J Geograph Syst (2000) 2:287–306

Journal of Geographical Systems

Seamless integration of spatial statistics and GIS: The *S-PLUS* for *ArcView* and the *S+Grassland* Links

Shuming Bao[1], **Luc Anselin**[2], **Doug Martin**[3], **Diana Stralberg**[1]

[1] University of Michigan, Suite 3630, 1080 S. University Ave., Ann Arbor, MI 48109-1106, USA (e-mail: sbao@umich.edu)
[2] University of Illinois at Urbana-Champaign, 326 Mumford Hall, Urbana IL, 61801-3608, USA (e-mail: anselin@uiuc.edu)
[3] MathSoft, Inc., 1700 Westlake Ave N., Seattle, WA 98109-3044, USA

Received: 14 January 1999 / Accepted: 11 May 1999

Abstract. The extension of the functional capacity of geographic information systems (GIS) with tools for statistical analysis in general and exploratory spatial data analysis (ESDA) in particular has been an increasingly active area of research in recent years. In this paper, two operational implementations that combine the functionality of spatial data analysis software with a GIS are considered more closely. They consist of linkages between the *S-PLUS* software for data analysis and two different GIS implementations, the *ArcView* desktop system, which is mostly vector-oriented, and the primarily raster-based *Grassland* open GIS environment. We emphasize conceptual and technical issues related to the software implementation of these approaches and suggest future directions for linking spatial statistics and GIS.

Key words: ESDA, *S-PLUS*, *ArcView*, *Grassland*, spatial statistics

JEL classification: C60, C63, C88

1 Introduction

With the increasing sophistication and accessibility of commercial geographic information systems, desktop software applications like ESRI's *ArcView* GIS have become common tools for viewing and analyzing many aspects of spatial

S-PLUS for ArcView and the *S+Grassland* Link are products of MathSoft, Inc. Research reported in this paper was supported by MathSoft, Inc. and funded by a U.S. Army SBIR project. Shuming Bao was involved in the development of the *S+Grassland* Link and the initial beta version of *S-PLUS for ArcView* while a Research Scientist at MathSoft, Inc. from 1996 to 1997. Luc Anselin's research was supported in part by Grant SBR-9410612 from the US National Science Foundation. The opinions expressed in this paper are solely those of the authors and do not imply an endorsement by MathSoft, Inc. Special thanks to Juergen Symanzik for his valuable comments on an earlier version of the paper.

data. In spite of these advances, however, most GIS packages still offer little in the way of statistical analysis, and state-of-the-art spatial statistical methods, in particular, have not been incorporated into their commonly provided spatial analytical toolboxes.

Nevertheless, the integration of spatial statistical methods and GIS has been an active topic of research in both the academic and commercial GIS communities, as evidenced by a growing number of publications on the topic, starting with Goodchild (1987, 1992), and, more recently, including among others, Fischer and Nijkamp (1993), Fotheringham and Rogerson (1994), Painho (1994), Fischer et al. (1996) and Fischer and Getis (1997). Several conceptual outlines have been formulated dealing with the set of techniques that should be incorporated and how the linkage between GIS and statistical methods should be established, for example in Openshaw (1991), Anselin and Getis (1992), Goodchild et al. (1992), Bailey (1994), Haining (1994), Openshaw and Fischer (1995), and Anselin (1998). Linkage strategies can be conceptualized as ways to combine the "traditional" spatial analysis functionality of the GIS (e.g., spatial queries, buffering, overlay) in a "GIS Module" with spatial statistical and data analysis methods in a "Spatial Data Analysis Module" (Anselin et al. 1993). One suggested approach to combining these two modules into an overall framework for spatial analysis consists of incorporating elements of one into the other (a so-called encompassing strategy), such as the addition of GIS capabilities to a statistical software package, or the extension of a GIS with statistical functionality. The latter is typically not included in the standard software release but is made possible by taking advantage of macro or script languages supported in the GIS software. Familiar examples include the use of AML (Arc Macro Language) for *Arc/Info* and the Avenue script language for *ArcView* to extend the functionality of a GIS with EDA tools (Batty and Xie 1994; Anselin and Bao 1997; Zhang and Griffith 1997) or descriptive spatial autocorrelation statistics (Ding and Fotheringham 1992; Bao et al. 1995; Can 1996). The advantage of such an approach is that the added functions are fully integrated into the familiar GIS data model and user interface. However, the scripting environments are somewhat limited in terms of the size of data sets that can be handled, and often seriously deficient in terms of speed.

Another strategy consists of developing an efficient linkage between existing commercially available GIS and statistical software packages. A number of taxonomies have been suggested to implement such linkages, for example loose coupling vs. seamless coupling, unidirectional vs. bi-directional links, and static vs. dynamic links (for reviews, see, e.g., Anselin and Getis 1992; Goodchild et al. 1992; Anselin et al. 1993; Symanzik et al. 1994, 1996, 1998; Anselin 1998). There are now several examples of such approaches, developed in academic environments (e.g., Farley et al. 1990; Williams et al. 1990; Symanzik et al. 1994, 1996, 1997; Anselin and Bao 1997; Bao and Anselin 1998), as well as in the commercial sector (e.g. the S-PLUS/GIS links, MathSoft 1996a).

In this paper, we review some recent extensions that deal with the integration of the statistical and spatial statistical functionality of the *S-PLUS* software (MathSoft 1997a) with the *ArcView* (ESRI 1996) and *Grassland* (L.A.S. 1996a) GIS, respectively. Both approaches are based on a concept of seamless linkage between the software packages, but they are implemented by means of different techniques. In the remainder of the paper, we briefly de-

scribe the overall architecture, linkage mechanism and operational implementation for each of these approaches and provide simple illustrations by means of an empirical example. We close with a comparison of the relative merits of the linkage strategies and some thoughts on future developments.

2 *S-PLUS for ArcView*

The link between *S-PLUS* and *ArcView* is an extension and implementation of the prototype suggested in Anselin et al. (1993). The *ArcView* GIS serves as the visualization engine, while *S-PLUS* is used for spatial data analysis. The main objective of the interface is to provide a comprehensive and efficient tool for spatial data analysis that can be accessed from within a GIS environment.

S-PLUS (MathSoft 1997a) is a modern object-oriented language and system for multi-purpose data analysis with over 2,000 functions. It provides powerful capabilities for graphical data analysis and statistical modeling. The optional module *S+SpatialStats* (MathSoft 1996b) provides additional analytical functionality to handle geostatistical data, lattice data, and point data, such as variogram estimation, kriging, Moran's I statistic, and the estimation of spatial regression models.

ArcView is currently one of the most popular desktop GIS software environments, primarily geared to the manipulation of vector data. *ArcView* can easily be customized by means of scripts written in the object oriented Avenue script language (supported by *ArcView* 2.1 and higher) and has recently been extended with optional modules for the analysis of raster data (Spatial Analyst extension), network data (Network Analyst extension) and three-dimensional data (3-D Analyst extension).

2.1 Architecture

S-PLUS for ArcView is characterized by the following features: (1) a seamless integration of *S-PLUS* with *ArcView*; (2) access to the full range of *S-PLUS* functions; and (3) a graphical user interface that hides the full complexity of the linkage mechanisms.

