w & p w & p

Table 5.2. Summary of DAM functions.

5.1.2.2. Data management (DM)

Data management refers to a set of supporting functions for the DV and DAM functions. Table 5.3 summarises the data management functions.

Category	functions
Feature and attribute data management	1) manage attribute and spatial feature data for data sets; 2) insert and remove attributes into and from data sets and retrieve attribute values; 3) relate attributes managed by other systems to data sets; 4) create new data sets; 5) manage W matrices; and 6) query data sets on its attributes, spatial features and adjacency.

Table 5.3. A list of data management functions.

5.1.3. Choosing a GIS for integration

The discussion so far has clearly indicated that a GIS appropriate for the integration should at least meet the following criteria:

1. It should support a vector-based data model, spatial processing functions such as dissolving polygons and cartographic mapping;

2. It should allow topological relationships between areas to be derived for specifying inter-relationships between areas;
3. It should provide a set of API or a programming language so that the access to any of its functions can be realised through API or programs written in that language. This is important if linking operations can be implemented effectively.

At the time the integration took place, there were a number of GIS packages that met the criteria above[2]. They could be divided into two groups – desktop mapping packages and spatial data process packages. MapInfo (MapInfo http://www.mapinfo.com) and ArcView (ESRI 1994) were two in the first groups available. MapInfo and ArcView have been available for PCs with Microsoft Windows, while ArcView is also available on the Unix platforms. Both packages were designed originally as mapping products. One common feature of both packages was that each provides an interpreting programming language, called MapBasic and Avenue respectively. With these languages, the developer could develop new functions.

These two packages have, however, the following drawbacks. First, they both support quite limited spatial data processing functions other than those for mapping. For example, they do not support such operations as converting a point set to an area-based data set. Second, because MapBasic and Avenue are interpreting languages, they are not ideal for programming computationally intensive SSA functions. Although MS Windows provides mechanisms allowing different programs to exchange data each other, these mechanisms were not fully supported by both packages at that time. Therefore, although SSA techniques could be implemented in low-level languages such as C and C++ to achieve high efficiency, it would be difficult to make them communicate with either of the packages at the time. This limitation, however, was removed in the later versions of both packages for Win32-based systems such as Windows 95/98 and NT. Third, both packages support only vector-based non-topological models. Therefore, it would require more computation to derive topological relationships

[2] The SAGE design started in the late of 1994.

between areas with either of the two packages than with another package that supports a vector-based topological model.

One of the best-known GIS packages in the second group is ARC/INFO with strong capabilities in spatial data acquisition, spatial data processing, spatial data management, and cartographic mapping. Basically, it consists of a number of subsystems including ARC, INFO and ARCPLOT. It supports a vector-based topological data model suitable for area-based data and provides a scripting programming language AML (ARC/INFO Macro Language). With it, one can write programs that call ARC/INFO functions (ISER 1994) and design graphic user interfaces.

Unlike MapInfo and ArcView, ARC/INFO was designed originally as a command-oriented GIS to be used on mainframe computers rather than desktop computers. AML was developed later to meet the needs for developing GUIs. Therefore, AML is loosely associated with other ARC/INFO modules. But this drawback could also be an advantage if ARC/INFO is mainly run in the background. Furthermore, it did not provide API for programming languages such as C or Fortran. This drawback was not removed until the arrival of the ARC/INFO version 7.2 for MS Windows (ESRI).

From the integration point of view, any of the three packages might be selected for integration. In this research ARC/INFO (version 7.0) was chosen for the integration for the following reasons. First, ARC/INFO has much stronger spatial data-handling capabilities than the other packages. This is still true today although MapInfo and ArcView have been extended to provide more spatial data handling functions. Second, ARC/INFO was selected as a primary GIS for UK Higher Institutions of Education (Wise 1990). A system developed around it is likely to be low cost. Third, at that time ARC/INFO was widely used not only in academia but also in GIS related industry. Therefore, it would be an ideal platform to promote the SSA techniques.

5.1.4. ARC/INFO - what functions does it support and does it not support?

ARC/INFO supports many functions useful for linking the SSA techniques for the analysis of area-based data. Underpinning these functions is the ARC/INFO vector-based topological model. The basis of the data model is the section of line running between two nodes - the points where lines meet. With this model, the physical locations of nodes and the identity of the start and end nodes and of the polygons on either side of the line are stored to capture the spatial structure of the set of lines. These two types of data are held in a set of related files referred to as a coverage. ARC/INFO stores the topological data for an area coverage in an INFO table, called the Arc Attribute Table, containing the from-node, to-node, left-polygon, right-polygon information for each line section. ARC/INFO provides functions able to manage area coverage correctly and efficiently. Also the way in which the topological information is managed allows the construction of adjacency matrices required by many SSA techniques.

Besides the functions for managing polygon coverages, ARC/INFO supports a rich set of functions for manipulating data sets. For example, a point data set can be converted to a polygon coverage. A new coverage can be constructed from a polygon coverage by polygon dissolving. ARC/INFO allows attributes in external databases to be related to its polygon coverage. It also supports a set of visualisation functions through one of its subsystems, ARCPLOT. These not only enable the generation of rich-featured overlay maps but also make it possible to query a data set according to attribute, location or their combination.

AML is a scripting language particularly designed for ARC/INFO. With AML, new functions can be constructed based on existing ARC/INFO functions and functions of a host computer environment. Simple GUIs can easily be built with it. A large number of useful AML programs exist as additional modules for ARC/INFO. ARCTOOLS is such a module written in an objected-oriented fashion, providing the user with easy to use GUIs and functions to access ARC/INFO. AML could be very useful for implementing linking operations (See

Section 2.32).

One limitation of ARC/INFO is that it does not provide API for other programs or packages to access its internal data structures and its functions. However, it does allow both attribute and spatial feature data to be exported to ASCII files and attribute data contained in ASCII files to be imported into INFO tables. These operations can be fulfilled quite easily using AML. Another major limitation is that, although the topological information can be obtained through data queries, ARC/INFO does not provide an efficient way to locate areas that are adjacent to each other at several lags away. This usually involves intensive computation. ARC/INFO does not provide functions for specifying and managing W matrices.

There are many other functions that either cannot be performed in ARC/INFO or could be performed better outside ARC/INFO. ARC/INFO does not support visualisation functions for drawing statistical plots directly although it does provide a set of drawing primitives. However, it would be very costly to utilise those functions using those primitives and to achieve the same level of sophistication what has been already achieved by many graphic packages. ARC/INFO lacks functions for displaying attribute data in a tabular form and does not support data analysis and modelling functions summarised in Table 5.2. Being a high level scripting language, AML is not suitable for developing many of them owing to the intensive computation needed. Indeed, there is little point in even considering doing this since many functions are either already implemented or can be easily implemented using many numerical packages such as NAG subroutines (NAG).

Based on the discussion above, the system functions can be divided into two components comprising those that can be performed in ARC/INFO or those which cannot be performed and would be better performed outside ARC/INFO. These two components are called the ARC/INFO component and the SSA component respectively for convenience.

Table 5.4 summarises the classes of the functions discussed above and indicates which component functions belong to.

Function		Component
DV	Mapping	ARC
	Graphing	SSA
	Tabulating	SSA
	Hot linking	SSA
DAM	Data preparation	SSA
	Analysis of univariate data	SSA
	Analysis of multivariate data	SSA
DM	Spatial features and attributes	ARC
	Connectivity	SSA

Table 5.4. Functions and the components to which the functions belong.

5.1.5. Non-functional requirements

There are some non-functional requirements concerning system behaviour that also need to be considered. First, one of the most important features of the system to be expected is the seamless link between the two components. That is, data transferring between the two components is done automatically and transparently as far as the user is concerned. Second, GUIs must be intuitive. If they are to be utilised separately, they must be controlled in such a way that the user can be directed to use only appropriate GUIs at any one time without suffering from "shifting" effects (Chou and Ding 1992). Third, it is desirable that the SSA and ARC/INFO components can be configured both on a single computer and on two different computers connected through networks. This is important in the situation where the GIS runs on a central computer managing valuable spatial data and non-spatial data.

5.2. System Modelling

Based on the system analysis, system modelling is intended to identify a model that would be the most appropriate for integrating SSA and ARC/INFO. As discussed in other publications (Haining *et al* 1996, Haining *et al* 1999, Wise *et al* 2000) and the following sub-sections, a client-server model is considered to be appropriate where the SSA component and ARC/INFO functions as a client and a server. Based on this model, another aim of system modelling is to identify a set of 'generic' client requests and server replies necessary for performing the area-based analysis.

5.2.1. A client-server model

As discussed in Section 2.4, research has been examining different integration approaches (Goodchild 1992, Chou and Ding 1992, Nyerges 1992, Abel *et al* 1994, Haining *et al* 1996). From the point of data exchange, the close-coupling and loose-coupling approaches are at the two extremes of this spectrum. With the former approach, data exchanges between components are implicit rather than explicit and facilitated by some kind of internal application interfaces in GIS. With the latter approach, components exchange data in the form of files and often require user intervention. To use the close-coupling, a GIS must provide API (application programming interface) for other components to access its data either directly or through some kind of internal transferring service. ARC/INFO does not provide API or a data transfer service directly for this purpose. Although AML could be used to utilise these to some extent, the degree of the coupling would be reduced. On the other hand, the loose-coupling approach does offer great flexibility for system integration. As discussed in Chapter 2, this approach is likely to result in a system that is less efficient in the sense of data transfer and less user-friendly in terms of user interfaces. Therefore, an alternative approach between these two extremes needs to be explored.

As suggested in Haining *et al* (1996), an intermediate approach based on the client-server model could not only overcome these disadvantages but also meet the system requirements better. In the client-server model (Smith and Guengerich 1994, Umar 1993), a component is considered as a client if it requests the services of other components to complete a certain task, or as a server if it provides services for clients. A component may function as a client at one time but as a server at another time. Under the client-server model, client and server are independent processes and may be run on the same or different computers in the network. The communications between the client and server processes are handled efficiently through a set of well-defined API based on such as Remote Procedure Calls (RPCs) (Simon 1996, p.65-8).

The client-server model is such a natural extension of the concept of modular programming that it recognises the role of each individual component in a system, and its specific requirements for computing resources. A client process is the front-end portion of an application, managing user interactions and data presentations, issuing requests to corresponding servers, and executing application logic accordingly. A client may require computing resources to handle graphical processes and displays, data manipulation and management. On the other hand, a server process is the back-end portion of the application, responding to the client requests by performing tasks and returning results to the client processes. Unlike a client process, a server process would not normally require graphical processing units but instead powerful CPU units, large memory and hard disks for managing and manipulating data.

The ARC/INFO and SSA components may be configured in three ways - client and server, client and server or server and client, and server and client. All but the last one is thought to be the most appropriate. ARC/INFO could function as a client to manage user interactions with GUIs utilised by AML, and invoke the SSA component with the requests through system calls and ARC/INFO inter-application communications (IAC ESRI 1994). However, this would lead to the separation between GUIs for the SSA functions and the SSA component and, therefore, reduce the reusability of the SSA component if SSA techniques are to be linked with another GIS. It would also inevitably increase synchronisation between the client and server processes since most tasks have to be done in the SSA component.

The second configuration allows synchronisation between the client and the server to be reduced to a minimum and is flexible for integrating ARC/INFO and the SSA component. However, this architecture is likely to lead to a system with two separate and uncontrolled GUIs. In order to be able to access functions in a logically correct order, one has to know which GUI he/she should use at a time. This may take the user a considerable amount of time in order to get familiar with the system.

The third configuration is more appropriate than the previous two because it suits the two components naturally in terms of their roles. First, it allows the SSA functions to be accessed more easily by the user without the need to pass the user requests and the results forward and backward as would be needed in the first one. Second it enables the GUI functions corresponding to the SSA functions to be made as part of the SSA component so that they form a self-contained component. This increases likelihood for the SSA component to be linked with another GIS. One implication of this model to integration is that a set of GUIs has to be developed to allow the user to interact with ARC/INFO's database management functions and cartographic functions. This would require quite a considerable amount of work since those functions are often complicated. One way to avoid this problem is to use GUIs that have already been developed for these purposes in ARCTOOLS. This means that two sets of GUIs would exist and therefore have to be controlled so that they can lead the user to access the system functions correctly. How this can be done will become clear in the later discussion.

In order for the SSA component and ARC/INFO to work seamlessly, in Abel and Kirly's terms, linking operations - constructor, accessory, filtering and transformation operations on data (see Section 2.3.1) - must be supported. Since these operations will be system-dependent, for the purpose of maximising the reusability of the system components, it would be better if another separate component can be dedicated to handle them so that the SSA component and ARC/INFO become totally independent of each other and can be updated separately. This component is referred to as the linking agent and transforms the two-tier client-server model described above to a three-tier model.

5.2.2. System architecture

Figure 5.2 shows the SAGE architecture with the three components, the SSA component, the linking agent and ARC/INFO, shown as rectangles. Each component is an independent process to run on either a stand-alone computer or networked computers. Networks are shown as the grey area in the middle. There are two data repositories associated with the SSA client and the ARC/INFO server respectively. The ARC/INFO server uses one to hold INFO tables that contain new attributes created during a session. The client data repository is essentially a cache holding all necessary data, including attribute values and W matrices, during a data analysis session. The W matrices could be used in a later session. The curved lines, originating from and ending on the rectangles, indicate the communications between processes. A display is linked with the client computer. The curved lines originating from and ending at the display indicate the communications between a process and the display where the visualisation and the user interactions take place.

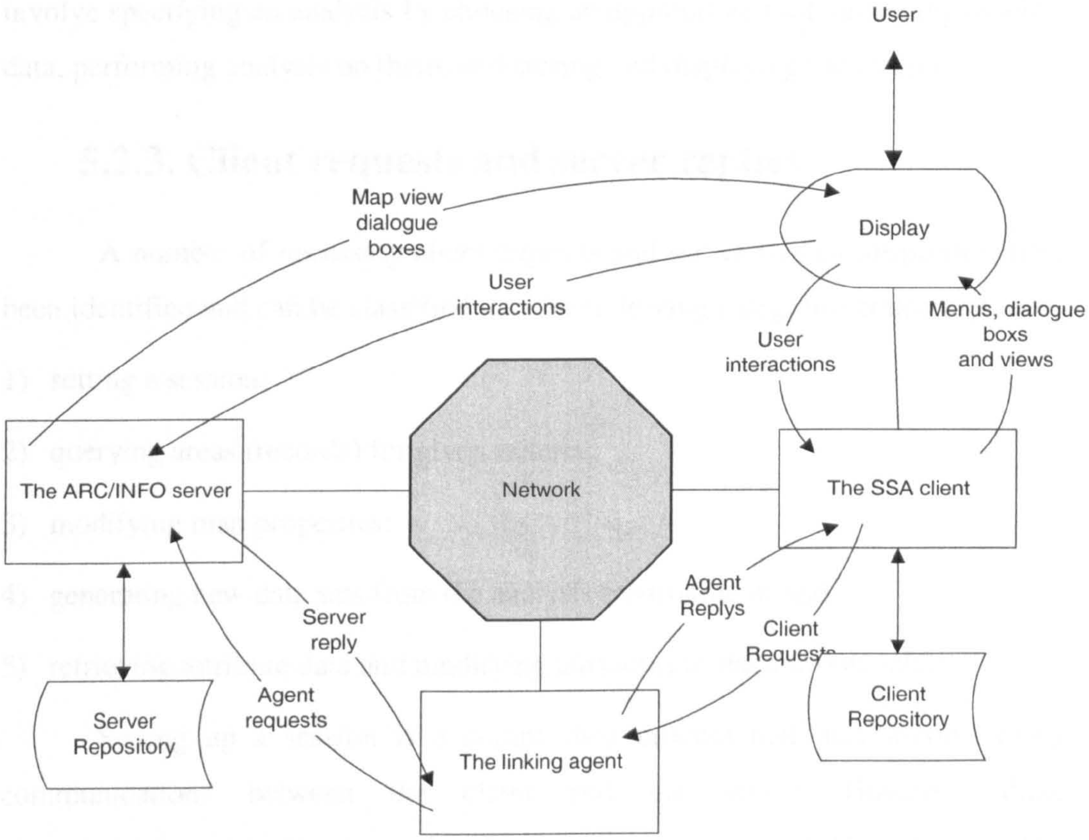


Figure 5.2. System architecture

Obviously, the communications have to be synchronised for the system to work in a logically correct order. Although different communication models may be applied, for simplicity a blocking model is chosen (Simon 1996). That is, the client process is blocked after sending a client request to the agent and the agent is blocked after passing the request to the server. The agent and the client remain blocked until they receive a reply from the server and the agent respectively. Clearly, this communication model guarantees the correctness of the logic between processes at any point in time. However, this model does not allow for any concurrence between the processes to achieve higher system efficiency than would be possible with an asynchronous model.

Under this system, a data analysis session can be viewed to include two main steps - setting up an analysis session (or environment) and performing data analysis in the environment. The former involves querying coverage, relating tables to it, choosing attributes for the analysis, caching all required attributes in the client repository, and constructing W matrices. Performing data analysis may involve specifying an analysis by choosing an appropriate tool, retrieving required data, performing analysis on them; and storing and displaying the results.

5.2.3. Client requests and server replies

A number of necessary client requests and server replies (responses) have been identified and can be classified into the following categories concerned with:

1) setting a session;

2) querying areas (records) for given criteria;

3) modifying map properties;

4) generating new data sets from the analysis environment; and

5) retrieving attribute data and modifying attributes in the attribute table.

Setting up a session is a complicated process that may involve many communications between the client and the server. However, these communications may be encapsulated in a single client request. How this could be possible is to be explained later.

The second category contains two requests. The first one takes a set of criteria – expressions of logical and spatial relations about data - as its argument and asks the server to return identifiers of areas that satisfy the criteria. The second request comprises an array of area identifies and requests the server to highlight the areas the identifiers define.

The third category comprises two requests. The first request, taking an attribute as an argument, asks the server to draw a map using the values of the attribute. The second request modifies the map properties such as shading scheme and pattern and is actually a compound of requests encapsulated in a single request.

The fourth category comprises only one request for creating a new coverage, which takes two arguments. The first argument is a special attribute, called a grouping index. The grouping index has a special property. That is, for any values of it if they are the same, the relating areas are part of the same area in the new data set. The second argument is values of attributes for each new area and the properties of the attributes (i.e. names and formats).

There are four requests in the last category for: 1) retrieving attributes, 2) adding new attributes to an INFO table, 3) deleting an existing attribute from a table, and 4) updating attribute values in a table. The first request takes an attribute name as its argument. The reply to it is the values of that attribute. The second one takes a new attribute name and its values, created by the client, as its arguments and asks the server to store them in a table. The third request takes an attribute name as its argument and requests the server to delete that attribute from a table. The last request takes the same arguments as the second request and requests the server to update modified attributes in the corresponding tables.

Although server replies may take different forms, each of them contains information on the status of the server processing the corresponding request.

5.3. System design

Based on the logical model for the integration discussed in Section 5.2, system design considers how this logical model can be realised in a given computer environment. The first part of this section therefore considers why a

UNIX-based computer environment running the X-window has been chosen for developing SAGE. The second part then considers the design of the client, the linking agent and the server. The main goal is to find appropriate structures for each component to increase their reusability and to make best use of techniques and computer resources available for implementation.

5.3.1. A computer environment for integration

A criterion for determining the appropriateness of a computer system is that it should allow the ARC/INFO presentations, including maps and GUIs, to be displayed on the client computer even if ARC/INFO is run on a different computer. One advantage of this is that no extra work is needed to make full use of ARC/INFO capabilities in presentation. Also compound client requests, such as setting up a session and modifying map properties mentioned above, could be simplified by invoking server-side programs with which the user interacts directly to accomplish the real requests. ARCTOOLS has already provided sufficient AML programs for doing these.

At the time the system was being developed, a UNIX system running the X window system is the most suitable computer environment, although ARC/INFO also runs on Microsoft Windows NT. The X window system itself is a distributed system and enables presentations generated by X window applications on a remote computer to be displayed on a local computer running an X window server. The X window is available on most UNIX-based computer systems. Although there are similar systems, such as WinFrame, enabling the presentations generated on a Windows NT server to be displayed on a local computer, they were not widely used and not available for this project. Since there were some implementations of the X window server for MS Windows, the system integration could be carried out in a hybrid computer environment. For example, ARC/INFO is run on a UNIX-based computer, while its presentations are displayed on a PC that runs an X server and the SSA component. However, this was not chosen because there were fewer software developing tools available to this project for this hybrid environment than for the UNIX system with the X window.

5.3.2. Design of the client

According to the classification of the system functions summarised in Section 5.1, it is natural to consider a structure for the client that includes three modules corresponding to three classes of functions DV, DAM and DM. Since SAGE is intended to support interactive spatial data analysis, graphical user interfaces (GUIs) must be provided to enable the user to access the functions and to control their behaviours. Because the implementation of GUIs often involves the use of some platform dependent graphic libraries, it would be desirable to shield GUIs from the DV, DAM and DM modules. Thus there is a further module - the GUIs module.

As indicated in the previous sections, the system functions may cause changes in the client repository and request the server to perform corresponding actions (e.g. highlighting areas in a map and inserting or removing attributes). It is clearly an advantage if only one module is responsible for doing these tasks. The DM module is more suitable than the others since it exists basically for managing data.

Figure 5.3 shows a layered structure with four modules for the client. The GUIs module sits at the top of the structure interfacing the user with the system. The DV and DAM modules underneath the GUIs module support the client-side data visualisation functions, and the data analysis and modelling functions respectively. The DM module manages the access to the client data repository, issues client requests to the server and handles the server replies on behalf of the other modules. Arrows in Figure 5.3 indicate the relationships between the modules. An arrow pointing from module A to module B implies that A calls one or more of B's functions to perform some operations. A function may return data to the caller.

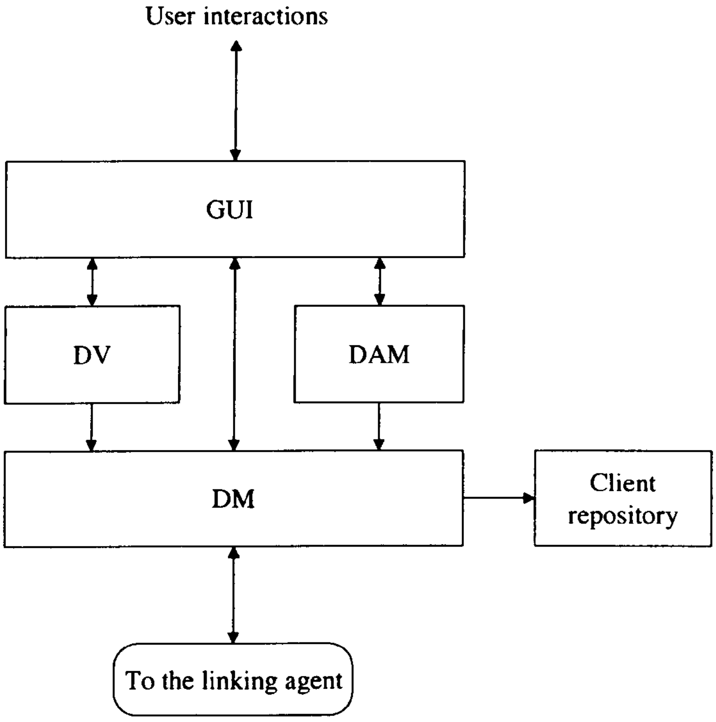


Figure 5.3. A layered structure for the client.

The following highlights the behaviours of each of four modules and the interplay between them. To support user interaction, the GUIs module provides three types of GUIs components - menus, dialogue boxes and windows. Menus provide a primary access to the system's functions while dialogue boxes allow the user to specify parameters controlling the behaviours of the functions. Dialogue boxes are also required to inform the user the state of operations, especially, long-lasting ones. There are two types of long-lasting processes - sending client requests to and receiving server replies from a remote machine over a busy network, and performing computationally intensive client functions. (e.g. regionalisation). A special technique is required for doing this and is to be discussed later in Section 5.4.2.

To simplify the management of different types of views, one window is chosen to hold the table view while one or more windows each holds a plot view of the same type. (The map view is managed by an ARCPLOT window.). Moreover, a text window is provided for displaying non-attribute results such as statistics in numeric form. Figure 5.7 shows a snapshot of these windows.

A table or plot window is associated with a set of functions for creating and deleting a view, editing the attribute data (for the table view only) or view properties, and selecting and highlighting view objects. A plot window is not responsible for 1) generating plot data and 2) deriving related areas for selected view objects and related view objects for selected areas. These responsibilities are delegated to the DV module. So the GUI module calls the DV module to perform these two actions. When the DV module is called to derive areas as a result of the user having selected view objects, it passes the identities of these areas, called area identifiers, to the DM module for the purpose of 'hot linking' views, as explained below. One benefit of splitting the responsibilities in such a way is that a plot view and the user controls on that view (e.g. how to colour objects) are separated from the underlying data models supported by the DV module. Thus the data models can be used without being changed even if the view and the controls may be implemented differently.

Under this client structure, the DAM module is greatly simplified. When it is called, the DAM module may perform the following actions. First, it calls the DM module to obtain the required data according to the parameters passed by the GUIs module. Second, it performs the chosen analysis on the data. Third, it calls the DM module to save the results into the client and/or server repositories if they are attributes or W matrices. Otherwise it calls the GUIs module to display them in the text output window.

As said above, the DM module is responsible for accessing the client data repository, issuing client requests to the server and handling server replies on behalf of the other modules. Another main function of the DM module is to handle 'hot linking' windows. Whenever area identifiers are passed to the DM module either by the DV module or by the server (in response to the first request in the third category) as a result of some view objects being selected, the DM module calls the GUIs module to highlight corresponding view objects in the table window and every plot window, and issues a client request to the server to highlight corresponding areas.

The GUIs module may call the DM module to issue a client requests regarding setting a session environment, modifying map properties, or querying areas interactively. For each of the first two cases, the server GUIs are invoked, and the user interacts with them to make specifications. For the last case, the user could query areas using one of the predefined tools (see the *SAGE User Guide* for details.). The server performs specific operations according to the specifications, and returns results - area identifiers. The GUIs module calls the DM module to update the client repository and to request the server to update its repository when the user modifies attribute values from the attribute table. The GUIs module may call the DM module to query available attributes, W matrices and the current area-identifiers in the client repository in response to user actions. If an error occurs during these processes, the DM module calls the GUIs module to show error messages.

5.3.3. Design of linking agent

The linking agent is divided into two portions – the client portion and the agent portion. The client portion comprises a library of API functions corresponding to the five categories of the client requests discussed in Section 5.2.4. The client calls on these functions to issue client requests to the agent, which, in turn, invokes the server to process the client requests. The purpose of the client portion is to maximise modulation between the client SSA and the agent. The client does not need to know how the API functions are implemented.

It is possible for all API functions to be mapped directly onto the agent portion. However, in order to simplify the agent, a single internal function is used instead and defined below (illustrated in C):

```
int RPC(char * command_argument, int mode)
```

The command_argument is a string of characters containing two parts: a unique request name and corresponding argument. The argument may be passed to the server in one or many times indicated by the mode. The server reply for the request is placed in an internal buffer and can be retrieved by the client. The status of processing a request is returned by the function as an integer.

When the client calls an API function, the client portion converts that function call into a call to the internal function on the client portion. This call then invokes the same call on the agent portion. Remote procedure call (RPC) is a mechanism that enables this (Simon 1996). On receiving a client request, the agent portion performs linking operations (Abel and Kirly 1994). It maps a request to a set of instructions in AML to be executed by ARC/INFO, transforms arguments into appropriate forms required by those instructions, invokes ARC/INFO to execute the instructions, and waits for the server to reply. The reply is a string of characters containing both the results and the status of processing the request. Upon receiving the reply from the ARC/INFO, the agent portion extracts the status and the results and returns them to the client portion. When a call to the RPC() returns, the client portion retrieves and transforms the results to match the prototype of the calling API function. Figure 5.4 illustrates the agent structure and the processes carried out in the agent.

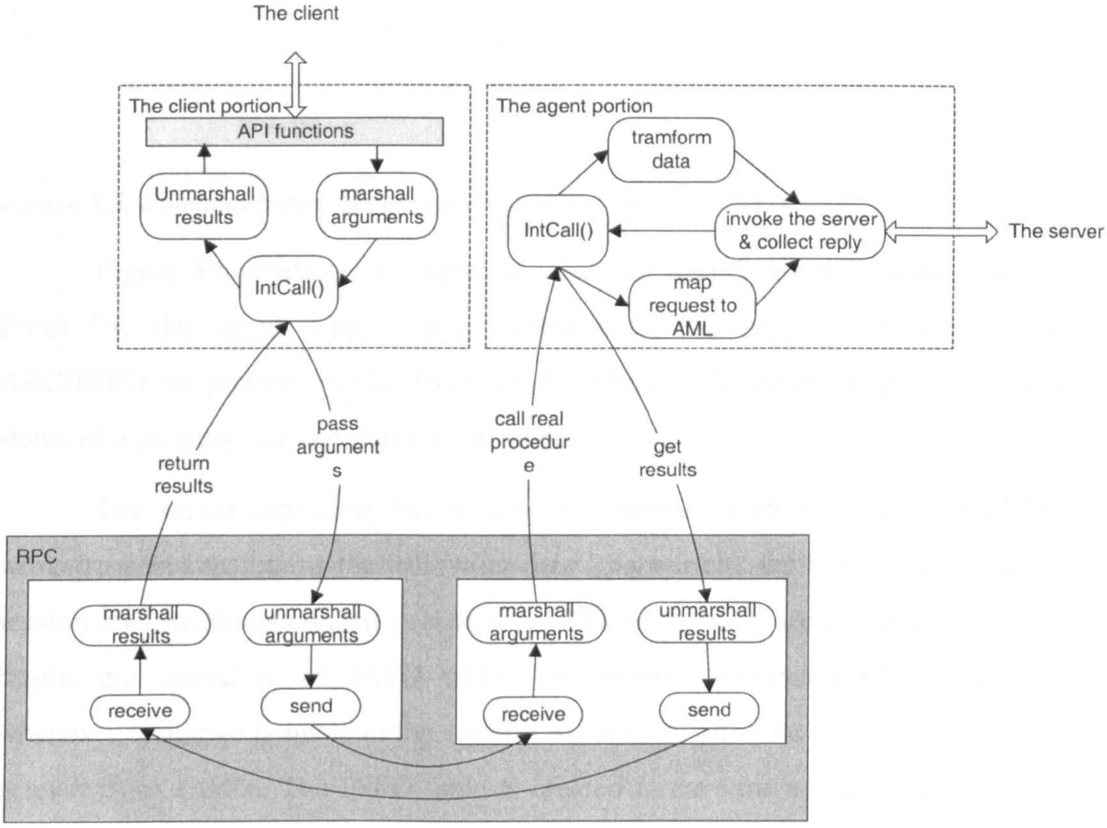


Figure 5.4. Structure of the linking agent.

5.3.4. Design of the server

The main concern in the design of the ARC/INFO server is how to make the linking agent and ARC/INFO communicate with each other and to organise the server repository. ARC/INFO is a command-oriented system, using standard input and output devices for issuing commands and outputting the results. Since UNIX treats the input and output devices as files and allows them to be re-directed to other files, an easy and simple way to enable their communications is to employ two FIFO (First In First Out) pipes to link the agent and ARC/INFO (Rochkind 1985).

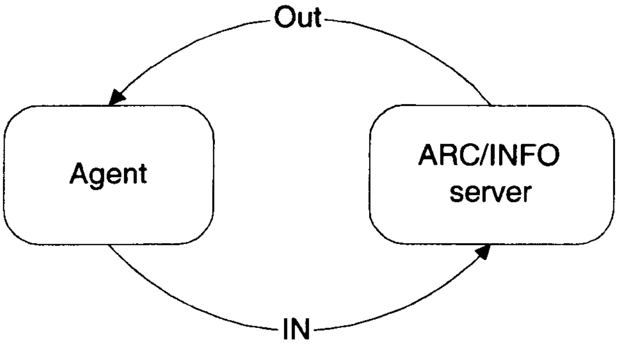


Figure 5.5. Communication between the agent and the ARC/INFO server.

Figure 5.5 illustrates the agent and the ARC/INFO server communications. From IN, the agent outputs (e.g. instructions and arguments) are sent to ARC/INFO to process, while from OUT, ARC/INFO output (e.g. results and status of a process) are sent back to the agent.

The server repository has a simple structure identical to an ARC/INFO workspace and maintains the following data - parameters defining a data analysis session and attributes created in a session. The parameters are in a file, while the results are stored in an INFO table. The session parameter file is updated whenever a change is made to the session, such as an attribute being added into or deleted from a table. The INFO table is created at the time when a session is set up and is related to the current polygon coverage.

5.4. System implementation and testing

This section summaries how each individual component was implemented and tested. It should be noted that there is no intention to cover every detail of implementation and testing but rather the procedure for doing these. Some of the key programming techniques and facilities used are discussed briefly. A computer environment used to implement this system is a SUN SPARC Workstation running the Solaris 2.5 operating system. The C language is the primary language used in the system implementation although other languages, such as FORTARN 77, are also used. The selection of C as the primary programming language but not an object-oriented language like C++ was largely determined by a factor that all required GUI components are programmed in C and wrapping all them into classes would require quite amount work without the assistance of professional programming tools.

5.4.1. Implementing and testing procedures

An incremental procedure was employed in the system implementation where the system grew gradually when the implemented functions were added into the system one by one. The linking agent was implemented first, followed by the implementation of the DM module. Then the DV module and the GUIs for it as part of the GUIs module were implemented. These resulted in an incomplete but working system. The DAM functions were implemented one by one along with GUIs, if required. Then both of them were added into the DAM and GUIs modules respectively.

Testing was carried out along with the implementation. Each function was validated in line with its requirements. For each function in the DAM module, its correctness was verified using a test data set enclosed in the SAGE distribution package. The results from many functions were found to agree with those results generated using SpaceStat (Anselin 1992).

The whole system has been tested on both a stand-alone computer and two networked SUN SPARC workstations. In the latter, the linking agent and

ARC/INFO are run on the same computer, whereas the client is run on another.

5.4.2. Implementation

The GUIs module was implemented based on Motif, an industrial standard of specification for the graphic user interfaces. A set of Motif widgets, Motif compliant GUI building blocks, was employed. Many of these widgets were available in a standard Motif widgets library shipped with the Solaris 2.5 development environment package. Two additional widgets, a plotting widget called Plotter and a table widget called Xbea, were obtained from the public domain. The plot windows and the table window were constructed using them respectively.

As mentioned in Section 5.3.2, a method is needed to prompt the running status for long-lasting operations. The key to the method is to be able to prevent the GUI from becoming blocked, which happens when a single thread is used to run the long-lasting task and to update GUI. One solution is to make a long-lasting operation as in a separate (child) process so that it can run independently of the (parent) process that handles the window updating. Figure 5.6 gives a schematic illustration of this multiprocessing computing. A memory space is assigned for the parent and child process to exchange information. The multiprocessing was implemented using the SUN multithreading library (thread.a).

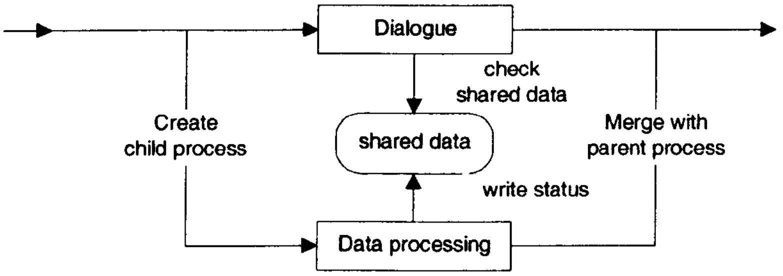


Figure 5.6. An illustration of multiprocessing for handling a long-lasting process.

The DM module was implemented straightforwardly in C. The function for querying high order adjacency was implemented based on an algorithm proposed by Anselin and Smirnov (1996).

The DAM was implemented in C also. Several functions were written from scratch, whereas others call NAG FORTRAN subroutines to perform internal and fundamental computations. Regionalisation was implemented based on a method described in Section 4.1 and by Wise *et al* (1997). This function is able to take any combination of three criteria, and to generate a grouping index. The user can randomly or manually select an initial partition. The algorithm underlying the implementation was that by Banfield and Bassill (1977). A hierarchical classification function was implemented using a number of NAG subroutines, while some simple classification schemes on the single variable were implemented from scratch.

An additional data preparation function is developed for creating new attributes. With it the user can specify any arithmetic expression to define and create new attributes. Specific grammars were defined in BNF. An expression can be defined using arithmetic operators (+, –, *, and /), and operands (numeric and attribute names). A set of pre-defined functions was made available. YACC (Yet A Compiler's Compiler) and LEX tools (Bennett 1990) were used to generate a C program to perform lexical and syntax checks on expressions and arithmetic calculations. Data preparation functions for creating W matrices were written based on the formulae in Haining (1993).

The data analysis and modelling functions were implemented by coding up formulae described in Sections 4.2 and 4.3 in C. Many NAG subroutines were called to calculate probabilities of random variables under different distributions. NAG subroutine G02DAF was used to perform the least squares fit of models 4.2.1, 4.3.1, 4.3.2 and 4.3.4. For fitting models 4.3.3, and 4.3.5 the maximum likelihood (ML) fitting procedures were implemented from scratch based on Anselin's (1988, p182-3) algorithms and facilitated by a number of NAG subroutines (searching optima, manipulating matrices, and solving linear equations). NAG routine G02GCF were used to fit a generalised linear regression model with Poisson errors using maximum likelihood (McCullagh and Nelder 1989).

In order to handle exceptions raised by the operating system, the UNIX signal mechanism was employed to detect a number of abnormal situations that are likely to arise. A C program was written to handle these situations. Upon catching an abnormal situation, the system calls that program to dump the client data into the client data repository and prompts the user to exit the system.

The linking agent was implemented using RPC and AML. The RPC was used to implement the remote procedure call RPC( ). An RPC protocol was defined and the C code for it was generated using rpcgen, an RPC protocol compiler (SunSoft 1994). The agent was implemented not to generate AML instructions for each request on the fly but to match each request to a set of pre-written AML programs. The AML programs were written in the same object-oriented approach as ARCTOOLS and call many ARCTOOLS programs directly. So very few programs were written to utilise setting up a data analysis session and modifying the map properties. In each pre-defined AML program, a pre-defined tag (a character string) is attached to every server reply after the server executes the program. That tag indicates whether the server succeeds or fails to complete the client request. The agent uses tags to identify the status of a requested process.

5.5. SAGE limitations

The previous sections have shown the benefits gained owing to the decisions taken for the system integration. This section looks at two limitations of SAGE that are also attributable to these decisions. The first mainly attributes to the selection of the client-server model where SSA and ARC/INFO are independent of each other. Since SAGE relies entirely on ARCPLOT to generate maps on the server, the communication between the client and the server tends to be very heavy and cannot be made efficient enough to satisfy the needs for highly dynamical data exploration, such as brushing (Goodchild et al 1992). This prevents a range of techniques, including those suggested by Brunsdon *et al* (1996) and Fotheringham *et al* 1(996) and implemented elsewhere (e.g. LiveMap Brunsdon (1998)), from being implemented in SAGE.

Another limitation is that SAGE does not support the generation of choropleth maps with legends. Although ARCTOOL provides a set of functions to

ARCPLOT for doing this, theses functions were not called by SAGE since the mapping process in SAGE is already too complicated for not only occasional users but also experienced users to use. Because of the limitation, some maps in the next chapter have to be produced using Arc/View. These two limitations could be remedied if there were a fully functional and user-friendly mapping component that can be integrated as part of the SAGE client.

SAGE has another major limitation. At present, SAGE only allows results for all areas to be saved in its repositories for later use. Results relating to part of a study region can only be printed out in the text window. To use such results in any further analysis, they have to be imported into SAGE. This limitation exists because SAGE lacks a mechanism that can record information on the conditions on which such results are derived. In the context of the analysis of non-spatial data, it is often sufficient to label items corresponding to the 'omitted' cases simply as 'missing' values. However, this may not be sufficient in the context of the analysis of spatial data, because results for part of the region may be yielded under different assumptions of inter-area relationships due to the necessity of re-coupling some remaining areas when others are excluded (Haining 1994). Therefore, whether results for part of the study region could be used meaningfully in a subsequent analysis relies on the availability of related information. A mechanism for recording information needs to be implemented if this limitation is to be removed effectively.

Besides these limitations above, two further drawbacks are the *ad hoc* nature of system integration in terms of application interfaces and the use of a non-object-oriented approach and programming language to the system analysis and implementation. Although efforts were made to increase the reusability of SSA techniques for other GIS packages by modularization, a large amount of work might still be required to modify the linking agent if a GIS to be linked is very different from ARC/INFO. As SAGE client is not implemented using object-oriented language, this makes the extension of SAGE SSA functions difficult. The *ad hoc* nature would not be removed until there is a standard that defines universal spatial data models, operations for them and API. Developing such a standard has been the ongoing work of the Open GIS Consortium (http://www.opengis.org).

5.6. SAGE - a user perspective

The previous sections in this chapter discussed SAGE from a system developer point of view. It showed how those SSA techniques considered in Chapter 4 were integrated with ARC/INFO. This section presents an overview of SAGE from a user perspective in order to show the user:

1. how to create different plots of data;
2. how to query the data from plots, the map and the table as well as using SQL-like expressions; and
3. how to perform statistical tests and fit a regression model.

A pre-prepared data set is used throughout this discussion. The data set is in the form of an ARC/INFO polygon coverage and contains 48 polygons. Variables included (not all used) are:

PAR30_85 – total population between 30 and 85 years old;

OBS – the number of observed incidences of a rare form of cancer.

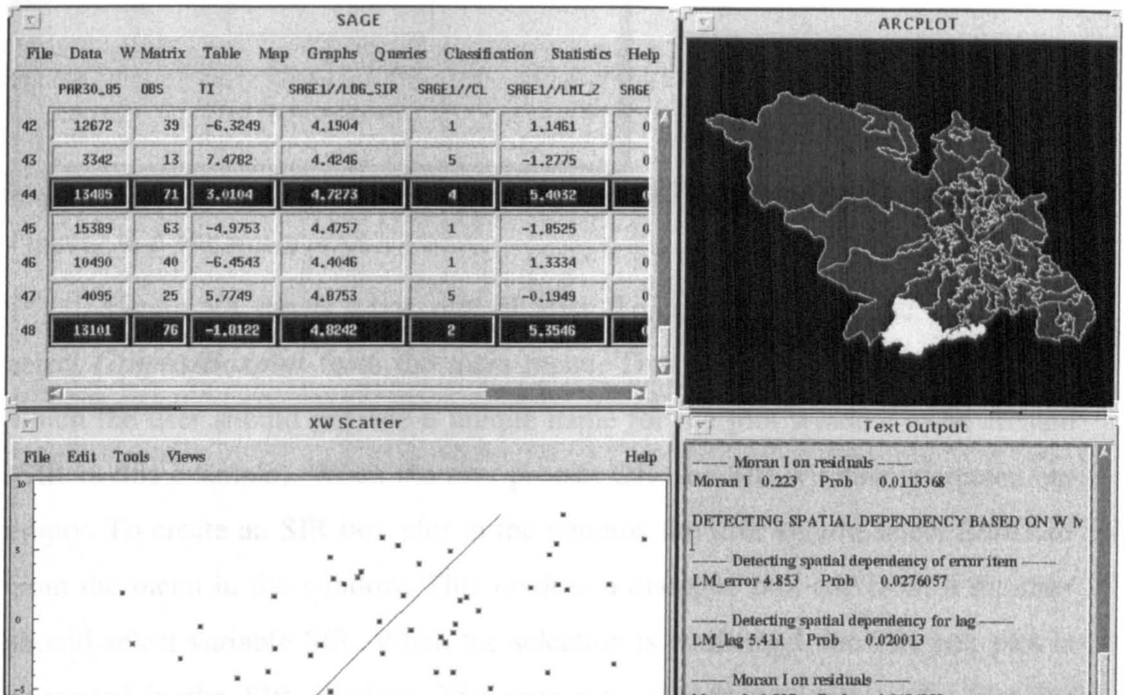
TI – Townsend Index - an index of deprivation (positive values - relative deprivation, negative values - relative affluence)

SIR - standardised incidence rate for the cancer (standardised by age and sex).

This data set is one of two synthetic data sets distributed with the SAGE package.

To be consistent with the *SAGE User Guide*, this chapter uses the same notations as in that document. The ***Bold italic underlined*** words are reserved for menu and dialogue elements. ***File/Open*** indicates that ***Open*** is a sub-element in the ***File*** menu.

Figure 5.7 shows a snapshot of a typical SAGE session. Of four windows, SAGE, ARCPLOT and Text Output exist throughout a SAGE session. A statistical plot window is displayed whenever the user requests the system to draw a statistical plot. The following discussion assumes a session has been set up successfully. For more details on this, the user is referred to Chapter 3 in the *SAGE User Guide*.


SAGE
File Data W Matrix Table Map Graphs Queries Classification Statistics Help
ARCPLOT
XW Scatter
File Edit Tools Views Help
10
5
0
-5
-10
-4
-3
-2
-1
0
1
2
3
Text Output

	PAR30_85	OBS	TI	SAGE1//LOG_SIR	SAGE1//CL	SAGE1//LMI_Z	SAGE
42	12672	39	-6.3249	4.1904	1	1.1461	0
43	3342	13	7.4782	4.4246	5	-1.2775	0
44	13485	71	3.0104	4.7273	4	5.4032	
45	15389	63	-4.9753	4.4757	1	-1.8525	0
46	10490	40	-6.4543	4.4046	1	1.3334	0
47	4095	25	5.7749	4.8753	5	-0.1949	0
48	13101	76	-1.8122	4.8242	2	5.3546	

```
----- Moran I on residuals -----
Moran I   0.223    Prob     0.0113368

DETECTING SPATIAL DEPENDENCY BASED ON W M

----- Detecting spatial dependency of error item -----
LM_error 4.853    Prob     0.0276057

----- Detecting spatial dependency for lag -----
LM_lag 5.411    Prob     0.020013

----- Moran I on residuals -----
Moran I   0.237    Prob     0.0126252
```

Figure 5.7. A snapshot of a SAGE session.

5.6.1. Graphical tools

This section shows the user how to create plots and a choropleth map. SAGE enables the user to create seven types of plots: histogram, XY scatterplot, XW scatterplot, rankit plot, boxplot, lagged boxplot and matrix plots. As discussed in Section 5.3.1.3, one or more plots of the same type may be displayed in a window. Each window consists of two main components: a menu for accessing the tools particular to the window and an area for drawing the graphs. This section will illustrate only how to create a box plot and an XW plot. The user is referred to Chapter 8 in the *SAGE User Guide* for details on creating other types of plots.

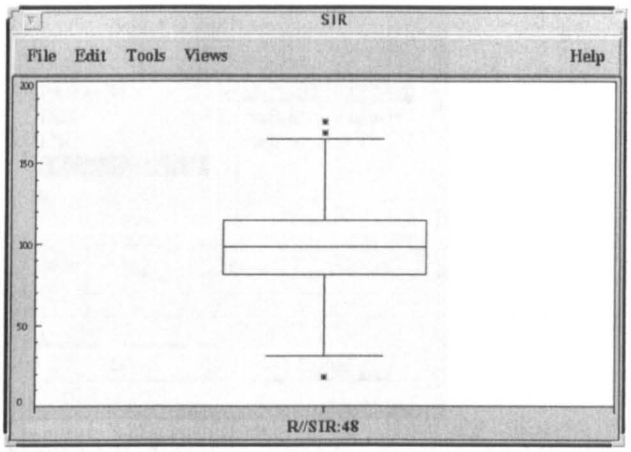
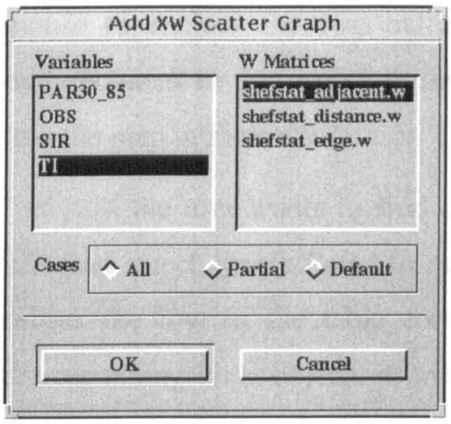


Figure 5.8. A box plot of SIRs.

Figure 5.8 shows a box plot of SIRs. To create this plot, the user should select ***Graphs/Boxplot*** from the main menu. This invokes a dialogue box from which the user should provide a unique name for the plot window to be created (SIR in this example). When the user presses OK, the SIR window is created but empty. To create an SIR box plot in the window the user should select ***Edit/Add*** from the menu in the window. This invokes a dialogue box and from it the user should select variable SIR. When the selection is confirmed, the SIR box plot is generated in the SIR window. The user may choose to generate the plot for selected cases. Details for doing this are given in the *SAGE User Guide*. As shown in Figure 5.8, SIR has a fairly uniform distribution with three outliers, two above the upper whisker and one below the lower whisker. The user can see how to identify areas with which outliers are associated later.

Suppose that the user wants create an XW plot for TI for all areas, similar to one shown in the XW Scatter window at the bottom left corner in Figure 5.7, to identify spatial outliers, or relative disadvantaged areas with respect to their neighbours. The user should follow the same procedure shown in the previous example to create an empty XW Scatter window by selecting ***Graphs/XW Scatter***. When the user selects ***Add/Edit*** from the window, a dialogue box as shown in Figure 5.9 is invoked. In order to create an XW plot for TI, the user should select TI and a W matrix. In this example, shefstat_adjacent.w, the adjacency matrix, is chosen. When the user confirms the settings, an XW plot of TI is added into the window.



Add XW Scatter Graph

Variables: PAR30_85, OBS, SIR, TI (selected)

W Matrices: shefstat_adjacent.w (selected), shefstat_distance.w, shefstat_edge.w

Cases: ● All ○ Partial ○ Default

Figure 5.9. A dialogue box for creating an XW plot.

SAGE is able to create a choropleth map for any variable using ARCPLOT. Areas can be shaded according to the mapping of values of the variable to the colour or grey indices in the current colour map. In SAGE, the user is encouraged to map the values to indices (of an index variable, integers starting from 1 onwards) explicitly with SAGE classification functions. The user is referred to Chapter 7 in the *SAGE User Guide* for details on how to use the classification functions.

Suppose that the user has created an index variable CL. To create a map of CL, the user needs to select it by pressing the right-hand mouse button and the Shift key from the keyboard at the same time and then select ***Map/Map Item*** from the system menu. This causes ARCPLOT to create a choropleth map.

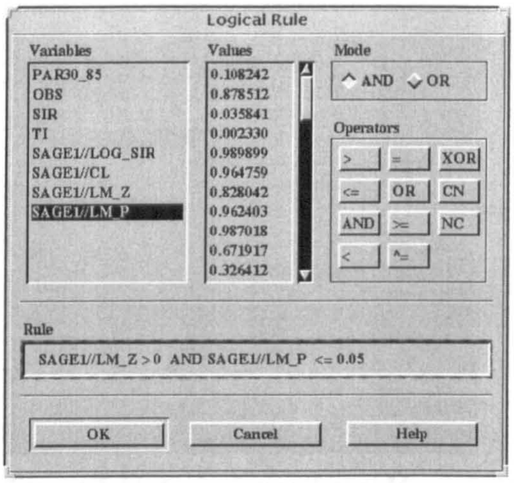
5.6.2. Data query

SAGE allows the user to query data by picking up graphic elements from plots, the map, and the table. It also allows the user to query data using SQL-like expressions. As discussed in Section 5.3.1, the data querying in SAGE is fully integrated under the 'hot' linking mechanism; that is, a query updates current area identifiers and causes all graphical elements associated with the identifiers in any graphic windows to be highlighted.

As shown in the box plot of SIRs, there are two large outliers with respect to 100. To see where these two outliers are on the map and in the table, the user should select ***Tools/Select*** from the box plot window, then click and drag the mouse to enclose the two outliers. When the selection is accepted, the two outliers, areas in the map and rows in the table relating to them, are highlighted (see the map in Figure 5.7).

If the user wants to find out attributes of a specific area in the map, the user can select ***Query/Point*** from the main menu and then click that area. This causes the row in the table corresponding to that area, as well as any other corresponding view objects in plots, to be highlighted. Then the user can locate the row and examine the attribute of interest. Note that SAGE allows the user to pick up more than one area. In reverse, the user can select table rows to find which graphical elements they are associated with in different statistical plots and the map. Detailed information for doing this is given in Chapter 9 in the *SAGE User Guide*.

SAGE enables the user to query data using logical expressions alone or in conjunction with the kinds of queries illustrated above. Figure 5.10 shows a dialogue box that allows the user to specify a logical expression. To invoke this dialogue, the user should select ***Query/SSQL*** from the system menu. This invokes a dialogue box for the user to compose a set of queries. When the user presses the ***Logical*** button in that dialogue box, it invokes the dialogue as shown in Figure 5.10. The expression shown is that LM_Z is positive and LM_P is less then 0.05, where LM_Z and LM_P correspond to two variables: the local Moran's I values of the SIRs and their significance levels. This expression is specified for identifying clusters of areas with similar SIRs. A tool for creating the two variables is discussed below.



Variables:
PAR30_85
OBS
SIR
TI
SAGE1//LOG_SIR
SAGE1//CL
SAGE1//LM_Z
SAGE1//LM_P

Values:
0.108242
0.878512
0.035841
0.002330
0.989899
0.964759
0.828042
0.962403
0.987018
0.671917
0.326412

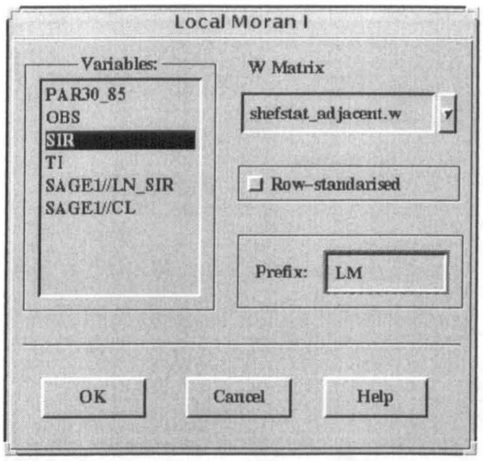
Rule: SAGE1//LM_Z > 0 AND SAGE1//LM_P <= 0.05

Figure 5.10. Logical query dialogue box.

5.6.3. SAGE statistical functions

This section considers two statistical functions available in SAGE. The first enables the user to calculate the local Moran's I values and their significance levels for each area which can be used to identify clusters. The second allows the user to fit a classical linear regression model to data using least squares.

Figure 5.11 shows a dialogue box for the first function. This dialogue box is invoked when the user selects ***Statistics/LISA/Local Moran I***. As indicated by the settings in the box, the local Moran's I values for SIRs and their corresponding significance levels are calculated using the adjacency matrix shefstat_adjacent.w and saved as variables LM_Z and LM_P respectively. When the user presses ***OK***, those two variables will be created and shown in the table. Then clusters can be identified as shown in Section 5.9.



W Matrix: shefstat_adjacent.w

□ Row-standarised

Prefix: LM

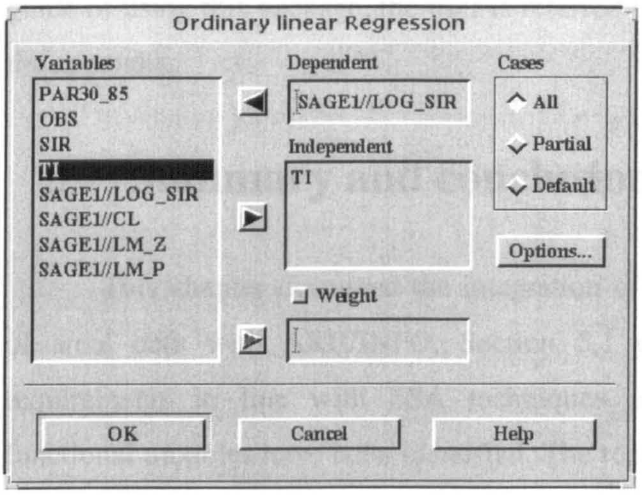
Figure 5.11. Calculating Local Moran's I values and significance levels.

SAGE enables the user to fit a classical linear regression model to data as well as other spatial forms of the regression model. All these tools perform statistical tests on the goodness-of-fit of the models and coefficient estimates. A test of the spatial dependence of the model residuals is optional. Residuals, fitted values and leverages may be saved as variables. Studentised and standardised residuals and Cook's distances may also be saved for further analysis.

Suppose that the user wants to fit a linear regression model of LOG_SIR, the dependent variable, on TI, the independent variable. LOG_SIR is the logarithm of the SIR constructed with a function accessible from ***DATA/Arithmetic Variable***. Note that variables created during a session are prefixed with the name of an attribute table (in the server repository) in which the variables are actually stored (e.g. SAGE1/ in this example) (see Chapter 4 in the *SAGE User Guide* for details).

Figure 5.12 shows the dialogue box for the user to specify a linear regression model. To invoke this dialogue, the user should select ***Statistics/Linear Regression/OLR***. The settings shown in the dialogue box specify the model stated above.

Figure 5.13 shows a dialogue box that the user uses to choose optional statistics to perform and new variables to create. This dialogue is invoked when the user clicks ***Options*** in the previous dialogue box. As shown, three spatial dependence tests on the model residuals are selected, as is the adjacency matrix.



Dependent: SAGE1//LOG_SIR

Independent: TI

Figure 5.12. Specification of linear regression model.

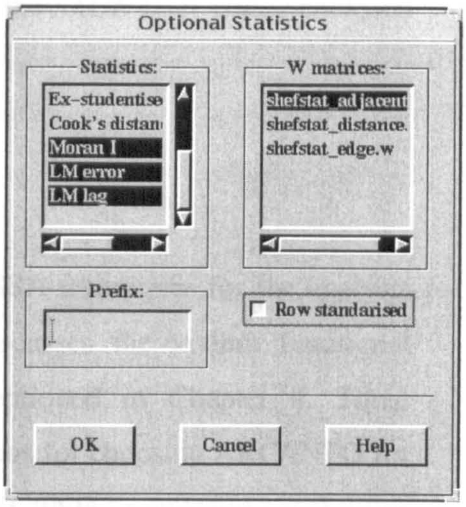


Figure 5.13. Selection of optional statistics.

```
        FITTING MUTIPLE ORDINARY LINEAR REGRESSION
        --- Variables used in OLR ---
        Dependent Variable:     SAGE1//LOG_SIR
        Independent Variables: Constant   TI
        OBS       48       VARS       2        DF        46
        R2      0.040    R2-adj   0.019
        LIK     22.688    AIC     -41.375   SC      -37.633
        SIG_SQ    0.024      SIG_SQ(ML)  0.023
        F-test   1.919     Prob      0.173
        ---- Shapiro-Wilk Normality Test ----
        Values    0.820     Prob       0.000
        --- Ordinary Least Squares Estimation ---
        Variable  Coeff      S.D.      t-value   Prob
        Constant  1.979      0.022     89.009    0.000
        TI        0.008      0.006     1.385     0.173

        DETECTING SPATIAL DEPENDENCY BASED ON W MATRIX,
shefstat_adjacent.w (row-standardised)
        ---- Detecting spatial dependency of error item ----
        LM_error 0.361      Prob       0.548149
        ---- Detecting spatial dependency for lag ----
        LM_lag 0.573      Prob       0.44889
        ---- Moran I on residuals ----
        Moran I   0.061      Prob       0.398774
```

Table 5.5. Outputs of fitting a linear regression.

Table 5.5 shows the outputs from fitting the regression model, displayed in the ***TextOutput*** window.

This section shows a number of SAGE functions in the hope of giving the user a feeling of what SAGE looks like and of the way SAGE works. For a full guide of using this package, the user is referred to the *SAGE User Guide* enclosed the appendix.

5.7. Summary and conclusion

This chapter discussed the integration of SSA techniques for the analysis of areal data with ARC/INFO. Section 5.1 discusses the system functional requirements in line with SSA techniques mentioned in Chapter 4. Three functional modules have been identified. The reasons for choosing ARC/INFO for integration have been given on the basis of the ARC/INFO's superior capabilities in handling spatial data and its availability to UK higher education institutions as well as its popularity in the GIS community. What ARC/INFO does and does not support for integration has been examined. Non-functional requirements on the behaviours of the integrated system have also been considered.

Section 5.2 discussed the modelling of the system. Based on a client-server model, a three-tier system architecture is considered to be appropriate for the system. The middle tier, the linking agent, functions as a mediator between the SSA component and ARC/INFO, which function as client and server tiers respectively. An advantage of this architecture is that it increases the component reusability and extensibility. However, the architecture has a disadvantage since ARCPLOT is run as part of the server but not the client. As a result, some highly dynamic SSA techniques could not be implemented to work as effectively as would be expected. A set of client requests and server replies was identified for SAGE, applicable to integration of the same SSA techniques and other GIS.

Section 5.3 considered the design of SAGE components in a networked UNIX-based computing environment running the X window system. The reason for choosing such an environment was that there were more facilities required at that time for integration for it than for others. It should be noted that this has been changed dramatically since then largely as a result of the advances in computing technologies, particularly on Microsoft Windows platforms. The design of the client identified a three-layered structure comprising four modules. The advantage of such a structure is that it maximises the reusability of SSA functions because they are effectively separated from those for handling the user interactions and access to spatial data and GIS functions. The linking agent was split into the client portion and the agent portion, which communicate with each other using the remote procedure calls (RPC). The agent portion performs a set of linking operations and talks to the server through two UNIX named pipes.

Section 5.4 summarised the SAGE implementation and testing. Under an incremental procedure, SAGE functions were implemented one by one, and tested using a data set. Some results were found to agree with those derived from SpaceStat. Some key programming facilities and techniques were considered.

The whole process of system integration demonstrated how a set of state-of-the-art SSA techniques might be linked with ARC/INFO. By so doing, the functionality of ARC/INFO has been extended beyond the spatial data management and basic spatial data operations to supporting exploratory and confirmatory spatial data analysis. Experience gained from this project and summarised above, especially at the system analysis, modelling and design stages, could be useful in other similar projects.

Section 5.5 considered SAGE limitations. One is that SAGE does not support efficiently hot-linking between the map and other windows because the high overhead in communication between the client and the server. Another is that SAGE does not support choropleth mapping with legends. These two limitations, however, may be remedied by integrating a suitable mapping component as part of the SAGE client. A third limitation is that SAGE cannot store results for part of a study region in an attribute table for later use. This is due to the lack of mechanism of tracking conditions on which the results are obtained. Further study into this problem is then required. The development of SAGE is subject to two further drawbacks – the *ad hoc* nature of system integration and the use of a non-object-oriented approach. Section 5.6 showed the user how she/he may use SAGE in a study.

For the purposes of SAGE evaluation, the next chapter will present a case study of colorectal cancer incidence for the city of Sheffield using SAGE.

CHAPTER 6. SAGE EVALUATION

This chapter presents an evaluation of SAGE through a case study of colorectal cancer (CRC) incidence on a small scale in the city of Sheffield. Since CRC is a disease with low occurrence, it is likely to give rise to those problems that have been discussed in the previous chapters and are common in small area studies of rare diseases. Therefore, this case study offers an opportunity to examine how well SAGE meets the need of analysing such health-related data, and to uncover its limitations.

This chapter is divided into five sections. Section 6.1 briefly describes the background to this case study including questions about CRC in which some health professionals are interested, data sets used in this study and some problems that they raise for SSA, in particular the small number problem and heteroscedasticity. The construction of a regional framework using SAGE for the subsequent analysis is summarised also. Section 6.2 summarises the use of SAGE tools to describe the properties of the incidence of CRC in line with the questions of concern. Section 6.3 summarises the use of SAGE modelling tools to model the spatial variation of incidence of CRC and the relationships between it and socio-economic deprivation. Section 6.4 discusses SAGE limitations, followed by conclusions.

6.1. Background, data sets and regionalisation

CRC is the second most frequent cause of cancer death in England and Wales (ONS). As with cancer at other sites, causes are not yet known. Advances in diagnosis technology, however, have made it possible to diagnose the disease at its early stage, while advanced treatments can significantly extend the survival of the early-diagnosed patients (Mandel *et al* 1996, Richards *et al* 2000). The possibility of providing a screening programme for the population in Britain has been investigated (Gavican 1989).

In response to this, health authorities have called for the geographical studies of CRC to provide information about it so that they may use this information to plan the deployment of screening resources in the most cost-effective ways. For this purpose, a health authority may want to know how CRC incidence is distributed across its administrative region. It may be interested in knowing any clusters where the population is likely to have high risks of CRC. Areas in each cluster may be dealt with together in later resource delivery, perhaps in association with any screening programme. The health authority may also want to know which areas are likely to be 'problem areas'. These areas may need more resources than others. It may also like to know any possible relationships that may exist between CRC and some socio-economic and/or environmental factors. A strong relationship might be used to 'predict' where CRC is likely to arise later.

One geographical study in this context is that carried out by the University of Sheffield in collaboration with the Sheffield Health Authority (SHA) as part of a project intended to investigate the ways in which GIS and SSA might contribute to the work of the SHA. Results of this study have been reported elsewhere (Haining *et al* 1994). That project has greatly influenced a great deal the development of SAGE. Therefore, it is appropriate for this study to analyse CRC data for the purpose of evaluating how helpful SAGE may be to obtain useful information.

Two data sets were used in this study: 1981 ED-based census data and postcoded colorectal cancer incidences from 1979 to 1983 for the city of Sheffield. The former was obtained from OPCS 1981 small area statistics, while the latter from the Trent Cancer Registry. CRC data for that period was chosen deliberately: 1) to increase the number of CRC cases and, therefore, the reliability of estimates of relative risks; and 2) to use 1981 census data to estimate the population not available for years other than 1981. The data sets were used in the previous study mentioned above and stored in the form of ARC/INFO polygon coverage and point coverage respectively (Haining *et al* 1994). Although it would be ideal to use much more recent data sets, the CRC data set for that period was the latest available for this project. Since the case study is not intended to be a contemporary study of the CRC incidence but rather an evaluation of SAGE, these two rather old data sets are thought to be still appropriate.

In the pre-prepared 1981 census data set, there was a total of 1158 EDs. Each ED was attached to a set of attributes including the size of the population for each age group, the number of households, the number of households without a car, the number of people not being home owners, living in overcrowded conditions and being economically active age who are unemployed. Other attributes also exist but were not used in this study. Of 1158 EDs, 12 contain no data on any attributes and thus each of them was merged arbitrarily with one of its adjacent EDs.

The CRC data for each year was stored in an ARC/INFO point coverage. For each incident, a set of attributes was recorded and some of the attributes useful for this study include full address, postcode, age and sex. In order to relate incidences to EDs, each address was assigned with a pair of co-ordinates using the PinPoint Address Code. Work was also done, in the previous project, to trace unmatched full addresses and addresses given multiple address codes. Thus the data set was 90% and almost 100% complete and accurate for the 1979-1980 and 1981-1983 periods respectively (Haining *et al* 1994).

For the five years, there was a total of 1622 incidences recorded for the city of Sheffield. Of them, 1605 were assigned to EDs after a point-in-polygon search. 17 fell outside the boundary of the city and were therefore excluded from the further analysis. This process resulted in a single ARC/INFO polygon coverage including attributes: the population in each of five-year age bands starting from 30 to 85+, the count of incidences decomposed to the same age bands, and the socio-economic data mentioned above. The total population in these bands was just over 300,000 out of the total population of just above 500,000.

The population under 30 years old was excluded because there were no incidences. The population over 30 was included since there was no evidence of problems attributable to the older population in this data set that would require the population over a specific age to be excluded. Since there is no clear difference between the number of incidences in males and in females, the sex composition is not considered either (Haining *et al* 1994).

Table 6.1 lists the number of incidences for each of five years in five year cohorts. The last row records the number of incidences falling outside the boundary. Figure 6.1 shows a bar plot of these data. Figure 6.2 is an overlay map of the CRC incidences across Sheffield EDs.

	Year 79	Year 80	Year 81	Year 82	Year 83
AGE30-34	0	0	1	1	0
AGE35-39	0	1	1	0	1
AGE40-44	3	2	3	3	4
AGE45-49	4	7	2	5	8
AGE50-54	13	7	6	11	8
AGE55-59	9	14	15	14	16
AGE60-64	32	25	34	19	28
AGE65-69	42	34	44	51	39
AGE70-74	63	67	52	44	55
AGE75-79	57	73	64	61	69
AGE80-84	35	56	43	43	60
AGE85+	47	53	53	47	56
Sum	305	339	318	299	344
Outsider	3	2	6	3	3

Table 6.1 Counts of colorectal cancer incidences for each of five years for the age groups.

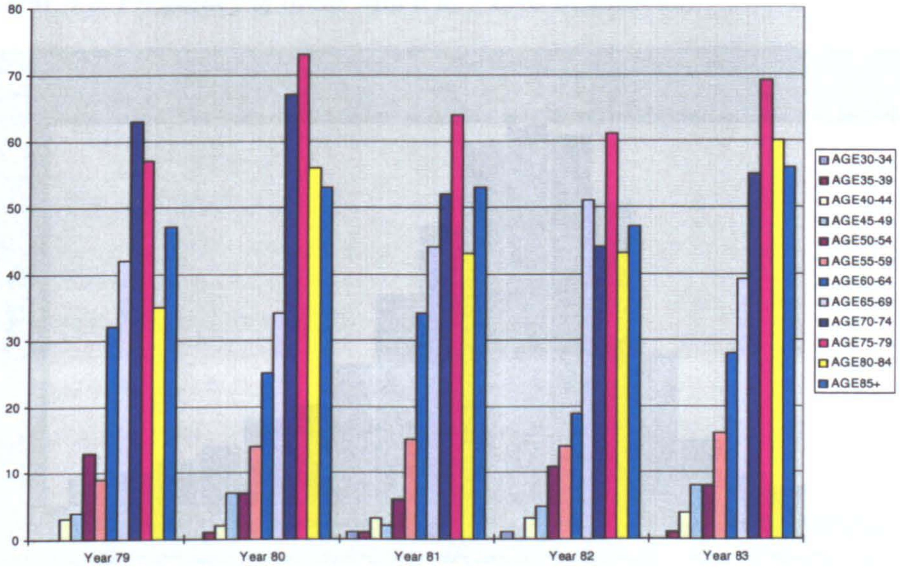


Figure 6.1. A graphic presentation of Table 6.1.

Table 6.2 shows the number of areas that have 0 to 7 incidences and the percentage of that number with respect total 1146 areas. 93% percent of EDs have 0 to 3 incidences. Figure 6.3 shows the existence of a large variation in the size of the population across EDs. Clearly, if EDs were used as a framework in the subsequent analysis, as discussed in Chapter 2, the small number problem and the heteroscedasticity problem would be substantial and affect SSA.

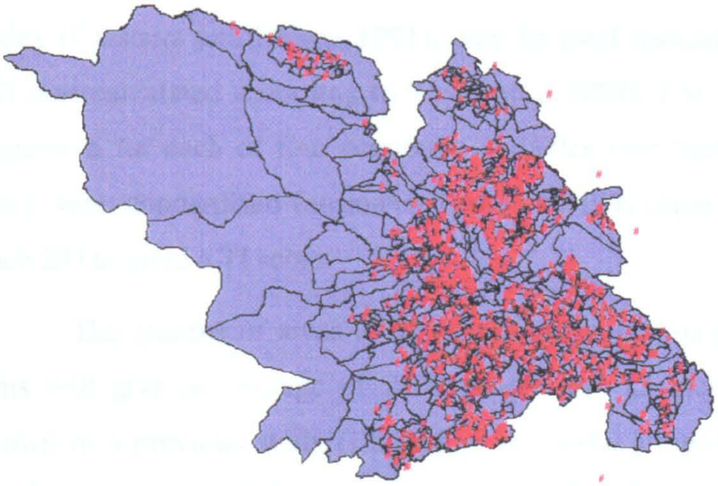


Figure 6.2. Colorectal cancer incidences for the years 1979 to 1983 shown as dots.

Case	0	1	2	3	4	5	6	7	SUM
Area	344	348	250	112	59	17	12	4	1146
Percentage	30.02	30.37	21.81	9.77	5.13	1.48	1.06	0.36	100.0

Table 6.2. Frequency of areas with 0 to 7 CRC incidences.

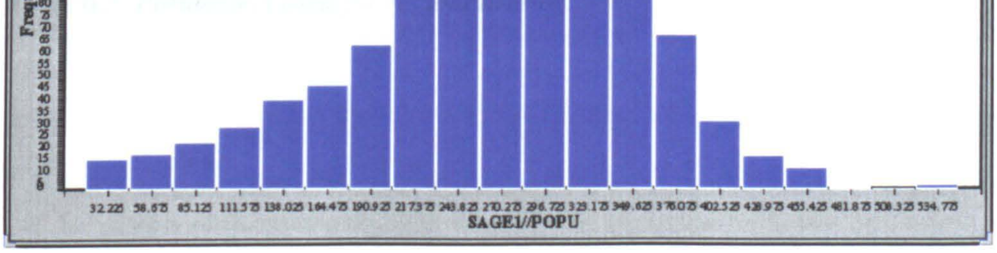


Figure 6.3. Distribution of the population over EDs.

In order to overcome these problems, this study chose to construct a new framework on EDs. For the purpose of this study, criteria for constructing the framework were equality in the population at a certain size and homogeneity with respect to socio-economic circumstances as well as contiguity. The homogeneity criterion is required for investigating the relationship between CRC incidence and socio-economic deprivation. The Townsend index (TI) was chosen as the measure of socio-economic circumstances although other measures, such as the Carstairs index (Carstairs and Morris 1991), may be used instead. A score of TI for each ED was calculated according to Townsend (1989). For each ED, four rates were computed for each of four economic variables (see Section 6.1). Then, the four rates were standardised (normalised) over all EDs separately and summed up for each ED to yield a TI score.

The number of areas in the new framework was chosen to be around 120. This will give an average of 2600 people and 13 incidences in each area. As shown in a previous study (Haining *et al* 1994), a regional framework with this number of areas might be able to preserve 85% to 90% of the original information on socio-economic circumstances measured by an information statistic (Johnston and Semple 1983).

Table 6.3 lists parameters chosen for the SAGE regionalisation tool (see Chapter 10 in the *SAGE User Guide*). The weights and thresholds were determined through many experiments. An initial partition was selected randomly by SAGE.

	Homogeneity	Equality
Variables	Townsend Index	Population
Weight	1.0	0.000005
Threshold	5%	5%

Table 6.3. Parameters used for regionalisation.

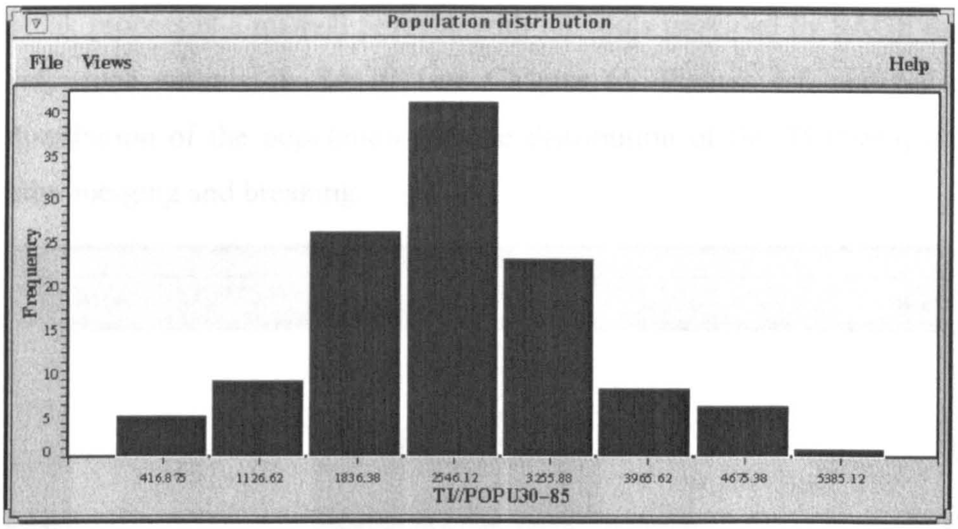


Figure 6.4. The frequency of 120 new areas with respect to the population.

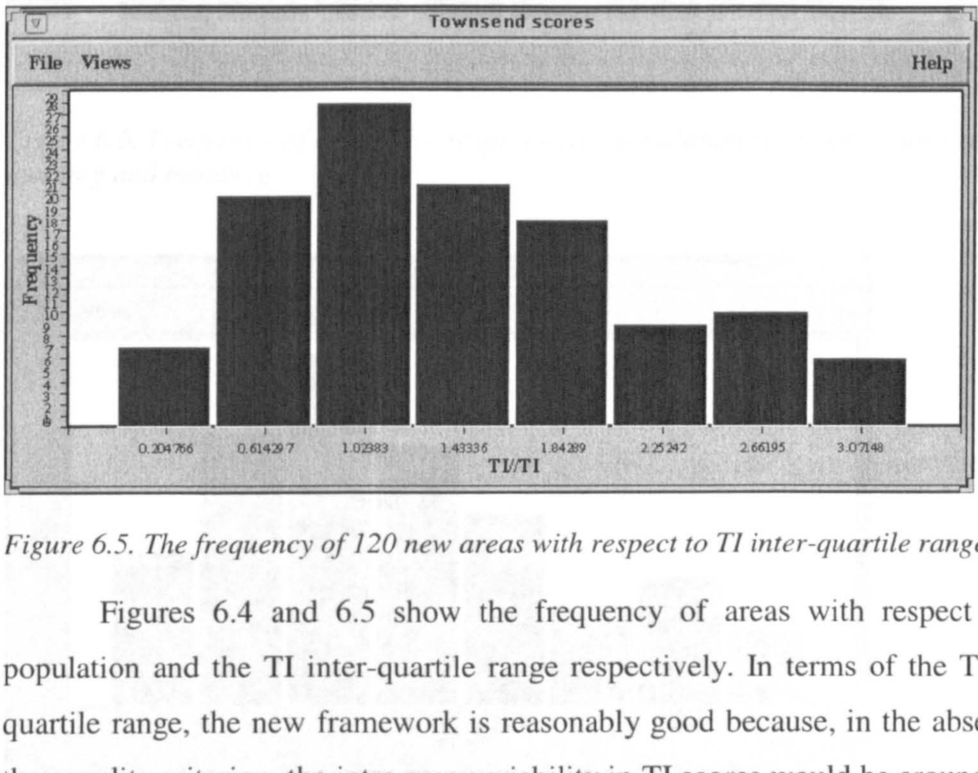


Figure 6.5. The frequency of 120 new areas with respect to TI inter-quartile range.

Figures 6.4 and 6.5 show the frequency of areas with respect to the population and the TI inter-quartile range respectively. In terms of the TI inter-quartile range, the new framework is reasonably good because, in the absence of the equality criterion, the intra-area variability in TI scores would be around 4.0 to 5.0.

However, in this framework, five areas had a population of less than 500 and one more than 5000. Therefore, a "merge" and "break" process was carried out to merge each of the 'small' areas with its adjacent areas and break up the 'large' area into a few smaller areas. This process was carried out for a number of iterations, and produced a framework with 119 areas. Although the merge-and-break process is a manual process, with the tools provided by SAGE this process was done entirely in SAGE (see Chapter 6). Figures 6.6 and 6.7 show the distribution of the population and the distribution of the TI inter-quartile range after merging and breaking.

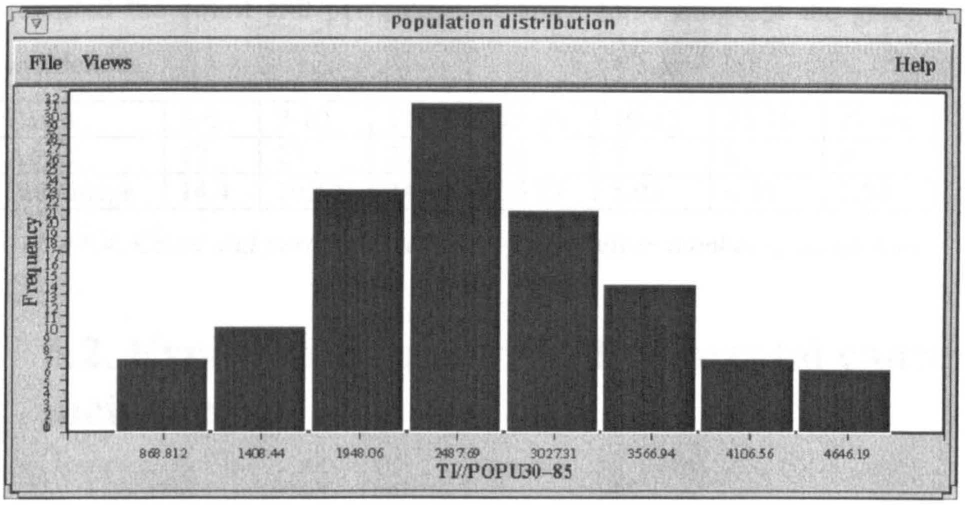


Figure 6.6. Frequency of areas with respect to the population in the new framework after merging and breaking.

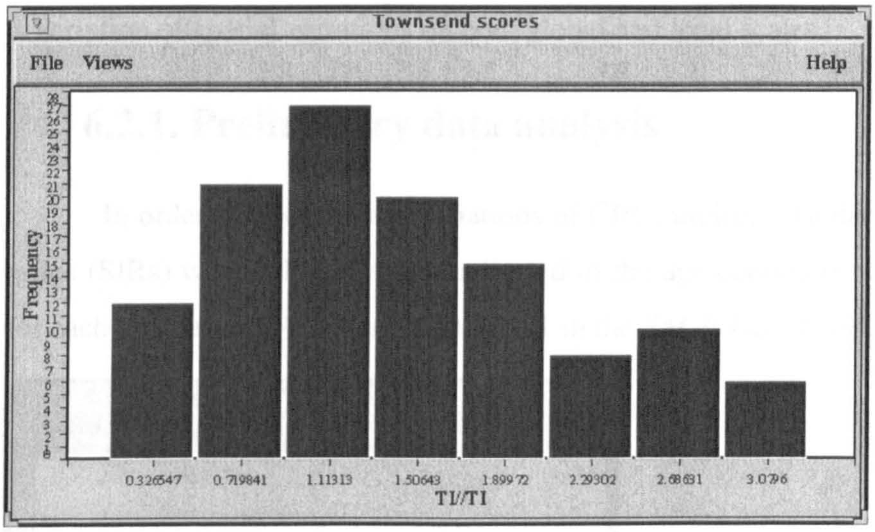


Figure 6.7. Frequency of areas with respect to the TI inter-quartile ranges in the new areas framework.

After the new framework had been constructed, CRC incidences were allocated to the new areas. There was one area without any incidence. Although it is perfectly valid to have this area, its existence will prevent the use of such techniques that involve a logarithm transformation on the data. Therefore, this area was merged with an area adjacent to it. The processes described above result in an areal framework with 118 areas. Based on this framework, ED-based counts and the number of the CRC incidences were aggregated using a SAGE tool, while the Townsend scores for areas were recomputed for every new area. Table 6.4 reported the count and percentage of areas for a range of the given number of incidences.

Cases	3-6	7-10	11-14	15-18	19-22	23-26	27-30
Areas	17	23	29	31	7	8	3
Percentage	14.4	19.49	24.58	26.27	5.93	6.78	2.54

Table 6.4. Count and percentage of areas for the given number of incidences.

6.2. Exploratory analysis of colorectal cancer incidence

This section shows how SAGE exploratory techniques can be used to analyse the spatial variations of the CRC data to answer the health questions of concern. It starts with a preliminary analysis of the CRC data, followed by the description of spatial variations on both global and local scales.

6.2.1. Preliminary data analysis

In order to describe the variations of CRC, indirect standardised incidence ratios (SIRs) were calculated, and adjusted to the age cohorts (see Section 4.3.1), for each area using SAGE (see Section 4.1 in the *SAGE User Guide*).

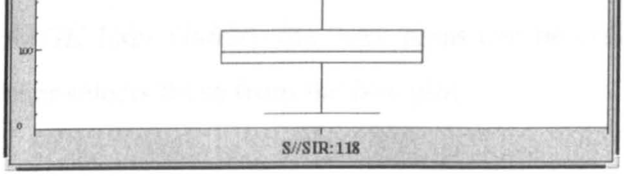


Figure 6.8. A box plot for the SIRs.

Figure 6.8 is a box plot of the SIRs across areas showing a skewed distribution of SIRs. There are five outliers all above the upper whisker. As discussed in Chapters 2 and 3, the interpretation of the SIRs may be complicated if the population varies across areas. Figure 6.9 shows an XY plot of the population against the SIRs across areas. Clearly, there is a negative linear relationship between them and so there is a need to down-weight estimates of the SIRs for less populated areas.

