View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

Integrating Spatial Regression into Decision Support Systems

Vijayan Sugumaran
Oakland University, sugumara@oakland.edu

Lee Rivers Mobley
Oakland University, mobley@oakland.edu

Follow this and additional works at: http://aisel.aisnet.org/amcis2000

Recommended Citation
Sugumaran, Vijayan and Mobley, Lee Rivers, "Integrating Spatial Regression into Decision Support Systems" (2000). *AMCIS 2000 Proceedings*. 249.
http://aisel.aisnet.org/amcis2000/249

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2000 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Integrating Spatial Regression into Decision Support Systems

Vijayan Sugumaran, Lee Rivers Mobley
School of Business Administration, Oakland University, Rochester, MI, 48309
{sugumara, mobley}@oakland.edu

Abstract

Spatial Information Management is gaining popularity and managers are beginning to appreciate the power and application of spatial modeling in decision support. A new breed of DSS called Spatial Decision Support System (SDSS) is emerging, which supports spatial data and spatial modeling. This paper highlights the importance of spatial modeling with an example, and presents an architecture for a SDSS, and discusses it components. A prototype of the environment is under development using ArcView, SpaceStat, Jess, and Visual Basic.

Introduction

Decision Support Systems (DSS) have been an important area of IS research. There is a vast literature examining various aspects of the DSS, such as its nature, effectiveness in decision making, framework or architecture, group support, etc. While various DSS have been used in managerial decision making, a major limitation of these systems is their inability to exploit spatial and temporal data. Because much useful business data is spatially referenced, ignoring this additional information is shortsighted. Accordingly, a new breed of DSS is emerging, the Spatial Decision Support System (SDSS).

The extension of the functional capacity of GIS with tools for spatial analysis has been a very productive area of research in recent years (Anselin and Bao, 1997). But most of this effort has focused on linkages which are useful for the *exploration* of geo-statistical data, which assumes that observed locations are a sample drawn from an underlying continuous distribution that can be conceptualized as a spatial mat represented by grid data. These linkages have been used primarily for univariate analysis, rather than multivariate spatial modeling. More recently, Environmental Systems Research Institute (ESRI, 2000) has developed a spatial modeling component which enables multivariate modeling using geo-statistical (grid) data. These things have enabled the development of SDSS which are geared toward the analysis of geo-statistical (grid) data. However, there is another sort of spatial data which may be more useful for business and economic analysis - lattice data. Much less progress has been made with linkages geared toward the lattice (or neighborhood) view of spatial data, which assumes that observed locations are single realizations from a spatial stochastic process, similar to the approach taken in the analysis of time series data (Anselin and Bao, 1997). This lattice approach is much more useful in the exploration of economic and business data, where each location conceivably interacts with neighbors. Examples of lattice-type data are data from the U.S. Census of Populations, customer databases, supplier databases, etc.. In this context, spatial regression models can be developed which are useful in understanding spatial interaction and in forecasting expected spatial patterns from business and economic data.

This research focuses on developing spatial regression models within a Spatial Decision Support System (SDSS) for managerial decision making. This linkage has considerable value in facilitating the analysis and forecasting of spatially-oriented business data. Similar to a DSS, our SDSS consists of the following subsystems: a) data management, b) model management, c) knowledge management, d) dialog management, and e) display and report generators. Ideally, SDSS have to be flexibly integrated systems that could be built on a GIS platform to deal with spatial data and manipulations, along with an analysis module, which could switch from exploration to explanation in an interactive, iterative and participatory way. Just like a DSS, SDSS have to support "what-if" analysis and also provide a range of tools to help the user in understanding the results, modifying the results, and determining the implications of each new iteration (Goodchild et al., 1992).

Spatial Regression and the GIS

Spatial regression is concerned with the analysis of spatially-referenced data. This differs from classical (a-spatial) regression in that the observations analyzed are not independent. Observations are correlated with others that are spatially proximate. The usual methods for correcting autocorrelation in linear models are not sufficient with spatially-referenced data, because the autocorrelation is not linear - it is multi-dimensional (geographic neighborhood-specific). If we ignore the statistical (spatial) dependence, we are ignoring information about potential data complexities.