The interface between *ArcView* and *S-PLUS* is based on a close coupling approach by means of a bi-directional linkage, as illustrated by the data flow chart in Fig. 1. Data are passed between the two environments using an automation technique. Specifically, attributes and spatial information, along with the *S-PLUS* commands for analysis, are moved from *ArcView* to *S-PLUS*, and location-specific results are passed back to *ArcView* for visualization. Types of spatial information include *X-Y* coordinates (e.g. points or polygon centroids), as well as topological information on the spatial arrangement of selected points or areal units (such as spatial neighbor contiguity). Spatial neighbor objects must be constructed explicitly from the locational information in *ArcView* shape files, and are stored in the *S-PLUS* workspace. Location-specific results, including estimated values from kriging, spatial regression, and local spatial statistics, can be integrated with *ArcView* tables. *S-PLUS* graphs are incorporated into layouts, for output with other *ArcView* elements in the usual fashion.

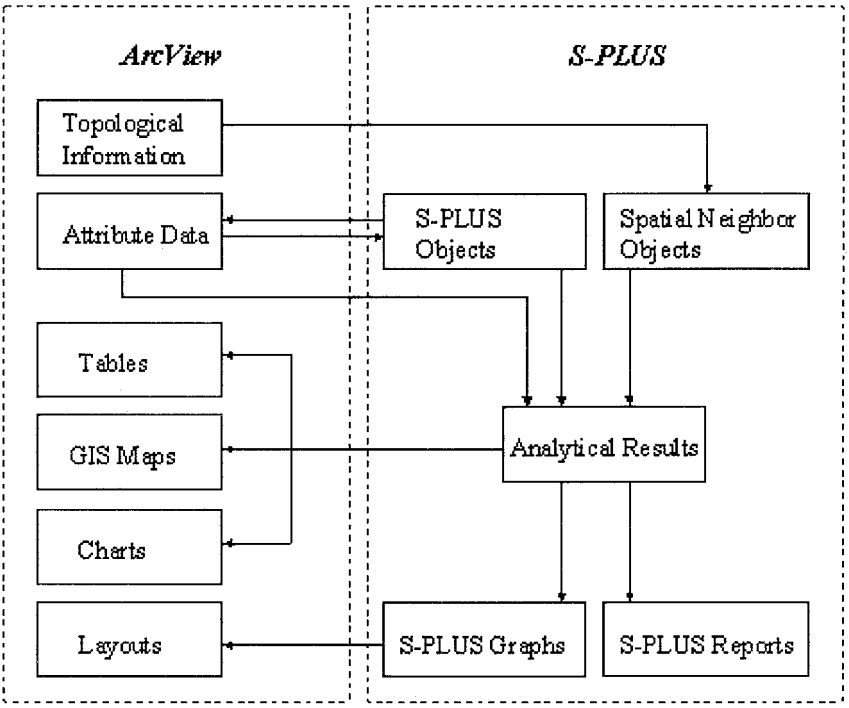

Fig. 1. Chart of data flows in *S-PLUS for ArcView*

2.2 Linkage mechanism

S-PLUS for ArcView is implemented as an "extension" to the standard *ArcView* GIS software. The customary user interface is augmented with two new menus and user interaction is carried out by means of a set of dialogs, constructed using ESRI's *ArcView* Dialog Designer (ESRI 1997). Each dialog invokes a number of Avenue scripts and *C* programs contained in a DLL (Dynamic Link Library). Once the *S-PLUS* extension is loaded into *ArcView*, the *S-PLUS* window is launched and a connection between *S-PLUS* and *ArcView* is established.

The linkage between *S-PLUS* and *ArcView* is dynamic and bi-directional. Data transfer occurs primarily between *ArcView* shape files and *S-PLUS* data frame objects. Within *ArcView*, selected fields and records are extracted from a shape file and exported to *S-PLUS* as a data frame object. Conversely, data from an *S-PLUS* data frame object can be imported into *ArcView* as a generic table and joined with a selected theme's attribute table. Spatial weight matrices are constructed using scripts that exploit the geo-locational information from *ArcView* shape files and are saved in *S-PLUS* as spatial neighbor objects, which can then be used in spatial statistical analyses such as spatial autocorrelation, local spatial association, and spatial regression. Similarly, results are saved in *S-PLUS* as data frame objects that can be imported into *ArcView* for visualization. Summary reports from *S-PLUS* analyses are

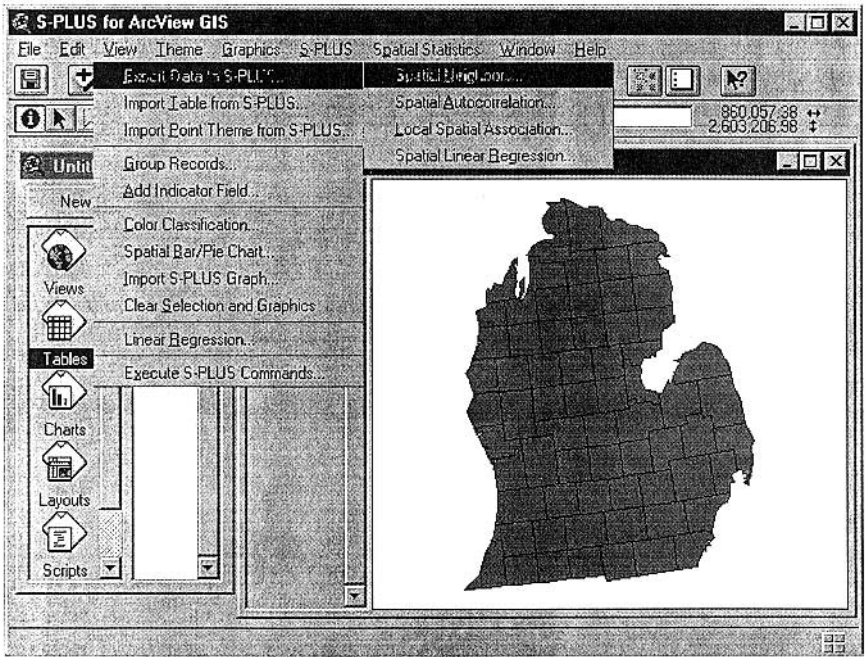

Fig. 2. *S-PLUS for ArcView* interface

output to ASCII text files and displayed in Notepad, a standard text editor for Microsoft Windows.

2.3 Operational implementation of S-PLUS for ArcView

S-PLUS for ArcView is implemented primarily in the *ArcView* environment, by means of a set of Avenue programs that call special purpose functions included in a DLL (Dynamic Link Library) and that are associate with menu items (see Bao and Martin 1997). Data export from *ArcView* to *S-PLUS* is carried out by packaging it into a SafeArray which can then be sent to *S-PLUS* via Automation. Once *S-PLUS* receives this data, it is written out to the current *S-PLUS* working directory. Conversely, data import from *S-PLUS* is implemented by creating a text file which is then added into *ArcView* as a new table and linked with the current *ArcView* attribute table. Two additional menus (*S-PLUS* and Spatial Statistics) with a total of fifteen functions are added to the standard View window.

The *S-PLUS* menu consists of eleven functions divided into five categories: (1) data transfer between *ArcView* and *S-PLUS*; (2) auxiliary manipulation of spatial information; (3) spatial data visualization; (4) linear regression; and (5) executing *S-PLUS* commands from *ArcView*.