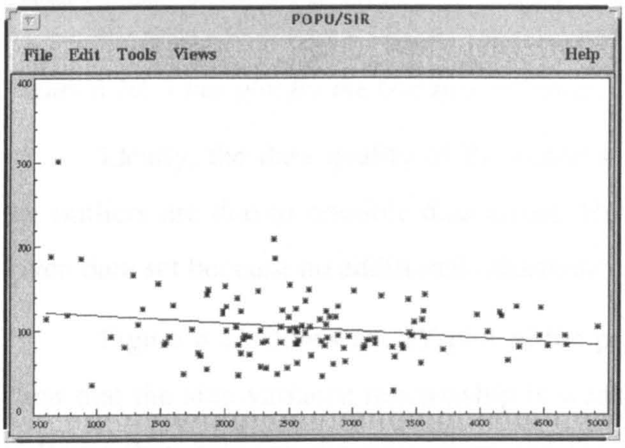


Figure 6.9. An XY plot of the population against the SIR.

Bayes adjusted estimates with a gamma model, called G-Bayes for short, were computed for each area (see Section 11.3.1 in the *SAGE User Guide*). The distribution of the G-Bayes estimates is shown in Figure 6.10. By comparing Figures 6.8 and 6.10, the effects of the Bayes adjustment can clearly be seen. Three out of five outliers above the upper whisker in the SIR box plot were shrunk because they correspond to areas with a small population of 635, 698 and 873 respectively. One area, which was not an outlier in the SIR box plot, became one below the lower whisker in the box plot of the G-Bayes. The outliers in the G-Bayes box plot correspond to area 6 (below the lower whisker) and 19 and 117 (above the upper whisker). Area 117 had the largest G-Bayes value. Since the plot and the map are hot linked, although this is not shown here (see Section 8.5 in *the SAGE User Guide*), the three areas can be easily identified on the map when the user selects them from the box plot.

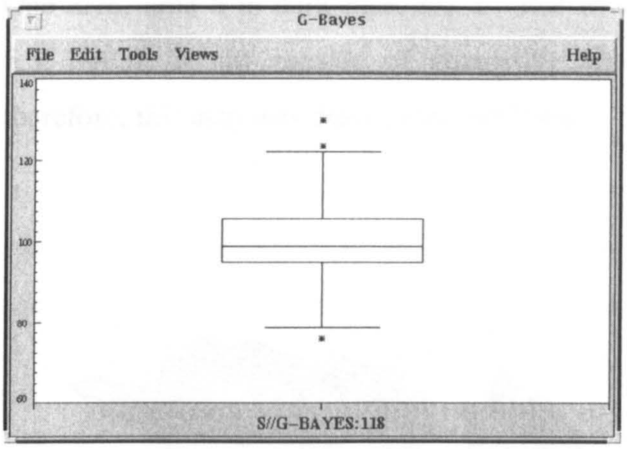


Figure 6.10. A box plot for the G-Bayes estimates.

Ideally, the data quality of the outliers should be checked to see whether the outliers are due to possible data errors. However, this cannot be done for the given data set because no additional information is available on data accuracy.

Figure 6.11 shows an XY plot of the population against the G-Bayes. It is clear that the size-variance relationship is weaker between the population and the G-Bayes than between the population and the SIRs. In other words, the G-Bayes values are more reliable for interpretation than the SIRs values.

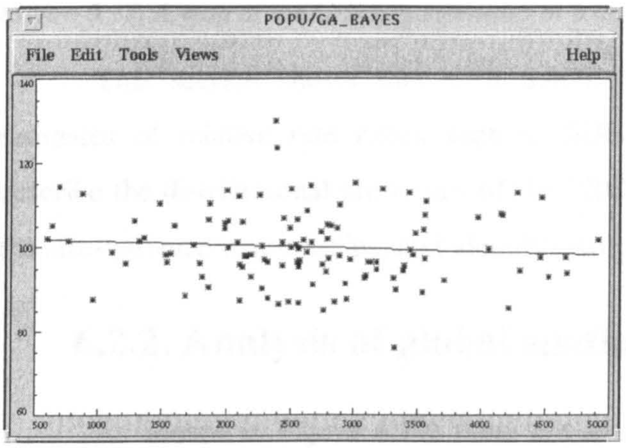


Figure 6.11. An XY plot of the population against the G-Bayes.

Figure 6.12 is a map of the G-Bayes classified into 6 classes indicated by the legend. It can be seen that from the middle to the west of the map, CRC is modest or low, but high in the east part, particularly the middle east part which comprises the rundown industrial areas and the city centre. Two areas, area 19 (in the middle) and 117 (at the bottom) drawn in red have values greater than 120.

One area, area 6 in dark blue, has a value less than 80. As discussed in Section 5.5, SAGE is not capable of generating choropleth maps with legends, and, therefore, this map was drawn with ArcView.

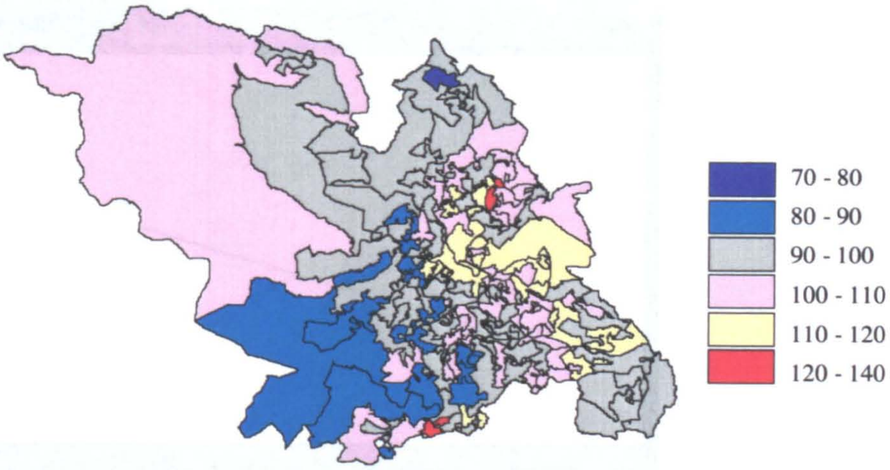


Figure 6.12. A map of the G-Bayes estimates in 6 classes.

This section shows that with SAGE tools the user can easily derive estimates of relative rate ratios such as SIRs and the G-Bayes and can also describe the distributional properties of the CRC data using statistical plots, assess the size-variance relationships, and identify outliers.

6.2.2. Analysis of global spatial properties

As shown in Figure 6.12, there are some indications that areas close to each other tend to have more similar G-Bayes than those far away. To investigate this, a spatial correlogram was drawn in Figure 6.13 where the values of Moran's I at the first to sixth orders (or lags) were plotted against the orders. The Moran's I values were positive for the first four orders but negative for the last two orders. Under the assumption of normal distribution (see Section 4.2.2.3), the probabilities of them are 0.000, 0.030, 0.006, 0.449, 0.003, and 0.011 respectively.

So the G-Bayes appeared to be positive autocorrelated at the first three orders and negatively autocorrelated at the last two orders but not autocorrelated at the fourth order. This indicates that the CRC incidence showed a patterning, which was possibly limited to local scales, and both the first and second order effects of variation may be responsible for the observed spatial variation.

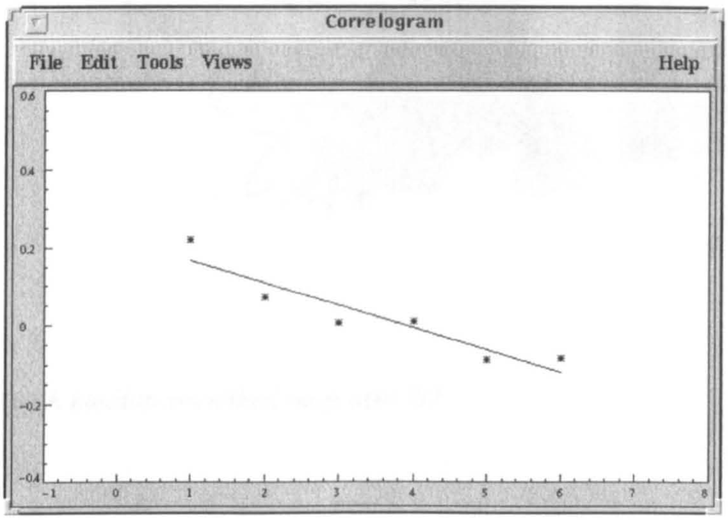
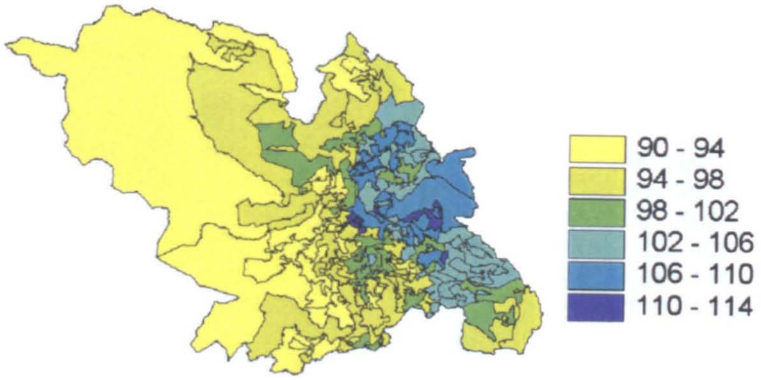


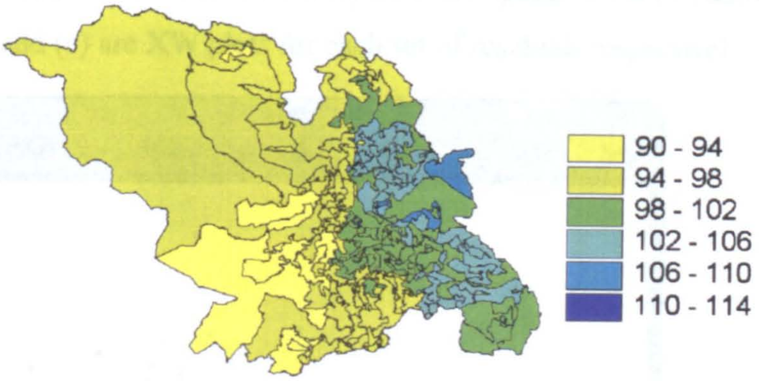
Figure 6.13. A spatial correlogram for the G-Bayes estimates.

Since SAGE does not provide a function for drawing the spatial correlogram directly, the following procedure was taken. First, adjacency matrices for each of the sixth orders were created using a SAGE tool (see Section 5.1 in the *SAGE User Guide*). It should be noted that the first order adjacency matrix is created automatically. Then, for each order, the Moran's I value for the G-Bayes was calculated using the SAGE Moran's I test function. Finally, these values were entered into the SAGE table and drawn using the SAGE XY plotting function.

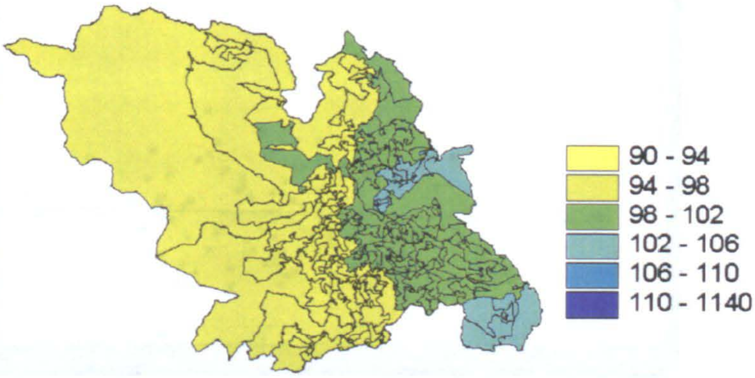
To highlight spatial trends of the CRC incidence, a median smoother was passed through the G-Bayes three times, each with a different window (See Section 11.3.2 in the *SAGE User Guide*). These windows were chosen to be the first, second and third order adjacency respectively, W1, W2, and W3 for short. For each window, a map of corresponding smoothed values was drawn. Figures 6.14 (a), (b) and (c) show the three maps. A west-to-east increasing gradient is clearly shown up in all three maps.



(a) A median smoothed map with W1.



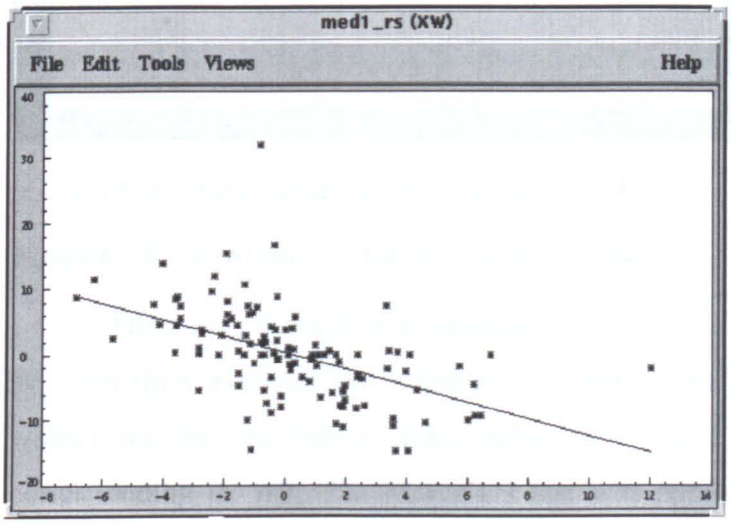
(b) A median smoothed map with W2.



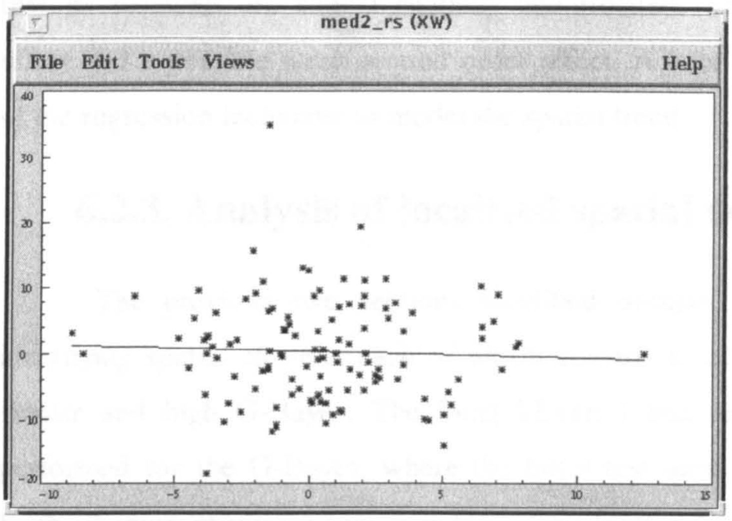
(c) A median smoothed map with W3.

Figure 6.14. (a), (b) and (c). Smoothed maps for the G-Bayes.

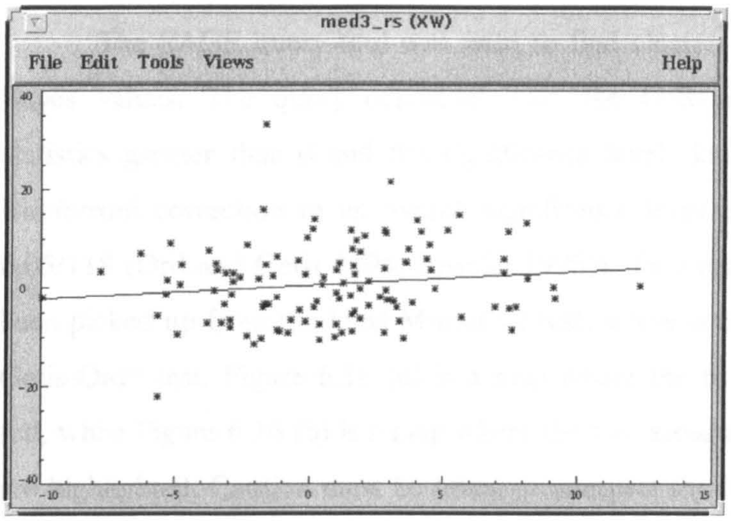
Three sets of residuals were extracted corresponding to the three sets of smoothed values and were examined for spatial autocorrelation. Figures 6.15 (a), (b), and (c) are XW plots for each set of residuals respectively.



(a) An XW plot for residuals of the smoother with W1.



(b) An XW plot for residuals of the smoother with W2.



(c) An XW plot for residuals of the smoother with W3.

Figure 6.15. (a), (b) and (c). XW plots for the residuals.

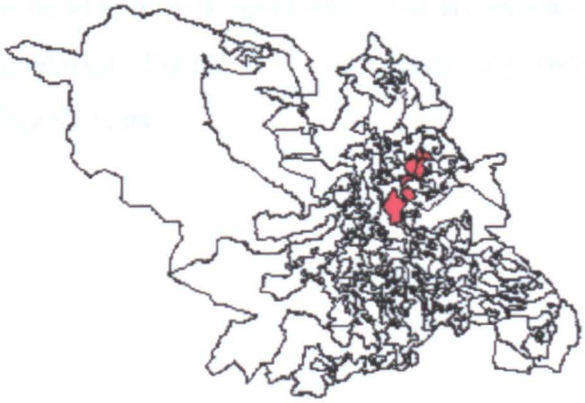
There was a negative relationship between the residuals corresponding to W1 and their averages on neighbouring areas. This relationship became much weaker for the residuals corresponding to W2, and positive for the residuals corresponding to W3. The Moran's I test was applied to the three sets of the residuals and gave values of -0.216, -0.020, and 0.064 respectively. Only the first is significance at 5%. Two points were identified to be outliers corresponding to areas 6 and 117.

The median smoother with the second order adjacency window may be considered to be the best simple representation of the spatial trends in the sense that its residuals are almost spatially uncorrelated. This indicates a first order effect and a possible weak second order effect. A later section will show the use of the regression technique to model the spatial trend.

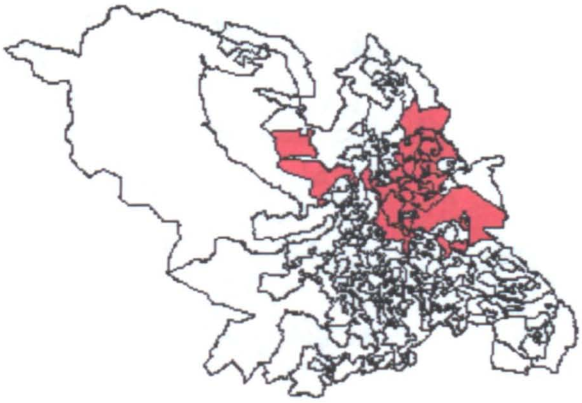
6.2.3. Analysis of localised spatial dependence

The previous two sections identified outliers. This section focuses on identifying spatial clusters, each of which consists a set of contiguous areas with similar and high G-Bayes. The local Moran I test and Getis-Ord* test, were performed for the G-Bayes, where the latter test used the first order adjacency. Each of them produced two variables - the local indicators and the significance levels of the indicators. The variables were stored in the SAGE table.

The SAGE query tool was used to find clusters with similar and large G-Bayes values. The query condition was: the G-Bayes greater than 110, the statistics greater than 0 and the significance levels less than 0.0004237 (after a Bonferroni correction to an overall significance level of 0.05 for 118 areas, i.e. 0.05/118 (Ord and Getis 1994, Anselin 1995)). Two areas, areas 17 and 19, have been picked up from the local Moran's I test, whereas none has been found in the Getis-Ord* test. Figure 6.16 (a) is a map where the two areas are highlighted in red, while Figure 6.16 (b) is a map where the two areas and areas adjacent to them are highlighted. Caution must be taken to interpret the maps because the presence of global spatial dependence may reduce the power of the test (Anselin 1995).



(a)



(b)

Figure 6.16. (a) and (b). Two areas picked up from the local Moran's I test and a possible cluster.

This section showed the use of SAGE tools to explore and analyse the spatial distribution properties of CRC. A correlogram of the G-Bayes (Figure 6.13) indicates the tendency of CRC to be positively spatially correlated at lower orders or short lags. The use of the smoothing technique reveals a spatial trend of CRC (Figure 6.14 (b)). There seems to be a first order effect on the spatial variation of CRC. No indication of a strong second order effect was shown. Area 6 and area 117 have been found to be outliers with low and high relative risks. By applying the local Moran's I test to the G-Bayes, area 17 and area 19 have been found to be significant and form a sub-regional cluster in the middle-east of the map as shown in Figure 6.16 (b). It is worth noting that techniques above would not be able to pick up clusters that are smaller than a single area or lie across area boundaries. Figure 6.17 summarise the two outliers picked up by the local Moran's tests.

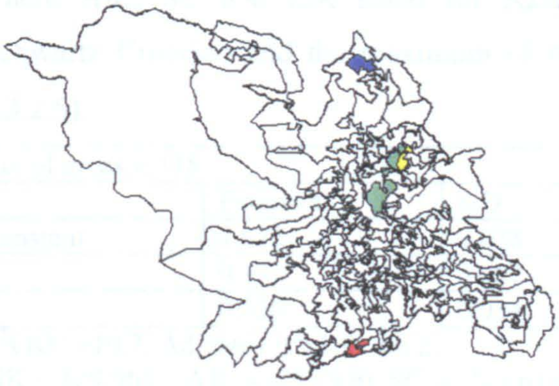


Figure 6.17. A summary map of the two outliers and the two areas picked up by the local Moran's I test. Areas 6, 117, 17 and 19 are coloured in blue, red, green and yellow respectively.

6.3. Statistical Modelling

The previous section showed how SAGE could be used to describe the spatial properties of CRC data. This section summarises how SAGE modelling techniques may be used to obtain a model-based representation of the spatial variation of CRC and the relationships between CRC and socio-economic deprivation.

6.3.1. Modelling spatial trends

Two surface models, the first and second order surface models (see Section 4.2.2.2), were fitted to the logarithm of the G-Bayes on the co-ordinates of area centroids. The co-ordinates were scaled to unity first. The choice of the logarithmic transformation was intended to stabilise the variance of the G-Bayes and help normalise the data. An attempt to fit a third order surface model failed because the coefficient matrix is co-linear.

Table 6.5 summarises ordinary least squares estimates for the first and second order trend surface models. The former appears to be a better fit than the latter. The former has an adjusted R^2 of 9.2, whereas the latter has 8.6, although the value of R^2 for the latter is slightly larger than that for the former. The former also has larger absolute values of AIC and SC but not of LIK than the latter has, where AIC, SC and LIK stand for Akaike's Information Criterion and the Schwartz Criterion and the maximum of the log likelihood (LIK) (see Section 4.3.2.5).

No. of areas = 118				
	Estimates	S.D.	t-value	Prob.
Constant	4.472	0.038	117.557	0.000
X	0.191	0.052	3.708	0.000
Y	0.036	0.034	1.063	0.290

R^2x100 =10.7; Adjusted R^2x100 = 9.2;
LIK = 129.965, AIC = -253.930 SC = -245.618;
Moran I = 0.116, Prob. = 0.013;
Shapiro-Wilk = 0.992, Prob. = 0.979;				

(a) Results for the firs order surface model.

No. of areas = 118				
	Estimates	S.D.	t-value	Prob.
Constant	4.451	0.147	30.304	0.000
X	0.155	0.408	0.379	0.705
Y	0.251	0.204	1.229	0.222
XX	0.039	0.267	0.147	0.884
YY	-0.195	0.128	-1.524	0.130
XY	-0.069	0.275	-0.251	0.802

R2X100 = 12.5; Adjusted R2x100 = 8.6;
LIK = 131.114, AIC = -250.389 SC = -233.765;
Moran I = 0.104; Prob. = 0.0075;
Shapiro-Wilk = 0.991, Prob. = 0.968;				

(b) Results for the second order surface model.

Table 6.5. (a) and (b). Ordinary least squares estimates and model diagnostics for the first and second order surface models

From Table 6.5 (a), it can be seen that the X term is highly significant in the first order surface model. This indicates there is a significant east-to-west component of variation in the CRC incidence (see also Figure 6.14). For the purpose of model diagnostics, the studentised residuals, the fitted values, and leverages were calculated. Figure 6.18 shows an XY plot of the studentised residuals and the fitted values. This plot did not indicate any problems regarding the form of the model. (Weisberg 1980, pp. 121). Two cases can be found to be outliers with the studentised residuals less than -3.0 and greater than 3.0 and correspond to areas 6 and 117 respectively.

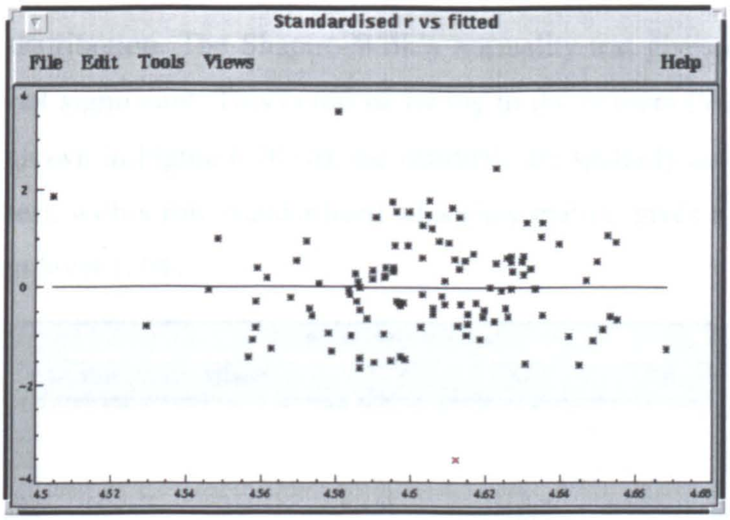


Figure 6.18. An XY plot for the studentised residuals and the fitted values.

When the leverages were examined, two areas, areas 1 and 4, were found to have leverage greater than (3*p/n) where n and p are equal to 118 and 3 (Haining 1990, pp. 311). This means that the fit of the model in this part of the map is based on a relatively small amount of information compared to the rest of the map. Figure 6.19 summarises these diagnostics for the first order surface model. Area 6 and 117, which are outliers, are shaded in red, while areas 1 and 4, with largest leverages, are shaded in blue.

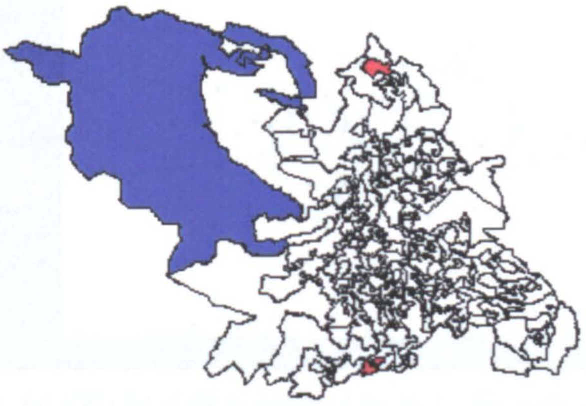
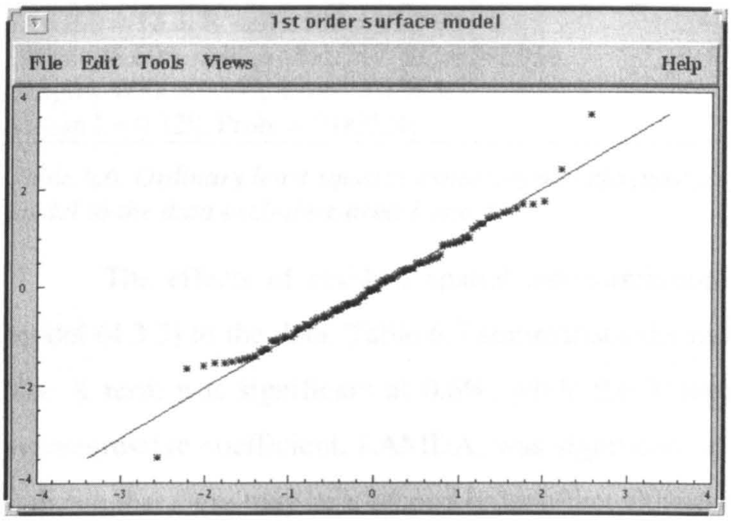
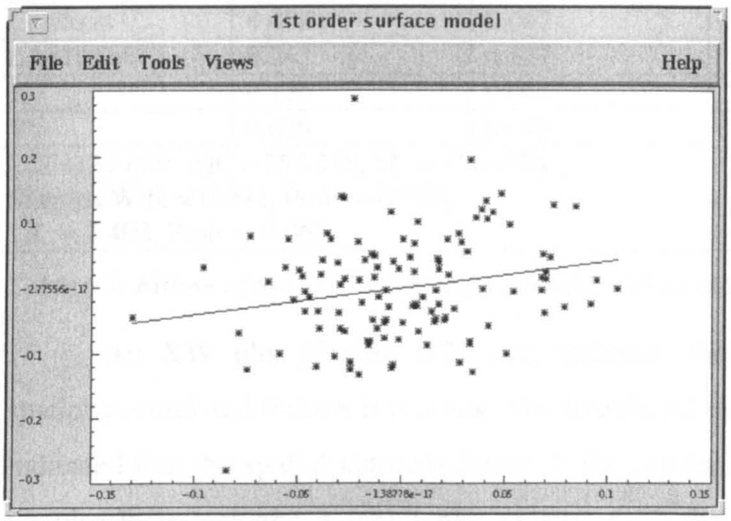


Figure 6.19. High leverage areas and outliers.

The normality and spatial dependence of the model residuals were assessed. Figures 6.20 (a) and (b) show a rankit plot and an XW plot for the residuals. As shown in Figure 6.20 (a), the residuals depart from the normal distribution. The Shapiro-Wilk's normality test gives a value of 0.992, which is not significant. This could be owing to the outliers (Weisberg 1980, pp. 134). As shown in Figure 6.20 (b), the residuals are spatially autocorrelated. The Moran's I test, with a row-standardised adjacency matrix, gives a value of 0.116, significant at level 1.3%.



A rankit plot of the residuals of the first order surface model.



(a) An XW plot of the residuals of the first order surface model.

Figure 6.20. (a) and (b). A rankit plot and an XW plot for the residuals of the first order surface model.

In order to assess the influences of the two large leverage areas on the fit of the model, the model was refitted with these two areas excluded. Table 6.6 summarises the fit of and diagnostics for this model. Estimates of model parameters were almost unchanged. Both R^2 and adjusted R^2 increased slightly.

The residuals still departed from the normal distribution as indicated by the Shapiro-Wilk test. The Moran's I test remained significant at less than 1%.

No. of Areas = 116				
	Estimates	S.D.	t-value	Prob.
Constant	4.451	0.039	112.675	0.000
X	0.231	0.056	4.154	0.000
Y	0.025	0.035	0.722	0.472

R^2 x100 = 13.3, R^2-adj x100 = 11.7
LIK= 128.608 AIC = -251.217, SC=-242.956
Shapiro-Wilk = 0.992, Prob. = 0.985;
Moran I = 0.128, Prob. = 0.00724;				

Table 6.6. Ordinary least squares estimates and diagnostics for the first order surface model to the data excluding area 1 and 4.

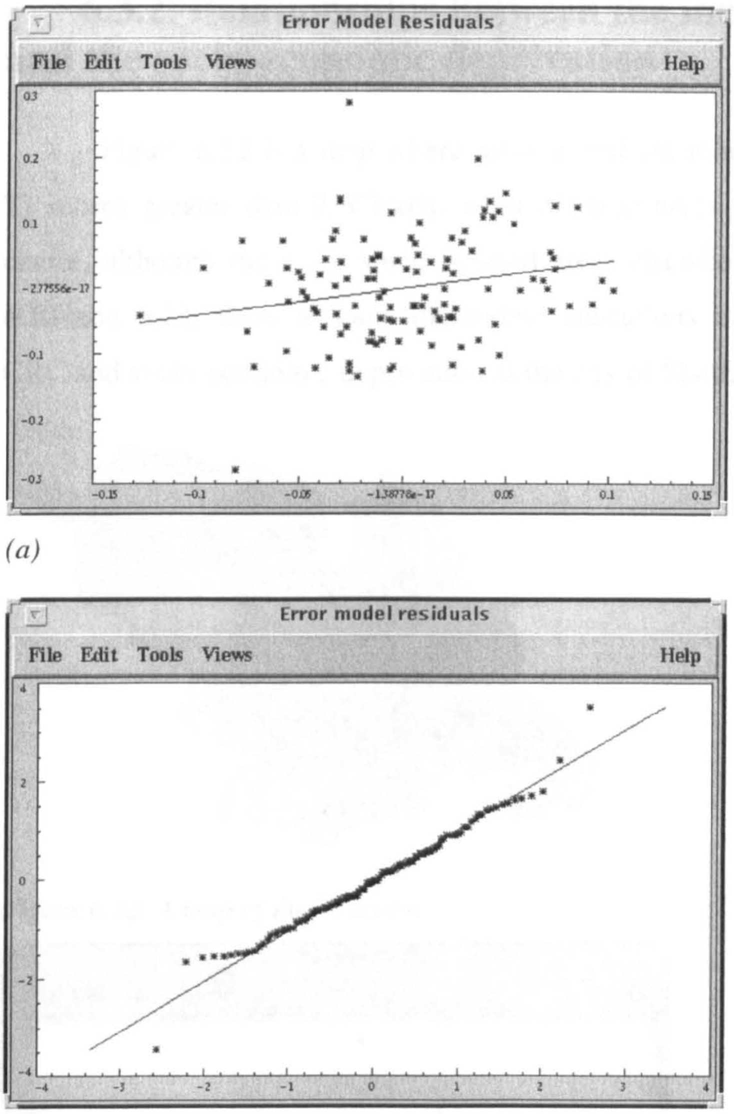
The effects of residual spatial autocorrelation were examined by fitting model (4.3.3) to the data. Table 6.7 summarises the model fitting and diagnostics. The X term was significant at 0.6%, while the Y term was not significant. The autoregressive coefficient, LAMDA, was significant at level 5.7%. This seems to indicate that there may be a second order effect although it seemed to be weak.

No. of areas = 118				
	Estimates	S.D.	Z-value	Prob.
Constant	4.484	0.047	95.111	0.000
LAMBDA	0.242	0.127	1.903	0.057
X	0.176	0.064	2.773	0.006
Y	0.029	0.042	0.702	0.483

LIK=131.669, AIC=-257.338, SC = -249.026;
Shapiro-Wilk = 0.992, Prob. = 0.978;
LR = 3.408, Prob. = 0.065;				

Table 6.7. Fitting of the first order surface model with spatially correlated error terms.

An XW plot (Figure 6.21 (a)) indicates that the residuals were still spatially correlated though it is weak. The likelihood ratio test was performed and indicated that the spatial autocorrelation of the residuals was not significant. The Shapiro-Wilk test and a rankit plot (Figure 6.21 (b)) indicated that the model residuals departed from the normal distribution. When the residuals of the XW plot were examined, no other areas were found to be spatial outliers besides areas 5 and 117.



(a)

Figure 6.21. (a) and (b). An XW and a rankit plots for the model residuals.

When the LIK AIC and SC statistics were compared, the first order surface model without and with spatially correlated error terms gave values (129.965, -253.930, -245.618) (see Table 6.5 (a)) and (131.669, -257.338, -249.026). The latter appears to fit better to the data than the former.

The analyses above show that CRC in the city of Sheffield tends to show a strong west-to-east trend with rates increasing from the relatively rural and suburban west to the inner city and industrial east. There appears to be a weak second order autocorrelation element to the spatial variation around that trend.

6.3.2. Relationships between the incidence of CRC and the socio-economic deprivation

Figure 6.22 is a map where areas in red are relatively deprived areas with TI scores greater than 0. Clearly, most of these areas are found around the city centre, although there are some isolated areas elsewhere. By comparing Figures 6.12 and 6.22, there is some qualitative indications of the association between CRC and socio-economic deprivation in the city of Sheffield.

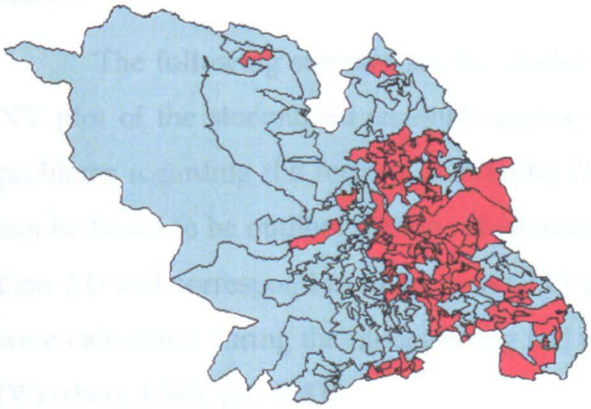


Figure 6.22. A map of the TI scores.

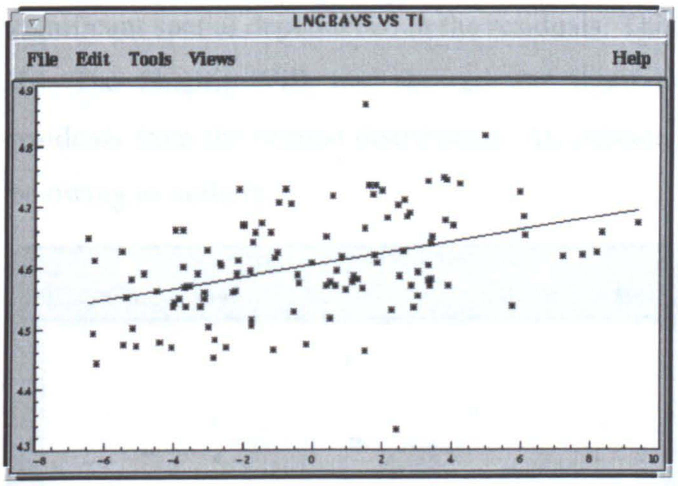


Figure 6.23 An XY plot the logarithm of G-Bayes (Y) and socio-economic deprivation TI (X).

In order to examine this, an XY plot is drawn for the logarithm of the G-Bayes on the TI scores as shown in Figure 6.23. The logarithmic transformation was performed for the same reasons as before (see Section 6.3.1). Table 6.8 reports the least squares fit indicated by the line in that figure. Both coefficients for the constant and TI were significant. The socio-economic deprivation explained only 16% of the variation of CRC.

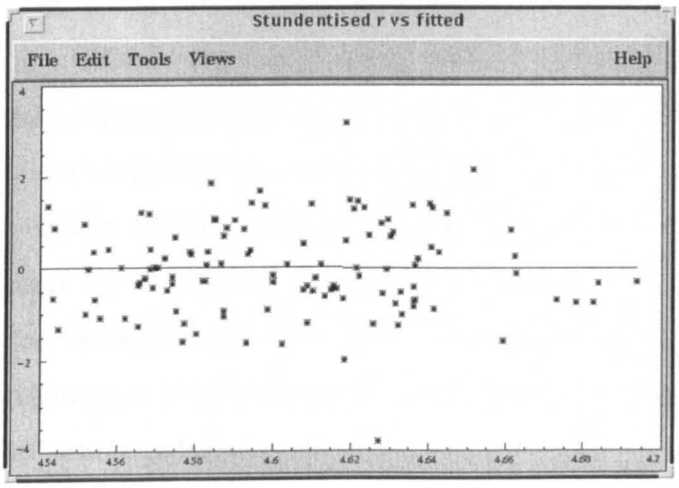
	Estimates	S.D.	t-value	Prob.
Constant	4.604	0.007	636.150	0.000
TI	0.010	0.002	4.725	0.000

R^2 x100 = 16.1, Adjusted R^2 x100 = 15.4;
Shapiro-Wilk = 0.990, Prob. = 0.944;
Moran I = 0.015, Prob = 0.636;
No. of area involved are 118;				

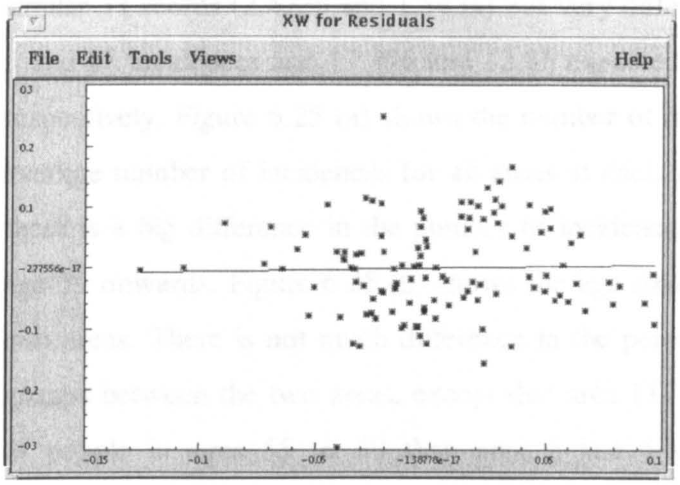
Table 6.8. The fit of the regression model of the logarithm of the G-Bayes on Townsend scores.

The following summarises the model diagnostics. Figure 6.24 (a) shows a XY plot of the studentised residuals against the fitted values. It did not indicate problems regarding the form of the model (Weisberg 1980, pp. 121). Two cases can be found to be outliers with the studentised residuals less than -3.0 and greater than 3.0 and correspond to areas 6 and 117 respectively. Cook-distance statistics were calculated during the model fitting and none of them can be considered large (Weisberg 1980, pp. 108).

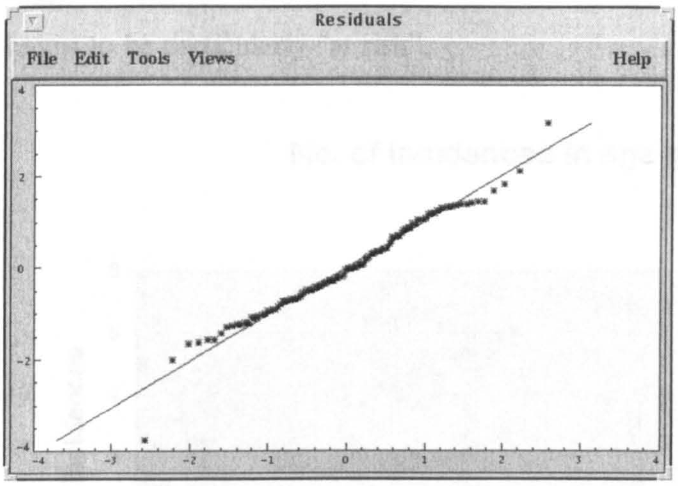
The Moran's I test on the model residuals indicated that there is no significant spatial dependence in the residuals. This was also shown in Figure 6.24 (b). The Shaprio-Wilk test showed that there was a serious departure in the residuals from the normal distribution. As indicated by Figure 6.24 (c), this might be owing to outliers.



(a). A XY plot for the studentised residuals and the fitted values.



(b). An XW plot for the model residuals.

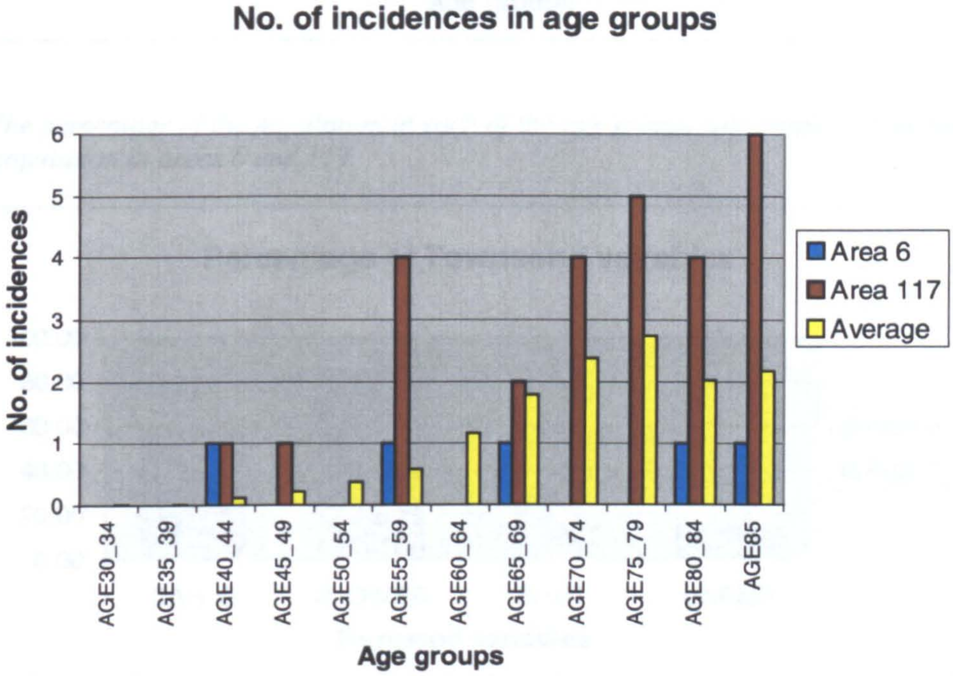


(c). A rankit plot for the model residuals.

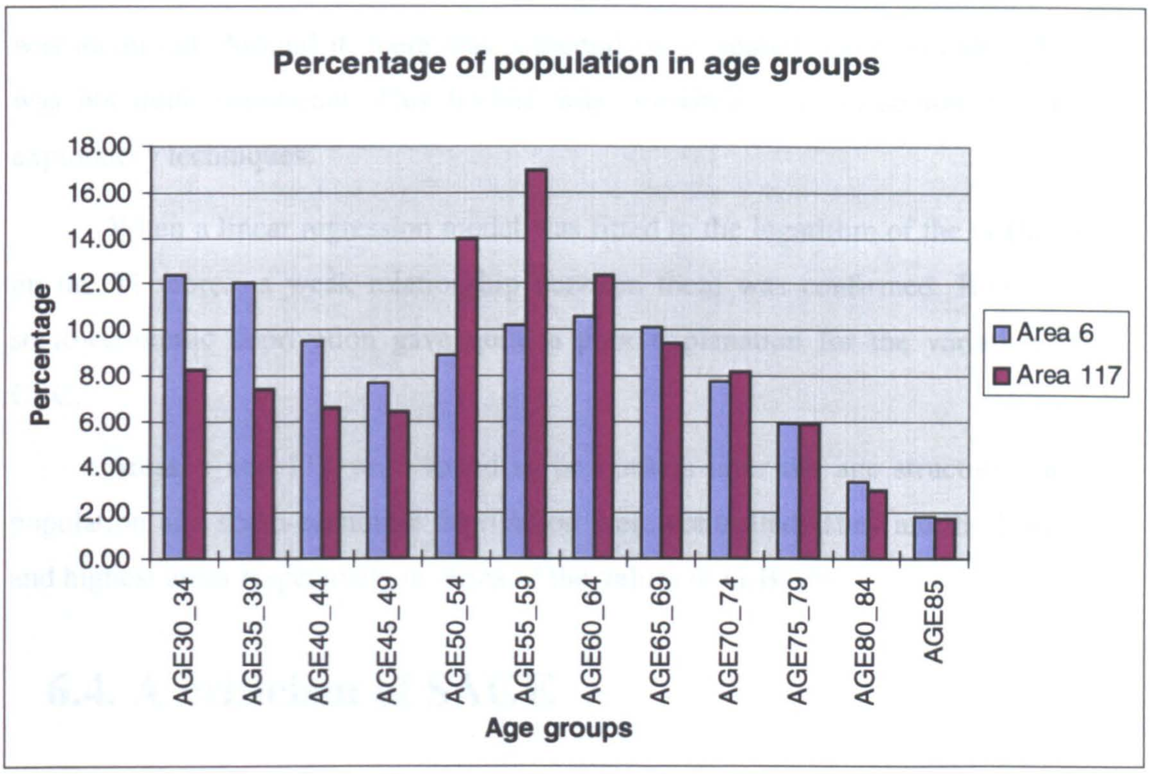
Figure 6.24. Diagnostic plots for the model.

Given the results above, there is some evidence of an association between the incidence of CRC and socio-economic deprivation, although it is not strong. It is not surprising that socio-economic deprivation offers little explanation for the variation of CRC because colorectal cancer is likely to be the outcome of many other factors beyond socio-economic deprivation, which cannot be assessed with the data available to this study. This agrees with the finding of the previous work (Haining *et al* 1994) that CRC is not predominately a disease of either deprived or affluent populations.

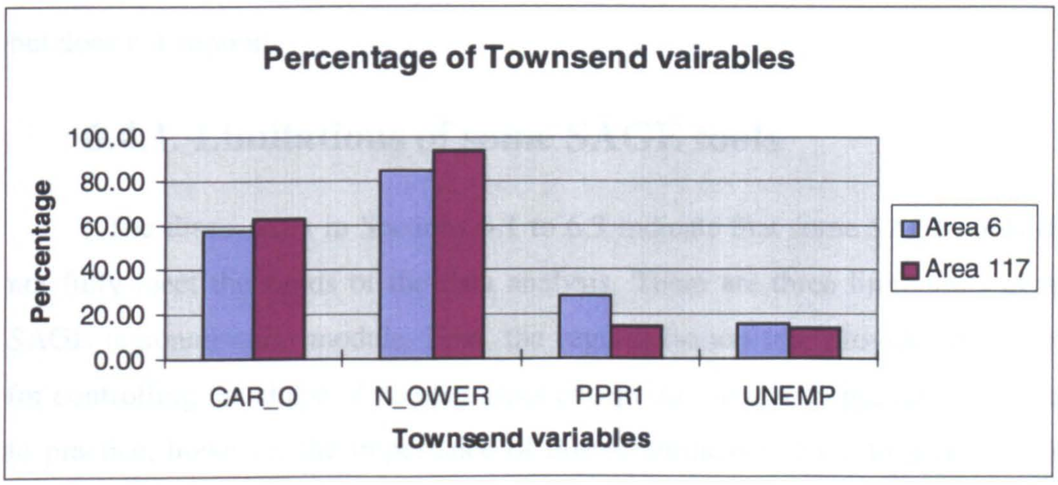
Areas 6 and 117 were identified to be outliers throughout the analyses above, so it may be of interest to look at them more closely. Areas 6 and 117 have similar TI scores (2.4235 and 1.5434) but very different incidence rates. There are 5 and 27 incidences and 17.496 and 12.86 expected incidences in areas 6 and 117 respectively. Figure 6.25 (a) shows the number of incidences for the areas and the average number of incidences for all areas at each age cohort. It can be seen that there is a big difference in the number of incidences between the two areas from age 55 onwards. Figure 6.25 (b) shows the age structures of the population in the two areas. There is not much difference in the percentage of populations in those groups between the two areas, except that area 117 has a much higher percentage of people in ages 55 to 59 than area 6 has. Figure 6.25 (c) shows that the percentage of the four Townsend variables between two areas is similar. This suggests that in addition to an area effect there is a particular age cohort that seems to be particularly "at risk".



(a) The number of CRC cases in each of the age groups between areas 6 and 117.



(b) The percentage of the population in each of the age groups with respect to the total of the population in areas 6 and 117.



(c) The percentage of the four Townsend variables between areas 6 and 117.

Figure 6.25. Comparison for areas 6 and 117.

This section has reported the use of SAGE to model the spatial variations of CRC and the relationships between CRC and socio-economic deprivation. A first order surface model with spatially correlated error terms was found to give a better representation of the spatial variation of CRC. A strong west-to-east trend was identified. Around it, there was a second order spatial variation although it was not quite significant. This finding was consistent with those found using exploratory techniques.

When a linear regression model was fitted to the logarithm of the G-Bayes on the TI scores, a weak relationship between them was confirmed. However, socio-economic deprivation gave quite a poor explanation for the variation of CRC.

Areas 6 and 117 were found to be outliers after the age structures, the population and socio-economic deprivation were controlled. They are the lowest and highest areas respectively in terms of the values of G-Bayes.

6.4. A criticism of SAGE

Sections 6.1 to 6.3 mainly showed how useful SAGE is for analysing CRC data. This section will consider the limitations of SAGE tools - what SAGE can and should do better, and limitations in SAGE functionality - what SAGE should but does not support.

6.4.1. Limitations of some SAGE tools

The discussions in Sections 6.1 to 6.3 indicate that some SAGE tools did not fully meet the needs of the data analysis. There are three limitations in the SAGE regionalisation module. First, the regionalisation tool provides no means for controlling the shape of merged areas except through a compactness criterion. In practice, however, the importance of this criterion may have to give away to that of other criteria and it is difficult to find a set of weighting factors appropriate for all criteria. As a result, the regionalisation tool may not be able to prevent elongated and curved areas from being formed. This problem is quite clearly in a part of the whole map shown in Figure 6.26. The odd-shaped areas would make a map difficult to interpret because the 'neighbours' may have to be interpreted in an unusual way. For example, two fringe areas (also see Figure 6.16) in the far north-west of the cluster would not be regarded as neighbouring areas in the usual sense. Thus a result may be anti-intuitive and so contradict common knowledge. Geddes and Flowerdew (2000 and see the conference presentation) report their work of designing policy-relevant regions for Europe carried out as part of the EuroStat project (EuroStat 1997). They employed the SAGE regionalisation tool and found problems similar to that above. One way that may help prevent odd-shaped areas from being generated may be to impose an upper limit on the number of areas in each group.

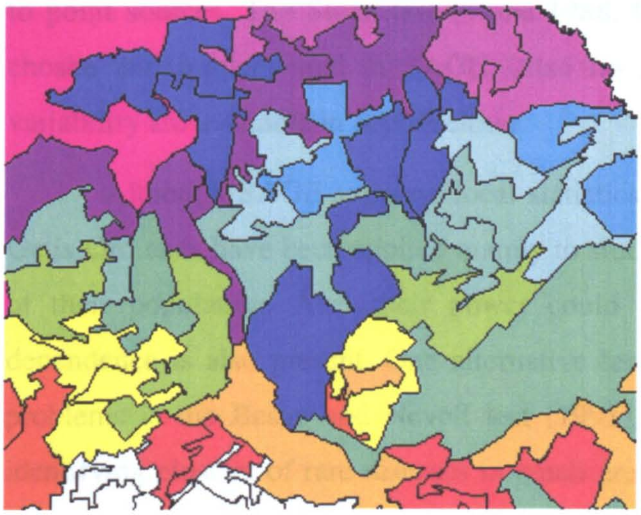


Figure 6.26. A map showing the part of the region with elongated areas.

Second, SAGE does not offer tools for the purpose of 'breaking' and 'merging' areas. At present, the user has to identify areas from all eligible areas to which a 'small' area can be merged and to measure the consequence of a 'merge' not in terms of the objective function, but rather in terms of the variability of values in each individual group. If many iterations of 'breaking' and 'merging' were required, doing this manually would be difficult. This was shown to be the case in this study. A tool would be useful if it could automatically identify areas to which a selected small area can be merged and to report them along with relevant properties in a certain order. Therefore, the user can pick appropriate areas and manipulate them.

Third, SAGE provides no 'intelligent' support for deriving appropriate weighting factors for regionalisation. Since the user often has no prior knowledge of which space he/she should look for the parameters, many runs of the regionalisation module may be wasted. Indeed, this may undermine the potential of the SAGE regionalisation as a swift tool for exploring spatial variation.

6.4.2. Limitations in SAGE functionality

This case study also indicated that there are some functions that should have been supported by SAGE. SAGE does not support the drawing of grouped bar plots. SAGE does not provide any function for testing clustering with respect to point sources. The Stone test (Stone 1988, Bithell and Stone 1989) could be chosen and implemented in SAGE. Also no statistical tests for extra-Poisson variability are available in SAGE either.

Although SAGE supports local statistics, the local Moran's I test and the Getis-Ord tests have been applied mainly to studies where areas are large in terms of their population. Also their power could be affected when global spatial dependence is also present. One alternative test that does not suffer from these problems is the Besag and Nevell test (1991), which is particularly useful for identifying clusters of rare diseases in small areas. This test is essentially a point pattern analysis technique treating incidences as points, but it is able to make use of censuses at EDs for estimating the population at risk. It can be implemented in SAGE if each disease incidence is considered to occur at the centroid of the area in which it falls. This test may be useful for identifying clusters of colorectal cancer incidences at EDs. A preliminary implementation of this test has been made available in the latest version of SAGE.

This section summarised limitations of some SAGE tools and gave some suggestions for remedying them. It also pointed out some SSA functions that can widen the SAGE application areas if they were to be implemented in SAGE.

6.5. Conclusion

The chapter presents an evaluation of SAGE through a case study of colorectal cancer incidences in the city of Sheffield. The SAGE regionalisation functions were shown to be useful in constructing a regional framework to overcome the small number problem, to reduce the heteroscedasticity problem and to control the intra-area variability in socio-economic deprivation across the areas at the same time. SAGE was shown to be able to compute estimates of relative risks such as SIRs and the G-Bayes - Bayes estimates with a Gamma prior. Both SAGE exploratory and confirmatory tools were shown to be capable of identifying spatial variations on global and local scales.

During this case study, a west-to-east spatial trend of CRC was identified, as was a spatial pattern around the trend. Two areas were found to have large and similar values compared with their adjacent areas respectively. These two areas and their adjacent areas formed a possible cluster. A relationship was found between CRC and socio-economic deprivation, although it was not strong. Two areas, areas 6 and 117 were found to be small and large outliers respectively even when the age structures, the inter-area variability in the population and the deprivation were allowed for.

This chapter also considered the limitations of SAGE. The main limitations exist in regionalisation. Possible ways forward to remedy these limitations were considered. A set of techniques thought to be useful to extend SAGE capability to cover a wider range of health studies was also suggested.

Data used in this case study, excluding CRC data for individuals, is enclosed in the CDROM attached to this thesis along with SAGE software. The next chapter will sum up each chapter discussed so far and consider directions for future work.

CHAPTER 7. SUMMARY, CONCLUSIONS AND FUTURE WORK

7.1. Summary and Conclusions

This thesis has shown how the two research objectives, constructing the SAGE software package to enable the user to undertake a coherent analysis of area-based health-related data, and evaluating and demonstrating the capabilities of this package, have been met. It also shows the systematic approach that has been used in this research project to identify and address related questions.

Chapter 2 reviewed two sets of general issues concerned with the analysis of area-based data and the provision of SSA in a GIS environment. Section 2.1 considered briefly statistical methodologies for analysing area-based data which recognise the features of such data characterised by spatial dependence and heterogeneity, heteroscedasticity, area configuration, and data accuracy. It emphasised the role of ESDA not only in identifying spatial properties but also in assessing CSDA for the purposes of model evaluation and diagnostics. Section 2.2 discussed why these features exist in area-based data and what problems they may raise for the analysis of such data and for the interpretation of results derived from the analysis.

Section 2.3 considered the strengths and the weaknesses of GIS for the analysis of spatially referenced data. It argued that GIS is an appropriate vehicle for integrating SSA techniques. Section 2.4 summarised some issues concerned with the system integration and common approaches to integrating SSA and GIS. It reviewed some recent work in this area and examined the advantages and disadvantages of different approaches.

Chapter 3 considered the subject field - health research - for which SAGE has been developed and the important roles SSA and GIS played in the field.

Section 3.1 summarised some common types of health studies in the areas of the spatial epidemiology and the health services research and typical questions they ask. Some examples have been drawn to show how SSA may help address these and the kinds of problems likely to arise especially in a small area study of rare disease. One particular aspect of these types of health studies emphasised in this section is that both ESDA and CSDA techniques need to be called in order to fulfil a coherent study.

Section 3.2 considered some major sources of routinely collected mortality and morbidity statistics and measurements on the social and physical environment. Data confidentiality was considered along with the practical implications it raises. One of them is that data collected originally from individuals may not necessarily be available to the public at individual level but only in the form of aggregates at area levels. In this case, lattice data analysis is likely to be more appropriate than point pattern analysis. Even if some data is available at individual level to projects for research purposes, the researcher is not allowed to publish the results in forms that may lead to individuals being identified. Issues of health-related data quality were discussed including the ways in which errors enter data sets and their possible impacts on small area studies of rare disease.

Section 3.3 considered the role of GIS in health research and argued the kind of GIS suitable for health research. It showed how GIS functions for managing, manipulating and visualising spatial data may be used to answer some of the common health questions that would be difficult to answer otherwise. It argued that GIS is likely to be most useful for health research if it provides a range of SSA techniques not only capable of answering each health question but also with various statistical complexity suitable for multidisciplinary users.

The discussion given in Chapter 3 leads to a review of specific lattice data analysis techniques useful for health research in Chapter 4. These techniques were classified into three groups in respect of data preparation, the analysis of univariate data and the analysis of multivariate data. Section 4.1 considered data preparation techniques for constructing regional frameworks and specifying inter-area relationships. Reasons for regionalisation were given. Three criteria - homogeneity, equality and compactness- were considered to be important in health studies. A K-means-based regionalisation procedure that is able to combine these criteria was discussed in detail.

Section 4.2 focused on some useful SSA techniques for analysing area-based data of a single variable. It showed how EDA techniques like histogram and box plots can be linked with maps to serve the purposes of identifying distribution properties with reference to their geography in the context of dynamic visualisation. It also showed many ESDA and CSDA techniques for the purpose of identifying spatial properties, including empirical Bayes estimation techniques for deriving reliable estimates of relative risks, map-smoothing techniques based on kernel estimation, techniques for identifying global and local spatial dependence as well as spatial outliers. A surface modelling technique was also considered.

Section 4.3 considered some SSA techniques for analysing multivariate data. Besides classical correlation and regression analysis techniques, it discussed a number of spatial regression models, catered specially for spatial data, along with model fitting and inference procedures as well as model diagnostics.

Chapters 3 and 4 together addressed one of two questions related to the first objective, that is, identifying lattice data analysis techniques for health research. This leads to the consideration, in Chapter 5, of the second question - how to link these techniques together and to integrate them with a GIS - ARC/INFO - seamlessly to form SAGE for health research. Sections 5.1 to 5.4 summarised the development of SAGE at the following stages: system analysis, modelling, design, and implementation and testing. Section 5.1 summarised functional requirements for SSA, which fall into three classes: DV (data visualisation), DAM (data analysis and modelling) and DM (data management). It discussed the reasons of choosing ARC/INFO for integration and examined those functions that ARC/INFO does and does not support. Some non-functional requirements on the behaviour of SAGE were also considered.

Section 5.2 discussed the modelling of SAGE. It argued that a three-tier client-server model is the most appropriate for SAGE. The SSA component functions as the client and communicates with the server - ARC/INFO - through the middle tier called the linking agent. This model results in a SAGE architecture that makes it possible to meet the system requirements. Five classes of client requests and corresponding server replies have been identified. Many of the requests are rather generic and may be applicable in other integrations.

Section 5.3 summarised how each SAGE component has been designed under a networked system where the computers run UNIX and the X Window systems. Reasons for choosing this environment were considered. A three-layer structure was thought appropriate for the client component. A GUIs module forms the top layer of this structure enabling the user to interact with other system functions. The second layer, in the middle, comprises DV and DAM modules. The DV module provides data models for a number of statistical plots, while the DAM module offers a range of data analysis and modelling functions. The DM module, at the bottom, manages access to attribute data and W matrices, sends the requests to the server and handles corresponding server replies for other modules. It is also responsible for maintaining the 'hot linking' of different windows. This design separates core SSA functions in the DAM and DV modules from other ancillary functions provided by the GUIs and the DM module, and therefore increases the reusability of the DAM and DV modules. The linking agent consists of a client portion handling client requests and server replies on the client side, and an agent portion performing actual linking operations on the server. The DM module calls the API to communicate with the linking agent. The linkage between two portions is underpinned by the RPC mechanism. A pipe-based structure provides a base for the server and the linking agent to communicate with each other.

Section 5.4 summarised some important aspects involved in the implementation and testing of SAGE. An incremental procedure for the implementation was illustrated, as were some important techniques and the development of the tools and packages employed.

Section 5.5 considered two specific drawbacks of SAGE as a direct result of the system design and implementation. One of these is that SAGE cannot support techniques that require highly dynamic hot-linking like brushing. This owes to the fact that ARCPLOT that provides SAGE mapping functions runs as part of the server rather than the client. Another is that SAGE is incapable of generating choropleth maps with legend because of taking a decision to reduce the complexity of mapping facilities. Also SAGE is unable to save results for part of the study region in repositories. This limitation is owing to the difficulty in recording and maintaining inter-area relationships based on which the results are obtained. It also noted that the development of SAGE was done in an *ad hoc* manner and using a non-object-orient approach which together may degrade the usability of the SAGE client.

Sections 5.1 to 5.5 elaborated SAGE from a software developer perspective in order to share the experience gained from this work with other developers. Section 5.6, on the other hand, illustrated SAGE from a user point of view. It shows the user some basic SAGE functions often required for performing the ESDA and CSDA of area-based data.

Having discussed how the first objective of this research has been accomplished in Chapters 2 to 5, Chapter 6 presented a small-scale case study of a rare disease - colorectal cancer for the city of Sheffield, UK, using SAGE. The purpose of doing this, in line with the second objective of this research, was to evaluate and demonstrate the capabilities of SAGE. The main focuses of this study were to describe and model spatial variation in the incidence of colorectal cancer and to seek the relationships between CRC incidence and socio-economic deprivation.

Section 6.1 considered the background to this study and the questions in which health professionals may be interested. It summarised two data sets used in this study - postcoded colorectal cancer (CRC) cases for the years 1979 to 1983 and ED-based small area statistics from the 1981 UK census. Both data sets have been used in a previous study and are in the form of ARC/INFO coverages. Because the data exhibits the small number problem and the possibility for heteroscedasticity at the ED level, the SAGE regionalisation function was called to construct a regional framework on the EDs. Equality in the population of sufficient size was involved to deal with the small number problem and partly heteroscedasticity. Homogeneity in socio-economic deprivation was employed to minimise the loss of information due to area aggregation.

Based on the regional framework illustrated in Section 6.1, Section 6.2 showed how SAGE may be used to perform exploratory data analysis. The SAGE techniques were employed to derive standardised incidence ratios and the relative risk ratios using a Gamma-based empirical Bayes estimation which further compensates for the effect of heteroscedasticity associated with still unequal population size, and to describe both distribution properties and spatial properties of CRC incidence. The Bayes estimates showed a west-to-east trend when a spatial median filter was applied to them. They also showed spatial dependence at low orders of adjacency. Two areas were found to have similar and large values with their neighbours under the local Moran's I test, and they and their neighbouring areas form a possible cluster. Two areas were found to be outliers.

Section 6.3 showed how SAGE modelling tools can be used in conjunction with ESDA tools to obtain a model-based representation of the spatial variation of the CRC data and to model the relationships between the CRC data and socio-economic deprivation. A first order trend surface turned out to be representative. As shown with this model, there is a west-to-east trend and possibly a pattern around the trend. The results drawn from fitting a linear regression to the logarithm of the Bayes estimates on the Townsend scores confirmed the existence of a very weak relationship between the CRC incidences and deprivation.

Section 6.4 considered some practical weaknesses of SAGE, some of which were uncovered in this case study. The SAGE regionalisation tools were found not to be intuitive and difficult to use to create ideal regional frameworks where areas are compact in terms of the shapes. Besides, a number of functions that are not supported yet by SAGE but useful for health research were mentioned.

Clearly, this work shows that there is a range of spatial statistical techniques useful for health professionals and researchers. Because of their simplicity, the ESDA techniques would appeal to a wide spectrum of professionals, while CSDA techniques to a minority which, nevertheless, could be significantly large in some specific fields such as spatial epidemiology. It demonstrates an approach to integrating these techniques with a GIS to meet demands for functions not only to handle but also to analyse spatially referenced data effectively. The SAGE development encouraged one to consider the construction of a generic SSA module for not only ARC/INFO but also for other GIS packages. It also extended our understanding of what SSA techniques are needed for GIS and how they may be integrated with GIS in a cost-effective way.

The first version of SAGE was made available for academic use in 1997 and was updated in November 1999. The CDROM enclosed in this thesis contains SAGE source code and executable code (for SUN workstations with the Solaris 2.5 or higher and ARC/INFO 7.0.x) as well as part of the data used in the case study.

Before finishing the thesis conclusion, a brief comparison of SAGE (version 1.0) and SpaceStat is given to show their relative strengths and weaknesses. The reason to compare these two systems is that they offer a similar range of ESDA and CSDA functions and SpaceStat is well known in regional science. All the information given on SpaceStat can be found from the following documents (http://www.spacestat.com/about.htm, Anselin 1992, Anselin 1999).

SpaceStat here refers to a cluster of four packages: the latest version of SpaceStat, version 1.90, ArcView, and the SpaceStat extension for ArcView and DynESDA – an ArcView extension. Both extensions, together with ArcView, extend the SpaceStat package beyond CSDA to ESDA. The SpaceStat extension for ArcView provides some specialised mapping functions, as well as some data handling functions, while DynESDA offers capabilities for drawing statistical plots and maintaining the 'hot-linking' between plots and ArcView maps. It should be noted that, unlike SAGE in which client and server share the same data at any time, SpaceStat's components rely on the user to manually invoke import and export functions for data exchange.

The two systems are compared based on the functional classification used in Chapter 4. Under this classification, a function may be considered to fall into one of three categories: data preparation, analysis of a single variable and analysis of multivariate variables, to fulfil one or more tasks. For a comparison of the ESDA functionality of SAGE and SpaceStat, along with two others (cdv (Dykes 1996) and MANET (Unwin *et al* 1996)), from a data visualisation point of view, the reader is referred to Wise et al (1999).

CATERGORY	TYPES OF FUNCTIONS		SAGE	SPACESTAT
Data preparation	Classification and regionalisation		Yes*	No
	Specifying inter-area relationships		Yes	Yes
	Creating new and manipulating existing variables		Yes	Yes
Analysis of a single variable	Identifying non-spatial properties		Yes*	Yes
	Identifying spatial properties	Trend (with and without reference to a specific location)	Yes*	Yes
		Global association	Yes	Yes
		Local clusters and outliers	Yes	Yes
Analysis of multivariate variables	Correlation analysis		Yes	Yes*
	Regression modelling	Classic model	Yes	Yes
		Models for spatially heterogeneous data	Yes	Yes*
		Model with spatially correlated errors	Yes	Yes*
		Models for spatial interactions	Yes	Yes*
	Model evaluation and diagnostics		Yes	Yes*

Table 7.1.A summary of functions supported by SAGE and Space Stat.

Table 7.1 summarises the types of functions that SAGE and SpaceStat support where an asterisk (*) indicates that one package has a significantly wider range of tools or techniques for doing the corresponding task than the other package. In the category of data preparation, SpaceStat, unlike SAGE, does not support classification and regionalisation[3]. An implication of this is that SpaceStat may not be able to tackle the heteroscedasticity problem, the small number problem and computational intractability problem (which arises when data volumes are large) through regionalisation as discussed in Chapter 6[4]. Both systems provide a set of functions useful for specifying W matrices and creating and manipulating variables, although they differ slightly in sophistication. For example, SpaceStat allows the user to choose area contiguity according to whether two areas share the same corner (Rook criterion), the same boundary (Bishop

[3] ArcView does offer some simple classification functions for creating choropleth maps.

[4] Note that SpaceStat provides a model with heteroscedasticity error terms, but the heteroscedasticity problem may also arise in testing spatial autocorrelation using the Moran's I test.

criterion), or both (Queen criterion), and stores a W matrix in a sparse form. On the other hand, SAGE allows the user to create a new variable by specifying an algebraic expression on existing variables with a set of functions (e.g. power and logarithm).

In the category of the analysis of a single variable, both systems provide similar functions for describing non-spatial properties of a variable. Both systems employ histogram, box plots and choropleth maps for summarising non-spatial and spatial properties. Unlike SpaceStat (which allows many maps to be created by ArcView) SAGE allows only one map to be created by ARCPLOT at a time. Both systems offer smoothing functions for exploring spatial trends based on both kernel and empirical Bayes estimation and support trend surface modelling. In addition, SAGE provides a function, the lagged box plot defined from a use specified site, for the exploratory identification of spatial trends. SpaceStat does not offer any similar function. For testing global spatial dependence, SAGE provides the Moran's I test and the Getis-Ord test, but SpaceStat also provides for example the Geary test. For identifying local spatial clusters, both systems support the local Moran's I test, Gi and Gi* and allow the results to be saved for further analysis. As far as mapping such results is concerned, SpaceStat will overlap two boxes or a pie chart (of two regions) on each area on a map for the visual comparison of spatial dependence. The more similar the height of the two boxes or the size of the two regions in the pie chart, the stronger the value of that area is correlated with the values of the neighbouring areas. For identifying spatial outliers, SAGE provides the XW plot of $(x_i, (Wx)_i)$, whereas SpaceStat uses the Moran's I plot of $((Wx)_i, x_i)$ for the same purpose. The advantage of using the latter is that the slope of the regression line (related to the Moran's I) gives some indications of the global spatial association. The former is a scatter-plot for fitting the model $x_i=\beta_0 + \beta_1(Wx)_i + e_i$ (the spatially lagged response variable model) and is, therefore, a natural choice for identifying local spatial clusters and outliers.

In the category of the analysis of multiple variables, SpaceStat provides not only functions for analysing correlation (e.g. drawing XY scatter plots and computing Pearson correlation coefficients supported also by SAGE), but also additional functions for analysing principal components and multivariate spatial autocorrelation. Both systems support the classic regression model and spatial models with autocorrelated error terms, lagged independent and/or dependent variables. In addition, SpaceStat offers a spatial model with heteroscedastic error terms. Moreover, SpaceStat allows the spatial regime and expansion to be taken into account in model specification, whereas SAGE allows for the former only through introducing a dummy variable. SpaceStat does not, however, support any generalised regression model with non-normally distributed error terms. By contrast, SAGE provides a generalised regression model with Poisson errors. Unlike SAGE, which employs only the maximum likelihood (ML) estimation for model parameters, SpaceStat employs a number of different fitting procedures.

For model evaluation, SAGE offers mainly ESDA functions, drawing rankit, XW and XY plots of the corresponding data, to check statistical assumptions although it also provides a few tests for doing this and checking the significance of a spatial model with respect to the corresponding classic model. By contrast, SpaceStat offers many statistical tests for doing these and some of them can be used to test a spatial model with autocorrelated errors, heteroscedastic errors or spatial lagged dependent variable when one of these terms may already be present in the spatial model to be tested against. For model diagnostics, SAGE can compute studentised and standardised residuals, leverages and Cook's distances for the classic model, while SpaceStat computes the last two only.

In summary, SAGE and SpaceStat support a set of similar functions for ESDA. SAGE offers more types of statistical plots but less types of maps than SpaceStat does. Such an enhancement in ESDA to the SpaceStat package is attributed to the recent development of two ArcView extensions. The advantage of SpaceStat over SAGE in CSDA is clearly evidenced.

7.2. Future Work

There are at least three areas on which future work may focus. First, there is a need to enhance SAGE in order to remedy some weaknesses of SAGE identified in Chapter 6. Some suggested SSA techniques should and could be implemented in SAGE to extend its SSA functionality. Improving SAGE regionalisation is a high priority. Research is required to look at the ways to enable prioritisation of different criteria. As acknowledged SAGE is unable to find a set of appropriate weighting factors to use when more than two criteria need to be considered. How to prevent 'snake-like' areas from being formed could be another but related research field. In order to achieve this, further understandings to the behaviour of the SAGE regionalisation function are required. Research is also needed to look at how to handle attributes for part of a study region. As argued previously, the main problem is to record and maintain the circumstances of the spatial references on which the attributes are obtained. Unless this problem is resolved, such data could not be guaranteed to be used correctly in subsequent analysis even if they were stored in the repositories.

Second, there is a need to explore the possibility to enhance the SAGE cartographic capability. Currently, SAGE relies on ARCPLOT to perform choropleth mapping. However, ARCPLOT is a package too complex and complicated for dynamic visualisation. This is made even worse in SAGE where ARCPLOT is run as part of ARC/INFO. This has limited SAGE to offer swift update of the map. Indeed, SAGE is not suitable to implement techniques requiring highly dynamic visualisation. Clearly, what is really needed for SAGE is an efficient cartographic package that can work on the SAGE client side. As acknowledged in Chapter 5, the latest version of ARC/INFO offers a mapping component that can be implemented. This component is much better catered for the Microsoft Windows environment than for the UNIX owing to the well-developed component object model (COM) in the former. In this regard, one point needs to be made clear. That is, the chosen package should conform to a standard application interfaces (API) in order to make it possible to work with other components which also conform to the same API. The Open GIS Consortium (OGC) is currently developing such an API that deserves a close examination (McKee and Kottman 1999).

Third, it is necessary to develop a pluggable SSA component for different

GIS packages. The efforts of OGC in developing standard API currently endorsed by many GIS companies will guarantee the API to be supported by major GIS products. This provides an opportunity to develop a SSA component that can be plugged into any GIS. Given the three-layer structure used for the SAGE client, it would not be difficult to modify the DM module so that it can access GIS data management and manipulation operations, including data query, through the standard API. However, since a 'real' pluggable SSA is likely to be plugged into an existing GIS application system, further application interfaces are required. Thus work is needed to specify APIs for GIS packages to support SSA. This might be seen to be important because, unless an appropriate API can be identified for the application systems, the SSA techniques implemented in SAGE are likely to remain separate from most other GIS packages.

REFERENCES

1. Abel, J., P. J., Kilby, and J. R. Davis (1994) "The system integration problem". *International Journal of Geographical Information Systems*, Vol. 8, No 1, pp. 1 -12.

2. Abel, J., S.K. YAP, R. Ackland, M. A. Cameron, D.F. Smith and G. Walker (1992) "Environmental decision support system project: an exploration of alternative architecture for geographical information systems". *International Journal of Geographical Information Systems*, Vol.88, No 3, pp. 193 -204.

3. Acheson, D. (1998) *Independent inquiry into inequalities in health*. Stationary Office, London, 1998.

4. Akaike, H. (1981) "Likelihood of a model and information criteria". *Journal of Econometric*, Vol. 16, pp. 3-14.

5. Alexander, F. E. (1991) "Investigation of localised spatial clustering and extra-Poisson variation". In G. Draper *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain*, 1966-83, Studies on Medical and Population Subjects No 53. HMSO, London, pp. 69-76.

6. Alexander, F. E. and J. Cuzick (1992) " Methods for the assessment of disease clusters". In P. Elliott, J. Cuzick, D. English and R. Stern (EDs) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies*, Oxford University Press, Oxford, pp. 238-247.

7. Anderberg, M. R. (1973) *Cluster analysis for applications*. Academic Press, New York.

8. Anselin, L. (1988) *Spatial Econometrics: Method and Models*. Kluwer Academic Publishers, Dordrecht.

9. Anselin, L. (1990) *SpaceStat: A Program for the Statistical Analysis of*

Spatial Data. Dept. of Geography, University of Califomia, Santa Barbara.

10. Anselin, L. (1992) "SpaceStat TUTORIAL - A Workbook for Using SpaceStat in the Analysis of Spatial Data", University of Illinois, Urbana-Champaign, Urbana, IL 61801, http://www.spacestat.com/.

11. Anselin, L. (1995a). "Local Indicators of Spatial Association -LISA". *Geographical Analysis*, Vol. 27, No. 2, pp. 93-115.

12. Anselin, L. (1999) "Spatial Data Analysis with SpaceStat and ArcView Workbook" (3 rd Edition), Department of Agricultural and Consumer Economics, University of Illinois, Urbana, IL 61801 http://www.spacestat.com/.

13. Anselin, L. and S. Bao. (1997) "Exploratory spatial data analysis linking SpaceStat and Arc View". In M. Fischer and A. Getis (EDs) *Recent Developments in Spatial Analysis: Spatial statistics, behavioural modelling and neuro-computing*. Berlin, Springer-Verlag, p35-59.

14. Anselin, L. and Can, A. (1986) "Model comparison and model validation issues in empirical work on urban density functions". *Geographical Analysis*, 18 pp. 179-97.

15. Anselin, L. and D.A. Griffith, (1988) "Do Spatial Effects Really Matter in Regression Analysis?". *Paper of the Regional Science Association*, 65, pp. 11-34.

16. Anselin, L and O. Smimov (1996) "Effective algorithms for constructing proper higher order spatial lag operators". *Journal of Regional Science*, Vol. 36, No. 1, pp. 67-89.

17. Arbia, G. (1986) "The modifiable areal unit problem and the spatial autocorrelation problem: towards a joint approach". *Metron*, 44. Pp. 325-41.

18. Arbia, G. (1991) "GIS-based sampling design procedures". In *proceedings EGIS' 91 Second European Conference on Geographical Information Systems*, Brussels, 2-5, April 1991, 1, (Eds) by J. Harts, H. Offens and H. Scholten, pp. 27-35.

19. Bailey, T. C. (1994) "A Review of Statistical Spatial Analysis in Geographical Information Systems". In Fotheringham S and Rogerson P. (EDs) *Spatial Analysis and GIS*. Taylor and Francis, London. pp. 11-44.

20. Bailey, T. C. and A. C. Gatrell (1995) *Interactive spatial data analysis*. Longman, Essex, UK.

21. Batty, M. and Y. Xie (1994) "Urban Analysis in a GIS Environment: Population Density Modelling using ARC/INFO". In Fotheringham, S. and Rogerson, P. (Eds) *Spatial Analysis and GIS*, Taylor and Francis, London, pp. 189-220.

22. Banfield, C. F. and L. C. Bassill (1977) "Algorithm AS 11: a transfer algorithm for non-hierarchical classification". *Applied Statistics*, Vol. 26, pp. 206-210.

23. Ben-Shlomo, Y. and N. Chaturvedi (1995) "Assessing equity in access to health care provision in the UK: does where you live affect your chances of getting a coronary artery bypass graft". *Journal of Epidemiology and Community Health*, 49, pp. 200-204.

24. Benzeval, M. K. Judge (1996) "Access to health care in England: continuing inequalities in the distribution of GPs". *Journal of Public Health Medicine*, Vol. 18, No. 1, pp. 33-40.

25. Benzeval, M. K. Judge and M. Whitehead (1995) *Tackling Inequalities in Health*. The King's Fund, London.

26. Besag, J. (1989) "Contribution to the discussion on cancer near nuclear installations". *Journal of the Royal Statistical Society, Series A*, 152, pp. 367-368.

27. Besag, J. and J. Nevell (1991) "The detection of clusters in rare disease". *Journal of the Royal Statistical Society, Series A*, 154, pp. 143-155.

28. Bithell, J. (1990). "An application of density estimation to geographical epidemiology". *Statistics in Medicine*, Vol. 9, pp. 691-701.

29. Bithel, J. F. and R. A. Stone (1989) " On statistical methods for analysing the

geographical distribution of cancer cases near nuclear installations". *Journal of Epidemiology and Community Health*, 43, pp. 79-85.

30. Bithell, J. F., S. J. Button, N. M. Neary, and T. J. Vincent (1995) "Controlling for socio-economic confounding using regression methods", *Journal of Epidemiology and Community Health*, 49 (Suppl. 2), pp. S15-19.

31. Black, D (1984) *Investigation of the possible increased incidence of cancer in West Cumbria.* HMSO, London.

32. Black, D. and J. N. Morris, C. Smith and P. Townsend (1980) *Inequality in health: report of a research working group.* Department of Health and Social security, London.

33. Boyle, P., C. S. Muir, and E. Grundmann (1989) *Cancer mapping*, Recent results in cancer research; 114, Springer-Verlag, Berlin.

34. Boyle, P., J. A. Gatrell and O. Duke-Williams (1999) "The effect on morbidity of variability in deprivation and population stability in England and Wales: an investigation at small-area level". *Social Science and Medicine*, Vol. 49, No. 6, pp. 791-799.

35. Brunsdon, C. (1998) "Exploratory spatial data analysis and local indicators of spatial association with XLISP-STAT". *The Statistician*, 47(3), pp. 471-484.

36. Brunsdon, C., A. S. Fotheringham and M. E. Charlton (1996) "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity". *Geographical Analysis*, Vol. 28, No. 4, pp. 281-298.

37. Bullen, N., G. Moon and K. Jones (1996) "Defining localities for health planning: a GIS approach". *Social Science and Medicine*, Vol. 42, pp. 801-816.

38. Carstairs, V. (1981) "Small area analysis and health services research". *Community Medicine* 3(2), pp. 131-139

39. Carstairs, V. and M. Lowe (1986) "Small area analysis: creating an area base for environmental monitoring and epidemiological analysis". *Community*

Medicine 8(1), pp. 15-28.

40. Carstairs, V. and R. Morris (1989) "Deprivation, mortality and resource allocation". *Community Medicine,* Vol. 11, No. 4, pp. 364-372.

41. Carstairs, V. and R. Morris (1991) "Deprivation and health in Scotland". Aberdeen University Press.

42. Chou, H. C., and Y. Ding (1992) "Methodology of integrating spatial analysis/modelling and GIS". In *Proceedings of 5th International Symposium on Spatial Data Handling*, Charleston, South Carolina. pp. 514-523.

43. Choynowski, M. (1959) "Maps based on probabilities". *Journal of the American Statistical Association*, 54, pp. 385-388.

44. Clayton, D. and J. Kaldor (1987). "Empirical Bayes Estimates of Age-standardised Relative Risks for Use in Disease Mapping". *Biometrics* 43, pp. 671-681.