The first group of menu functions implements simple data transfer between *ArcView* and *S-PLUS*. To export data from *ArcView* to *S-PLUS*, users select a subset of records from the attribute table associated with the current theme or coverage. The selected fields and records are extracted from the shape files,

and transferred to *S-PLUS* as a data frame object. Geographical information (such as *X-Y* coordinates for a point coverage and centroids for a polygon coverage) may also be included in the exported object. To import data from *S-PLUS*, users can select one or more columns from a given *S-PLUS* data frame. The selected data are imported into *ArcView* as a new table. Finally, an *S-PLUS* data frame containing *X-Y* coordinates may be imported into *ArcView* as a new point theme.

The second group of menu functions is used to create new character fields for selected records from the current theme table. With the Group Selected Records function, users can categorize records into several groups identified by a new field. The newly-created field is joined with the current theme table, and can be exported to *S-PLUS* for further analysis. Similarly, the Add Indicator Field function allows numeric fields to be created, which can then be used in *S-PLUS* to subset observations or as a dummy variable in statistical analysis.

The third group of menu functions is used for visualization of spatial data and analytical results. Variables from *S-PLUS* data frame objects may be displayed in a GIS map using the Color Classification and Spatial Bar/Pie Chart options. These functions are especially useful for comparing analytical results, such as fitted values and residuals from classical or spatial regression analyses. In addition, the rich graphical features of *S-PLUS* can be accessed with the Import Graph function. *S-PLUS* graphs (e.g. histograms, boxplots, variograms and various 3-D graphs) can be imported into an *ArcView* layout, where they may be combined with *ArcView* maps and charts for output.

The Linear Regression function allows users to build a regression equation using variables from either the current *ArcView* theme or an *S-PLUS* data frame. A summary report of the regression is saved in a text file, and predicted values and residuals are saved in an *S-PLUS* data frame object, which can easily be joined to the current theme table.

Finally, the Execute *S-PLUS* Commands feature provides ready access to all *S-PLUS* commands from the *ArcView* environment. All *S-PLUS* objects are listed on this dialog window, and users can type in *S-PLUS* commands on the command line, or make use of several auxiliary button functions.

The Spatial Statistics menu contains four functions: (1) Spatial Neighbor; (2) Spatial Autocorrelation; (3) Spatial Association; and (4) Spatial Regression [See Bao and Martin (1997) and MathSoft (1998) for technical details].

The Spatial Neighbor function is designed to construct spatial neighbor objects (spatial weights) for *ArcView* shapefile data, to be used in *S-PLUS* spatial statistics and modeling. Spatial weights must be defined externally and saved in text files so that they can be imported into *S-PLUS* neighbor objects. The spatial weights can be constructed by using topological information (adjacency criteria) or geo-spatial distance (distance criteria) by means of Avenue's buffering and spatial query functions. Several options are provided, such as First Order Neighbor Weights, Adjusted First Order Neighbor Weights (a combination of adjacency and distance criteria), and Higher Order Neighbor Weights. In addition, the Spatial Neighbor function provides several distance-based methods for both point and polygon shapefiles. The user can specify distance values, using border-to-border or centroid-to-centroid measurement options. The distance units are specified in the *ArcView* properties dialog.

The Spatial Statistics menus provide direct access to *S+SpatialStats*

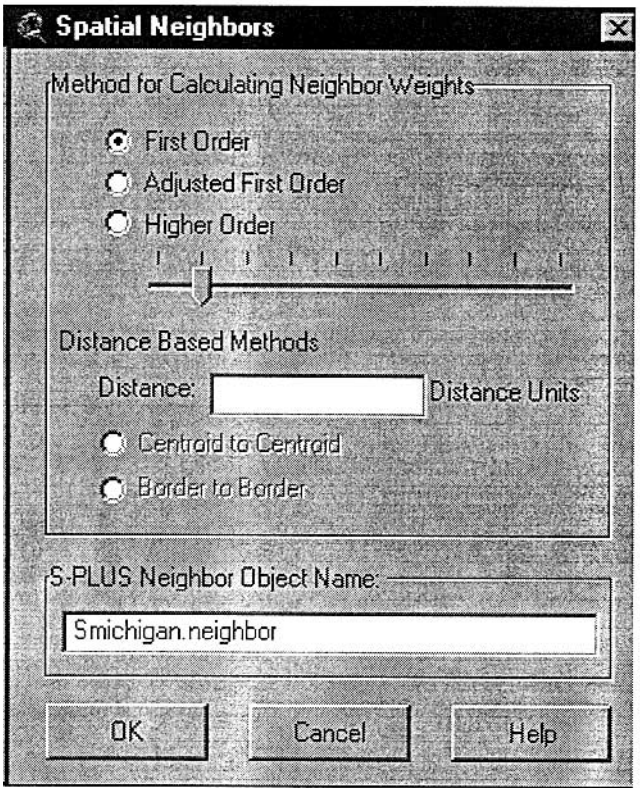

Fig. 3. Defining a spatial weight matrix

functions such as such as spatial autocorrelation (Moran's *I* and Geary's *C*), spatial association (General Local Indicators of Spatial Association – GLISA), and spatial linear regression. Spatial association options include General Local Moran and Local Geary by Bao and Henry (1996), which are derived from the LISA statistic by Anselin (1995). Spatial linear regression include three types of spatial error models (Cressie 1993): SAR – *Simultaneous Autoregressive Model* (Whittle 1954), CAR – *Conditional Autoregressive Model* (Bartlett 1971; Besag 1974), and MA – *Moving Average Model* (Cliff and Ord 1981). The variables can be selected from either an *ArcView* theme table or an *S-PLUS* data frame object, but a spatial neighbor object must have been pre-defined and be consistent with the selected variables for those spatial statistics. The summarized results are output to text files and the estimates are saved in *S-PLUS* objects that can then be joined with an *ArcView* theme table for map visualization.

2.4 Illustration

Large-scale ecological studies often have an inherent spatial component, given that the independence of sampling locations throughout a landscape generally

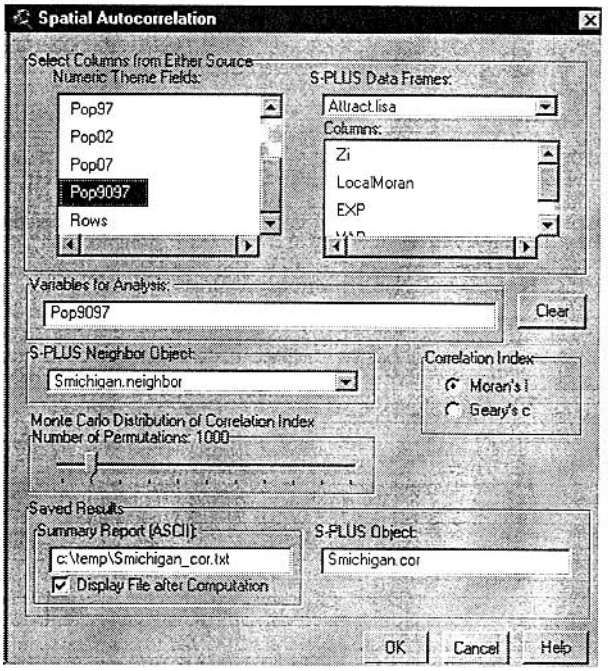

Fig. 4. Estimating spatial autocorrelation

cannot be assumed. In such situations, an array of spatial data analysis and modeling tools are needed to model the interactions between spatially-distributed variables, while controlling for this geospatial component. This constitutes an ideal area of application for an integrated GIS-spatial statistics framework.