45. Cleveland, W.S. and McGill, M. E. (1988) *Dynamic Graphics for Statistics*, Wadsworth and Brooks/ Cole, Pacific Grove, CA

46. Cliff, A. D., P. Haggett, J.K. Ord, K. Bassett, and R. B. Davies (1975) *Elements of Spatial Structure: A Quantitative Approach.* Cambridge University Press, Cambridge.

47. Cliff, A. D. and K. Ord (1981) *Spatial Processes: Models and Applications*, Pion, London

48. Cliff, A. D., P. Haggett and J.K. Ord (1985) *Spatial Aspects of Influenza Epidemics*, Pion, London.

49. Clifford, P. and Richardson, S. (1985) "Testing the association between two spatial processes". *Statistics and Decisions, Suppl.* No 2, pp. 155-60.

50. Collins, S (1998) "Modelling Spatial Variations in Air Quality using GIS". In A. C. Gatrell and M. Löytönen (Eds.) *GIS and Health,* GISDATA IV, Taylor & Francis, London, pp. 81-96.

51. Cook, D. G and S. J. Pocock (1983) "Multiple Regression in Geographical Mortality Studies, with Allowance for Spatial Correlation Errors".

Biometrics 39, pp. 361-371.

52. Cook, D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman & Hall.

53. Cook, D., J. J. Majure, J. Symanzik and N. Cressie (1996) "Dynamics graphics in a GIS: exploring and analyzing multivariate spatial data using linked software". *Computational Statistics,* Vol. 11, 467-480.

54. Cook-Mozaffari *et al* (1989) "Geographical variation in mortality from Leukaemia and other cancer in England and Wales in relation to proximity to nuclear installation 1969-1978". *British Journal of Cancer,* 59, pp. 476-85.

55. Cressie, N. A. C. (1993) *Statistics for Spatial Data.* (Revised Edition) John Wiley & Sons, New York 1993.

56. Cromley, R. G. (1996) " A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data". *International Journal of Geographical Information System,* Vol. 10, No. 4, pp. 405-424.

57. Cuzick, J and P Elliott (1992) "Small-area studies: purpose and methods". In P. Elliott, J. Cuzick, D. English and R. Stern (EDs.) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies.* Oxford University Press, Oxford, pp. 10-21.

58. Cuzick, J. and R. Edwards (1990) "Test for spatial clustering of events for inhomogeneous populations". *Journal of the Royal Statistical Society B,* 52, pp. 73-104.

59. Dangermond, J. (1984) "A classification of software components commonly used in geographic information systems". In Peuquet, D.J. and Marble, D.F. C. (Eds.) *Introductory readings in Geographic Information Systems,* pp. 30 – 51.

60. de Lepper, C. H J. Scholten and R. M. Stern (1994) *The Added Value of Geographical Information Systems in Public and Environmental Health,* Kluwer, Dordrecht.

61. Department of the Environment (1987) *Handling Geographic Information.*

Report of the Committee of Enquiry chaired by Lord Chorley. HMSO, London

62. Diggle, P. T., A. C. Gatrell and A. A. Lovett (1990) "Modelling the prevalence of cancer of the larynx in part of Lancashire: a new methodology for spatial epidemiology". In Thomas, R. W. (Ed) *Spatial Epidemiology*, London papers in regional science 21, a Pion Publication, pp. 35-47.

63. Diggle, P.J. (1990) "A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of prespecified point". *Journal of Royal Statistical Society A,* 153, pp. 349-62.

64. Ding, Y. and S. Fotheringham (1992) "The Integration of Spatial Analysis and GIS". *Computers, Environment and Urban Systems*, 16, pp. 3-19.

65. Draper, G. J., C. A. Stiller, C. M. O'Connor and T. J. Vincent (1991) "Introduction and objectives". In G. Draper *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain*, 1966-83, Studies on Medical and Population Subjects No 53. HMSO, London. pp. 1-6.

66. Draper, G. (1991) *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain*, 1966-83. Studies on Medical and Population Subjects No 53. HMSO, London.

67. Dykes J. (1996) "Dynamic maps for spatial science: a unified approach to cartographic visualisation". In D.Parker (Ed), *Innovation in GIS 3*, Taylor and Francis, London, pp. 177-187.

68. Eames, M. Y Ben-Shlomo, and M. G. Marmot (1993) "Social deprivation and premature mortality: regional comparison across England". *British Medical Journal.* Vol. 307, pp. 1097-1102.

69. Elliott, P., J. A. Beresford, D. J. Jolley, S. H. Pattenden and M. Hills (1992) "Cancer of the larynx and lung near incinerators of waste solvents and oils in Britain". In P. Elliott, J. Cuzick, D. English and R. Stern (Eds.) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies.* Oxford University Press, Oxford, pp. 359-367.

70. Elliott, P., J. Cuzick, D. English and R. Stern (1992) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies*, Oxford University Press, Oxford.

71. Everitt, B. S. (1979) "Unresolved problems in cluster Analysis". *Biometrics*, Vol. 35, pp.169-181.

72. English, D. (1992) “Geographical epidemiology and ecological studies”. In P. Elliott, J. Cuzick, D. English and R. Stern (Eds.) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies*, Oxford University Press, Oxford, pp. 1-9.

73. ESRI (1995) ARC/INFO online documentation. Environmental Systems Research Incorporated, Redlands CA.

74. ESRI (1996) *ArcView 2.1 The Geographic Information System for Everyone*, Redlands, Environmental Systems Research Institute, CA.

75. Eurostat by Lancaster University North West Regional Research Laboratory & Department of Mathematics and Statistics, Sheffield University Sheffield Centre for Geographic Information & Spatial Analysis and Nene College Nene Centre for Research (1997) “Final report to Geographical Information System for the Commission (GISCO)”, Eurostat for project SUP.COM 95, Lot 15 'Geographical Information Systems in Statistics'.

76. Ferris, G., P. Roderick, A. Smithies, S. George, J. Gabbay, N. Couper and A. Chant (1998) “An epidemiological needs assessments of carotid endarterectomy in an English health region. Is the need being met?” *British Medical Journal* 317, pp. 447-451.

77. Flowerdew, R. and M. Green (1991) “Data Integration: Statistical Methods for Transferring Data Between Zonal Systems”. In I. Masser and M. Blackmore (Eds.), *Handling Geographical Information*, Longman, London, pp. 38-54.

78. Fotheringham, A.S. (1999) “Trends in quantitative methods III: stressing the visual”. *Progress in Human Geography*, Vol. 23, No. 4, pp. 597-606.

79. Fotheringham A. S. and D. W. S. Wong (1991) "The modifiable areal unit problem in multivariate statistical analysis". *Environmental Planning,* A, 23, pp. 1025-1044.

80. Fotheringham, A. S., M. Charlton and C. F. Brunsdon (1996) "The Geography of Parameter Space: An Investigation into Spatial Non-stationarity". *International Journal of Geographical Information Systems*, Vol. 10 No. 5, pp. 605 – 627.

81. Fotheringham S and P. Rogerson (1993) "GIS and spatial analytical problems". *International Journal of Geographical Information Systems*, Vol. 7, No. 1, 3-19.

82. Frank, A. U. (1988) "Requirements for a database management system for a GIS". *Photogrammetric Engineering and Remote Sensing*, Vol. 54, pp. 1557-64.

83. Gatrell, A. C. (1989) "On the spatial representation and accuracy of address-based data in the United Kingdom". *International Journal of Geographical Information Systems*, Vol. 3, No. 4, pp. 335-348.

84. Gatrell, A. C. (1997) "Structures of Geographical and Social Space and Their Consequences for Human Health". *Geographiska Annaler* 79 (B) pp. 141-154.

85. Gatrell, A. (1999) " GIS and Health: from Spatial Analysis to Spatial Decision". In Craglia, M. and Onsrud, H. (Eds.) *Geographic Information Research: Trans-Atlantic Perspectives*, Taylor & Francis, London, pp. 143-159.

86. Gatrell, A. C and C. E. Dunn (1995) "Geographical Information Systems and Spatial Epidemiology: Modelling the Possible Association Between Cancer of the Larynx and Incineration in North-West England." In M. J. C. de Lepper, H J. Scholten and R. M. Stem *The Added Value of Geographical Information Systems in Public and Environmental Health*, Kluwer, Dordrecht, pp. 215-235.

87. Gatrell, A. C., S. Garnett, J. Rigby, A. Maddocks and M. Kirwan (1998)

"Uptake of screening for breast cancer in South Lancashire." *Public Health*, 112, pp. 297-301.

88. Gatrell, A. G. and M. Löytönen (1998) *GIS and Health,* GISDATA IV, Taylor & Francis, London.

89. Gatrell, A.C. and Rowlingson, B. (1994) "Spatial Point Process Modelling in a GIS Environment. In Fotheringham S and Rogerson P. (Eds.) *Spatial Analysis and GIS.* Taylor and Francis, London. pp. 147-164.

90. Garvican, L. "Planning for a possible national colorectal cancer screening programme". *Journal of Medical Screening*, Vol. 5, pp. 187-194.

91. Geddes A and Flowerdew R (2000) "Geographical considerations in designing policy-relevant regions". In the 3rd AGILE conference on Geographic Information Science, Helsinki/Espoo, Finland, May 25th-27th 2000, pp. 80-82 (Abstract).

92. Getis, A. and K. Ord (1996) "Local spatial statistics: an overview", in P. Longley and M. Batty (Eds.) *Spatial Analysis: Modelling in a GIS environment,* Geographical International, Cambridge, England, pp. 261-277.

93. Getis, A. and K, Ord (1992). "The analysis of Spatial Association by Use of Distance Statistics". *Geographical Analysis*, 24, pp. 189-206

94. Glick, B. J. (1982) "The spatial organisation of cancer mortality". *Annals of the Association of American Geographers*, 72, pp. 471-481.

95. Goodchild, M. (1989) "The Issue of Accuracy in Global Databases". In M. Goodchild and R. Gopal (Eds.) *Accuracy of Spatial Databases*, Taylor and Francis, London, pp. 31-48.

96. Goodchild, M., R. Haining and S. Wise, *et al* (1992) "Integrating GIS and spatial data analysis problems and possibilities". *International Journal of Geographical Information Systems*, Vol. 6, No. 5, pp. 407- 423.

97. Goodchild, M. and S. Gopal (1989) *Accuracy of Spatial Database*, Taylor and Francis, London.

98. Goodchild, M. F. and N. S-N. Lam (1980) "Areal interpolation: a variant of

the traditional spatial problems". *Geo-Processing*, Vol. 1, pp.297-312.

99. Goodchild, M.G. (1987) "A spatial analytical perspective on geographical information systems". *International Journal of Geographical Information Systems*, Vol. 1, pp. 327-334.

100. Green M. and R. Flowerdew (1996) "New evidence on the modifiable areal unit problem". In P. Longley and M. Batty (Eds.) *Spatial Analysis: Modelling in a GIS environment,* Geographical International, Cambridge, England, pp. 41-54.

101. Greenland, S. and H. Morgenstern (1989) "Ecological bias, Confounding, and Effect Modification". *International Journal of Epidemiology*, Vol. 18, No. 1, pp. 269-274.

102. Haining, R.P. (1993) *Spatial data analysis in the social and environmental sciences*, Cambridge University Press.

103. Haining, R.P. (1994a) "Designing Spatial Data Analysis Modules for Geographical Information Systems". In Fotheringham, S. and Rogerson, P. (Eds.) *Spatial Analysis and GIS*, Taylor & Francis, pp. 45-64.

104. Haining, R. P. (1994b) "Diagnostics for regression modelling in spatial econometrics". *Journal of Regional Science*, Vol. 34, No. 3, pp. 325-341.

105. Haining. R. P. (1998) "Spatial Statistics and the Analysis of Health Data." In A. C. Gatrell and M. Löytönen (Eds.) *GIS and Health,* GISDATA IV, Taylor & Francis, London, pp. 29-48.

106. Haining R. P. J. Ma and S. M. Wise (1996) "The design of a software system for interactive spatial statistical analysis linked to a GIS". *Computational Statistics*, 11, pp. 449-466.

107. Haining R.P., S. M. Wise and M. Blake (1994) "Constructing regions for small area analysis: material deprivation and colorectal cancer". *Journal of Public Health Medicine* 16(4), pp. 429-438.

108. Haining R.P., S. M. Wise and J. Ma (1998) "Exploratory data analysis in a geographic information systems environment". *The Statistician*, 47 (3), pp.

457-469.

109. Haining, R. P., S. M. Wise and J. Ma (2000) "Designing and implementing software for spatial statistical analysis in a GIS environment". *Journal of Geographical Systems*, Vol. 2 (3), pp. 257-286.

110. Haining, R.P, and S.M. Wise (1991) *GIS and Spatial Data Analysis: Report on the Sheffield Workshop.* Regional Research Laboratory Initiative Discussion Paper No.11. Department of Town and Regional Planning, University of Sheffield.

111. Haggett, P. (1976) "Hybridizing alternative models of an epidemic diffusion process". *Economic Geography*, 52, pp. 136-146.

112. Hart, J. T. (1971) "The inverse care law". *Lancet*, Vol. 7, pp. 80-84.

113. Haslett, J. (1992) "Spatial data analysis - challenges". *Statistician*, Vol. 41, pp. 271-284.

114. Haslett, J. Wills, G. and Unwin, A. R (1990) "SPIDER - an interactive Statistical Tool for the Analysis of Spatially Distributed Data". *International Journal of Geographical Information Systems*, Vol. 4, No. 3, pp. 285-296.

115. Hayes, M. (1999) "Man, disease and environmental associations': from medical geography to health inequality". *Progress in Human Geography*, Vol. 23, No. 2, pp. 289-296.

116. Heuvelink, G. B. M., and P. A. Burrough (1989) "Propagation of errors in spatial modelling with GIS". *International Journal of Geographical Information Systems* Vol. 3, No. 4, pp. 303-322.

117. Heywood, I., S. Cornelies, and S. Carver (1998) *An Introduction to Geographical Information Systems*. Longman, Essex.

118. Hills, M. (1992) "Some comments on methods for investigating disease risk around a point source". In P. Elliott, J. Cuzick, D. English and R. Stern (Eds.) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies*, Oxford University Press, Oxford, pp. 231-238.

119. Hirschfield, A., P. J. B. Brown and P. Bundred (1995) "The spatial analysis

of community health services on the Wirral using Geographic Information Systems". *Journal of the Operational Research Society*, 46, pp. 14-59.

120. Holland, W. W. and Stewart, S. (1990) "Screening in Health Care". The Nuffield Provincial Hospital Trust.

121. Horn, M. E. T. (1995) "Solution techniques for large regional partitioning problems". *Geographical analysis* 27(3), pp. 230-248.

122. Jessop, E. G. (1988) "Equity of access? Small area variations in surgery". *Community Medicine*, Vol. 10, No. 1, pp. 1-7.

123. Johnson, I. S., P. C. Milner and J. N. Todd (1987) "An assessment of the effectiveness of cervical cytology screening in Sheffield". *Community Medicine*, vol. 9, No. 1, pp. 160-170.

124. Jones, K. and G. Moon (1987) "Health, disease and society", London, Routledge.

125. Judge, K. and N. Mays (1994) "Allocating resources for health and social care in England". *British Medical Journal*, Vol. 308, pp. 1363-1366.

126. Kendall, M.G. (1939) "The geographical distribution of crop productivity". *Journal of the Royal Statistical Society*, Vol. 102, pp. 21-48.

127. Kennedy S. (1989) "The small number problem and the accuracy of spatial databases". In M. F. Goodchild and S. Gopal (Eds.) *The accuracy of spatial databases*. Taylor and Francis, London, pp. 187-196.

128. Kievell, P. T., B. J. Turton and B. P. Dawson (1990) "Neighbourhood for Health Service Administration". *Social Science and Medicine*, Vol. 30, No. 6, pp. 701-711.

129. Kreuger, FAF, HAM van Oers and HGT Nijs (1999) "Cervical cancer screening: spatial association of outcome and risk factors in Rotterdam", *Public Health*, 113, pp.111-115.

130. Kulldorff, M. (1998) "Statistical Methods for Spatial epidemiology: Tests for Randomness". In A. C. Gatrell and M. Löytönen (Eds) *GIS and Health*, GISDATA IV, Taylor & Francis, London, pp. 49-62.

131. Last, J. M. *A Dictionary of Epidemiology*, Third Edition, Oxford University Press, Oxford.

132. Law, M. R. and J. K. Morris (1998) "Why is mortality higher in poorer areas and in more northern areas of England and Wales?". *Journal of Epidemiology and Community Health*, Vol. 52, pp. 344-352.

133. Lazar, P. (1981) "Geographical Correlations between Disease and Environmental Exposures". In J. F. Bithell and R. Coppi (Eds.) *Perspectives in Medical Statistics,* Academic Press, London, pp. 21-38.

134. Lepoz, A. D. (1992) " Mortality data". In P. Elliott, J. Cuzick, D. English and R. Stern (Eds.) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies,* Oxford University Press, Oxford, pp. 37-50.

135. Levine, N. (1996) "Spatial Statistics and GIS". *Journal of the American Planning Association*, Vol. 62, pp. 381-391.

136. Lovett, A. R. Haynes, G. Bentham, S. Gale, and J. Brainard (1998) "Improving Health Needs Assessment using Patient Register Information in a GIS". In A. C. Gatrell and M. Löytönen (Eds.) *GIS and Health,* GISDATA IV, Taylor & Francis, London, pp. 191-204.

137. Maes, J and M. H. Cornaert (1994) "Data Aspects of Geographical Information Systems". In M. J. C. de Lepper, H J. Scholten and R. M. Stern (Eds.) *The Added Value of Geographical Information Systems in Public and Environmental Health*, Kluwer, Dordrecht, pp. 99-114.

138. Macmillan, W. and T. Pierce (1994) "Optimization modelling in a GIS Framework: the problem of political redistricting". In Fotheringham, S. and Rogerson, P. (Eds.) *Spatial Analysis and GIS,* Taylor & Francis, pp. 221-246.

139. Maguire, D. J. (1991) "An overview and definition of GIS." In D. J. Maguire, M. F. Goodchild and D. W. Rhind (Eds.) Geographical Information Systems – principles and applications, Longman, Essex, pp. 9-20.

140. Maguire, D. J. and J. Dangermond (1991) "The functionality of GIS". In D. J. Maguire, M. F. Goodchild and D. W. Rhind (Eds.) *Geographical Information Systems – principles and applications*, Longman, Essex, pp. 319-35.

141. Majeed, F. A. N. Chaturvedi, R Reading and Y Bren-Shlomo (1994) "Equity in the NHS Monitoring and promoting equity in primary and secondary care". *British Medical Journal*, 308, pp. 1426-29.

142. Majeed, F. A., D. G. Cook, H. R. Anderson, S. Hilton, S. Bunn and C. Stones (1994) "Using patient and general practice characteristics to explain variations in cervical smear uptake rates". *British Medical Journal,* 308, pp. 1272-1276.

143. Martin, R. J. (1987) "Some comments on correction techniques for boundary effects and missing value techniques". *Geographical analysis*, 19, pp. 273-82.

144. Martin, R. J. (1992) "Leverage, Influence and Residuals in regression Models when Observations are correlated". *Communications in Statistics,: theory and methods*, Vol. 21, pp. 1183-1212.

145. Martuzzi, M. and M. Hills (1995) "Estimating the Degree of Heterogeneity between Event Rates Using Likelihood", *American Journal of Epidemiology*, Vol. 141, No. 4, pp. 369-374.

146. Marshall, R. (1991) "Mapping Disease and Mortality Rates using Empirical Bayes Estimators". *Applied Statistics*, Vol. 40, No. 2, pp. 283-294.

147. Masser, I. (1990) "The Regional Research Laboratory initiative: an update". *The association for Geographic Information Yearbook 1990*, Talyor & Francis, London, pp. 259-263.

148. MathSoft (1996) *S+GisLink*, MathSoft, Inc. Seattle.

149. McCullagh, P. and J.A. Nelder, (1989). *Generalised Linear Models*, Second Edition, Chapman & Hall, London.

150. McKee, L. and C. Kottman (1999) "Inside the OpenGIS Specification".

[Online document] http://www.opengis.org/info/gisworld/PERSArticle9910LMCK2.htm. [visited on 26th Aug. 2000]

151. McQueen, J. (1967) "Some Methods for Classification and Analysis of Multivariate Observations". *Proceedings 5th Berkeley Symposium on Mathematical Statistics and probability*, Vol. 1, pp. 281-297.

152. Meade, M. S., J. Florin and W. Gesler (1985) *Medical Geography*, The Guilford Press, New York.

153. Muir, C. and (1987) *Cancer incidence in five continents*, International Agency for Research on Cancer.

154. Muir, C. (1989) "Cancer Mapping: Overview and Conclusion". In P. Boyle, C. S. Muir, and E. Grundmann (Eds.) *Cancer Mapping: recent Results in Cancer Research*, Vol. 114, Sprint-Verlag, Berlin, pp. 269-273.

155. NCGIA (1989) "The research plan of the National Center for Geographic Information and Analysis". *International Journal of Geographical Information Systems*, 3, pp. 117-136.

156. NAG (Numerical Algorithms Group), http://www.nag.co.uk

157. Nyerges, T. L. (1992) "Coupling GIS and spatial analytical models". In *Proceedings of 5th International Symposium on Spatial Data Handling*, Charleston, South Carolina. pp. 535-543.

158. Oden, N (1995) "Adjusting Moran's I for Population Density". *Statistics in Medicine*, Vol. 14, pp. 17-26.

159. Office for the National Statistics, Cancer Statistics: Registrations, England and Wales Series MB1, Stationary Office, London.

160. Olsen, S. F., M. Martuzzi, and P. Elliott (1996) "Cluster analysis and disease mapping -- why, when and how?, a step by step guide". *British Medical Journal*, 313, pp. 863-866.

161. Openshaw, S. (1978) "An optimal zoning approach to the study of spatially aggregated data". In I. Masser and Brown, P.J. (Eds.) Spatial Representation

and Spatial Interaction, Martinus Nijhoff, Leiden, pp. 95-113.

162. Openshaw, S. (1984) *The modifiable areal unit problem.* Concepts and Techniques in Modern Geography 38. GeoAbstracts, Norwich.

163. Openshaw, S. (1990) "A spatial analysis research strategy for the Regional Research Laboratory initiative". *Regional Research Laboratory Initiative Discussion Paper Number 3*, Department of Town and Regional Planning, University of Sheffield.

164. Openshaw, S. (1990) "Automating the search for cancer clusters: a review of problems, progress and opportunities", In R. W. Thomas (Ed) *Spatial epidemiology*, London, Pion, pp. 48-78.

165. Openshaw, S. and P. J. Taylor (1978) "The modifiable area unit problem". In N. Wrighley and R. J. Nennet, (Eds.) Quantitative Geography, London, Routledge, & Regan Paul. pp. 60-69.

166. Openshaw, S. and A. Craft (1991) "Using Geographical Analysis machines to search for evidence of clusters and clustering in childhood leukaemia and non-Hodgkin lymphomas in Britain". In G. Draper (Ed) *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain*, 1966-83, Studies on Medical and Population Subjects No 53. HMSO, London. pp. 109-122.

167. Openshaw S., M. Charlton, C. Wymer and A. Craft (1987) "A mark 1 Geographical Analysis Machine for the Automated Analysis of Point Data Sets". *International Journal of Geographical Information Systems,* Vol. 1, pp. 335-358.

168. Openshaw S. and Rao L. (1995) "Algorithms for re-engineering 1991 census geography". *Environment and Planning* A, 27(3), pp. 425-446.

169. Ord, K. and A. Getis (1995). "Local Spatial Autocorrelation Statistics: Distribution Issues and an Application". *Geographical Analysis*, 27, No.4, pp. 286-306.

170. Ousterhout, J. K. (1994) *Tcl and the Tk Toolkit.* Addison-Wesley, Reading,

MA.

171. Payne, N. and C. Saul (1997) "Variations in use of cardiology services in a health authority: comparison of coronary artery revascularisation rates with prevalence of angina and coronary mortality". *British medical Journal*, Vol. 314 pp. 257-261.

172. Peuquet, D.J. (1984) " A conceptual framework and comparison of spatial data models". *Cartographica*, 2, pp. 66-113

173. Phillimore. P, A. Beattie and P. Townsend (1994) "The widening gap. Inequality of health in northern England 1981-1991". *British Medical Journal*, 308, pp. 1125-1128.

174. Pocock, S. J., D. G. Cook and S. A. A. Beresford (1981) "Regression of Area Mortality Rates on Explanatory Variables: What Weighting Is Appropriate?" *Applied Statistics*, Vol. 30, No. 3, pp. 286-295.

175. Pollock, A. M. and N. Vickers (1998) "Deprivation and emergency admissions for cancer of colorectum, lung and breast in south east England: ecological study". *British Medical Journal*, Vol. 317, pp. 245-252.

176. Potthoff, R. F. and M. Whittinghill (1966) "Testing for homogeneity: II. The Poisson distribution". *Biometrika*, pp. 183-190.

177. Pukkala, E. (1992) "Use of record linkage in small-area studies". In P. Elliott, J. Cuzick, D. English and R. Stern (Eds.) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies,* Oxford University Press, Oxford, pp. 125-131.

178. Quinn, M. J. (1992) "Confidentiality". In P. Elliott, J. Cuzick, D. English and R. Stern (Eds.) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies,* Oxford University Press, Oxford, pp. 132-140.

179. Quinn, M. P. Babb, J. Jones and E. Allen (1999) "Effect of screening on incidence of and mortality from cancer of cervix in England: evaluation based on routinely collected statistics". *British Medical Journal*, 318, pp. 904-908.

180. Raper, J and M. Bundock (1993) "Development of a generic spatial language interface for GIS". In P. M. Mather (Ed) *Geographical Information Handling Research and Applications*, John Wiley & Sons Ltd, New York.

181. Raper J. K. and D. W. Rhind and J. W. Shepherd (1992) *Postcodes: the New Geography*. Longman, Harlow, UK.

182. Rhind, D. (1983) *A Census User's handbook*. Methuen, London.

183. Richards, M. A., D. Stockton, P. Babb, and M. P. Coleman (2000)"How many deaths have been avoided throughout improvement in cancer survival*?"*. *British Medical Journal*, Vol. 320 pp. 895-898.

184. Rochkind, M.J. (1985) *Advanced UNIX programming*, Prentice-Hall, New York.

185. Rossiter, D. J. and Johnston, R. J. (1980) "Program GROUP: the identification of all possible solutions to a constituency-delimitation problem". *Environment and Planning A*, Vol. 13, pp. 231-38.

186. Shapiro, S. (1977) "Evidence on screening for breast cancer from a randomised trial." *Cancer* 39, pp. 2772-2782.

187. Sheth, A. P., and J. A. Larson (1990) "Federated database systems for managing distributed, heterogeneous and autonomous databases". *ACM Computing Surveys*, Vol. 22, pp. 183-236.

188. Simon, E. (1996) *Distributed Information System - from client/server to distributed multimedia*, McGraw-Hill Companies, London.

189. Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

190. Smith, P. and Guengerich, S (1994), *Client-server Computing —All-in-one reference for total systems development*, Second Edition, Sams Publishing.

191. Spath, H. (1980) *Cluster Analysis Algorithms*, John Wiley and Sons, New York.

192. Stone, R. A. (1988) "Investigation of excess environmental risk around putative sources: statistical problems and a proposed test". *Statistics in*

Medicine, 7, pp. 649-660.

193. Sui, D. (1999) "GIS, Environmental Equity Analysis, and Modifiable Areal Unit Problems (MAUP)". In Craglia, M. and Onsrud, H. (Eds.) *Geographic Information Research: Trans-Atlantic Perspectives*, Taylor & Francis, London, pp. 41-54.

194. SunSoft, (1994), *Solaris 2.4 — Network Interfaces Programmer's Guide*. Sun Microsystems Inc.

195. Swaldows, A. J. (1992) "Cancer incidence data for adults." In P. Elliott, J. Cuzick, D. English and R. Stern (Eds) *Geographical and Environmental Epidemiology - Methods for Small-Area Studies*, Oxford University Press, Oxford, pp. 51-62.

196. Twigg, L. (1990) "Health-based Geographical Information Systems: their existing potential examined in the light of existing data sources". *Social Science and Medicine*, Vol. 30, pp. 143- 155.

197. Thomas, R. W. (1990) *Spatial epidemiology*, Pion, London.

198. Thomas, R. W. (1991) "Quantitative methods: clines, hot spots and cancer clusters". *Progress in Human Geography* 15, 4, pp. 444-455.

199. Thunhurst, C. (1985) "The analysis of small area statistics and planning for health". *Statistician*, 34, pp. 93-106.

200. Tjøstheim, D. (1978) "A measure of association for spatial variables." Biometrika, 65, pp. 109-14.

201. Tobler, W. R. (1976) "Spatial Interaction Patterns". *Journal of Environmental Systems*, Vol. 6, pp. 271-301.

202. Tomlin, C. D. (1990) *Geographical Information Systems and Cartographic Modelling*. Prentice-Hall, Englewood Cliffs, New Jersey.

203. Townsend, P. (1987) "Deprivation". *Journal of Social Policy*, Vol. 16, No. 2, pp. 125-146.

204. Turkey, J. (1977) *Exploratory Data Analysis*. Addison-Wesley Publishing Co. Reading, MA

205. Umar, A. (1993) *Distributed Computing and Client-Server Systems.* Prentice Hall, New York.

206. Unwin, A. R. (1996) "Exploratory spatial analysis and local statistics" *Computational Statistics,* Vol. 11, pp. 387-400

207. Unwin A., G. Hawkins, H. Hofman and B. Siegl (1996) "Interactive graphics for data sets with missing values – MANET". *Journal of Computational and Graphical Statistics,* Vol. 5, pp. 113-122.

208. Unwin, D. J. (1981) *Introductory Spatial Analysis,* Methuen, London.

209. Upton, G. J. G., and B. Fingleton (1985) *Spatial Data Analysis by Examples.* Wiley, New York.

210. Waldhor, T. (1996) "The spatial autocorrelation coefficient Moran's I under heteroscedasticity". *Statistics in Medicine,* Vol. 15, pp. 889-892.

211. Walter, S. D. (1992a) "The analysis of regional patterns in health data. I Distributional considerations". *American Journal of Epidemiology,* 136, pp. 730-741.

212. Walter, S. D. (1992b) " The analysis of regional patterns in health data. II. The power to detect environmental effects". *American Journal of Epidemiology,* 136, pp. 742-759.

213. Wakeford, R. (1990) "Some Problems in the Interpretation of Childhood Leukemia Clusters". In Thomas, R. W. (Ed) *Spatial Epidemiology,* London papers in regional science 21, a Pion Publication, pp. 79-89.

214. Waugh, T.C. (1986) "The Geolink system; interfacing large systems". In Blakemore (Ed) *Proceedings of Auto-Carto* Vol. 1, pp. 76-85, London:RICS

215. Wilkinson RG. (1996) *Unhealthy societies: the afflictions of inequality,* Routledge, London.

216. Wilkinson, P., C. Grundy, M. Landon and S. Stevenson (1998) "GIS in Public Health". In A. C. Gatrell and M. Löytönen (Eds.) *GIS and Health, GISDATA IV,* Taylor & Francis, London, pp. 179-189.

217. Williams, R. and J. Wright, and J. R. Wilkinson (1998) "Health needs

assessment-epidemiological issues in health needs assessment". *British Medical Journal*, Vol. 316, pp. 1379-1382.

218. Wise S.M., J. Ma and R. P. Haining (1996) "SAGE - a system for the interactive analysis of area-based health data". *Proceedings of 11th ESRI European User Conference,* ESRI(UK) http://www.esri.com/base/common/userconf/europroc96/PAPERS/PN51/PN51F.HTM.

219. Wise S. M., R. P. Haining and J. Ma. (1997) "Regionalisation tools for the exploratory spatial analysis of health data". In M.Fischer and A.Getis (Eds.) *Recent Developments in Spatial Analysis - Spatial Statistics, Behavioral Modelling and Neurocomputing, Springer*, Berlin, pp. 83-100.

220. Wise, S., R. Haining and J. Ma (2001) "Providing Spatial Statistical Data Analysis functionality for the GIS user: the SAGE project". *International Journal of Geographical Information Systems*, Vol. 15 (3), pp.239-254.

221. Wise S.M. (1990) "Evaluating GIS software for use in Higher Education". *Mapping Awareness,* 4(7), pp. 41-43.

222. World Health Organisation (1977) "International classification of diseases. Manual of the international statistical classification of disease, injuries, cause of death", Vol. 1, (the 9th edition), WHO, Geneva.

223. Wright, J. and R. Williams (1998) "Health needs assessment – Development and importance of health needs assessment". *British Medical Journal*, Vol. 316, pp. 1310-1313.

224. Wrigley, N. T., D. Holt, D. Steel and M. Tranmer (1996) "Analysing, modelling and resolving the ecological fallacy". In Longley, P. and M. Batty (Eds.) *Spatial Analysis: Modelling in a GIS environment*, Geographical International, Cambridge, England., pp. 23-40.

225. Yule, G. U. and M. G. Kendall (1950) *An introduction to the theory of statistics*. Griffin, London.

APPENDIX

The following materials have been produced during this project and are enclosed for reference.

1. *SAGE User Guide*

This document provides a complete guide for using SAGE. It teaches the user how to set up a SAGE session and how to use menus and dialogue boxes to interact with the system. For each SAGE tool, a brief introduction to its underlining techniques is given.. This document is also distributed with SAGE source and executable code and is recommended to be used in conjunction with its companion document *Getting Started with SAGE*, which is also available in the packages.

2. Copies of publications

Copies of six papers published or to be published papers concerning SAGE, in which the author of this thesis is also a co-author, are listed in the chronicle order.

3. SAGE source and executable code and some data used in the thesis.

The SAGE source code and a copy of SAGE executable code for Sun Solaris 2.5 are enclosed in a CDROM. A copy of data used in the case study, excluding CRC data for individuals, is also enclosed. Three files in the CDROM are the UNIX tar files. The user is advised to copy them to a UNIX system before doing anything on them. The user should consult the README file enclosed in the CDROM for instructions. It should be noted that SAGE works only with ARC/INFO version 7.0.x but not higher versions.

SAGE User Guide

Jingsheng Ma, Robert P. Haining and Stephen M. Wise

Department of Geography
The Sheffield Centre for Geographic Information and Spatial Analysis
University of Sheffield
S10 2TN, England

TABLE OF CONTENTS

List Of Diagrams

List of Tables

Acknowledgement

SAGE was written as part of a project funded by the UK Economic and Social Research Council (Grant R00234470) and the authors gratefully acknowledge this support.

Copyright

SAGE may be freely used for the purposes of academic research. Its use should be acknowledged in any publications, and the authors would be pleased to receive copies of any publication in which SAGE is used. The publicly available version of SAGE cannot be used for any other purpose (including contract research, consultancy or commercial exploitation). Anyone wishing to use SAGE for any such purpose must contact the authors first.

Introduction

SAGE (Spatial Analysis in a GIS Environment) is a spatial statistical analysis (SSA) software package for interactively studying area-based data. The latest release of SAGE is version 1.0. SAGE has been implemented on Sun workstations with either the Solaris 2.4 or 2.5 operating system. It also requires ARC/INFO 7.0.x with Arctools 7.0.x and Motif 1.2 shared library. The SAGE has been compiled only for running on a single SUN workstation, even though SAGE has been implemented for running in a networking environment.

The SAGE distribution package is available from the Department of Geography, the University of Sheffield, free for research purposes. For details of obtaining a copy of it, visit the web site: http://www.shef.ac.uk/Geography

SAGE User Guide (this document) is intended to introduce prospective users to the SAGE software package by providing detailed information about the capabilities of SAGE. It also gives users an overview of where tools fall within SAGE and how to access these tools using the graphical user interface.

For convenience, the ***SAGE User Guide*** is divided into three parts:

Part 1: Preliminaries–outlining SAGE;

Part 2: SAGE tools–discussing each SAGE tool;

Part 3: Appendices–providing more information on SAGE;

The ***SAGE User Guide*** is the second of two documents for the SAGE software package. The user who wants to get a quick start with SAGE, should read the first document - ***Getting Started with SAGE***, which is available, as well as this document, in the SAGE distribution pack.

Preliminaries

SAGE is a spatial statistical analysis (SSA) software package for interactively studying area-based data. As a result of integrating spatial statistical analysis (SSA) techniques with a state-of-the-art GIS, ARC/INFO, SAGE not only extends the usefulness of ARC/INFO for analysing spatial data, it also demonstrates how SSA can be incorporated into a GIS environment.

In this chapter, we outline SAGE by summarising its architecture, the structure of spatial data that it handles and its main features.

1. SAGE's Client-Server Architecture

Unlike other similar systems (Bailey 1994, Haining *et al* 1996), SAGE has been implemented as an application of client-server technology. It is composed of two separate but co-operative programs: a client program and a server program. The server program is based on ARC/INFO, while the client is made from a set of SSA tools. The server is the provider of services, while the client is the consumer of the services. When the client needs services, it sends requests to the server. After receiving and decoding the requests, the server performs a series of operations accordingly and sends the results of the operations to the client. With the results, the client performs further operations. The communication between the server and the client has been implemented by using remote procedure calls (RPC) (see Haining *et al* 1996). SAGE could be configured on either a networking environment or a stand-alone environment.

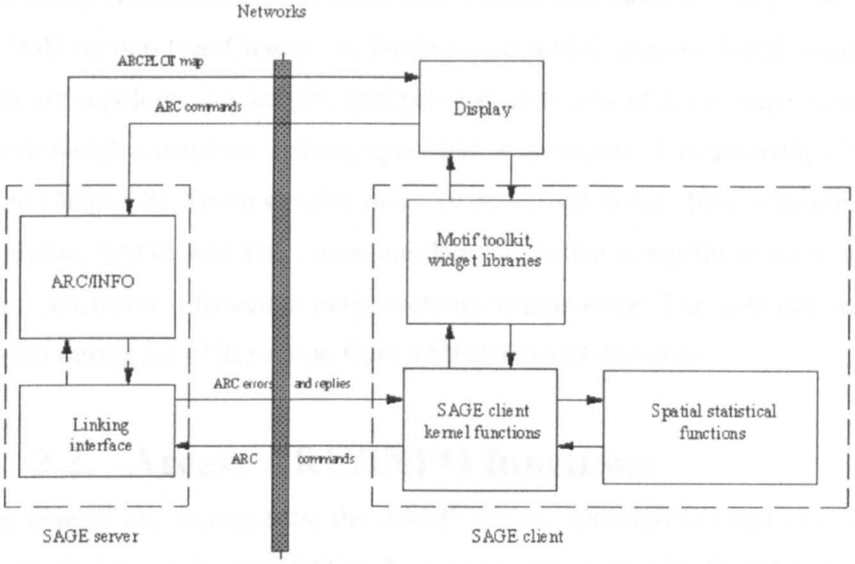


Diagram 2.1. An illustration of SAGE

2. Features of SAGE

2.1. Data management

SAGE is designed to handle area-based data sets. Each area, together with its associated attributes, is considered as a single object by SAGE, and is identified internally by a unique integer - object identifier.

SAGE uses data stored in standard ARC/INFO polygon coverages. The simplest situation is where all the data for a SAGE view (see Chapter 3) is stored in a single coverage, with the attributes held in a single Polygon Attribute Table (PAT). It also possible to attach additional attributes held in separate INFO tables to this PAT using the ARC/INFO RELATE mechanism, and this can be done when a SAGE view is created. In either case it is also possible to set up the view so that only a subset of the attributes are used - these are considered the active attributes. In order to speed up processing, the active attributes are read from ARC/INFO into the SAGE client.

SAGE requires two working directories as its working space, one for the server and one for the client, to manage temporary files[1]. Each SAGE session also creates a working space inside the SAGE server working space, named as sageat#####view (# is a number), and a number of INFO tables in this directory. One of these tables is TMPINFO storing

[1] The SAGE server working space is a sub-directory of the client working space, named as sagetry.

all newly created attributes during that session[2] and those attributes can be used by a later SAGE session (see Chapter 3). During each SAGE session, SAGE reads information on the arc topology, arc lengths and polygon centroids of a coverage from which it builds three weights matrices defining quantitative measures of relationships between the areas (see Chapter 5). These weights matrices are stored in the client working space as files. In addition, SAGE will also create another file, called dumpfile to store all active attribute data whenever it detects a unrecoverable system error. The user can recover the results saved during a SAGE session from TMPINFO and dumpfile.

2.2. Access ARC/INFO functions

As objects are managed by the SAGE server, ARC/INFO functions can be applied to them directly. Although SAGE does not employ all ARC/INFO functions, it includes functions for managing, mapping and querying the objects as well as carrying out spatial operations on objects, for instance, dissolving polygons.

2.3. Manipulation of data

2.3.1. Create new attributes

A weights matrix is a N by N matrix which defines the analyst's assumptions about the spatial relationships existing between the N objects. This matrix is required by many SSA tools (see Cliff 1981, Haining 1990, 1994a). It is of interest for example to explore the effect of different definitions of inter-area relationships on analysis findings (see Haining 1990, Haining *et al* 1996, Wise *et al* 1997, Haining *et al* 1997). SAGE provides tools which create weights matrices in pre-defined forms by using feature attributes but it also allows the user to modify pre-defined weights matrices and build weights matrices from scratch.

2.3.2. Create new attributes

Often we are interested in not only the original data themselves but also the data that can be computed from them. SAGE provides a set of tools allowing the user to perform a set of arithmetic and statistical computations on the original data and save results as new attributes.

[2] An active attribute must be in one of the types used by INFO: character, integer, or floating, and a newly created attribute can only be in either integer or floating.

2.3.3. Modify attribute data

Besides creating new data, SAGE allows the user to interactively modify attribute data from the table window. This enables the user to deal with the "missing values" problem and to perform a sensitivity analysis by modifying data values albeit in a simple way.

2.3.4. Import and export

SAGE allows the user to import data from ASCII files as new attributes. It also offers tools to export attributes and weights matrices to ASCII files so that they can be used by other software packages such as SpaceStat (see Anselin 1992), SPSS and Minitab. In addition, SAGE is able to create a set of new data (ARC/INFO polygon coverage and tables) from the data it uses.

2.4. Visualisation, linked windows and GUIs to SAGE tools

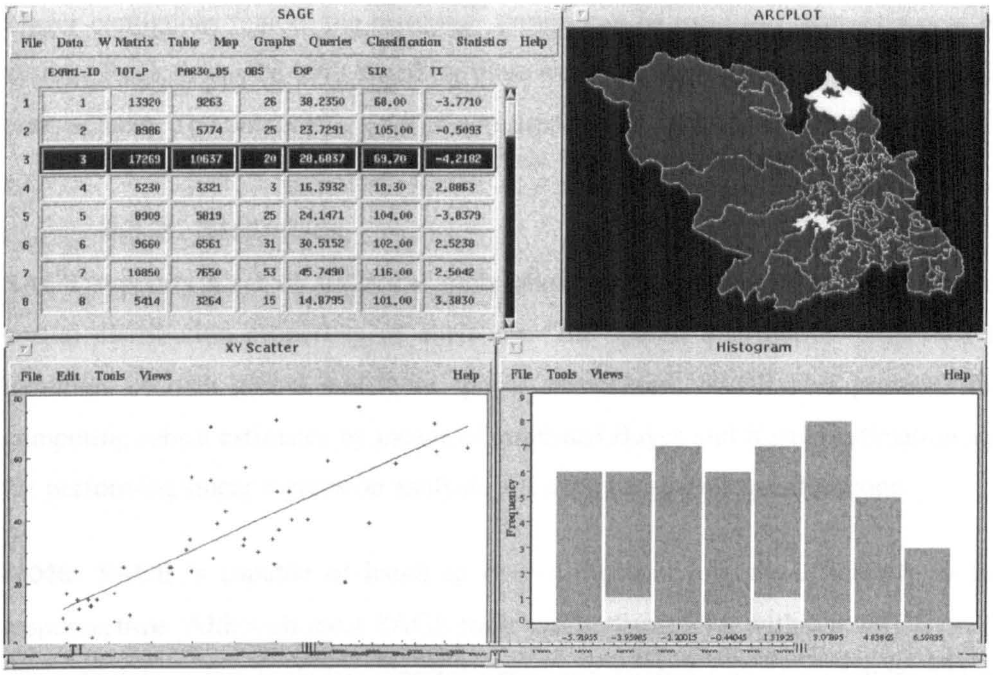
2.4.1. Visualisation

SAGE provides such GUI components as a map window, a spreadsheet-like window, and one or more graph windows. The properties of each object can be displayed as such visual items as a shaded polygon in the map window, a row in the table window or a point or bar in each graph window. The items are called map, table and graph items respectively.

A map item shows the shape of an object, while a row in the table window presents all the active attributes of an object. A point or a bar gives a view of the partial properties of an object such as an item in a histogram, an XY scatter, an XW scatter, a rankit plot, a boxplot, a lagged boxplot or a matrix graph.

2.4.2. Linked windows

In SAGE, visual items of an object are automatically associated with each other through object identifiers. This is done by an object matching mechanism implemented inside SAGE. Therefore, for each object, selecting any one of its visual items results in all other items of it being highlighted. In this sense, the windows are linked to each other. Diagram 2.3 shows the appearance of SAGE with highlighted polygons, rows and points and bars, which are associated with the selected points in the graph window displaying the XY Scatter.



	EXAM1-ID	TOT_P	PAR30_05	OBS	EXP	SIR	TI
1	1	13920	9263	26	38.2350	68.00	-3.7710
2	2	8986	5774	25	23.7291	105.00	-0.5093
3	3	17269	10637	20	28.6837	69.70	-4.2182
4	4	5230	3321	3	16.3932	18.30	2.0863
5	5	8909	5819	25	24.1471	104.00	-3.8379
6	6	9660	6561	31	30.5152	102.00	2.5238
7	7	10850	7650	53	45.7490	116.00	2.5042
8	8	5414	3264	15	14.8795	101.00	3.3830

Diagram 2.3 Appearance of SAGE

2.5. SSA Tools

Besides the GUI components mentioned above, SAGE provides a text output window for reporting SSA results. Except for this window and the map window, each of the other windows has a menu allowing the user access to SAGE tools.

The spreadsheet-like table window serves as a main window, and most SAGE tools can be accessed through its pulldown menu which is called the system menu. On the other hand, each graph window has its own menu to allow operations particular to the graphs it displays.

2.5.1. Querying objects

SAGE offers a set of tools for querying objects intuitively by selecting their visual items. It also provides a tool that allows the user to define a set of spatial and/or logical rules interactively and then to query the objects to find those that satisfy the rules. In both cases, all visual items associated with the selected objects are highlighted.

2.5.2. Classification and regionalisation

SAGE employs classification techniques for assigning objects to classes. Three tools are supported for classifying N objects into K groups under criteria that include either/both their attribute or/and spatial features. In the case where each group is required to form a bigger contiguous region, the resulting groups can be used to construct a new regional system. Regions can be built based on three weighted criteria: homogeneity, equality and compactness. The outcome is a set of new area-based data on which to apply SSA.

2.5.3. Statistics

SAGE supports a set of statistical tools allowing the user to study both classical and spatial distribution properties of attributes. The spatial distribution properties include measures of both global and local spatial association. SAGE also provides tools for computing robust estimates by means of empirical Bayes and Kernel estimation and tools for performing linear regression analysis allowing for spatial specifications.

Note: SAGE is capable of handling over a thousand objects efficiently in terms of response time. Although most SAGE tools can perform SSA within a very short time on this size data, a few tools like ***Spatial Error*** and ***Spatial Lag*** (see Chapter 11) perform slowly. However, tests have shown that they work efficiently on data with up to 500 objects.

SAGE Tools

SAGE tools are classified into nine categories by functionality as shown in Table 1 below. Each row in the table refers to a category of tools. As the system menu is organised based on the classification, the tools in each category can be accessed through the menu commands under the menu option corresponding to the category. The name of the menu option corresponding to each category is listed in the third column.

Category	Functionality	Option Name
Management of SAGE environments	manage SAGE environments on which SAGE tools work.	File
Data Management	create new attributes from existing data; import/export data from/to ASCII files; create new area-based data sets.	Data
Connectivity	create weights matrices; change the system default weights matrix (SDWM).	W Matrices
Tabular operations	delete attributes from the current SAGE environment; highlight all visual items associated with the selected rows; update INFO tables which have the values associated with the modified cells.	Table
Map Operations	shade the map; edit map legends; zoom in and out of a part of the map; save the map as a formatted dump file and display it.	Map
Drawing Graphs	create graph windows for drawing graphs in one of seven types: histogram, XY scatter, XW scatter, rankit, box, lagged box and matrix. In each window, query objects by selecting graph items.	Graphs
Querying Objects	query objects using individual or a set of spatial and logical rules. The items satisfying criteria are highlighted in the windows.	Queries
Classification	classify all N objects into K groups or regions using different methods with a variety of criteria.	Classification
Statistics	perform basic statistics for selected variables; compute Chi-square and Kolmogrov-Smirnov tests; compute local indicators of spatial association; compute robust Bayes estimates and kernel estimates; fit linear regression models with or without allowing for spatial specification; fit a generalised linear regression model with Poisson errors;	Statistics

Table 2.1. Classification of SAGE tools

The remainder of Part 2 is divided into nine chapters, one for each category. Each chapter covers information on the functionality and usage for every tool in each category.

For convenience, a SAGE tool is named in ***bold italic***, while the menu command for accessing the tool is in the same name but underlined. In addition, **bold** words are reserved for the names of items in dialogue boxes. Finally, the words, *observation*, *case* and *area* are used as synonymous to the term *object*, while *variable* and *item* are synonymous with *attribute* unless otherwise stated.

3. Management of SAGE Environments

Before being able to analyse a set of area-based data with SAGE tools, the user must set up a SAGE environment for it. To do this, the user must assign a number of environmental variables and then load the relevant data into the SAGE system with the tools provided. This process is described as follows:

Phase 1:

Set SAGE environmental variables pointing to:

- the location of an ARC/INFO polygon coverage with which a set of area-based data is associated;
- the location of external INFO tables which are related to the coverage;
- the relationship between the coverage and the INFO tables -- how the INFO tables are related to the coverage (if there is an external INFO table);
- a list of selected attributes (called active attributes) stored in the coverage and/or the external INFO tables to be used for analysis;
- cartographic properties;

Save the environmental variables in a file called *view* and named as *xx_view*[3];

Phase 2:

Perform the following operations to completely set up the SAGE environment:

- open a view created in Phase 1;
- load the data of all active attributes defined in the view into the spreadsheet-like table window;
- create a weights matrix from those topological and geometric data indicating the adjacency of every pair of areas and another two matrices defining spatial

[3] A SAGE view is also an ARCTOOLS view, but has only one theme relating to a polygon coverage.

relationships between the cases. For more information about weights matrices, see Chapter 5.

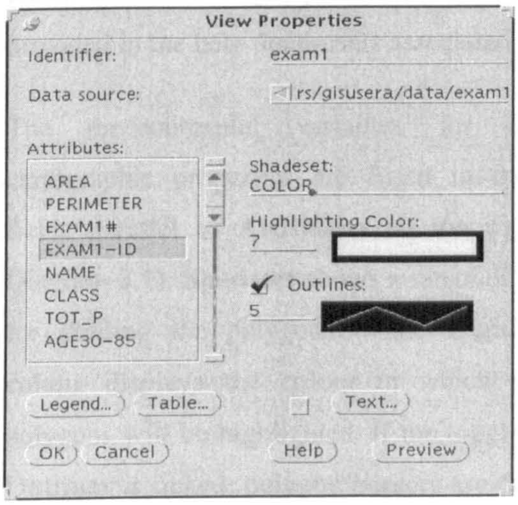
SAGE provides a set of eight tools for setting up and managing a SAGE environment. The tools are named as ***New, Open, Edit, Load, Save, Save As, Close*** and ***Exit***, and can be accessed through the menu commands under the ***File*** option in the system menu. The following sections will discuss each tool and its usage in turn. The discussion is based on the assumption that the user is familiar with creating a view in ARCTOOLS. Although the following materials are adequate for the user to create a simple view, for a complete understanding of creating a complicated view the user should read ARC/INFO and ARCTOOLS manuals (ESRI 1994).

3.1. Creating a New View File for SAGE

Tool, ***New***, allows the user to create a new view by interactively specifying the environmental variables. Before it allows the user to do so, ***New*** will check whether there still exists a SAGE environment. If so, it will call ***Save*** or ***Save As*** to save the environmental variables to a view and ***Close*** to clean up the environment. See sections 3.5, 3.6 and 3.7 for details about ***Save***, ***Save As*** and ***Close***.

Usage

Select ***New*** under ***File*** to call the dialogue box **View Properties** (see Diagram 3.1) and specify the environment variables in the dialogue box.



View Properties

Identifier: exam1

Data source: rs/gisusera/data/exam1

Attributes:
AREA
PERIMETER
EXAM1#
EXAM1-ID
NAME
CLASS
TOT_P
AGE30-85

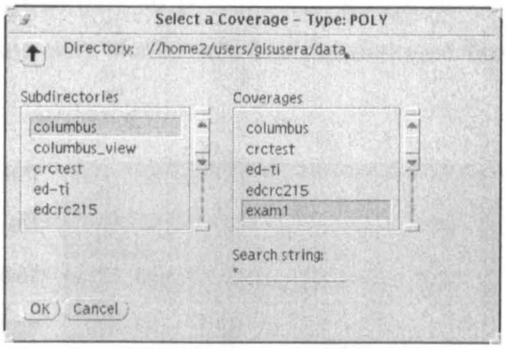
Shadeset: COLOR

Highlighting Color: 7

☑ Outlines: 5

Diagram 3.1. View properties

The first environmental variable that needs to be specified is the coverage to be used. To set this variable, provide the full name (including path name) of the coverage in the text field **Data Source**. Type the name from the keyboard or interactively select the name from a list of coverages in the dialogue box **Select a Coverage –Type POLY** (Diagram 3.2). To invoke this dialogue box, press the right mouse-button with the cursor in the text field **Data Source**. After a coverage is selected, its full name appears in the field. If the coverage name is valid, the scroll list **Attributes** will list all feature attributes in the *.PAT table of the coverage.


Select a Coverage - Type: POLY

Directory: //home2/users/gisusera/data

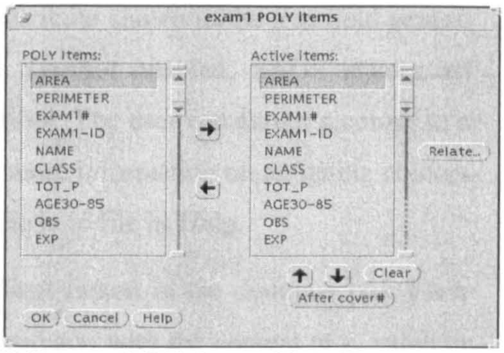
Subdirectories: columbus, columbus_view, crctest, ed-ti, edcrc215

Coverages: columbus, crctest, ed-ti, edcrc215, exam1

Search string: *

Diagram 3.2. Select a coverage

If not all feature attributes are of interest, the user can discard some of them from the list of active attributes as follows: firstly, press the push-button **Table** to invoke a dialogue box as shown in Diagram 3.3; secondly, select an attribute to discard from the scroll list **Active Items**; finally press the arrow pointing to the left to discard it. The user can also add an attribute to the list of active attributes by reversing the procedure. In order to relate external INFO tables to the coverage, press the push-button **Relate** and then follow the instructions provided in the help documents associated with **Help** in each dialogue box.

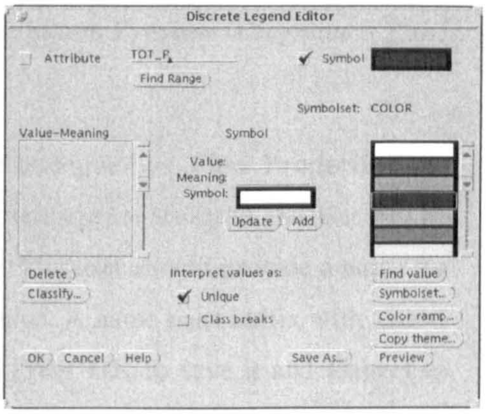

exam1 POLY Items

POLY Items: AREA, PERIMETER, EXAM1#, EXAM1-ID, NAME, CLASS, TOT_P, AGE30-85, OBS, EXP

Active Items: AREA, PERIMETER, EXAM1#, EXAM1-ID, NAME, CLASS, TOT_P, AGE30-85, OBS, EXP

Diagram 3.3.Relate tables

The environmental variables for defining cartographic properties are listed in the three fields parallel to **Attributes** on the right (see Diagram 3.1). **Shadeset** shows a range of colours for shading the polygons, while **Highlighting colour** displays the colour in which selected polygons will be highlighted. If the toggle button **Outlines** is ticked, polygon borders are drawn in the colour displayed in the text field underneath.


Discrete Legend Editor

□ Attribute TOT_P

✓ Symbol

Symbolset: COLOR

Value-Meaning

Symbol

Value:

Meaning:

Symbol:

Interpret values as:

✓ Unique

□ Class breaks

Diagram 3.4 Discrete legend editor.

Each field is associated with a dialogue box which lists all possible options for it. The user can invoke each dialogue box in the same way she/he invokes the dialogue box **Select a Coverage –Type POLY**.

SAGE draws a coverage as a map and shades each polygon in a colour assigned to it. By default, all polygons are shaded in a single colour. The user can, however, override the default settings through the dialogue box **Discrete Legend Editor** (Diagram 3.4), which can be invoked by pressing the button **Legend**. With the toggle button **Attribute** checked, SAGE shades each polygon using the colour which can be automatically or manually assigned to it based on the value of the attribute shown in the text field next to **Attribute**. By contrast, with the toggle button **Symbol** enabled, SAGE shades all polygons in the same colour as shown next to **Symbol**. The user can assign a colour to a group of objects using classification schemes. For more information on assigning colours to polygons and using classification schemes, see the help file in **Help**.

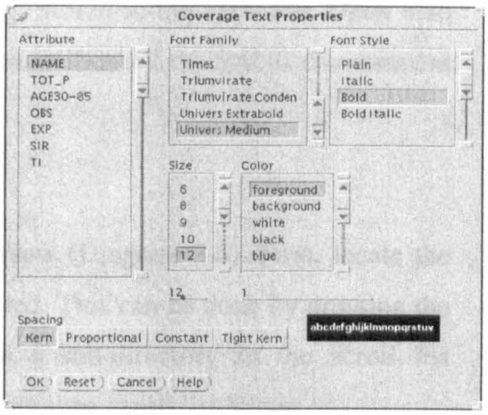


Diagram 3.5.Text properties.

With the toggle button next to the push button **Text** ticked in the dialogue box **View Properties** (Diagram 3.1), SAGE will label each polygon with the content (e.g. value or name) of a selected attribute. The user can set the properties for labelling in the dialogue box **Coverage Text Properties** (Diagram 3.5) that can be invoked by pressing **Text**. For further details, see the help documents in the dialogue box.

The user can check the settings before they are confirmed by looking at the map defined under the settings. This can be done by pressing the button **Preview** (Diagrams 3.1 and 3.4).

Before confirming the settings pressing **OK** in the dialogue box **View Properties**, the user MUST close all other dialogue boxes. After the settings are accepted, the user will be asked to save them in a view file (see Diagram 3.7). The user should provide a name for the view, otherwise SAGE will assign it a default name. A name must suffix with *_view*, otherwise SAGE won't automatically recognise it. Press **OK** to save it and **Cancel** to abandon it.

3.2. Setting SAGE Environmental Variables from a View File

Open allows the user to select a view, created in an earlier session through the use of ***New***, and set a SAGE environment using the settings stored in the view file. ***Open*** also displays a map defined by the view. Like ***New***, ***Open*** will check the SAGE environment first and close the current view as necessary.

Usage

Select ***Open*** to invoke the dialogue box **Select a View** (Diagram 3.6). First, locate the proper directory which contains the view to be selected. This can be done by pressing the arrow button (upper left corner) and/or click on a sub-directory in the scroll list **Subdirectories**. Then the user should select a view from the scroll list **Views**.

3.3. Modifying SAGE Environmental Variables

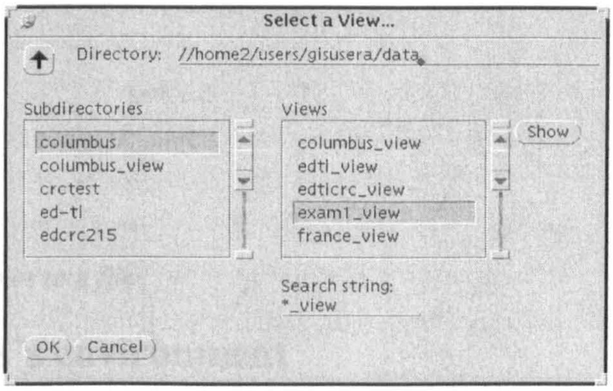
Edit allows the user to modify the current settings of environmental variables opened using ***Open***.

Usage

Select ***Edit*** to invoke the dialogue box **View Properties** (see Diagram 3.1) and modify environmental variables. See the usage for ***New***.

3.4. Loading Data into SAGE

After a view has been opened through ***Open***, the user MUST use ***Load*** to load the data to complete a SAGE environment. First, ***Load*** loads the active attributes defined in the view into the spreadsheet-like table window. Then it starts to extract topological and geometric data to create three weights matrices. See Chapter 5 for details.



Select a View...

Directory: //home2/users/gisusera/data

Subdirectories

columbus
columbus_view
crctest
ed-tl
edcrc215

Views

columbus_view
edtl_view
edtlcrc_view
exam1_view
france_view

Search string:
*_view

Diagram 3.6. Select a view

Usage

Select ***Load***.

3.5. Saving the Changes of Environmental Variables

Save allows the user to save the changed environmental variables back to the original view file from which the current SAGE was set up. A change could happen due to an explicit modification to any variable defined in the view through the tool ***Edit*** or an implicit modification when a new attribute has been added into the active attribute list or an active attribute has been deleted from the list. Note that the saving can only be successfully done if the user has writing access to the view file.

Usage

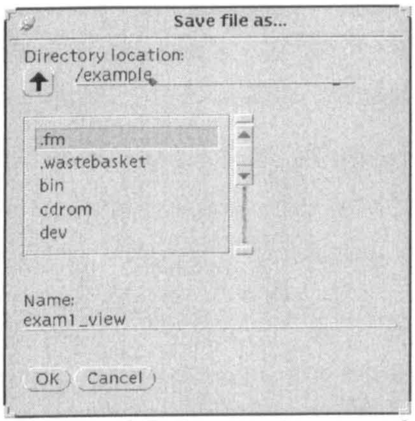
Select ***Save***.

3.6. Saving Environmental Variables to a New View File

Save As enables the user to save the environmental variables of the current environment as a new view.

Usage

Select ***Save As*** and provide a new view name in the dialogue box **Save File As** (Diagram 3.7).



Directory location: /example

Name: exam1_view

Diagram 3.7. Save environmental variables to a file

3.7. Cleaning up SAGE's environment

Close cleans out the resources used by the current SAGE environment.

Usage

Select ***Close***.

3.8. Exiting SAGE

Exit calls Close to clean up the SAGE environment and exit SAGE.

Usage

Select ***Exit***.

4. Data Management

This chapter discusses those SAGE tools for:

1. constructing new variables for a SAGE environment;
2. importing data in ASCII form into a SAGE environment as new variables and exporting variables and weights matrices to ASCII files;
3. creating new data sets; and
4. reporting properties about a SAGE environment.

Four tools, named as ***Arithmetic Variable***, ***Wx Variable***, ***Dummy Variable*** and ***Prob. Variable***, are provided for the first purpose, where the newly created variables are stored in a temporary INFO table. Tools, ***Import*** and ***Export***, are designed for the second, while ***Create Dataset*** and ***Properties*** are for the last two purposes respectively. These tools can be accessed under ***Data***.

4.1. Creating New Variables from Arithmetic Expressions

Arithmetic Variable creates variables from arithmetic expressions through the combination of arithmetic operators, numeric variables, constants and functions. Arithmetic operators are +, -, * and /. Functions could be any one listed in Table 4.1. Each function may take one or two arguments, and each argument may be a numeric variable, a constant or an expression. A constant is a special numeric variable that takes on the same value for each case.

Name	Description[4]	Example
SUM	$nv_i = \sum_j^N v1_j, i = 1,2,\ldots,N$	@SUM($(v1))
FABS	$nv_i = \|v1_i\|, i = 1,2,\ldots,N$	@FABS($(v1))

[4] nv_i, $v1_i$ and $v2_i$ are *i*th item in the vectors nv, v1 and v2 respectively.

EXP	$nv_i = e^{v1_i}$	@EXP($(v1))
LN	$nv_i = \log_e v1_i$	@LN($(v1))
LOG	$nv_i = \log_{10} v1_i$	@LOG($(v1))
POW	$nv_i = v1_i^{v2_i}$	@POW($(v1), $(v2))
SQRT	$nv_i = \sqrt{v1_i}$	@SQRT($(v1))
HYPOT	$nv_i = \sqrt{v1_i^2 + v2_i^2}$	@HYPOT($(v1), $(v2))
ASIN	$nv_i = \arcsin(v1_i)$	@ASIN($(v1))
ACOS	$nv_i = \arccos(v1_i)$	@ACOS($(v1))
ATAN	$nv_i = \mathrm{arctg}(v1_i)$	@ATAN($(v1))
SIN	$nv_i = \sin(v1_i)$	@SIN($(v1))
COS	$nv_i = \cos(v1_i)$	@COS($(v1))
TAN	$nv_i = \mathrm{tg}(v1_i)$	@TAN($(v1))
FLOOR	nv_i the greatest integer value less than or equal to $v1_i$.	@FLOOR($(v1))
ZSCORE	$nv_i = \frac{(v1_i - vm)}{sd};$ $vm = \frac{\sum_j^N v1_j}{N}; sd = \sqrt{\frac{\sum_j^N (v1_j - vm)^2}{N}}$	@ZSCORE($(v1))
RANGE	$nv_i = \frac{v1_i}{range};$ $range = \max(v1) - \min(v1)$	@RANGE($(v1))

UNIT	$nv_i = \dfrac{(v1_i - \min(v1))}{range};$ $range = \max(v1) - \min(v1)$	@UNIT($(v1))
POISSON	$nv_i = 1 - \sum_{j=0}^{v1_i} \dfrac{\exp(-v2_i)v2_i^j}{j!};$ $v1_i$ is an integer and $v2_i > 0$.	@POISSON($(v1), $(v2))
CHI	$nv_i = (v1_i - v2_i)^2 / v2_i;$	@CHI($(v1), $(v2))
INDEX	$nv_i = v1_i + (i-1)$	@INDEX($(v1))

Table 4.1.Functions supported by the tool Arithmetic Variable.

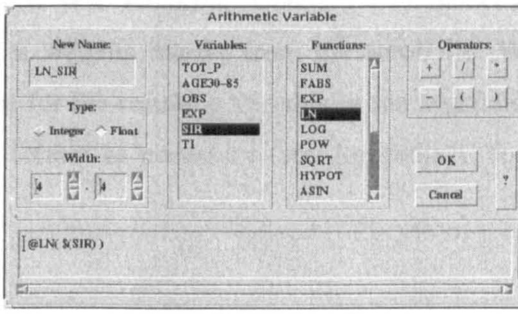
The first column of Table 4.1 names functions which appear in the dialogue box shown in Diagram 4.1. The second column gives a description of the function, while the last column shows an example of using the function. A variable name in an expression is prefixed by $ and enclosed with a pair of brackets, and a function is prefixed by @. A constant can appear at any place where a variable can but without $ and a pair of brackets. A constant could be in any format as the following example shows.

Suppose that there are three active numeric variables, v1, v2, v3, we can define an expression as:

```
1.23e3 - 2 + $(v1) + @FABS(@ZSCORE($(v2))) + @POW($(v3), 4.0);
```

Usage

Select ***Arithmetic Variable*** to invoke the dialogue box **Arithmetic Variable** (Diagram 4.1) to define an expression.



New Name: LN_SIR

Width: 4 . 4

@LN($(SIR))

Diagram 4.1. Dialogue box for creating new variable.

The user must assign a new name for the resulting variable in the field **New Name** and select the correct data type: integer or floating. This can be done by checking on the appropriate toggle button **Integer** or

Float. The user should make adjustments to the format of the variable by setting the number of digits before the decimal point and after the decimal point (for floating data only) in the area **Width**. Appendix C explains the importance of the format.

To define an expression, click on variables in the scroll list **Variables**, operators in the push button group **Operators**, and functions in the scroll list **Functions** in the correct order. The expression is shown in the text area in the bottom. The user can also add to the expression by typing directly from the keyboard.

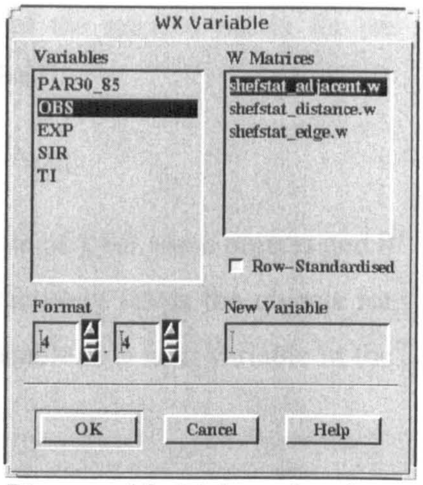


Diagram 4.2. Dialogue box for calculating the multiplication of W and a variable.

SAGE will report any syntax error which occurs in the expression in both a message box and the text output window. It will also provide message if the evaluation of the expression fails due to a mathematical inconsistency (e.g. trying to take the logarithm of zero). However, the user should check the correctness of the expression before submitting it for evaluation. Press **OK** to start the evaluation.

4.2. Creating WX Variables

WX Variable computes the product of a weights matrix, *W*, and a numeric variable, *x* and saves it (i.e. *WX*) as a new variable. By using different weights matrices, this facility enables the use to compute different types of weighted spatial sums of a numeric variable. The user can select to use either the raw elements of a weights matrix or its row-standardised version for the computation[5].

Usage

Select ***WX Variable*** to invoke the dialogue box **WX Variable** (Diagram 4.2). Select a variable from the scroll list **Variables** and a weights matrix from the scroll list **W Matrices**. The user should assign a new name for the resulting variable in the text field **New variable** and change the output width format as necessary (see Appendix C for

[5] A row-standardised weights matrix is a matrix where the sum of the elements of each row is equal to 1.0.

details). To use the row-standardised weights matrix of the selected matrix for the computation, check on the toggle button **Row-Standardised**[6].

4.3. Creating Dummy Variables

A dummy variable is a variable which takes on a value of 1 for some objects and 0 for others. ***Dummy Variable*** allows the user to interactively select the objects for which a dummy variable takes on 1, and add the variable as a new variable in the current SAGE environment.

Usage

Select ***Dummy Variable*** to invoke the dialogue box **Dummy Variable** (Diagram 4.3). Assign a new name for the dummy variable. Follow the instructions to select objects. See Appendix A for more information.

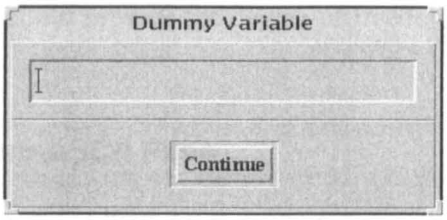


Diagram 4.3. Defining a dummy variable.

4.4 Computing Probabilities of Variables

Prob. Variable is designed to compute the probabilities of a variable for each object when the variable values are samples from one of the following distributions: normal, central Students t, F or χ^2. Probabilities are computed with respect to the standard forms of Normal distribution (mean = 0, standard deviation = 1). For the other distributions the user must provide the appropriate degree(s) of freedom. For more information, see NAG (1995) for subroutines G01EAF, G01EBF, G01ECF, and G01EDF.

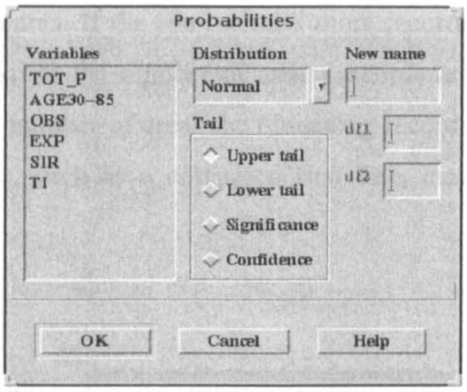


Diagram 4.4. Computing probabilities.

Usage

Select ***Prob. Variable*** to invoke the dialogue box **Probability** (Diagram 4.4). Then select a variable from the scroll list **Variables** and assign a name for the new variable in the text

[6] If the sum of the values in a row is 0.0, the value in the product corresponding to the row will be assigned to 0.0.

field **New Name**. The user should choose a distribution from the combo box **Distribution** and the required from the toggle button group **Tail** for computation.

The probabilities computed according to different tails are:

Upper tail: $y_i = \text{Prob}\{X>= x_i\}$;

Lower tail: $y_i = \text{Prob}\{X<= x_i\}$;

Significance $y_i = \text{Prob}\{X>= |x_i|\} + \text{Prob}\{X<= -|x_i|\}$;

Confidence: $y_i = \text{Prob}\{X<= |x_i|\}- \text{Prob}\{X<= -|x_i|\}$;

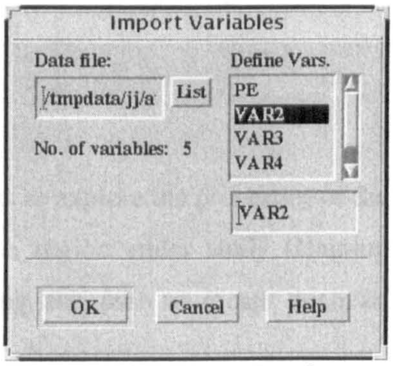
where x_i and y_i are the *i*th values of the selected attribute and the resulting attribute. For those distributions other than the normal, the user should provide the degree of freedom in **df1** and/or **df2** as required.

4.5 Importing Attributes Into SAGE from ASCII Files

Import allows the user to import numeric data from an ASCII file into SAGE as new attributes. The data in the file must be organised in columns separated by a comma, a space or a tab, and each column corresponds an attribute. The number of attributes to be imported into SAGE is determined by the number of columns in the first line in the file, and the *i*th record in the file is attached to the *i*th area. If the file includes more records than the number of areas only the first N (N equal to the number of areas) records are imported. If the file includes less records than the number of areas the remaining records will be filled with 999.00. Any error in the file, such as a corrupted first line, may produce unpredictable results.

Usage

Select ***Import*** to invoke the dialogue box **Import Variables** (Diagram 4.5). Press **List** to invoke a file selection box and then select a file from it. SAGE assigns each attribute with a default name (i.e. Var1, ...Var#). In order to change any default name, click on it within the **Define Vars.** list. This name appears in the text area below the list. Modify the name in the text field. The change is confirmed when the user presses the Return/Enter key from the keyboard or the user moves the mouse cursor out of the text field. Press **OK** to start importing the data. Note


Data file: /tmpdata/jj/a

No. of variables: 5

Define Vars.: PE, VAR2, VAR3, VAR4

VAR2

Diagram 4.5. Importing data into INFO table.

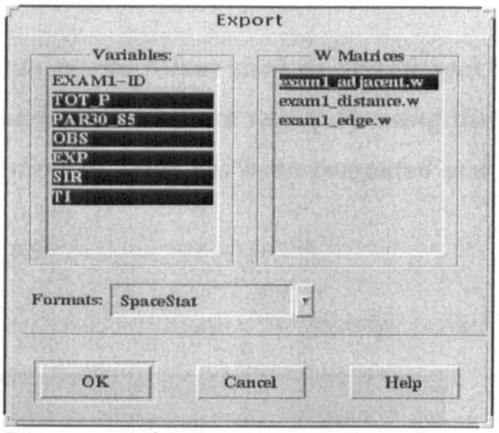
that all attributes must be imported at the same time and the user is not allowed to select particular attributes to be imported.

4.6 Exporting Attributes and Weights Matrices to ASCII Files

Export allows the user to export variables and weights matrices to ASCII files that can be used by other packages such as SpaceStat, SPSS, and Minitab.

Usage

Select ***Export*** to invoke the dialogue box **Export** (see Diagram 4.6). Select one or more variables from the scroll list **Variables**, and then specify the target package from the combo box **Formats**. If SpaceStat is the target format, the user is allowed to select weights matrices to be exported from the scroll list **W Matrices** as well. After pressing **OK**, the user is asked to assign a new directory to which the data are exported.



Formats: SpaceStat

Diagram 4.6. Exporting attribute data and W matrices for other packages.

In the case that the format is SpaceStat, the selected variables are saved in the file att_file.asc, and each selected W matrix is saved as a single file named as *W_matrix_name.asc*. There is another file in that directory, bi_adj.asc which saves the SDWM (see Chapter 5). For more information about the format of the SpaceStat file, see the SpaceStat User's Guide (Anselin 1995). If the user selects the format Tab separated, only the values of attributes are exported into a file called att_file.asc and are stored in columns separated by Tab. This file can be imported into statistical packages like SPSS and Minitab.

4.7 Creating New Data Sets

Often, in the analysis of area-based data, the analyst wants to explore the properties of the data aggregated according to different partitions of the region under study (Haining 1990). SAGE provides a tool ***Create Dataset*** allowing the user to create such an aggregated data set in an ARC/INFO coverage.

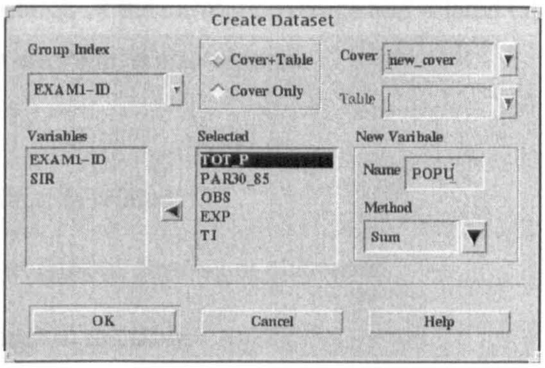
In SAGE, partitioning a region is regarded as a special type of grouping of objects where the polygons of the objects in each group form a contiguous larger area. SAGE creates an integer attribute, *grouping index*, for each grouping of objects where its values indicate to which group an object belongs. SAGE is able to automatically validate whether an integer attribute is a grouping index and whether a grouping index forms a partition of the region.

Given a grouping index which forms a partition of a region, ***Create Dataset*** aggregates the original data set into a new data set by dissolving the shared boundary of the objects in each group and merging the values of their attributes. The boundaries are merged using the ARC/INFO function Dissolve, while the values are merged by taking either the mean, the median, the standard deviation, the inter-quartile range or the sum of them.

Note that SAGE does not provide any function to merge values such as percentages, proportions or ratios. In these case, it is necessary to re-compute values by using the source data from which the percentage, proportion or ratio has been computed and aggregating these before re-computing.

Usage

Select ***Create Dataset*** to invoke the dialogue box **Create Dataset** (see Diagram 4.7). The combo box **Group Index** lists all available grouping indices, each of which forms a partition of the whole region. The user should select a grouping index on which a new set of data is to be created, otherwise SAGE will use the current one in that box.



Group Index: EXAM1-ID

◆ Cover+Table
◇ Cover Only

Cover: new_cover

Table:

Variables: EXAM1-ID, SIR

Selected: TOT_P, PAR30_85, OBS, EXP, TI

Name: POPU

Method: Sum

Diagram 4.7. Creating a new data set.

A new set of data can take one of two forms: either new attributes stored in a separate INFO table or in an ARC/INFO polygon PAT file. To make a choice, check **Cover+Table** or **Cover only**. In the case that the **Cover+Table** is checked, the user must provide a coverage name and a table name. In the case that the **Cover only** is checked, the user is required to provide only a coverage name.

The text fields **Cover** and **Table** display a user specified coverage name and a table name. The user should follow the following procedure to assign new names. First, click on the arrow next to the text fields **Cover** or **Table** to invoke a dialogue box (named as either **Coverages** or **Tables**) which lists all exiting coverage names or table names in a scroll list. Second, type a new name for either the coverage or the INFO table in the text field **Input New Name** below the list.

After assigning a coverage name (and a table name), the user needs to select variables to be included in the new data set. Select variables one by one from the scroll list **Variables** and move each selected variable into the scroll list **Selected** by pressing the arrow. The user also needs to give a name for each variable and attach an aggregation method to the variable. Select each variable in **Selected** in turn, then assign a name to it in the text field **Name** and choose the required method for aggregation from the combo box **Methods**. After doing this for all the selected variables, press **OK** to create the new data set.

In order to use the new data set, the user needs to go through the procedure discussed in Chapter 3.

4.8 Reporting Properties of SAGE Environments

Properties lists information about the current SAGE environment in the text output window. The information includes the location of the current coverage and related tables, and the number of observations. For each attribute, it also reports:

- attribute name;
- source data file to which the attribute is related;
- type of data in the attribute such as integer, floating or character;
- exterior width of the attribute values in the table;
- whether the attribute is a grouping index; and;
- if so, the number of classes and number of regions.

Usage

Select ***Properties***;

5. Connectivity

A weights matrix W is a N by N matrix which defines the analyst's assumptions about the spatial relationships existing between the N areas. This matrix is required by many SSA tools (see Cliff 1981 and Haining 1990).

SAGE stores a weights matrix as a file named with a suffix ".w" (e.g. adjacent.w) where each element is a double floating, not negative value. The structure of such a file is illustrated by the following diagram where the length of the first three items is fixed.

Coverage Name	No. Cases	Comments	Weights Matrix Data

A weights matrix file could be used in another SAGE environment if:

- the coverage name in the file matches the name of the coverage in the current SAGE environment; and

- the number of cases recorded in the file is equal to the number of cases in the current SAGE environment.

As discussed in 3.4, SAGE creates three matrices during Phase 2 of the configuration of a SAGE environment. The first is a weights matrix and contains information indicating the adjacency of every pair of cases. If a pair of areas C_i and C_j are adjacent to one another, the elements of the matrix, $w_{ij} = w_{ji} = 1.0$, otherwise $w_{ij} = w_{ji} = 0.0$. The second matrix stores the distance between the centroid of every pair of areas, while the third one records the length of the shared boundary of every pair of adjacent areas. Information for constructing these matrices is extracted from the feature attribute tables (AAT and PAT) of the coverage defined in the current SAGE environment. These matrices are stored as files, XX_adjacent.w, XX_distance.w and XX_edge.w respectively, where XX is the coverage name in the current SAGE environment.

Note that the elements in XX_distance.w and XX_edge.w are not, directly, appropriate measures of the strength of adjacency, but they are used for constructing weights matrices.

Since some tools rely on the adjacency, SAGE always keeps a binary weights matrix in the computer memory. This weights matrix is called the SDWM (System Default Weights Matrix) and is symmetric. So in SAGE any two areas can be either adjacent to each other or not adjacent at all. However, the user should be aware a weight matrix is not necessarily a symmetric matrix.

SAGE allows the user to convert any weights matrix to the SDWM by assigning the element at the *i*th row and *j*th column to 1 if the elements, say w_{ij} and w_{ji}, in the weight matrix file are not zero, otherwise to 0. The first SDWM is obtained by SAGE automatically from xx_adjacent.w during the phase 2 configuration of a SAGE environment. Note that it is up to the user to decide whether a SDWM makes sense with a tool.

Table 5.1 lists those tools which behaviours depend on the SDWM.

Name	Reference chapter
Export (with SpaceStat format only)	4
Properties	4
Pre-defined	5
Self-defined	5
Lag	9
Heuristic Classification	10
Kernel (median smoother only)	11
Select Cases	Appendix A

Table 5.1. Tools access the SDWM

SAGE provides three tools ***Pre-defined***, ***Self-defined*** and ***Set Adjacency*** allowing the user to:

1. create five types of pre-defined weights matrix;
2. create a self-defined matrix from scratch or by editing an existing matrix;
3. convert a weights matrix to the SDWM.

The remainder of this chapter discusses each tool in turn. These tools can be accessed from the corresponding menu commands under the menu option ***W Matrices***.

5.1. Creating Pre-defined Weights Matrices

Pre-defined allows the user to create weights matrices of the following types:

1. Inverse distance: $w_{ij} = d_{ij}^{-\gamma}$; Where d_{ij} is the Euclidean distance between the centroids of objects *i* and *j* and γ is a user specified positive constant.

2. Exponential distance: $w_{ij} = \exp(-d_{ij}^{\gamma})$; where d_{ij} is the same as given above.

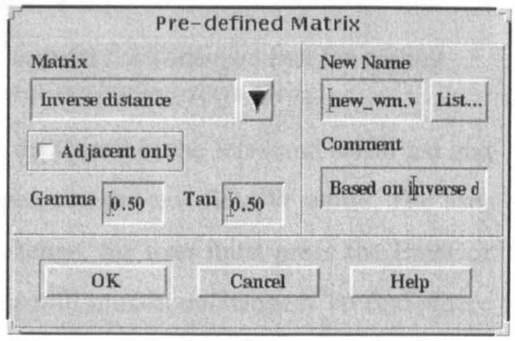
3. Shared boundaries: $w_{ij} = (l_{ij}/l_i)^{\tau}$ ($\tau > 0$); where l_{ij} is the length of the shared-edge between objects *i* and *j* and l_i is the perimeter of objects *i*, and τ is a user specified positive constant.