To illustrate the application of the *S-PLUS for ArcView* interface, we briefly describe a study of the effects of landscape-level urbanization patterns on breeding bird abundance (see also Bao et al. 1998). Located in the Santa Monica Mountains, a coastal mountain range experiencing urban encroachment by the surrounding metropolitan Los Angeles, the study area represents a fragmentation gradient, increasing in urbanization from west to east (Fig. 7). The Wrentit, a common chaparral bird, was selected for closer examination due to its high abundance levels and strong association with the chaparral vegetation type. In 1997, a randomly-selected set of chaparral-vegetated sites across the landscape was censused for birds and georeferenced with Global Positioning System (GPS) readings. Using a 1000 m radius, the urbanization proportion surrounding each site (point) was calculated from a GIS land use coverage and linked with the bird census data. The resulting point coverage was then converted to an *ArcView* shape file for analysis with *S-PLUS for ArcView*.

An important part of the exploratory analysis was to check for spatial autocorrelation in Wrentit abundance. With the Spatial Neighbor function provided by *S-PLUS for ArcView*, spatial weights were calculated for the set of points at neighborhood distances of 500 m, 1000 m and 2000 m, using the

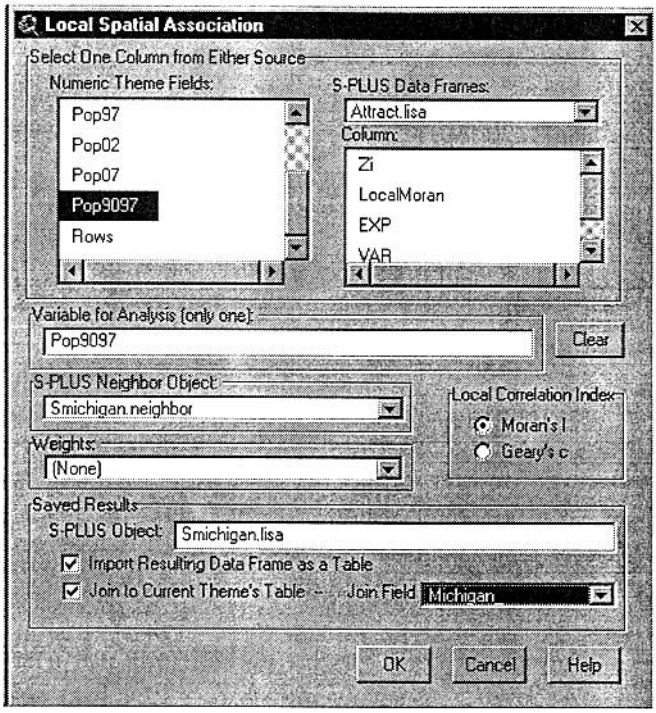

Fig. 5. Estimating local spatial statistics

border-to-border distance option. The Spatial Autocorrelation function was then used to test for spatial dependence of Wrentit abundance, using Moran's I statistic. For each neighborhood distance used, these tests indicated a highly positive and significant degree of spatial autocorrelation. In order to visualize the nature of this spatial autocorrelation, an empirical variogram was constructed from the Wrentit abundance data, using the *S-PLUS* Spatial Statistics module, which is accessed from *ArcView* via the Execute *S-PLUS* Commands feature. The resulting variogram (Fig. 8) increases somewhat linearly with distance, indicating the presence of a large-scale spatial trend or gradient, and suggesting non-stationarity.

Given the increase in urbanization from west to east, it was thought that the spatial trend in wrentit abundance might be attributable to surrounding urbanization levels. To assess the effect of urbanization (URB1000) on Wrentit abundance (WREN), an ordinary least-squares (OLS) regression model was estimated, using the Linear Regression function in *S-PLUS for ArcView*. The results indicated a statistically significant negative linear relationship between WREN and URB1000 (regression coefficient -5.86, t-value -6.19, $R^2 = 0.269$), but the question remained as to whether the residuals (the model "error" term) from this regression exhibit any spatial correlation. The empirical variogram for OLS residuals (Fig. 9) suggests an initial peak around 2000 meters (gamma ≈ 2.0), and a maximum near 9000 meters (gamma ≈ 3.0), after which the function steadily decreases. Depending on the underlying model, a theoretical variogram for the OLS residuals would have a

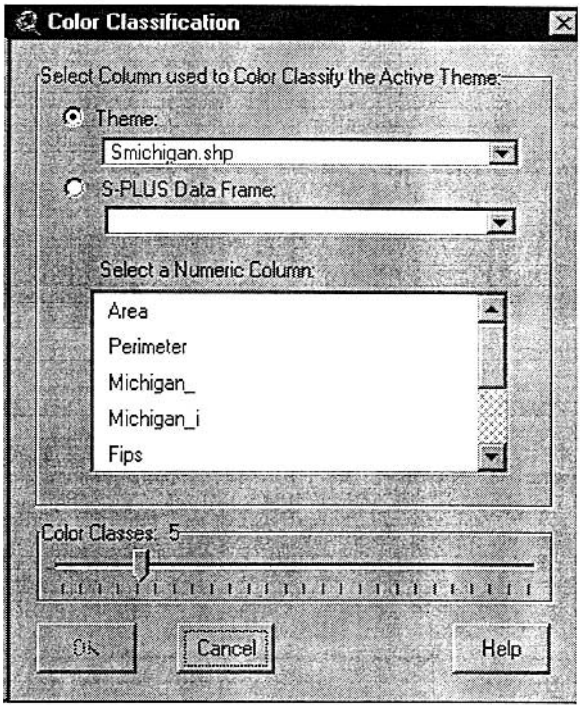

Fig. 6. Displaying a classified map

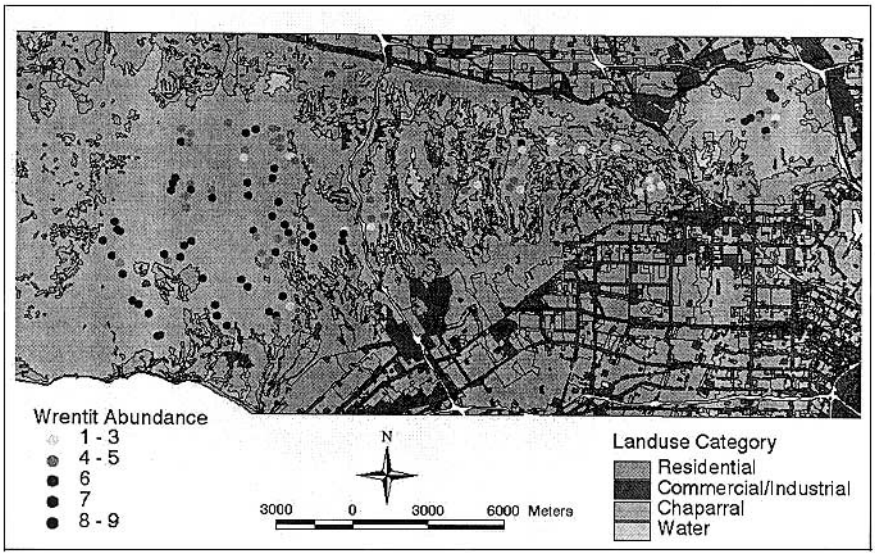

Fig. 7. Santa Monica Mountains study area. Wrentit abundance overlaid on landuse

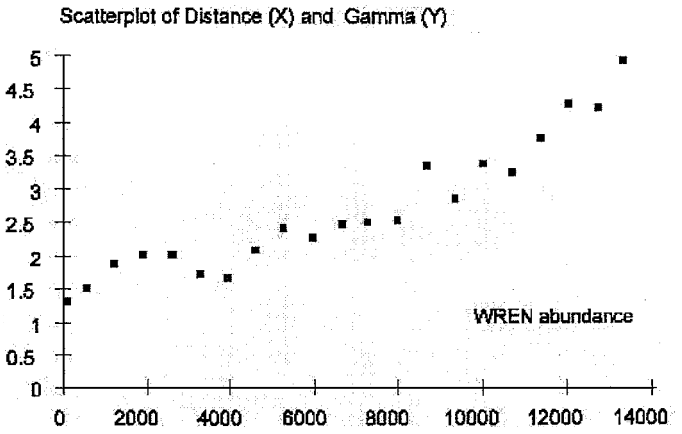

Fig. 8. Variogram of wrentit abundance

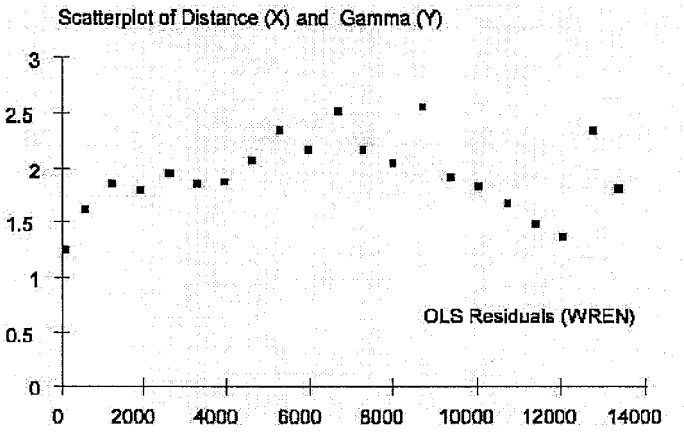

Fig. 9. Variogram of residuals from OLS regression of wrentit abundance on urbanization

sill between 2.0 and 3.0, and a range falling within 2000 and 9000 meters. The presence of a sill indicates that Wrentit abundance not explained by urbanization (i.e., the model error term) is stationary – i.e., constant throughout the study area. The strong correlation of residuals in the initial distance range (0–2000 meters) suggests that spatial neighborhood distances of up to 2000 meters are most appropriate.

Using the Spatial Linear Model (SLM) option provided by *S-PLUS for ArcView*, the spatial variability not encompassed by URB1000 was incorporated statistically, using a spatial autoregressive specification for the error term based on a 2000 m spatial weight. Regression coefficients for this model were similar to the OLS model coefficients (regression slope of -5.782, t-value of -5.39), with an autoregressive parameter of 0.045 (for non-row standardized spatial weights), significant at $p < 0.01$.

In summary, the statistical analysis and modeling illustrated the usefulness

of an integrated GIS and spatial analysis framework to discover the complex spatial patterns in the relationship between Wrentit abundance and urbanization.

3 The *S+Grassland* Link

As with *S-PLUS for ArcView*, the main objective behind the *S+Grassland* Link is to combine the functionality of the *Grassland* GIS with the statistical modeling and visualization capabilities of *S-PLUS* by means of a seamless integration or close coupling. In particular, the goal is to develop a product with a graphical user interface for statistical analysis of multilayer geo-spatial data such as raster, vector and associated attribute data, rather than being limited to a single layer.

Grassland (L.A.S. 1996a) is a commercial, Windows-based version of the *GRASS* GIS, a command line-based GIS developed by the US Army CERL (USACERL 1993). The *Grassland* Graphical User Interface (GUI) was developed in the Sun Microsystems *Tcl/Tk* (Tool Command Language/Tool Kit) scripting language (Ousterhout 1994), a modular, open-architecture public domain programming language. All *Grassland* user interfaces are Tcl/Tk scripts so that the user can modify and customize their application windows. Each *Grassland* infrastructure component has a corresponding *Tcl* extension loaded dynamically at boot time. *Grassland* provides a tool for the manipulation of raster and point data and several standard spatial analysis functions such as overlay, buffer zone analysis, and terrain analysis. In addition, it contains a powerful built-in map builder, such that the user can drag and drop overlays onto the map background to display rivers, tree cover, geological layers, topography, wildlife populations or any other GIS element to create fast, customized visual decision-assistance tools.

3.1 Architecture

The *S+Grassland* Link is characterized by the following features: (1) a seamless integration of *S-PLUS* functions with *Grassland*; (2) an application of statistical analyses to both raster and vector data; and (3) an application of statistical analyses to multiple data layers.

In contrast to most other GIS software, *Grassland* implements interoperability by providing an URL (Uniform Resource Locator) connection to geospatial data sources using the *OGDI* (Open Geospatial Datastore Interface) technique for standardized access and transfer of geospatial data (L.A.S. 1996b). Provided as a *C* utility library by L.A.S., *OGDI* is an application programming interface (API) that resides between a GIS software package (the application) and various geodata products, to provide standardized geospatial access methods. *OGDI* uses a client/server architecture to facilitate the dissemination of spatial data over the Internet, and a driver-oriented approach to facilitate access to a variety of spatial data and formats. It handles the most important spatial data integration needs such as conversion of various formats into a uniform data model, adjustment of coordinate systems and cartographic projections, and retrieval of geometric and attribute data. Specifically, geospatial data are obtained by means of a specific data driver and moved

into buffer memory, which can then be accessed through the utility functions provided by the *OGDI* library. The *OGDI* library in *Grassland* therefore allows an application to connect to any geographic "datastore" (such as *GRASS*, *Arc/Info*, *ADRG*, *VRF*, and *DTED*) in a transparent manner.

The *S+API* (Application Programming Interface), developed by MathSoft and provided as a *C* utility library starting with *S-PLUS* 4.0 (MathSoft 1997b, 1997c), allows *S-PLUS* to function as a server for a wide variety of clients. The *S+API* provides a set of functions to allow simultaneous and uniform access to *S-PLUS* from multiple applications, possibly residing on heterogeneous machines that are connected via a network. It provides three categories of functions: control, messaging, and access. Control functions allow for administrative control of the client/server connection. Messaging functions pass commands and data to and from the *S-PLUS* server. Access functions make it easy for client programs to build *S-PLUS* commands and to fill and retrieve data from complex *S-PLUS* data structures. Clients can send data or commands to a specific *S-PLUS* server and receive the results of the requested calculations.

Built on the *Tcl/Tk*, *OGDI* and *S+API* techniques, the linkage between *Grassland* and *S-PLUS* is a close coupling with a bi-directional conversation. Selected data are exported from *Grassland* to *S-PLUS*, and *S-PLUS* objects and functions are accessed directly from within the *Grassland* environment.

3.2 Linkage mechanism

The *S+Grassland* Link is accomplished by means of several DLL utility functions. These include functions to export data and pass commands from *Grassland* to *S-PLUS*, and an *S-PLUS* data driver to access *S-PLUS* objects from within *Grassland*. The *GRASS* data can be retrieved and transformed into the *S-PLUS* object structure via the *S+API*. The *S-PLUS* data driver is composed of *Tcl/Tk* script programs and *C* routines, which allows a URL connection to be established between *Grassland* and the *S-PLUS* workspace from within the *Grassland* Library window. Once a URL connection is established, *S-PLUS* objects and results from statistical analyses can be accessed directly from *Grassland*, which yields a truly dynamic link (for technical details, see Bao 1997).