4. Shared boundaries over inverse distance: $w_{ij} = (l_{ij}/l_i)^{\tau} / d_{ij}^{-\gamma}$ ($\tau > 0$); where l_{ij}, l_i, d_{ij}, γ and τ are the same as given above.

5. Higher order adjacency: for given integer k, w_{ij} = 1.0 if object i is adjacent to object j at any order, up to kth order inclusively or at the order, otherwise w_{ij} = 0.0;

Usage

Select ***Pre-define*** to invoke the dialogue box **Pre-defined Matrix** (see Diagram 5.1). The user should select the type of matrix to be created from the combo box **Matrix** otherwise SAGE will use the default one, Inverse distance. If a weight matrix is to be created conditional on the adjacency defined by the SDWM, the user should check the toggle button **Adjacent only**. The user should provide proper settings for γ and τ in **Gamma** and **Tau**. When Higher order adjacency is the selected type, SAGE will create a higher order adjacent matrix based on the SDWM and a given order that the user is asked to provide after pressing **OK**. Meanwhile, **Adjacent only** will change to **Up to the order** and SAGE will create a weights matrix by taking into account all objects up to the given order or at the order according to whether the toggle is set or not.


Pre-defined Matrix

Matrix: Inverse distance

New Name: new_wm.v

☐ Adjacent only

Comment: Based on inverse d

Gamma: 0.50

Tau: 0.50

Diagram 5.1. Creating pre-defined W matrices.

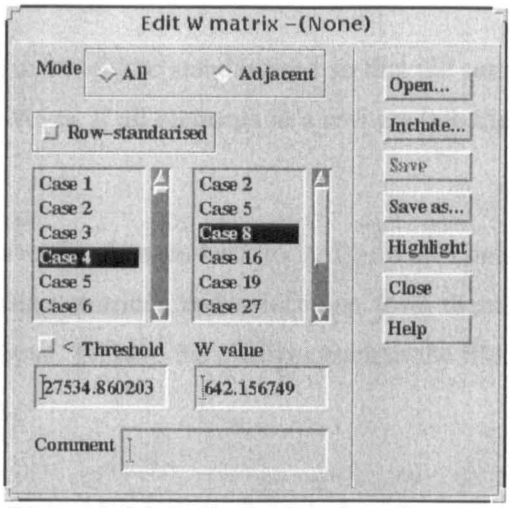
Before a new weights matrix can be created, the user must give a name to the matrix. This is done as follows. First, select **List...** to list all existing matrices from a list dialogue box. Assign a new name in the text field **New Name**. The text field **Comment** (Diagram 5.1) allows the user to put some comments into the matrix file.

5.2. Creating Self-defined Weights Matrices

Self-define allows the user to create a new weights matrix from an empty matrix and by editing an existing one.

Usage

Select ***Self-define*** to invoke the dialogue box **Edit Matrix** (Diagram 5.2). An initial matrix with elements set to 0.0 is created at this stage. The user can work with this empty matrix. Alternatively, the user can load an existing matrix as an initial matrix. To select a matrix, click the push-button **Include**, and select that weights matrix.



Threshold: 27534.860203

W value: 642.156749

Diagram 5.2.Dialogue box for editing relationship between areas.

Two scroll lists in the dialogue box list cases. To edit an element, say w_{ij}, the user should click on **Case i** in the left-hand scroll list and **Case j** in the right-hand list. The value of w_{ij} appears in the text field **W value**. The user can edit the value in the field. To confirm the change, the user must press the Enter or Return key from the keyboard, otherwise the value will remain unchanged. To find where the two selected cases are on the map, press the push-button **Highlight**.

To edit all elements w_{ij}, the user should check on the radio button **All** set by default. To edit only elements w_{ij} where **Case j** is adjacent to **Case i**, check the radio button **Adjacent**. In this case, the right-hand scroll list will only list those cases which are adjacent to the one being selected in the left-hand scroll list. All other elements excluded from editing are set to 0.0. Note that the adjacency is defined by the SDWM.

The user can set those elements of the matrix being edited to 0.0 which are either greater or smaller than a threshold shown in the text field **Threshold**. The user can edit the threshold and should confirm each editing by pressing the Enter/Return from the keyboard. The greater or smaller than the threshold is indicated by the sign ">" or "<"

which can be changed by pressing the toggle button. By default, if the sign is ">", the threshold is set to the maximum of all elements, otherwise the threshold is 0.0. Note that the elements beyond the threshold appear in the text field **W value** as 0.0. However, a real change to each element is only made when the user confirms to save the matrix.

After editing elements, the user can save this unnamed matrix to a file using **Save As** or the user may be asked to do so when the user tries to edit another one and close the dialogue box. In order to help remember the way the matrix has been edited, the user can type key words in the text field **Comment** before saving it.

If **Row-standardised** is checked, the weights matrix will be standardised so that the sum of all elements for each row is equal to 1.0. However, if all elements in a row are 0.0, the row will be left unchanged.

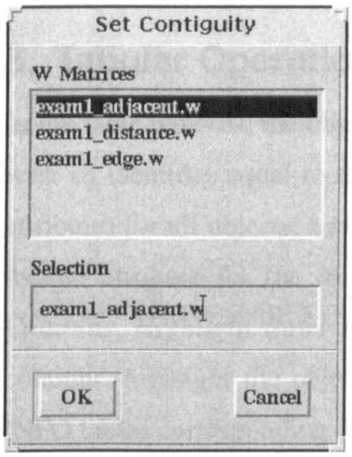
The user can also edit an existing matrix and save the changes back to it. To do so, click on the push-button **Open** to get a list of existing matrices and select one from them. Editing elements follows the same procedure above. To save the changes back to the file, press **Save.** Press **Close** to exit the dialogue box.

5.3. Setting the System Default Weights Matrix

Set Adjacency allows the user to set a matrix to the SDWM. Since the SDWM affects tools listed in Table 5.1, ***Set Adjacency*** should be used cautiously.

Usage

Select ***Set Adjacency*** and then select one matrix from the matrices listed in the dialogue box shown in Diagram 5.3



Set Contiguity

W Matrices

exam1_adjacent.w
exam1_distance.w
exam1_edge.w

Selection

exam1_adjacent.w

Diagram 5.3.Setting the current connectivity with a W matrix.

6. Tabular Operations

In the table window, the objects are presented as follows: a row corresponds to an object with an identifier equal to the row number; a column represents an active variable (i.e. attribute) for all objects; and a cell of the *i*th row and *j*th column holds a value which is the *j*th attribute for the *i*th object. SAGE provides the user with three tools ***Delete Columns***, ***Highlight Rows*** and ***Update Data***, which can be used to delete a number of columns, highlight the objects associated with the selected rows and update values in INFO tables corresponding to the cells being modified. These tools can be accessed from the menu commands under the menu option ***Table***.

6.1. Deleting Columns

Delete Columns allows the user to delete a number of variables associated with selected columns from the current SAGE environment. As a consequence, it changes the setting of the list of active attributes of the current SAGE environment.

Usage

Firstly, select columns to be deleted from the table. This is done as follows: select the first column to be deleted by moving the mouse cursor into the column, and then press the Shift key and click the right-hand mouse button (i.e. Shift + Right-Mouse) at the same time. To select an additional column, move the mouse cursor in the column, and then press the Ctrl key button and the right-hand mouse button (i.e. Ctrl + Right-Mouse) at the same time. The selected columns are highlighted in the table. After selecting columns, click ***Delete Columns*** to delete them.

Note that the cell where the arrow cursor is positioned must not be an currently active cell for the purpose of editing, otherwise, the selection will not be performed.

6.2. Highlighting Rows

Highlight Rows highlights those visual items on the map and on graph windows which are associated with selected rows.

Usage

Follow the same procedure described in 6.1 to select rows, but use Shift + Left-Mouse and Ctrl + Left-Mouse instead of Shift + Right-Mouse and Shift + Right-Mouse respectively. After selecting rows, the user needs to click ***Highlight Rows***.

6.3. Updating Data

SAGE allows the user to modify the values of variables by editing them in the corresponding cells in the table. However, the user is allowed to edit only the values of those variables which are created after a SAGE environment is set up[7]. After modifying the values in cells, the user must use ***Update Data*** to update them in the INFO tables. In addition, ***Update Data*** also updates all graphs which use the updated values.

Usage

After modifying the values in cells, select ***Update Data*** to make the updating.

[7] In order to modify the values of any other variable, the user could use the ***Arithmetic Variable*** tool to copy the variable as a new attribute.

7. Map Operations

SAGE provides the user with eight tools for redrawing the map, shading each polygon of the map based on the value of a selected attribute, and zooming in and out of the map. The tools are named as ***Redraw Map***, ***Map Item***, ***Map in Mono-colour, Edit legend***, ***Drag Zoom***, ***Zoom In***, ***Zoom Out***, and ***Reset***, and can be accessed from the menu commands under the menu option Map.

7.1. Redrawing the Map

Redraw Map redraws the map according to the environmental variables defined in the current SAGE environment.

Usage

Select ***Redraw Map***.

7.2. Colouring the Map with Attribute Values

Map Item allows the user to shade the polygon of each object in a colour which is assigned to the object according to the value of a selected variable This is done by calling ARCPLOT functions. ARCPLOT sorts polygons in descending order of the values of the selected variable. Then it assigns a colour in the shade set (see chapter 3) to a polygon if the polygon ranks the same as the colour does in the shade set. (Two polygons will be assigned in the same colour if their values are equal.) For those polygons which rank outside the range of the colours in the shade set, they are assigned the background colour.

SAGE allows the user to shade a group of polygons in the same colour. This can be done in two ways. The user can use classification tools (see chapter 10) to classify the values into groups. This will generate a new attribute called a grouping index which indicates to which group a polygon belongs. Then the user can apply ***Map Item*** to this index. Alternatively, the user can use ***Edit Legend*** or ***Edit*** tool to archive a similar result. Moreover, the user could assign a polygon any colour manually. For more information on the use of the latter two, see ***Edit Legend*** below.

Usage

Select a column (Shift + Left-Mouse) and choose ***Map Item***.

7.3. Mapping in mono-colour

Map in Mono-colour allows the user to shade the map using a single colour which is specified in the dialogue box **Discrete Legend Editor** (see Diagram 3.4).

Usage

Select ***Map in Mono-colour***.

7.4. Editing Legends

Edit Legend allows the user to modify the legend properties.

Usage

Select ***Edit Legend*** to invoke the dialogue box **Discrete Legend Editor** (see Diagram 3.4). Edit the properties in the box.

7.5. Dragging to Zoom into the Map

Drag Zoom allows the user to define a square area on the map and amplify it to the size of the map window.

Usage

Select ***Drag Zoom*** first. Then move the mouse cursor into the map window and click the Left-Mouse button to define the upper-left corner of an area, and drag the cursor down and click again to define lower-right corner.

7.6. Zooming into a Map

Zoom In allows the user to amplify the map 4 times around a selected centre.

Usage

Select ***Zoom In*** first. Then, move the mouse cursor into the map windows and click the Left-Mouse button once to define a zooming centre.

7.7. Zooming out from the Map

Zoom Out allows the user to contract the map 4 times towards a selected centre.

Usage

Select ***Zoom Out*** first. Then, move the mouse cursor into the map window and click the Left-Mouse button once to define a zooming centre.

7.8. Resetting the Map

Reset allows the user to set the zoomed map back to the original size.

Usage

Select ***Reset***.

8. Drawing Graphs

SAGE enables the user to visualise objects as histograms, XY scatterplots, XW scatterplots, rankit plots, boxplots, lagged boxplots and matrix graphs in graph windows. Each graph window can display one or more graphs of one type, and consists of two main components: a menu for accessing the tools particular to the window, and a plotter for drawing the graphs.

SAGE provides the user with six tools, ***Histogram***, ***XY Scatter***, ***XW Scatter***, ***Rankit***, ***Box***, ***Lag Box*** and ***Matrix*** for creating graph windows in the corresponding graph types. These tools can be accessed from the menu commands under the menu option ***Graphs***.

The following section discusses the specification for each type of graph, the procedures to create graph windows and to plot graphs, and the tools for manipulating the graphs.

8.1. Histogram Graphs

Histogram allows the user to create a window to draw a single histogram for a set of values. A histogram graph consists of a number of consecutive bars (say *m* bars) of the same width. Each bar is made from a group of the values which fall into one of the *m* consecutive sub-intervals over the interval from the minimum to the maximum of the values. The height of a bar is equal to the number of values in a sub-interval.

A histogram can be presented in two formats according to the two ways of grouping the values:

1. Equal interval: divide the interval from the minimum to the maximum value into *m* equal intervals and assign the values in the same interval to a group;

2. Mean pivot: divide both sides pivoting at the mean into *m*/2 equal intervals. The length of an interval is equal to half the standard deviation of the values. Assign all values falling into the same interval to the same group. For values beyond either of the intervals at the two ends, assign each of them to the group corresponding to the interval it is closest to.

A histogram may be constructed from either the original values of a variable or the aggregates of the values based on a grouping index (See tool ***Create Dataset*** in Chapter 4 for details). In the latter case, ***Histogram*** can be used to assess the results of classification or regionalisation.

Diagram 8.1 shows a histogram window with a histogram in format 2. The x axis shows the intervals specified and the name of the attribute, while the y axis shows the frequency.

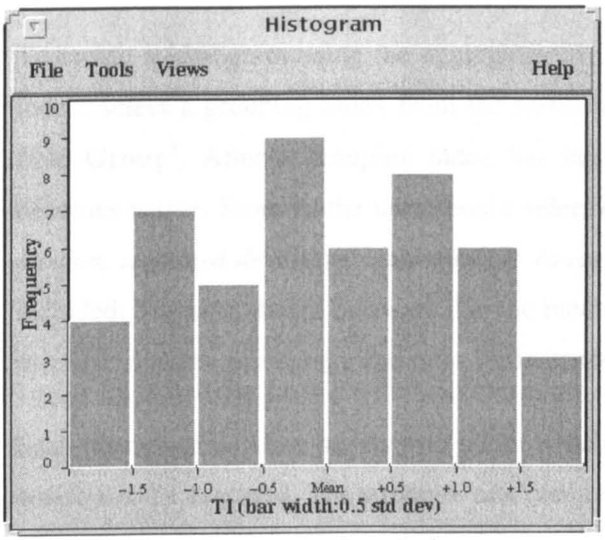
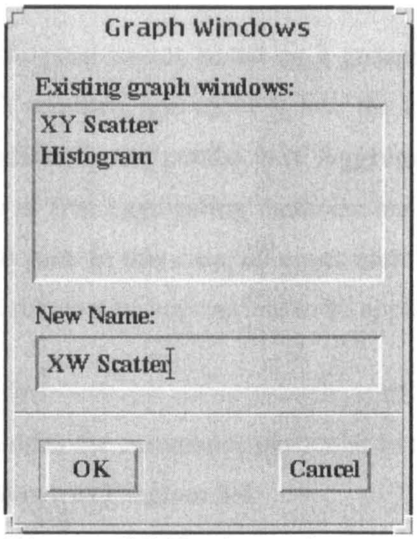


Diagram 8.1.A Histogram plot window.



Existing graph windows:

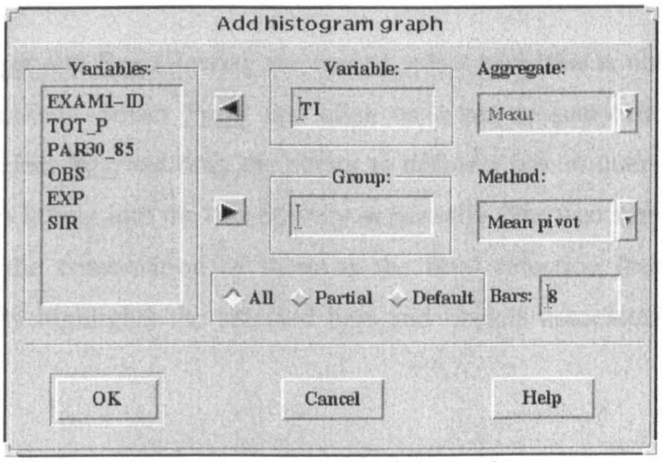
XY Scatter
Histogram

New Name: XW Scatter

Diagram 8.2.Naming a plot.

Usage

After selecting ***Histogram***, the user is asked to give a new name for the coming window in the dialogue box **Graph Windows** (Diagram 8.2). The scroll list **Existing Graph Windows** lists the names of existing graph windows, and the user must assign a new name for the coming window in the text field **New Name**. Note that this is required for every command under ***Graphs***.



Variables:

EXAM1-ID
TOT_P
PAR30_85
OBS
EXP
SIR

Variable: TI

Group:

Aggregate: Mexu

Method: Mean pivot

◆ All ◇ Partial ◇ Default

Bars: 8

Diagram 8.3. Dialogue box for defining histograms.

After the new name is accepted, the user is asked to define the histogram graph in the dialogue box **Add histogram graph** (Diagram 8.3). Select a variable from the scroll list **Variables** and a method, Equal interval or Mean pivot from the combo box **Method**. The user may change the number of bars in the text field **Bars**. With these settings, a histogram of the original values will be created after the user presses **OK**.

By default, all the original values are used in constructing the histogram. However, the user is allowed to select from the original values to form a histogram. To do this, check on **Partial**. Then after pressing **OK**, the user is led to select the cases. To use the values associated with the cases during the last query, check on **Default**. See Appendix A for interactively selecting cases.

To create a histogram using the aggregates of the original values based on a grouping index, select a grouping index from the scroll list **Variables,** and move it into the text field **Group**[8]. After a grouping index has been selected, the combo box **Aggregate** becomes active. From it, the user should select one of five aggregating methods: *mean, median, standard deviation, inter-quartile range* and *sum*. In this case, all cases must be included. The user should be aware that the methods may not be appropriate to be applied to values such as percentage and ratio for instance.

Each histogram window has its own menu which enables the commands particular to the window to be executed. The structure of a menu is shown as Diagram 8.4

File	*Tools*	*Views*	*Help*
Info	*Point*	*Grid*	
Exit	*Box*		

Diagram 8.4.Menu of the histogram plot window.

Under ***File***, there are two commands ***Info*** and ***Exit*** allowing the user to print out information about the graph in the text output window and to close the window respectively.

Tools includes two commands ***Point*** and ***Box*** allowing the user to select bars (this is not available when **Aggregate** is checked). Select ***Point*** and click on a bar to query the objects associated with the bar. Select ***Box*** and drag the cursor to define a box to query the objects associated with the bars falling into the box entirely or partially. The user may make many selections and take the combination of them as the final selection (see Appendix A). SAGE automatically highlights the selected bars and objects associated with them in other windows.

[8] There may or may not exist a grouping index in the scroll list **Variables**. To find out which variable is a grouping index, use the ***Properties*** tool under the ***Data*** menu to list the properties of each variable.

Grid under ***View*** allows the user to show grids over the histogram in the window. The grids are crossing lines drawn from ticks on both the X and Y axes.

8.2. XY Scatter Graphs

XY Scatter allows the user to create a window in which one or many XY scatter graphs can be plotted. An XY scatter graph consists of a number of points (x_i, y_i) where x_i and y_i are two values of two variables for the same case. An XY scatter graph also includes a regression line fitted to the points using least squares estimation.

Usage

Select ***XY Scatter***, and give a new name for the coming window in the same way as for ***Histogram***. Unlike ***Histogram***, after the user gives a unique name, ***XY Scatter*** creates an empty window with default X-Y axes. Now, the user needs to use the menu commands in the window to add graphs.

Each XY scatter window has its own menu as shown in Diagram 8.5.

File	*Edit*	*Tools*	*Views*	*Help*
Info	*Add*	*Select*	*Grid*	
Exit	*Delete*	*Drag Zoom*	*Legend*	
	Show	*Zoom In*		
	Hide	*Zoom Out*		
	Set Active	*Reset Zoom*		
	Propertie			

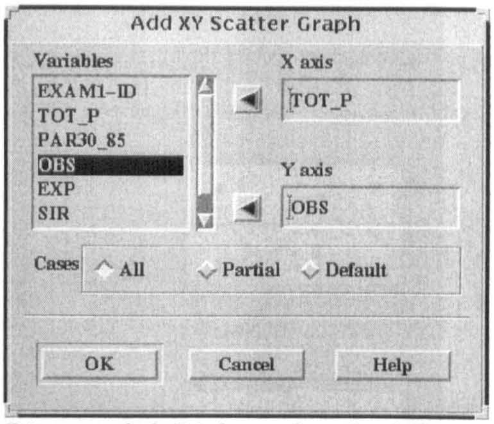
Diagram 8.5. Menu for the XY plot window.

File

Under ***File*** there are two menu commands ***Info*** and ***Exit*** allowing the user to print out information about every graph in the text window, and to close the window respectively.

Edit

There are six menu commands under ***Edit*** allowing the user to manage the graphs in the window. They will be discussed in turn as follows.



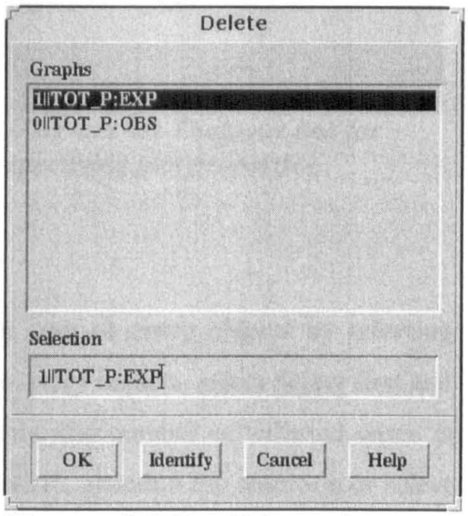
X axis: TOT_P

Y axis: OBS

Diagram 8.6. Dialogue box for defining XY plots

Add — add a graph into the window.

Select ***Add*** to invoke the dialogue box **Add XY Scatter Graph** (see Diagram 8.6) to define a scatter graph to be drawn. The text field **X axis** accepts a variable name for constructing the X dimension of a scatter graph. To assign the name of a variable into the field, select a name from the scroll list **Variables** and click on the arrow button (pointing to the right) next to **X axis**. Assign the name of another variable into the text field **Y axis** following the same procedure. To discard the variable named in a text field, say **X axis**, click on the arrow button (pointing to the left) next to **X axis**. In order to use the attribute values on selected cases, the user should check **Partial** or **Default**.



Selection: 1||TOT_P:EXP

Diagram 8.7. Specifying a plot to delete from a window.

Delete — delete a graph from the window.

Select ***Delete*** to invoke the dialogue box **Delete** (see Diagram 8.7). In order to delete a graph, the user needs to select its name from the scroll list in the dialogue box, and press **OK**. Before the selected graph is deleted, the user is asked to confirm the deletion. In order to make sure that the selected graph is the one the user wants to delete, press **Identify** to highlight it before clicking on **OK**.

Hide — hide a graph without deleting it.

Repeat the procedure above to hide a graph. Only unhidden graphs are listed in the scroll list.

Show — show a hidden graph.

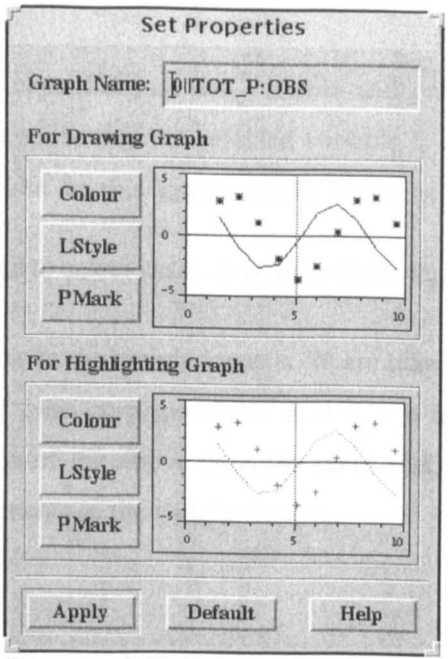
Repeat the procedure above to show a graph. Only hidden graphs are listed in the scroll list.

Set active — set a graph as the current active one on which the command ***Select*** (see ***Tools*** below) works.

Repeat the procedure above to set a graph as the current active one. Only unhidden graphs are listed in the scroll list.

Properties — set such graph properties as colours, line styles and point marks for drawing and highlighting the graph.

Repeat the procedure above to select a graph. After the selection is confirmed, the dialogue box **Set Properties** (see Diagram 8.8) appears and allows the user to change the default properties for drawing and highlighting in corresponding areas.



Set Properties

Graph Name: 0||TOT_P:OBS

Diagram 8.8. Dialogue box for specifying plot properties.

Tools

There are five commands under ***Tools*** allowing the user to query objects by selecting graph items (points) on them and to zoom graphs. To query objects, select ***Select*** first and then drag the cursor to include the points. After this, the number of selected cases is displayed in a message box. The user is asked to confirm, abandon the selection or select more. If not all points have been included, the user should continue to select more graph items. By accepting the selection, the objects associated with the selected graph items are highlighted.

In addition to the ***Select***, there are other commands to zoom on the plotter. ***Drag Zoom*** allows the user to drag the cursor to define a portion of the plotter and zoom it out. ***Zoom In*** and ***Zoom Out*** allow the user to zoom into the centre of the plotter and zoom out from the centre respectively. ***Reset Zoom*** enables the user to set the plotter to its original size.

Views

Two menu commands ***Grid*** and ***Legend*** are available. The first allows the user to turn on or off the grid associated with tics on both the x and y axes, while the second allows the user to show or hide an area (the legend area) which lists every pair of variables used for all graphs. The first named variable is the variable on the X axis, while the second is the variable on the Y axis.

8.3. XW Scatter Graphs

XW Scatter does the same thing as ***XY Scatter*** except that just one variable and one weights matrix are required to construct a graph. Suppose that the selected variable is x and the selected weights matrix is W, a scatter point in this graph would be (x_i, y_i), $y_i = \frac{1}{w_{i.}} \sum_{j=1}^{N} w_{ij} x_j$, $w_{i.} = \sum_{j=1}^{N} w_{ij}$. The user is allowed to select cases. For those cases that are not selected, the corresponding rows and columns in the weights matrix W are taken out of the computation. Moreover, if any selected case corresponds to $w_{i.} = 0.0$, it is treated as being deselected. Note that the evaluation of the above equation takes place after taking out all such cases and corresponding elements in the weights matrix.

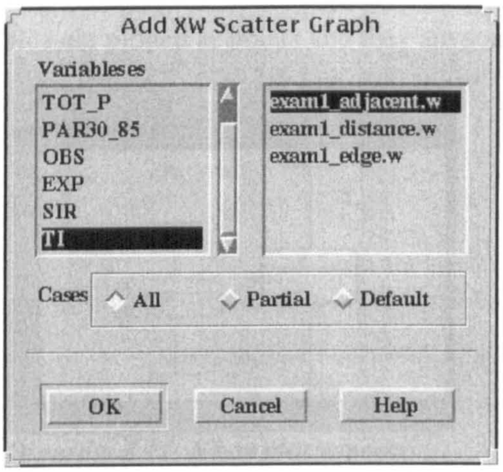


Diagram 8.9. Dialogue box for defining XW plots

The menu in a XW scatter window is identical to one in a XY scatter window except that ***Add*** under ***Edit*** invokes a slightly different dialogue box as shown in Diagram 8.9.

8.4. Rankit Graphs (visual test for normality)

Rankit does the same thing as XY Scatter except that just only one variable is required for a graph. The scatter points are constructed with the Z-scores of the variable on the Y axis and a sample of normal scores with the same parameters on the x axis. The menu in a rankit graph window is identical to one in a XY scatter window except ***Add*** under ***Edit*** invokes a slightly different dialogue box shown in Diagram 8.10.

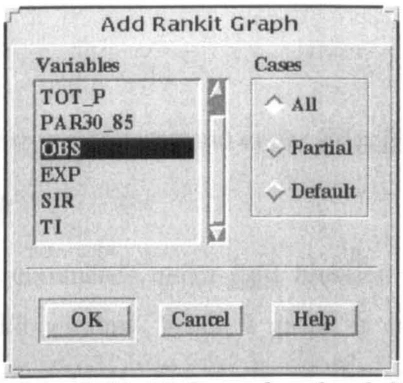


Diagram 8.10. Dialogue box for defining rankit plots

8.5. Box Graphs

Box allows the user to create a window in which one or many Box and Whisker graphs can be drawn. Each graph requires attribute values of at least five cases to be constructed, and consists of a number of components showing the distribution properties of a variable for selected cases (see Diagram 8.11). The values which are beyond the whiskers at either side are plotted as points and they are sometimes regarded as extreme values or outliers.

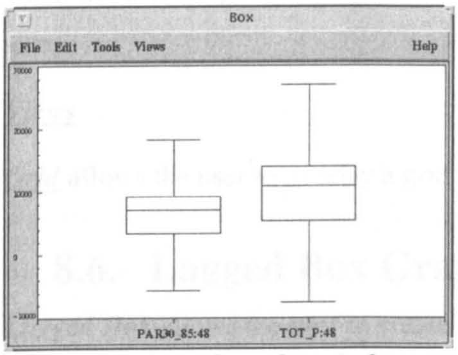


Diagram 8.11. A box plot window

Usage

Like ***XY Scatter***, ***Box*** creates an empty window when ***Box*** is selected. Then the user must use its menu (see Diagram 8.12) to add graphs.

File	*Edit*	*Tools*	*Views*	*Help*
Exit	*Add*	*Select*	*Grid*	
	Delete	*Zoom Out*		
	Show	*Zoom In*		
	Hide			
	Set Active			
	Properties			

Diagram 8.12.Menu for the box plot window.

File

There is one command under ***File***, ***Exit*** allowing the user to close the window.

Edit

All commands under ***Edit*** function exactly the same as their counterparts in ***Edit*** in a rankit window. When a graph is added in the window, it appears to the right to the existing plots. The name of the selected variable followed by the number of cases involved are displayed as X axis labels underneath.

Tools

Select allows the user to select objects from the current box plot in the same way as ***Select*** in a XY scatter window. So dragging an area crossing the central line of the current active box will pick up all values of the variable within that interval. There are only two commands, ***Zoom In*** and ***Zoom Out***. These allow the user to amplify part of the plotter when the user drags the cursor to define a segment of it and setting it back to its original size.

Views

Grid allows the user to overlay a grid over the plotter.

8.6. Lagged Box Graphs

Lagged Box allows the user to create a window on which one or many lagged box graphs can be drawn. A lagged box consists of a number of box plots for a variable, and they are arranged from the left to the right in the plotter. For a given case, the *i*th box from the left corresponds to the attribute values for those cases which are adjacent to the case at the *i*th order (Haining 1990 p 86 and 224). The user is asked to provide the order and select a case. Note that the number of boxes actually to be drawn is equal to the highest order at which the number of cases adjacent to the selected one is not less than five. A lagged box may help reveal the trend of a variable radiating from a selected case. Diagram 8.13 shows a lagged box window.

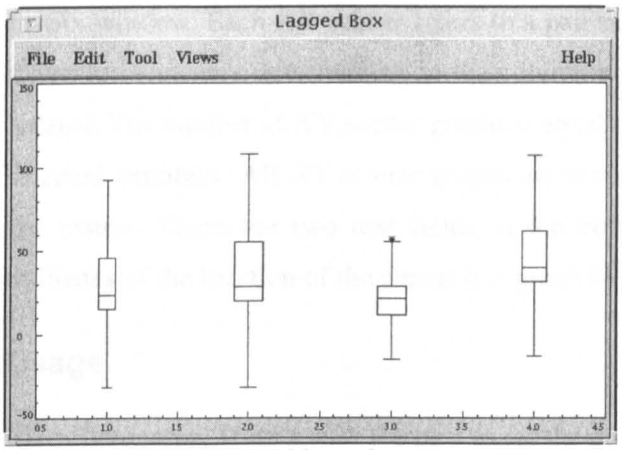
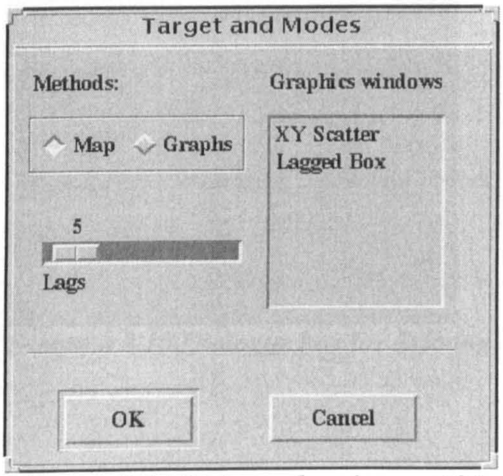


Diagram 8.13. A lagged box plot.

Usage

Like ***Box***, ***Lagged Box*** creates an empty window first. Then the user has to use commands in its menu to add graphs. A menu in a lagged box window is identical to a menu in a box window except that ***Add*** functions slightly differently.

To add a graph, select ***Add*** to invoke the dialogue box to select a variable. The user should select a variable. After accepting the setting, the user is asked to select a case from a window (either the map window or a selected graph window) and a lag order through the dialogue box shown in Diagram 8.14. By checking **Graphs**, the user must select one of the names of all available graphics windows listed in the scroll list **Graphics Windows**. The user should drag the scale **Lags** to define a lag order. After confirming this by pressing **OK**, the user should select a case from the target window.



Target and Modes

Methods: ◇ Map ◇ Graphs

Graphics windows: XY Scatter, Lagged Box

Lags: 5

Diagram 8.14.Dialogue box for specifying a lagged plot.

8.7. Matrix Graphs

Matrix allows the user to create a window in which a number of XY scatter graphs are drawn in a matrix fashion and displayed at the same time. Diagram 8.17 shows a typical matrix window. Each XY scatter refers to a pair of variables, the names appearing at the diagonal elements corresponding to the column and row where the XY scatter graph is located. The number of XY scatter graphs is equal to N*(N-1)/2 where N is the number of selected variables. All XY scatter graphs are arranged in the lower off-diagonal part of the matrix. There are two text fields in the bottom of the window showing the co-ordinates of the location of the cursor in a graph block.

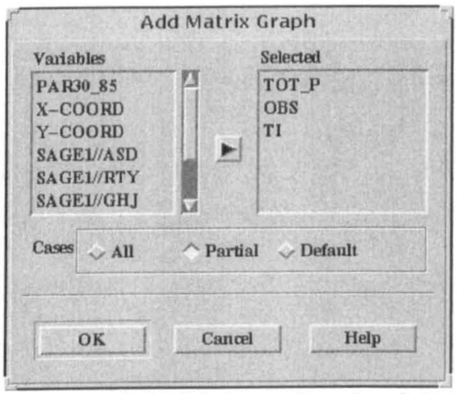
Usage

Like ***Histogram***, ***Matrix*** asks the user to set the parameters in the dialogue box shown in Diagram 8.17. The user has to select a number of variables (at least two) from the scroll list **Variables** and move them to the scroll list **Selected**. The user is allowed to select cases too.

A matrix window has its menu as shown in Diagram 8.15.

File	*Tools*	*View*
Exit	*Select*	*Motion*
	Zoom Out	
	Zoom In	

Diagram 8.15. Menu for the matrix plot window.

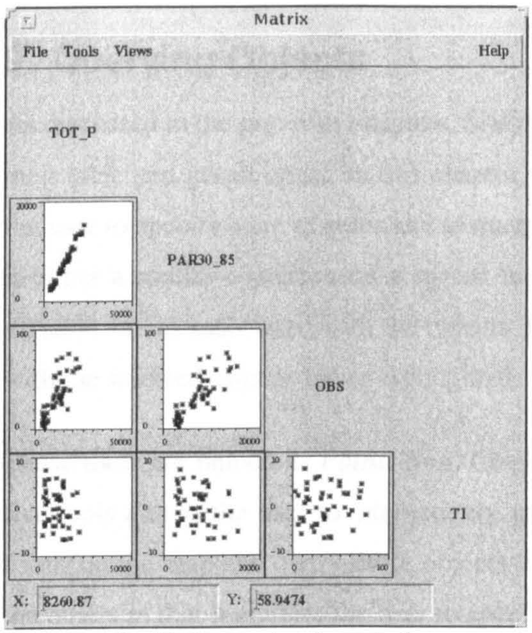

Add Matrix Graph

Variables
PAR30_85
X-COORD
Y-COORD
SAGE1//ASD
SAGE1//RTY
SAGE1//GHJ

Selected
TOT_P
OBS
TI

Cases ◇ All ◆ Partial ◇ Default

Diagram 8.16.Dialogue box for defining a matrix plot.



X: 8260.87 Y: 58.9474

Diagram 8.17. A matrix plot window.

There is only one command ***Exit*** under ***File*** allowing the user to dismiss the window. Three commands under ***Tools*** function the same as their counterparts under ***Tools*** in a box window except that they work on each individual XY scatter graph (Diagram 8.15). ***Motion*** under ***View*** allows the user to turn on or off the facility to show co-ordinates relating to the selected attributes in the text fields **X** and **Y** in the bottom of the window.

9. Querying Objects

As discussed in the previous chapters, SAGE allows the user to query objects by selecting their table and graph items. In this chapter, we discuss another six SAGE tools allowing the user to specify a set of rules and to query objects satisfying the rules. In SAGE, a rule is either a spatial condition on a spatial feature of the objects or a logical condition on variable values associated with the objects. Each tool results in all visual items associated with the selected objects being highlighted.

These tools are named as ***Point***, ***Box***, ***Circle***, ***Polygon***, ***Lag*** and ***SSQL***. Each of the first five tools allows the user to interactively specify one spatial rule in one of five different forms on the map and query those objects which meet the rule. The last tool differs from the others in that it enables the user to specify a set of both spatial and logical rules in the form of expressions and to query objects accordingly. Each tool can be accessed from the corresponding menu command under the menu option ***Queries***.

9.1. Querying Objects with *Point*

Point allows the user to click at a location on the map and query the object(s) which cross the point.

Usage

Select ***Point*** and click the right mouse button at a location in the map.

9.2. Querying Objects with Box

Box allows the user to define a square area on the map and query all the objects which fall either entirely within or partially within the area.

Usage

Select ***Box*** and define the upper left-hand corner of a square area at the first click, drag the cursor down and then click to define the lower right-hand corner of the area.

9.3. Querying Objects with Circle

Circle allows the user to define a round area on the map and query all the objects which fall either entirely within or partially within the area.

Usage

Select ***Circle*** and define a centre at the first click, drag the cursor and define the circle's radius at the second click.

9.4. Querying Objects with Polygon

Polygon allows the user to define a polygon area on the map and query all the objects which fall either entirely within or partially within the area.

Usage

Select ***Polygon*** and click consecutively on the map to define each vertex and type "9" from keyboard to close the polygon.

9.5. Querying Objects with Lag

Lag allows the user to select an object and query all the objects which are adjacent to it either up to a given order inclusively (Within) or at the given order (At Order).

Usage

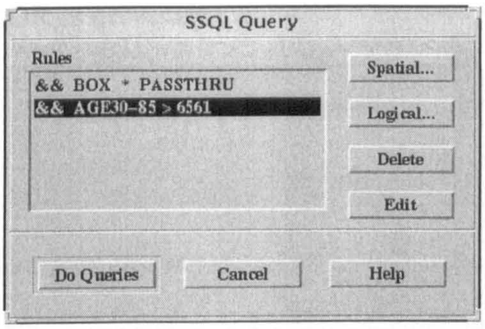
Select either ***Within Order*** or ***At Order*** under ***Lag***. Then the user is asked to define a lag in a dialogue box. Finally, click on the map to select an object.

9.6. Querying Objects with *SSQL*

Unlike the first five commands which allow the user to query objects satisfying a single spatial rule, ***SSQL Query*** allows the user to define a set of spatial and logical rules and query the objects satisfying the rules. The rules are processed one by one in the order they are defined. The objects satisfying all rules previous to the current rule and the current rule are combined based on the merging mode AND or OR. If the mode is AND, those objects are selected if they satisfy all rules previous to the current rule and the current rule. If the mode is OR, those objects are selected if they meet either all rules previous to the current rule or the current rule. Appendix B gives a summary of the syntax of both spatial and logical rules and provides some examples. For a full description of the syntax, see ESRI (1994).

Usage

Select ***SSQL Query*** to invoke the dialogue box **SSQL Query** (Diagram 9.1). The push buttons, **Spatial**, **Logical**, **Delete** and **Edit** allow the user to define a spatial and logical rule, delete and edit an existing rule listed in the scroll list **Rules**.

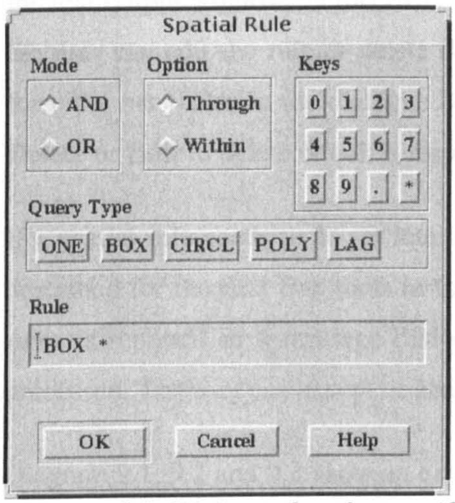


&& BOX * PASSTHRU

&& AGE30-85 > 6561

Diagram 9.1. Dialogue box for specifying a query.

In order to define a spatial rule, the user needs to invoke the dialogue box **Spatial Rule** (see Diagram 9.2) by pressing **Spatial**. A spatial rule is equivalent to one of the first five tools discussed above but in expression form. The user can use the push buttons in **Query Type**[9] and **Keys** to define a rule. With the radio buttons either **Through** or **Within** set, objects are selected if they fall either entirely or partially (Through) or entirely (Within) inside the area. Click on the radio buttons **AND** or **OR** allow to define how the result of the current rule is to be combined with the result of the previous rules.



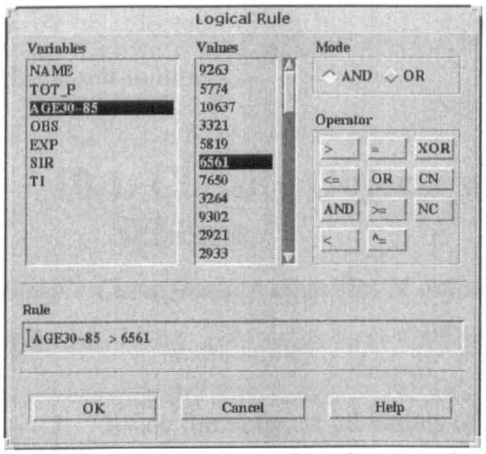
Rule: BOX *

Diagram 9.2. Dialogue box for specifying a spatial query.

In order to define a logical rule (i.e. a logical expression), the user needs to invoke the dialogue box **Logical Rule** (see Diagram 9.3) by pressing **Logical**. A logical rule is combination of variable names, values, arithmetic and logical operators or logical rules,

[9] Button **One** is used to define a rule equivalent to the one defined by ***Point***.

and it has the same syntax as a logical expression in ARC/INFO (see ESRI 1994). When a variable is selected in the scroll list **Variables**, its name is inputted into the text field **Rule** and its values are listed in the scroll list **Values**. The user can select values from the list to input them into the rule, or type the values from the keyboard. The user can select a button in **Operators** and type from the keyboard to input a logical operator. Set **AND** or **OR** as necessary.



Variables: NAME, TOT_P, AGE30-85, OBS, EXP, SIR, TI

Values: 9263, 5774, 10637, 3321, 5819, 6561, 7650, 3264, 9302, 2921, 2933

Rule: AGE30-85 > 6561

Diagram 9.3. Dialogue box for specifying a logical query.

Note that a space is required between an operator and an operand.

If there is a syntax error in a rule, SAGE will indicate in which rule the error occurs in a message box when the rule is processed. Some further information may be printed out in the text output window. In this case, the dialogue box **SSQL Query** will come back, and the user can edit the rule or delete the rule. This can be done by first selecting the rule from the list **Rules** in the dialogue box **SSQL Query** and then click on the push-button **Delete** or **Edit** to delete or edit it respectively.

If a spatial rule needs to be set interactively, the user should do this in the same way as described for the first five tools in this chapter. The number of objects satisfying all the rules is reported in a message dialogue box after all the rules have been successfully processed. The user can accept or abandon the query.

Diagram 9.1, 9.2 and 9.3 show an example of defining a set of two rules. The first rule is spatial whilst the second rule is logical. The first rule is also interactive so that the user will be asked to define a box on the map. As the merging mode for the second rule is AND (shown as &&), so all areas which have over 6561 people in the age 30 to 8 and which fall partially or entirely into the box on the map will be selected.

10. Classification

SAGE provides three tools, ***Simple Classification***, ***Hierarchical Classification*** and ***Heuristic Classification*** for classifying N (= the number of all objects) objects into K (K < N) groups. Each tool produces a grouping index (see ***Create Dataset*** in Chapter 4 for details) and saves it as a new variable. However, the last tool is able to produce such a grouping index which also defines a partition of the whole region. All three tools do NOT work on selected objects but ALL objects. In the remainder of this chapter, each tool is discussed in turn.

10.1. Classify Objects with a Simple Classification Scheme

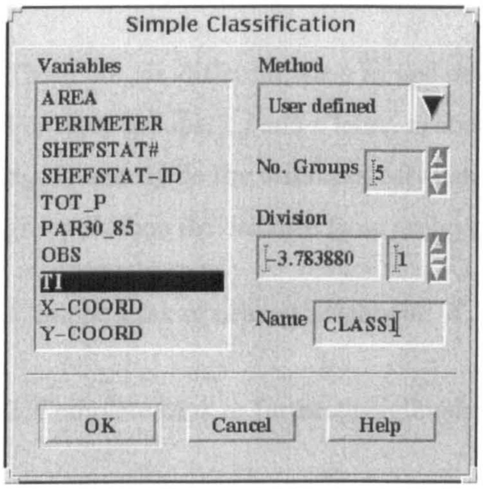
Simple Classification classifies N objects into K classes on the basic of single attribute as follows:

1. divide the range of all values of a given attribute into K consecutive intervals;
2. assign a unique integer (<=K) to those objects whose attribute take on values falling into a given interval;

The range of values of a given attribute can be divided by means of an *equal interval* method, an *equal number* method or a *user-defined interval* method. The first method divides the range into K intervals of equal length, while the second method splits the range into K intervals such that the number of objects assigned to one group is (almost) the same as the number in any other group. The third method allows the user to define the boundary of intervals.

Usage

Select ***Simple Classification*** to invoke the dialogue box **Simple Classification** (Diagram 10.1). The user must select a variable from the scroll list **Variables**, and assign an integer value for the number of groups required in the text field **No. of Groups**. One of three methods described above can be selected from the combo box **Method**. If the user defined interval method is selected, the user can define all K-1 divisions in the text field **Division**, where K is the number of groups. A set of initial divisions are provided by SAGE dividing the range of the values equally. The user can only change a division, say *i*th division, within the range of the division (or the minimum) on its left (i.e. (*i*-1)th division), and the division (or the maximum) on its right (i.e. (*i*+1)th division). The user may provide a name for the new grouping index, otherwise SAGE will assign one automatically.


Simple Classification

Variables

AREA
PERIMETER
SHEFSTAT#
SHEFSTAT-ID
TOT_P
PAR30_85
OBS
TI
X-COORD
Y-COORD

Method: User defined

No. Groups: 5

Division: -3.783880 1

Name: CLASS1

Diagram 10.1. Dialogue box for performing simple classification.

Note that the number of groups will be reduced accordingly if there are intervals, derived from the first or third method, in which no value falls.

10.2. Classify Objects with a Hierarchical Classification Scheme

Hierarchical Classification performs a multivariate classification and classifies N objects into K groups in hierarchical fashion. Given l variables, ***Hierarchical Classification*** first constructs a similarity matrix, D, each item of which, d_{ij} (i,j = 1 to N), is a distance between the ith and jth objects. This is obtained by taking the variable values for an object as locations in a l dimensional space. A distance D(u,v) between any two locations say u and v in this space is defined using one of three types of distance as follows:

1. Euclidean distance D(u,v) = $\sqrt{\sum_{i}^{l}(u_i - v_i)^2}$;

2. Euclidean squared distance D(u,v) = $\sum_{i}^{l}(u_i - v_i)^2$; and

3. Absolute distance D(u,v) = $\sum_{i}^{l}|u_i - v_i|$.

With agglomerative methods, ***Hierarchical Classification*** produces a hierarchical tree starting with N groups each with a single object. At each of N-1 stages, it then merges two groups, which are nearest to each other, to form a new group. The distances between the new group to others are updated. This process goes on until all objects are in a single group. Finally, K groups are reconstructed from the tree information.

There are six different ways to update the distances between the new group and others. For three groups, *i, j* and *k* let n_i, n_j and n_k be the number of objects in each group and let d_{ij}, d_{ik} and d_{jk} be the distances between the groups. Let groups *j* and *k* be merged to give group *jk*, then the distance from group i to *jk*, $d_{i,jk}$ can be computed in the following ways.

1. Single Link or nearest neighbour: $\boldsymbol{d_{i,jk} = min(d_{ij}, d_{ik})}$;

2. Complete Link or furthest neighbour: $\boldsymbol{d_{i,jk} = max(d_{ij}, d_{ik})}$;

3. Group average: $\boldsymbol{d_{i,jk} = \frac{n_j}{n_j + n_k} d_{ij} + \frac{n_k}{n_j + n_k} d_{ik}}$;

4. Centroid: $\boldsymbol{d_{i,jk} = \frac{n_j}{n_j + n_k} d_{ij} + \frac{n_k}{n_j + n_k} d_{ik} - \frac{n_j n_k}{(n_j + n_k)^2} d_{jk}}$;

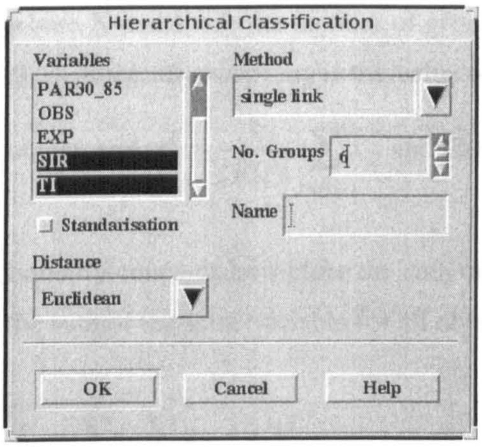
5. Median: $\boldsymbol{d_{i,jk} = \frac{1}{2} d_{ij} + \frac{1}{2} d_{ik} - \frac{1}{4} d_{jk}}$;

6. Minimum variance: $\boldsymbol{d_{i,jk} = \{(n_i + n_j)d_{ij} + (n_i + n_k)d_{ik} - n_i d_{jk}\} / (n_i + n_j + n_k)}$;

For further details see Everitt (1974) and NAG manual for subroutines G03EAF, G03ECF and G03EJF.

Usage

Select ***Hierarchical Classification*** to invoke the dialogue box **Hierarchical Classification** (see Diagram 10.2). Click on variables to select them for classification. To deselect any of them, click again. To use values which are scaled by the standard deviation of the values rather than the original, check the toggle **Standardisation**. The user should select a proper merging method from the combo box **Method** and assign a name for the grouping index, otherwise SAGE will use the default merging method and assign a name for the index.


Hierarchical Classification

Variables: PAR30_85, OBS, EXP, SIR, TI

Method: single link

No. Groups: 6

Name:

□ Standarisation

Distance: Euclidean

Diagram 10.2. Dialogue box for performing hierarchical classification.

10.3. Classify Objects with a Heuristic Classification Scheme

Heuristic Classification classifies N objects to K groups (either classes or regions) using a K-means method. The method requires two basic elements - an initial partition with K groups from N objects and a combined objective function derived from the criteria for measuring whether one partition is better than another. Starting from an initial partition, the method aims to find a partition so that the objective function is 'optimal' through an iterative procedure as follows. Firstly, it selects each object in turn, allocates the object experimentally to every other group and works out the value of the objective function; secondly, it finds out such an allocation that results in the greatest or the first improvement in the objective function; thirdly, if there is such an allocation, it confirms the allocation (called a swap) to form a new partition. The steps above are repeated till no allocation or swap can be found to improve the objective function, or the number of iterations reaches a given limit. An iteration stands for that every object from the first to the last has been tried just once. This procedure is also described in a diagram in Appendix D.

An objective function is an analytical form of one or more criteria. There are three criteria for the classification procedure: homogeneity, equality and compactness. Homogeneity relates to how similar variable values are within any group found by the classification. It is measured by the sum of squares of within group variances measured on the variables as follows:

$$f_h = \sum_{i=1}^{K} \sum_{l=1}^{L} \sum_{j \in R(i)} (x_{jl} - x_{l.}^{i})^2$$

where K and L are the number of groups and selected attributes respectively, $R(i) = \{k \mid O_k$ is in the ith group$\}$, x_{jl} is the value of the lth selected attributes associated with the jth object, and $x_{l.}^{i} = \frac{1}{|R(i)|}\sum_{j\in R(i)} x_{jl}$ and $|R(i)|$ is the number of objects in the ith group.

Equality concerns how close the sum of a selected variable for all objects of a group is to the sum of the same variable for all objects across all groups. It is measured by:

$$f_e = \sum_{i=1}^{K}\left(\sum_{j\in R(i)} x_j - x.\right)^2$$

where K is the number of groups, $R(i) = \{k \mid O_k$ is in the ith group$\}$, x_j is the value of a selected attribute associated with the jth object, and $x. = \frac{1}{K}\sum_{j=1}^{K} x_j$.

Compactness concerns how close each object is to other members of the same group in a spatial sense. Compactness is measured by the sum of squares of within group variances measured on the X and Y location co-ordinates defining the centroid of each object.

$$f_c = \sum_{i=1}^{K}\sum_{l=1}^{2}\sum_{j\in R(i)} (x_{jl} - x_{l.}^{i})^2$$

where K is the number of groups, $R(i) = \{k \mid O_k$ is in the ith group$\}$, $x_{j1} = x_j$ and $x_{j2} = y_j$ and x_j and y_j are the X and Y co-ordinates of the jth object, and $x_{l.}^{i} = \frac{1}{|R(i)|}\sum_{j\in R(i)} x_{jl}$ and $|R(i)|$ is the number of objects in the ith group.

An objective function may be defined as the sum of weighted homogeneity, equality or compactness functions as follows: $f_o = w_h f_h + w_e f_e + w_c f_c$, where w_h, w_e, and w_c are weights corresponding to the individual functions. The decision rule is based on the assumption that the smaller the objective function f_o is, the better the final partition.

Due to the fact that the criteria are competitive, it is often the case that although the objective function appears to be improved by a swap, one of the individual functions may get worse. However, if a swap is only considered acceptable when the objective function and all the individual functions improve, the process often stops at a very poor partition.

In order to tackle this situation, a threshold, *thrhd*, is introduced for each individual function to relax the strict constraints. Therefore, a swap is said to be acceptable if

1. $f_o^a - f_o^b < 0.0$ and

2. $(f_s^a - f_s^b) / f_s^b < thrhd$

where s takes on h, e, or c, and f^b and f^a are the value of an objective function "before" and "after" a swap.

The classification requires an initial partition. SAGE provides methods to produce different initial partitions. There are three methods available. The first method produces an initial partition by randomly selecting K objects as seeds, then takes the seeds as an initial uncompleted partition, and lets K initial groups be "grown" so that every object is assigned to a group. The principle of a growing procedure is illustrated as follows. Suppose that there are *l* objects still to be allocated to K uncompleted groups which consist of the remaining N-*l* objects. Take each of the *l* objects in turn and allocate it to one of K groups to which the object is closer than any other group. The "closer" is measured by the value of the one of the individual functions on given objects. If the homogeneity is taken into account, the individual function is the one associated with it. If not and the equality is taken into account, the function is the one associated with the equality, otherwise the function is the one associated with the compactness.

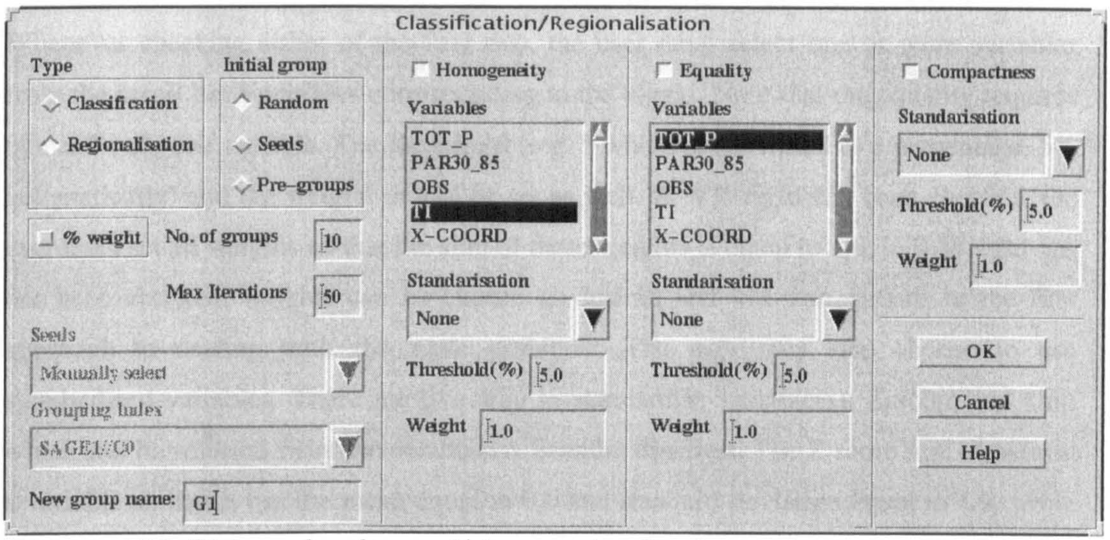
If regions rather than clusters are required, every group to which the object is assigned must form a region after the allocation. Repeat above till every object has been allocated to a group. The second method employs the same growing procedure used by the first method but allows seeds to be selected manually. The third method uses the previous grouping index as an initial partition.

The scales of different variables may cause another problem, that is, one individual function dominates the objective function and therefore the swapping decision. The first approach to deal with this problem is to weight the individual functions. The second approach is to standardise variables so that the individual functions are at the same scale. The third approach is similar to the first method. However, a swap is accepted if

1. $\sum_s w_s (f_s^a - f_s^b) / f_s^b < 0.0$ and

2. $(f_s^a - f_s^b) / f_s^b < thrhd$

where s takes on h, e, or c, and f^b and f^a are the value of an objective function "before" and "after" a swap. These approaches may be used together as indicated below. For further details, see Wise *et al* 1997.



Classification/Regionalisation

Type
○ Classification
● Regionalisation

Initial group
● Random
○ Seeds
○ Pre-groups

□ % weight
No. of groups 10
Max Iterations 50
Seeds: Manually select
Grouping Index: SAGE1//C0
New group name: G1

□ Homogeneity
Variables: TOT_P, PAR30_85, OBS, TI, X-COORD
Standarisation: None
Threshold(%) 5.0
Weight 1.0

□ Equality
Variables: TOT_P, PAR30_85, OBS, TI, X-COORD
Standarisation: None
Threshold(%) 5.0
Weight 1.0

□ Compactness
Standarisation: None
Threshold(%) 5.0
Weight 1.0

Diagram 10.3. Dialogue box for specifying a regionalisation session.

Usage

Select ***Heuristic Classification*** to invoke the dialogue box **Classification/ Regionalisation** (see Diagram 10.3). The user should decide whether to group the units to clusters or regions by checking on the correct toggle **Classification** or **Regionalisation**. Choose **Random**, **Seeds** or **Pre-groups** for selecting the initial partition. If **Random** is checked, the number of groups in the text field **No. of Groups** must be provided. If **Seeds** is checked, the user will be guided to select seeds after pressing **OK**. The number of groups will be equal to the number of the seeds being selected. **Seeds** is only available when **Regionalisation** is checked. If the **Pre-groups** is checked, the user should select a previous grouping index variable in the combo box **Grouping Index**[10], otherwise the first one in the box will be used. The user may set an integer value in the text field **Max Iterations** instead of using the default 50. The user should provide a unique name for the new grouping index variable, otherwise SAGE will assign one automatically.

[10] A grouping index which does not represent a region partition will be converted to a grouping index which does by simply assigning objects, which are in the same group and form a contiguous region, into the same group. To find out how many regions a grouping index forms, see **Properties** in chapter 4.

In order to use the sum of weighted percentage changes to decide each swap (the third approach to dealing with the scale problem described above), check the toggle button **%Weight**.

After setting all the above, the user needs to set the criteria to be used. The user can set one or more criteria by checking toggles, **Homogeneity**, **Equality** or **Compactness**. Whenever checking either of the first two, the user must select one or more variables from the scroll list **Variables** corresponding to the toggle. Note that the equality requires one and only one variable. The **threshold** (e.g. 5 which is converted to a percentage, 5% automatically) and the **weight** should be set as well. If **%Weight** has been checked, the user must set all weights so that the sum of these weights is equal to 1.0. If **%Weight** has not been checked, weights can be chosen as desired and this corresponds to the first approach to dealing with the scale problems. The user may also choose to use standardised variables. There are two way to standardise variable(s): Z score and Unit which can be selected from the combo box **Standardisation**. The Z score will transform a variable so that it has the mean equal to 0.0 and standard deviation equal to 1.0, while the Unit will transform a variable so that every value of it is within 0.0 and 1.0 and these two values correspond to its minimum and maximum values before transformation. See Table 4.1 for details.

As the heuristic classification normally takes a while to complete for a large data set, feedback is provided to inform the user of the current state of the classification process. A graphical message box shows the performance of the regionalisation process at regular intervals during the process. The user is able to stop the process if desired.

Diagram 10.4 shows an example of the graphical message box for a classification which takes into account homogeneity and equality. Three graphs show the performance in the objective function, the homogeneity function and the equality function respectively[11]. For each graph, its X axis is labelled as **Sequence** and its Y axis is labelled as **Relative change**, and a point (x,y) in the graph is defined as follows: x is equal to m - standing for the *m*th time period and y is equal to V_m/V_s, where V_m is the value of the corresponding function at the beginning of the *m*th time period and V_s is the value of the function at the first time period.

[11] Note that if **%Weight** is set, no total objective function will be shown in the box.

It is evident that although the objective function and the equality function are improving at each step, the homogeneity function does not follow the right track at all. It is clear that the equality function is dominating the process. In this case, the user should stop the process by pressing the **Stop** button, redefine weights and thresholds and start again. Although the process has been terminated, a grouping index is still produced and can be saved as a new variable for later use.

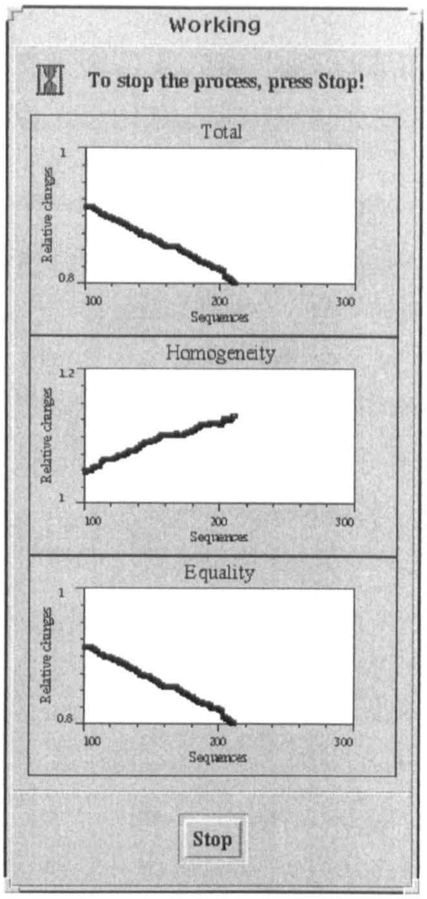

To stop the process, press Stop!

Diagram 10.4. Dialogue box for reporting the progress of a regionalisation session..

In order to make reasonable adjustments of thresholds and weights, the user should refer to the values of each individual function and the objective function which are printed out in the text output window. The user should be aware that these values may only suggest rough scales of the functions since they depend on the initial partition. If another run with the new settings shows satisfactory improvement in functions, the user may leave the process till it finishes. In this case, the graph message box will remain on the screen. However, the **Stop** button in the message box will change to **Done** to allow the user to discard the window. How long the process takes partly depends on the maximum number of iterations defined in the text field **Max Iterations**.

Note that in the case of regionalisation this tool may fail to generate an initial partition if the whole region is actually divided into several disconnected parts and in each apart there is no area being selected as a seed. If this is the case, SAGE will show messages to inform the user. To overcome this problem, the user could modify the SDWM to actually link the separate parts together and start the regionalisation again. Alternatively, the user could manually select seeds so that for each disconnected part there is at least one area as a seed[12].

[12] To avoid any potential problem relating to creating a new data set, the user is advised to use the second method.

11. Statistics

This chapter will discuss tools performing the following types of statistical analysis:

1. computing distribution properties in both the classical and spatial senses, and performing correlation analyses for pairs of variables;
2. computing local indicators of spatial association for variables;
3. computing robust estimates by means of empirical Bayes and Kernel estimation;
4. performing linear regression analysis with or without taking spatial specification into account;
5. fitting a generalised linear regression model with Poisson errors;

The remainder of this chapter is divided into five sections. Each section discusses the tools performing one group of statistical analyses mentioned above and their usage.

11.1. Basic Statistics

There are four tools for performing basic statistics. They are ***Descriptive***, ***Correlation***, ***Chi-Square*** and ***K-S tests***. These tools can be accessed under the menu ***Statistics/Basic Statistics***.

11.1.1. Descriptive statistics

Descriptive computes a number of statistics for each selected variable and selected cases, and reports the statistics in the text output window. The statistics include the sample mean, median, skew, upper and lower quartiles, standard deviation, Moran I (see Cliff and Ord 1973) and Getis-Ord statistics (see Getis and Ord 1992) in the global sense as well as the Shapiro-Wilk's W statistic. The significant values for both the Moran I and the Getis-Ord statistics (z-values) under assumption of a normal distribution are also reported. The Getis-Ord statistics assume that the variable only takes on non negative values.

Unlike other statistics, the Moran I and Getis-Ord measure the global spatial association of a variable under a spatial structure specified by a weights matrix. For the Moran I, a positive and significant z-value indicates positive spatial autocorrelation, whereas a negative and significant z-value indicates negative spatial autocorrelation. For Getis-Ord, a positive and significant z-value indicates spatial clustering of high values, whereas a negative and significant z-value of it indicates spatial clustering of low values (see Anselin 1995). The Shapiro-Wilk's W statistic can be used to provide a formal test for the normality of a given variable.

Except the last three statistics which are optional, other statistics are always reported whenever the tool is applied.

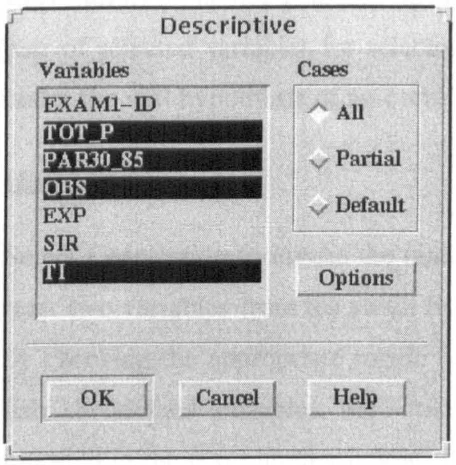


Diagram 11.1. Dialogue box for describing properties of data..

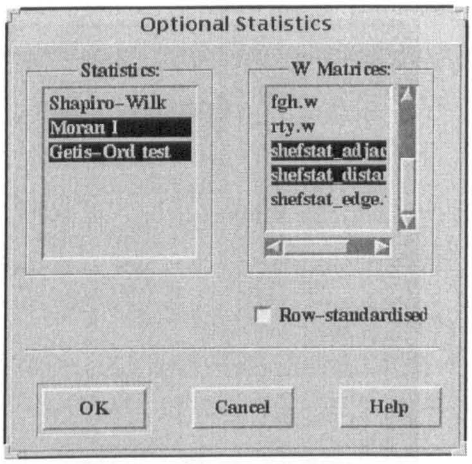


Diagram 11.2. Dialogue box for specifying optional statistics.

Usage

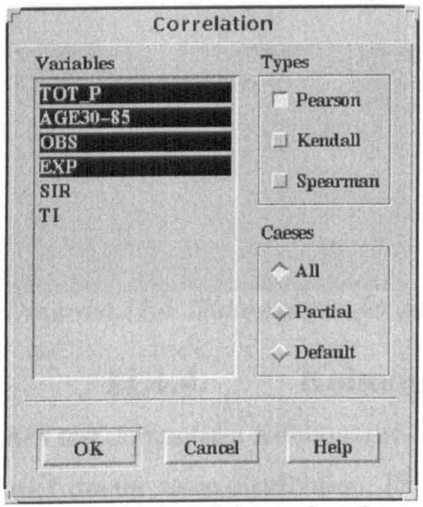
Select Descriptive to invoke the dialogue box Descriptive (see Diagram 11.1). The user should select one or more variables from the scroll list Variables. In order to perform optional statistics, the user must selects them from the scroll list Statistics in the dialogue box Optional Statistics (see Diagram 11.2) which can be invoked by pressing Options. If either Moran I or Getis-Ord test has been selected, the user must select at least one weights matrix from the W Matrices. If more than one weights matrix is selected, these two statistics are computed for each weights matrix in turn. To select cases interactively or the cases of the last selection, the user must switch on the radio button Partial or **Default** (See Appendix A).

11.1.2. Correlation analysis

Correlation computes Pearson, Spearman and Kendall correlation coefficients for every pair of selected variables for selected cases. It also performs significance tests for them under the null hypothesis of no-correlation.

Usage

Select ***Correlation*** to invoke the dialogue box **Correlation** (see Diagram 11.3). Select at least two variables from the scroll list **Variables**. The user should select the coefficients by checking the appropriate toggle buttons **Pearson**, **Spearman** and **Kendall**. For each pair of selected variables, the coefficient and the significance of the coefficient are reported in the text output window. To select cases interactively, the user must switch on the radio button **Partial**.



Correlation

Variables

TOT_P
AGE30-85
OBS
EXP
SIR
TI

Types

□ Pearson
□ Kendall
□ Spearman

Caeses

◇ All
◇ Partial
◇ Default

Diagram 11.3. Dialogue box for performing correlation analysis.

11.1.3. χ^2 test

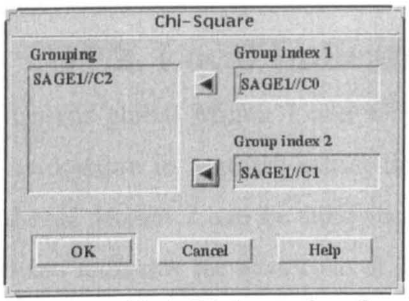
Chi_Square computes χ^2 for testing relationship between two attributes that have been classified into groups (see Diagram10.1). The null hypothesis is of no relationship between the two attributes.

Suppose one attribute has been classified into K1 classes, the other into K2 classes, so that each case falls into one of the K1xK2 cells in the cross tabulation as determined by its class grouping on each of the two attributes. Let O_{ij} be the number of cases falling into the cell of the table on the *i*th row and *j*th column, and let $E_{ij} = \dfrac{\sum_{j=1}^{K1} O_{ij} \times \sum_{i=1}^{K2} O_{ij}}{\sum_{j=1}^{K1}\sum_{i=1}^{K2} O_{ij}}$, then

$$\chi^2 = \sum_{j=1}^{K1}\sum_{i=1}^{K2}(O_{ij} - E_{ij})^2 / E_{ij}$$

Usage

Select Chi-Square to invoke the dialogue box Chi-Square (see Diagram 11.4). Select two grouping indices listed in the scroll list Grouping. Press OK to start the computation. Besides the $\chi 2$ value, the degrees of freedom (i.e. (K1-1)(K2-1)) and the significance, this tool also reports the count and its expectation (i.e. Eij) for each cell in the text output window.



Grouping: SAGE1//C2

Group index 1: SAGE1//C0

Group index 2: SAGE1//C1

Diagram 11.4. Dialogue box for performing Chi-Square test.

11.1.4. Kolmogorov-Smirnov test

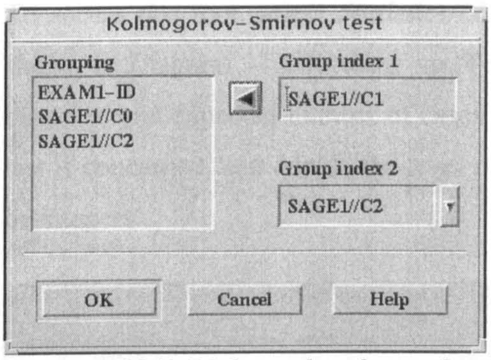
K-S test computes a Kolmogorov-Smirnov test on two samples to determine whether they are from the same distribution. The test requires that one attribute has been classified into two classes representing the two samples and the other attribute has been classified into K ordered classes. Each case is allocated to one of the 2K cells of the 2xK table. Let N1 and N2 be the number of cases falling into class 1 and class 2 of the attribute that has been classified into two classes. So N1 + N2 = N where N is the total number of cases. For a one-tailed Kolmogorov-Smirnov test a statistics D is computed as follows:

$D = \max[S_{1,j} - S_{2,j}]$, j = 1,2,..,K; $S_{1,j} = {}^{n1}\!/_{N1}$, $S_{2,j} = {}^{n2}\!/_{N2}$.

where n#(# = 1,2) equal to the total number of cases from the first class to the *j*th class (the cumulative sum) corresponding to the first sample and the second sample respectively.

Usage

Select ***K-S test*** to invoke the dialogue box **Kolmogorov-Smirnov test** (see Diagram 11.5). Select a grouping index from the scroll list and move it into the text field **Group index 1**. Then choose a grouping index of 2 groups in the combo box **Group index 2**. Press **OK** to perform the Kolmogorov-Smirnov test. The results to be reported include D value, N1, N2, and significance.



Group index 1: SAGE1//C1

Group index 2: SAGE1//C2

Diagram 11.5. Dialogue box for performing Kolmogorov-Smirnov test.

11.2. Local Indicators of Spatial Association

Unlike global Moran I and Getis-Ord statistics that are designed to spot the spatial association in a global sense, three tools named as ***Besag-Nevell test, Local Getis*** and ***Local Moran I*** can be used to explore local spatial association. Each tool computes a local indicator for each case of a selected variable and the significance of the z-value of each local indicator under the assumption of a normal distribution. The inferences based on these statistics should be made against the null hypothesis of no corresponding local spatial association (see Besag and Nevell 1991, Getis and Ord 1992, Anselin 1995).

Note that the validity of the assumption of a normal distribution may not be held in a number of circumstances. For a detailed discussion and a test for significant local spatial association based on a conditional randomisation approach, see Getis and Ord (1992) and Anselin (1995). The user should also be aware that the presence of global association may have an influence on the moments of the distribution of the local indicators (Anselin 1995).

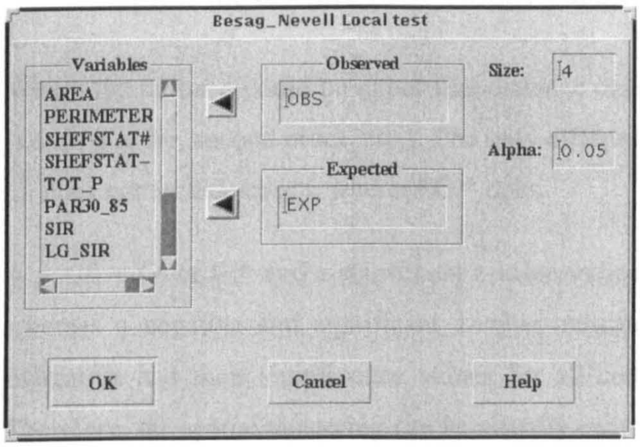
To access the tools, select menu commands ***Statistics/LISA/Besag_Nevell local test*** ***Statistics/LISA /Local Getis*** and ***Statistics/LISA/Local Moran I***.

11.2.1. Besag-Nevell Test

The Besag-Nevell test tests whether points, to a given order, are close enough to be statistically significant. In SAGE the points are located in centoids of the polygons, and so the test should be used for testing rare events at fine geographical scales. For details of this test, the user is referred to Besag and Nevell (1991).

Usage

To access this tool, select ***Statistics/LISA/Besag_Nevell test***. This causes the dialogue shown in Diagram 11.6 to show up. The user is required to specify two variables, the observed and expected number of cases and may override Size, the order with which the test is concerned, and Alpha, the level of significance. Note that the observed cases must be integers.



Besag_Nevell Local test

Observed: OBS

Expected: EXP

Size: 4

Alpha: 0.05

Diagram 11.6. Dialogue Box for the Besag-Nevell local test.

This tool will produce four variables in the SAGE table, BESAGP#, BESAGN#, BESAGOBS#, and BESAGEXP#, where # is a number used for generating unique variable names. For each area (corresponding to a row in the table),

BESAGP# -- the Poisson probabilities of the Besag and Nevell test;

BESAGN# -- the number of areas involved in computing the test;

BESAGOBS# -- the accumulated cases minus 1 over the involved areas;

BESAGEXP# -- the accumulated expected cases minus 1 over the involved areas.

A file, called besa-index.dat, will be created in the user's working space (e.g. /tmpdata/gg1jsm.sage) contains the indexes of areas involved in computing the test for each area.

Note that this will only work for all cases.

11.2.2. Local Getis indicators

Local Getis computes two forms of local indicators Gi and Gi* based on the following two equations respectively (see Getis and Ord 1992, 1995).

$$Gi(d) = \frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j} \quad j \neq i;$$

$$Gi^*(d) = \frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j};$$

$$w_{ij}(d) = \begin{cases} 1, \text{ if the distance between case j and case i is not greater than } d; \\ 0, \text{ if the distance between case j and case i is greater than } d; \end{cases}$$

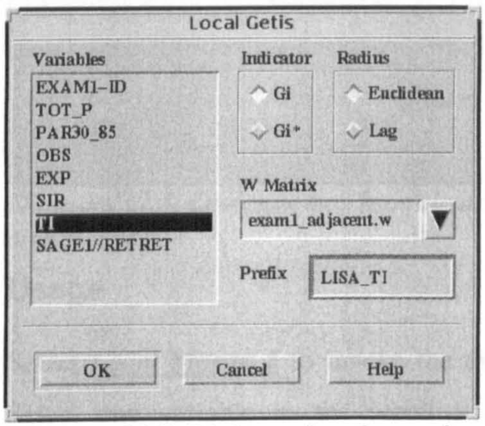
Where the radius ***d*** could be either real distance or topological distance, that is lag order (i.e. first order, second order, etc.). The only difference between the two indicators is that Gi does not include case i, whereas Gi* does.

A positive Gi or Gi* and a significant z-value indicates spatial clustering of high values, whereas a negative and significant z-value indicates spatial clustering of low values. Indicators and their significance values for all cases are saved as two new variables. Therefore, the spatial clustering can be visually examined on the map by using commands such as ***SSQL Queries***.

Usage

Select ***Local Getis*** to invoke the dialogue box **Local Getis** (see Diagram 11.7). Select one variable from the scroll list **Variables**. The user can choose to compute either Gi or Gi* indicators by checking on either **Gi** or **Gi*** and a type of distance to be used. The user must select a weights matrix from the combo box **W Matrix** The indicators and significance values are saved under names XX_Z and XX_P respectively. By default, SAGE gives a name GIn (GI0, GI1,...) to substitute XX. The user can provide a name in the text field **Prefix**.

After pressing **OK**, the user is prompted to provide either a lag or a real distance (radius) accordingly. In the case of using the lag order, this tool first converts the selected weights matrix to the SDWM, and then computes the statistics based on the SDWM. Just before finishing the computation, it restores the previous SDWM. See section 5.4 for details about the SDWM.



W Matrix: exam1_adjacent.w

Prefix: LISA_TI

Diagram 11.7. Dialogue box for performing the local Getis-Ord tests.

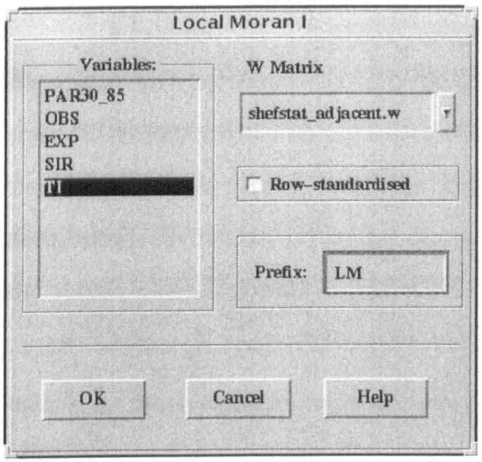
11.2.3. Local Moran I indicators

For each case of a variable, ***Local Moran*** computes a local Moran indicator based on the following formula (see Anselin 1995).

$$\boldsymbol{I}_i = z_i \sum_j w_{ij} z_j$$

Where *i* and *j* range from 1 to N. The z_i is the *i*th value after standardisation (the mean equal to 0 and the standard deviation equal to 1). The w_{ij} is the element of a weights matrix corresponding to case i and case j.

The local Moran I can be interpreted as an indicator of local pattern. A positive value indicates a spatial clustering of *similar* values, a negative value indicates a spatial clustering of *dissimilar* values. The local Moran I indicator for each case and its significance value are computed and saved as two new variables.



Variables: PAR30_85, OBS, EXP, SIR, TI

W Matrix: shefstat_adjacent.w

□ Row-standardised

Prefix: LM

Diagram 11.8. Dialogue box for calculating the local Moran's I values and their significance.

Usage

Select ***Local Moran I*** to invoke the dialogue box **Local Moran I** (see Diagram 11.8). Select one variable in the scroll list **Variables** for computing the indicators and significance statistics. The user should select a weights matrix in the combo box **W Matrix**, otherwise the default one appearing in the box will be used. The indicators and significance values are saved in the same way as discussed in section 11.2.1, except that the default prefix is in the form of LMn (LM0, LM1,...). The user may provide his/her own prefix in the text field **Prefix**.

11.3. Robust Estimation

Mapping relative rates of disease incidence or mortality plays an important role in studying area-based health data. For area i, a commonly mapped estimate of the relative rate θ_i is (O_i /E_i), where O_i and E_i are the observed and expected incidence or mortality counts. Alternatively, the standardised incidence or mortality ratio $SMR_i = (O_i /E_i)*100$ is computed. However, SMR_i (i = 1,2,...,N) does not take into account differences in the reliability of estimates that derive from different base populations. In the case of areas with small populations, SMR_i (i = 1,2,...,N) becomes very sensitive to individual cases. The same may be true for other ratios.

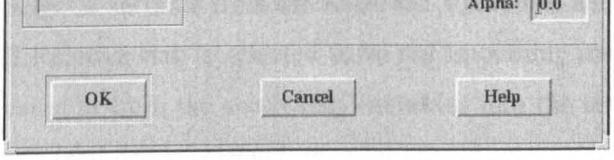
SAGE offers two tools, ***Bayesian*** and ***Kernel*** for computing robust relative rates by means of empirical Bayesian estimation and Kernel estimation. The resulting estimates are saved as new variables so that they can be mapped by mapping tools. To access the tools, select menu ***Statistics/Robust Estimation/Bayesian*** and ***Statistics/ Robust Estimation/Kernel***.

11.3.1. Empirical Bayes estimation

Bayesian allows the user to calculate two types of empirical Bayes estimates of relative risks (O_i/E_i) assuming different distributions for the relative risks. The first assumes that the relative rates are independent and identically distributed (iid) following a gamma distribution with scale parameter α and shape parameter ν, while the second assumes that the relative rates are iid following a multivariate log-normal distribution, that is, $\theta_i = \exp(\beta_i)$ where $\{\beta_i\}$ follows a multivariate normal distribution with mean μ and dispersion matrix Σ. Estimates are computed using the procedures described in Clayton and Kaldor (1987).

Usage

Select ***Bayesian*** to invoke the dialogue box **Empirical Bayesian Estimation** shown in Diagram 11.9. Select a variable from the scroll list **Variables** and assign it into the text field **Observed**. Then select another and assign it into the text field **Expected**. The user should select the appropriate prior distribution for calculating the empirical Bayesian estimates, otherwise SAGE will use the gamma distribution as the default. If the gamma distribution is selected, the user should provide initial values for α and ν in the **Nu** and **Alpha** text fields for estimation (see Clayton and Kaldor 1987). The results times 100 are stored as a new variable (the Bayes adjusted SMR).