In *Grassland*, geo-spatial data types include Raster, Area, Point, Line, and Text. To make the *S-PLUS* object structure compatible with this GIS, several new types of spatial objects (or geo-objects) needed to be defined, such as Raster (an integer matrix object that contains the raster map information), Point (a data frame that contains the *X* and *Y* coordinates information and other attributes), Text (a single vector – *.out – that contains the attribute information or the results from an *S-PLUS* analysis), and Probability Map (pmap, an integer matrix object – *.map – that contains the predicted probabilities from a generalized linear regression analysis).

This interface also introduces the concept of an *S-PLUS* Geospatial Workspace (SGW), a subdirectory containing a series of *S-PLUS* workfiles for spatial objects. Each SGW contains a special *S-PLUS* object (_GRS) that holds the parameters for the geospatial reference information such as the projection and global boundaries. This special *S-PLUS* object is automatically created by the *S+Grassland* interface. In addition to this

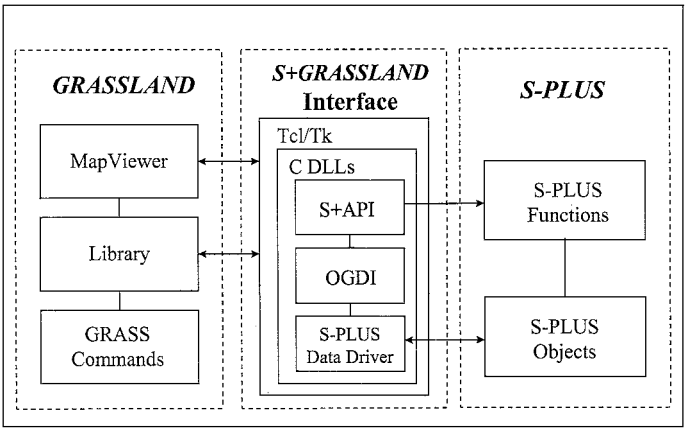

Fig. 10. System diagram of the *S+Grassland* Link

global geospatial reference object, each individual *S-PLUS* spatial object also has an associated object (*.grs) that stores individual geospatial reference information.

3.3 Operational implementation of the S+Grassland Link

The *S+Grassland* link is integrated within the standard *Grassland* windows user interface. All new functions needed to establish the *S+Grassland* Link are implemented by editing existing *Grassland Tcl/Tk* scripts and adding new menus to its Librarian and Mapviewer windows. A system diagram of the *S+Grassland* Link is given in Fig. 10. By applying the *OGDI* and the *S+API* techniques, some new *Tcl/C* extensions (DLLs) needed to be developed for the interface. These introduce a number of new procedures in the *Tcl* interpreter, such as launching an *S-PLUS* window session, exporting GIS data from *Grassland* to *S-PLUS*, constructing spatial neighbor weight objects, and estimating a probability map by applying a general linear regression to sampled data from multiple selected layers.

Selected geographical data can be exported from *Grassland* to *S-PLUS* through either the Librarian or Mapviewer windows. Raster data are exported to *S-PLUS* as matrix objects, and the other geospatial objects (area, point and text data) as data frames containing *X-Y* coordinates and associated attribute data.

In the other direction, using the *S-PLUS* data driver specific to the *Grassland/OGDI*, *S-PLUS* objects can be accessed directly from *Grassland* via a URL connection established in the Librarian window in the usual fashion (using the Open Connection dialog in the File menu). All *S-PLUS* geospatial objects (Raster, Point, Text, Pmap) in the targeted workspace then become available to be dragged into the *Grassland* Mapviewer window for display.

Consequently, a user can export selected data from *Grassland* to *S-PLUS*, send commands to *S-PLUS*, and get results back from the statistical analysis for visualization. This link between the two processes is less transparent than the *S-PLUS for ArcView* interface and involves a number of explicit steps: (1)

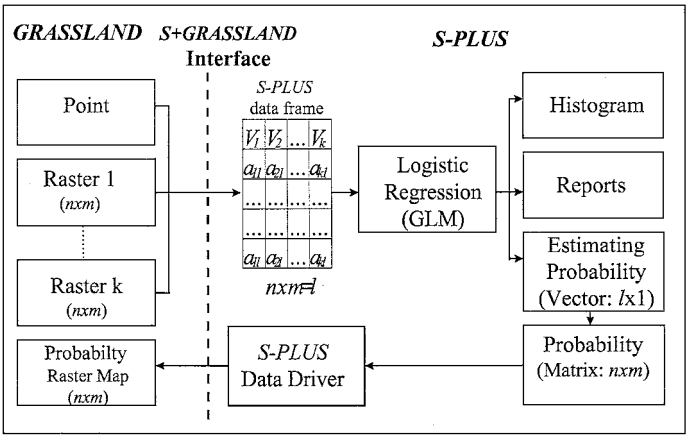

Fig. 11. GLM analysis with multilayer information

identify a coverage (data layer) inside *Grassland*; (2) choose a region of interest (a sub-regional boundary); (3) open an *OGDI* data source via an URL connection; (4) extract a spatial coverage using *OGDI*; (5) establish a connection with *S-PLUS*; (6) transform the *OGDI* data structure to an *S-PLUS* object structure; (7) write the data to an *S-PLUS* workfile; (8) write the geospatial reference data to a related *S-PLUS* workfile; (9) write the global geospatial reference data to a *S-PLUS* object (_GRS) in the case of a new SGW; (10) send *S-PLUS* commands to *S-PLUS*; (11) return the analytical results to *Grassland*; (12) close the URL connection; and (13) close the *S-PLUS* connection.

One important feature of the *S+Grassland* Link is its open structure, which allows the interface to be customized by adding new menus or buttons for additional functionality. For example, the menu item dealing with Estimating a Probability Map, which has been added to *Grassland*'s Mapviewer window, illustrates how statistical analyses can be applied to multiple data layers (see Fig. 11). Data are sampled from the selected point and raster layers and passed to *S-PLUS* as a data frame. A Generalized Linear Model (GLM) is applied to estimate a logistic regression equation, predicted values are obtained as a vector of probabilities, which is then converted into a matrix to be loaded into *Grassland*'s MapViewer window for visualization as a raster map. The output also includes a histogram for predicted probabilities, regression reports, ANOVA reports, and a list of those observations with high probabilities. As before, the essence of the link is that the statistical operations are carried out by *S-PLUS*, while the GIS is used as a spatial visualization device.