Variables: EXAM1-ID, TOT_P, PAR30_85, SIR, TI

Observed: OBS

Expected: EXP

Distribution: Gamma

Nu: 0.0

Alpha: 0.0

Diagram 11.9. Dialogue Box for performing Empirical Bayesian estimations.

11.3.2. Kernel estimation

Kernel allows the user to perform three types of smoothing: *Relative risk, Mean smoother* and *Median smoother*. The first one requires two variables, an observed variable and an expected variable, and calculates the relative risks for each case which are the results of the smoothed observed over the smoothed expected as:

$$\frac{O_i}{E_i}\bigg|_{smooth} = \frac{\sum_j w_{ij} O_j}{\sum_j w_{ij} E_j}$$, where j = 1, 2, ... N and w_{ii} is set to 1.0.

This formula can be derived from the work of Bethell (1990) and Cislaghi (1996), which is based on a kernel estimation method (Silverman 1986), with a fixed smoothing factor to first order adjacent elements of the weights matrix for each case.

The second smoothing method takes only one variable and passes a mean filter through the variable across every case. The smoothed result is defined as:

$$O_i\big|_{mean} = \frac{\sum_j w_{ij} O_j}{\sum_j w_{ij}}$$, where j = 1, 2, ... N and w_{ii} is set to 1.0.

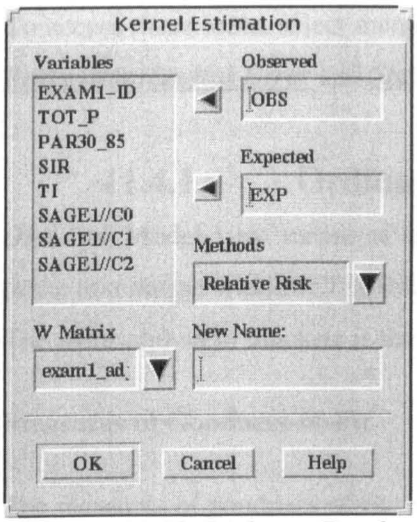
Like the second method, the third smoothing method takes only one variable but passes a median filter through the variable across every case. The smoothed result is defined as:

$O_i\big|_{median} = Median(O_j)$ for j = i and j adjacent to i up to a specified order inclusively.

Note that if the number of adjacent areas is less than five the mean of them is used instead of the median. The results of each method will be saved as new variables named as KERN# where # is an ordinal number.

Usage

Select ***Kernel*** to invoke the dialogue box **Kernel Estimation** shown in Diagram 11.10. Select a variable from the scroll list **Variables** and assign it into the text field **Observed**. If Relative risk is selected to be the smoothing method, the user must select an expected variable from the scroll list **Variables** into the text field **Expected**. If either of the first two methods is selected, the user should select a weight matrix. If the Median smoother is selected, the user will be asked to provide a lag (e.g. 1, 2, 3,...) after pressing OK.



Observed: OBS

Expected: EXP

Methods: Relative Risk

W Matrix: exam1_ad

New Name:

Diagram 11.10. Dialogue Box for specifying smoothing operations.

11.4. Linear Regression Analysis

SAGE provides three tools ***OLS***, ***Spatial Error*** and ***Spatial Lag*** allowing the user to perform linear regression analysis by fitting three models respectively:

Model One – Ordinary Linear Regression Model

$$y = X\beta + \varepsilon$$

Model Two – Linear Regression Model with Spatially Correlated Error Term

$$y = X\beta + \eta$$

$$\eta = \lambda W \eta + \varepsilon$$

Model Three – Linear Regression Model with Spatially Lagged Dependent Variable

$$y = \rho W y + X\beta + \varepsilon$$

Where y is a dependent variable (Nx1 vector). X is a NxP matrix with N cases on P explanatory (including the constant term) variables. β is a Px1 vector with P regression coefficients, and ε is random error with mean 0.0 and variance matrix $\sigma^2\mathbf{I}$. These models are discussed in Anselin (1988) and Haining (1990).

Note that a model with spatially lagged *independent* variables can be fitted using Model 1 after first creating Wx variables using the tool in the ***Data*** menu. These variables then become additional independent variables (see Haining 1990).

To access these tools, select menu ***Statistics/ Linear Regression/OLS***, ***Statistics/ Linear Regression/Spatial Error*** and ***Statistics/ Linear Regression/Spatial Lag***

11.4.1. Ordinary linear regression model

OLS fits Model 1 by means of least squares estimation and reports a series of statistics in the text output window. The statistics could be classified into the following three types. The name shown in brackets is the name displayed in the text window.

Measures of Goodness-of-Fit

The measures of goodness-of-fit reported include: R2 (R-Square), adjusted R2[13], the sums of squares of the residuals, the value of the maximised log likelihood (ML), Akaike's Information Criterion (AIC) and the Schwartz Criterion (SC) (Akaike 1981). The last three measures based on the maximum likelihood approach to estimation are not necessary when only Model One is fitted. They become useful when they are used to compare the fit of this model to Model Two or Model Three.

Hypothesis tests

A significance test by means of Student's t-test is reported for each individual regression coefficient under a null hypothesis that the population regression coefficient is zero. The F-test (the value and probability) is provided to assess the significance of the model. Both the least squares estimate (OLS-SGM) and the maximum likelihood estimate (ML_SQM) of σ^2 are reported, that is $\frac{e^t e}{N-P}$ and $\frac{e^t e}{N}$, where e is the set of residuals in vector form. N and P are the number of cases and the number of parameters in the model respectively.

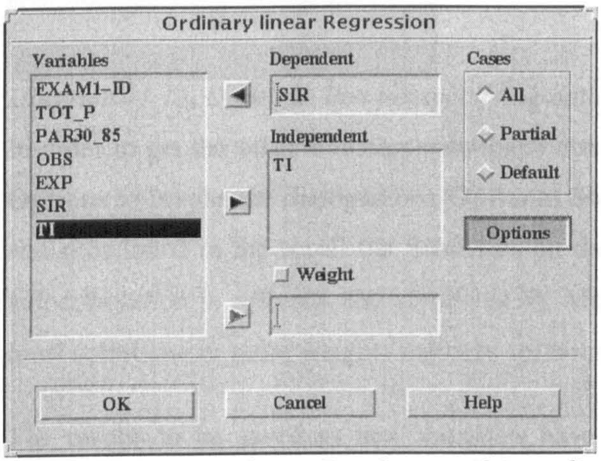
Model Specification

Only one statistic, Shapiro-Wilk's W is reported for testing normality of residuals, whereas three tests are reported for diagnosing mis-specification of the model due to the presence of spatial dependency. The null hypothesis for the test is that Model One has uncorrelated errors. The first test is the Moran I test (Moran I) for measuring spatial autocorrelation in regression residuals. The second test is a Lagrange Multiplier test (LM error) for measuring the spatial dependency formalised by Model Two with a null

[13] Adjusted R2 = R2 - (1 -R2)(P-1)/(N-P).

hypothesis λ =0, while the third one is a Lagrange Multiplier test (LM lag) for measuring the spatial dependency formalised by Model three with a null hypothesis ρ = 0.

Besides reporting these statistics, ***OLS*** allows the user to save such results as residuals, predictors, leverages, internal and external studentised residuals and Cook's distances for all cases as new variables. They can be used to test the validity of assumptions of ordinary least square estimation, detect outliers and study the influence of cases. Heteroscedastic and correlated errors might also be detected by plotting external studentised residuals against predicted values of the dependent variable or predictors with the ***XY Scatter*** tool (see Weisberg 1985, p130-133).



Variables: EXAM1-ID, TOT_P, PAR30_85, OBS, EXP, SIR, TI

Dependent: SIR

Independent: TI

Cases: ● All ○ Partial ○ Default

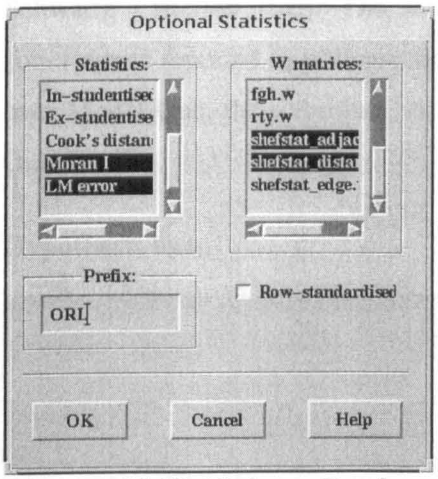
□ Weight

Diagram 11.11. Dialogue Box for specifying a linear regression model.

Usage

Select ***OLS*** to invoke the dialogue box **Ordinary Linear Regression** as shown in Diagram 11.11. To select a variable as a dependent variable into the text field **Dependent**, click on the variable in the scroll list **Variables** first and press the arrow button next to the **Dependent**. Select independent variables and move them into the scroll list **Independent** one by one in the same way.

With the toggle **Weight** ticked, the user could select a variable into the text field underneath and let ***OLS*** perform weighted least squares estimation where the variance-covariance matrix is a diagonal matrix, and each diagonal element (i.e. a weight) takes the inverse of the corresponding value of the variable. The larger the weight attached to a case the smaller the variance and the more importance attached to that case in the estimation. An example of this is where the weight reflects, say, the population of the area. The larger the population of an area, the more importance the user might want to attach to that area in the estimation. If there is a value equal to 0.0, the case related to it will be automatically excluded in the estimation. See Weisberg (1985) and NAG manual for subroutine G02DAF for further details.



Prefix: ORI

☐ Row-standardised

Diagram 11.12. Dialogue Box for specifying optional statistics.

In order to get the other statistics mentioned above, the user must select the push-button **Options** to invoke the dialogue box **Optional Statistics** (see Diagram 11.12) and choose statistics listed in the scroll list **Statistics** on the left. If any statistic for testing spatial autocorrelation is selected, the scroll list **W Matrices** becomes available, and the user must select one or more weights matrices for computing the statistics.

The results to be saved as new variables have names XX_Y. XX is a prefix for all variables, while Y takes RES (residuals), FIT (fitted values) LEV (leverages), ISR (internal studentised residuals), ESR (external studentised residuals) or COOK(Cook's distance). Although ordinary linear regression works with selected cases, it can only save these results as new variables if all cases are included.

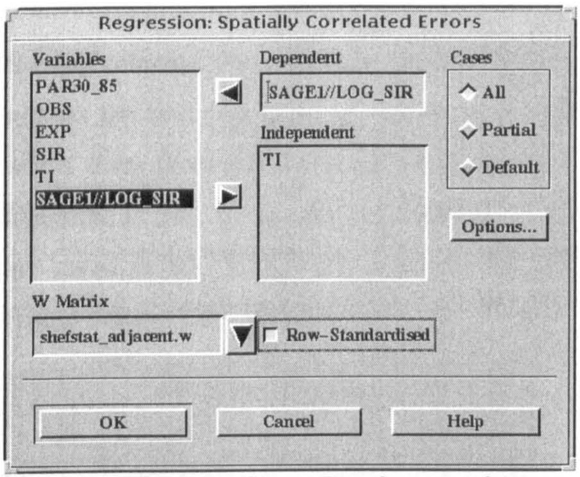
11.4.2. Linear regression model with spatially correlated error term

Spatial Error fits Model Two by means of maximum likelihood estimation and reports a series of statistics in the output text window. The statistics could be classified into the following three types. The names shown in brackets are the names displayed in the text window. The procedure adopted for fitting the model is similar to the one discussed by Anselin (1988 pp182-183).

Measures of Goodness-of-Fit

The measures of goodness-of-fit reported include: pseudo-R2 (R-Square)[14], the value of the maximised log likelihood (ML), Akaike's Information Criterion (AIC) and the Schwartz Criterion (SC). The last three measures should be compared with their counterparts reported in ordinary linear regression. If Model Two is an improved model over Model One, their absolute values will be greater than their counterparts in Model One.

Hypothesis tests



Regression: Spatially Correlated Errors

Variables: PAR30_85, OBS, EXP, SIR, TI, SAGE1//LOG_SIR

Dependent: SAGE1//LOG_SIR

Independent: TI

Cases: ● All ○ Partial ○ Default

W Matrix: shefstat_adjacent.w

□ Row-Standardised

Diagram 11.13. Dialogue Box for specifying a regression model with an autocorrelated error term.

A significance statistic by means of an asymptotic t-test is reported for each individual regression coefficient including the autoregressive coefficient λ under a null hypothesis that the population regression coefficient is zero. The maximum likelihood estimate (ML_SQM) of σ^2 is reported as well.

Spatial Dependency

A Likelihood Ratio (LR) test on the spatial autoregressive coefficient λ is included. This LR test corresponds to twice the difference between the log likelihood in Model Two and the log likelihood in Model One with the same data, and indicates the improvement of the fit.

Besides reporting these statistics, ***Spatial Error*** can also report the variance-covariance matrix for coefficients and Shapiro-Wilk's W for residuals. Residuals, fitted values and

[14] The ratio of the sums of squares of the predicted values over the sums of squares of the observed values for the dependent variable.

P-leverages (Haining 1994b) for all cases can be saved as new variables. All these statistics are optional.

Usage

Select ***Spatial Error*** to invoke the dialogue box as shown in Diagram 11.13. Select a dependent variable and independent variables in the same way as described in the usage of ***OLS***. The user selects a weights matrix from the comb box **W Matrix**. Check **Row-Standardised** on, if a row standardised weight matrix is required.

Having selected the variables, ***Spatial Error*** can fit the model to the data, but it only reports the first two types of statistics. In order to get other statistics, the user needs to select them from the scroll list **Statistics** in the dialogue box **Optional Statistics** (see Diagram 11.14). To invoke the dialogue, press the push-button **Options**. New variables are saved as XX_Y. XX is a prefix for all variables, while Y takes RES (residuals), FIT (fitted values) LEV (P-leverage).

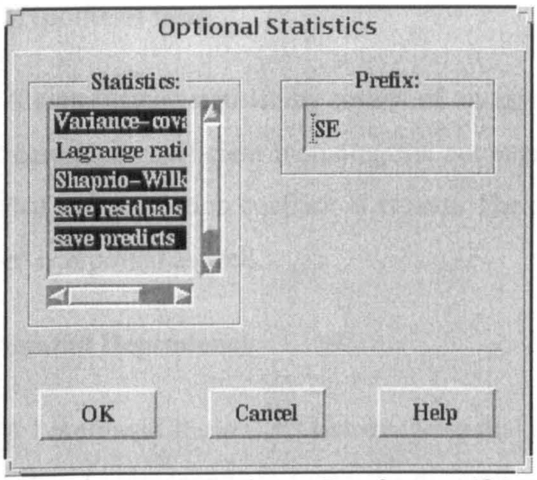


Diagram 11.14. Dialogue Box for specifying optional statistics.

Spatial Errors also works with selected cases. In this case, the weights matrix selected for the fitting is tailored so that only row and column elements corresponding to the selected cases are remained. The user is not allowed to edit this tailored matrix. So the user should modify the weights matrix before calling this tool (see Chapter 5). Moreover, ***Spatial Errors*** will not save the results (e.g. residuals) as new variables but report them in the text output windows.

11.4.3. Linear regression model with spatially lagged dependent variable

Spatial Lag fits Model Three by means of maximum likelihood estimation and reports a series of statistics in the text output window. The statistics are also classified into the following three types. The names shown in brackets are the names displayed in the text window. The procedure adopted for fitting the model is similar to the one discussed by Anselin (1988 pp182).

Measures of Goodness-of-Fit

The measures of goodness-of-fit reported include: pseudo-R2 (R-Square), the value of the maximised log likelihood (ML), Akaike's Information Criterion (AIC) and the Schwartz Criterion (SC). The last three measures should be compared with their counterparts reported in ordinary linear regression. If Model Three is an improved model over Model One, their absolute values will be greater than their counterparts in Model one.

Hypothesis tests

A significance statistic by means of an asymptotic t-test is reported for each individual regression coefficient including the autoregressive coefficient ρ under a null hypothesis that the population coefficient is zero. The maximum likelihood estimate (ML_SQM) of σ^2 is reported as well.

Spatial Dependency

A Likelihood Ratio (LR) test on the spatial autoregressive coefficient ρ is included. This LR test corresponds to twice the difference between the log likelihood in Model Three and the log likelihood in Model One with the same data.

Besides reporting these statistics, ***Spatial Lag*** can also report the variance-covariance matrix for coefficients and Shapiro-Wilk's W for residuals. Residuals, fitted values and leverages[15] for all cases can be saved as new variables. All these statistics are optional.

Usage

See the usage of ***Spatial Error***.

[15] The leverages are the diagonal elements of $(I-\rho 1W)^{-1}X(X^TX)^{-1}X^T(I-\rho 1W)$, where $\rho 1$ is the estimates of ρ.

11.5. Generalised Linear Regression

SAGE provides a tool ***Poisson Errors*** allowing the user to fit a generalised linear regression model with Poisson errors by means of maximum likelihood estimation[16]. The estimates of the β coefficients are obtained by iterative weighted least squares. A generalised linear model is specified by following three stages:

1. The *random component*: a dependent variable y, from a distribution from the exponential family with $E(y) = \mu$;

2. The *systematic component*: a set of independent variable $x_1, x_2, \ldots x_k$ produces a linear predictor given by $\eta = \sum_{1}^{k} x_j \beta_j$;

3. The *link* between the random and systematic components: $\eta = g(\mu)$, where g() is a monotonic differentiable function;

The maximum likelihood estimates of the β coefficients are obtained by iterative weighted least squares.

A generalised linear regression model with Poisson errors is a model in which the values of a dependent variable, y_i, i = 1,2,...N, are from a Poisson distribution:

$$\frac{\mu^y e^{-\mu}}{y!}$$

and a link function could take one of the following five forms:

1. exponent link: $g(x) = x^\alpha$ for a constant α;

2. identity link: $g(x) = x$;

3. log link: $g(x) = \log(x)$;

4. square root link: $g(x) = x^{1/2}$; and

5. reciprocal link: $g(x) = 1/x$;

[16] This model is provided only of all the generalised linear regression models that exist because the Poisson model is a natural model in the context of the occurrence of disease counts.

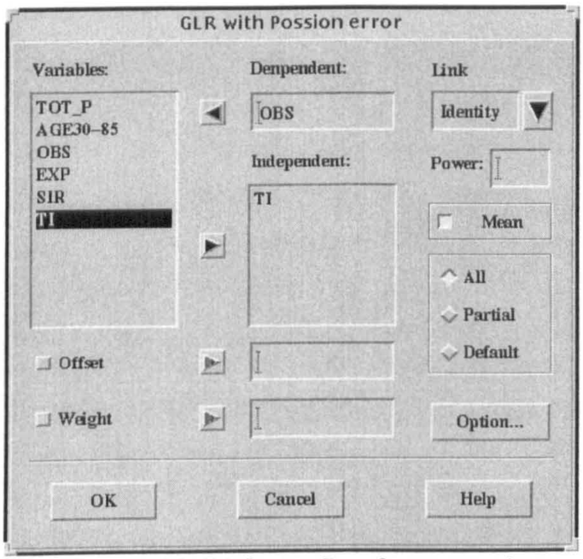
Poisson Errors reports the deviance which can be used to assess the fit of the model by comparing the difference in deviance between nested models. The difference has, asymptotically, a Chi-square distribution with degrees of freedom given by the difference in the degree of freedom associated with the two deviances. It also reports the estimate for each coefficient and its significance.

Besides these, the user is allowed to save such results as deviance residuals, fitted values and leverages for all cases as new variables. For further details, see the NAG manual for G02GCF, and McCullagh and Nelder (1989).

To access the tools, the user needs to select menu ***Poisson Errors*** under the menu command ***G-Linear Regression*** which is in turn under the menu option ***Statistics*** in the system menu.

Usage

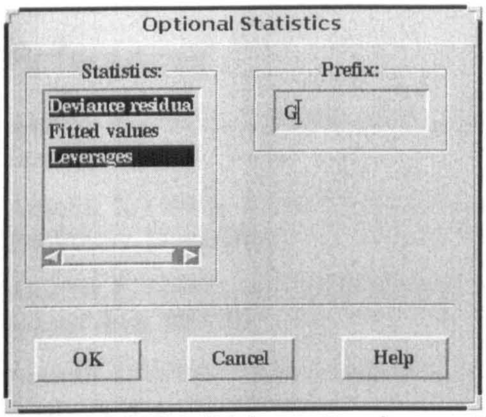
Select ***Poisson Errors*** to invoke the dialogue box **GLR with Poisson Errors** shown in Diagram 11.15. Select a dependent variable and independent variables in the same way as described in the section 11.4.1.



Denpendent: OBS
Independent: TI
Link: Identity

Diagram 11.15. Dialogue Box for regression model with a Poisson error term.

With the toggle **Weight** ticked and a variable assigned in the text field, ***Poisson Errors*** can perform weighted least squares estimation with the values of the variable as the weights. The user is allowed to specify linear predictors with a known offset variable. To do this, check **Offset** and then select a variable as an offset variable. See Weisberg (1985) and the NAG manual Mark 16 for further details.



Prefix: G

Diagram 11.16. Dialogue Box for specifying optional statistics.

In order to store deviance residuals, fitted values or leverages as new variables, the user needs to specify them in the dialogue box **Optional Statistics** (see Diagram 11.16). To save them as new variables, the user should tick **All**. The results to be saved as new variables have the names as XX_Y. XX is a prefix for all variables, while Y takes RES (residuals), FIT (fitted values) LEV (leverages). Generalised linear regression works with selected cases too, but it can only save these results as new variables if all cases are involved.

References

Akaike, H. (1981). "Likelihood of a model and information criteria", Journal of Econometrics 16, 3-14.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordercht.

Anselin, L. (1992). *SpaceStat tutorial: A workbook for using SpaceStat in the analysis of spatial data*. NCGIA.

Anselin, L. (1995). "Local Indicators of Spatial Association -LISA", Geographical Analysis, Vol. 27, No. 2.

Bithell, J. (1990). "An application of density estimation to geographical epidemiology", Statistics in Medicine, Vol. 9, 691-701.

Bailey, T. C. (1994). "A review of statistical spatial analysis in geographical information systems", *Spatial Analysis and GIS*, Fotheringham, S. and Rogerson, P., (Eds), Taylor & Francis Ltd.

Besag, J. and J. Nevell (1991) "The detection of clusters in rare disease". Journal of the Royal Statistical Society, Series A, 154, pp. 143-155.

Cislaghi, C. et al (1996). " Methodological aspects of ecological study on the association between two biological indicators", Working paper.

Cliff, A.D. and J. K. Ord (1973). *Spatial autocorrelation*, Pion Limited, London.

Cliff, A.D. and J. K. Ord (1981). *Spatial Processes: Models and Applications*, Pion Limited, London.

Clayton, D. and J. Kaldor (1987). "Empirical Bayes Estimates of Age-standardised Relative Risks for Use in Disease Mapping.", Biometrics 43, 671-681

ESRI (1994). *ArcDoc 7.0* (on-line help), ESRI.

Everitt, B. (1974). *Cluster Analysis*, Heinenmann

Getis, A. and K, Ord (1992). "The analysis of Spatial Association by Use of Distance Statistics." Geographical Analysis, 24, 189-206

Getis, A. and K, Ord (1995). "Local Spatial Autocorrelation Statistics: Distribution Issues and an Application", Geographical Analysis, 27, No.4.

Haining, R. P. (1990). *Spatial data analysis in the social and environmental sciences*, Cambridge University Press.

Haining, R. P. (1994a). "Designing spatial data analysis modules for geographical systems", *Spatial Analysis and GIS*, Fotheringham, S. and Rogerson, P., (Eds), Taylor & Francis Ltd.

Haining, R. P. (1994b). "Diagnostics for regression modelling in spatial econometric", Journal of Regional Science, Vol. 34, No. 3, pp 325-341.

Haining, R. P., J. Ma and S. M. Wise (1996). "The design of a software system for interactive spatial statistical analysis linked to a GIS", Computational Statistics, 11, 449-466.

Haining, R. P., S. M. Wise and J. Ma (1998). "Exploratory Spatial Data Analysis in a Geographic Information Systems Environment", *The Statistician*, 47 (3), pp. 457-469.

McCullagh, P. and Nelder, J.A. (1989). *Generalised Linear Models*, Second Edition, Chapman & Hall, London.

NAG (1995). *NAG FORTRAN Library Routine Document, Mark 16*, NAG.

Siegel, S. (1956). *Nonparameteric Statistics for the behavioural sciences*, McGRAW-HILL, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Weisberg, S. (1982). *Applied Linear Regression*, Second Edition, John Wiley & Sons, New York.

Wise, S. M., R. P. Haining and J. Ma (1997). "Regionalisation tools for the exploratory spatial analysis of health data". In M. Fischer and A. Getis (Eds) *Recent Developments in Spatial Analysis: Spatial statistics, behavioural modeling and neuro-computing*, Springer-Verlag, Berlin, pp. 83-100.

Appendices

Appendix A: Selecting Partial Cases

SAGE provides a tool ***Select Cases*** which allows the user to interactively select cases. This tool is embedded inside those tools which are able to perform operations on partial cases. As mentioned in the previous sections, each tool has a toggle button, **Partial**, in the dialogue box associated with it. With **Partial** checked, the user is led to select cases by a series of dialogue boxes.

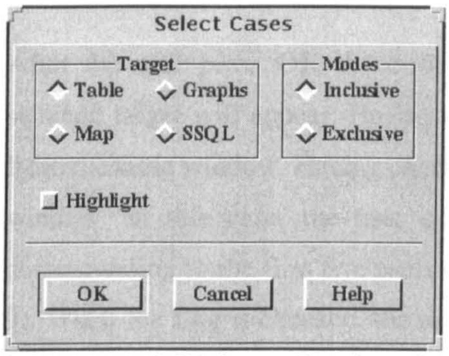


Diagram A. 1.Dialogue Box for selecting cases.

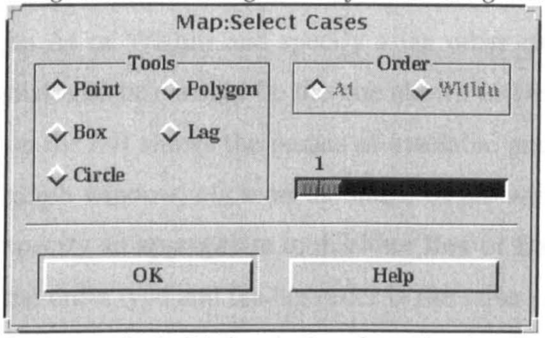


Diagram A. 2. Dialogue Box for selecting cases from the map.

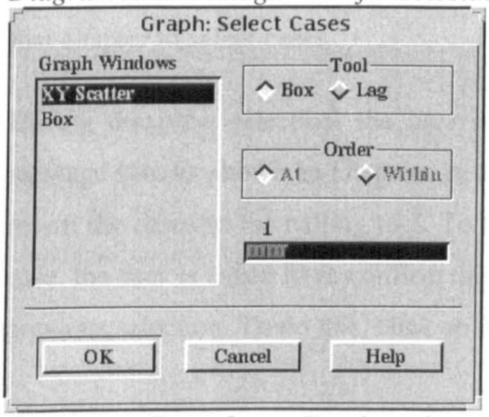
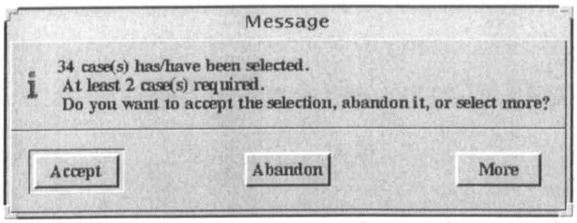


Diagram A. 3. Dialogue Box for selecting cases from a plot.

The first dialogue box is Diagram A.1. The user needs to check the appropriate toggle button in **Target** to specify a target window from which cases are to be selected. To select cases from visual items, the user should check one of **Table**, **Map** or **Graphs**. By checking **SSQL** , the user will be provided with the *SSQL* tool.



34 case(s) has/have been selected.
At least 2 case(s) required.
Do you want to accept the selection, abandon it, or select more?

Diagram A. 4. Dialogue Box for confirming the selection of cases.

The user should specify whether the cases will form the final selection (choose **Inclusive**) or whether all other cases will form the final selection (choose **Exclusive**). With **Highlight** set, the cases in the final selection will be highlighted in every window.

After the user press **OK** to confirm the setting, a dialogue box associated with the selected target will appear. Having checked **Table**, the user is prompted to select rows from the table window. Having checked **Map**, the user is led to select cases from the map window. In this case, the user should select an appropriate tool of the five tools corresponding to the first five tools discussed in Chapter 9 in Diagram A-2 (see Chapter 9). When the **Lag** is checked, the user should specify whether all cases, at that lag order or up to a certain lag order inclusively, adjacent to a cases are to be included by checking on **At** or **Within** and specify a lag order using the scalar. With **Graphs** checked[17], the dialogue box would be the one shown in Diagram A.3. The scroll list **Graph Windows** on the left shows the names of available graph windows. In order to select cases from a graph window, click on the name of the window in the scroll list. The user should also specify an appropriate tool, either **Box** or **Lag**. In the latter case, the user should specify the order type and the lag order in the same as discussed above. If the user sets **SSQL**, the user actually invokes the tool *SSQL* (see Chapter 9). Follow the procedures discussed in that chapter to select cases.

Having done the selection, the user is asked to accept, abandon or select more by a message box as shown in Diagram A.4. Click on **Accept** to accept the selection and to return the cases to the calling tool. To abandon the selection, click on **Abandon**. In this case, the user is asked to re-confirm this. The user is allowed to add more cases over the previous selection. To do this, click on **More** and select cases from the same target. If the

[17] Histograms have not been linked with this tools.

number of cases does not meet the requirements of a calling tool, the user is only allowed to abandon the selection or select more.

Appendix B: Defining Spatial and Logical Rules

A spatial rule could be specified as any of those listed below.

- ONE * or ONE x, y, where (x, y) refers to a point on the map;
- BOX * or BOX x1, y1, x2, y2, where (x1, y1) and (x2, y2) refer to two diagonal corners of a square area on the map.
- CIRCLE * or CIRCLE x, y, r, where (x,y) and r refer to the centre and radius of a circle respectively on the map.
- POLYGON * or POLYGON x1, y1, ...,xk, yk, where (x1,y1) to (xk, yk) are a set of points on the map which define the vertices of a polygon.

"*" indicates that a rule is need to be specified interactively when it is processed.

A logical rule is an arithmetic or logical expression on a set of operands. An operand could be the name of an attribute, constant or even an expression based on them. The arithmetic operators permitted are +, -, *, / and **, while the logical operators are shown in Table B below. Note that a space is required between each operator and operand.

Operators	Description	Example
>, <, =, ^=, <= or >=	the left-side operand greater than, less than, equal to, not equal to, not greater than or not less than the right-side.	attr1 < 9, where attr1 is numeric attribute.
CN or NC	the left-side string includes or does not include a character specified by the right-side.	attr2 CN 'ct', where att2 is in character form.
OR or AND	either (OR) or both (AND) the left-side and right-side operands must be true.	attr1 ^= 9 OR attr2 NC 'ct',

Table B. 1. Logical operators.

Appendix C. How does SAGE handle extreme "values"?

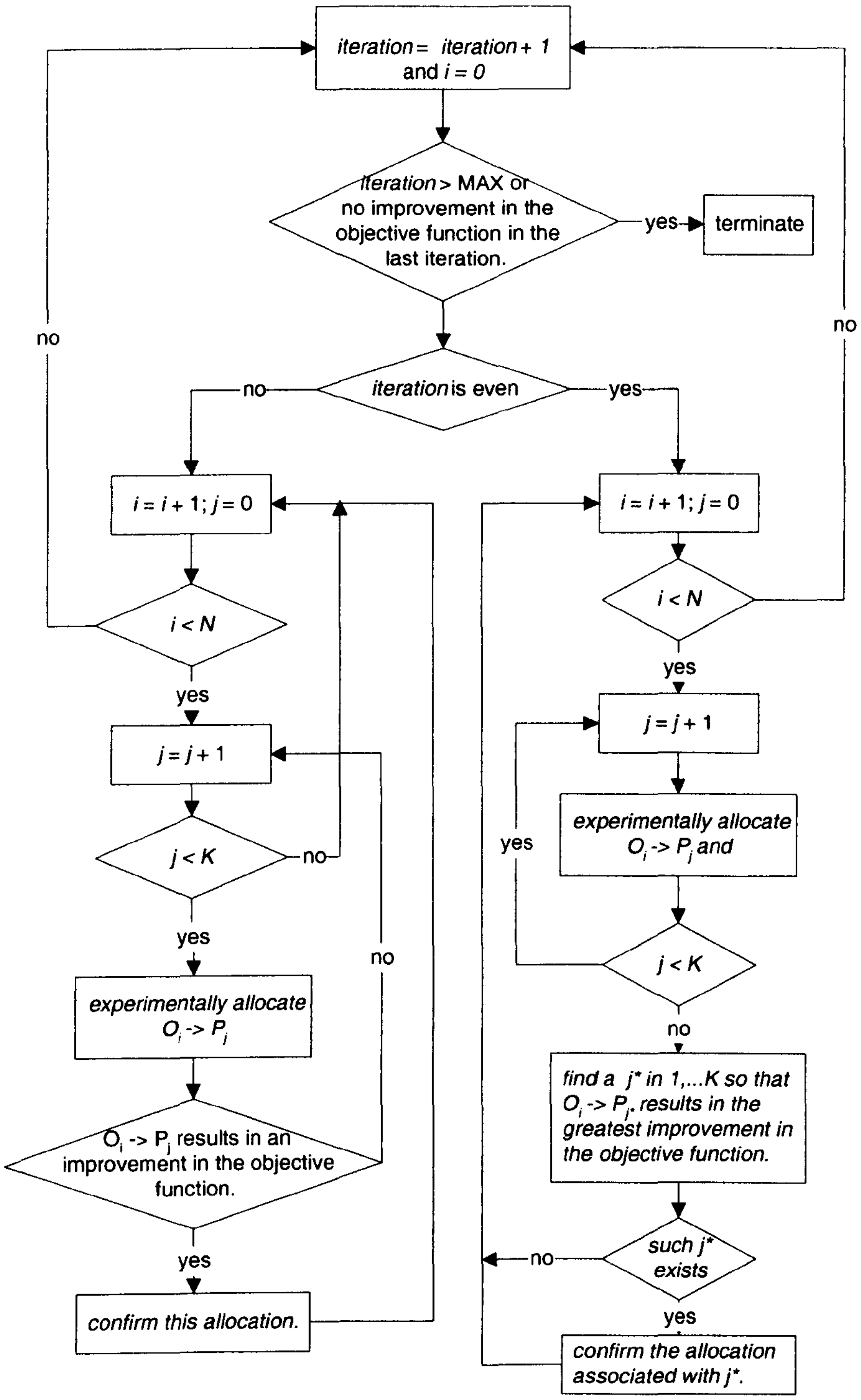
SAGE (version 0.1) relies on ARC/INFO to manage area-based data and it needs to transfer attribute data between the SAGE client and the SAGE server through RPC. The data of an attribute under transfer are in the form of ASCII strings with a format which is the same as the external format of the attribute to be or being defined in an INFO table (ESRI 1994). Therefore, SAGE has to handle those data which are beyond their external formats. Those data are referred as extreme "values". For example, if a format for a floating attribute requires 4 digits before and 3 digits after the decimal point and a value of the attribute is 12345. 678, then this value is an extreme value because it requires 5 digits be for the decimal point. Note that a value which only loses the precision during a transfer is not regarded as an extreme value.

In SAGE, the SAGE client is in charge of dealing with extreme "values" during data transfer processes. When the SAGE client sends a request of retrieving an attribute to the SAGE server, ARC/INFO calls INFO commands, OUTPUT and PRINT, to extract the attribute out of an INFO table. Then the server returns them to the client. If there are extreme "values", they will be indicated by a special string ******. The client interprets this special string as "a missing value" and replaces it with 0.0.

Before the client calls the server (i.e. INFO command ADD) to add a newly created attribute with an external format into a temporary INFO table, TMPINFO (see chapter 2), it searches through the data for any one that can not be fitted into the format defined by the user or the system. If the attribute is floating and either its largest value or its smallest value can not be fitted in the format, the client tries to modify the format for it by increasing the number of digits before the decimal point. If total number of digits is not greater than 24, this format is applied to the attribute, otherwise the client scales down the data so that it can be fitted into a 24 digit format. SAGE reports the number of digits being scaled down in text output window. If the attribute is integer and either its largest value or its smallest value can be fitted in a 10 digit format, the client applies the format to the attribute, otherwise it converts all values of the attribute to floating and treats them in the same way described above.

Appendix D: An illustration of the algorithm used in the heuristic classification

Let denote O_i the ith object and $\bigcup_{j=1}^{K} P_j$, $P_j = \left\{O_{j_l}, l = 1,2,..\left|P_j\right|\right\}$, $\left|P_j\right|$ = the number of objects in P_j. Start with *iteration* = 0, the algorithm is illustrated in the following diagram.



iteration = *iteration* + *1* and *i = 0*
iteration > MAX or no improvement in the objective function in the last iteration.
yes
terminate
no
no
iteration is even
no
yes
$i = i + 1; j = 0$
$i < N$
yes
$j = j + 1$
$j < K$
no
yes
no
experimentally allocate O_i -> P_j
O_i -> P_j results in an improvement in the objective function.
yes
confirm this allocation.
$i = i + 1; j = 0$
$i < N$
yes
$j = j + 1$
experimentally allocate O_i -> P_j *and*
yes
$j < K$
no
find a j^* *in 1,...K so that* O_i -> P_{j^*} *results in the greatest improvement in the objective function.*
such j^* *exists*
no
yes
confirm the allocation associated with j^*.

Copies of six papers

1. Haining R. P. J. Ma and S. M. Wise (1996) "The design of a software system for interactive spatial statistical analysis linked to a GIS". *Computational Statistics*, 11, pp. 449-466.

2. Wise S.M., J. Ma and R. P. Haining (1996) "SAGE - a system for the interactive analysis of area-based health data". *Proceedings of 11th ESRI European User Conference,* ESRI(UK) http://www.esri.com/base/common/userconf/europroc96/PAPERS/PN51/PN51F.HTM.

3. Wise S. M., R. P. Haining and J. Ma. (1997) "Regionalisation tools for the exploratory spatial analysis of health data". In M.Fischer and A.Getis (Eds.) *Recent Developments in Spatial Analysis - Spatial Statistics, Behavioral Modelling and Neurocomputing, Springer*, Berlin, pp. 83-100.

4. Haining R.P., S. M. Wise and J. Ma (1998) "Exploratory data analysis in a geographic information systems environment". *The Statistician*, 47 (3), pp. 457-469.

5. Haining, R. P., S. M. Wise and J. Ma (2000) "Designing and implementing software for spatial statistical analysis in a GIS environment". *Journal of Geographical Systems*, Vol. 2 (3), pp. 257-286.

6. Wise, S., R. Haining and J. Ma (2001) "Providing Spatial Statistical Data Analysis functionality for the GIS user: the SAGE project". *International Journal of Geographical Information Systems*, Vol. 15 (3), pp.239-254.

Computational Statistics (1996) 11:449-466

Design of a Software System for Interactive Spatial Statistical Analysis Linked to a GIS

Robert Haining, Jingsheng Ma and Stephen Wise

Department of Geography, The University of Sheffield, Sheffield
S10 2TN, England

Summary

We review the forms of spatial statistical analysis (SSA), discuss the benefits of developing software systems to implement SSA through linkage to a Geographic Information System (GIS) and alternative ways of creating the link and then describe the approach being followed by the authors which will add visual, interactive SSA capability to the ARC/INFO GIS using a client-server architecture.

Keywords: Client-Server, Linked Windows, Spatial Analysis, Visualisation, Geographic Information Systems

1 Introduction

This paper describes the first stages of a project to write a software package that will add statistical analysis capability to the ARC/INFO GIS. If geographical data sets are to be analysed rigorously, a range of techniques are needed that require numerical, graphical and cartographical capabilities. The challenge is to provide these in an environment where the user can interact quickly with the data and where these capabilities operate in different but linked windows. Part of the challenge is to draw wherever possible on existing software. However as the next section shows, spatial statistical analysis (SSA) comprises numerous specialist techniques that, whilst well established in the academic literature, are not routinely available in most statistical packages.

0943-4062/96/4/449-466 $2.50 © 1996 Physica-Verlag, Heidelberg

A prototype version called SAGE (Spatial Analysis in a GIS Environment) has now been developed using a client-server architecture. ARC/INFO acts as a server, while a program, which consists of a number of visual and non-visual SSA functions complementing those in ARC/INFO, acts as a client. Functions in the client requiring graphs, spreadsheets (etc) cannot be implemented effectively inside ARC/INFO due to functional and performance limitations. Instead they have been implemented through the X window system and other software packages. The linkage between the client and server has been implemented through RPCs (Remote Procedure Calls) to allow them to co-operate with each other seamlessly and efficiently so that ARC/INFO SSA capability can be extended as necessary. The difference of this approach from other earlier linking approaches will be discussed in a later section.

2 Spatial Statistical Analysis

Within the Geographic Information Systems (GIS) literature the term spatial analysis embraces three types of analysis: map based analysis (sometimes called cartographic modelling) in which layers of spatial information are manipulated to produce new information (Tomlin 1990), modelling, such as network analysis or hydrological modelling, where the systems are spatially distributed (Goodchild et al. 1995) and spatial statistical analysis (Wise and Haining 1991). It is in the third category, spatial statistical analysis (SSA), that GIS are currently weak. SSA is an area of research, independent of the field of GIS, which comprises an underpinning theory of analysis and a set of associated statistical techniques for analysing events where the form of analysis and the interpretation of results depends on the arrangement of the events in some defined space. An event in the present context consists of an object (a point, line or area) located in the defined space to which are attached a set of attribute values. In fact because of the linkage with GIS we will be concerned only with "geographical" events where the term geographical should be taken merely as defining (in a rather general way) the nature of the space within which analysis takes place (such as a city, a region, a national area or set of nations) and the types of attributes that are recorded (particularly attributes of human populations and environmental characteristics). Many of the methods of SSA could be applied in any of a number of fields of research where data values are attached to locations and many of the texts in this field draw attention to the breadth of application of these methods both in terms of spatial scale and range of attributes (see for example Cressie 1991; Haining 1990).

SSA comprises four main types of analysis all of which may find application at geographical scales of analysis (Cressie 1991). These four types are distinguished by the nature of the stochastic model that is responsible for the data (the located objects plus their associated attributes). The attribute values may be continuous or discrete, aggregates or individuals whilst the geographical pattern of the locations may be regular or irregular from a continuous or discrete set of locations. The four types are (Cressie 1991, p 8-9):

i Geostatistical analysis where attribute values (such as soil pH values) are recorded at (sampled) locations on a continuous surface;

ii Lattice data analysis where attribute values (such as disease mortality counts) are recorded for fixed locations (points or areas that may be regular or irregular);

iii Point pattern analysis where the location of events is the outcome of a random process (and a distinction is drawn between processes where all locations on a continuous surface are available as in the case of seed dispersal and processes where only specific locations are available for the event, such as diseased trees in an orchard);

iv Object analysis where the location of events is the outcome of a point process and attributes are random sets. Vegetation patches are one example of such events.

In this paper we are concerned only with lattice data analyses since the software system under development is for area based health data where the (irregular) areas refer to regions that represent population aggregates and attached to each of these areas is a set of attributes referring to population and health characteristics.

The aims of analysis are usually to find summary descriptions of the data or to build a statistical model that represents the static (cross sectional) features of what might be a dynamic process. The problem for SSA is to develop statistical procedures that allow the analyst to detect spatial properties and patterns in the data, perhaps through exploratory methods, and to fit models that recognize that each attribute value may, in part at least, be a consequence not only of the absolute position of each case within the space and the characteristics of the space at that position but also the relative position of each case with respect to some or all of the other cases. One of the implications of these remarks is that "classical" statistical analysis techniques, where each case is assumed to be independent and obtained by a process of random sampling, do not provide appropriate techniques for detecting spatial pattern or models for representing spatial variation. This class of problems shares some common ground with that arising in the analysis of temporal data (see for example Haining 1990).

It is easy to demonstrate that spatial statistical analysis raises new questions for both descriptive analysis and modelling. Classical univariate data summaries include measures of central tendency and spread and criteria for identifying extreme values (outliers). When data values are associated with locations, interest may focus on making descriptive comparisons between different areas so that the facility to subdivide the data set using location criteria is often important. Where the full data set is being analysed, interest may focus not on the overall mean (which may actually be of little interest) but on whether there are any spatial *trends* in the values. Also of interest may be the existence of any spatial pattern (or *spatial autocorrelation*) in the variation of data values about that trend. Finally an extreme value may not be an outlier in a distribution sense but may be a *spatial outlier* when its value is compared with values at adjacent locations. A software system to support a programme of exploratory data analysis and data description that identifies both global and local properties of the data needs not only numerical capability but also the capability to visualise both raw data and data analysis outcomes. Whilst graphs fulfil this role in classical statistics, maps are needed as well in the case of exploratory spatial data analysis in order to display data values and analysis outcomes in their geographical context. Moreover it may be helpful to be able to move rapidly between numerical summaries, graphs and maps in order to reveal data patterns (Haslett et al. 1990).

Spatial modelling also raises issues not encountered in classical statistical modelling. Consider the regression model as an illustration. The classical regression model expresses a dependent variable (Y) as a linear function (with parameters $\beta_0, \beta_1, \beta_2, ...$) of a set of independent variables ($X(1), X(2), \ldots, X(k)$). The level of Y in any case (i) is a function of the levels of $X(1), X(2), ..., X(k)$ in the same case. Using lower case to denote measured values the classical regression model is:

$$y(i) = \beta_0 + \beta_1 \cdot x(1,i) + \beta_2 \cdot x(2,i) + \cdots + \beta_k \cdot x(k,i) + \epsilon(i) \quad i = i, ..., n$$

where ϵ, an independent and identically distributed normal random variable that measures the variation in Y not explained by the linear influence of the set of X variables, is often referred to as the model error. Suppose cases are regions and Y is the number of asthma sufferers. The level of Y (in any region i) may be a function of the levels of an X factor (air quality) in other regions because of population mobility across the boundaries of the regions exposing them to air pollution levels in other areas. In the case where Y refers to numbers of sufferers from an infectious disease then the level of Y (in any region i) may be a function of the levels of Y in other regions as a result of population mixing and disease transmission associated with this mixing. The specification of the regression model needs to reflect the fact that in terms of the underlying processes, case boundaries are both modifiable (regions are artificial constructs) and permeable. Specifications of the sort described above can only use those classical methods for parameter estimation, hypothesis testing and model assessment that are routinely found in classical statistical packages by expanding the set of independent variables and adopting complicated parametric forms. A better approach is to develop alternative, more direct specifications (see Ord 1975, Martin 1992, Haining 1994a).

Even if the classical regression model (in the sense defined above) is appropriate, some of the underlying statistical assumptions require checking in special ways. Least squares fitting assumes that errors are independent. In the case of a regression model fitted to regions, if the model has been misspecified and an important variable omitted from the set of X factors then the errors may not be independent but rather spatially autocorrelated particularly if there is spatial pattern or continuity in the omitted X factor at a scale larger than the areal units. If material deprivation is important in explaining levels of an acute disease but is omitted from the specification then given that similar levels of deprivation are found in adjacent areas of a city then errors are likely to be spatially autocorrelated. A test for residual spatial autocorrelation is required just as time series modelling requires tests (such as the Durbin Watson test) for temporal autocorrelation (Cliff and Ord 1981). In the absence of a clear explanation for pattern in the regression residuals, or where it is suspected that there are many factors responsible, a solution may be to fit a regression model where a specific autocorrelation model is used to summarize the non independent error in order to ensure that parameter estimates and hypothesis tests are valid (Anselin and Griffith 1988). There are other diagnostic checks that are influenced by the presence of spatial effects and which call for specially defined measures (see for example Haining 1994).

There are other analysis capabilities that are needed to underpin the SSA of area data.

- Since regions are artificial constructs it is important to construct areal frameworks that are relevant to the problem being studied. For example if the analyst wishes to explore the link between levels of deprivation and measures of ill health using regression, then an areal framework in which each spatial unit is broadly homogeneous in terms of deprivation provides a sounder basis on which to explore this question than one in which the level of deprivation in each spatial unit is an average of a very diverse set of levels. This requires that the software can build a set of k regions from a set of m $(m > k)$ smaller spatial units according to a set of defined criteria (Haining et al. 1994).

- Where rates are required (such as incidence or mortality rates for example) these should be robust and comparable across spatial units where base population levels may be small in some cases and highly variable between spatial units.

- For any given set of spatial units there is no "unique" or "right" way to specify spatial relationships between the units. The issue does not arise in classical statistical analysis and even in the case of time series analysis the relationships between events in time are only one not two dimensional with, in many cases a strong directionality (the past affects the future and the reverse is not possible). In the case of SSA there is a particular need to explore the effects of different definitions of inter-area relationships on analysis findings (Haining 1990, p69-74).

3 Why link GIS and SSA?

There are two reasons for linking SSA to GIS:

1. to facilitate SSA;
2. to enable GIS to realise its potential as a general purpose tool for handling spatial data since currently there is a lack of statistical analysis capability in most GIS.

SSA currently entails high start up costs in terms of inputing the spatial configuration of the sample data and of writing the software to perform the analysis, since there are not many widely available packages which offer SSA capabilities. It is also important to be able to manipulate data in many routine but arduous ways as described in the previous section and also visualise data in different ways combining maps, graphs and numerical measures in order to explore the spatial data (or defined subsets of the spatial data). Visualisation needs to be done interactively, sometimes simultaneously (in the case of more than one subset of data) and, of course, using appropriate methods. All this is time consuming to do by hand and in many cases impossible to do within standard statistical software either because of limitations in the software or because of data complexity. If GIS becomes a standard medium for holding spatial data it seems natural to interface analysis routines directly to the GIS both to facilitate analysis and take advantage of the visualisation and data manipulative capabilities of GIS that are important for SSA.

In order to carry out a program of SSA, the user needs to access software that can store and manipulate spatially referenced information, can execute appropriate statistical analysis and finally allow good interactive visualisation of the raw data or analysis outcomes. In each case there already exist packages which provide these facilities - in the order listed above GIS, statistical packages (although the large commercial statistical packages are not designed to facilitate SSA) and visualisation packages. However, no current package provides all three. At issue is whether it is best to attempt to write a special package for SSA, or link together existing software. In section five we describe an approach that is based on linking together existing software where it is efficient to do so rather than re-developing software. However in the next section we review previous work in this area from the point of view of the type of linkage used. For another review but from the perspective of different forms of SSA see Bailey (1994).

4 Approaches to producing SSA software

Previous work on developing software for SSA has used one of three approaches: purpose written software; loose coupling of existing packages (for example the various packages simply exchange files); close coupling of existing packages (one package calls routines from another).

4.1 Purpose written software

An early example of a purpose written package was INFOMAP which provided the features of a simple thematic mapping package but with the ability to perform simple statistical analyses of the data (Bailey 1990). The features provided included on-screen statistical summaries, correlograms and gravity modelling. Probably the best known example of a purpose written package is REGARD (formerly called SPIDER) produced by Haslett and co-workers (Haslett et al. 1990). This is written for the Apple Macintosh, and takes advantage of the excellent graphical and windowing capabilities which are built into the Macintosh operating system. The key feature of REGARD is the ability to produce multiple linked views of the same data – thus with a map in one window it is possible to open a second window containing a histogram of one of the variables associated with the map, or a scatterplot showing the relationship between two of the variables. What adds extra power to this approach is that the windows are linked – highlighting an area on the map for example will cause the data points relating to those areas on the scatterplot to be highlighted. Outliers on the scatterplot can be identified and selected and the corresponding map points will be highlighted allowing the user to determine if the outliers are spatially clustered. A similar system called Polygon Explorer, based on polygon rather than point data, has been described by MacDougall (1992). By contrast with the systems that have focussed on the visualisation side of SSA, Anselin (1990) has developed a system, SpaceStat, which provides comprehensive facilities for the statistical analysis of spatial data and in particular has the ability to fit different types of spatial regression model.

Openshaw and colleagues (Openshaw et al. 1990) provide another example of work in this area, based on a different approach to the problem of spatial data analysis. They argue that GIS has created the potential for large databases, bringing together basic demographic statistics with information on crime, health and consumer behaviour, for example, to form a single GIS database. In the face of such a richness of data, existing techniques of analysis, they argue, are inadequate. They propose that what is needed is a suite of automatic 'pattern spotters' and detectors which can browse through GIS databases looking for interesting patterns and relationships. The earliest of these was the Geographical Analysis Machine (GAM) (Openshaw et al. 1987) which detected clusters of leukaemia cases by repeating a standard cluster test at a very fine spatial resolution over the whole of the area of interest, and detecting regions which gave multiple positive results. This general idea has since been extended to looking for patterns in the temporal and attribute dimensions (Openshaw et al. 1990) and using animation to illustrate the progress of the search for meaningful patterns in the database (Openshaw and Perre 1995). It is worth commenting that like others working in this field, we regard this as an interesting approach to data exploration and hypothesis building but reject Openshaw's argument that the data-rich world of GIS renders more traditional approaches obsolete (Bailey 1994, Gattrell and Rowlingson 1994).

The main advantage of writing software from scratch is the complete control over what the software will do. Where the techniques being developed are very new, as with the work of Openshaw, there is possibly no real alternative. It is noticeable that many of the systems developed so far this way have concentrated on one aspect of spatial data analysis to the exclusion of others. However, one of the problems of this approach is the effort required to write a complete package from scratch. Recent work by Brunsdon and Charlton (1995) and Dykes (1995) has reduced this problem by the use of tools designed for the development of graphical user interfaces. Dykes (1995) describes a system based upon the Tcl and Tk toolkit of Ousterhout (1994), which provides a basic 'canvas' widget which can draw graphical primitives such as lines, circles and text in a graphics window. In the Cartographic Data Visualiser (cdv) system (Dykes 1995) this is used to draw choropleth maps of the data, producing a system with similar functionality to REGARD or Polygon explorer, but with only a small amount of new code needed. Using the Tcl programming environment it is also possible to link this map window with other graphical windows. Although this approach takes much of the drudgery out of writing software from scratch, it is still necessary to modify the existing tools to produce cartographic output, an area where GIS is already very strong.

The other major disadvantage with this approach is that the spatial data must be imported from elsewhere, usually a GIS. This leads to a number of problems: it creates a copy of the database which is likely to get out of step with the data held in the original GIS database; it is inconvenient to have to move data between systems in this way; it limits the type of analysis which can be undertaken, since it is no longer possible to modify the spatial data. In the case of point data this may be less important, but with area-based data the construction of new areal bases from the original one can be an important analytical tool (Haining et al. 1994).

Rather than start entirely from scratch, many workers have opted to use existing packages which provide much of what they need and augment it as necessary. The question then becomes one of which strategy to use to add on the extra facilities needed. In the Goodchild et al. (1992) classification there are two approaches, termed loose and close coupling.

4.2 Loose coupling

Here several packages may be used, and the linkage is made simply by passing data files between them. The logic is that since in many cases software already exists to do much of what is needed it makes sense to use this rather than rewrite it. A good example of this approach is the work of Kehris (1990,1990a), who used the ARC/INFO GIS and the statistical package GLIM. A similar approach was taken by Anselin et al. (1993) who also used ARC/INFO but used SpaceStat, already referred to above, for the statistical calculations. In this case the two packages were not even running on the same machine - ARC/INFO was a workstation version running on a Sun while SpaceStat was a PC package.

The main advantage of this approach is clearly the flexibility of being able to link together a wide range of packages. However, this is a very cumbersome way of actually doing any real spatial analysis - not only is it necessary to learn more than one package, as Anselin et al. (1993) point out it will often be necessary to switch between viewing the maps in the GIS and doing calculations in the statistical packages. The other problem is that the visualisation tools available are limited to those within the GIS, so that it will not necessarily be possible to view a map and related graph at the same time, let alone have the two dynamically linked in the way that is possible in some of the earlier examples.

One solution is to base the system around one package, and modify this to add in the extra features required, if necessary by calling other packages or user-written routines, a technique called close coupling.

4.3 Close coupling

One of the earliest examples of this was the work of Ding and Fotheringham (1992) who like Kehris (1990) were trying to calculate spatial autocorrelation statistics for spatial data held in ARC/INFO. However, rather than pass data out to GLIM, they used the programming language within ARC/INFO (called ARC Macro Language or AML) to do the necessary calculations, which could then be saved back into the GIS database for use in later analysis and viewing. A more sophisticated example is the work of Batty and Yichun (1994) who constructed a whole software suite for running urban population density models based on ARC/INFO. Again AML was used to do some of the work, but certain routines were written in FORTRAN which could also be called from within AML which means that the user is still presented with a single interface. They were also able to link together a map and associated graphical display of population density values at different distances from the city centre. However this is a much more limited capability than is possible with a system such as REGARD - the different views are both in the same graphical window and depend upon programming ARC/INFO to draw the graphical display.

A number of workers have explored the possibility of using existing statistical software as the starting point rather than a GIS and here there are essentially two modifications needed: statistical modifications which involve adding spatial statistical techniques (as described above); graphical modifications which add the ability to draw maps of the data or the results. Brunsdon and Charlton (1995) use XLisp-Stat an object-oriented programming language which already includes facilities for performing basic statistical calculations (Tierney 1991). As with the work of Dykes (1995) the advantage is that it is possible to modify existing window-drawing widgets to produce ones which can draw maps. It is also possible to use the Lisp programming language to calculate spatial statistics. Gatrell and Rowlingson (1994) describe the use of the S-Plus system as the basis for their work. S-Plus is a programming language which provides a wide range of standard statistical calculations, plus powerful facilities for writing new code, and the authors showed how it was possible to produce a system for analysing point patterns, including the calculation of kernel estimates of density and special (K) functions. S-Plus includes facilities for graphical output as both maps and charts, although it does not appear to support linked windows.

The main advantage of this second approach is in the use of existing statistical software. This means that certain standard techniques are immediately available, and in code which is numerically robust and statistically sound. In addition, such packages often provide a programming language capability so that it is possible to add extra facilities within the framework of the package. However many of the disadvantages quoted above for the purpose written software apply with equal force here, in particular the need to import the spatial data from elsewhere and the inability to modify that data once it is imported.

4.4 Limitations of these approaches

Each of the approaches described above has the capability to deliver some of the desired features of a spatial analytical system, but not all of them. Using free standing software makes it possible to provide excellent visualisation facilities, but means that all the data has to be imported into the system, and features such as mapping or statistical calculations have to be written from scratch. If existing packages are used to supply the latter features, then the flexibility of the visualisation tools is greatly reduced. What is needed is a system which can take advantage of the excellent facilities which already exist in GIS and statistical packages, but provide these in an environment which allows for dynamic visualisation of the results and within an architecture of "seamless" integration.

5 A fully integrated architecture based on the client-server model

The choice for the form of software integration depends critically on the facilities of the chosen GIS (in the present case ARC/INFO) and the computing environment as well as the intended application area which defines the other types of packages to be linked to the GIS. In the present context there are two possible models for integration. The first is close coupling, in which ARC/INFO calls an external program to carry out a set of spatial operations that are not available within ARC/INFO itself and waits for the results to be returned. This is the approach which has been used by a number of other workers (Kehris 1990, Ding and Fotheringham 1992). The second is the client-server model (Smith and Guengerich 1994, Umar 1993) which has not previously been used in this context. Here, ARC/INFO functions as a server process and external programs which use other facilities of the computer environment function as client processes running independently of the server process. The clients request services from the server; the server processes the requests and returns the results to the clients. This second model is the route that is explored here since the first cannot be adopted when a required external program must be active along with an ARC/INFO session.

The client-server model is a form of loose coupling, since it allows existing software to be linked together to produce a new application. In the current case, since ARC/INFO is weak at manipulating graphical and tabular data, these functions are performed by existing X windows packages, which are linked to ARC/INFO using the client-server architecture. As well as providing a more powerful, flexible and rapid means of linkage than a simple exchange of data files, client-server processes can be distributed around the computer network (Smith and Guengerich 1994, SunSoft 1994). So, the external program can perform operations on the data residing at remote computers through the services of ARC/INFO that may itself be running on another remote computer. Moreover, the client-server model provides a chance to specify a common interface, on which many client and/or server processes could work together, so that the system is extensible.

The objective is to link SSA tools with GIS to the mutual benefit of both domains of analysis. These tools take a number of generic forms. They are visually interactive and are presented in the form of maps, graphs, and spreadsheet-like tables. Although the spatial objects take different forms in different tools (such as areas in maps; points, lines or a group of points in graphs; rows in tables), each of the tools should allow users to identify spatial objects in the system, allow objects in different tools to be simultaneously linked and allow SSA to be carried out on objects as required.

Mapping tools are required for displaying spatial data and for interactive querying of spatial relationships among objects. Maps are the fundamental tools for the presentation of geographical data. Graphical tools are important for a range of exploratory statistical procedures and the examination of data properties through such devices as scatter plots, box plots and "autocorrelation detecting" plots for example (Cressie 1991). One and two dimensional graphics are often used but so are three dimensional graphics and matrix graphics (Tukey 1977, Cleveland 1988). For some types of plots such as scatter plots and box plots, the graphical tools should allow plots to be automatically overlaid in the same graphics window in order to facilitate the comparison of SSA results. Table-handling tools are required to display, edit, query and do statistics on object attributes and store results. In addition, there are at least two other kinds of tools that are important to support the functions of those already described : regionalization tools and connectivity tools. Regionalization tools are needed to construct alternative areal systems on which SSA can be conducted. In addition a number of SSA techniques require information on the connectivities between areas (usually in the form of the so-called **W** matrix used for calculating spatial autocorrelation statistics (Haining 1990)), and the connectivity tools are needed to automatically derive this information and to allow the user to edit the **W** matrix, to explore the effects of different connectivities on the analysis.

The data model of ARC/INFO, provides a good working environment for the system operations. The integration will extend the already powerful capabilities of ARC/INFO in the areas of cartography and database management as well as data capturing and editing and the fact that ARC/INFO already contains these capabilities makes a strong argument for using it as the starting point. Where ARC/INFO is weak is in not providing any effective interactive graphical package with sophisticated data-manipulating capabilities, nor is it possible to effectively implement such tools within ARC/INFO because of the weaknesses of its AML language and because of its very primitive drawing capabilities. Moreover, ARC/INFO does not have functions for implementing spreadsheet-like tables for tabular data manipulation. These unsupported capabilities must therefore be complemented by other facilities outside ARC/INFO. For the X-window and UNIX systems on which our version of ARC/INFO runs, there are some graphical and table packages (widget libraries) that meet our project requirements and which provide a real opportunity to build an integrated architecture using the client-server model.

The architecture for integrating SSA tools with ARC/INFO consists of three system components: ARC/INFO, a linking interface (LI) and a SSA module shown as rows in Figure 1. Each of the three system components is an independent process. The SSA module is an executable program written in C, which provides facilities for data exploration via a series of graphical user interfaces (GUIs) as well as carrying out a variety of spatial and aspatial statistical operations. The LI, also written in C, is designed as a connecting program to allow ARC/INFO and the SSA module to communicate with each other.

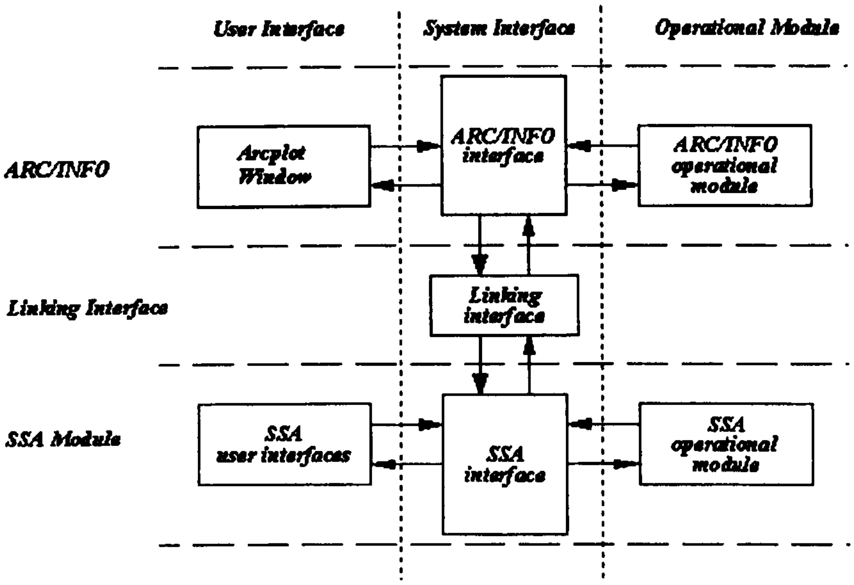


Figure 1: Architecture for integrating ARC/INFO and SSA
(Rows for system components; Columns for functional components)

In terms of functional components, the architecture consists of a user interface, a system interface and an operational module, shown in separate columns in Figure 1. The user interface consists of an 'Arcplot' window from ARC/INFO and the graphical and tabular windows of the SSA module. There are menus for launching operations peculiar to each of these interfaces. The system interface includes an ARC/INFO interface, (i.e. ARC/INFO command interface), a LI and a SSA interface (SSA management system). The main functions of the system interface are to manage the requests/responses between client and server, and the operations issued from user interfaces. The operational module contains the ARC/INFO operational module and the SSA operational module.

Under this architecture, a request issued by the SSA module to ARC/INFO is received, translated by the LI into a sequence of operations predefined in the form of AML files, and sent to ARC/INFO. Then ARC/INFO performs these operations and sends the results back to the LI which sends the appropriate responses back to the SSA module. Communication between the LI and both the SSA module and ARC/INFO, is currently handled using the standard Unix facility of named pipes and RPCs. ARC/INFO also provides an Inter-Application Communication mechanism (ESRI, 1994) which could be used instead. This gives the advantage of allowing communica-

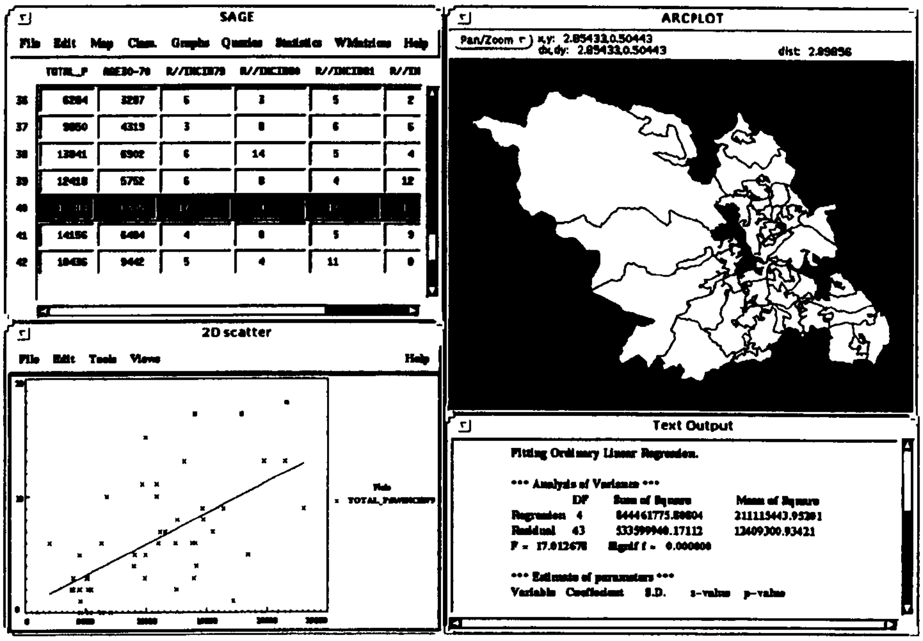
Category	Functions
Management	Create, open and edit a SAGE analysis view; load data; edit data table; create new coverages.
Display	Create map; graphic displays; highlight objects in windows.
Query	Select objects; query objects according to spatial and logical criteria.
Tabular	Perform arithmetic operations.
Classification	Classify objects by attribute scores; hierarchical and non hierarchical classification; use additional criteria based on contiguity to construct regions.
Obtain **W** matrices	Identify regional adjacencies according to various criteria.
Statistics	Descriptive and inferential statistical techniques; standard and spatial statistics; regression.

Table 1: Summary of SAGE functions

tion across the network, so that an SSA module running on one workstation could communicate with ARC/INFO running on another one. However the current version of IAC only has a small communications buffer making it unsuitable for transferring the results of complex requests between the LI and ARC/INFO.

The spatial datasets are managed by ARC/INFO in the form of polygon coverages, which contain all the locational and attribute data for the polygons. The SSA also keeps a copy of some of this information in order to speed up certain statistical analysis – for example, for each line forming part of the polygon boundaries, ARC/INFO maintains topological information including the identity of the polygons on each side of the line, which can be used for deriving the W matrix. When the user identifies attributes which are to be used in the analysis, these are retrieved from ARC/INFO and then held and manipulated within the SSA. SSA results such as residuals after fitting regression models are automatically stored in ARC/INFO database files so that they may be used by the ARC/INFO-based mapping tools. An object-matching mechanism is implemented in the SSA module to automatically link objects in different tools. Thus, selecting an area on a map, will cause the data point relating to that area on a scatterplot, and the relevant row in a tabular display of area attributes, to be highlighted too.

Table 1 lists the main categories of functions that have been implemented in SAGE together with a brief description of them. Figure 2 shows a typical session in which the user has opened the spreadsheet-like table, a map window, a graphics window and an analysis window. The windows are linked and rows of the table are highlighted both on the map and graphics windows. The prototype, as currently implemented, seems reasonably rapid in its re-



SAGE
File Edit Map Class. Graphs Queries Statistics WMatrices Help
2D scatter
File Edit Tools Views Help
TOTAL_P
ARCPLOT
Pan/Zoom x,y: 2.854330.50443
dx,dy: 2.854330.50443
dist: 2.89856
Text Output

	TOTAL_P	AGE30-70	R//INCID79	R//INCID80	R//INCID81	R//IN
36	6204	3207	6	3	5	2
37	9050	4313	3	0	6	6
38	13041	6302	6	14	5	4
39	12418	5752	6	0	4	12
40	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
41	14156	6404	4	0	5	9
42	18436	9442	5	4	11	0

```
Fitting Ordinary Linear Regression.

*** Analysis of Variance ***
             DF    Sum of Square        Mean of Square
Regression    4    844461775.80804      211115443.95201
Residual     43    533599940.17112      12409300.93421
F =  17.012678     Signif f =   0.000000

*** Estimate of parameters ***
Variable  Coefficient   S.D.    t-value   p-value
```

Figure 2: Analysis in SAGE

sponses when working on a data set with over 1000 areal units and running on a Sun Sparc 20 workstation.

6 Conclusions

In this paper we have described the design and implementation of a software system to undertake spatial statistical analysis linked to a GIS. Such software must be capable of providing a variety of techniques for both displaying and analysing spatial data including cartographic and graphical displays and the calculation of both aspatial and spatial statistics. Some researchers have taken the approach of writing standalone packages, which have the advantage of providing speed and the ability to link information in different windows on the screen, but fail to take advantage of the powerful facilities provided in existing GIS, statistics and visualization packages. Researchers who have tried to link existing packages, have used one of two approaches: loose coupling, where the link is via files passed between packages, or close coupling where one set of software is called from within the other. Loose coupling provides flexibility, but is often cumbersome, and does not provide the ability to link views of the data produced in separate packages. Close coupling provides a better interface, but still only provides limited facilities for linking windows.

The approach used in SAGE is to use the architecture of client-server computing to provide a form of loose coupling between packages. In contrast to other loose-coupling approaches, however, the system is controlled by a client program, which can keep track of the spatial objects being analysed, and hence link the representation of these objects in windows produced by separate packages. A further advantage of this approach is that the client-server approach offers the possibility of distributed working.

A problem with the current version of SAGE, which is in the prototype stage, is the limited facilities for client-server working provided in existing GIS software. However, as client-server computing becomes more widespread, this limitation should be removed.

The authors acknowledge receipt of ESRC research grant Number R000234470 "Developing spatial statistical software for the analysis of area based health data linked to a GIS".

7 References

Anselin, L. and Griffith, D.A. (1988) Do Spatial Effects Really Matter in Regression Analysis? Papers of the Regional Science Association, 65, 11-34.

Anselin, L. (1990) Space Stat: A Program for the Statistical Analysis of Spatial Data. Dept. Geography, University of California, Santa Barbara.

Anselin, L., Dodson, R.F. and Hodak, S. (1993) Linking GIS and Spatial Data Analysis in Practice. Geographical Systems 1 (1), 3-23.

Bailey, T.C. (1990) GIS and Simple Systems for Visual Interactive Spatial Analysis. The Cartographic Journal, 27, 79-84.

Bailey, T.C. (1994) A Review of Statistical Spatial Analysis in Geographical Information Systems. In Fotheringham S and Rogerson P. (ed) Spatial Analysis and GIS. Taylor and Francis, London. 11-44.

Batty, M. and Yichun, X. (1994) Urban Analysis in a GIS Environment : Population Density Modelling using ARC/INFO in Fotheringham, S. and Rogerson, P. (ed) Spatial Analysis and GIS, Taylor and Francis, London, 189-220.

Brunsdon C and Charlton M. (1995) A Spatial Analysis Development System using LISP. Proc. GISRUK '95, 155-160.

Cleveland, W.S. and McGill, M.E. (Eds.) (1988) Dynamic Graphics for Statistics, Wadsworth and Brooks/Cole, Pacific Grove, CA.

Cliff, A.D. and Ord, J.K. (1981) Spatial Processes. Pion, London.

Cressie, N.A.C. (1991) Statistics for Spatial Analysis, John Wiley and Sons, New York.

Ding, Y. and Fotheringham, S. (1992) The Integration of Spatial Analysis and GIS. Computers, Environment and Urban Systems, 16, 3-19.

Dykes J. (1995) Pushing Maps Past their Established Limits: a Unified Approach to Cartographic Visualization. Proc. GISRUK '95, 78-95.

ESRI (1994) ArcDoc version 7.0. (online help) ESRI, Redland, CA.

Gattrell, A.C. and Rowlingson, B. (1994) Spatial Point Process Modelling in a GIS Environment. In Fotheringham, S. and Rogerson, P. (ed) Spatial Analysis and GIS, Taylor and Francis, London, 147-164.

Goodchild, M.G., Haining, R.P. and Wise, S.M. (1992) Integrating Geographic Information Systems and Spatial Data Analysis: Problems and Possibilities. Int. Journal of Geographical Information Systems, 16, 407-24.

Goodchild, M.F., Parks, B.O. and Steyaert, L.T. (1993) (eds) Environmental Modelling with GIS. Oxford University Press, New York.

Haining, R.P. (1990) Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press.

Haining, R.P. (1994) Designing Spatial Data Analysis Modules for Geographical Information Systems, In: Fotheringham, S. and Rogerson, P. (eds) Spatial Analysis and GIS, Taylor and Francis, pp. 45-64.

Haining, R.P. (1994a) Diagnostics for Regression Modelling in Spatial Econometrics. Journal of Regional Science, 34, 3, 325-341.

Haining, R.P., Wise, S.M. and Blake, M. (1994) Constructing Regions for Small Area Analysis: Material Deprivation and Colorectal Cancer. Journal of Public Health Medicine, 16, 429-438.

Haslett. J, Wills G. and Unwin A.R. (1990) SPIDER — an Interactive Statistical Tool for the Analysis of Spatially Distributed Data. Int. J. Geographical Information Systems 4(3), 285-296.

Haslett, J., Bradley, R., Craig, P.S., Wills, G. and Unwin, A.R.
(1991) Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies. American Statistician, 45, 234-42.

Kehris E. (1990) A Geographical Modelling Environment Built Around ARC/INFO. North West Regional Research Lab Report 13.

Kehris E. (1990) Spatial Autocorrelation Statistics in ARC/INFO. North West Regional Research Lab Report 16.

MacDougall E.B. (1992) Exploratory Analysis, Dynamic Statistical Visualisation and Geographic Information Systems. Cartography and Geographic Information Systems 19(4), 237-246.

Martin, R.J. (1992) Leverage, Influence and Residuals in Regression Models when Observations are Correlated. Communications in Statistics : Theory and Methods, 21, 1183-1212.

Openshaw S., Charlton M., Wymer C. and Craft A. (1987) A mark 1 Geographical Analysis Machine for the Automated Analysis of Point Data Sets. Int.J.Geographical Information Systems 1,335-358.

Openshaw S., Cross A., Charlton M. (1990) Building a Prototype Geographical Correlates Exploration Machine. Int.J.Geographical Information Systems 4(3)297-311.

Openshaw S.and Perree T. (1995) User-centred Intelligent Spatial Analysis of Point Data. Proc. GISRUK '95, 153-154.

Ord, J.K. (1975) Estimating Methods for Models of Spatial Interaction. Journal of the American Statistical Association, 70, 120-6.

Ousterhout (1994) Tcl and the Tk Toolkit. Addison-Wesley, Reading, MA.

Smith, P. and Guengerich, S. (1994), Client-server Computing All-in-one reference for total systems development, second edition, Sams Publishing.

SunSoft (1994), Solaris 2.4 Network Interfaces Programmer's Guide. Sun Microsystems Inc.

Tierney, L. (1991) LISP-STAT: An object-oriented environment for Statistical Computing and Dynamic Graphics. Wiley; New York.

Tomlin, C.D. (1990) Geographic Information Systems and Cartographic Modelling. Prentice-Hall, Englewood Cliffs, J.J.

Tukey J. (1977) Exploratory Data Analysis. Addison-Wesley Publishing Co. Reading, MA.

Umar, A. (1993) Disturbed Computing and Client-Server Systems. Prentice Hall, New York.

Wise, S. and Haining, R. (1991) The Role of Spatial Analysis in Geographical Information Systems. Proceedings of the 3rd National Association for Geographic Information Conference, 3.24.1-3.24.8.,13

Steve Wise, Jingsheng Ma, and Bob Haining

SAGE - A System for the Interactive Analysis of Area-based Health Data linked to ARC/INFO

GIS packages, such as ARC/INFO, have very good facilities for many types of analysis, but are currently weak in the statistical analysis of spatial data and the use of scientific visualisation techniques. This paper describes the development of a system based around ARC/INFO, which provides a wide range of facilities for the analysis of area-based health data. The system, called SAGE (Spatial Analysis in a GIS Environment), consists of purpose written graphical and statistical software which calls ARC/INFO running as a server, to perform certain operations, such as the provision of the data to be analysed, cartographic display and some GIS operations such as polygon dissolve. The maps produced by ARC/INFO can also be linked to other graphs and data displays produced by SAGE. For example highlighting a polygon on the map not only causes the relevant row in the attribute display to be highlighted, but also the data points on any graphs drawn using the data. SAGE exploits the ability of ARC/INFO to use client/server computing to produce a flexible system, providing a range of analytical tools to analyse spatial data held in ARC/INFO - the facilities provided include a range of graphical display techniques, the ability to create purpose built regions from basic spatial units such as census areas and a range of statistical techniques ranging from simple summary statistics to the fitting of regression models specially modified to deal with spatial data. Although developed with health applications in mind, the system would probably be of use in the analysis of many other area-based datasets.

Introduction

The analysis of spatial data has always been one of the principal strengths of Geographic Information Systems. For instance the earliest operational GIS, the Canadian GIS, was developed to allow the analysis of large amounts of environmental data. The commonest type of analytical operation in GIS involves the manipulation of coverages to produce new information or further coverages (Burrough 1986). The simplest example is where a single coverage is modified in some way to produce new information - for example a DTM might be used to generate a map of slope steepness, which might in turn be processed to select those areas above a certain angle. More complex operations can be performed by combining coverages, as in the various types of overlay operation, which are commonly used in sieve mapping.

ARC/INFO is particularly rich in these kinds of cartographic modelling techniques, especially as it has the capability of performing vector or raster-based analyses. In addition, it contains functions for more specialised types of analysis, the two best examples being the hydrological modelling tools within GRID, and the network routing and location/allocation tools in NETWORK.

However the analytical capabilities of current GIS, including ARC/INFO, are still limited in two areas:

- Scientific Visualisation

This term describes the use of interactive graphical techniques for the analysis of data, and was originally coined in respect of the analysis of vast amounts of information created by simulation programmes in the physical sciences. Although many of the display techniques used in visualisation, such as statistical graphs, maps, 3D views and block diagrams, are standard graphical devices, what makes visualisation different from simply producing graphs of data is that these tools are provided in a highly interactive environment which allows the user to explore the data by using a wide range of different graphical techniques. Some of the techniques are by their very nature unsuitable for the production of hard copy graphics - for example the use of animation or the ability to rotate 3D views in real time. A particularly powerful feature of some visualisation packages is the ability to link different views of the same data, so that highlighting an extreme data point on a scatter plot will highlight the location of that sample on a map.

- Statistical techniques.

The statistical capabilities of many GIS packages are quite poor, even for producing standard summary statistics such as measures of central tendency and spread. ARC/INFO actually has slightly better facilities for raster data, since GRID can calculate some statistical measures, and calculate correlations between coverages. However, there are also a wide range of more sophisticated statistical techniques for analysing patterns in spatial data and relationships between variables, very few of which are available in standard GIS systems (Haining 1990, Bailey and Gatrell 1995).

These limitations of the current analytical toolkit of GIS have been most strongly noted by academic researchers (Goodchild et al 1992) and some have argued that such facilities are too specialised to be of interest to the majority of GIS users. While it may be true that for many, a GIS is essentially a data management tool, there are also important GIS application areas where the analysis of data is important. One such area is in crime pattern analysis (CPA), where the analysis of patterns in the vast amounts of information stored by the police on the occurrence of crime can often lead to useful insights in tracking down the perpetrators of crime, or in planning preventative measures, and some work has been done to develop CPA facilities linked to ARC/INFO (Cross and Openshaw 1991).

Another area where analysis is important is in the field of health which is the focus of this paper. There is a long history of analysing the spatial patterns of ill-health as a guide to causative factors, dating from Snow's pioneering work on cholera in London (Snow 1854) right up to present day studies of the evidence of the link between environmental pollution and some cancers suggested by the clustering of cases in certain areas (Bailey and Gatrell 1995). For this kind of analysis it is not enough to be able to plot a map of the cases of a disease. The human eye is notoriously prone to seeing clusters in randomly scattered points, and numerical techniques are needed to allow for the variation in the location of the background population for example and test whether a cluster genuinely exists.

The mapping of incidence rates of disease, or of uptake rates of services is also important in the management of health resources, a topic which is becoming increasingly important with the growing pressure on health services in many countries. In the UK, the need for efficient management has been increased by the separation of the health sector into providers and purchasers, which has generated a need for both groups to manage their resources as cost effectively as possible.

This is not to suggest that all health data analysis requires sophisticated techniques - for example the production of standard choropleth or dot density maps for standard reporting areas is still a very useful tool for many health professionals. However, we would suggest that the use of interactive visualisation techniques and spatial statistics is not limited to academic research, but has a real relevance to all those with an interest in health data.

In this paper, we describe the development of a software system called SAGE (Spatial Analysis in a GIS Environment) which is designed to assist in the analysis of area-based health data. In the next section we briefly describe some of the other work which has been done on the provision of extra analytical facilities for spatial data. This is followed by a description of the approach we have taken which we believe combines the best elements of previous work. The architecture of SAGE is briefly described, followed by an outline of the facilities provided, together with some examples of their use.

Linking GIS and Spatial Analysis

A number of reviews of the work on providing GIS with better spatial analysis facilities have already been written, and the reader is referred to these for a complete review of the field (Bailey 1994, Haining et al 1996). Here we concentrate on outlining the main approaches taken, identifying their strengths and weaknesses.

One of the earliest systems to illustrate the potential benefits of interactive visualisation tools for the analysis of spatial data was developed by Haslett et al (1990). The system ran on the Macintosh and was a purpose written package which provided the ability to produce multiple linked views of a set of data. In an early example, a map of geochemical sample locations in one window was linked to a scatter plot of the copper and zinc contents of the samples - a series of particularly anomalous values on the scatter plot were highlighted and the relevant locations on the map were automatically highlighted, indicating that all the samples came from the same region. Since then, a number of workers have developed a series of similar systems using different programming environments to speed up the development process (Dykes 1995, Brunsdon and Charlton1995).

Although such systems can produce a wide range of graphic displays, and can link these together, they all suffer the drawback that the data must be imported into them from the GIS. This has a number of problems:

- It is inconvenient to have to export all the locational and attribute data from the GIS. If the cartographic facilities of the GIS are going to be used to produce presentation quality graphics of the results, then information may need to be re-imported (for example if new variables have been created in the process of the analysis).

- Such systems need to provide a means of producing cartographic displays of the data, which is wasteful of effort when GIS already contains these capabilities.

- The systems cannot perform any analysis which requires the actual manipulation of the spatial data. This means polygon dissolve or overlay operations still have to be carried out in the GIS.

- Therefore a number of workers have investigated ways of linking any new analytical facilities more closely to a GIS package. Two types of approach are common here.

- One method is to call a routine from within the GIS which can perform the necessary analysis, and this has been used to make various statistical techniques available. Ding and Fotheringham (1992) used this approach to create SAM, a set of routines for the spatial analysis of data held in ARC/INFO. This used ARC/INFO AML and the &CALL facility to call a set of FORTRAN routines for calculating Moran's I, a measure of the spatial autocorrelation present in a set of data. This calculation requires a matrix identifying the connections between a set of areas, which can be constructed using the topological information held in an ARC/INFO polygon coverage.
  This approach effectively makes the new facilities appear like an addition to the GIS and means there is no need to export data. However it is difficult to use this mechanism to build the sort of interactive visualisation facilities provided by a purpose written system, since one is limited by the visualisation facilities of the GIS. One attempt to do this is the work of Batty and Yichun (1994) in their implementation of an urban population density model, also in ARC/INFO. This provided a map of the urban area, split into a series of zones, with a graphical display of some of the zone attributes. These two views of the data were linked, such that selecting a zone caused the relevant data points on the graph to be highlighted, but in fact the two views were being drawn in the same ARCPLOT window. The system was therefore special-purpose and could not be extended to other types of data.

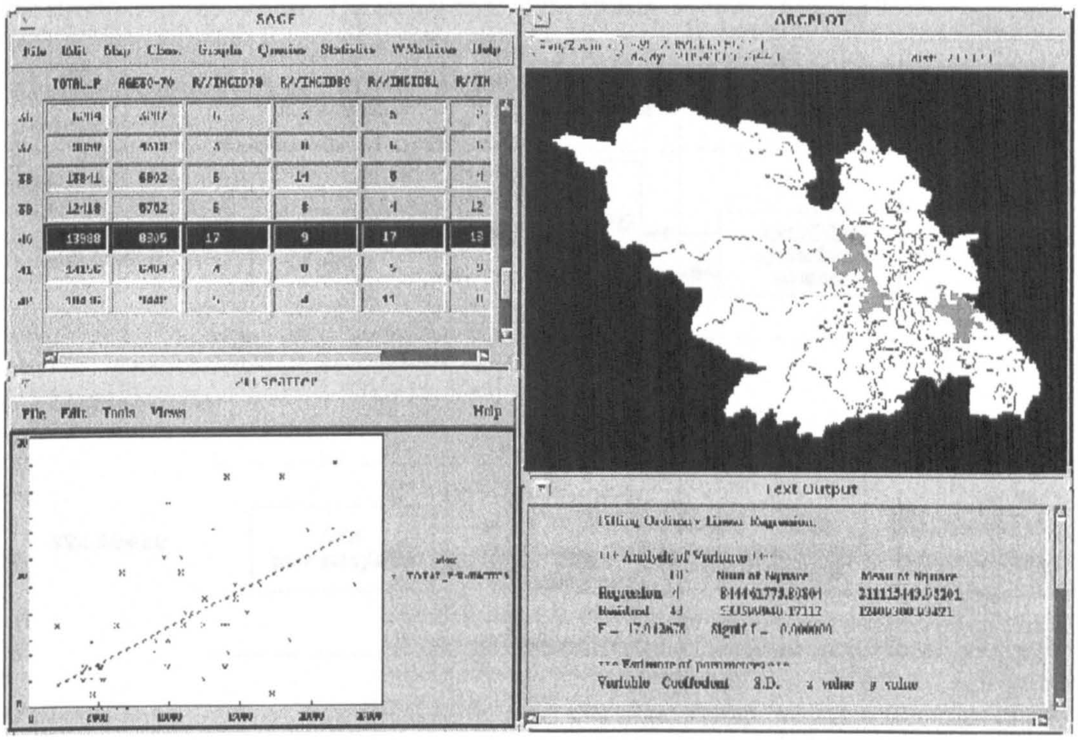
- The other method of linking new facilities to GIS has been to use a completely separate package to provide the extra functionality, and create some sort of link between this and the GIS. At the simplest level, the link can consist of the manual transfer of files (Anselin et al 1993) although it would clearly be possible to automate this and create a uniform user interface. This approach is clearly potentially flexible, but since the two packages are now completely separate, it is not possible to link a map being drawn by ARC/INFO with a graph being drawn by another package.

- The approach we have taken in writing SAGE is to keep the new analysis facilities in a separate package which interfaces with ARC/INFO using the client/server approach. This allows the flexibility to provide a range of visualisation and analytical techniques while still being able to call upon the facilities of the GIS when necessary.

The design of SAGE

In designing SAGE, our aim was to provide a comprehensive set of tools to assist in the analysis of health data. The tools are described in more detail below, but they include both interactive visualisation methods, including the linking of graphical, tabular and cartographic displays, plus more quantitative tools such as exploratory and confirmatory statistics. In addition we wanted to make use of existing software capabilities wherever possible, which meant that it was planned to use the GIS for the storage and cartographic display of the data, and use other packages for some of the other functions, such as tabular display of the attribute data, statistical calculations etc.

These different packages are presented to the user via a consistent user interface as shown in Figure 1.


SAGE
File Edit Map Class. Graphs Queries Statistics Windows Help
ARCPLOT
File Edit Tools Views Help
TOTAL_P
Text Output

	TOTAL_P	AGE60-70	R//INCID79	R//INCID80	R//INCID81	R//IN
36	[illegible]	[illegible]	6	3	5	[illegible]
37	[illegible]	4513	3	8	6	5
38	13841	6902	5	14	5	4
39	12418	6702	5	8	4	12
40	13988	8305	17	9	17	13
41	14156	6404	4	0	5	9
[illegible]	[illegible]	[illegible]	[illegible]	4	11	11

```
Fitting Ordinary Linear Regression.

*** Analysis of Variance ***
             DF    Sum of Square      Mean of Square
Regression   4     844461773.80804    211115443.45201
Residual     43    533599940.17112    12409300.93421
F = 17.012678     Signif F = 0.000000

*** Estimate of parameters ***
Variable   Coefficient   S.D.   t value   p value
```

Figure 1: Typical screen layout during a SAGE session.

Four windows are shown. In the top left hand corner is the window from which the operation of SAGE is controlled - this is a tabular display of the attributes associated with the ARC/INFO polygon coverage. The coverage itself is displayed in map form in the ARCPLOT window on the upper right. Notice that two of the polygons on the map are highlighted - these correspond to two of the rows in the table which are also highlighted (one is visible in Figure 1 - the other is further down the table). Selecting one or more rows of the table will cause the relevant polygons to be highlighted on the map - the selection can be a manual one or the result of an SQL-like query. Conversely, selecting one or more polygons on the map will cause the relevant rows in the table to be highlighted - again the selection can be manual or the result of a spatial query. All these functions can of course be performed within ARC/INFO itself - however in SAGE, this linkage extends to other windows such as the scatter plot in the lower left hand corner - although it may not be very clear, the three uppermost points on the plots have been selected, and it is this selection which has actually caused the rows in the table and the polygons on the map to be highlighted. Other windows could also be opened up and these would also be linked to the existing views. For instance we might decide to see whether the outliers on the scatter plot related to polygons with small population values (and hence potentially unreliable incidence rates) and this could be checked by creating a window with a box plot of the population values for each polygon.

As explained above, the system consists of a number of pieces of software, linked together using the client/server mechanism. The architecture of the system is shown in Figure 2.

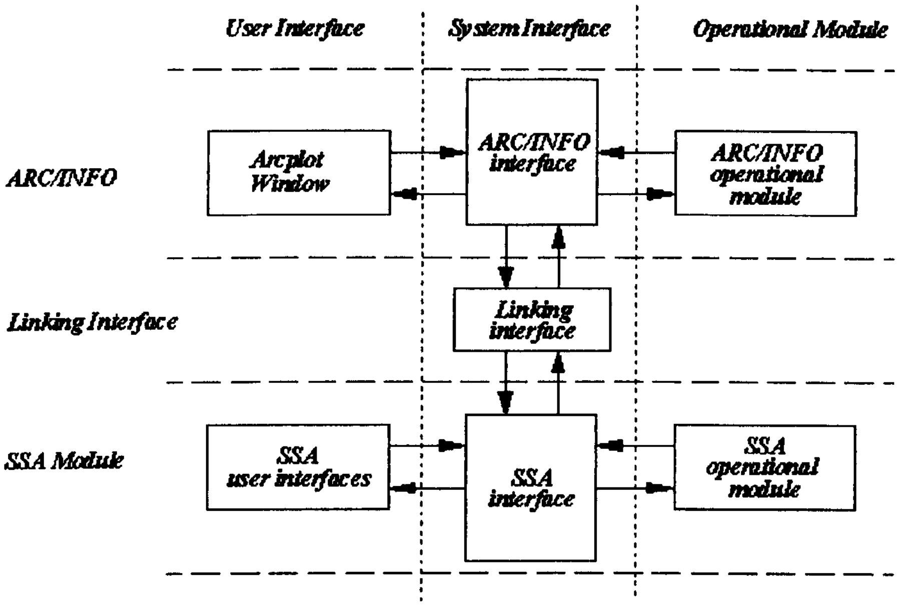


Figure 2: Architecture of SAGE

In all, three pieces of software are involved, each shown as a row in the diagram: ARC/INFO, running in server mode, a spatial statistical analysis (SSA) package and a Linking Interface (LI) both running as clients. As shown by the columns, both ARC/INFO and the SSA can be thought of as having three elements - a user interface, an operational module that performs any computation required and an interface which allows the packages to communicate both internally and with external processes.

The system is controlled from the SSA, the heart of which is a purpose written C program. However a great deal of use has been made of existing software in the construction of the SSA - for example the software which produces the tabular display in Figure 1 is a modified version of a public domain package and many of the graphical plots are produce by public domain code. This software re-use is possible because all these packages are written using object-oriented programming languages which allows them not only to be used by other packages, but for their basic functionality to be modified as well.

When a user selects a row from the table, this information is passed on to the SSA interface. This keeps track of all the other displays which are currently active, and sends out the appropriate instructions for these to be updated to highlight the selected set of records. In the case of the map display, this means that the SSA sends a request to the LI, which translates this into a set of ARC/INFO commands (largely calls to purpose-written AMLs) which are transmitted as requests to the ARC/INFO server. Communication between the LI and both the SSA and ARC/INFO is performed using the standard UNIX facility of named pipes and RPC. ARC/INFO provides an Inter-Application Communication mechanism (ESRI, 1994) which could be used instead, and which would allow true distributed computing in that an SSA module running on one computer could communicate with an ARC/INFO server running on another one. However, the current version of IAC only has a small communications buffer making it unsuitable for transferring the results of complex requests.

One of the advantages of the SAGE architecture is that it uses data which is stored in ARC/INFO, which means that all the normal functions of data entry, data editing and manipulation are automatically available (although not via SAGE itself - the point is that the data does not need to be exported from ARC/INFO in order to be used by SAGE). However, when SAGE is running it does in fact take a local copy of some of the information held by ARC/INFO, namely selected attribute values from the Polygon Attribute Table and the topology data from the Arc Attribute Table. This is done for three reasons:

1. It speeds the system up because it can work with data held in memory.
2. It allows changes to be made to the data without necessarily altering the actual data held in ARC/INFO. A temporary INFO table is established at the start of the session, which is linked to the coverage PAT using a RELATE. Any changes, such as new attribute columns, are thus added to this table, and only added to the original PAT if requested by the user. This is important because many of the analytical techniques create new attribute columns which are only useful for the purpose of the analysis and don't need to be kept - one example would be the column of residual values which can be created when a regression is fitted between two of the variables.
3. To allow the construction of contiguity information required by certain spatial analysis functions.

Functionality of SAGE

As stated above, SAGE has been designed to support the analysis of area-based health data and so a wide range of analytical tools are provided. The main ones can be grouped together under three headings for the purposes of discussion, although there is some overlap between these categories and they do not actually cover every element of the system.

Exploratory tools

These include the facility of the linked graphical windows which has already been described above. Three types of graphical display are supported:

Cartographic

This makes use of the excellent facilities of ARCPLOT, and one of the strengths of SAGE is that there has been no need to write any software for this purpose.

Tabular

This is the spreadsheet-like display shown in the top left of Figure 1. This package holds a local copy of information from the coverage PAT.

Statistical

A wide range of statistical graph types are provided, including histograms, box plots, scatter plots .

In a sense these facilities can be used as analytical tools in their own right, to explore the data, identify patterns, relationships and outliers and suggest hypotheses which can then be tested using some of the other techniques. In addition they support many of the other types of analysis, as will be described below.

Classification and Regionalisation

A common problem with the analysis of area-based data is that the basic spatial units (bsus) for which the data are available are not well suited to the type of analysis. In the case of health data, the basic spatial units are usually census areas such as the UK Enumeration Districts (EDs). These are relatively small, which can cause sensitivity problems in some cases. For example, in a study of colorectal cancer in Sheffield, Haining et al (1994) were dealing with a disease with approximately 300 cases per year distributed over nearly 1100 EDs. Simply calculating incidence rates on an ED basis was not advisable because the results would be very sensitive to small errors in the data - a single case assigned to the wrong ED because of a mis-diagnosis or an error in geocoding could effectively double the apparent rate in that ED.

One solution to this problem is to use larger areal units. However the next standard unit in the UK, the ward, is too heterogeneous and may be too large. What is needed is the ability to group the basic spatial units into purpose-built regions and so SAGE includes a suite of regionalisation tools. These are described in more detail elsewhere (Wise et al, 1996) but they provide the ability to construct regions which satisfy any combination of the following three criteria:

Homogeneity

It will usually be important that the bsus which are merged into regions are similar in terms of one or more attributes. If studying the relationship between health and deprivation for example, it makes sense to do this using areas which have relatively uniform levels of deprivation, rather than being made up of a mixture of affluent and deprived areas.

Equality

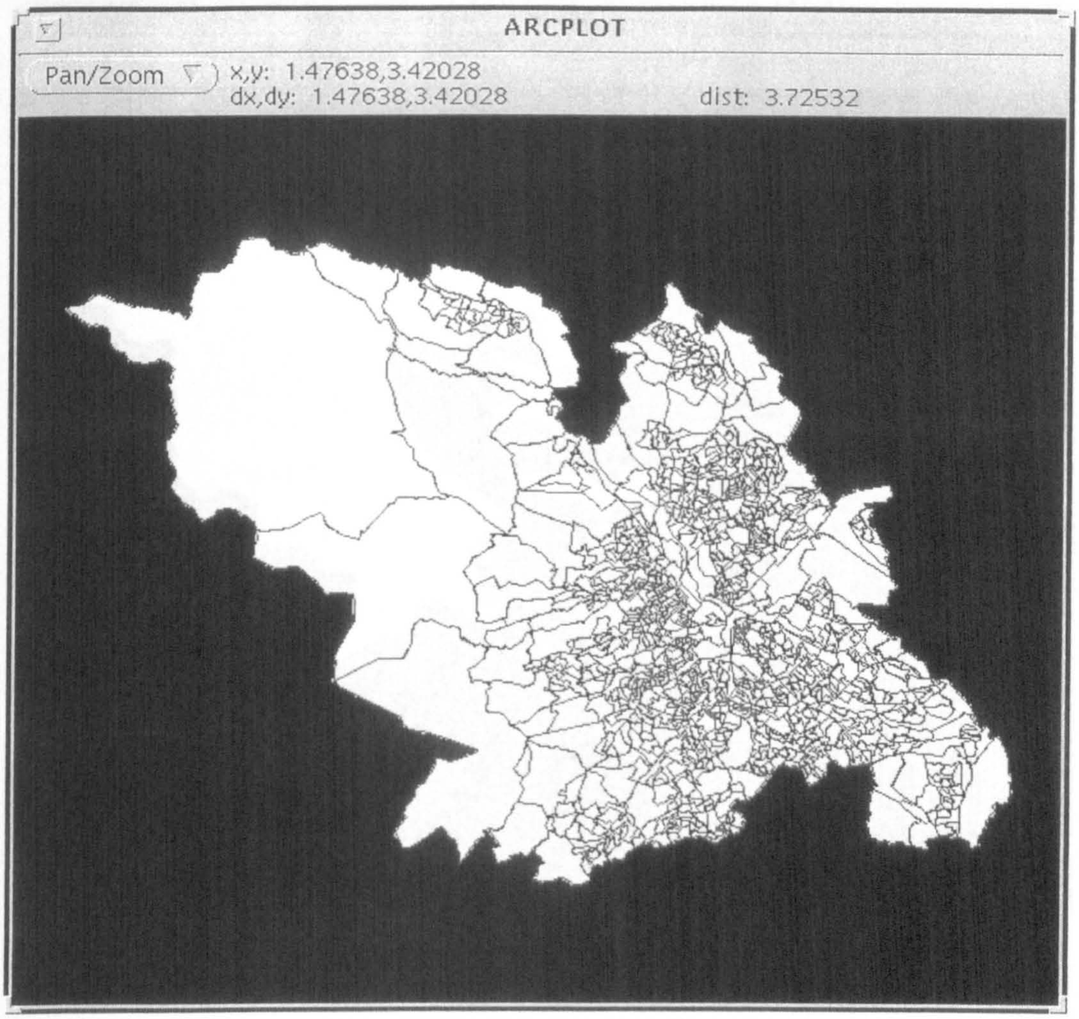
When calculating incidence rates for health data, it is useful if the regions have similar populations, and so SAGE has an option to create regions which have equal values of some attribute - population is the clearest example but the user is given complete freedom to select any of the columns in the attribute table to be equalised.

Compactness

Simply grouping similar bsus may produce regions which have strange shapes. In some cases it may be desirable to constrain the shape, and try and produce regions which are reasonably compact - this will be important if the regions may be used for administrative purposes for example, but it may also accord with one's intuitive notion that natural regions within cities for example will form compact zones, possibly centred on some focal area.

Since these criteria are often competitive - forcing regions to be compact will almost certainly mean merging bsus which differ more than if homogeneity is the only criterion - they may each be weighted from 0 to 100% in importance.

Regionalisation is a long-standing research topic in many areas and it is well known that it is very difficult to find the best possible regionalisation given a set of bsus (partly because the number of possible regionalisations is enormous, so that it is impossible to try them all to find the best (Cliff et al 1975)). One of the strengths of using SAGE is that it is possible to construct several different regionalisations and compare them using some of the exploratory tools described above.



ARCPLOT

Pan/Zoom x,y: 1.47638,3.42028
dx,dy: 1.47638,3.42028 dist: 3.72532

Figure 3: 1981 Census Enumeration Districts of Sheffield

Figure 3 shows a map of the 1981 EDs for Sheffield, for which infomation on deprivation measured using the Townsend index and population was available. The regionalisation tools were used to produce the 30 regions shown in Figure 4, based on two equally-weighted criteria: (1) homogeneity in terms of deprivation (2) equality in terms of population.

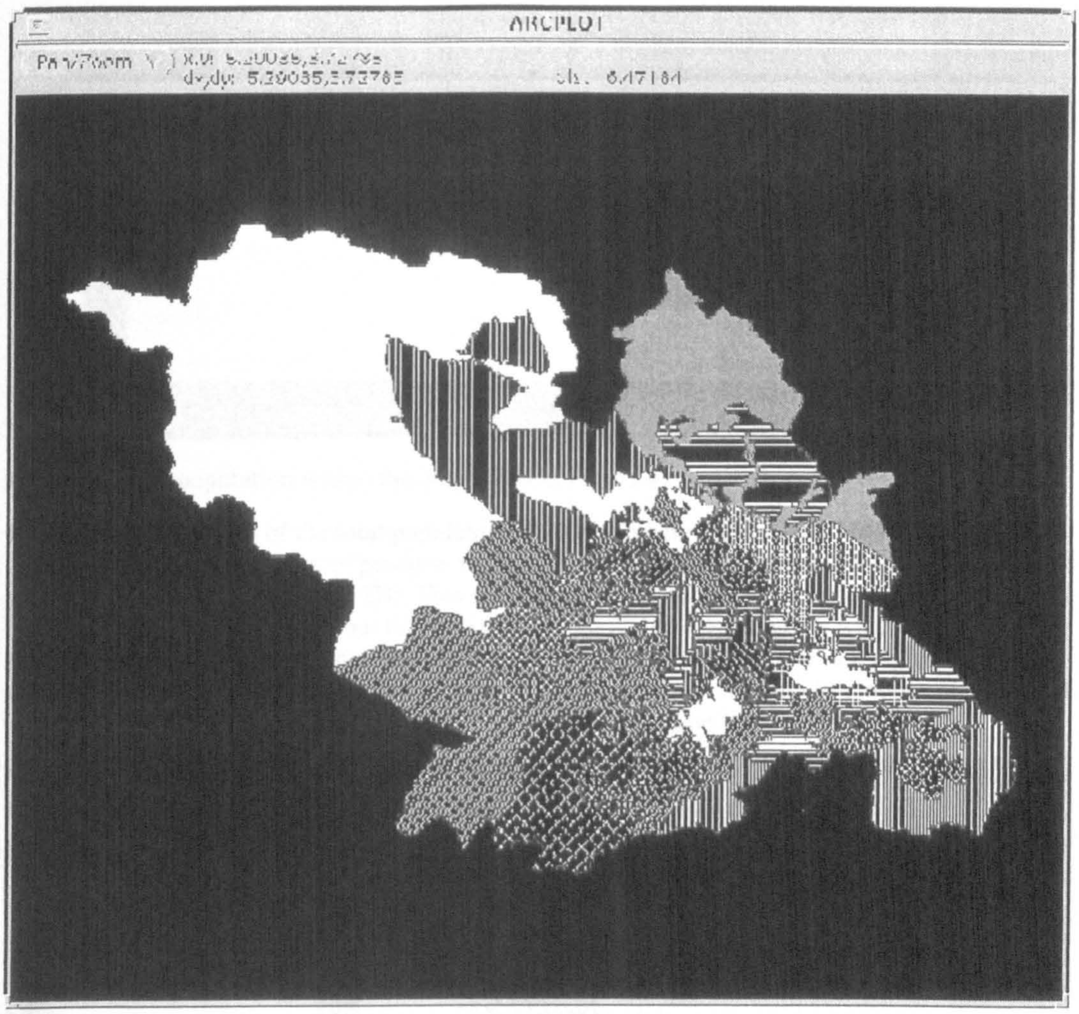


Figure 4 : 30 regions constructed from EDs. Criteria used were that deprivation within regions should vary as little as possible and that regions should have equal population.

The regionalisation process produces a new column in the attribute table held in SAGE, which indicates which region each ED belongs to. At this stage the polygons are not dissolved to form new polygons for the regions which means it is possible to use the graphical tools of SAGE to look at the results of the regionalisation before creating a new coverage. From the map it can be seen that without a compactness criterion, the regions are rather strange shapes in some cases.

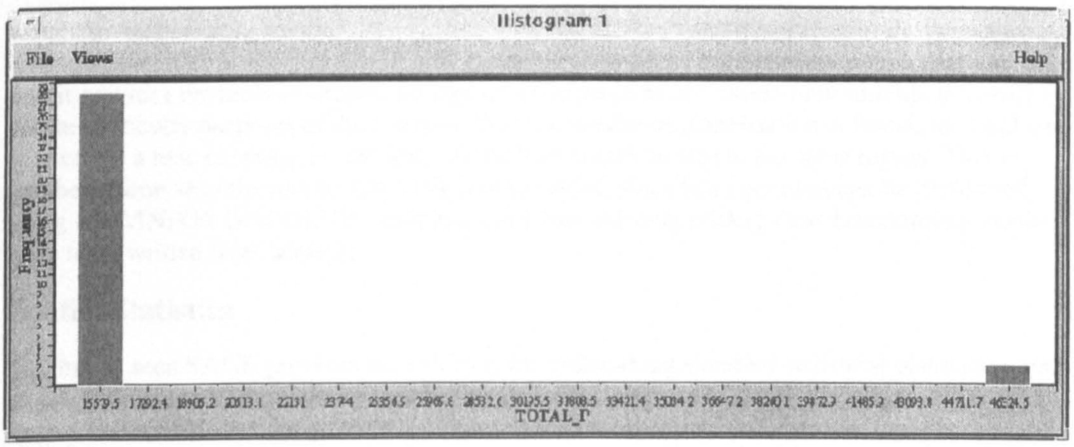


Figure 5: Total population within the 30 regions in Figure 4.

Figure 5 is a histogram of the total population which each new region would have - the graph drawing tools have an option to produce a graph of one variable (population in this case) grouped according to another (the region ID). This shows that the majority of the regions have population counts which are very similar, but that two regions would have populations which are much larger. If this was a problem, then it would be possible to go back and redo the regionalisation with the population equality criterion given a greater weighting - this would produce another new column in the attribute table which could be graphed in the same way for comparison.

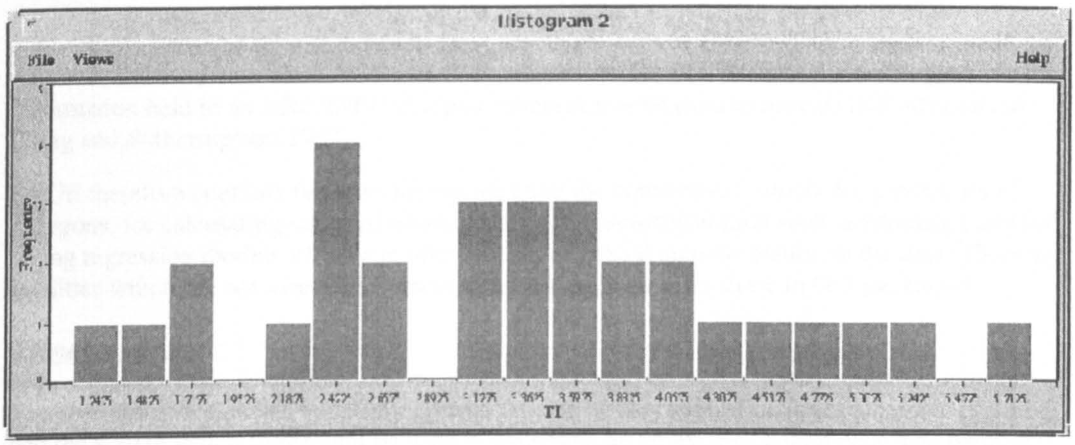


Figure 6: Inter quartile range of deprivation levels in 30 regions in Figure 4.

Figure 6 shows a graph designed to assess the homogeneity of each new region. The inter-quartile range of the ED deprivation scores within each region has been calculated and plotted, again using the SAGE histogram tool. The inter quartile range for the whole of Sheffield was 4.66, Many of the new regions have values of below 2 which indicates that they are reasonably homogeneous, although there are a few with rather large values. Again depending on the nature of the analysis to be undertaken, it might be desirable to try the regionalisation with a greater

weight given to homogeneity.
The key point here is not that SAGE will guarantee to produce the optimum regionalisation, but that it provides the tools to explore a range of different ones and assess how suitable they will be for the particular purposes of the analysis. When a suitable regionalisation is found, the final step is to create a new coverage by merging all the bsus which belong to the same region. This is another reason why the link to ARC/INFO is so useful, since this operation can be performed using ARC/INFO's DISSOLVE command - in a free standing package this functionality would have to be written from scratch.

Spatial Statistics

As already seen SAGE provides the facilities for undertaking standard statistical operations, such as calculating summary statistics, producing statistical graphs and fitting regressions. It is well known that spatial data has particular properties which can create problems for certain standard statistical methods (Haining 1990). It is commonly the case that values for a particular variable will be very similar for neighbouring areas reflecting the underlying spatial trends in the phenomenon in question. For example, deprivation levels will vary in broad patterns across a city, and will not change abruptly at ED or ward boundaries. This tendency of neighbouring areas to have similar characteristics is called spatial autocorrelation, and where it is present it violates one of the basic assumptions of most classical statistical techniques, that the sample values should be independent of one another. This in turn can render many of the significance tests which are normally performed invalid.

A range of methods have been developed to deal with these problems (Haining 1990). Many of the methods require some knowledge of the spatial autocorrelation in the data, or of the connectivity between areas (i.e. which areas are neighbours of which). This is another good reason to link a system like SAGE to a GIS such as ARC/INFO, because the topological information held in an ARC/INFO polygon coverage can be used to provide this information (Ding and Fotheringham 1992).

SAGE therefore contains facilities for constructing the connectivity matrix for a given set of polygons, for calculating standard measures of spatial autocorrelation such as Moran's I and for fitting regression models which can take account of spatial autocorrelation in the data. These are facilities which are not available in most statistical packages, let alone in GIS packages!

Discussion

The preceding section has hopefully given a flavour of the range of facilities available in SAGE, and the way in which they might be used to undertake an analysis of health data. The types of analysis which such a system allows will range from largely exploratory approaches, using the graphical tools and the ability to link the different views to simply look at patterns and relationships in the data, through to a more formal approach which might begin by constructing a set of areal units with desirable properties of homogeneity and equality of population, establish the strength of a relationship between health and other factors by fitting a regression model, and then use the graphical tools to look at outliers from this model.

There are often calls for extra facilities to be added to GIS, which is partly a reflection of the widespread use of spatial data in many areas of research and commerce, and the great interest in the power of GIS software for handling spatial data. However, it is not always possible or even advisable simply to add in extra functionality to an existing package such as ARC/INFO. Apart from anything else, there is the risk of turning an already large system into an unwieldy giant, too complex for anyone to hope to master.

The approach described here for adding extra functionality to ARC/INFO has a number of advantages, which may make it of interest as a general mechanism for linking specialised software to a general purpose GIS. There are two advantages in any approach which links new software to GIS, rather than simply importing data from it:

1. It avoids duplicating the GIS database.
2. It means that all the standard facilities of the GIS, such as cartographic display and standard GIS analysis operations such as polygon dissolve, are automatically available.

Performing the link via a client/server architecture can also make the resulting system very flexible, as has been illustrated in the case of SAGE, and provides the potential for a distributed strategy with interactive graphical software running on a desktop PC and communicating with a GIS server running on a powerful central computer. This type of approach is likely to become more common in the future, but there are currently two impediments to its widespread adoption:

1. The lack of GIS packages offering the possibility of client/server working. ARC/INFO has taken a lead in this respect, although there are still limitations in the current version such as the small communications buffer with the IAC mechanism.
2. The lack of a standard GIS language. When linking packages to relational databases, for example, it is possible to use a standard language, SQL, which makes it easy to interface software to more than one database. The same is not true of GIS - if SAGE were to be linked to a GIS other than ARC/INFO, this would require the rewriting of the parts of the system which translate user requests into commands for the GIS, and which deal with the information which is returned.

It will be interesting to see whether future developments remove these impediments.

Conclusion

In this paper we have described the design and implementation of a software system for the analysis of area-based health data which combines the power of ARC/INFO with new software for visualisation and statistical spatial analysis. We have tried to show that the facilities this system offers will be of potential interest to many people working with health data, and not just to a handful of academic researchers, but whether this is true will only become clear once the system has been tested in practice. The system is still a prototype version, but if anyone is interested in discussing its possible use in their own work, they should contact the team as indicated below.

Acknowledgements

The authors acknowledge receipt of grant R000234470 from the UK Economic and Social Research Council which has made the work reported here possible.

References

Anselin, L., Dodson, R.F. and Hodak, S. (1993) Linking GIS and Spatial Data Analysis in Practice. *Geographical Systems* 1 (1), 3-23.

Bailey, T. C. (1994) A Review of Statistical Spatial Analysis in Geographical Information Systems. In Fotheringham S and Rogerson P. (ed) *Spatial Analysis and GIS.* Taylor and Francis, London. 11-44.

Bailey T.C and Gatrell A.C. Batty, M. and Yichun, X. (1994) Urban Analysis in a GIS Environment : Population Density Modelling using ARC/INFO in Fotheringham, S. and Rogerson, P. (ed) *Spatial Analysis and GIS*, Taylor and Francis, London, 189-220.

Brunsdon C and Charlton M. (1995) A Spatial Analysis Development System using LISP. *Proc. GISRUK '95*, 155-160.

Burrough P.A. (1986) *Principles of geographical Information Systems for Land Resources Assessment.* Oxford University Press. Cliff, Haggett, Ord, Bassett and Davies. (1975) *Elements of spatial structure.* Cambridge University Press. Cross A. and Openshaw S. (1991) Crime pattern analysis: the development of ARC/CRIME. *Proc. AGI 91* 3.28.1-3.28.6. Westrade Fairs, London.

Ding, Y. and Fotheringham, S. (1992) The Integration of Spatial Analysis and GIS. *Computers, Environment and Urban Systems*, 16, 3-19.

Dykes J. (1995) Pushing Maps Past their Established Limits : a Unified Approach to Cartographic Visualization. *Proc. GISRUK '95*, 78-95.

ESRI (1994) *ArcDoc version 7.0 (online help).* ESRI, Redlands, CA.

Goodchild, M.G., Haining, R.P. and Wise, S.M. (1992) Integrating Geographic Information Systems and Spatial Data Analysis: Problems and Possibilities. *Int. Journal of Geographical Information Systems*, 16, 407-24.

Haining, R.P. (1990) *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge University Press.
Haining, R.P., Wise, S.M. and Blake, M (1994) Constructing Regions for Small Area Analysis: Material Deprivation and Colorectal Cancer. *Journal of Public Health Medicine*, 16, 429-438.

Haining R.P., Wise S.M. and Ma.J. (1996) The design of a software system for the interactive spatial statistical analysis linked to a GIS. *Computational Statistics (in press)*

Haslett. J, Wills G. and Unwin A. (1990) SPIDER - an Interactive Statistical Tool for the Analysis of Spatially Distributed Data. *Int.J.Geographical Information Systems* 4(3),285-296.

Snow J. (1854) *On the mode of communication of cholera.* Churchill Livingstone, London.

Wise S.M., Haining R.P and Ma.J. (1996) Regionalisation tools for the exploratory spatial analysis of health data. Paper presented at the 28th International Geographical Congress, The Hague, August 4th-10th, 1996.

Steve Wise, Jingsheng Ma, Bob Haining
Sheffield Centre for Geographic Information and Spatial Analysis
Department of Geography
University of Sheffield
Sheffield S10 2TN
UK
Telephone : +44 (0)114 282 4749
Fax: +44 (0)114 272 7919
Email : R.Haining@shef.ac.uk

5 Regionalisation Tools for the Exploratory Spatial Analysis of Health Data

Steve Wise, Robert Haining and Jingsheng Ma

Sheffield Centre for Geographic Information and Spatial Analysis, Department of Geography, University of Sheffield, Sheffield S10 2TN, UK.

5.1 Introduction

This paper considers issues associated with the construction of regions as part of a programme of exploratory spatial data analysis in the case of what Cressie (1991) refers to as "lattice data". Lattice data arise where a study area has been partitioned into a set of zones or regions attached to each of which is a vector that describes the set of attributes for that zone. The focus of this paper will be the analysis of health data so the attributes in question may be health related but may also include demographic, socio-economic and environmental attributes.

Exploratory spatial data analysis (ESDA) comprises a set of statistically robust techniques that can be used to identify different forms of spatial variation in spatial data. ESDA represents an extension of exploratory data analysis (EDA) to handle spatially referenced data where in addition to the need for techniques to identify distributional properties of a set of data there is also a need for techniques to identify spatial distributional properties of the data. Typically these techniques comprise numerical summaries (e.g. measures of central tendency, measures of dispersion, regression) and graphic displays (e.g. boxplots, histograms, scatterplots) but in the case of ESDA cartographic displays make a vital and distinctive additional contribution enabling the analyst to see each attribute value in its geographical context relative to other attribute values. Like EDA therefore, ESDA exploits different methods of visualisation. Moreover like EDA, ESDA is not associated with any one stage in the process of data analysis for the techniques can be appropriate both for preliminary stages of analysis (pattern detection; hypothesis formulation) and later stages (model assessment).

EDA is underpinned by a conceptual data model in which data values comprise an element that is "smooth" (sometimes called the "fit") and an element that is "rough" (sometimes called the "residual"). In the case of ESDA the "smooth" is often associated with large scale patterns (such as spatial trends, patterns of autocovariation or concentration) whilst the "rough" may be outliers, that is individual areas with values that are higher ("hot") or lower ("cold") than found in the neighbouring areas. To identify these data properties in the case of ESDA a number of techniques have been brought together, either adapted from EDA (e.g. median polish to detect trends (Cressie 1984)) or custom developed for the identification of spatial properties (e.g.

spatial autocorrelation tests (Cliff and Ord 1981)). A distinction is also drawn between "whole map" techniques which generate aggregate measures (e.g. the Cliff and Ord statistics for spatial autocorrelation) and "focused" or "local" tests which only treat subsets of the data (e.g. the G statistics of Getis and Ord (1992), Ord and Getis (1995) and others (Anselin, 1995). For a longer review of these techniques and underlying models see Haining (1996).

There are several fundamental properties of the spatial system (the zones) that will influence the results of the application of ESDA techniques. One of these is the definition of the assumed spatial relationships between the zones. This is represented by the so-called "connectivity" or "weights" matrix and represents the analyst's chosen definition of inter-zonal relationships. The analyst may well wish to explore the robustness of ESDA findings to alternative definitions and hence alternative constructions of the weights matrix. At least as fundamental as this, however, is the construction of the zones themselves - the regionalisation (or spatial filter) through which events are observed. An important feature of any system that purports to offer spatial data analysts ESDA facilities ought to be a capacity to look at the spatial distribution of values in many different ways and to observe the robustness, or conversely the sensitivity, of findings to alternative (but equally plausible) regional frameworks. Regionalising by aggregating smaller spatial units is not a sine qua non for ESDA since a variety of smoothing techniques (e.g. kernel smoothing (Bithell 1990)) might be preferable, since they will allow the analyst to stay close to the original data whilst still making it possible to detect spatial properties, but, as will be argued in section 5.2, there are times when aggregation is essential. However in those situations where regionalisation through aggregation is part of the analysis there is a fundamental problem. Whilst the analyst engaged in ESDA may need to be able to construct alternative regionalisations under defined criteria, the analyst will also want to obtain different regionalisations reasonably quickly. This will require trade offs between many things but particularly (at the present time on most generally available hardware platforms) between optimality according to the specified criteria and the time taken to arrive at a an acceptable solution.

The spatial analysis of health data is normally undertaken with one of two objectives in mind:

a. To detect or describe patterns in the location of health-related events such as incidence of a disease or the uptake rate of a service. The focus here is on seeking explanations or causative factors - spatial patterns can rarely provide direct evidence for either but can be suggestive of particular mechanisms (Elliott et al 1992)

b. To assist in the targeted delivery of health services, by identifying areas of need. Here the emphasis is not on explanation but on producing a reliable picture of the variation in health needs across an area, allowing for the variations in other factors such as population age structure.

Both types of analysis require information about the population in the area of interest, whether this is simply a count of people for calculating incidence or mortality rates, or more detailed information relating to the age structure or levels of economic deprivation. Such data is usually available for small areas - in Great Britain for example the reporting areas, known as Enumeration Districts (EDs), normally contain about 150 households. This very detailed level of spatial resolution can cause problems for many types of analysis, as described in detail below, and it is common practice to undertake analysis at more aggregated spatial scales. However, larger standard statistical areas may also be unsuitable because they are too large or are too inhomogeneous. There is therefore a need to be able to merge the basic spatial units into tailor-made groups by means of a classification or regionalisation procedure.

As Openshaw and Rao (1994) point out, the availability of digital boundaries and Geographical Information System software capable of manipulating them has made this task far easier than before, and opened up the possibility of making regionalisation an intrinsic part of the analysis of spatial data, rather than a one-off, time-consuming exercise.

This paper describes work to develop a suite of classification and regionalisation tools suitable for use in the analysis of health data. The research forms part of a larger project which is developing a software system for the analysis of health data, with the emphasis on rapid interactive visualisation of data supported by a suite of cartographical, graphical and analytical tools. The system uses a GIS for the storage, manipulation and cartographical display of the spatial data and is called SAGE - Spatial Analysis in a GIS Environment. The architecture of the whole system is described elsewhere (Haining et al, 1996) but two of the key design elements of SAGE are important in what follows:

a. The focus of the software is on the provision of tools for data visualisation. A range of display techniques are provided including map display, histograms, box plots and scatter plots. More importantly, these different views of the data are linked so that highlighting an area on the map will also cause the equivalent point on the scatter plot to be identified. This is in keeping with much of the software that has been developed following the pioneering work of Haslett et al (1990,1991).

b. The emphasis is on building a system by utilising existing software and techniques wherever possible, rather than writing it from scratch. One of the important consequences of this is that users can explore data which they already hold in a GIS without having to export it to another software package (although the facility to export to some specialised packages is being provided). It also means that all the standard tools of data input and management are provided by the GIS.

The structure of this paper is as follows. We firstly elaborate on the importance of classification and regionalisation in the analytical process, and then describe the criteria governing the design of the classification and regionalisation tools in SAGE. This is followed by a brief description of the methods used, some initial results and a discussion of future research directions.

5.2 The Importance of Classification and Regionalisation in Analysis

A common problem with area-based analyses is that the results of the analysis may be sensitive to the choice of spatial unit. It is well known that census EDs are designed to minimise enumerator's workloads, and to nest within higher order areas such as wards and districts. As a result these basic spatial units are quite varied in terms of areal extent, population composition and population size and do not reflect the underlying variation in socio-economic conditions. If analysis reveals a link between levels of deprivation and levels of ill-health or leads to the identification of disease "hot-spots" there remains the possibility that if the boundaries of the areas were drawn differently, then different results might be found. This is the well-known modifiable areal unit problem or MAUP (Openshaw 1978, Fotheringham and Wong 1991) although we would emphasise that in a well defined analytical context only a subset of possible zoning systems would be considered a plausible basis for analysis.

Despite these problems, there are a number of reasons why it is often beneficial to group the basic spatial units to form larger regions as the framework for the analysis:

a. To increase the base population in each area, so that incidence rates for example will be based upon a larger sample size and hence more robust to small random variations in the number of cases. This will be particularly important when the events being studied are rare (such as incidences of a rare cancer).

b. To reduce the effect of suspected inaccuracies in the data. When dealing with small areas, an error of one or two in the count of health events could have a large effect on the calculated rates, whereas such errors will have far less effect for larger areas (assuming the positive and negative errors for each ED tend to cancel out when the areas are aggregated - any systematic bias will simply carry forward into the aggregated data). In the UK census, counts are routinely modified by the random addition of -1,0 or +1 as a further protection of confidentiality thus producing a level of 'error' in the population data which could similarly affect the calculation of expected rates in small areas.

c. To reduce the effect of suspected inaccuracies in the location of the health events. A common form of locational reference for health data in the UK is the unit postcode, which is shared by 15 addresses on average (Raper et al 1992) and therefore has an inherent precision of the order of 100 m. This problem is compounded by the fact that the file which is widely used to assign UK National Grid references to postcodes (the Postzon file) gives the location of the first address in the postcode to the nearest 100m, and is known to contain errors. In a study in Cumbria, Gatrell et al (1989) found that using the Postzon file to allocate postcodes to EDs (using a point in polygon operation) resulted in the incorrect assignment of 39% at the ED level, but only 3% at the ward level (the first level of aggregation of EDs to standard statistical units). (This is quoted as an example of the effects of scale on the problem - for the 1991 UK census more accurate data is available than was the case for the 1981 census). Using larger areal units essentially means that less of the cases are near the edges of units, and hence less are likely to be assigned to the wrong unit.

d. To make the analysis computationally tractable. Some spatial analytical techniques require the manipulation of an NxN connectivity matrix, where N is the number of regions in the analysis (and element (ij) in the matrix defines the spatial relationship between regions I and j). In the study of Sheffield by Haining et al (1994) the initial dataset was based on 1000 EDs, which would give a contiguity matrix with a million elements.

e. To facilitate visualisation. An important element of visualisation techniques is that they must operate quickly enough that the user is encouraged to explore different views of the data and use different techniques. Dealing with large numbers of small areas may cause the system to be too slow. In addition, it may be difficult for a user to see trends and patterns in the data when confronted by maps and graphs relating to large numbers of areas (although techniques such as kernel smoothing may also be employed to deal with this problem).

All these are situations where regionalisation might be performed prior to some other form of analysis, such as fitting a model to the data. However, as noted in the introduction, regionalisation can also be seen as an intrinsic part of the analytical process, in the sense that it can be considered as a means of visualising broad scale patterns in the data, and as a technique for testing the robustness of model results.

The construction of new regions can be achieved by classifying the basic spatial units, and merging together those that fall into the same class. A map of these new regions will often appear very fragmented however, and it is often desirable to use a regionalisation approach which will ensure that the regions form contiguous areas on the map.

5.3 The Design of a Regionalisation System for the Analysis of Health Data

Both classification and regionalisation are topics of long standing in many areas of science, and a large number of methods now exist for both. The intention was to make use of existing techniques wherever possible, but before reviewing the main approaches described in the literature we describe the design decisions which governed the choice of methods for SAGE.

Different applications of regionalisation will have different criteria. Cliff et al (1975) suggested that, in general, an 'optimal' regionalisation should simultaneously satisfy the following criteria:

a. It should be simple, in the sense that a solution which produces few regions is better than one which produces many.

b. Regions should be homogeneous in terms of the characteristics of the zones which comprise them.

c. Regions should be compact.

As they point out, these criteria are competitive. For example, the simplest regionalisation would group all zones into one region, but this would be the least homogeneous. These are not universal criteria - in political redistricting for example homogeneity may well be very undesirable (since it smacks of gerrymandering) and compactness plus equality in terms of population are more likely to be important (Horn 1995, Sammons 1978).

In terms of the analysis of health data, the following three criteria seem most appropriate as objectives:

Homogeneity. The new regions should be made up of zones which are as similar as possible in terms of some characteristic which is relevant to the analysis. In a study relating ill health and deprivation, for example, it clearly makes sense to use regions in which deprivation levels are relatively uniform.

Equality. In some cases it may also be important that the new regions are similar to one another in certain respects. The most obvious example of this is in terms of their population sizes. It is known that there are problems in comparing incidence rates which have been calculated on the basis of different base populations, and adjustment techniques have been developed to allow for this (e.g. Bayes adjustment (Clayton and Kaldor 1987). One way to minimise this problem is to ensure that the populations in the new regions are of sufficient size and as similar as possible.

Contiguity. There are two aspects to this. Firstly there is the distinction between classification, in which the location of the original zones is not considered, and regionalisation, in which only neighbouring zones may be merged to form the regions. Secondly, in the case of regionalisation there is the question of the shape of the regions, since merely enforcing contiguity may lead to long thin regions rather than compact ones. This is sometimes regarded as merely a cosmetic aim (e.g. Openshaw 1994) but Horn (1995) argues that it accords with our intuitive understanding that regions within cities for example, form tight units of economic and social activity and hence are spatially compact.

It is not necessarily the case that all three of these will be important in all applications. For example, in the design of regions for the administration of health delivery homogeneity may be unimportant, whereas equality of population size and compactness would be very desirable. Conversely for any statistical analysis homogeneity is almost certain to be desirable, and perhaps equality, but contiguity may be undesirable, especially since enforcing it will probably reduce the level of homogeneity and equality. This implies that it should be possible to specify any combination of these three criteria in the classification or regionalisation process, and since the objectives may be competitive, it must be possible to decide on the balance between them.

5.4 Classification and Regionalisation Methods in SAGE

The aim of both classification and regionalisation is to group N initial observations into M classes (where M<N) in a way which is optimal according to one or more objective criteria. The terminology varies depending on the discipline, but here we will use the term zones to refer to the initial small areas and groups or regions to refer to the larger units. One of the key problems is that the number of possible solutions with even modest numbers of zones is astronomical (Cliff et al 1975), so that enumerating all possible solutions to find the optimal one is not possible. It is often not even clear what the number of classes or regions should be, unless this is dictated by external requirements (as is often the case with political redistricting for example). Most methods are therefore heuristics in the sense that they cannot guarantee to find a global optimum solution, but employ a range of techniques to attempt to find a good local optimum.

Methods are generally based upon some measure of similarity (or dissimilarity) between the zones - the simplest is Euclidean distance between the values of one or more attributes measured for each zone, although other metrics have also been used. Early approaches used a hierarchical (Lankford 1969) in which zones are merged one at a time. The distance metric is calculated for all possible pairs of zones, and held in a dissimilarity matrix. The two zones which are most similar are then merged to form a group, and the matrix updated by calculating the distance between this group and the remaining zones. This process is repeated until all the original zones have been allocated to a group.

The earliest attempts to produce contiguous regions simply included the coordinates of the zone centroids as additional variables. This will tend to merge adjacent zones, but cannot guarantee to produce M regions when grouping to M classes. This can be guaranteed by building in a contiguity constraint such that only distances between neighbouring zones are considered in deciding which zones to add in next.

Hierarchical methods are very fast since each zone is only considered once, although storage of the whole matrix means memory requirements can be high. However, because each zone is only considered once it is known that these methods produce sub-optimal solutions (Semple and Green 1984).

More recent work has therefore concentrated on iterative techniques, of which the best known is the k means method McQueen 1967, Anderberg 1973). The number of classes (k) must first be specified, and the first k observations are assigned to a class each. All the other observations are then assigned to the class to which they are closest and the class mean is calculated for all classes. On each iteration, each case is examined to see if it is closer to the mean of different class than its current one - if it is then it is swapped. This process continues until the process converges on a solution (i.e. no more swaps are possible). As with the hierarchical methods, the k-means method can be used for regionalisation by building in a contiguity constraint, and this is essentially the method used by Openshaw (1978) in his Automatic Zoning Procedure (AZP).

The k-means approach is rapid and will often converge on a solution in a few iterations (Spath 1980). The main problem is that this solution is often only a local optimum, and so the method is sensitive to the initial allocation of observations to classes - subsequent runs can produce quite different results. One solution to this is simply to run the algorithm several times and save the best result. This seems inelegant and potentially rather slow and cumbersome, and a number of other techniques have been tried to improve the method. In a recent development of the AZP approach, Openshaw (1994) has studied techniques such as simulated annealing and tabu which are designed to prevent the search getting stuck in a local optimum. Simulated annealing does this by using the analogy of a material slowly cooling. In the initial stages when the temperature is high, swaps are allowed even when they make the objective function worse. As the temperature is cooled, the probability of allowing this is reduced until by the end only improving exchanges are allowed. It has been shown that as the rate of cooling approaches zero, the method will find the global optimum. However all these methods result in much slower execution, as far more iterations are necessary to converge to a solution.

An alternative approach to regionalisation is to modify the original dissimilarity matrix to take account of the relative positions of pairs of zones. The idea is to weight the differences such that ones which are close in space are allowed to differ more in terms of their attributes than zones which are distant in space. Some workers have used a simple exponential distance decay but Oliver and Webster (1989) used the estimated semi-variogram to base the weightings on the spatial autocorrelation present in the data. The modified matrix can then be used in a normal classification such as k-means or even a hierarchical approach. This will produce a rapid result, but the initial stage of fitting a variogram adds an extra complication and it is difficult to see how it would be possible to allow the user to control the degree of importance to be attached to the compactness criterion.

It was therefore decided to base the regionalisation tools in SAGE on a k-means classification because this is relatively rapid, and can be adapted to deal with more than one objective function. Its main weakness is that it is known to be sensitive to the initial allocation of zones to regions, and a series of methods have been developed to try and reduce this problem as described in the next section.

5.5 Implementation

The heart of the SAGE regionalisation tools is a k-means based classification procedure which can classify zones according to any of the following functions:

- Homogeneity - the objective is to minimise the within group variance of one or more attributes.

- Equality - the objective is to minimise the difference between the total value of an attribute for all zones in a region and the mean total of this value for all regions.

•Compactness - this is currently achieved simply by treating the X and Y coordinates of the zone centroids as two variables. Although this will tend to create compact regions, it is not ideal because the location of the centroid is to some extent arbitrary, and especially in large zones, different positions for the centroid could lead to quite different results with no obvious way to select which is the 'right' one. A better alternative may be to base the measure on the length of perimeter between regions (Horn 1995) or to use some measure of area shape such as the perimeter/area ratio.

In the case of trying to satisfy competing objectives, an interesting issue is what to do when a swap improves one objective but makes another one worse. The problem is complicated by the fact that the objective functions are measured in different units as described above.

Cliff et al (1975) used two objective functions which took values from 0 (the worst solution for that objective) to 1 (the best). The optimum regionalisation was therefore the one which maximised the sum of these functions. This would also be amenable to weighting, although it relies on finding suitable objective functions which can be scaled in this way.

A alternative solution to the problem was used by Sammons (1987) who found that insisting that all objective functions must improve with every swap lead to poor results because of premature convergence i.e. the method tends to stabilise quickly but at a local optimum which gives a poor result. He therefore adopted a scheme whereby a swap was accepted if it improved any of the functions.

In SAGE the problem is tackled by expressing the changes in the objective functions due to a swap in percentage rather than absolute terms. The percentage changes are summed, and the swap accepted if this sum is positive (i.e. the gains outweigh the losses). One advantage of this scheme is that it is then easy to weight each objective function in calculating the sum. This provides an intuitive means for the user to control the balance of objectives for the regionalisation - any objective can be weighted from 0 (i.e. not considered) to 100%. One option would be to make 100% mean that the objective must be maximised at all costs i.e. no swap to be accepted unless it improves this objective function. However, this would certainly constrain the regionalisation to certain parts of the solution space and may lead to poor solutions (even in terms of the objective set at 100%).

As mentioned above the k-means method is sensitive to the initial allocation of zones into regions. When this is done randomly, different runs of the method will produce very different results. Our solution to this problem is to regard the k means regionalisation as simply one in a set of procedures which a user may use so that the process of regionalisation in SAGE can be visualised as shown in Figure 5.1.

The k means regionalisation forms stage 2 in the process. Prior to this the user may decide how to make the initial allocation of zones to regions. The default is a random allocation, but Sammons (1978) found it improved the results of his political redistricting algorithm if this was started with a 'good' allocation of zones to regions. In the context of political re-districting the main objective was of equal population, and hence the initial allocation could be based on the population distribution in the area.

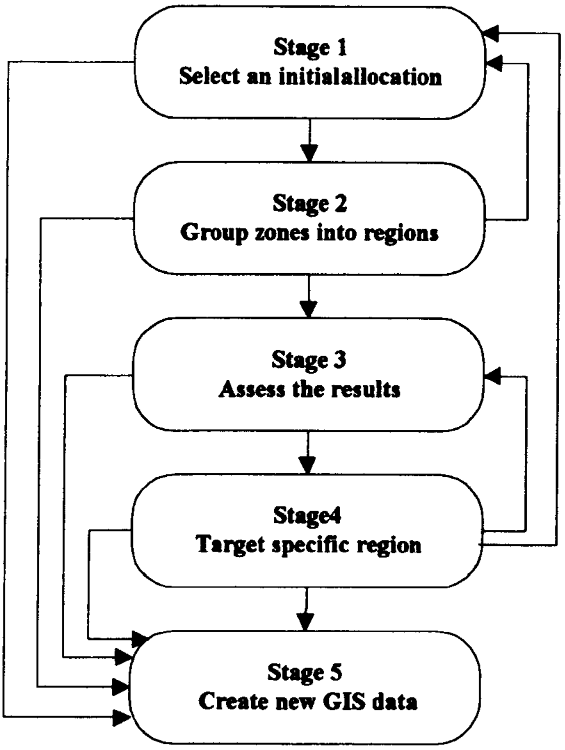
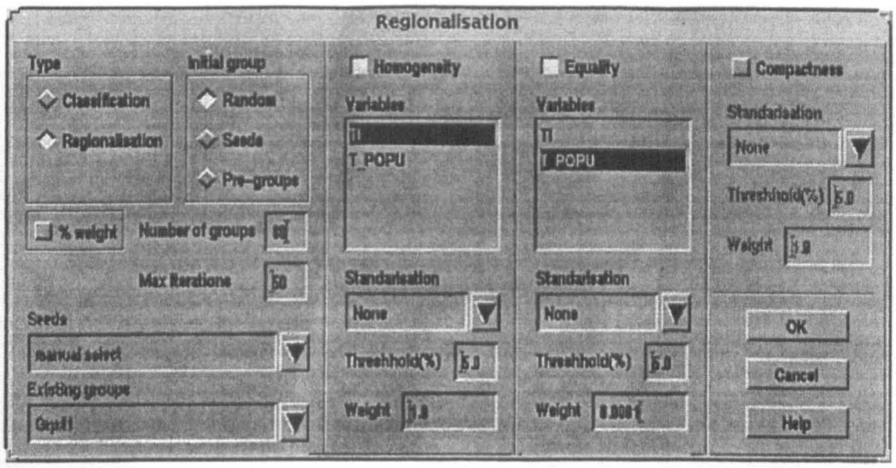


Fig. 5.1. Sequence of Steps for Performing Regionalisation in SAGE

For a more general solution, SAGE employs an adaptation of a method devised by Taylor (1969) in which a regionalisation is based on first identifying 'nodes' - zones which are typical of their local area and which can be regarded as the centre of uniform regions. In Taylor's method each zone was visited in turn, and a count made of the number of zones within a certain distance (D_g) whose attributes differed by less than a specified threshold (D_S). Given the number of regions required (M), the M zones which had the greatest number of similar neighbours were selected as region seeds. In successive steps the remaining zones were added on to the region which they bordered and which they most resembled. As a regionalisation method this has a number of drawbacks - it is not clear how to select suitable values for D_g and $D_{S,}$ and it is a hierarchical method. What is interesting though is the idea of identifying relatively uniform areas of the map from which to grow regions, an idea which is similar to some of the methods used in image recognition to identify objects in images (Rosenfeld and Kak 1982). In the initial version of SAGE, we have simply implemented the first step of Taylor's method as a means of identifying a starting point for the regionalisation.



Number of groups 60
Max Iterations 50
Seeds: manual select
Existing groups: Grp11

Homogeneity
Variables: TI, T_POPU
Standarisation: None
Threshhold(%) 5.0
Weight 1.0

Equality
Variables: TI, T_POPU
Standarisation: None
Threshhold(%) 5.0
Weight 0.0001

Compactness
Standarisation: None
Threshhold(%) 5.0
Weight 1.0

Fig. 5.2. User Interface for Stage 1 and 2

Future work will focus on better methods for identifying uniform regions in which to locate seed points, possibly using some of the LISA statistics which are local measures of spatial association which do not require the choice of threshold values (Getis and Ord 1992, Barkley et al 1995).

The selection of the initial allocation of zones is done on the main menu screen for the regionalisation package which is shown in Figure 5.2. The type of initial group is chosen from the following options:

•Random - this is the default.

•Seeds - selecting this option activates the seeds option lower down, giving a choice on how the seed points are to be chosen. Currently the options are for manual selection and Taylor's method.

•Pre-groups - the starting point for regionalisation can be an existing regionalisation, where each zone has already been allocated to a region, but the boundaries within the regions have not been removed. This could have been produced by an earlier run of SAGE, or by some other software package, with the results imported into an ARC/INFO coverage.

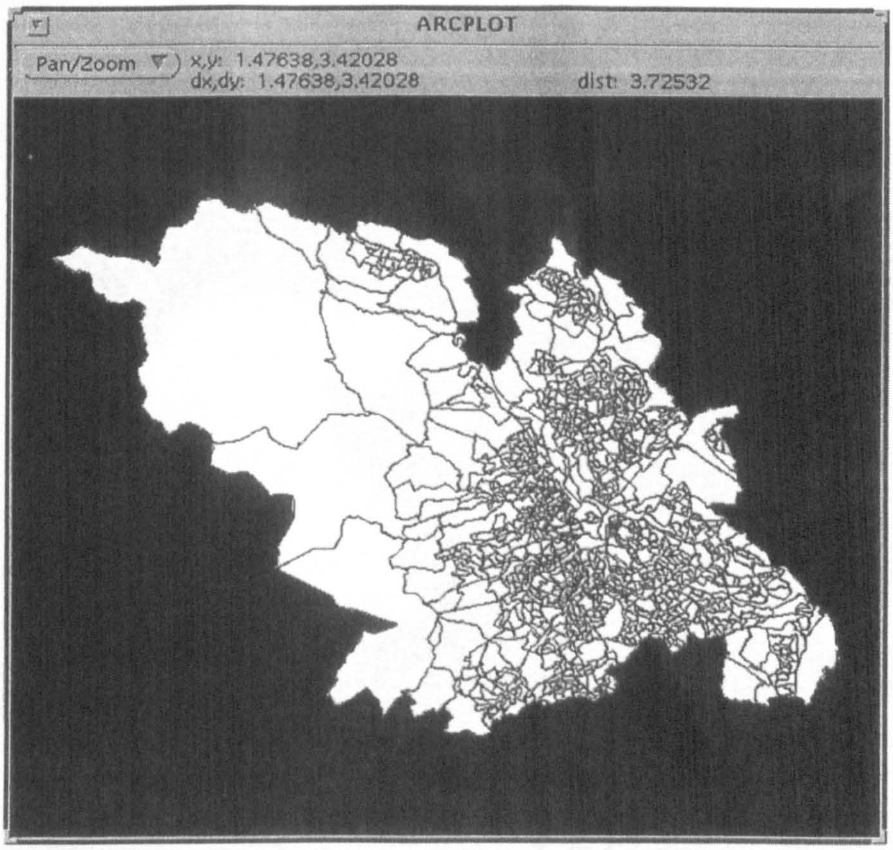


Fig. 5.3. 1991 Enumeration districts of Sheffield

This menu also allows the user to decide whether to perform a classification or a regionalisation and how many groups or regions are to be created. The other menu items allow the user to select which of the three objectives are to be considered, and where more than one is chosen, what weighting is to be attached to each. Two types of weighting are supported, selected by means of the % weight option on the left hand side.

•When %weight is selected, the weights under each of the objective functions are used to weight the percentage change in the objective functions when a

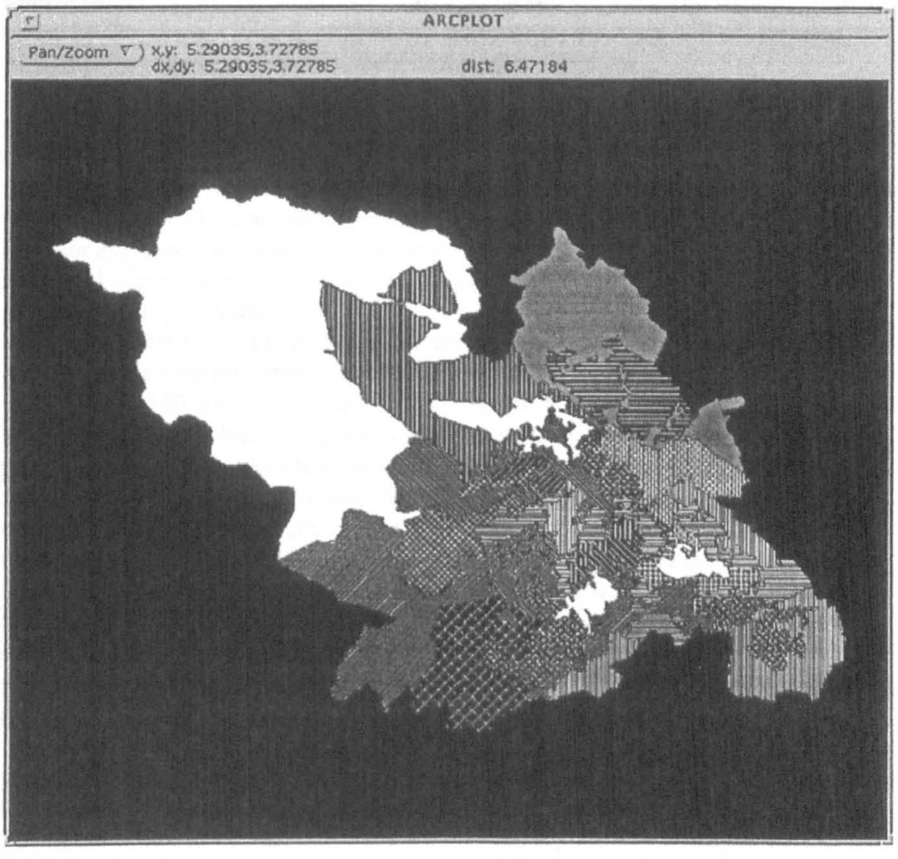


Fig. 5.4. Regions produced from the EDs using the options indicated on Figure 5.1.

swap is being considered.

•When %weight is not selected the weights under the objective functions are used to weight the objective functions directly. Since these are measured in different units, the user has the option of either allowing for this in assigning values to the weights, or standardising all the objective functions using the appropriate options for that function - in the case of the homogeneity variables for example, the values can be converted to z scores.

There is an overall weighting set between 1 and 100%, which is used to weight the percentage changes in each function when a swap is considered. This will allow some functions to get worse while others improve as long as there is an overall improvement. However, a threshold can also be set, such that if a function is made worse by this amount (also expressed as a percentage) the swap will not be accepted, no matter how great the improvement in the other functions.

A key element of the interactive approach taken in SAGE is the importance attached to feedback to the user on the performance of the regionalisation, which takes two forms. As the k means procedure is running, a graph is displayed showing the improvement in the objective function - this allows the user to terminate the process if repeated iterations are making little improvement for example. The main tools for assessing the output of the k means procedure are the range of graphical and statistical visualisation techniques provided in SAGE which are best illustrated by means of a simple example.

Figure 5.3 shows the 1159 EDs which were used for the 1991 census in Sheffield. Associated with the boundary data is a table of attributes, which contains one row for each of the 1159 zones. This table is initially stored as an internal attribute table within ARC/INFO, but SAGE takes a local copy to use for display and analysis purposes. These EDs were regionalised using the options shown in Figure 5.2. The objectives for the new regions were homogeneity in terms of the Townsend index of deprivation and equality of population, each objective being given equal importance (weights of 1). Each was also allowed to become worse by 5% on any given swap. The map of the resulting regions is shown in Figure 5.4, as it would appear in the ARCPLOT window when drawn from SAGE. The results of the regionalisation are stored as a new column in the attribute table, each zone being given the number of the region to which it belongs. Note that at this stage the ED boundaries are not altered - there are still 1159 zones in the coverage.

In order to assess the success of the method in achieving the objectives, the graphical capabilities of SAGE can be used to draw the two histograms shown in Figure 5.5. The lower graph is a histogram of the interquartile ranges of the Townsend Index values for the zones within each region. The interquartile range of values for the Townsend Index for the original EDs was 4.66 , whereas many of the new regions have interquartile ranges of around 1, although some still have high values . The upper graph shows a histogram of total population in the new regions, and shows that most of the new regions have very similar values. Again though one or two regions have populations which are far larger than the rest.

If there are problems with the regions which are constructed, one option is clearly to go back and re-run the regionalisation with modified objectives. Since the results will be saved into the same table, it is relatively easy to compare the effects of

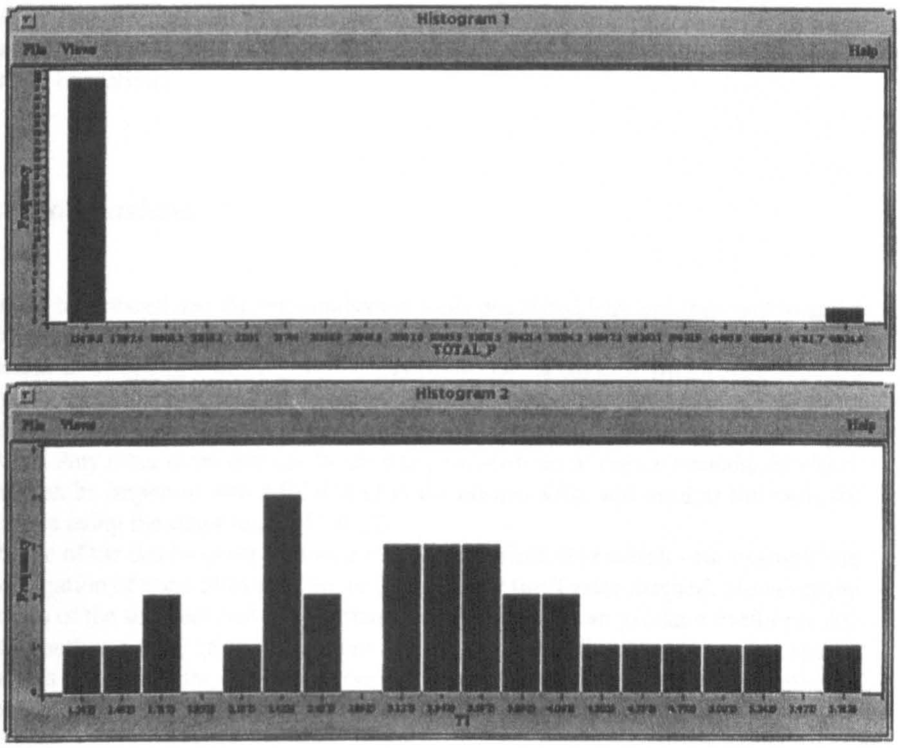


Fig 5.5. Graphical Assessment of Regionalisation Results.
The upper graph is a histogram of the total population in each region, the lower one is a histogram of the interquartile range of Townsend deprivation index values for zones making up each region.

different runs.

Alternatively, it is possible to move on to step 4 in Figure 5.1, and use a series of tools which will split large regions or merge small ones in order to improve the equality or homogeneity of a set of regions. In performing these split and merge operations, constraints can also be built in to avoid making the overall objective function much poorer - in trying to equalise population for example a threshold can be set on how much this is allowed to reduce the performance in terms of homogeneity.

These procedures can result in a change in the number of regions. If the best result seems to be to split one large region no attempt is currently made to merge two others to compensate. It is felt that this is justified by the fact the number of regions is often a relatively arbitrary decision, possibly guided by some idea of what the minimum population should be in the regions for the calculation of reliable rates, rather than by some fixed notion of a number of regions.

The final stage of the process (stage 5 on Figure 5.1) is to create a new set of boundaries by removing the boundaries between zones in the same region and merging their attribute values together, which in a GIS such as ARC/INFO is a very simple operation.

5.6 Conclusions

It must be stressed that the regionalisation tools described here are intended to assist with the interactive spatial analysis of data. Speed has therefore been one of the key criteria in selecting methods, which has ruled out techniques which may produce superior results. However, the design of SAGE, as a system linked to a GIS package means that the user is not restricted to the regionalisation methods provided within SAGE. Any other technique can be used to produce a set of region boundaries which can then be imported into ARC/INFO in the normal way, and used as the basis for analysis using the other tools of SAGE.

Some of the details of the system are still the subject of research - for example the investigation of ways of identifying nodal zones for the Taylor method. However the success of the software will largely stand upon whether it can produce useful results, and whether SAGE provides an environment suitable for the interactive spatial analysis of health data, which will only become apparent after it has been tested by others, which is the next stage of this research project.

Acknowledgments
The authors acknowledge receipt of ESRC grant R000234470 which has made the work reported here possible.

References

Anderberg M.R. 1973. *Cluster analysis for applications.* Academic Press. New York

Anselin L. 1995. Local indicators of spatial association - LISA. *Geographical Analysis.* 27:93-115.

Barkley D.L., Henry M.S.. Bao S. and Brooks K.R. 1995. How functional are economic areas? Tests for intra-regional spatial association using spatial data analysis. *Papers in Regional Science* 74:297-316.

Bithell J.F. 1990. An application of density estimation to geographical epidemiology. *Statistics in Medicine* 9:691-701.

Clayton D. and Kaldor J. 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**:671-681.

Cliff, Haggett, Ord, Bassett and Davies. 1975. *Elements of spatial structure*. Cambridge University Press.

Cliff A.D. and Ord J.K. 1981. *Spatial processes : models and applications*. Pion, London.

Cressie N. 1984. Towards resistant geostatistics. In G.Verly, et al (eds) *Geo-statistics for natural resources characterization*, 21-44. Reidel, Dordrecht.

Cressie N.A.C. 1991. *Statistics for spatial analysis*. John Wiley and Sons, New York.

Elliott P., Cuzick J., English D. and Stern R. 1992. *Geographical and environmental epidemiology : methods for small area studies*. Oxford University Press.

Fotheringham A.S. and Wong D.W.S. 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23:1025-1044.

Gatrell A.C. 1989. On the spatial representation and accuracy of address-based data in the United Kingdom. *International Journal of Geographical Information Systems* **3**:335-48.

Getis A. and Ord J.K. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* **24**:75-95.

Haining R.P., Wise S.M. and Blake M. 1994. Constructing regions for small area analysis: material deprivation and colorectal cancer. *Journal of Public Health Medicine* **16**:429-438.

Haining R.P, Wise S.M. and Ma J. 1996. The design of a software system for interactive spatial statistical analysis linked to a GIS. *Computational Statistics* (in press).

Haining R.P. 1996. *Spatial statistics and the analysis of health data*. Paper presented to the GISDATA workshop on GIS and health. Helsinki, June 1996.

Haslett J., Wills G. and Unwin A.R. 1990. SPIDER - an interactive statistical tool for the analysis of spatially distributed data. *International Journal of Geographical Information Systems* **4**:285-296.

Haslett J., Bradley R., Craig P.S.,Wills G. and Unwin A.R. 1991. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *American Statistician*, **45**:234-42.

Horn M.E.T. 1995. Solution techniques for large regional partitioning problems. *Geographical Analysis* **27**:230-248.

Lankford P.M. 1969. Regionalisation: Theory and Alternative Algorithms. *Geographical Analysis* **1**:196-212.

McQueen J. 1967. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. **1**:281-297.

Oliver M.A. and Webster R. 1989. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology* **21**:15-35.

Openshaw S. 1978. An optimal zoning approach to the study of spatially aggregated data. in Masser I. and Brown P.J. (eds) *Spatial Representation and Spatial Interaction*. Martinus Nijhoff,Leiden, 95-113.

Openshaw S. 1984. *The modifiable areal unit problem*. Concepts and Techniques in Modern Geography 38. GeoAbstracts, Norwich.

Openshaw S. and Rao L. 1994. Re-engineering 1991 census geography: serial and parallel algorithms for unconstrained zone design.

Ord J.K. and Getis A. 1995. Local spatial autocorrelation statistics : distributional issues and an application. *Geographical Analysis* **27**:286-306.

Raper J., Rhind D.W and Shepherd J.W. 1990. *Postcodes: the new geography*. Longman, Harlow.

Rosenfeld A. and Kak A. 1982. *Digital picture processing*. Academic Press, London.

Sammons R. 1978. A simplistic approach to the redistricting problem. in Masser I. and Brown P.J. (eds) *Spatial Reprsentation and Spatial Interaction*. Martinus Nijhoff, Leiden71-94.

Semple R.K. and Green M.B. 1984. Classification in Human Geography. in G.L.Gaile and C.J.Wilmott (eds) *Spatial statistics and models*, Reidel, Dordrecht. 55-79.

Spath H. 1980. *Cluster Analysis Algorithms*. John Wiley and Sons. New York.

Taylor P.J. 1969. The location variable in taxonomy. *Geographical Analysis* 1:181-195.

The Statistician (1998)
47, *Part* 3, *pp.* 457–469

Exploratory spatial data analysis in a geographic information system environment

Robert Haining†, Stephen Wise and Jingsheng Ma
University of Sheffield, UK

[Received June 1996. Revised January 1998]

Summary. The paper describes SAGE, a software system that can undertake exploratory spatial data analysis (ESDA) held in the ARC/INFO geographical information system. The aims of ESDA are described and a simple data model is defined associating the elements of 'rough' and 'smooth' with different attribute properties. The distinction is drawn between global and local statistics. SAGE's region building and adjacency matrix modules are described. These allow the user to evaluate the sensitivity of results to the choice of areal partition and measure of interarea adjacency. A range of ESDA techniques are described and examples given. The interaction between the table, map and graph drawing windows in SAGE is illustrated together with the range of data queries that can be implemented based on attribute values and locational criteria. The paper concludes with a brief assessment of the contribution of SAGE to the development of spatial data analysis.

Keywords: Adjacency matrix; Area data; Brushing; Local and global statistics; Regionalization

1. Introduction

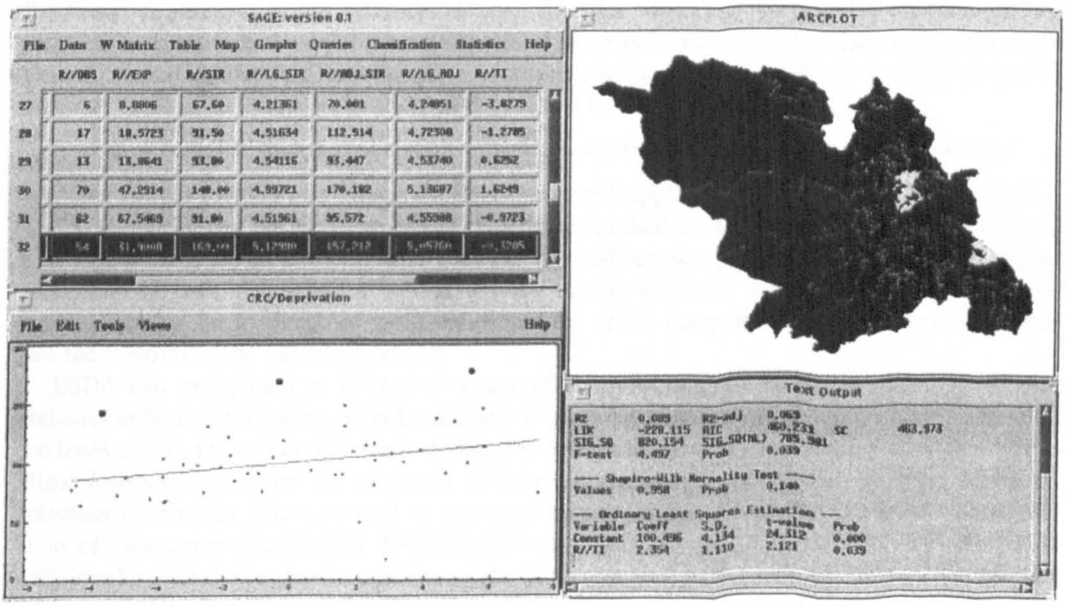
Exploratory spatial data analysis (ESDA) is the extension of exploratory data analysis (EDA) to the problem of detecting spatial properties of data sets where, for each attribute value, there is a locational datum. This locational datum references the point or the area to which the attribute refers. Examples include rainfall measurements taken at a number of sample sites in a region or mortality rates for a set of wards or counties. EDA is a collection of descriptive techniques for detecting patterns in data, identifying unusual or interesting features (including detecting errors), distinguishing accidental from important features and for formulating hypotheses from data. EDA may also be employed after data modelling to assess aspects of model fit. The set of exploratory techniques combines techniques that are visual (including charts, graphs and figures) and numerical but statistically robust. Exploratory techniques generally stay 'close' to the original data, meaning that they use relatively simple intuitive manipulations of the data.

ESDA is an extension of EDA to detect spatial properties of data: to detect spatial patterns in data, to formulate hypotheses which are based on, or which are about, the geography of the data and to assess spatial models. The class of techniques that are used is, as in EDA, visual and robust. However, it is important to be able to link numerical and graphical procedures with the map and to be able to answer questions such as 'where are those cases on the map?', 'where do attribute values from this part of the map lie in the data summary?' or 'which areas on the map lie in this subregion of the map and meet specified attribute criteria?'. The map is an essential additional tool for exploring spatial data.

†*Address for correspondence*: Department of Geography and Sheffield Centre for Geographic Information and Spatial Analysis, University of Sheffield, Sheffield, S10 2TN, UK.
E-mail: R.Haining@sheffield.ac.uk

0039–0526/98/47457

This paper reports on the development of a software system for carrying out ESDA linked to the ARC/INFO geographical information system (GIS). Because there are many types of spatial data we focus only on what Cressie (1991), pages 7–10, called 'lattice' data, a term which includes the general case where the regions that partition the map may be irregular in shape. Here the attribute values must be standardized in some way so that values in different regions are comparable. The denominator is usually a measure of area in the case of a spatially continuous variable like crop yields or a count of households or individuals (for example) in the case of a spatially discrete variable like population. The wish to link spatial data analysis to the GIS is because the GIS has become widely used for geographical data management and cartographic modelling and because it has functionality that facilitates the development of many spatial analysis techniques including spatial data analysis. Recent papers have considered the types of analytical capabilities that are most suited to a GIS environment (Goodchild *et al.*, 1992; Fotheringham and Charlton, 1994). The arguments and illustrations in this paper are drawn from one such project that has led to the development of the SAGE package (Haining *et al.*, 1996). The development of SAGE has been based on the assumption that even in the typical GIS environment which is characterized by very large data sets there is still an important role for simple and familiar statistical methods. For an alternative view see, for example, Openshaw (1994). SAGE has also been built using wherever possible existing, well-tested, software. All the processes of data input, data management and data analysis are provided within the GIS without the need to export or import data files during the analysis. Fig. 1 shows SAGE with all the four types of window open: the table window (which has limited spreadsheet capability) displaying the current set of data and any new variables created during a session, the map window, a graph window and the text output window that returns statistical output such as model parameters. Note that the linked windows facility is being used with selected data cases identified in the table, map and graph windows. More than one graph window can be opened and linked with the other windows.


SAGE: version 0.1
File Data W Matrix Table Map Graphs Queries Classification Statistics Help
CRC/Deprivation
File Edit Tools Views Help
ARCPLOT
Text Output

	R//OBS	R//EXP	R//SIR	R//LG_SIR	R//ADJ_SIR	R//LG_ADJ	R//TI
27	6	8,8806	67,60	4,21361	70,001	4,24051	-3,8279
28	17	18,5723	91,50	4,51634	112,514	4,72308	-1,2785
29	13	13,8641	93,80	4,54116	93,447	4,53740	0,6252
30	70	47,2914	148,00	4,99721	170,102	5,13607	1,6249
31	62	67,5469	91,80	4,51961	95,572	4,55988	-0,9723
32	54	31,9000	169,[illegible]	5,12990	157,212	5,05266	-0,3205

```
R2          0,089      R2-adj       0,063
LIK         -228,115   AIC          460,231   SC         463,973
SIG_SQ      820,154    SIG_SQ(ML)   785,981
F-test      4,497      Prob         0,039

---- Shapiro-Wilk Normality Test ----
Values      0,958      Prob         0,140

--- Ordinary Least Squares Estimation ---
Variable  Coeff     S.D.      t-value   Prob
Constant  100,496   4,134     24,312    0,000
R//TI     2,354     1,110     2,121     0,039
```

Fig. 1. SAGE: displaying the four types of window and the linked windows facility

The next section defines a data model for the patterns which ESDA may be used to detect, making the distinction between 'whole map' and 'local' statistics. Section 3 considers the importance of the regional partition and the definition of adjacency in ESDA. Section 4 describes ESDA techniques in SAGE for detecting properties of a single mapped attribute. Section 5 comments briefly on the availability of other software packages for implementing spatial data analysis.

2. A data model for spatial pattern in a single attribute data set

One simple data model for EDA distinguishes between the 'smooth' component of the data which derives from some summary of the data and the 'rough' component which is the residual (Tukey, 1977). Thus

$$\text{data} = \text{smooth} + \text{rough}.$$

A spatial data set comprises for each case an attribute value and its locational identifier. If we disregard the locational identifier initially this leads to the association of smooth and rough with just the (non-spatial) attribute values of the data set. Such non-spatial smooth properties include the central tendency of the distribution measured by the median, the dispersion of the distribution measured by the interquartile range and the shape of the distribution depicted by box plots or histograms. The non-spatial rough property is the difference between the data value and the smooth value and outliers are defined as cases with particularly high levels of rough. Outliers are identified as data values that are more than a certain distance above or below the upper or lower quartile respectively. With some modification this decomposition can also be adapted to spatial data.