3.4 Illustration

To illustrate the features of the *S+Grassland* link, we apply the Probability Map function to a data set of archaeological developed by the U.S. Army Corps of Engineers, Construction Engineering Research Laboratories

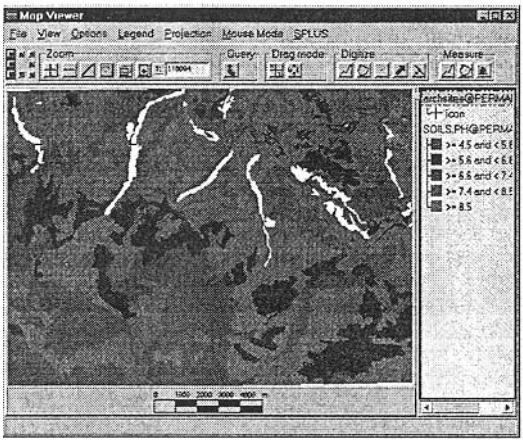

Fig. 12. The PH value of soils

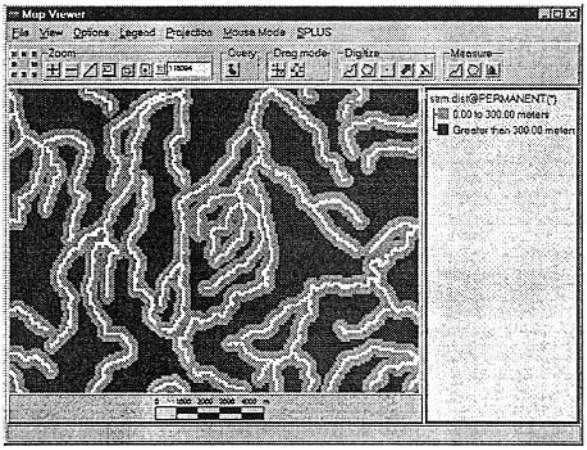

Fig. 13. The distance to streams

(CERL) for GRASS. There are 48 sample sites in the study region, 25 of which were classified as archaeological sites. In addition, there are data on soils, streams, vegetation, elevation, roads and land use. The archaeological sites are stored as point data with binary attributes (one if classified as a site, zero otherwise), while the other GIS data layers are stored as raster data.

The purpose of this illustration is to predict the location of potential archaeological sites by using the sampled site data and an estimated GLM with geographical attributes as explanatory variables, such as soil pH (see Fig. 12) and distance to streams (see Fig. 13). *Grassland* serves as the map viewer and spatial data manager from which the data and appropriate statistical commands are sent to *S-PLUS*. The resulting predicted probabilities are accessed directly from *Grassland* and viewed in the program's Map Viewer.

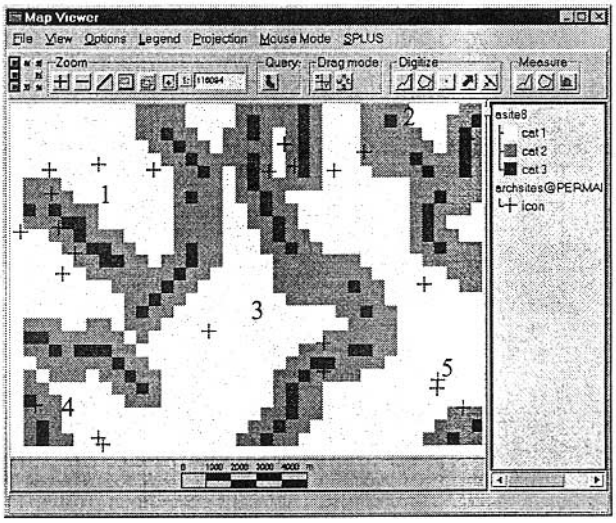

Fig. 14. Probability map for predicted archaeological sites

In this example, all selected data layers are stored in the *GRASS* data format and loaded into the *Grassland* Map Viewer. The archaeological site data (archsites) are sent to *S-PLUS*, along with the attribute data. The latter are extracted from the raster data and converted to an *S-PLUS* object (data frame). Binomial general linear regression is applied to estimate the relationship between the presence of archaeological sites and the associated geographical attributes. The results suggest that the potential for finding an archaeological site is higher if the soil has a low pH value ($\beta = -0.0901$) or if the site is closer to streams ($\beta = -0.4947$), and both coefficients were found to be significant. To obtain probabilities for the location of potential archaeological sites over the entire region, all observations extracted from the raster data set are used to calculate predicted values, which are then converted into a matrix to constitute a probability map, as illustrated in Fig. 14 (this suggests a grouping into five local areas labeled as 1–5 in Fig. 14).

4 Comparison of linkage strategies

An important shared feature of the two approaches is the seamless link between *S-PLUS* and the GIS. This is accomplished by adding new menu items or buttons to the respective graphical user interfaces in the GIS, which convert data transfer (to *S-PLUS*), function execution (in *S-PLUS*) and data visualization into one-click operations, transparent to the user.

Both links also include the added functionality needed to construct spatial weight matrices based on the topology of the data in the GIS. An explicit implementation of this is required, since, in contrast to say Arc/Info, the intrinsic data models in *ArcView* and *Grassland* are non-topological.

A major difference between *S-PLUS for ArcView* and the *S+Grassland* Link lies in their technical implementation. *S-PLUS for ArcView* uses an

automation technique, while the *S+Grassland* Link is implemented using the combination of *S+API* and *OGDI*. In addition to this distinction in design, this also results in operational differences. For example, *S-PLUS for ArcView* will launch an *S-PLUS* window automatically when it is loaded, which may create more overhead than necessary, especially if the user only needs functions implemented in the extension. In the *S+Grassland* Link, *S-PLUS* only functions as an utility library and runs in the background.

Secondly, all the *S-PLUS for ArcView* functions are based on data from a single layer (theme), while the *S+Grassland* Link can be applied to multiple layers of information. Finally, *S-PLUS for ArcView* can only be applied to attribute data associated with points and polygons, while the *S+Grassland* Link can be applied to spatial data in mutiple geospatial formats, such as point, polygon, and raster data.

5 Conclusions

The linkages between spatial statistical functionality and GIS outlined in this paper are still fairly rudimentary. They illustrate some important concepts, however. In order to establish an effective linkage, it is necessary to develop efficient formats and data structures that enable a bi-directional data exchange between the GIS and the statistical software. These data structures must respect the complexities incorporated in spatial data, such as location, projection and topology, as they are formalized in the existing GIS. Alternatively, new explicitly geospatial objects must be created, as in *S+Grassland* Link. More importantly, to establish a real-time functional integration between the two software packages, an indirect conversation by means of loose coupling is clearly limited, at least in principle. Instead, the elegant solution obtained by the use of the *OGDI* techniques combined with the *S-PLUS API* provide the tightest possible coupling. In practice, however, one still needs to consider performance issues related to the communication between the packages, as illustrated by the many steps encompassed in a single menu item in the *S+Grassland* GUI. Moreover, even in the *S-PLUS for ArcView* link, there is still a degree of loose coupling involved in order to transfer intermediate results (e.g., by means of Notepad) or data structures (to construct the Spatial Neighbor objects).

An altogether different issue pertains to the types of statistical techniques that are most effectively included in an integrated framework. Clearly, the temptation will exist to use any technique that is available, even though many/most standard statistical approaches (such as classical linear regression) become inappropriate in the presence of spatial autocorrelation, which is predominant in the spatial data sets manipulated by GIS. Instead of linking a comprehensive statistical (or spatial statistical) module with the GIS as a single piece of software, it may perhaps be more effective to implement selected methods in small self-contained software components that can be invoked from within a GIS. Alternatively, selected GIS components could be added to the functionality of a statistical package using automation technologies such as ActiveX (e.g., MapObjects components). An API such as *S+API* may provide the first step towards such a decentralized approach that would allow the individual user to customize the spatial data analysis "toolbox" for each application. In this respect, it still remains to be seen whether "traditional"

statistical packages such as *S-PLUS* and comprehensive GIS packages such as *ArcView* or *Grassland* will be the basis for the tools of the future, or instead become replaced by JAVA-based or similarly conceived free-standing web-based components. This constitutes a very promising area of future research.