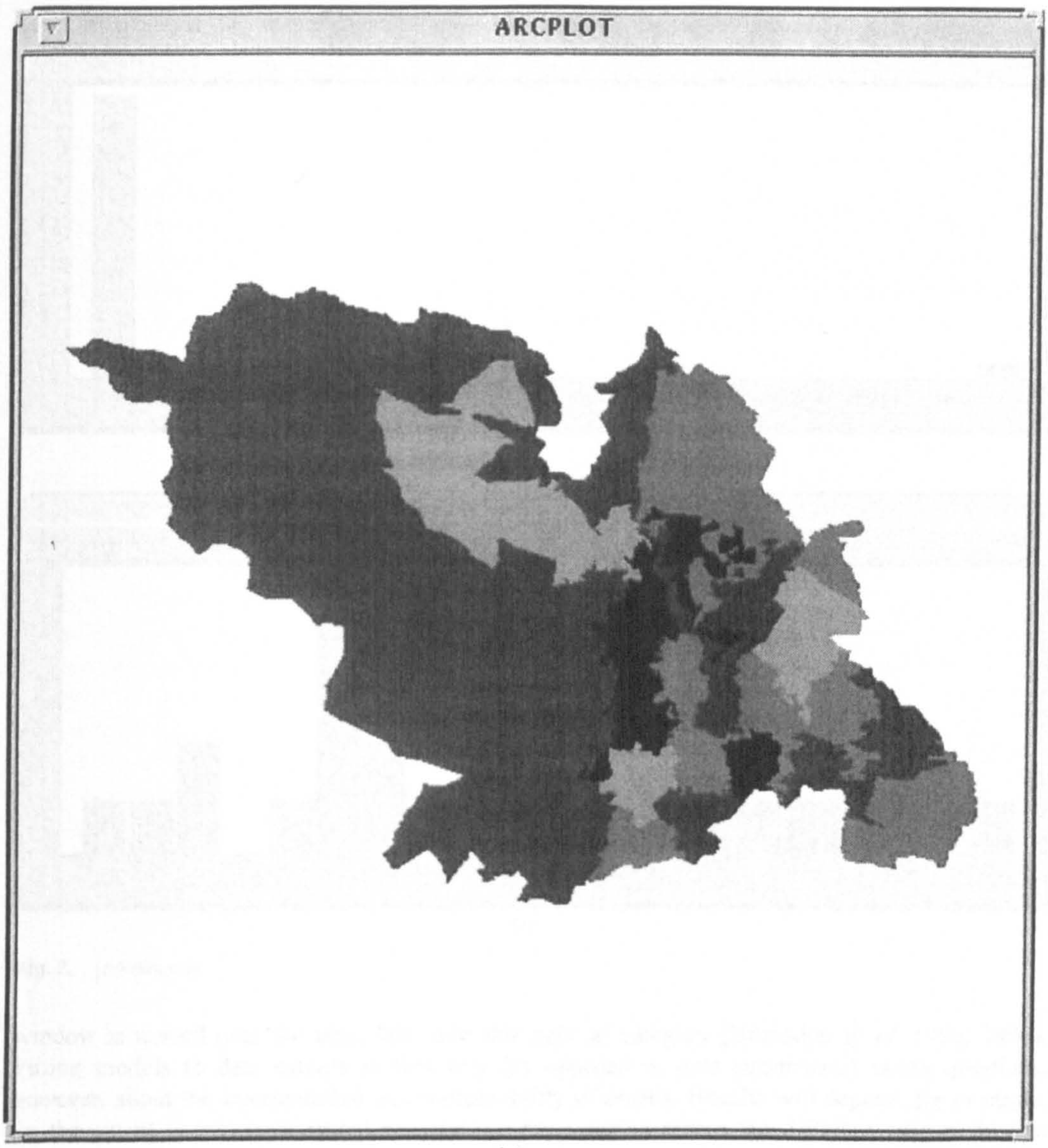
When the locational identifier is included smooth and rough properties need to be defined in terms of *where* on the map the cases are found. Smooth spatial properties include spatial trends, spatial autocorrelation (the propensity, in the case of positive autocorrelation, for similar values to be found together across the whole map) and spatial concentrations (the propensity for large values to be found together and/or low values to be found together across the whole map). Again the rough component is the distance between the data value and the smooth component. The residuals may show evidence of localized patterns of spatial autocorrelation and spatial concentration. A spatial outlier is a case where the attribute value is very different from neighbouring attribute values. This suggests a modified data model for ESDA of the form

$$\text{data} = (\text{trend} + \text{spatial covariation} + \text{concentration}) + (\text{residuals and spatial outliers}).$$

A model similar to this model, though differing in detail, applies to data ordered in time as well.

ESDA techniques fall into two broad categories: 'global' or whole map statistics, which process all the cases for an attribute, and 'focused' or local statistics, which process spatially defined subsets of the data, one subset at a time, and which may involve a sweep through all the defined subsets looking for evidence of localized properties of the mapped data—or the residuals after, say, the removal of the trend component.

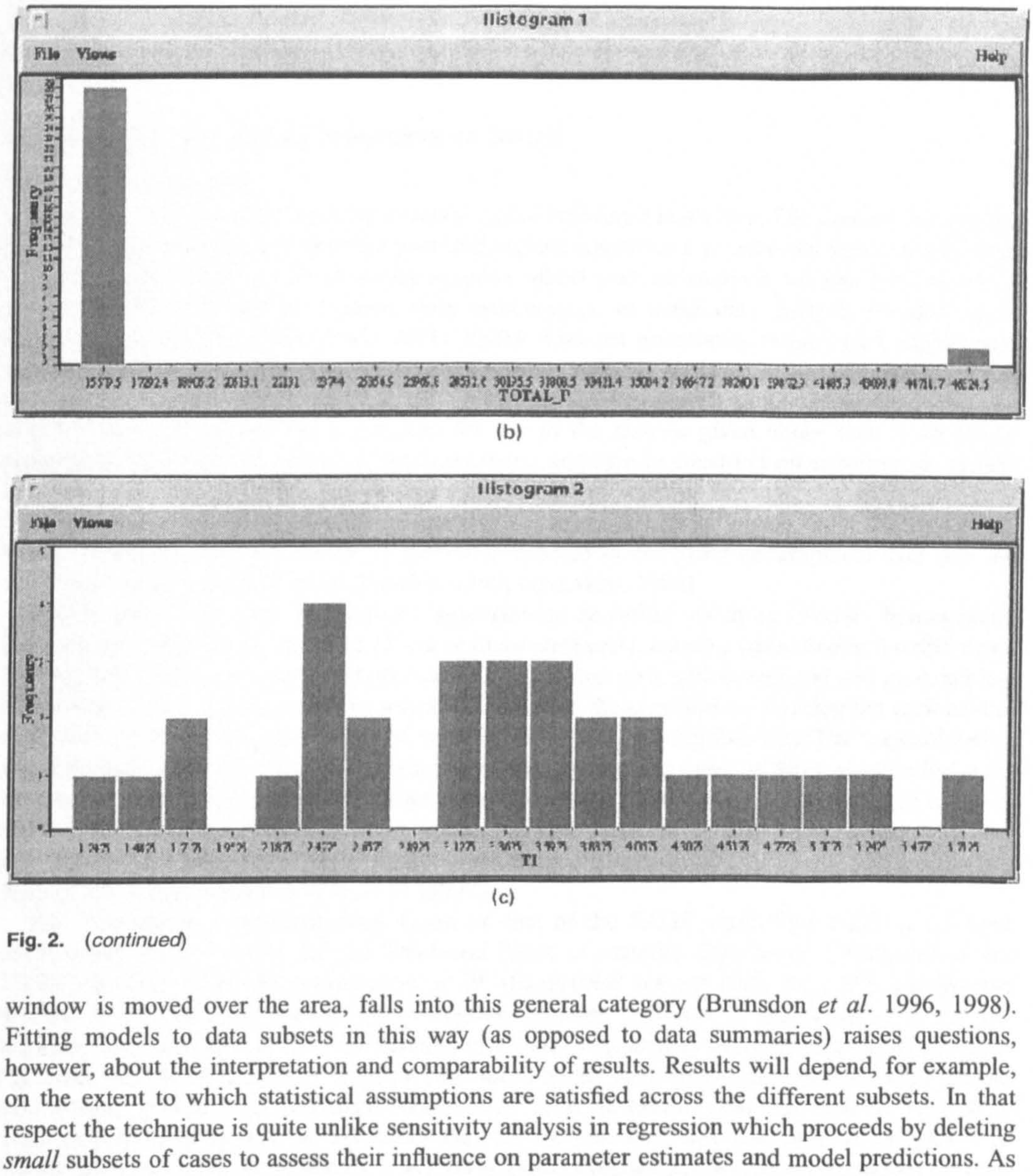
ESDA can be applied to defined subareas of the map, e.g. by applying global or focused statistics only to cases falling within a defined window. Different methods can be distinguished on the basis of how the subset of cases is defined. The majority of systems that are currently available allow *brushing*, in which the selection of areas is made interactively on the map. Once the selection is complete, the graphical or statistical results for this subset are displayed. This is the style of interaction provided by SAGE, where the brushing can take place in any one of the cartographic, tabular or graphical views of the data, resulting in the identification of the selected cases in all the other views. Subsequent calculations (e.g. of summary statistics) can be restricted to the currently selected subset of cases.



(a)

Fig. 2. (a) Regionalization of Sheffield, aggregating enumeration districts into 29 regions on the basis of deprivation scores, (b) histogram window showing the population sizes of the 29 regions (equality criterion) and (c) histogram of the interquartile ranges of the enumeration district level deprivation scores for the 29 new regions (homogeneity criterion)

Craig *et al.* (1989) suggested and implemented an extension to this idea in which the calculation of statistical results would be done as the brushing was done. If the selection of areas was made using a fixed window (e.g. a circle), then as this was moved across the map the statistics would be recalculated and graphics redisplayed, allowing the user to explore differences across the region. The technique known in the geographical literature as *geographically weighted regression*, in which a regression model is fitted to data in a defined fixed window and then refitted as the



(b)

(c)

Fig. 2. (*continued*)

window is moved over the area, falls into this general category (Brunsdon *et al.* 1996, 1998). Fitting models to data subsets in this way (as opposed to data summaries) raises questions, however, about the interpretation and comparability of results. Results will depend, for example, on the extent to which statistical assumptions are satisfied across the different subsets. In that respect the technique is quite unlike sensitivity analysis in regression which proceeds by deleting *small* subsets of cases to assess their influence on parameter estimates and model predictions. As far as we are aware, this general form of spatial data analysis has not yet been implemented in software linked to any GIS although it is being implemented in other computing environments (see Bivand (1998), Dykes (1998) and Unwin and Hofmann (1997)).

The model for map pattern described in this section is not formal and the distinction between what is trend, spatial autocorrelation or spatial clustering is deliberately not well defined. The techniques that will be discussed may be used to identify attribute properties but cannot be said to estimate the various components of map pattern.

Any spatial analysis based on area data must recognize that results are dependent on the form of the regional partition. One of the elements of SAGE is a simple region building module that is appropriate for ESDA. Spatial properties may also depend on definitions of the pseudo-ordering of the regions. SAGE allows for alternative definitions of adjacency between regions. We discuss these now.

3. Handling the spatial framework in SAGE

3.1. Region building

Spatial data analysis often starts from small spatial building-blocks (e.g. UK census enumeration districts), aggregating these until the resulting regions constitute a satisfactory basis for statistical analysis. Aggregation may be necessary to create robust rates for analysis, to reduce the effects of any suspected locational or attribute data inaccuracies, to make data analysis tractable or to facilitate visualization (Wise *et al.*, 1997). ESDA does not necessarily require such aggregation and in some spatial data sets (e.g. analysing electoral outcomes by constituency) the spatial unit is naturally defined and relevant both to the underlying process as well as to subsequent interpretation. However, if aggregation is required for any of the reasons given above then it should be possible to aggregate according to specified criteria and then to construct other similar or equally plausible aggregations fairly quickly and easily to assess whether findings change significantly. This amounts to allowing the user to examine for possible effects arising from the modifiable nature of the areal units, a matter of particular concern in analysing geographical data and one which has a long history of study (Kendall, 1939; Openshaw, 1984).

SAGE allows the user to construct aggregations according to three criteria: homogeneity (minimizing within-group variance of one or more attributes), equality (minimizing the difference between the total value of an attribute, such as population size, across regions) and geographical compactness. The importance to be attached to each of these criteria in forming the regionalization can be adjusted through the use of weights within an objective function. The regionalization is a *k*-means-based classification that allows the user to start from one of many initial allocations of zones to regions and then allows swaps at the boundaries. Swaps may be allowed even if one or two of the individual criteria become worse, provided that the overall function improves and provided that those that do become worse do not exceed a user-defined threshold. There is further description of this module in Wise *et al.* (1997).

Fig. 2(a) shows a regionalization based on one of the SAGE algorithms building up from enumeration district scores for the Townsend index of material deprivation (Townsend *et al.*, 1988). Fig. 2(a) shows the construction of 29 'deprivation' regions from the 1159 enumeration districts in the Sheffield region. The histograms provide evidence of the extent to which the algorithm has been able to meet the equality criterion (measured by regional population counts—Fig. 2(b)) and the homogeneity criterion (measured by the intra-region interquartile range for the enumeration district level Townsend scores—Fig. 2(c)). It appears that the equality criterion is quite well satisfied except for two areas that are far too large and will need to be split. The homogeneity criterion shows that there is intra-regional variation in deprivation. However, the new regionalization is still a considerable improvement over other partitions at the same scale such as wards (there are 29 in Sheffield) in terms of demarcating areas of similar deprivation. (For a discussion of this see Haining *et al.* (1994).)

3.2. Adjacency measures

Many spatial analysis techniques require the analyst to define the set of neighbours of each region in the map partition and to define the relative weights to be attached to each paired neighbour. Unlike time, geographic space has no natural order and with irregular regional units there may be a need to explore the sensitivity of results to many alternative definitions of neighbourhood. As it is loaded into memory SAGE automatically creates a definition and creates the measures needed for two other neighbourhood or adjacency matrices (**W**). These are derived from the stored adjacency relationships held by ARC/INFO in which each line segment or arc has a direction and a list is maintained of the polygons (regions) that lie on the left-hand and right-hand sides of each arc (Ding and Fotheringham, 1992). The adjacency matrix automatically generated by SAGE is a simple binary adjacency matrix determined by whether regions share a common boundary (1) or not (0). Two other matrices are constructed using intercentroid distances and the length of the shared common boundary. These can be converted by the user into an appropriate adjacency matrix (Haining (1993), pages 73–74). There is a further module in SAGE that allows the user to create other matrices or to modify the automatically generated matrices.

4. Exploratory spatial data analysis for identifying properties of a univariate data set

As illustrated by the following, EDA summaries and graphics that do not depend on any spatial referencing have important roles to play in ESDA.

(a) *Median*—ESDA query: which areas have attribute values above (or below) the median? Do they show any evidence of pattern?

(b) *Quartiles*—ESDA query: which areas lie in the upper (or lower) quartile? If F_U and F_L denote the upper and lower quartiles then which cases have attribute values that are greater than $F_U + 1.5(F_U - F_L)$ or less than $F_L - 1.5(F_U - F_L)$ and may be defined as outliers?

(c) *Box plots*—ESDA query: where do cases that lie in particular areas of the box plot occur on the map? Where are the outlier cases located on the map? The two previous queries can be subsumed within this query.

(d) *Histograms*—ESDA query: where do cases that relate to particular bars of the histogram occur on the map?

Fig. 3 shows a box plot of standardized incidence rates of a form of cancer in Sheffield displayed in the graphics window and all the cases lying above the median are 'brushed' and highlighted in the map window. Note that most of the areas with high rates are to be found in the eastern and central area of Sheffield which includes many of the more deprived parts of the city.

ESDA techniques for identifying spatial properties of the attribute data usually require a definition of adjacency. Here we define a general $n \times n$ (n corresponding to the number of regions on the map) adjacency matrix **W** with non-negative elements $\{w_{ij}\}$ where the subscripts reference regions i and j and by definition we set $w_{ii} = 0$. In some cases the row sums of **W** are standardized to a constant (usually 1). It is important to recognize that many of the techniques for ESDA can (and probably should) be replicated with different definitions of **W** for there is no natural ordering. In addition it is often appropriate to replicate the analysis by taking a sequence of distances or 'lag' (step) orders on the graph of regions to detect properties at different spatial distances.

Where the map consists of many small areas a simple smoothing method may reveal general patterns (such as a trend) that are not apparent from the mosaic of values. Kernel estimation, in its simplest form, involves passing the equivalent of a moving average or 'local mean' filter across the surface:

$$\mathrm{MA}_i = \left(X_i + \sum_j w_{ij} X_j\right) \bigg/ \left(1 + \sum_j w_{ij}\right)$$

where the weight w_{ij} is 1 if region j shares a common boundary with region i and is 0 otherwise

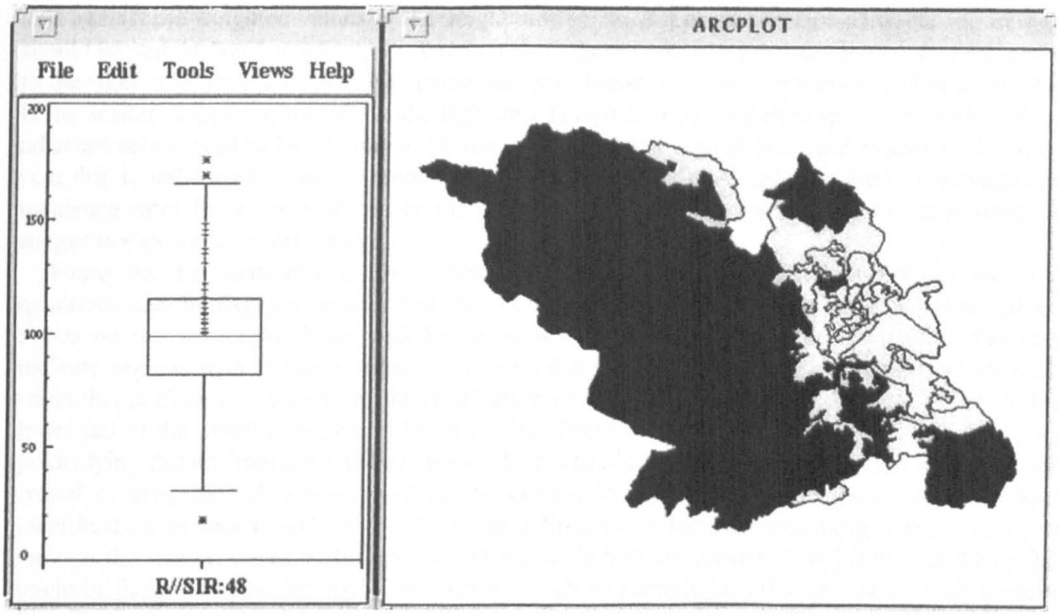


Fig. 3. Box plot of standardized incidence rates for a cancer by regions of Sheffield linked to a map of the regions and highlighting all cases with higher than expected rates (rates greater than 100)

($w_{ii} = 0$). Other weights and constructions for kernel estimation can be used which are also implemented in SAGE. A slight modification to this method for smoothing and hence detecting trends in spatial data would be to replace the value in a region i with the median value from the set that includes X_i and the values in the adjacent regions—a moving median or 'local median' filter (MM_i). This would still further reduce the effect of extreme values on the smoothed surface. The smoothed component of the map can then be extracted from the map by computing $X_i - \mathrm{MA}_i$ or $X_i - \mathrm{MM}_i$. Using the median smoother rather than the mean results in areas with particularly high rates standing out even more strongly as areas with a large element of rough. This last stage, using MA_i, is similar to the process of smoothing by spatial differencing described by Cliff and Ord (1981), p. 192, provided that the weights are defined in the same way. The principal distinction lies in whether the value at i is or is not included in the term MA_i.

Where it is thought that attribute values might decrease (or increase) away from a specific area such as the centre of a city then a transect of values might be helpful and can be implemented in SAGE. SAGE also allows the construction of a series of 'lagged' box plots in the graphics window where the first box plot is the first-order neighbours of the selected region, the second box plot is the second-order neighbours and so on (Haining (1993), p. 224). This second method is only likely to be useful provided that all the areas are of similar size and shape but in those cases can indicate the presence of trend and dispersal around the trend.

Whole map statistical tests have been developed for testing for global spatial autocorrelation (e.g. Moran's I) and spatial concentration (e.g. the Getis–Ord G- and G^*-statistics) (Cliff and Ord, 1981; Getis and Ord, 1992) and these techniques are available in SAGE. However, these tests are not based on robust estimators (of the centre of the distribution of values); nor could they be described as exploratory. Values of the statistic do not have any intuitive interpretation. They are really more appropriate for confirmatory work. A simple ESDA tool in SAGE that can explore for these properties is based on the scatterplot. Values of an attribute (X_i) are plotted on the vertical axis against the weighted values of the neighbours ($\Sigma_j w_{ij}X_j$) on the horizontal where the weights should be standardized to sum to 1. A scatterplot where there is a general upward sloping scatter to the right is indicative of positive spatial autocorrelation, i.e. adjacent values tend to be similar. If the scatter slopes downwards to the right this is indicative of negative spatial autocorrelation; adjacent values tend to be dissimilar. (If the scatter is linear and shows little evidence of dispersion this is indicative of spatial trend). Fig. 4 illustrates the scatterplot applied to standardized incidence rates for a form of cancer for Sheffield. There is a general trend in the scatterplot, suggesting spatial autocorrelation.

Points on the scatterplot in the extreme parts of the top right-hand or bottom left-hand quadrants may be flagging regions that show a concentration or clustering of high or low values. Points on the scatterplot lying well below or well above any part of the general scatter may indicate regions with attribute values that make them spatial outliers. For example, an attribute value that is close to the mean of the distribution of values, encircled by values at or close to the lower tail of the distribution, could be an outlier. There are no very clear cases in Fig. 4, but six points lying distant from the line have been selected to illustrate that, as the histogram shows, such spatial or geographical outliers need not be outliers in the statistical distributional sense. This identification of spatial outliers can be made a little more formal by running a regression line through the scatter. Cases with standardized residuals that are greater than 3.0 or less than −3.0 might be flagged as possible spatial outliers although this simple test, if based on the least squares fit, will tend to overstate the number and size of outliers (see Haining (1993), pages 214–215). As noted earlier it is possible to brush any part of the scatterplot to identify where the regions are on the map and the corresponding values are also highlighted on the spreadsheet.

Local statistics, available in SAGE, can be used to assess the presence of localized spatial autocorrelation or concentration. The Getis–Ord (G_i^*-) statistic for detecting localized con-

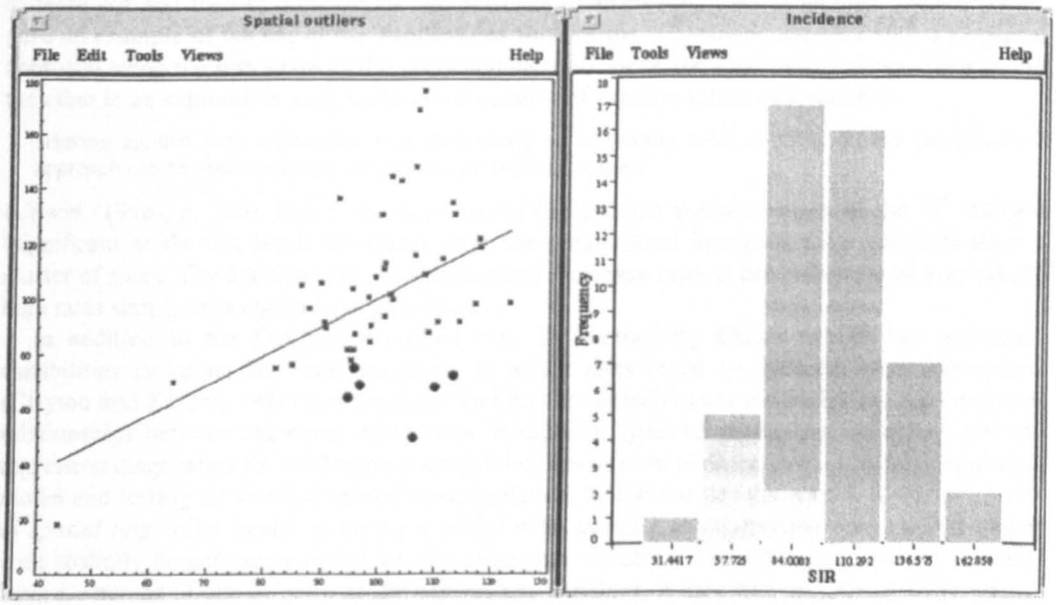


Fig. 4. Scatterplot of standardized incidence rates of a cancer against the average of the rates in adjacent regions: cases with low rates but surrounded by regions the average of whose rates is at or near the expected rate (100) are highlighted; these cases are also highlighted in the histogram

centrations (or localized clusters) in an attribute which is positive valued with a natural origin is defined:

$$G_i^* = \sum_j w_{ij}^* X_j \Big/ \sum_j X_j$$

where w_{ij}^* is the entry in the weights matrix $\mathbf{W}^*$ where $w_{ii}^* > 0$ and the statistic is computed for each region in turn (Getis and Ord, 1992). A large value of G_i^* signals a clustering of high values around region i; a small value signals a clustering of low values around i. The local Moran statistic is defined (Anselin, 1995):

$$I_i = x_i \sum_j w_{ij} x_j$$

where x_i and x_j signify deviations from the mean. A large positive value of I_i signals a local set of similar values in the neighbourhood of region i; a large negative value signals a local set of dissimilar values at i.

The G_i^*-values are comparable (same mean and variance) if the weights matrix $\mathbf{W}^*$ is standardized so that row sums equal a constant. In this case the set of n regional values for each of these statistics could be rank ordered to signal where localized clusters might exist on the map, or treated as a distribution and examined (as suggested above) for extreme values which can then be brushed to identify the cases on the map. Formal tests of significance are available in SAGE and the standardized form of the statistic will be required to allow for non-constant means and variances if $\mathbf{W}^*$ has not been standardized. For the local Moran statistic, standardization of the statistic is always required since, although the expected values are constant if $\mathbf{W}^*$ is standardized so that row sums equal a constant, the variances are not.

There is often an advantage to simultaneously using spatial and non-spatial statistical methods to tease out and then to demonstrate the presence of interesting data properties. Unwin (1996) gave an example of the use of the standardized form of the G_i^*-statistic together with a graphical approach using the histogram of the original data to illustrate the way that each can complement the other in an exploratory analysis to locate clusters of extreme values of a variable:

> 'Having applied both approaches it is then easier to understand what is going on and the graphical approach can be used to present and explain the results to others'

(Unwin (1996), p. 396). Fig. 5 shows a map of the extreme positive values of the G_i^*-statistic (significant at the 5% level) computed from the standardized incidence rates and indicating a cluster of cases. The histogram of the standardized incidence rates is not indicative of particularly high rates simply on a region-by-region basis.

In addition to the facilities described here for performing ESDA, SAGE has additional capabilities including Bayesian smoothing to adjust rates based on different base populations (Clayton and Kaldor, 1987) and graphical and numerical techniques for exploring and analysing relationships between attributes. SAGE can fit different types of regression model and provide regression diagnostics for confirmatory spatial data analysis. In addition to the standard regression model and testing for residual spatial autocorrelation, SAGE enables the user to fit various types of *spatial* regression model including a model with spatially autocorrelated errors and a model with spatially lagged terms in the set of explanatory variables. The latter may include spatially lagged versions of one or more of the explanatory variables; it may also include spatially lagged values of the response variable among the set of explanatory variables. All these models are described in, for example, Haining (1993), pages 339–341. The description of these facilities will be the subject of Haining *et al.* (1998). Some of the ESDA facilities described above can also be

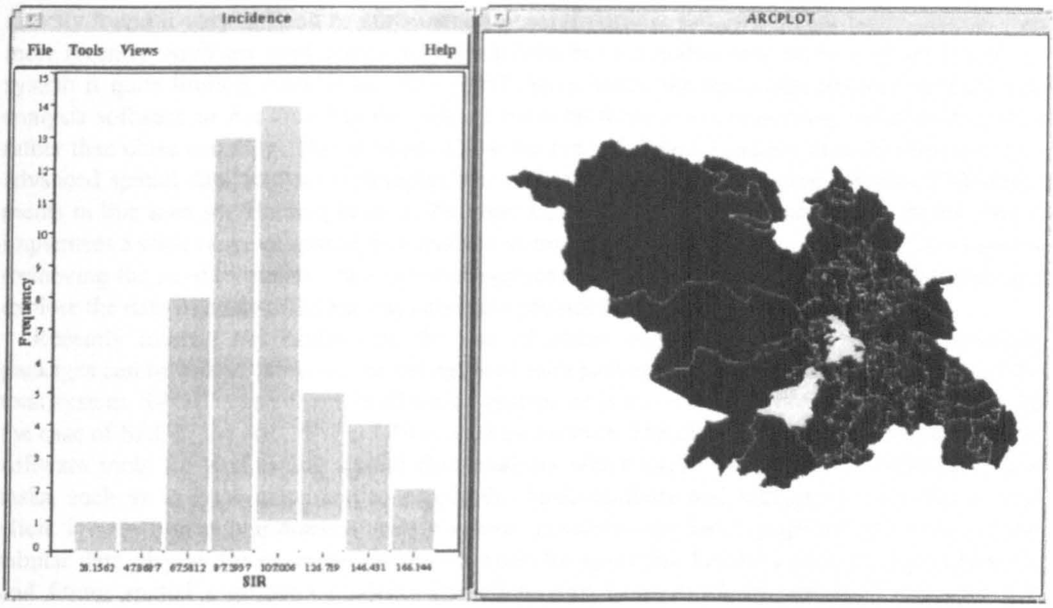


Fig. 5. Map showing regions that have high values of the Getis–Ord statistic (G_i^*) indicating a cluster of regions with high standardized incidence rates of a cancer: the histogram of the individual regional rates is shown in the adjacent window with the cases from the map window highlighted

employed for initial model assessment such as detecting autocorrelation in the residuals of a regression model. A full description of the range of statistical techniques that are available in SAGE is given in Ma *et al.* (1997).

5. Summarizing remarks: the contribution of SAGE to spatial data analysis

Any software for ESDA must provide at least the following capabilities in addition to those required for EDA:

(a) cartographic display capabilities;
(b) the ability to handle spatial data and to implement techniques that depend on the attributes of pairs of areas where the pairs are constructed on the basis of location criteria such as adjacency (Goodchild, 1987) (many ESDA applications require the ability to derive the contiguity information for a set of areas);
(c) the ability to link the tabular, graphical and cartographical views of the data.

It is arguable that since the analysis of area data is not usually based on a natural (right) geographic partition and no natural (right) definition of adjacency then it should be possible to assess the sensitivity of results to alternative, equally plausible, partitions and definitions of adjacency. This argument underlies the recommendation to make available the sort of capability discussed in Section 3.

Much of the work to provide computer-based spatial data analysis incorporating visualization techniques has been pursued independently of GISs (see for example the early innovative work by Haslett *et al.* (1990, 1991) and more recently Dykes (1996), Brunsdon and Charlton (1996) and the MANET software package developed by Unwin and Hofmann (1996) which includes the capability to handle missing values). Early attempts to add spatial analytical capabilities to GISs directly found it very difficult to implement a general purpose linked window facility. Developed more recently, ArcView implements linked windows but the spatial data analysis capability of the system is quite limited. Anselin and Bao (1997) have linked the SpaceStat advanced spatial data analysis software to ArcView but the linkage between them is via importing and exporting data rather than close coupling. This does not allow the kind of linked windows visualization based on advanced spatial data analysis techniques that is found in SAGE. For recent reviews of developments in this area see Haining *et al.* (1996) and Levine (1996). SAGE enables the statistician to implement a wide range of spatial data analysis techniques within a single computing environment (removing the need to transfer data between applications). SAGE allows the user to visualize and explore the data in many different ways and also provides tools for data modelling.

Recently interest has centred on the use of client-server computing, by which existing packages can be linked, allowing the strengths of each package to be exploited to the benefit of the total system. SAGE is one example of such a system, as is that described by Cook *et al.* (1996). In the case of SAGE, the ARC/INFO GIS is used as a server. The client is a purpose-written suite of software tools for performing spatial data analysis which calls ARC/INFO to perform certain tasks, such as to draw maps and to supply the basic attribute and contiguity information. The client is based on public domain code wherever possible—the basic graphical routines and the tabular data display for example—with new code for specialist facilities such as regionalization and fitting spatial regression models. The client also keeps track of which of the areas are currently selected and can therefore update all the open windows whenever the selection is changed (including the map window which is drawn by ARC/INFO).

This approach has produced a system which manages to combine the benefits of interactive linked windows with the capabilities of GIS software. SAGE extends ARC/INFO functionality very effectively into ESDA because ARC/INFO has the facility to integrate desk top computing and visualization both of which are fundamental to ESDA. This may be a useful route for adding other statistical analysis facilities to GIS packages.

6. Postscript

An introduction to SAGE, as well as a copy of the SAGE software package for downloading and details of its operating requirements, is available at the Sheffield Centre for Geographic Information and Spatial Analysis Web site:

`http://www.shef.ac.uk/~scgisa`

Acknowledgement

The authors acknowledge receipt of Economic and Social Research Council research grant R000234470 'Developing spatial statistical software for the analysis of area based health data linked to a GIS'.

References

Anselin, L. (1995) Local indicators of spatial association—LISA. *Geogr. Anal.*, 27, 93-115.

Anselin, L. and Bao, S. (1997) Exploratory spatial data analysis linking SpaceStat and Arc View. In *Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling and Neuro-computing* (eds M. Fischer and A. Getis), pp. 35-59. Berlin: Springer.

Bivand, R. S. (1998) Software and software design issues in the exploration of local dependence. *Statistician*, 47, 499-508.

Brunsdon, C. and Charlton, M. E. (1996) Developing an exploratory spatial analysis system in XLisp-Stat. In *Innovations in GIS 3* (ed. D. Parker), pp. 135-145. London: Taylor and Francis.

Brunsdon, C., Fotheringham, A. S and Charlton, M E (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr Anal.*, **28**, 281-298.

———(1998) Geographically weighted regression—modelling spatial non-stationarity. *Statistician*, **47**, 431-443.

Clayton, D. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics*, **43**, 671-681.

Cliff, A. D. and Ord, J. K. (1981) *Spatial Processes*. London: Pion.

Cook, D., Majure, J. J., Symanzik, J. and Cressie, N. (1996) Dynamic graphics in a GIS: exploring and analyzing multivariate spatial data using linked software. *Comput Statist.*, **11**, 467-480.

Craig, P., Haslett, J., Unwin, A. R. and Wills, G. (1989) Moving statistics—an extension of brushing for spatial data. In *Computing Science and Statistics: Proc 21st Symp Interface*, pp. 170-174. Alexandria: American Statistical Association.

Cressie, N. (1991) *Statistics for Spatial Data*. New York: Wiley.

Ding, Y. and Fotheringham, A. S. (1992) The integration of spatial analysis and GIS. *Comput. Environ. Urb. Syst.*, **16**, 3-19.

Dykes, J. (1996) Dynamic maps for spatial science: a unified approach to cartographic visualization. In *Innovations in GIS 3* (ed. D. Parker), pp. 177-187. London: Taylor and Francis.

———(1998) Cartographic visualization: exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv. *Statistician*, **47**, 485-497.

Fotheringham, A. S. and Charlton, M. E. (1994) GIS and exploratory spatial analysis: an overview of some research issues. *Geogr. Syst.*, **1**, 315-328.

Getis, A. and Ord, J. K. (1992) The analysis of spatial association by use of distance statistics. *Geogr. Anal.*, **24**, 189-206.

Goodchild, M. G. (1987) A spatial analytical perspective on geographical information systems. *Int. J. Geogr. Inform. Syst.*, **1**, 327-334.

Goodchild, M. G., Haining, R. P. and Wise, S. M. (1992) Integrating geographic information systems and spatial data analysis: problems and possibilities. *Int J Geogr Inform Syst.*, **16**, 407-424.

Haining, R. P. (1993) *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press.

Haining, R. P., Ma, J. and Wise, S. M. (1996) Design of a software system for interactive spatial statistical analysis linked to a GIS. *Comput. Statist.*, **11**, 449-466.

Haining, R. P., Wise, S. M. and Blake, M. (1994) Constructing regions for small area analysis: material deprivation and colorectal cancer. *J Publ. Hlth Med.*, **16**, 429-438.

Haining, R. P., Wise, S M. and Ma, J. (1998) SAGE—an interactive package for spatial statistical analysis in a GIS environment. Submitted to *Int J Geogr Syst.*

Haslett, J., Bradley, R., Craig, P. S., Wills, G. and Unwin, A. R. (1991) Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *Am. Statistn*, **45**, 234-242.

Haslett, J., Wills, G. and Unwin, A. R. (1990) SPIDER—an interactive statistical tool for the analysis of spatially distributed data. *Int. J. Geogr. Inform. Syst.*, **4**, 285-296.

Kendall, M. G. (1939) The geographical distribution of crop productivity in England. *J. R. Statist. Soc.*, **102**, 21-48.

Levine, N. (1996) Spatial statistics and GIS. *J Am Planng Ass.*, **62**, 381-391.

Ma, J., Haining, R. P. and Wise, S. M. (1997) *SAGE Users Guide*. (Available from http://www.shef.ac.uk/~scgisa, within the SAGEV01.tar file.)

Openshaw, S. (1984) The modifiable areal units problem. *Concepts and Techniques in Modern Geography*, no. 38. Norwich: GeoAbstracts.

———(1994) Two exploratory space-time-attribute pattern analysers relevant to GIS. In *Spatial Analysis and GIS* (eds S. Fotheringham and P. Rogerson), pp. 83-104. London: Taylor and Francis.

Townsend, P., Phillimore, P. and Beattie, A. (1988) *Health and Deprivation: Inequality and the North*. London: Croom Helm.

Tukey, J. W. (1977) *Exploratory Data Analysis*. Reading: Addison-Wesley.

Unwin, A. R. (1996) Exploratory spatial analysis and local statistics. *Comput. Statist.*, **11**, 387-400.

Unwin, A. R. and Hofmann, H. (1996) *MANET*. (Available from http://www1.math.uni-augsburg.de/Manet/.)

———(1997) New interactive graphics tools for exploratory analysis of spatial data. In *Innovations in GIS 5* (ed. S. Carver). London: Taylor and Francis.

Wise, S. M., Haining, R. P. and Ma, J. (1997) Regionalisation tools for the exploratory spatial analysis of health data. In *Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling and Neuro-computing* (eds M. Fischer and A. Getis), pp. 83-100. Berlin: Springer.

J Geograph Syst (2000) 2:257–286

Journal of Geographical Systems
© Springer-Verlag 2000

Designing and implementing software for spatial statistical analysis in a GIS environment

Robert Haining, Stephen Wise, Jingsheng Ma

Department of Geography, Sheffield Centre for Geographic Information and Spatial Analysis, University of Sheffield, UK (e-mail: R.Haining@Sheffield.ac.uk)

Received: 20 March 1999 / Accepted: 13 September 1999

Abstract. This paper provides a description of SAGE, a software package linked to the Arc/Info GIS that can be used to undertake spatial statistical analysis of area based data. The paper is written from the perspective of the user who wishes to undertake exploratory and confirmatory spatial data analysis. The paper discusses design aspects of the package and also the statistical analysis philosophy underlying its contents. The paper describes the statistical analyses which SAGE can perform and details on how it performs them together with some illustrative examples. Detail on visualisation aspects of SAGE are discussed in a separate paper (Haining et al. 2000). The wider contribution of the paper is to build on earlier developments in this area and identify the needs of software packages if they are to enable users to implement effective spatial statistical analysis.

Key words: Exploratory analysis, regionalisation, visualisation, autocorrelation, spatial regressions, client-server computing, brushing

JEL classification: C21, C51, C52, C87, C88

1 Introduction

Computer software to facilitate statistical analysis of spatial data has often been called for (Upton and Fingleton 1985; Anselin 1988; Haining 1990a; Levine 1996). Although commercial packages like MINITAB and SPSS contain facilities that either undertake or can be adapted to undertake certain forms of spatial statistical analysis (see for example Griffith 1988), they provide only limited support to the data analyst working with spatially referenced data. This is because they do not contain the necessary specialist techniques or diagnostic tools. The source of the problem is their limited capacity to handle information on the location of attributes and to model spatial dependence. Geographical Information Systems, that can manage locational data, contain only very rudimentary statistical facilities (see for example Goodchild et al. 1992).

A number of software modules or systems have been developed to implement specific spatial statistical analysis techniques (e.g. Stacas: Ding and Fotheringham 1992) and carry out exploratory spatial analysis (e.g. REGARD and MANET: Haslett Wills et al. 1990; Haslett Bradley et al. 1991; Unwin 1996). Others, like the spatial statistics modules in the programming and statistical language *S*-Plus, have been extended and directly interfaced to GIS to enable cartographic display (Levine 1996; Mathsoft 1999). SpaceStat provides a spatial statistical analysis package giving the user a comprehensive set of tools to undertake data description and modelling for area based data (Anselin 1992). It has recently been interfaced with ArcView to enable the user to map or graph output (Anselin and Bao 1997). For a recent review of progress in developing spatial statistical analysis software see Haining et al. (1996).

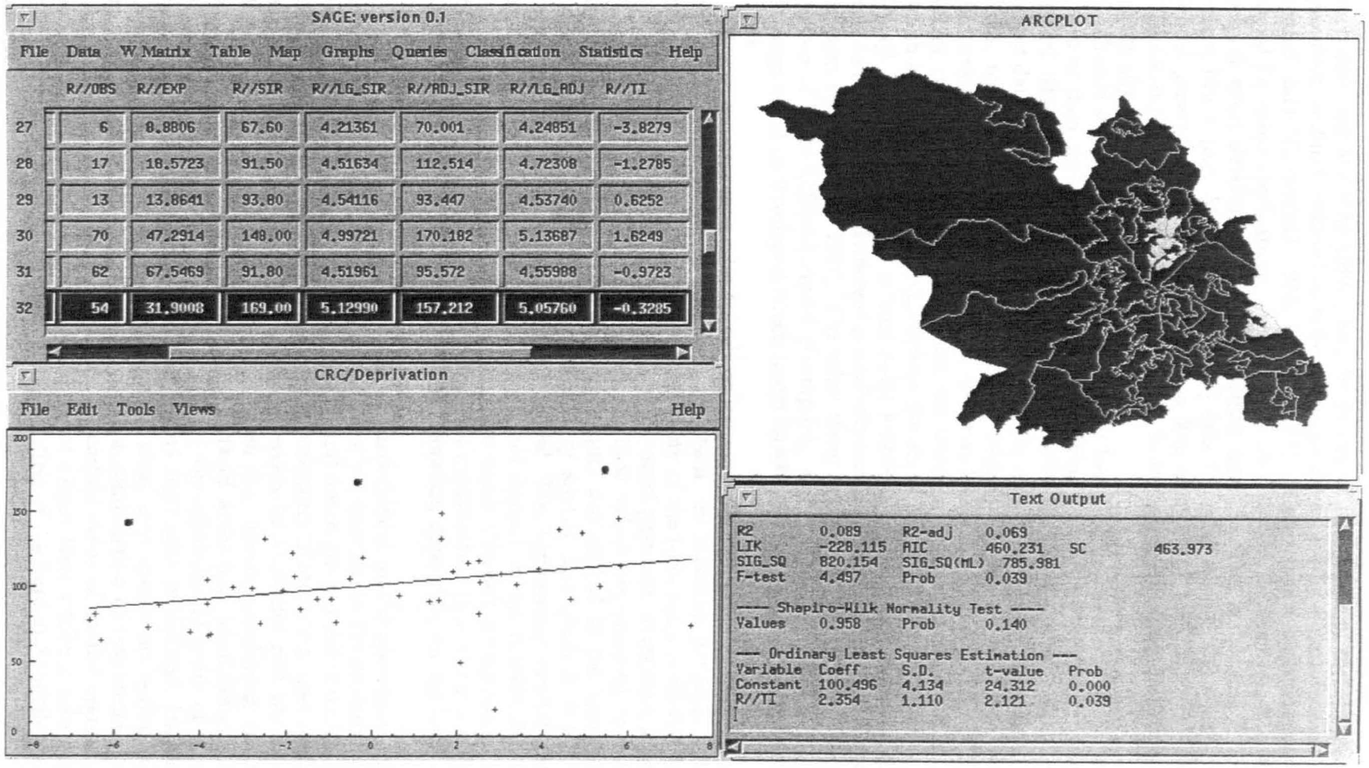
The purpose of this paper is to report the development of the software package SAGE (Spatial Analysis in a GIS Environment) as a contribution to the evolution of software packages for undertaking effective spatial statistical analysis. SAGE was developed for analysing area-based data within a GIS environment with Arc/Info operating as the server in a seamless client-server architecture. The reason for linking spatial statistical analysis software to a GIS was first, because of the importance of GIS as a general purpose platform for managing spatial data and second, because GIS has a number of facilities that are needed for spatial data analysis, to re-invent them seemed wasteful. The general capabilities of GIS that are of value to the development of a spatial statistical analysis package are its database management capabilities (including for example its ability to transform data between different scales and spatially aggregate data through facilities like the polygon dissolve operation); the locational and topological information it holds about the areal units or points to which attributes are attached; its cartographic display capabilities (Haining et al. 1996).

The seamless integration within SAGE of Arc/Info with graphics packages and specialist data analysis modules brings additional benefits. A crucial benefit lies in the fact that the various elements of the system (database, graphics windows and cartographic display) are linked so that for example attribute values in the database, including those generated during analysis, can be mapped or graphed without the need to transfer data files. File updating is handled automatically avoiding the danger of duplicate copies of a database getting out of step. Integration also allows the user to link cases in different windows so that a case or cases identified in a graph window, for example, will be automatically highlighted in the map, table and other graph windows.

The paper is organised as follows. Section 2 discusses the design criteria for SAGE and describes the range of facilities. Section 3 describes SAGE's confirmatory techniques, Sect. 4 describes the system design and operating limitations whilst Sect. 5 contains conclusions. Throughout the paper we illustrate the use of SAGE using data from a recent study of the geography of the uptake of breast cancer screening services in Sheffield. The Appendix describes how the reader can access SAGE software and documentation.

2 The design criteria for SAGE

Spatial data can be classified according to whether they represent discrete objects (points, lines or areas) or fields of continuous variation (surfaces). In


SAGE: version 0.1
File Data W Matrix Table Map Graphs Queries Classification Statistics Help
CRC/Deprivation
File Edit Tools Views Help
200
150
100
50
0
-8
-6
-4
-2
0
2
4
6
8
ARCPLOT
Text Output

	R//OBS	R//EXP	R//SIR	R//LG_SIR	R//ADJ_SIR	R//LG_ADJ	R//TI
27	6	8.8806	67.60	4.21361	70.001	4.24851	-3.8279
28	17	18.5723	91.50	4.51634	112.514	4.72308	-1.2785
29	13	13.8641	93.80	4.54116	93.447	4.53740	0.6252
30	70	47.2914	148.00	4.99721	170.182	5.13687	1.6249
31	62	67.5469	91.80	4.51961	95.572	4.55988	-0.9723
32	54	31.9008	169.00	5.12990	157.212	5.05760	-0.3285

```
R2         0.089     R2-adj      0.069
LIK        -228.115  AIC         460.231   SC          463.973
SIG_SQ     820.154   SIG_SQ(ML)  785.981
F-test     4.497     Prob        0.039

---- Shapiro-Wilk Normality Test ----
Values     0.958     Prob        0.140

--- Ordinary Least Squares Estimation ---
Variable  Coeff     S.D.        t-value   Prob
Constant  100.496   4.134       24.312    0.000
R//TI     2.354     1.110       2.121     0.039
```

Fig. 1. A SAGE session showing table, graphic, text and map windows and the highlighting facility in operation

each case, the attribute values may be at any of four levels of measurement (nominal, ordinal, interval or ratio). This produces a 4-by-4 classification of spatial data (Goodchild 1992; Unwin 1981). While such a classification is useful for many applications, and is widely used in GIS for example, it is less helpful when developing spatial statistical analysis. This is because it and others like it (e.g. MacEacheren 1995) fails to make explicit the stochastic model generating the data, which is a key element of statistical analysis. "Traditional" statistics, for example, is based upon the independent and identically distributed model.

Cressie sub-divides spatial statistics into four different problem areas and types of data that arise from a single unifying and broadly defined spatial model. These data are: geostatistical data, lattice data, point pattern data and object data (Cressie 1991, p8–9). Lattice data refer to attributes recorded at a fixed, regular or irregular, collection of objects (points or zones) in two dimensional space. **X(s)** is a random vector at location s with each row of the vector corresponding to an attribute for which a value is recorded at s. A graph is usually specified which defines the set of neighbours at each location (N(**s**)). Data thus refer to n fixed point locations or zones that partition the area into n subregions. Different probability models can be constructed for the random vector (Besag 1974). The other three types of data arise as a consequence of quite different model assumption, and it is unlikely that a single package could be developed which might handle the full range of spatial data types.

SAGE handles area data (pixels or irregular areas) and point data where the point events are fixed locations to which are attached polygons. The polygon definition is important because many of the techniques in SAGE use the adjacency information, generated by the areal partition, to create a graph (Cressie 1991, p. 384). In both cases attribute values are attached to each discrete areal object and analytical techniques are applied to the attributes across the set of objects. The confirmatory spatial data analysis (CSDA) techniques are from Cressie's lattice model. The exploratory spatial data analysis (ESDA) techniques, since they do not depend on any explicit statistical model, can be considered to come from either Cressie's lattice model or from the point and area elements of the first classification (providing a graph is also defined) with the validity of any operation depending on the level of measurement of the attribute.

Goodchild (1987, p. 332) identifies six fundamental spatial operations that underpin spatial data analysis. Four of these are relevant to the construction of the technical capabilities within SAGE and hence are embedded within its design: operations requiring access only to attributes of areas (e.g. non-spatial statistical analyses); operations requiring access to attributes and the locational identifiers of the corresponding areas (e.g. queries combining spatial and logical rules); operations creating pairs of areas (e.g. identifying pairs of adjacent areas); operations that analyse the attributes of pairs of areas (e.g. spatial autocorrelation tests and spatial regression modelling). Haining (1994a, p. 58–61) provides illustrations of these four operations noting that in addition some areas of analysis require a combination of operations. For example when computing diagnostics for certain types of spatial regression modelling, individual areas are deleted one at a time. Here attribute pairs need to be redefined, and this requires a combination of the third and fourth operations.

2.1 ESDA and CSDA in SAGE

ESDA is the extension of EDA to spatial data (Haining 1990a, p. 197–228; Fotheringham and Charlton 1994). Resistant techniques are used but visualisation includes cartographical as well as graphical tools and some new types of graphical tools. The aims of ESDA include those of EDA (Hoaglin et al. 1983, 1985) but extend to detecting spatial patterns in data, identifying unusual cases given their location on the map (spatial outliers), formulating hypotheses based on the geography of the data and assessing spatial models. It is important to be able to link cases to map objects (such as the corresponding area or point location) in order to answer queries such as: "where are those cases on the map?" or "where are those cases in relation to one another on the map or in relation to other map features?" (Haining et al. 1998).

The underlying data model for EDA distinguishes between the "smooth" and "rough" components of data, that is:

$$\text{DATA} = \text{SMOOTH} + \text{ROUGH}$$

(Tukey 1977). Spatial data also can be conceptualised as comprising smooth and rough properties either with or without reference to the locational identifier attached to each attribute value. The methods of EDA are applicable to the attribute data but when the locational identifier is added smooth and rough properties are defined in terms of where on the map the cases are found. Smooth properties on a map include spatial trend, spatial autocorrelation and spatial concentration. They are sometimes referred to as global properties of the data set and the associated techniques for identifying them, global statistics. The rough properties of the map are the differences between individual data values and the smooth component – however that is defined. A "spatial outlier" is a data case where the attribute value is very different from neighbouring attribute values. Such rough map properties may themselves show evidence of trend, autocorrelation or concentration within subsets of the data. These are sometimes referred to as local properties and the associated statistics as local or focused statistics (Anselin 1995). For a fuller discussion of ESDA see Haining et al. (1998).

Confirmatory data analysis (CDA) involves parameter estimation and hypothesis testing, based on a probability model for the data. CSDA extends CDA methods to testing hypotheses and fitting models whilst recognizing the locational properties of the attribute values – for example testing for spatial autocorrelation in regression residuals or testing for non-constant variance as a function of the population size of areas (Anselin and Griffith 1988; Haining 1990a). CSDA also extends CDA to testing hypotheses and fitting models that are explicitly spatial in the sense that spatial dependency is incorporated in the model specification (Anselin 1988; Haining 1990a; Haining 1990b).

It is in the area of ESDA and model criticism (diagnostic and sensitivity analysis) that SAGE's range of database management and mapping tools available through GIS are most valuable and where graphing tools are also particularly valuable. Spatial statistical analysis includes not just parameter estimation (model fitting in the narrow sense) but a wider process that starts with the identification of data properties that then leads to model specification, model estimation and model assessment in an (iterative) process of data adaptive modelling (Martin 1987). The analyst moves from a "soft" model to one which is made progressively "firmer" by the evidence of data analysis. Database management and mapping facilities are also valuable in a "model driven" methodology where the analyst is confronted by a number of competing representations of the data deriving partly or wholly from subject matter theory (Anselin 1988).

2.2 General facilities in SAGE that underpin ESDA and CSDA

The input to SAGE is a polygon coverage which contains four types of data: the geometry of area boundaries (i.e. their locations); the location of the centroid of each area; topological relationships between areas; attribute values for each area. SAGE contains a number of generic facilities that draw on these data.

The connectivity or weights matrix is a fundamental tool in spatial analysis of area data. It specifies the pairwise spatial relationships between areas (the graph) and can be constructed in many different ways to reflect different assumptions (Haining 1990a). SAGE sets up three basic forms of the matrix during the creation phase of the SAGE environment. One matrix is based on adjacency (0 if areas *i* and *j* do not share a common boundary and 1 if they do). Two further matrices contain inter-centroid distances and the length of the shared common boundary between two areas. These can be converted directly to connectivity matrices by specifying parameter values (Haining 1990a, p. 74). Further forms can be created by starting from an empty matrix or by modifying a pre-existing matrix or by standardizing a matrix so that row sums equal unity. Higher order versions of certain matrices can be automatically generated. The analyst is in a position to examine data properties based on different assumptions about inter-area relationships and assess the sensitivity of findings to these different assumptions.

The analyst may wish to examine sub areas in isolation from the rest of the study area. A second generic facility allows the user to define spatial subsets of the data either by selecting areas one at a time or by selecting a contiguous set based on a user defined box, circle, polygon or spatial lag distance (from a specified case). This facility allows the user to apply techniques or fit models to geographic subsets. Where such techniques call the weights matrix, the rows and columns of the excluded areas are automatically deleted from the matrix. This facility may help in the identification of non-stationarities in the data (Fotheringham and Charlton 1994).

A third generic facility allows the user to reconfigure the internal partitioning through aggregation. It is well known that the results of certain types of analyses depend on the scale of the areal partition that subdivides the study region and, at any given scale, the particular configuration of the partition (Openshaw 1984). A region building module is provided that allows the user to construct a new regional partition or group starting from the initial set of basic spatial units (bsu's). Examples might be enumeration districts or census tracts which are then aggregated. The user can specify the number of new regions (K) to be constructed and grouping proceeds based on three criteria:

1. Intra-region **homogeneity** i.e. the similarity of selected variables within a new region.

2. Inter-region **equality** i.e. the new regions are similar to one another in terms of the sum of one variable.
3. **Compactness**. This is measured by the within-group variance of the X and Y coordinates of the area centroids.

The process begins by defining an initial partition of the bsu's into K regions, which can be done randomly, starting from user selected 'seed' points or by using an existing partition. The algorithm then visits each region in turn and tests whether swapping a bsu to a neighbouring region would improve the objective function which is defined as the sum of the functions for the three criteria:

$$f_O = w_H f_H + w_E f_E + w_C f_C$$

f_H, f_E and f_C are the objective functions for homogeneity, equality and compactness respectively and w_H, w_E, w_C are weights that the user can set to reflect the relative importance to be attached to each criteria in the region building process. The weights can be adjusted to allow for the differing scales between variables, or the change in objective function can be calculated in percentage rather than absolute terms (Wise et al. 1997).

Swaps are allowed as long as the overall objective function improves i.e.

$$f_O^a - f_O^b < 0.0$$

where f^b and f^a are the values of the function "before" and "after" a swap. It is possible that one or two of the individual criteria could be made worse by a swap, but allowing the swap in this situation helps ensure that the algorithm does not stop too soon, which would be likely to produce a poor partition. However, the user can control this process by specifying a threshold, such that a swap will not be accepted if the objective function for that criterion worsens by more than the threshold amount.

In designing a regionalization algorithm, there is always a trade off between speed and optimality. With even a small number of bsu's, the number of possible regionalizations is astronomical, and it is not possible to determine the optimal one by enumerating all the possibilities. A number of methods, such as simulated annealing and tabu, have been developed which attempt to search a large proportion of the solution space before converging on a solution, but this means that such algorithms may take many hours to run (Openshaw and Rao 1995). The regionalisation algorithm in SAGE is designed for use with exploratory analyses where the analyst wishes to assess quickly the possible effects of other, equally plausible, partions on the results of analysis, and so runs quickly, but may stop at a poorer solution than other methods. One of the problems of regionalization, is that it is not possible to say how much poorer one regionalization is compared with another relative to the optimal solution since the latter is unknown. One advantage of linking SAGE to a GIS, is that if desired another regionalization package can be used, and as long as the result can be saved as an Arc/Info coverage, it can be used in SAGE.

Analysing the Sheffield breast cancer screening data is an example of where the ability to create new regions is important. The original data are the proportion of women who choose to take up the service of a precautionary screening for breast cancer of women. At the Enumeration District Level (there are 1159 EDs in Sheffield), the number of women eligible ranges from 2 to 140, which presents two problems: firstly the numbers are very small, making the calculation of rates potentially unreliable, and secondly there is a wide variation, making the level of unreliability very variable across the city. The total number of women eligible for the service in Sheffield was about 40,000, so creating 40 regions should provide a base population in each of approximately 1,000. The regionalisation algorithm was run with equality of numbers of eligible women as one criterion, the other two being homogeneity in terms of material deprivation (measured by the Townsend Index of material deprivation) and compactness. As the algorithm proceeds, a graphical display of the changing objective functions is produced, as shown in Fig. 2. It will be seen from this that the overall trend is an improvement in the objective function, but that after about 10 sequences compactness is allowed to become slighly worse, since this is balanced by a large improvement in the equality criterion.

This regionalization took about 10 minutes to run. Once the regionalisation has finished, a new column is added to the attribute table in which the identifier of the new region to which each area belongs is stored. This column can be used to produce a new polygon coverage for the new set of regions if desired, by performing a polygon dissolve operation in Arc/Info. Equally important, storing the results of the regionalization in this way makes it possible to make an assessment of the results before producing a new coverage, and even to compare several different runs of the regionlization process (since each will be stored in a separate column).

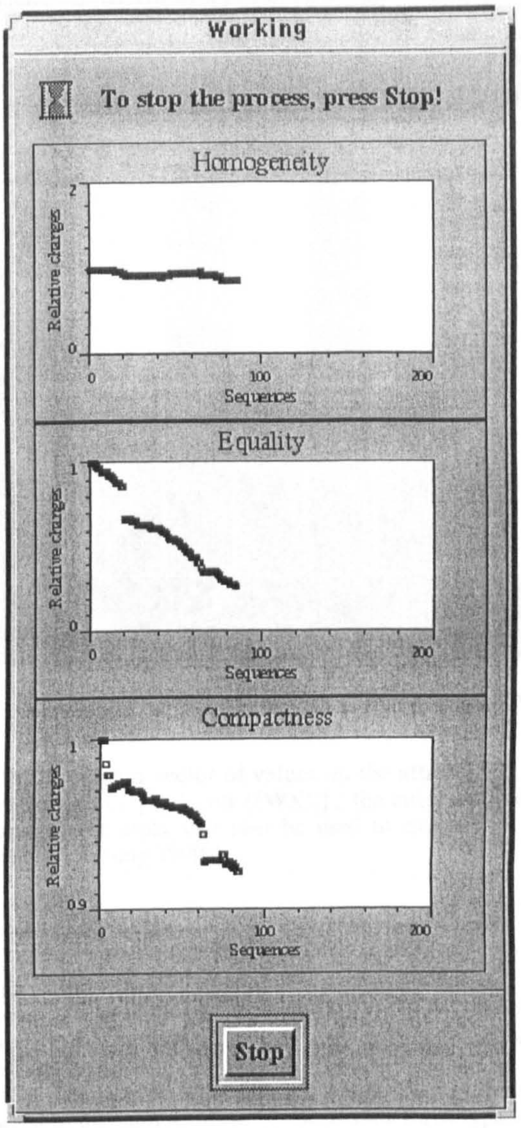
The results of the regionalisation can be assessed by drawing a map of the new regions (i.e. shading the existing areas using the new region identifiers) and by producing graphs which show some of the characteristics of the new regions. Figure 3 shows a histogram of the inter-quartile range of Townsend Index values in the 40 regions – the majority have a range of less than three, compared with an Inter-Quartile range for the original EDs of 5.64. Figure 4 shows the total number of eligible women in each region, which shows that the average of around 1000 has been achieved with a variation from 990 to 1150, providing a much more uniform set of base populations for the calculation of uptake rates. Both plots are produced using a facility in the graph drawing tool of SAGE which will produce a histogram for areas grouped according to an 'index' variable – i.e. the values graphed are summary statistics (sum, mean, interquartile range etc) for areas with the same value of the index.

2.3 Visualisation capabilities in SAGE for ESDA and CSDA

2.3.1 Graph windows

Cases can be visualised graphically in several ways. SAGE provides familiar graphical tools including histogram plots, box plots and rankit plots, to provide a visual test for normality, and various forms of scatterplot.

In addition to familiar graphical forms, SAGE provides a set of graphical tools that deal with the geographical arrangement of the data. There is a lagged box plot facility in which the user clicks on an area and a sequence of box plots are constructed at each lag (up to a user specified number) away from



To stop the process, press Stop!

Fig. 2. Window showing the changing objective functions as the regionalisation proceeds

the selected area. This can help to identify spatial trends in a data set (Haining 1990a, p. 224). The technique is similar to a trellis plot with lag order distance representing the categorical variable (Cleveland 1994). A simple visual check for first order spatial autocorrelation in a set of values is provided by the Moran plot $\{x_i, (\mathbf{WX})_i\}_i$ where $(\mathbf{WX})_i$ is the i^{th} entry in the vector obtained from the matrix product of the row standardised adjacency weights matrix

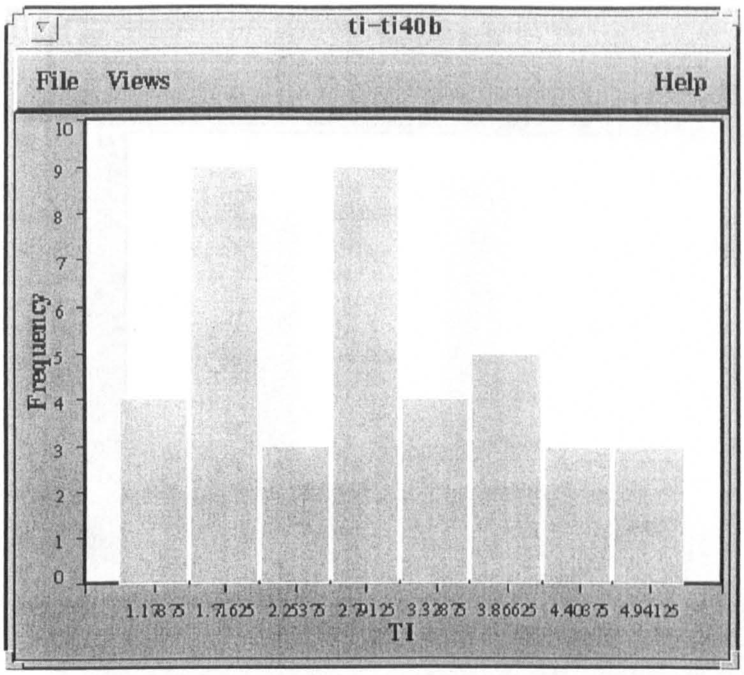


Fig. 3. Homogeneity criterion: Histogram of the inter-quartile range of the Townsend index values in the 40 regions

with the column vector of values on the attribute X. This plot together with a regression of $\{x_i,\}_i$, on $\{(\mathbf{WX})_i\}_i$, the latter available from one of the data management tools, can also be used to explore for the presence of spatial outliers (Haining 1990a).

2.3.2 Map window

This window displays the study area and its internal partition into sub areas. Areas can be shaded, according to values of a selected variable, calling ARCPLOT functions. The user can zoom into and out of the map and edit the map legend.

SAGE contains tools for analysing spatial patterns that are then best displayed in the map window. These include three types of smoothers which may help to reveal general map properties:

Mean smoother: $\sum_j w_{ij}x_j / \sum_j w_{ij}$

Median smoother: Median $\{x_j | j \varepsilon N(i), \text{ or } j = i\}$

Relative risk smoother: $\sum_j w_{ij}x_j / \sum_j w_{ij}y_j$

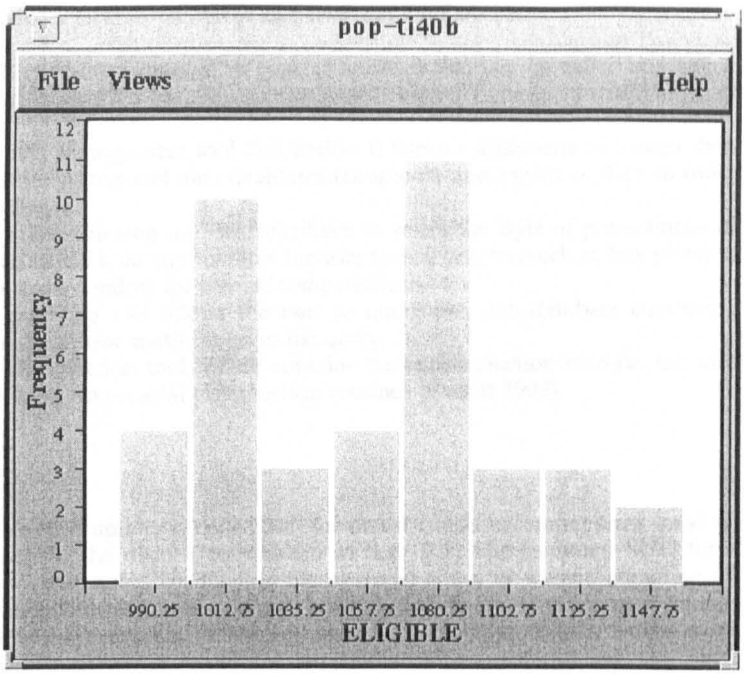


Fig. 4. Equality criterion: Histogram showing the total number of women eligible for screening in the 40 regions

where x_j is the value of the variable in area j, w_{ij} is the value in the connectivity matrix $\mathbf{W}$ where $w_{ii} = 1.0$ and $N(i)$ denotes the set of neighbours of i included in the smoothing operation. The median smoother is a robust statistic and likely to give a smoother representation than the mean smoother. The relative risk smoother, takes the sum of local observed values (x_j) and divides by the sum of local expected values (y_j) (Bithell 1990). When applied to counts of health events these three tools may, in some circumstances provide more reliable estimates of relative risk, by area, by incorporating local information into area estimates (Smans and Esteve 1992). This is discussed further below.

2.3.3 Linked windows

The table, graph and map windows are linked which means that any case or set of cases highlighted in one window will be highlighted in all the other windows (see Fig. 1). Any graph window has a set of tools that allow the user to select cases either by pointing or defining a box on the graph and subsequently to add to any selected set of cases. Only the text output window is "dead" in this sense. This linked windows capability is made possible by the architecture of the system, allowing the user to quickly and easily explore data properties and graphical and cartographical relationships which are of importance in ESDA and areas of CSDA concerned with assessing model fit.

2.4 General facilities in SAGE that underpin data analysis

SAGE contains a number of general facilities that can be called and which support its analysis functions. These facilities are:

1. A data management tool that performs various arithmetic and other data manipulations and also facilitates the import and export of data to other packages.
2. A graph drawing tool with facilities to select the style of presentation of graphs and a facility to allow the user to add graphs (such as box plots) to the same window for ease of comparison.
3. A querying tool allows the user to interrogate the database combining logical and/or spatial rules in the query.
4. A classification tool, which contains the regionalisation module, but also contains non-spatial classification routines (Everitt 1993).

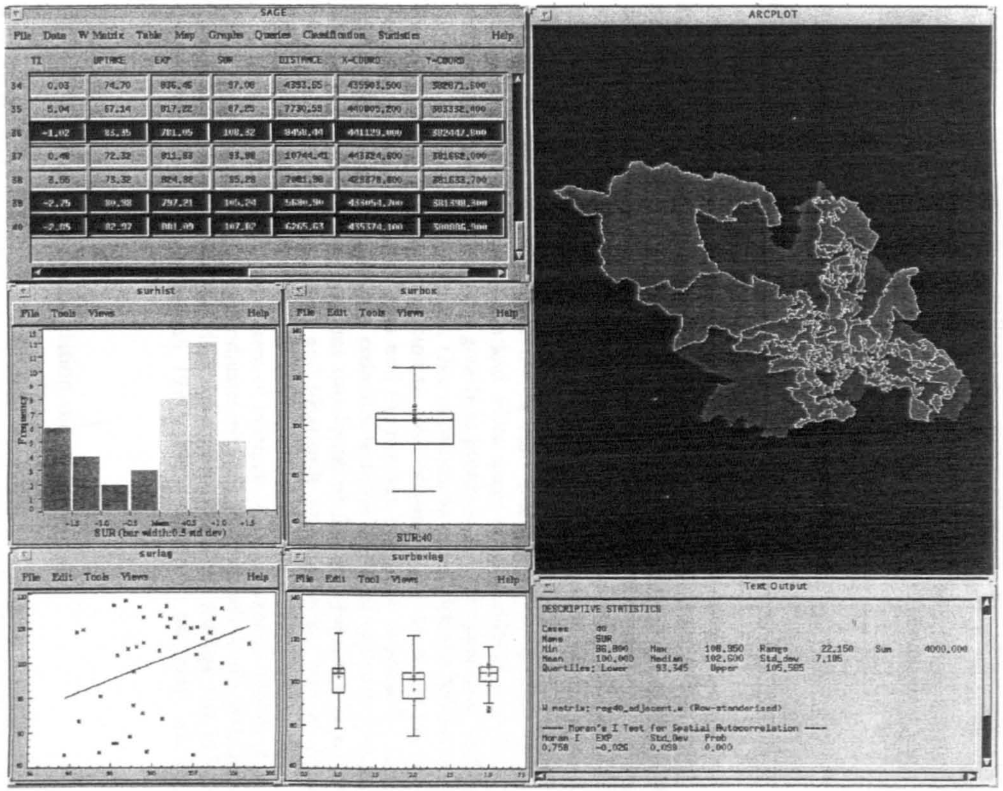
2.5 Example

Standardised uptake rates (SURs) for breast cancer screening were obtained for each of the regions constructed in Sect. 2.2. The (indirect) SUR for a region is computed by dividing the observed count of women attending for screening by the expected count and multiplying by 100. The expected count is obtained by computing the average uptake rate for Sheffield (the total number of women attending divided by the number of women eligible) and multiplying this by the number of eligible women in each region. An SUR over 100 signifies that in that region there is a greater than average uptake of the service by eligible women.

Figure 5 reveals some features of the geographical distribution of these SURs. The histogram is displayed and the boxes above the mean (which in this case is 100) have been highlighted to reveal the geography of these areas in the map window. Interestingly large areas of the inner city seem to enjoy relatively high uptake rates. The two lower graph windows contain the Moran plot and the lagged boxplot up to lag order three where the area containing the site of the (single) breast screening unit has been selected as the origin. There is evidence of spatial autocorrelation in the SURs (confirmed in the text window by the Moran test for spatial autocorrelation). The lagged box plot shows that whilst the median SUR level at lag two is lower than at lag one, the spread of rates is high and by lag three the median has risen. The highlighted cases from the histogram window are identifiable in each of the three boxplots and are well scattered across them. Taken together this evidence lends no support to the view that distance to the screening unit (as measured in this analysis) adversely affects attendance by area.

3 Confirmatory spatial data analysis in SAGE

This section reviews the CSDA capabilities of SAGE. The collection of tools have been assembled to enable the user to identify distributional properties and spatial properties of individual variables with a particular focus on the analysis of health data. SAGE provides tools that can be used to look for


SAGE
File Data W Matrix Table Map Graphs Queries Classification Statistics Help
ARCPLOT
surhist
File Tools Views Help
Frequency
SUR (bar width:0.5 std dev)
surbox
File Edit Tools Views Help
SUR:40
surlag
File Edit Tools Views Help
surboxlag
File Edit Tool Views Help
Text Output

	TI	UPTAKE	EXP	SUR	DISTANCE	X-COORD	Y-COORD
34	0.03	74.70	836.45	97.08	4393.65	435503.500	388671.100
35	5.04	67.14	817.22	87.25	7730.59	440805.200	383332.400
36	-1.02	85.35	781.05	108.32	8458.44	441129.000	382442.000
37	0.48	72.32	811.83	93.98	10744.41	443324.800	381652.000
38	3.55	73.32	824.92	95.28	7081.38	429278.800	381633.700
39	-2.75	89.38	797.21	106.24	5630.90	435054.700	381398.300
40	-2.05	82.37	881.09	107.82	6265.63	435374.100	380886.300

```
DESCRIPTIVE STATISTICS

Cases      40
Name       SUR
Min        86.800    Max        108.950   Range      22.150    Sum        4000.000
Mean       100.000   Median     102.600   Std_dev    7.105
Quartiles: Lower     93.345    Upper      105.585


W matrix: reg40_adjacent.w (Row-standardised)

---- Moran's I Test for Spatial Autocorrelation ----
Moran I   EXP       Std_Dev   Prob
0.758     -0.026    0.058     0.000
```

Fig. 5. A SAGE session illustrating several of the graph tools available for EDA (histogram and boxplot) and ESDA (Moran and lagged box plot) applied to standardised breast screening uptake rates. All regions with standardised rates over 100 are highlighted

statistically significant map pattern at a range of scales and investigate relationships between variables and fit models to the data. SAGE therefore aims to provide a range of tools that go beyond ESDA and which are of use to the data analyst and modeller.

3.1 Univariate analysis

3.1.1 Reliable relative risk estimation

Relative risk is the ratio of observed to expected disease counts (x_j/y_j). Indirect standardisation provides the maximum likelihood estimator (MLE) of relative risk, but when rates have to be shown together as for a group of contiguous areas the MLE is not, generally, the best estimator. If Poisson variation, present in the observed counts, is ignored this could lead to spurious conclusions being drawn about the variation in relative risk rates. However the observed dispersion of the risk estimates (across the map) gives information on this risk variability which can be used to adjust the MLE. The danger of misinterpretation is most marked when areas vary greatly in population size because then estimates vary greatly in precision with the precision least in the cases with small populations. One approach to estimating the relative risk for each area uses Bayes estimation to yield estimates that are a compromise between the MLE for each area and the average risk. The estimate for each area is like a mean of these two estimates weighted by their precision (Smans and Esteve 1992). SAGE contains two forms of Bayes adjustment based on the assumption of independent and identically gamma or log normal prior distributions (Clayton and Kaldor 1987).

This is a rapidly developing area of research (see for example Waller et al. 1997). Valid methods of Bayes adjustment include those where the estimates are a compromise between the MLE for each area and the average risk in just the *neighbouring* areas (Marshall 1991). These methods, however, are not currently available in SAGE.

3.1.2 Tests for spatial autocorrelation and clustering

The methods of Sect. 3.1.1 deal with within-area variation (particularly important with small area data sets). Now we turn to methods that handle between-area variation assuming that the problems associated with within-area variation have been dealt with (Richardson 1992).

Health data may display evidence of spatial autocorrelation and/or spatial concentration. These may be global or local properties of a data set. For example, the entire map of values may show a general (or global) tendency to display autocorrelation but even if no such general tendency is apparant then (local) subsets of the data set may show evidence of significant autocorrelation.

The first evidence may come from applying techniques like those in Sect. 2.3 but SAGE enables the user to test for statistical significance. Both the generalized Moran test (Cliff and Ord 1981, Eq. 1.15) and the Getis-Ord statistic for positive valued variables with a natural origin (Getis and Ord 1992, Eq. 5) are provided in SAGE:

$$I = \frac{n\sum_i\sum_j w_{ij}(d)z_iz_j}{S_o\sum_i z_i^2} \qquad G(d) = \frac{\sum_i\sum_j w_{ij}(d)x_ix_j}{\sum_i\sum_j x_ix_j} \, i \neq j$$

where S_o is the sum of the elements in the connectivity matrix $\mathbf{W}$ and z_i denotes the variable values with mean value subtracted. The term $w_{ij}(d)$ is 1 if the distance between case i and case j is less than or equal to d and 0 otherwise. In the Moran statistic the diagonal elements of the matrix $\mathbf{W}$ are zero.

These are both global statistics. The inference theory for the Moran test is based on the normality assumption (Cliff and Ord 1981, Eqs. 1.37 and 1.38) whilst the inference theory for the Getis-Ord test is based on randomization (Getis and Ord 1992, Eqs. 6 and 7 with the 1993 correction). The statistics are assumed to be normally distributed so the test is based on z-values. SAGE allows the user to specify the weights or connectivity matrix to be used by the test.

SAGE provides local versions of both these tests. The local Moran test is based on the standardized form of the variables (mean zero and unit standard deviation):

$$I_i = z_i\sum_j w_{ij}z_j$$

and the inference theory is based on a randomization hypothesis (Anselin 1995, Eqs. 13 and 14). The two forms of the local Getis-Ord statistics are avaliable (Getis and Ord 1992, Eq. 1 and p. 192):

$$G_i(d) = \sum_j w_{ij}(d)x_j \Big/ \sum_j x_j j \neq i \quad G_i^*(d) = \sum_j w_{ij}(d)x_j \Big/ \sum_j x_j$$

where $w_{ij}(d)$ is 1 if the distance between cases i and j is less than or equal to d, otherwise it is 0. The inference theory is from Getis and Ord (1992, Table 1 with the 1993 correction) with expectations and variances derived under the randomisation hypothesis. The significance of the local Getis-Ord statistics are tested as z-values. Anselin (1995) has drawn attention to the problems of testing in the case of both of these local statistics. Both sample size and number of neighbours affect whether the normal distribution assumption is tenable for the behaviour of the statistics. In addition the distribution moments assume that there is no global spatial association (Anselin 1995, p105). If the global Moran or global Getis-Ord statistics are significant, since the moments of the local statistics are based on the assumption that every value is equally likely at any location (the randomization hypothesis), this assumption is clearly violated. Furthermore, statistics for adjacent individual locations will be correlated since they use overlapping subsets of the data. Both Ord and Getis (1995) and Anselin (1995) discuss adjustments to conventional critical bounds for determining significance.

The presence of trend in a spatial data set or the simultaneous presence of trend and spatial autocorrelation can be tested in SAGE using the regression modelling facility to be described in the next section (Haining 1987, 1988).

Table 1. Summary of SAGE output on regression models

Model	Fitting procedure	Goodness of fit	Hypothesis testing		Diagnostics
			Model	Coefficients	
1	OLS and WLS (ML if errors are normal)	R^2; Adjusted R^2, Residual *SS*; MLL; AIC; SC ML estimate of σ^2 OLS estimate of σ^2	*F* test	*t* tests	Shapiro – Wilks *W* test on residuals; Moran I on the residuals; LM_{ERR}; LM_{LAG}; Residuals*; Fitted values*; Leverages*; Cooks distances*; Internal and external studentized residuals*
2 & 3	ML	Pseudo-R^2 MLL; AIC; SC; ML estimate of σ^2	LR test	Asymptotic tests; Variance-Covariance matrix of coefficients	Shapiro-Wilks *W* test on residuals; Residuals*: Fitted values: *P*-leverages* (for model 2) leverage measure* (for model 3)
GLM	Iterative weighted least square	Deviance	Difference in deviance between nested models	Asymptotically normal	Deviance residuals*; Fitted values*; Leverages*

* Saved as new variables
SS Sums of squares
LR Likelihood ratio
See text for explanation of other acronyms

3.2 Bivariate and multivariate analysis

3.2.1 Statistical tests for bivariate relationships

SAGE contains several tests for exploring relationships between pairs of variables. There are three bivariate correlation tests (Pearson, Kendall and Spearman) and partial correlations can be computed using the tools described in Sect. 3.2.2. The Chi-square test is avalable for exploring relationships between variables measured at the nominal level and the Kolmogorov-Smirnov test can be used on ordinal level data.

Care must be exercised in using these techniques if spatial dependence is present. Richardson (1992) reviews her work on providing a rigorous inference theory for the Pearson correlation test and there are further results (including for the Spearman rank correlation test) together with applications in health research in Haining (1990a, 1991b). These adjustments which involve the number of degrees of freedom for the test are not currently available in SAGE. It is arguable in the context of spatial data that regression (where the problem of what to do about any trends in the data does not arise) provides a more powerful suite of tools.

3.2.2 Modelling facilities

Three forms of the linear regression model are provided in SAGE. All these models are discussed in detail in amongst other sources Anselin (1988), Haining (1990b, 1991a, 1998) and Cressie (1991).

Model 1. The ordinary linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

where $\mathbf{Y}$ is the $N \times 1$ vector of observations on the response variable and $\mathbf{X}$ the $N \times P$ matrix of observations on the $P-1$ explanatory variables and with a first column of 1's. The vector β is the $P \times 1$ vector of parameter values and $\mathbf{e}$ is the $N \times 1$ vector of independent and identically distributed errors with mean 0 and variance matrix $\sigma^2\mathbf{I}$.

Model 2. The linear regression model with spatially correlated errors

$$Y = \mathbf{X}\beta + \eta$$

$$\eta = \lambda\mathbf{W}\eta + \mathbf{e}$$

where $\mathbf{W}$ is the connectivity matrix and λ is the autoregressive parameter on the spatially correlated errors.

Model 3. The linear regression model with a spatially lagged response variable as one of the explanatory variables.

$$\mathbf{Y} = \mathbf{X}\beta + \rho\mathbf{W}\mathbf{Y} + \mathbf{e}$$

where ρ is the autoregressive parameter on the spatially lagged response variable.

Other models can also be fit using these three different routines. Model 1, after first using the data management tool to construct $\{(\mathbf{WX})_i\}_i$, can be used to fit a regression model in which one or more of the explanatory variables is spatially lagged (Haining 1990a). Model 2 can be used to fit univariate models of trend plus spatial autocorrelation (Haining 1987) since as the SAGE environment is set up the two locational co-ordinates for each area centroid are stored in columns of the table.

Model 1 is a standard model and is fit using ordinary least squares (OLS). Fitting procedures for Models 2 and 3 are maximum likelihood (Anselin 1988, p182–183). The statistics generated for Models 2 and 3 have been cross checked using selected data sets and agree with the output from SpaceStat.

All models can be fit on all cases or a user selected subset of cases but new data columns (for example of regression residuals) will only be saved if the "all cases" option has been chosen. Model 1 contains the option to fit by weighted least squares (WLS) with a user defined set of weights. This is provided because if the user is working with areas of widely differing population size, the weights can be set to downweight the influence of areas with small populations in the regression fit (Weisberg 1985, p. 80–83). Each diagonal element of the variance covariance matrix takes on a value that is the inverse of the value of the variable that is specified as the weighting variable. The larger the weight attached to a case the more importance is attached to that case in the estimation.

Table 1 summarizes the statistics that are generated for each of the models. Under goodness of fit, the adjusted R^2 is the usual formula that adjusts for the number of explanatory variables in the model whilst the "pseudo-R^2" is the ratio of the sums of squares of the predicted values over the sums of squares of the dependent variable. The maximized log likelihood (MLL), Akaike's information criterion (AIC) and Schwartz Criterion (SC) are useful for comparing model fits when the sample size is large (Anselin 1988). AIC and SC penalize the MLL assessment by a factor reflecting the number of parameters fit. If the absolute value of one of these three discriminators is smaller for say Model 1 compared to Model 2, this implies that Model 1 provides a better fit than Model 2 according to the selected discriminator.

Details on standard model diagnostics are available, for example in Weisberg (1985), but there are specialist diagnostics provided as well. The Moran test for spatial autocorrelation in the residuals from an ordinary least squares regression (model 1) is from Cliff and Ord (1981; Eqs. 8.12, 8.21 and 8.29) which assumes that the residuals are normally distributed. The Lagrange Multiplier tests for Model 2 (LM$_{\mathrm{ERR}}$) and Model 3 (LM$_{\mathrm{LAG}}$) are given by:

$$\mathrm{LM}_{\mathrm{ERR}} = \{\mathbf{u}^T\mathbf{Wu}/S^2\}^2/\mathrm{trace}[\mathbf{W}^T\mathbf{W} + \mathbf{W}^2]$$

$$\mathrm{LM}_{\mathrm{LAG}} = \{\mathbf{u}^T\mathbf{Wy}/S^2\}^2/\{(\mathbf{WXb})^T\mathbf{MWXb}/S^2 + \mathrm{trace}[\mathbf{W}^T\mathbf{W} + \mathbf{W}^2]\}$$

where $\mathbf{W}$ denotes the connectivity matrix (Anselin 1988) $\mathbf{u}$ is the vector of residuals, $S^2 = \mathbf{u}^T\mathbf{u}/N$ which is the ML estimator for error variance. Here, as elsewhere in the fitting procedure, the user can specify the form of the connectivity matrix to be used and whether it is to be row standardised. The P-leverage diagnostics for Model 2 are from Haining (1994b; Eq. 3) after Martin (1992). The leverages for Model 3 are the diagonal elements of the matrix:

$$(\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{X}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}(\mathbf{I}-\rho\mathbf{W})$$

Where ρ is the estimate of the parameter ρ. Many of the facilities described earlier are valuable here in assessing model fits, for example through graph plots and map displays of the regression residuals.

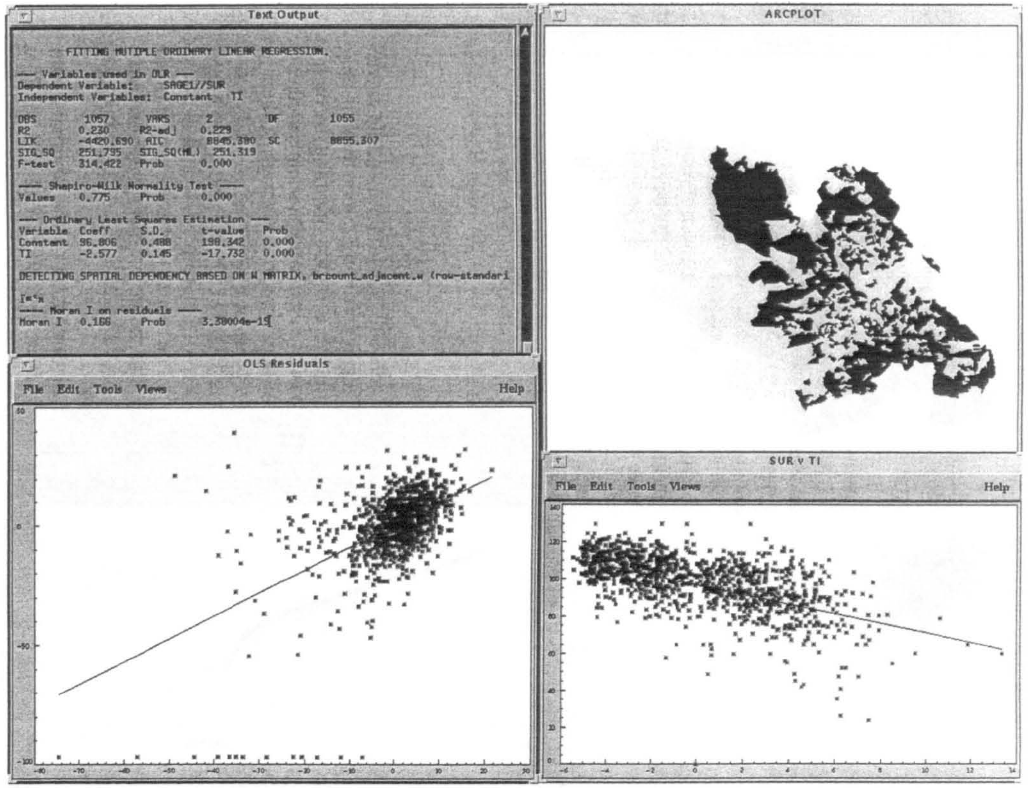
SAGE also enables the user to fit a generalized linear regression model (GLM) with poisson errors by iterative weighted least squares using NAG routine G02GCF (McCullagh and Nelder 1989). The model is appropriate for use in modelling count data such as disease events. Model fit can be assessed by testing the deviance between nested models. The difference in deviance is asymptotically chi-squared with degrees of freedom given by the difference in the degrees of freedom associated with the two deviances. The model does not allow the user to test formally for residual autocorrelation although residual plots can be examined to make an informal assessment. Nor does the specification of the model make allowance for spatial correlation amongst the poisson errors if it exists (Besag 1974). However the fitting procedure does allow the user to include an offset to account for differences in area size and to estimate the effects of over-dispersion which is to be expected in geographical situations.

3.3 Example

To illustrate the use of the modelling tools in SAGE, consider a simple model to explain the pattern of uptake of Breast Cancer screening. Since there is only one screening unit, distance from this might be a factor affecting women's decision whether to attend or not, but the earlier visual analysis suggested this was not the case. It is known that material deprivation is strongly linked to many health outcomes in Sheffield, and so a regression analysis is used to test this hypothesis. For this analysis the original EDs (1159) are used. The graph in the lower right of Fig. 6 shows the scatterplot between SUR and deprivation (measured using the Townsend index). The visual impression of a moderately strong association is supported by the results of fitting an ordinary least squares linear model, which gives an R^2 value of 23%. Although this is not high the model is statistically significant.

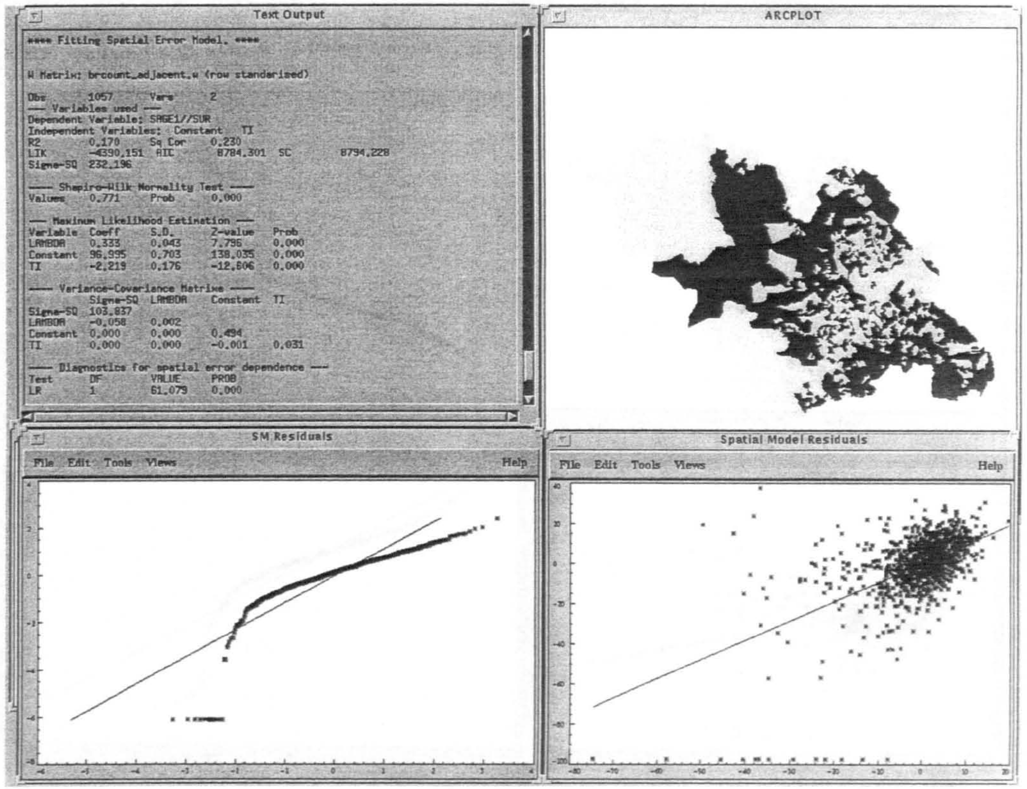
It is important to check for spatial autocorrelation in the residuals, and Fig. 6 illustrates three ways that this can be done in SAGE. The text window shows the value of Moran's I on the residuals, which is one of the optional outputs from the regression. The plot in the lower left is a Moran plot of the residuals, and the map has been drawn by selecting all the positive residuals. The graphical tools suggest the presence of spatial autocorrelation in the residuals, which is supported by a low but significant value for Moran's I.

Models 2 and 3 (see Sect. 3.2.2) were fit. Model 2 is fit because there may be many, hard to specify, factors that also influence uptake rates and fitting Model 2, whilst not throwing any light on what they might be, at least copes with the misspecification present in the OLS regression. Model 3 reflects a possible underlying interaction effect with high levels of attendance in one area associated with high levels of attendance in neighbouring areas perhaps through processes of inter-personal communication. The models were fit in about 15 minutes (real time) and in both cases the additional spatial parame-



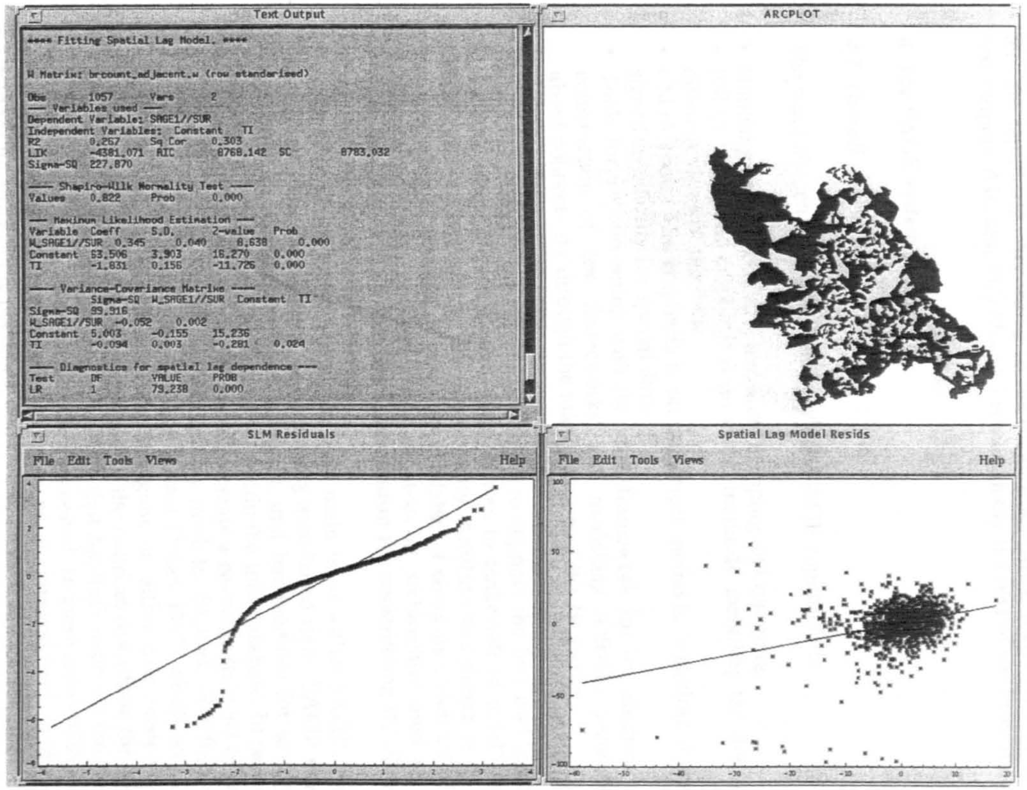
```
        FITTING MUTIPLE ORDINARY LINEAR REGRESSION.

--- Variables used in OLR ---
Dependent Variable:      SAGE1//SUR
Independent Variables:  Constant   TI

OBS        1057      VARS       2         DF         1055
R2         0.230     R2-adj     0.229
LIK        -4420.690  AIC        8845.380  SC         8855.307
SIG_SQ     251.795   SIG_SQ(ML)  251.319
F-test     314.422   Prob       0.000

---- Shapiro-Wilk Normality Test ----
Values     0.775     Prob       0.000

--- Ordinary Least Squares Estimation ---
Variable  Coeff      S.D.       t-value    Prob
Constant  96.806     0.488      198.342    0.000
TI        -2.577     0.145      -17.732    0.000

DETECTING SPATIAL DEPENDENCY BASED ON W MATRIX, brcount_adjacent.w (row-standari

I*^x
---- Moran I on residuals ----
Moran I   0.166     Prob       3.38004e-15
```

Fig. 6. A SAGE session illustrating the results of modelling the relationship between standardised breast screening uptake rates and material deprivation at the ED level. Windows (clockwise from top left): diagnostic output from regression, map of positive/negative residuals, scatter plot and best fit line, Moran plot of residuals



```
**** Fitting Spatial Error Model. ****


W Matrix: brcount_adjacent.w (row standarised)

Obs         1057        Vars       2
--- Variables used ---
Dependent Variable: SAGE1//SUR
Independent Variables:  Constant   TI
R2          0.170       Sq Cor     0.230
LIK         -4390.151  AIC          8784.301  SC          8794.228
Signa-SQ  232.196

---- Shapiro-Wilk Normality Test ----
Values     0.771       Prob        0.000

--- Maximum Likelihood Estimation ---
Variable  Coeff       S.D.       Z-value    Prob
LAMBDA     0.333       0.043      7.796      0.000
Constant  96.995      0.703      138.035    0.000
TI         -2.219      0.176      -12.606    0.000

---- Variance-Covariance Matrixe ----
           Signa-SQ  LAMBDA     Constant  TI
Signa-SQ  103.837
LAMBDA    -0.058     0.002
Constant  0.000       0.000      0.494
TI         0.000       0.000      -0.001     0.031

---- Diagnostics for spatial error dependence ---
Test       DF          VALUE      PROB
LR         1           61.079     0.000
```

Fig. 7. A SAGE session showing the results of using a spatial regression model to deal with spatial autocorrelation in the residuals from the standard linear regression. Windows (clockwise from top left) diagnostic output from regression, map of positive/negative residuals, Moran plot of residuals, rankit plot of residuals



```
**** Fitting Spatial Lag Model. ****


W Matrix: brcount_adjacent.w (row standarised)

Obs       1057      Vars      2
--- Variables used ---
Dependent Variable: SAGE1//SUR
Independent Variables:  Constant   TI
R2        0.267     Sq Cor    0.303
LIK       -4381.071  AIC       8768.142  SC        8783.032
Sigma-SQ  227.870

---- Shapiro-Wilk Normality Test ----
Values    0.822     Prob      0.000

--- Maximum Likelihood Estimation ---
Variable  Coeff     S.D.      Z-value   Prob
W_SAGE1//SUR  0.345    0.040     8.638     0.000
Constant  53.506    3.903     16.270    0.000
TI        -1.831    0.156     -11.726   0.000

---- Variance-Covariance Matrixe ----
          Sigma-SQ  W_SAGE1//SUR  Constant  TI
Sigma-SQ  99.916
W_SAGE1//SUR  -0.052   0.002
Constant  5.003     -0.155    15.236
TI        -0.094    0.003     -0.281    0.024

---- Diagnostics for spatial lag dependence ---
Test      DF        VALUE     PROB
LR        1         79.238    0.000
```

Fig. 8. A SAGE session showing the results of using a lagged spatial regression model to deal with spatial autocorrelation in the residuals from the standard linear regression. Windows as in Fig. 7

ters are significant. According to the diagnostics (MLL (denoted LIK in the text output), AIC and SC) Model 3 is marginally the best of the three models.

4 The SAGE system design

4.1 Overall system design

The main elements of functionality which SAGE supports are:

- Handling area-based data, including mapping and querying.
- ESDA tools. The emphasis is on visual methods, including the linking of different views of the data.
- CSDA tools. The emphasis is on numerical methods, including those designed specifically for spatial data.
- Tools for experimenting with the areal framework for the analysis – the construction of new frameworks, and modelling different assumptions about connectivity through the manipulation of the W matrix.

Since GIS packages contain functionality to support the first and fourth of these categories, a basic design decision was to make use of a GIS within SAGE, rather than re-write this functionality. Arc/Info was chosen as the GIS for a number of reasons: it is widely available; it contains a rich set of GIS functionality; it can be used in the client-server architecture used to build SAGE (see Sect. 4.2); it provides a mechanism for constructing the W matrix from the original area boundaries.

Arc/Info functionality is used in three main ways within SAGE. First, it provides a basic mechanism for managing area-based data. SAGE reads its data from an Arc/Info polygon coverage, and has facilities for saving new data as new coverages or as new tables within the Info database. In particular, the polygon dissolve operation is used to create a coverage for a set of regions defined using the SAGE regionalisation module. Second, Arc/Info stores polygon boundaries in a topological format (Jones 1997) which means that each link has the identifier of the polygons on either side recorded. This information is used in the construction of the matrices describing the relationships between the areas. It is worth noting that Arc/Info is almost unique in the ease with which this information can be accessed. In more recent GIS systems, the topological information is regarded as purely internal data, used to maintain the integrity of the polygons, and in systems such as ArcView, it is not even explicitly stored but calculated on the fly as needed. Third, ArcPlot is used for drawing the map window, and to enable spatial queries within this window.

However other functions needed in SAGE cannot be supported by Arc/Info. Although Arc/Info's macro language (AML) could be used to provide some of them, they would almost certainly have poor performance since AML is an interpreted language. Therefore the other elements of functionality, covering SSA were provided by other specialist software packages. Where suitable public domain code was available (as was the case for software for drawing basic statistical charts) this was used, but certain code (such as for regionalisation and spatial regression modelling) had to be purpose written. The other key decision in designing SAGE was therefore how best to link the GIS and SSA elements.

4.2 SAGE architecture

A number of workers have already considered the issue of how to integrate GIS and Spatial Statistical Analysis (SSA) software (Goodchild 1992; Chou and Ding 1992; Nyerges 1992; Abel et al. 1994; Haining et al. 1996) and two main approaches have been identified: close-coupling and loose-coupling. With the former, the SSA element is called from within the GIS, while with the latter the two components exchange data in the form of files (Goodchild 1992). Although there are integrated systems based on both approaches (Ding and Fotheringham 1992, and Anselin et al. 1993), the use of either is unlikely to lead to both high-performance and a user-friendly Graphical User Interface (GUI). Hence, an intermediate approach was used for SAGE based on the client-server computing model (Haining et al. 1996; Smith and Guengerich 1994; Umar 1993) an approach which has also been used by Cook et al. (1996). In this approach the user interacts with a client, which makes calls to the server to perform certain operations as necessary. Arc/Info could have acted as either client or server, and in the work of Cook et al. (1996) for instance performs both roles depending on the type of analysis required. However, it was decided that using Arc/Info as the server gave greater control over the deisgn of the user interface for the client.

Figure 9 illustrates the system architecture comprising three components (shown as rectangles): the SSA, the linking agent and Arc/Info. The SSA is

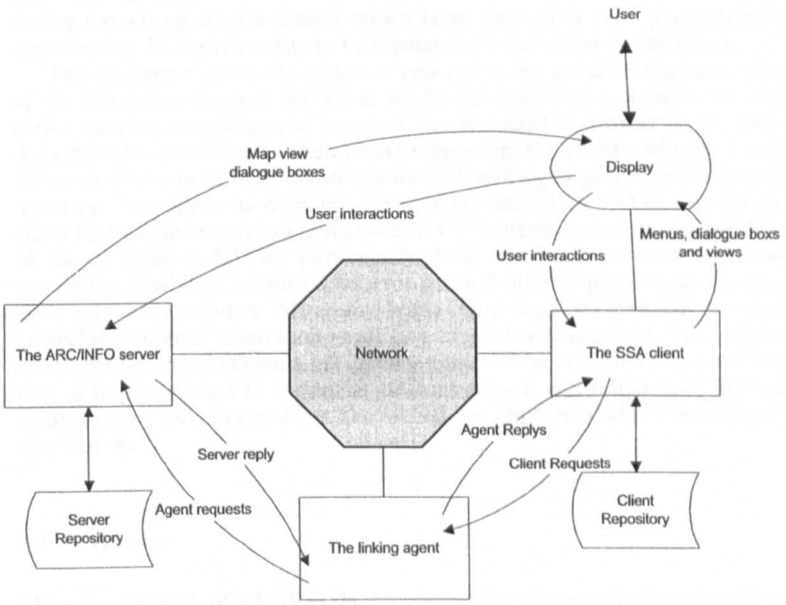


Fig. 9. The SAGE architecture
Notes: Replaces former figure 8

the client which provides the main user interface through the window containing a tabular view of the attribute data (the window labelled 'SAGE' in the top left of Fig. 5). Communication between the SSA and Arc/Info is not direct, but via the linking agent, an approach which will help to minimize the effects on SAGE of new releases of Arc/Info. These components can be run as three independent processes on either a stand-alone computer or on networked computers. Communications between processes are indicated by curved arrows. The display is where visualisation and user interactions take place.

In order to reduce the amount of communication between components and, hence, to increase system performance, two data repositories are employed. The data repository with the Arc/Info server is basically an Arc/Info workspace used to hold analytical results. The SSA client uses the client data repository to manage the connectivity data derived from the Arc/Info polygon coverage and to store all attribute data. All these data are transferred from the server to the client when a session is set up. Analytical results produced during a session are stored in the client data repository which also holds another type of data - area identifiers of currently selected areas. It is the area identifiers that SAGE uses to maintain the dynamic link. The area identifiers are updated whenever a data query is invoked.

With these two repositories, data transfer between the server and client processes is reduced and this is why highlighting cases is very quick. There is no need for the system to transfer any attribute data on the fly from the server to the client whilst the transfer of analytical results from the client to the server data repository is also quite quick since most SSA techniques produce only one variable at a time. However, intensive data transfers are needed during the setting up of a session since a large amount of data, particularly for constructing W matrices, has to be loaded from the server to the client.

The response time of the system is generally very good on test data sets of up to 500 areas. Beyond 500 areas whilst the system is responsive to interactive queries, mapping and most of the standard statistical tools, SAGE does slow down markedly when fitting regression models like Models 2 and 3 where there is a need to fit iteratively whilst handling large sparse connectivity matrices. The application of the client-server model in SAGE has led to a client that carries many complicated and CPU-intensive computations. Some of the facilities in SAGE, particularly those involving some of the spatial modelling capability, might be better separated off and performed on a powerful, separate computer. Alternatively the client could be operated on a more powerful computer when such modelling capability is required. The extent to which further improvements in response time can be achieved is more a function of improvements in certain algorithms than it is a function of improvements in how software systems like SAGE are implemented or the computers they run on.

4.3 Discussion

The main strength of SAGE is the provision of a wide range of statistical, and particularly spatial statistical tools, in combination with graphical facilities which support exploratory analysis. SAGE is also almost unique in providing a regionalization tool as part of a spatial analysis package. However, use of the system in a range of applications, and a comparative study of SAGE with some of the other systems developed for ESDA (Wise et al. 1999) have revealed some limitations, and it is useful to review these since they indicate directions for future work.

First, the forms of interaction (between user and data) provided within SAGE is restricted to brushing objects or groups of objects in a single window in order to see where the selected cases are in the other windows. However other forms of interaction can be implemented. Majure and Cressie (1997) have applied "dynamic graphing" to spatial data analysis. Their tools essentially allow the user to vary the parameters of a graph such as the angle class, lag distance and tolerance range in the construction of different types of univariate and multivariate spatial dependence plots. (Their application is to geostatistical data.) This facility allows the user to assess the effects of these parameters on graphical output for ESDA. In the context of lattice data, varying the class interval of a histogram is one example of a dynamic graphical tool that would be relatively easy to implement in SAGE. Far more problemmatic for a system like SAGE, however, would be the implementation of "dynamic brushing" in which a user defined shape is moved over a geographical region (in the map window) whilst the software returns statistics on the area covered (in one or more graph windows) and which are updated as the shape is slid over the map (Craig et al. 1989; Haslett et al. 1991; Haining and Wise 1991). Such a facility would be impossible to implement effectively in SAGE because as currently structured area identifiers have to travel between the client and the server (the map cannot be part of the client) and this makes it impossible to achieve a rapid response time.

Second, the decision to use the map drawing capabilities of ArcPlot has been a mixed blessing. It avoided the necessity to write more code, but compared with the cartographic facilities in packages which are not linked to a GIS such as MANET (Unwin et al. 1996), cdv (Dykes 1996) and Descartes (Andrienko et al. 1999), the map drawing in SAGE is limited and cumbersome. It is difficult to produce a greyscale choropleth map since the data must be classified and the class values saved, and then a greyscale palette with the appropriate number of shades must be produced. The problem is that ArcPlot, the part of Arc/Info which draws maps, was designed for the production of high quality paper maps, and is not well suited to the demands of interactive cartography and ESDA. On the other hand, those systems with better graphical facilities tend to have a much more limited range of statistical facilities, and there is a clear case for the development of a software package which combines these two elements.

Third, SAGE does not have a facility for handling missing data which means that the user must ensure that any data set imported into SAGE has had missing values estimated or the areas excluded from the file (Bennett et al. 1984). Furthermore, because of the architecture of the system if data generated in the course of a session does not refer to all the sub areas then it is sent to the "dead" text output window and cannot be passed to the table for further analysis. As currently implemented it is not possible to handle either of these aspects of data analysis within SAGE although there is nothing in the design of the system which precludes a missing values facility being added. This is largely a methodological question as to how spatial relationships should be redefined in the case of missing or deleted values (Haining 1994b).

5 Conclusions

This paper has described the capability of SAGE as a package for undertaking spatial data analysis in a GIS environment. In reviewing its capability the purpose has been to contribute to the evolution of such systems.

One of the strengths of SAGE is the combination of a wide range of numerical analytical methods together with graphical facilities for exploratory analysis. The examples above have shown how these two elements of the software complement one another – graphical plots can be used to explore some features of model fitting, and analytical methods can be used to test hypotheses suggested by graphical analysis. A particular strength is the range of explicitly spatial methods in the system, and these include both numerical (e.g. spatial regression) and graphical (e.g. lagged boxplots) methods.

These are possible because of the design decision to link SAGE to a GIS, which provides the basic topological data necessary to underpin such techniques. The general approach of building software by linking existing components is very much in line with current trends in the IT industry, and as the example of SAGE shows, can bring distinct benefits in terms of the wide range of facilities which can be offered. Our experience would suggest that this is a fruitful approach for future work in this area, particularly since many of the GIS vendors are providing basic GIS functionality in the form of re-useable software objects. Whether these are useful for the development of software to support spatial data analysis depends on whether their map drawing capabilities are flexible enough, whether they can support the linking of objects between different windows and whether they can supply topological information to other elements of the system.

Appendix

SAGE and its documentation are freely available from the SCGISA home page at the University of Sheffield:

http://www.shef.ac/uk/~scgisa

Simply follow the link labelled SAGE. Three items are available for downloading:

1. The User Guide in Word for Windows version 6.0 format.
2. Getting Started with SAGE in Word for Windows version 6.0 format.
3. The full package – software and all documents (including the two above)

SAGE has been designed and implemented to work on networked Sun workstations running Solaris 2.5 with the *X* window system. It can also be configured to run on a stand-alone Sun workstation.

Acknowledgements. The authors wish to acknowledge receipt of ESRC research grant R000234470 "Developing spatial statisitical software for the analysis of area based health data linked to a GIS". The authors also wish to thank Dawn Thompson, a PhD student in the Department of Geography at The University of Sheffield for kindly allowing them to use the breast cancer screening uptake data.

References

Abel J, Kilby PJ, Davis JR (1994) The system integration problem. *International Journal Geographical Information Systems* 8(1):1-12

Andrienko GL, Andrienko NV (1999) Interactive maps for visual data exploration. *International Journal of Geographical Information Science* 13:355-374

Anselin L (1988) *Spatial econometrics: Methods and models*. Kluwer Academic Publishers, Dorderchl

Anselin L (1992) *SpaceStat tutorial: A workbook for using SpaceStat in the analysis of spatial data*. NCGIA

Anselin L (1995) Local indicators of spatial association - LISA. *Geographical Analysis* 27:93-115

Anselin L, Bao S (1997) Exploratory spatial data analysis linking SpaceStat and ArcView. In: Fischer M, Getis A (eds) *Recent developments in spatial analysis: Spatial statistics, behavioural modelling and neuro-computing*. Springer, Berlin Heidelberg New York, p35-59

Anselin L, Dodson RF, Hodak S (1993) Linking GIS and spatial data analysis in practice. *Geographical Systems* 1:3-23

Anselin L, Griffith DA (1988) Do spatial effects really matter in regression analysis? *Papers, Regional Science Association* 65:11-34

Bennett RJ, Haining RP, Griffith DA (1984) The problem of missing data on spatial surfaces. *Annals, Association American Geographers* 74:138-156

Besag JE (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal, Royal Statistical Society* B36:192-236

Bithell J (1990) An application of density estimation to geographical epidemiology. *Statistics in Medicine* 9:691-701

Clayton D, Kaldor J (1987) Empirical Bayes estimates of age standardised relative risks for use in disease mapping. *Biometrics* 43:671-681

Cleveland WS (1994) *The elements of graphing data*. AT&T Bell Laboratories, Murray Hill NJ

Cliff AD, Ord JK (1981) *Spatial processes*. Pion, London

Cook D, Majure JJ, Symanzik J, Cressie N (1996) Dynamics graphics in a GIS: exploring and analyzing multivariate spatial data using linked software. *Computational Statistics* 11:467-480

Craig P, Haslett J, Unwin AR, Wills G (1989) Moving statistics - an extension of brushing for spatial data. In: *Computing science and statistics*. Proceedings of the 21st Symposium on the Interface, p. 170-174

Cressie N (1991) *Statistics for spatial data*. Wiley, New York

Chou HC, Ding Y (1992) *Methodology of integrating spatial analysis/modelling and GIS*. In: Proceedings of 5th International Symposium on Spatial Data Handling. Charleston, South Carolina

Ding Y, Fotheringham S (1992) The integration of spatial analysis and GIS. *Computers, Environment and Urban Systems* 16:3-19

Dykes J (1996) Dynamic maps for spatial science: a unified approach to cartographic visualization. In: Parker D (ed) *Innovation in GIS 3*. Taylor and Francis, London, pp. 177-187

ESRI (1998) Open Development Environment. Feb 1998. [Online document] http://www.esri.com. Environmental Systems Research Institute, Redlands, CA

Everitt B (1993) *Cluster analysis*. Edward Arnold, London

Fotheringham AS, Charlton M (1994) GIS and exploratory spatial analysis: an overview of some research issues. *Geographical Systems* 1:315-328

Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24:189-206 (With correction 1993, p. 276.)

Goodchild MG (1987) A spatial analytical perspective on geographical information systems. *International Journal of Geographical Information Systems* 1:327-334

Goodchild MG (1992) Geographical data modelling. *Computers and Geosciences* 18:401-408

Goodchild MG, Haining RP, Wise SM (1992) Integrating geographic information systems and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Systems* 16:407-424

Griffith DA (1988) Estimating spatial autoregressive model parameters with commercial statistical packages. *Geographical Analysis* 20:176-186

Haining RP (1987) Trend surface analysis with regioal and local scales of variation with an application to aerial survey data. *Technometrics* 29:461-469

Haining RP (1988) Estimating spatial means with an application to remotely sensed data. *Communications in Statistics: Theory and Methods* 17:573–597

Haining RP (1990a) *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge

Haining RP (1990b) Models in human geography: problems in specifying, estimating and validating models for spatial data. In: Griffith DA (ed) *Spatial statistics: Past, present and future*. Michigan Document Services Ann Arbor, pp 83-102

Haining RP (1991a) Estimation with heteroscedastic and correlated errors: a spatial analysis of intra-urban mortality data. *Papers in Regional Science* 70:223–241

Haining RP (1991b) Bivariate correlation with spatial data. *Geographical Analysis* 23:210-227

Haining RP (1994a) Designing spatial data analysis modules for geographical systems. In: Fotheringham S, Rogerson P (eds) *Spatial Analysis and GIS*. Taylor and Francis, London

Haining RP (1994b) Diagnostics for regression modelling in spatial econometric models. *Journal of Regional Science* 34(3):325–341

Haining RP (1998) Spatial Statistics and the Analysis of Health Data. In: Gatrell A, Loytonen M (eds) *GIS and health*. Taylor and Francis, London, pp 29–47

Haining RP, Ma J, Wise SM (1996) The design of a software system for interactive spatial statistical analysis linked to a GIS. *Computational Statistics* 11:449–466

Haining RP, Wise SM (1991) *GIS and Spatial Data Analysis: Report on the Sheffield Workshop*. Regional Research Laboratory Initiative Discussion Paper No. 11. Department of Town and Regional Planning, University of Sheffield

Haining RP, Wise SM, Ma J (1998) Exploratory spatial data analysis in a geographical information system environment. *The Statistician* 47:457–469

Haining RP, Wise SM, Signoretta P (2000) Providing scientific visualisation for spatial data analysis: Criteria and an assessment of *SAGE*. *Journal of Geographical Systems* (in press)

Haslett J, Bradley R, Craig PS, Wills G, Unwin AR. (1991) Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *American Statistician* 45:234–242

Haslett J, Wills G, Unwin AR (1990) SPIDER – an interactive statistical tool for the analysis of spatially distributed data. *International Journal of Geographical Information Systems* 4:285–296

Hoaglin DC, Mosteller F, Tukey JW (1983) *Understanding robust and exploratory data analysis*. Wiley, New York

Hoaglin DC, Mosteller F, Tukey JW (1985) *Exploring data tables, trends and shapes*. Wiley, New York

Jones CB (1997) *Geographical information systems and computer cartography*. Longman, Harlow

Levine N (1996) Spatial Statistics and GIS. *Journal of the American Planning Association* 62:381–391

Ma J, Haining RP, Wise SM (1997) SAGE users guide. Available at http://www.shef.ac.uk/~scgisa

MacEachren AM (1995) *How maps work: Representation, visualization and design*. The Guilford Press, New York

Majure JJ, Cressie N (1997) Dynamic Graphics for exploring spatial dependence in multivariate spatial data. *Geographical Systems* 4:131–158

Marshall RJ (1991) A review of methods for the statistical analysis of spatial patterns of disease. *Journal Royal Statistical Society* A 154:421–441

Martin RJ (1984) Exact Maximum Likelihood for Incomplete Data from a Correlated Gaussian Process. *Communications in Statistics: Theory and Methods* 13:1275–1288

Martin RJ (1987) Some comments on correction techniques for boundary effects and missing value techniques. *Geographical Analysis* 19:273–282

Martin RJ (1992) Leverage, influence, and residuals in regression models when observations are correlated. *Communications in Statistics: Theory and Methods* 21:1183–1212

Mathsoft (1999) MathSoft.com. *S*-Plus. http://www.mathsoft.com/splus. [Online document] Mathsoft Inc., Main Street, Cambridge, MA 02142-1521. Visited June 1999

McCullagh P, Nelder JA (1989) *Generalised linear models*, 2nd ed., Chapman and Hall, London

Nyerges TL (1992) *Coupling GIS and spatial analytical models*. In: Proceedings of 5th International Symposium on Spatial Data Handling, Charleston, South Carolina

Openshaw S (1984) The modifiable areal units problem. *Concepts and techniques in Modern Geography* 38. GeoAbstracts, Norwich

Openshaw S, Rao L (1995) Algorithms for re-engineering 1991 census geography. *Environment and Planning A* 27(3):425–446

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* 27:286–306

Richardson S (1992) Statistical methods for geographical correlation studies. In: Elliott P, Cuzick J, English D, Stern R (eds) *Geographical and Environmental Epidemiology: methods for small area studies.* Oxford University Press, Oxford, pp 181–204

Smans M, Esteve J (1992) Practical approaches to disease mapping. In: Elliott P, Cuzick J, English D, Stern R (eds) *Geographical and Environmental Epidemiology: methods for small area studies.* Oxford University Press, Oxford, pp 141–149

Smith P, Guengerich S (1994) *Client-server computing all-in-one reference for total systems development*, 2nd ed. Sams Publishing, Indianapolis, In.

Tukey JW (1977) *Exploratory data analysis.* Addison Wesley Reading, Mass

Umar A (1993) *Distributed computing and client-server systems*, Prentice Hall, New York

Unwin A (1996) Exploratory spatial analysis and local statistics. *Computational Statistics* 11:387–400

Unwin A, Hawkins G, Hofman H, Siegl B (1996) Interactive graphics for data sets with missing values – MANET. *Journal of Computational and Graphical Statistics* 5:113–122

Unwin D (1981) *Introductory spatial analysis.* Methuen, London

Upton GJ, Fingleton B (1985) *Spatial data analysis by example volume 1: Point pattern and quantitative data.* Wiley, New York

Waller LA, Carlin BP, Xia H, Gelfand A (1997) Hierarchical spatio-temporal mapping of disease rates. *Journal American Statistical Association* 92:607–617

Weisberg S (1985) *Applied linear regression.* Wiley, New York

Wise SM, Haining RP, Ma J (1997) Regionalisation tools for the exploratory spatial analysis of health data. In: Fischer M, Getis A (eds) *Recent developments in spatial analysis: Spatial statistics, behavioural modelling and neuro-computing.* Springer, Berlin Heidelberg New York, pp 83–100

Wise SM, Haining RP, Signoretta P (1999) Scientific visualization and the exploratory analysis of area-based data. *Environment and Planning A* 31:1825–1838.

Providing Spatial Statistical Data Analysis functionality for the GIS user: the SAGE project

Stephen Wise, Robert Haining and Jingsheng Ma

Sheffield Centre for Geographical Information and Spatial Analysis

Department of Geography

University of Sheffield

Sheffield S10 2TN

Abstract

Geographic Information Systems (GIS) are being used in a growing number of application areas. As a consequence there have been frequent calls to expand the range of spatial analysis tools available to users of GIS but a reluctance on the part of GIS software vendors to include such tools in standard software packages. An alternative approach is to link extra tools to GIS packages which raises a series of issues: what sort of tools should be included? how should the linkage be done? and to what extent can the functionality of the GIS be used? This paper draws on the results of a project in which software for statistical spatial data analysis (SSDA) was linked to ARC/INFO to produce a software system called SAGE. The statistical tools implemented included those which were felt to be useful to the general GIS user (as opposed to the specialist spatial statistician or econometrician), and they were linked to ARC/INFO using a client server architecture. The GIS was used within the context of SSDA for map drawing, spatial queries and operations on the topology of the spatial data, although it was found that the map drawing facilities of ARC/INFO were not well suited to the needs of this application. One of the conclusions of the project was that many of the techniques of exploratory spatial data analysis, such as providing graphical data summaries and linking these to cartographic views of the data could be easily integrated into existing GIS packages, providing a useful addition to their functionality for many GIS users. Many of the other SSDA facilities are probably still best provided in specialist software, but

there is a need for a robust and standardised means for such software to extract information about the topology of spatial data from within GIS packages.

1. Introduction: the case for developing SAGE for GIS users

A feature of the GIS research agendas of both the UK Regional Research Laboratories and the American National Centre for Geographic Information and Analysis (Masser 1988, NCGIA 1989, Openshaw 1990) was the perceived need for GIS users to have access to a greater range of facilities for undertaking spatial analysis. Spatial analysis techniques can be defined as those 'whose results are dependent on the locations of the objects or events being analysed' (Goodchild et al 1992) . For example, computing the arithmetic mean of a set of values located across an area would not be a spatial analysis technique (since any re-arrangment of the values on the map would leave the arithmetic mean unchanged) but fitting a trend surface to the set of values would be because the order of the surface and the parameter estimates would be affected by where values were located.

Wise and Haining (1991) identified three types of spatial analysis which might be of interest to those working with GIS: map-based analysis, spatial modelling and statistical spatial data analysis (SSDA). SSDA, which is the subject of this paper, may be described as the analysis of empirical spatial data using statistical methods. There is a considerable degree of concensus as to the types of techniques that ought to be made available to the GIS community to facilitate SSDA (see for example the papers by Bailey, Haining and

O'Kelly in Fotheringham and Rogerson (1994) and those by Haining and Anselin in Fischer et. al. (1996)). Two types of SSDA can be identified, although there is some overlap between them. Exploratory spatial data analysis (ESDA) is concerned with detecting spatial patterns in data, identifying unusual or interesting spatial features of the data (such as spatial outliers), formulating hypotheses which are based on or which are about the geography of the data and validating spatial models (Haining et.al. 1998). Confirmatory spatial data analysis (CSDA) is concerned with model building, which normally involves the estimation of parameters (and their errors) and usually includes hypothesis testing as part of the process of model specification (Haining et. al. 2000).

Current GIS contain only limited support for GIS-relevant SSDA (Goodchild 1987, Burrough 1990) which is not surprising since early developments in GIS were driven by a need for mapping, map-based analysis, and facilities management. However a number of factors have lead to a growing interest in facilities for more sophisticated analyses of spatial data. First, there is a growing number of large spatial databases, including topographic, environmental and socioeconomic data. It is possible to integrate datsets within GIS and this capability has significant practical implications for academic research in many areas where processes are multi-variate across social, economic and environmental explanatory variables. The analysis of fine grained geocoded multivariate datasets has the potential to make a uniquely important contribution to disentangling associations and assessing the importance of local conditions in fields like spatial epidemiology or environmental criminology.

Second, many organisations in both the public and private sector are under increasing pressure arising from increasing "demand" and spiralling costs to increase efficiency by targetting resources, managing and auditing more rigorously. Geographical targetting and geographically based auditing are elements in a resource allocation policy given greater impetus by government pressures for public agencies to work together ("joined up government"). In many cases spatial analysis can be a useful tool in supporting the strategic and operational needs of organizations. In the field of health services provision, people's health and use of the health service often correlate with patterns in the social and economic circumstances of the population. Obtaining a reliable picture of these patterns, and of areas with unusually raised incidence can lead to a more targetted use of resources, but requires more than simply mapping the basic incidence rates, because these can be affected by other factors such as the age and sex structure of the area and the size of the population at risk. Similar examples can be cited in the areas of criminology, housing and education. There is a potential market here of GIS users for whom certain aspects of SSDA functionality would be useful.

Since GIS were not defined with SSDA in mind, it can be argued that it is preferable to build specialist SSDA software, and import the necessary spatial data from the GIS (Haslett et al 1990, Unwin et al 1996, Dykes 1996 and Brunsdon 1998). However it is then not possible to use those features of GIS which can help support SSDA, such as the ability to draw maps and the handling of information on the geometrical and topological properties of the spatial objects (Wise and Haining 1991). In addition there is the inconvenience of transferring files and the danger of creating multiple versions of the same dataset within the different packages. An alternative view is that SSDA facilities are likely to be most widely and effectively used if they can be linked to the GIS packages used to store and manipulate these spatial databases. One way of doing this is to use the GIS as the main software platform, using its customization facilities to supply the additional functionality (Ding and Fotheringham (1992), Batty and Yichun (1994) and Kehris (1990a, 1990b)). Another approach is to link the GIS with another package (Anselin and Bao 1997, Cook et al 1996).

This paper reports on the main findings of a project, one of whose aims was to explore whether it was possible to develop an SSDA package linked to a GIS, which could provide general purpose facilities for both ESDA and CSDA and take advantage of the features of the GIS. To this end, a software system called SAGE (Spatial Analysis in a GIS Environment) was designed and implemented which included a direct link to the ARC/INFO GIS. The functionality of SAGE has been described elsewhere (Haining et al 2000, Wise et al 1997) and complete documentation for the package is available online (Ma et al 1997). The purpose of the present paper is to focus particularly on the role which standard GIS functionality plays in SAGE, and drawing on our experience of SAGE and other similar packages, to discuss some of the issues relating to the inclusion of SSDA into GIS

The paper next describes the criteria which governed the selection of the SSDA functionality to be included in SAGE, and the most appropriate means of incorporating ARC/INFO. The features of the system are illustrated using an example and the paper concludes with a discussion of some of the implications of this work for future developments in the provision of SSDA facilities for the GIS community.

2. Design criteria for SAGE

SAGE was designed for the analysis of what Cressie (1991) calls 'lattice data' in which the spatial units are fixed, and interest lies in the variability of the attributes across the units. Areas are the commonest example of such spatial units, but many of the techniques of area-based analysis can also be applied to the analysis of data for fixed points. Such data can be analysed by SAGE, by attaching a notional area to each point (normally by the generation of a Dirichlet tesselation around the points).

The intention was to provide a wide range of both ESDA and CSDA methods, allowing the analyst to proceed all the way from the initial exploration of a set of data, through to model specification, calibration and validation. ESDA is often used as a means of suggesting hypotheses about data which can then be tested more formally using CSDA methods. However, the division is not always so clear cut and the two approaches are best seen as complementary. For instance, although regression modelling uses formal tests of significance for model specification, visual, exploratory methods also have an important role to play in model specification, and can be used to check assumptions and examine the results of the analysis.

EDA methods (Tukey 1977) are characterised by an emphasis on visual and statistically robust means of exploring data, and this same emphasis has been carried over into ESDA.

What is particularly important in the case of spatial data however, is the connection between the attribute values and the location of the areal units in geographical space. If a boxplot reveals a distributional outlier, then one of the first questions the analyst is likely to ask is 'where is that case on the map?' This link is provided as standard functionality in most modern GIS software, but the connection is normally between the selection of one or more records from the database, and the selection of areas on a map. Haslett et al (1990) were among the first to demonstrate that by making a much richer set of links, between various forms of statistical graph and a map, a wide range of analyses become very easy to perform. This linked windows or brushing facility is now provided by most software packages which have been developed for ESDA (MacDougall 1992, Dykes 1996, Brunsdon and Charlton 1996, Cook et al 1996, Unwin et al 1996) and was regarded as an important element in the functionality of SAGE.

Area-based data have two characteristics which led to a number of other design criteria for SAGE. Firstly, many spatial phenomena are characterised by a degree of spatial autocorrelation, particularly positive spatial autocorrelation which is the tendency for nearby locations to have similar values of a given attribute. This means that samples taken from neighbouring locations cannot be regarded as independent. This invalidates one of the central assumptions of (parametric and nonparametric) classical statistical methods, and may render the results of significance tests and the estimation of confidence intervals unreliable. A variety of techniques and models have been developed to deal with this situation (Haining 1990) and one of the aims of writing SAGE was to make some of these available in an easily accessible fashion. Autocorrelation effects are also of interest in their own right, including the identification of first order effects (what trends exist in attribute values across space?) and second order effects (to what extent are attribute values in neighbouring areas correlated?). This type of analysis requires information about the spatial arrangement of the areal units which is normally handled as a matrix (often called the connectivity or **W** matrix) representing the nature of the links between all pairs of areas e.g. whether the areas are contiguous or not (Haining 1990). To construct this matrix requires information about the topology of the areal units, information which can often be extracted from a GIS, although as Goodchild (1987) points out, GIS are not generally designed to handle the information in matrix form. One of the aims was that SAGE would not only be able to handle the **W** matrix, but would allow the analyst to modify it in order to model different assumptions about the nature of the links between areas. A simple example is the situation where a single administrative zone is divided by a river, thus producing areas in the GIS data which are apparently not connected. One way of dealing with this is to modify the **W** matrix to reflect the fact that the areas are neighbours.

The second important feature is that the majority of area-based data is derived by aggregating values for individual items (people, households, houses etc). The areas are modifiable in that the aggregation could be done at any one of a number of different scales, and at any given scale in numerous different (but equally plausible) ways. It has long been known that different aggregations of the same data can lead to different analytical results (Kendall 1939, Openshaw 1984, Openshaw and Rao 1995). In practice, areas for which data are available often have a real significance in the sense that they represent divisions of responsibility for an organisation - examples are health and police areas. However for any analysis which is trying to investigate patterns and processes in the underlying variables which have been aggregated, there may be several reasons why it is useful to be able to re-aggregate the data into new zones. First, it is common in area-based analyses to convert absolute count data into rates, using the population of the areas as the demoninator. However, the reliability of such rates will vary since it is a function of the size of the populations (Kennedy 1989, Clayton and Kaldor 1987) and one solution to this is to aggregate the areas into larger units with approximately equal populations. Second, rates calculated for areas with small populations will be particularly sensitive to inaccuracies in the data, such as errors in the basic count variable, or errors in assigning individual events to areas. This sensitivity can be reduced by aggregating the original zones into a series of larger ones. Third, with a large number of small areas, broad trends may be lost in local detail, and so aggregation is one way of identifying broad scale trends in the data.

A number of other design criteria were adopted because of the particular nature of the project. Because of the level of resources made available, it was important to use existing software wherever possible, rather than writing code from scratch. Since funding was from a UK research council, it was important that any resulting software be freely available to academics in the UK (and as far as possible, elsewhere). ARC/INFO was chosen as the GIS, because of its widespread availability in the GIS community, and because it is available to UK academics at a heavily discounted price due to a central purchase deal (Wise 1990).

3. The architecture of SAGE

In this section we consider how much of the functionality needed to support the specific programme of SSDA could be provided by the GIS itself. We only consider ARC/INFO as this was the GIS used in the work. However because ARC/INFO has a particularly rich set of functionality (one of the reasons it was selected for widespread use in the UK academic community (Wise 1990)) this does not restrict the generality of the discussion although clearly other systems may well have particular strengths which would have lead to a slightly different division of labour between the GIS and the other pieces of software. We will return to the general issue of how best GIS might support spatial analysis in the concluding section.

The operations needed to support the functionality outlined in the design section can be classified as shown in Table 1

<Table 1 about here>

The table presents the functionality at two levels - as seen by the user (the High Level), and broken down into the underlying technical operations needed to support the high level functionality (the Fundamental Level).

Many of the high level operations cannot currently be performed by ARC/INFO, but in many cases it does possess the underlying technical capability to support them. This can be seen in the functions listed under data display for instance. ARC/INFO has comprehensive map drawing capabilities, some ability to display attribute data in tabular and graphical form, but no linked windows facility. However, using AML and the graphical drawing primitives of Arcplot, it is possible to implement these features entirely within ARC/INFO. For example, Batty and Yichun (1994) implemented a model of urban land use in ARC/INFO in which two different views of the model - a map and a graph - are drawn within the same Arcplot window. Using ARC/INFO's macro language (AML) the two views were linked so that when elements were selected from one view, the same elements were highlighted in the other view. Similar experiments were carried out during the writing of SAGE, but it was found that this was not an appropriate way to build a general-purpose linked windows facility. All the graphics must be contained in the same window, which is inflexible and cumbersome and the use of AML, an interpreted command language, makes the response of the system very slow.

To write SAGE it was necessary to provide some of the functionality outside ARC/INFO and to link this with ARC/INFO itself. Given the need for a rapid and responsive system, the majority of the graphical capabilities were provided outside ARC/INFO, simply using ARC/INFO for drawing the maps. ARC/INFO would also be used to perform the selection of spatial subsets of data (e.g. all areas within a defined polygon), again since it already has comprehensive capabilities in this area. Many of the numerical elements of SAGE are computationally intensive (especially the classification/regionalisation elements) and would perform more efficiently if written in a third generation language such as C++. Those parts of SAGE which related to the handling of topological data are split between those which are central to ARC/INFO and which it would be foolish to re-write (generating the **W** matrix from the polygon data and dissolving a set of polygons as a result of a reclassification) and those which it would be hard to do in ARC/INFO (editing the **W** matrix).

In summary, the major functions for which ARC/INFO was suitable within SAGE were map drawing and querying, generating the basic topological data and the polygon dissolve operation. All the other functions of SAGE would be provided using other software. The next question to be considered was how best to link ARC/INFO with the other elements of SAGE. Other workers have used a number of approaches for this task - loose coupling via the transfer of files (Anselin et al 1993), close coupling in which the GIS calls routines written in other languages (Ding and Fotheringham 1992) and client-server computing (Cook et al 1996). Since the intention was that SAGE should operate rapidly and responsively, especially when the graphical, exploratory tools were being used, methods such as loose coupling were discarded because they would be too slow. Close coupling would be quick, but it was felt that the capabilities which could be provided might then be constrained by what could be achieved from within ARC/INFO.

The client server architecture (Umar 1993) provides flexibility in the way that various software components may be linked, and for this reason this approach was chosen for SAGE (Haining et al 1996). A component is considered a client if it requests the services of other components to complete a certain task, or as a server if it provides services for clients. The communications between clients and servers are handled efficiently through a set of well-defined Application Program Interfaces (APIs) utilising, for example, remote procedure calls (RPCs) (Simon 1996, p.65-8).

As described above, SAGE has two major software components - ARC/INFO and a purpose written module for providing all the other functionality. These could have been implemented so that both could function as either client or server which is the approach used by Cook et al (1996) to link ArcView and Xgobi. However, this results in a system in which the user must know when to use each component as the client and it was decided that it would be simpler if the purpose written SSDA module was the client, calling ARC/INFO as a server for mapping and dataset management.

This approach has a number of advantages. Firstly, it allows the client and the server to be implemented independently communicating with each other only through pre-defined APIs. Secondly, it allows existing and tested code appropriate for their implementation to be re-used wherever possible without being constrained by each other. Therefore, the implementation workload could be reduced and the reliability of the system could be expected to be high. Thirdly, because the client and the server communicate through a pre-defined API, it offers the possibility of using a different GIS to replace ARC/INFO, thus making the SAGE client potentially portable. Fourthly, since the client-server model can be used for distributed computing (using RPCs for example) the SAGE server and the SAGE client could be run on different platforms on the network. This would be useful for analysing spatial data held remotely. All these advantages were exploited during the implementation of SAGE where networked SUN workstations running X-Windows were used.

The full architecture of the system is shown in Figure 1.

<Figure 1 about here>

It can be seen from Figure 1 that the main link between the SAGE client and ARC/INFO is via a linking interface which translates requests from the client into a series of ARC/INFO commands (some of which are actually purpose-written in AML) and converts the responses from ARC/INFO into a form suitable for the client.

4. An example SAGE session

A comprehensive description of the facilities provided in SAGE is available elsewhere (Ma et al 1997, Haining et al 2000). This section will attempt to give an overview of how the system operates by describing how a simple SAGE session might work. The example has been chosen to illustrate the sort of facilities which might be of interest to the general GIS user, and so the focus is on graphical, exploratory methods rather than the statistical modelling techniques which are also provided in SAGE.

The example is taken from work undertaken on behalf of the Trent Region of the UK National Health Service Executive, examining trends in the pattern of ill health in the region. The only health variable which will be considered here is the proportion of people in each area who have a limiting long term illness (LLTI) i.e. one which they consider limits their ability to work. It is well known that material deprivation is strongly linked to people's health, and so in an initial consideration of LLTI the questions which might be posed would include:

- What is the pattern of LLTI variation in the area?
- Is there a relationship between LLTI and deprivation?
- Is there any evidence of spatial clusters of high rates of LLTI?

Figure 2 shows a screen shot of a SAGE session. The window in the top left (labelled SAGE) is the user interface to the SAGE client . The map, which has been drawn by ArcPlot, shows the rates of LLTI in the region's 871 wards. Figure 3 shows the location of the region which roughly corresponds with the Eastern half of the English Midlands.

<Figure 2 about here>

<Figure 3 about here>

In order to draw the choropleth map in SAGE (Figure 2) several steps are needed. First the LLTI values must be classified, to identify which class each ward will fall in. This is done using the classification option on the main SAGE window, and the resulting classification is saved as a new variable (labelled LLTI-5) which is displayed in the main

SAGE window (Figure 2). In this case the variable was grouped into five classes, with equal numbers of wards in each class (a quintile classification), and so it was necessary to create a palette of five shades of grey, ranging from dark to light, in order to produce the final map.

What is of particular interest is the distribution of high rates of LLTI which can be identified in several different ways in SAGE. Figure 2 shows a boxplot of the LLTI rates with the points above the upper quartile selected graphically. This causes the corresponding areas on the map to be highlighted, in this case using a cross-hatch shading. There seem to be two main areas of high LLTI rates – along the coast north of Skegness, and in the main urban areas (Leicester, Nottingham, Sheffield, Rotherham). In order to check that these high values are not simply the result of wards with very small populations, a histogram of the population in each ward is also shown in Figure 2. When the high rates are selected in the boxplot window, the same areas are highlighted in the histogram window showing that some high values do occur in the least populated wards. Rates based on small populations are not robust so the regionalization module of SAGE was used to merge some of the smaller wards together. This module is fully described in Wise et al (1997) and Haining et al (1998). It allows areas to be combined together to produce new regions, which satisfy one or more of three criteria: homogeneity in terms of one or more variables, equality in the total value of one variable and compactness of shape. The original 871 wards, with population values between 837 and 30,450, were combined to produce 500 regions using the criteria of population equality, and homogeneity in values of the Townsend deprivation index. Although the minimum population in the new regions only rises to 1136, none of the high LLTI rates is now found in a zone with a small population.

In order to look at the relationship between LLTI and deprivation a scatter plot is drawn using these two variables as shown in Figure 4.

<Figure 4 about here>

The scatterplot suggests that deprivation is an important factor in the distribution of ill health in this area. The graph also shows three areas where the rate of LLTI is considerably higher than expected given the levels of deprivation. These are outliers from the regression fit (i.e. the standardized residuals are more than three standard deviations above the regression line) and to see where they are located, they have been selected in the scatter plot window. The map shows that one is located on the outskirts of Rotherham in the north, the other two on the coast.

The analysis so far has suggested that wards with high rates of LLTI tend to occur in urban and coastal areas but this has been based on considering each ward independently. By considering each ward in relation to its neighbours, it is possible to identify clusters of wards with high LLTI values, and this has been done in Figure 5.

<Figure 5 about here>

The Getis-Ord statistic (Getis and Ord 1992) has been calculated for the LLTI values and the histogram of the resulting values is shown on the right of Figure 5. A high positive value indicates a ward with a high LLTI value with neighbours which also have high LLTI values, although individually they need not necessarily fall in the upper tail of the LLTI histogram. Selecting the top three categories from the histogram reveals the location of these clusters on the map. None of these are located on the coast. There is a large cluster around the Sheffield/Rotherham conurbation particularly in those areas which house workers in the steel industry. The other major cluster is further south around Derby. Although this lies in one of the areas previously identified as having generally high rates (Figure 2) it is not clear why this particular area should appear to have a cluster of high values. Indeed, this may simply be an artefact of the area boundaries, something which could be explored by repeating the analysis using a different set of regions.

This is a fairly brief analysis of this set of data but even so the example illustrates how the use of relatively straightforward numerical and graphical techniques within SAGE can reveal useful information about spatial data.

5 Discussion: operational benefits and other approaches

Section one presented the case for building SAGE in terms of GIS user needs. In this section, we discuss the extent to which the SSDA facilities in SAGE benefited from being linked to a GIS. Within SAGE, ARC/INFO performs data management tasks, provides topological information and displays maps. The topological data is needed for the calculation of spatial statistics, such as the Getis-Ord statistic, and in the regionalization process. For both of these, the contiguity information about the area boundaries is extracted by SAGE from ARC/INFO. The regionalisation algorithm assigns each of the original areas a code which identifies which new region it will belong to, and this information is passed back to ARC/INFO so that it can perform a polygon dissolve operation to create a polygon coverage for the new regions.

With spatial data analysis the analyst should be able to manipulate the spatial framework and there is a strong case for making use of the functionality of existing GIS to do this. There is a large difference between the ability to import or calculate the topological information for a set of area boundaries, as is done by systems like cdv (Dykes 1996) or LiveMap (Brunsdon 1998), and the ability to alter those boundaries and update the topology and this seems one area of ESDA where a link to GIS is of great benefit.

The cartographic facilities of ARC/INFO work well in the case of selecting spatial subsets of the data and for implementing the brushing technique. However, it would be useful to make this facility more dynamic by being able to change the selection of areas interactively, with the related graphical plots changing simultaneously (dynamic brushing). However, given the need to communicate each change between the server and client in the current SAGE design, this would be far too slow using the current architecture. For this to work at an appropriate speed, the cartographic and graph drawing facilities would have to be provided in the same piece of software.

There are other respects in which the cartographic facilities of ARC/INFO are less suitable for a system like SAGE. ARC/INFO's Arcplot module was originally written with the production of paper maps in mind, not interactive visualisation. For example, it is difficult to draw a greyscale choropleth map of a variable unlike in systems such as cdv (Dykes 1996), MANET (Unwin et al 1996) and Descartes (Andrienko and Andrienko 1999).

In order to derive some general conclusions about the possible role for GIS in supporting spatial analysis, it is necessary to consider the extent to which our experience in building SAGE might have been different if we had used a different GIS package. Our view is that the majority of GIS packages provide a very similar mix of functionality to that provided in ARC/INFO - none appear to possess a wider range of standard statistical graphs, or numerical statistical facilities. Some, such as MapInfo, support linked windows, but this appears to be limited to a single link between the map and a tabular display of attribute values. However some other packages might have been more suited to providing some of the functionality provided by ARC/INFO. In the case of map drawing this is almost certainly the case. More recent desktop mapping packages, such as ArcView and MapInfo for example have been designed with interactive map viewing in mind, and provide methods which allow data to be viewed very quickly on screen.

The one area in which ARC/INFO is perhaps strongest is in the accessibility of the topological information. Many of the current generation of vector GIS packages are based on the topological link and node data structure (Peucker and Chrisman 1975) which stores the identity of the two polygons bordering each link. This information can be used to construct two of the three **W** matrices which SAGE uses - that based on simple contiguity, and that based on the length of the border between areas.

In the case of ARC/INFO the data structure can be held internally, or it can be stored in a relational table in the INFO database for both line and area coverages making it directly accessible to the user. ARC/INFO is unusual in allowing such easy access to this information, and there are at least two reasons why most other software systems do not. Firstly, since the topological data is fundamental to the integrity of the GIS database, it is not desirable to store it in such a way that it can potentially be corrupted by the user. Secondly, there should be no need for the user to have direct access to this low level data - all of the operations which require it should be provided as high level functions. However, as SAGE has shown, certain types of spatial analysis which are not currently supported by GIS also require access to this topological data. Given the reluctance of vendors to incorporate SSDA techniques into standard GIS (Maguire 1995) this makes it important that the mechanisms provided to link GIS to other software include some means of accessing the topological data.

The current trend in the GIS industry is towards the greater use of interoperability as a means of building software solutions, especially using an object oriented approach (Graham 1999), and the SAGE project and the work of others (Cook et al 1996) has shown that this can be used as a means of providing SSDA functionality to GIS users. However, in the current offerings from vendors, the topological information is still seen as something to be handled internally, rather than as information which might be transferred between objects. The OpenGIS consortium has produced a specification of what it calls an Essential Model and it is to be hoped that this process will lead to the development of a standard way of providing information on the spatial relationships between spatial objects. However this process is at an early stage, with the only proposal which has reached any sort of fruition being the one which states how systems will deal with simple, geometrical forms such as lines and whole polygons. (Buehler and Mckee 1998, OpenGIS Consortium 2000)

There is a strong case for incorporating some of the techniques from systems such as SAGE directly into mainstream GIS software, and in this context we identify two particular types of functionality: the first is the linked windows or brushing facility. This is a common feature of the majority of systems written to explore different methods of providing ESDA functionality (Andrienko and Andrienko 1999, Haslett et al 1990, Cook et al 1996, Dykes 1996, Unwin et al 1996, Anselin and Bao 1997, Brunsdon 1998) and has been found to support a wide range of analytical operations. The second, which is related to the first, is the provision of graphical and other techniques which support spatial, as opposed to non-spatial, exploratory analysis. Features such as lagged boxplots and Moran plots are simple, intuitive means of exploring spatial data and regionalization allows the user to experiment with the effects of altering the spatial framework (Haining 1990). None require a detailed knowledge of spatial statistics in order to use them, but they do provide useful insights into the spatial variations which may exist in a set of data.and there would seem to be a strong case for adding them directly into mainstream GIS software.

Conclusions

There are a number of conclusions which are of potential interest to both the GIS community and GIS software vendors. It should be clear from this work, and the work of other software developers, that there are a range of techniques which will be of interest to a wide range of GIS users, who already have GIS databases of area-based data and wish to do more with than than simply draw maps or make routine queries. In particular, the provision of relatively simple graphical displays - boxplots, scatterplots and choropleth maps - in a graphical environment in which the windows are dynamically linked has been shown by many researchers to provide a very powerful tool for undertaking many forms of ESDA. What is more, experience suggests that such techniques are intuitively simple to use, and so will be accessible to a wide range of users. There is a clear potential here for GIS software vendors to add such functionality into standard GIS packages and this would be a major benefit to many users.

The more esoteric, specialist statistical tools of SDA are likely to be of interest only to a minority of users, although we believe that this is a significant minority. As indicated in the introduction, the range of science-based and policy-based research issues that call for rigorous SSDA is growing in both the public and private sectors. In the past, vendors have been reluctant to invest in developing such methods as part of their existing products (Maguire 1995) , and this may continue to be the case. The examples of SAGE and other systems (Cook et al 1996) have demonstrated that this type of functionality can be linked to GIS software, so it is not strictly necessary for GIS vendors to provide it within the GIS. What is needed to facilitate this linking process is a means of passing the topological information between the two applications. The current offerings of vendors in the area of software objects which can be used to build special-purpose GIS systems do not appear to support this, viewing the topological information as something to be encapsulated within the spatial objects. Work in the OpenGIS consortium does appear to have recognised the importance of information about the relationships between spatial features as important data in its own right, and it is to be hoped that this leads to better access to this type of information.

In summary, as the range of GIS users grow, the case for incorporating certain types of SSDA techniques directly into GIS also grows. Whilst the list of appropriate techniques will undoubtedly evolve, the linkage is also technically appropriate since GIS have many features that facilitate the implementation of SSDA techniques.

Acknowledgements

Our thanks to two anonymous referees whose comments did much to improve the quality of this paper. The authors wish to acknowledge receipt of ESRC research grant R000234470 “Developing spatial statistical software for the analysis of area based health data linked to a GIS". The ward boundaries shown in the figures are from the EDLINE Digitised Boundary Data, and are Crown and EDLINE copyright (ESRC and JISC purchase). The urban areas in Figure 3 are from the Bartholomew Map Data Great Britain 1;200000 database, Copyright Bartholomew Limited. (CHEST Bartholomew agreement). The population data are derived from the 1991 Census, Crown Copyright (ESRC purchase). The health data were provided by the Trent Region of the NHS Executive.

References

Andrienko G.L. and Andrienko N.V. (1999) Interactive maps for visual data exploration. Int.J.Geographical Information Science (in press)

Anselin, L. and S. Bao. (1997) “Exploratory spatial data analysis linking SpaceStat and Arc View.” In M.Fischer and A.Getis (eds) *Recent Developments in Spatial Analysis: Spatial statistics, behavioural modelling and neuro-computing.* Berlin, Springer-Verlag p35-59.

Anselin, L., Dodson, R.F. and Hodak, S. (1993) Linking GIS and Spatial Data Analysis in Practice. *Geographical Systems* 1 (1), 3-23.

Batty, M. and Yichun, X. (1994) Urban Analysis in a GIS Environment : Population Density Modelling using ARC/INFO in Fotheringham, S. and Rogerson, P. (ed) *Spatial Analysis and GIS,* Taylor and Francis, London, 189-220.

Brunsdon C (1998) Exploratory spatial data analysis and local indicators of spatial association with XLISP-STAT. *Journal of the Royal Statistical Society Series D The Statistician* 1998, Vol.47,No.Pt3, pp.471-484

Brunsdon C and M.E.Charlton (1996) Developing an exploratory spatial analysis system in XLisp-Stat . In D.Parker (ed), *Innovation in GIS 3*, 135-145. Taylor and Francis, London.

Buehler K. and McKee L. (1998) *The OpenGIS guide.* 3rd edition. [Online document] http://www.opengis.org/techno/guide/guide/Guide980629.pdf. [Visited 5th May 1999]

Burrough P.A. (1990) Methods of spatial analysis in GIS. *Int.J.Geographical Information Systems* 4(3),221-223.

Clayton D. and Kaldor J. (1987) Empirical Bayes esimates of age-standardized relative risks for use in disease mapping. *Biometrics* 43,671-681.

Cook, D., J.J.Majure, J.Symanzik and N.Cressie (1996) “Dynamics graphics in a GIS: exploring and analyzing multivariate spatial data using linked software” *Computational Statistics* Vol 11, 467-480.

Cressie, N. A. C. (1991) *Statistics for Spatial Analysis*, John Wiley and Sons, New York.

Ding, Y. and Fotheringham, S. (1992) The Integration of Spatial Analysis and GIS. *Computers, Environment and Urban Systems*, 16, 3-19.

Dykes J. (1996) Dynamic maps for spatial science: a unified approach to cartographic isualization. In D.Parker (ed), *Innovation in GIS 3*, 177-187. Taylor and Francis, London.

Fischer M., Scholten H. and Unwin D (1996) *Spatial Analytical Perspectives on GIS.* Taylor and Francis, London.

Fotheringham, A.S. and Rogerson, P. (1994) *Spatial Analysis in GIS.* Taylor and Francis, London.

Getis A. and Ord J.K. (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24, 189-206.

Goodchild M.J., Haining R.P., Wise S.M. and 12 others (1992) Integrating GIS and spatial data analysis : problems and possibilities. *Int.J.Geographical Information Systems* 6(5), 407-423.

Goodchild, M.G. (1987) "A spatial analytical perspective on geographical information systems" *International Journal of Geographical Information Systems,* Vol 1, p327-334.

Graham L. (1999) NT-based GIS rises to the occasion. GeoEurope 8(7),34-39.

Haining, R.P (1990) *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge University Press.

Haining R.P., Ma J. and Wise S.M (1996) The design of a software system for interactive spatial statistical analysis linked to a GIS. *Computational Statistics*, 11, 449-466.

Haining R.P., Wise S.M. and Ma J. (1998) Exploratory spatial data analysis in a geographic information system environment. *The Statistician* 47(3), 457-469.

Haining R.P., Wise S.M. and Ma J. (2000) Designing and implementing software for spatial statistical analysis in a GIS environment. *J.Geographical Systems* (in press)

Haslett, J, G. Wills and A.R. Unwin (1990) "SPIDER - an interactive statistical tool for the analysis of spatially distributed data." *International Journal of Geographical Information Systems,* Vol 4, 285-296.

Kehris E. (1990a) *Spatial Autocorrelation Statistics in ARC/INFO.* North West Regional Research Lab Report 16.

Kehris E. (1990b) *A Geographical Modelling Environment Built Around ARC/INFO.* North West Regional Research Lab Report 13.

Kendall, M.G. (1939) "The geographical distribution of crop productivity" *Journal of the Royal Statistical Society,* Vol 102, 21-48.

Kennedy S. (1989) The small number problem and the accuracy of spatial databases. in Goodchild M.F. and Gopal S. (eds) *The accuracy of spatial databases.* Taylor and Francis, London, 187-196.

Ma, J., R.P. Haining and S.M. Wise (1997) *SAGE users guide.* [Online document] Available from Sheffield Centre for Geographic Information and Spatial Analysis homepage http://www.shef.ac.uk/~scgisa

MacDougall E.B. (1992) Exploratory Analysis, Dynamic Statistical Visualisation and Geographic Information Systems. *Cartography and Geographic Information Systems* 19(4), 237-246.

Maguire D.J. (1995) Implementing spatial analysis and GIS applications for business and service planning. In P.Longley and G.Clarke (eds) *GIS for Business and Service Planning*, 171-191.

Masser I. (1988) The Regional Research Laboratory Initiative: a progress report. *International Journal of Geographic Information Systems* 2(11), 11-22.

NCGIA (1989) The research plan of the National Centre for Geographic Information and Analysis.. *International Journal of Geographic Information Systems* 3(2), 117-136.

OpenGIS Consortium (2000*) Open GIS Consortium Inc. – Frequently Asked Questions (FAQs).* [Online document] http://www.opengis.org/FAQs.htm#q9 [Updated 18th April 2000]

Openshaw S. (1984) *The modifiable areal unit problem.* Concepts and Techniques in Modern Geography 38. GeoAbstracts, Norwich.

Openshaw S. (1990) *A spatial analysis research strategy for the Regional Research Laboratory initiative.* Regional Research Laboratory Initiative Discussion Paper Number 3, Department of Town and Regional Planning, University of Sheffield.

Openshaw S. and Rao L. (1995) Algorithms for re-engineering 1991 census geography. *Environment and Planning* A, 27(3), 425-446.

Peuker, T.K., and N. Chrisman (1975) Cartographic Data Structures, *American Cartographer* 2(1):55-69.

Simon E. (1996*) Distributed Information Systems – from client-server to distributed multimedia.* McGraw Hill, London.

Tukey,J.W. (1977) *Exploratory Data Analysis.* Reading, Mass. Addison Wesley.

Umar, A. (1993) *Disturbed Computing and Client-Server Systems.* Prentice Hall, New York.

Unwin A., Hawkins G. Hofman H and Siegl B. (1996) Interactive graphics for data sets with missing values - MANET. *J.Computational and Graphical Statistics*,5,113-122.

Wise S.M. (1990) Evaluating GIS software for use in Higher Education. *Mapping Awareness* 4(7), 41-43.

Wise S.M., Haining R.P. and Ma.J. (1997) Regionalisation tools for the exploratory spatial analysis of health data In M.Fischer and A.Getis (eds) *Recent Developments in Spatial Analysis - Spatial Statistics, Behavioural Modelling and Neurocomputing,* Springer, Berlin

Wise, S and Haining, R (1991) The Role of Spatial Analysis in Geographical Information Systems. *Proceedings of the 3rd National Association for Geographic Information Conference*, 3.24.1-3.24.8.

Table captions

Table 1 Operations needed to support ESDA functionality

Figure Captions

Figure 1 The architecture of SAGE

Figure 2 Screenshot of a SAGE session. The main user interface is via the menu options on the tabular view of the data. The map shows rates of Long Term Limiting Illness for wards in the Trent region on the NHS executive and the same rates are shown in the boxplot. The histogram shows total population in each ward. For an explanation of the highlighted features, see the text.

Figure 3 Location of the Trent region of the NHS. The larger map shows the ward boundaries, urban areas and places named in the text.

Figure 4 Relationship between Long Term Limiting Illness (LLTI) and Material Deprivation (measured using the Townsend Index) is shown on the scatterplot. The map shows LLTI rates, with the six outliers from the scatterplot identified by cross-hatching.

Figure 5 The detection of clusters using SAGE. The histogram shows values of the Getis-Ord statistic, a measure of local autocorrelation. Areas with high positive values have been selected, and are shown on the map by cross-hatching.

	High Level	Fundamental level
Data management		• Manage Spatial data • Manage Attribute data • Create **W** matrix
Data manipulation	• Calculation of rates • Statistical tests • Classification and regionalisation • Selection of subsets of data • Editing **W** matrix	• Calculations on attributes • Classification and regionalisation algorithms • Spatial element selection • Database queries • Polygon dissolve
Data display	• Map drawing • Tabular data display • Statistical graphs • Linked windows	• Graphical display • Window management

Table 1 Operations needed to support ESDA functionality

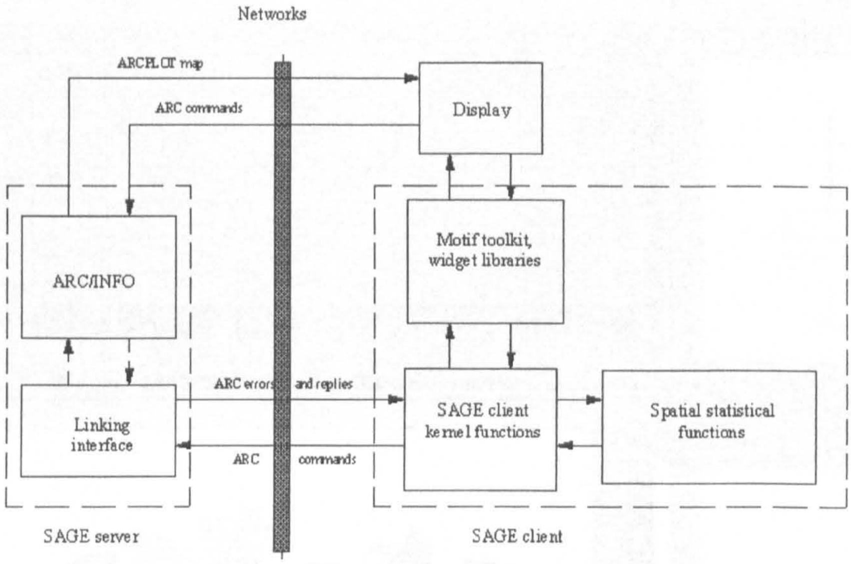
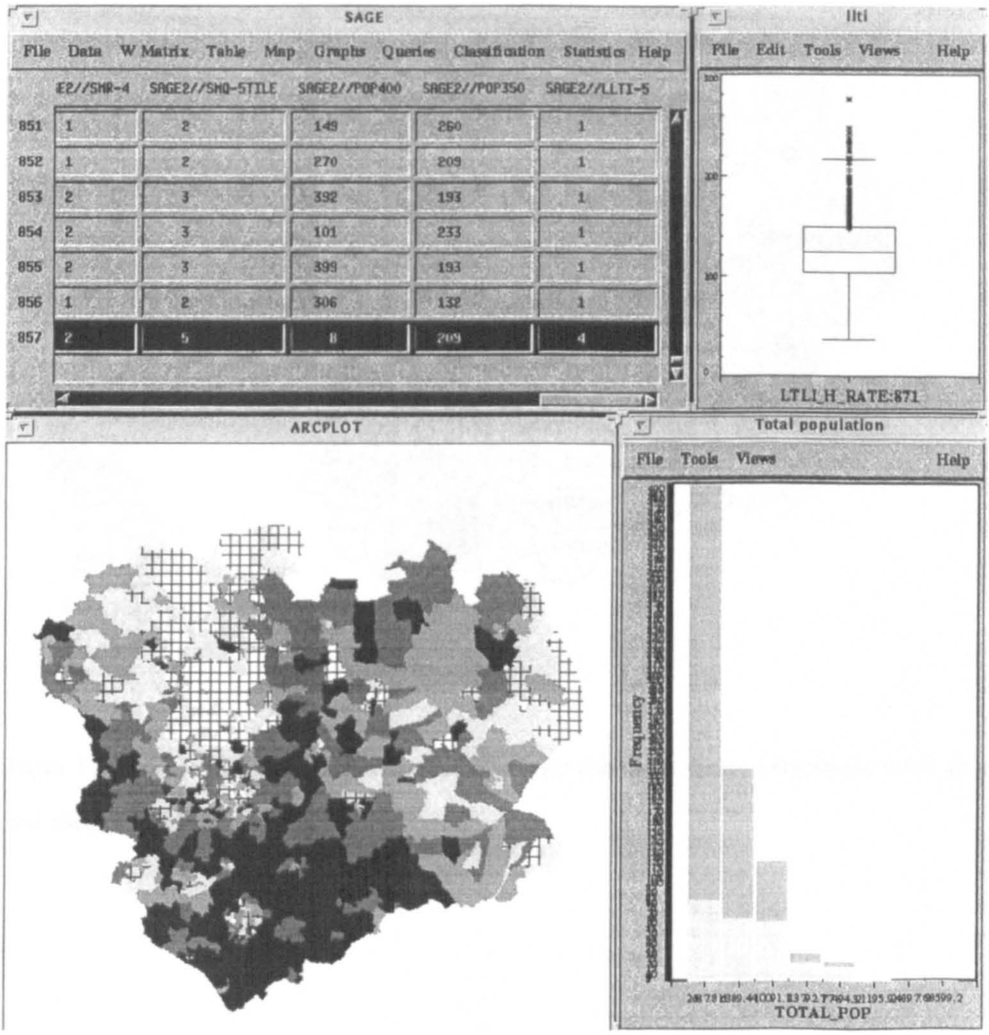


Figure 1 The architecture of SAGE


SAGE
File Data W Matrix Table Map Graphs Queries Classification Statistics Help
llti
File Edit Tools Views Help
LTLI_H_RATE:871
ARCPLOT
Total population
File Tools Views Help
Frequency
TOTAL_POP

	E2//SMR-4	SAGE2//SMQ-5TILE	SAGE2//POP400	SAGE2//POP350	SAGE2//LLTI-5
851	1	2	149	260	1
852	1	2	270	209	1
853	2	3	392	193	1
854	2	3	101	233	1
855	2	3	399	193	1
856	1	2	306	132	1
857	2	5	8	209	4

Figure 2 Screenshot of a SAGE session. The main user interface is via the menu options on the tabular view of the data. The map shows rates of Long Term Limiting Illness for wards in the Trent region on the NHS executive and the same rates are shown in the boxplot. The histogram shows total population in each ward. For an explanation of the highlighted features, see the text.

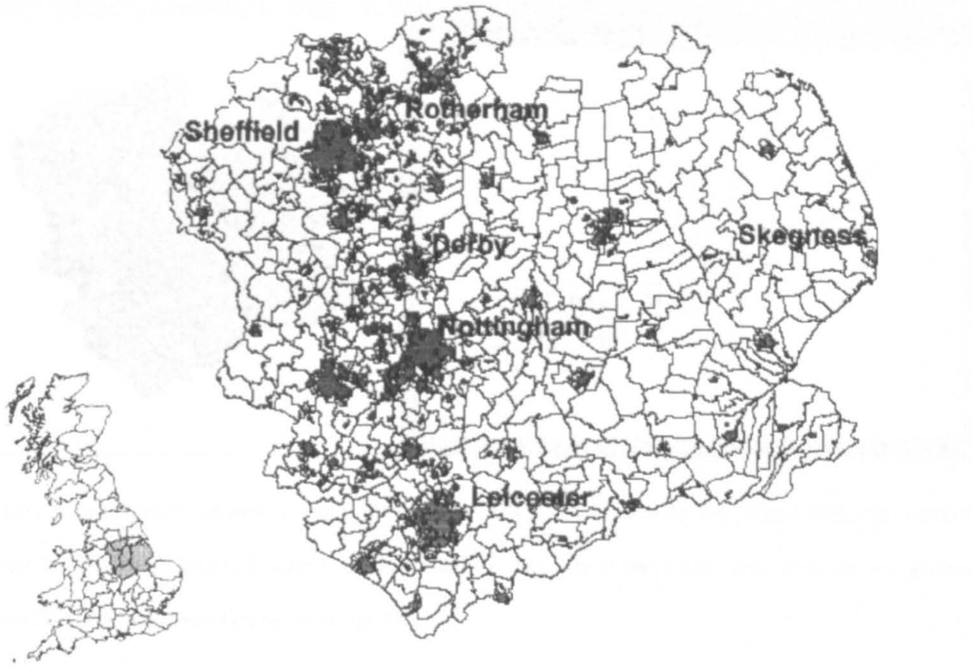


Figure 3 Location of the Trent region of the NHS. The larger map shows the ward boundaries, urban areas and places named in the text.

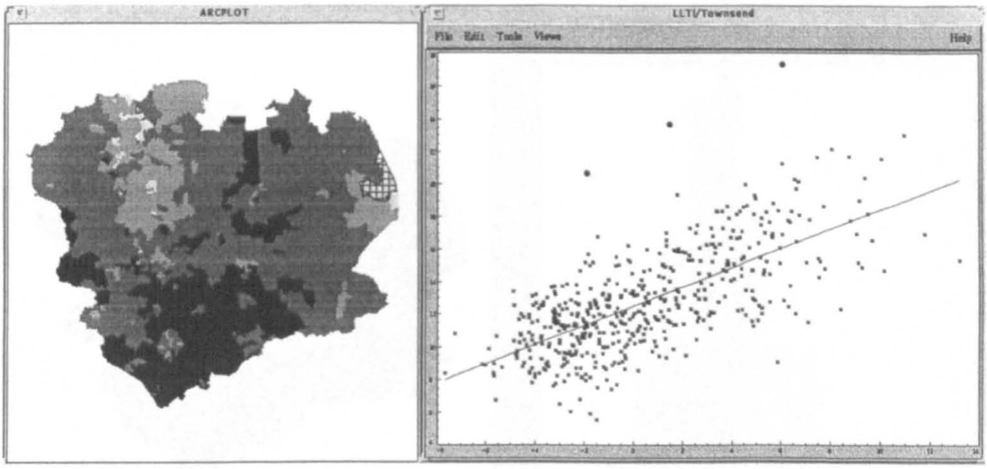


Figure 4 Relationship between Long Term Limiting Illness (LLTI) and Material Deprivation (measured using the Townsend Index) is shown on the scatterplot. The map shows LLTI rates, with the six outliers from the scatterplot identified by cross-hatching.

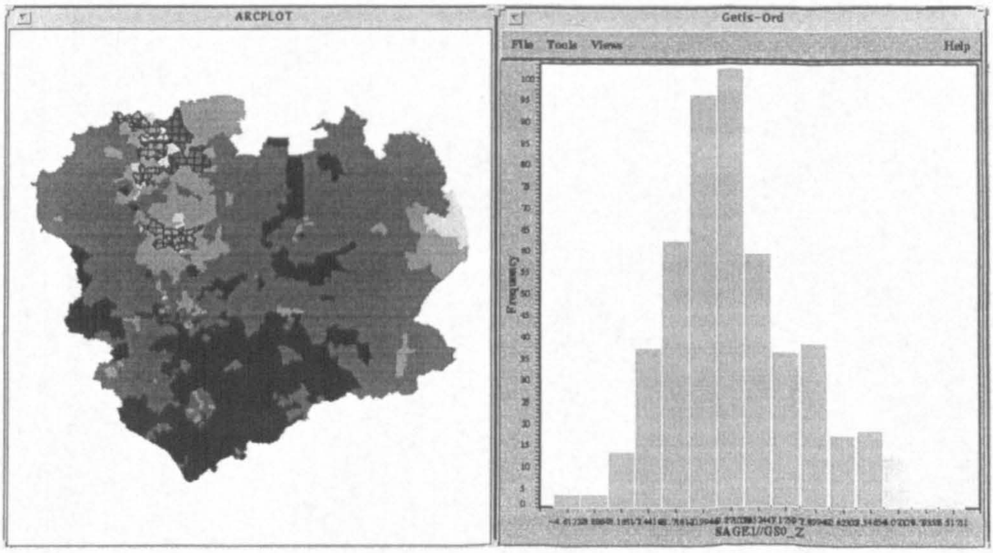


Figure 5 The detection of clusters using SAGE. The histogram shows values of the Getis-Ord statistic, a measure of local autocorrelation. Areas with high positive values have been selected, and are shown on the map by cross-hatching.