References

Anselin L (1995) Local Indicators of Spatial Association – LISA. *Geographical Analysis* 27:93–115

Anselin L (1998) Exploratory Spatial Data Analysis in a Geocomputational Environment. In: Longley P, Brooks S, McDonnell R, Macmillan B (eds) *Geocomputation, a primer.* Wiley, New York pp. 77–94

Anselin L, Bao S (1997) Exploratory Spatial Data Analysis Linking *SpaceStat* and *ArcView*. In: Fischer M, Getis A (eds) *Recent developments in Spatial Analysis*. Springer, Berlin Heidelberg New York, pp. 35–59

Anselin L, Getis A (1992) Spatial Statistical Analysis and Geographic Information Systems. *The Annals of Regional Science* 26:19–33

Anselin L, Dodson R, Hudak S (1993) Linking GIS and Spatial Data Analysis in Practice. *Geographical Systems* 1:3–23

Bailey TC (1994) A review of statistical spatial analysis in geographical information systems. In: Fotheringham S, Rogerson P (eds) *Spatial analysis and GIS.* Taylor & Francis, London, pp. 13–44

Bao S, Stralberg D, Martin D (1998) Spatial data analysis and statistical modeling get a boost with GIS. *ArcNews Summer Issue*, pp. 24

Bao S (1997) *User's reference for the S+Grassland Link.* Mathsoft, Inc Seattle, WA

Bao S, Anselin L (1998) *Linking Spatial Statistics with GIS: Operational issues in SpaceStat/ArcView interface and S+Grassland link.* American Statistical Association, 1997. Proceedings of the Section on Statistical Graphics VA: ASA Alexandria pp. 61–66

Bao S, Henry MS (1996) Heterogeneity issues in local measurements of spatial association. *Geographical Systems* 3:1–13

Bao S, Martin D (1997) *Integrating S-PLUS* with *ArcView in spatial data analysis: An introduction to the S+ArcView Link.* ESRI's Users Conference, San Diego, CA

Bao S, Henry MS, Barkley DL, Brooks K (1995) RAS – an integrated regional analysis system with ARC/INFO. *Computers, Environment, and Urban Systems* 19:1:37–56

Bartlett MS (1971) Physical nearest-neighbour models and non-linear time-series. *Journal of Applied Probability* 8:222–232

Batty M, Xie Y (1994) Modeling inside GIS: Part I. model structures, exploratory spatial data analysis and aggregation. *International Journal of Geographical Information Systems* 8:291–307

Besag JE (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society* B 36:192–225

Can A (1996) Weight matrices and spatial autocorrelation statistics using a topological vector data model. *International Journal of Geographical Information Systems* 10:1009–1017

Cliff AD, Ord JK (1981) *Spatial processes: Models and applications.* Pion Limited, London

Cressie N (1993) *Statistics for spatial data* (revised edition). Wiley, New York

Ding Y, Fotheringham AS (1992) The integration of spatial analysis and GIS. *Computers, Environment and Urban Systems* 16:3–19

ESRI (1996) *ArcView GIS.* Environmental Systems Research Institute, Redlands, CA

ESRI (1997) *Using the ArcView Dialog Designer.* Environmental Systems Research Institute, Redlands, CA

Farley JA, Limp WF, Lockhart J (1990) The archeologist's workbench: Integrating GIS, remote sensing, EDA and database management. In: Allen K, Green F, Zubrow E (eds) *Interpreting space: GIS and archaeology.* Taylor & Francis, London, pp. 141–164

Fischer M, Getis A (1997) Recent developments in spatial analysis. Springer, Berlin Heidelberg New York

Fischer M, Nijkamp P (1993) Geographic information systems, spatial modelling and policy evaluation. Springer, Berlin Heidelberg New York

Fischer M, Scholten H, Unwin D (1996) Spatial analytical perspectives on GIS in environmental and socio-economic sciences. Taylor & Francis, London
Fotheringham AS, Rogerson P (1994) Spatial analysis and GIS. Taylor & Francis, London
Goodchild MF (1987) A spatial analytical perspective on geographical information systems. *International Journal Systems* 1:327–334
Goodchild MF (1992) Geographical information science. *International Journal of Geographical Information Systems* 6:31–45
Goodchild MF, Haining R, Wise S, (1992) Integrating GIS and spatial analysis – Problems and possibilities. *International Journal of Geographical Information Systems* 6:407–423
Haining R (1994) Designing spatial data analysis modules for geographical information systems. In: Fotheringham S, Rogerson P (eds) *Spatial analysis and GIS.* Taylor & Francis, London pp. 45–63
Logiciels et Applications Scientifiques (L.A.S.) Inc (1996a) *Grassland user's guide for Window 95 and Windows NT Version 1.1.* L.A.S. Montreal, Canada
Logiciels et Applications Scientifiques (L.A.S.) Inc (1996b) *Open geographic datastore interface version 1.0 programmer's reference.* L.A.S. Montreal, Canada
MathSoft (1996a) *S+GISLink* MathSoft, Inc, Seattle
MathSoft (1996b) *S+ SpatialStats user's manual for Windows and Unix.* MathSoft, Inc, Seattle
MathSoft (1997a) *S-PLUS user's guide version 4.0 for Windows.* MathSoft, Inc, Seattle
MathSoft (1997b) *S-PLUS API object interface library, reference manual, beta version 0.3* MathSoft, Inc, Seattle
MathSoft (1997c) *S+API: External specifications.* MathSoft, Inc, Seattle
MathSoft (1998) *S-PLUS for ArcView GIS user' guide, version 1.0* MathSoft, Inc, Seattle
Painho M (1994) *New tools for spatial analysis.* Eurostat, Luxembourg
Openshaw S (1991) Developing appropriate spatial analysis methods for GIS. In: Maguire D, Goodchild MF, Rhind D (eds) *Geographical information systems: Principles and applications*, vol 1. Longman, London, pp. 389–402
Openshaw S, Fischer M (1995) A framework for research on spatial analysis relevant to geostatistical information systems in Europe. *Geographical Systems* 2:325–337
Ousterhout JK (1994) *Tcl and the Tk Toolkit.* Addision-Wesley Publishing Company
Symanzik J, Majure J, Cook D, Cressie N (1994) Dynamic graphics in a GIS: A Link between Arc/Info and XGobi. *Computing Science and Statistics* 26:431–435
Symanzik J, Majure J, Cook D (1996) Dynamic graphics in a GIS: A bidirectional Link between *ArcView* 2.0 and XGobi. *Computing Science and Statistics* 27:299–303
Symanzik J, Kötter T, Schmelzer S, Klinke S, Cook D, Swayne DF (1998) Spatial data analysis in the dynamically Linked ArcView/XGobi/XploRe environment. *Computing Science and Statistics* 29: (in press)
USACERL (1993) *GRASS 4.1 user's reference manual.* U.S. Army Corps of Engineers. Construction Engineering Research Laboratories, Champaign, Illinois
Whittle P (1954) On stationary process in the plane. *Biometrika* 41:434–449
Williams I, Limp W, Briuer F (1990) Using geographic informations and exploratory data analysis for archaeological site classification and analysis. In: Allen K, Green F, Zubrow E (eds) *Interpreting space: GIS and archaeology.* Taylor & Francis, London pp. 239–273
Zhang Z, Griffith D (1997) Developing user-friendly spatial statistical analysis modules for GIS; An example using *ArcView*. *Computers, Environment and Urban Systems* 21:5–29