Sometimes observations are correlated strictly due to their locational positions, resulting in spill-over of information from location to location. In cases of positive autocorrelation, this spill-over causes redundant

information to be present in data values, and the redundancy is an increasing function of the degree of locational similarity. This redundancy means that there is actually less information present in the sample than would be present in a sample of independent observations. In this case, the goodness-of-fit statistic (R squared) is overly optimistic - fit is not as good as it appears. For forecasting purposes, this is problematic, because the model may not predict as well as it seems to indicate.

There are (at least) two other possible types of statistical problems in spatial regression, with even more serious consequences. One type is created by the arbitrary way in which boundaries may be designated, which defines the unit of aggregation for the data. This problem exists when there is a mismatch between the spatial scale of the phenomenon under study and the spatial scale at which it is measured. This mismatch causes spatial measurement errors and spatial autocorrelation between these errors in adjacent locations (Anselin, 1988). Another type of statistical problem can arise from the failure to include (as explanatory variables) measures that fully model the spatial environment. Either problem can cause the Ordinary Least Squares (OLS) estimators to be biased and inconsistent. This means that the estimated impacts of explanatory variables on the dependent variable are not reliable - they may either over or understate influence, with direction of bias unknown. This is also problematic for managers using spatial regression analysis for planning and forecasting, as predictions can be misleading in both magnitude and direction.

Fortunately, sophisticated spatial regression software exists (SpaceStat), which can diagnose these problems. How to use these diagnostics is a crucial part of the knowledge base, which we build into our SDSS. Furthermore, a dynamic link between SpaceStat and ArcView GIS software allows the user to interact with the GIS as part of the diagnostic process. In what follows, we first briefly describe the set of protocols to be built into the knowledge base, and how they should be applied. We then provide a detailed example of this process using data and an example. In the next section we describe the linkage we develop between spatial regression and the GIS, and show how this can provide feedback to assist in the modeling process.

Specification Testing: Heteroskedasticity, Error and Lag Structures

The first sign of trouble in a spatial regression model is the presence of significant heteroskedasticity in the residuals from an ordinary least squares (OLS) regression. This signals that spatial heterogeneity exists in the data and that it has not been captured (modeled) in the model. The spatial heterogeneity may arise due to three situations: 1) spatial regimes (regions with significantly different spatial patterns in key variables), 2) similarities between neighbors due to some local phenomenon (i.e. all sick due to a poisoned well), or 3) similarities between neighbors due to some spreading phenomenon (i.e. all sick due to measles epidemic). The problem of finding the correct model specification - one which yields 'white noise' residuals - is complicated by the fact that either situation 2) or 3) can *create* heteroskedasticity as a by-product, making it difficult to distinguish between these three cases. Fortunately, SpaceStat has built-in diagnostics which are robust to these complexities (Anselin, 1995; Anselin and Bera, 1998).

Situation 2) above is considered the least problematic; if undiagnosed and left un-corrected, this leads to bias in the goodness-of-fit. When there is significant evidence from the *spatial error test statistic* that this sort of *spatial error* process is present in the data, the model should be re-estimated using a *spatial error* model form. Situations 1) and 3) are more serious - they can lead to biased and inconsistent parameter estimates for the impact variables. The first sort of situation (1) calls for a spatial regression model which treats each region as a separate subset of the data, and allows all parameters to vary across regions. A spatial regimes Chow test statistic is the diagnostic which enables identification of significantly different regimes. Finally, when there is significant evidence from the *spatial lag test statistic* that a *spatial lag* process is present in the data (situation 3), the model should be re-estimated using a *spatial lag* model form. These three different spatial models (among others) are available in SpaceStat.

To test for significant lag and error processes, the software must compare residuals among neighbors. Thus the modeler must choose a neighborhood set to use in constructing a spatial weights matrix. SpaceStat links dynamically with ArcView GIS to facilitate construction of these spatial weights, and many variations are possible (i.e. include 5 closest neighbors, all within 20 miles, everyone with influence diminishing with distance, etc.).

In summary, the steps for diagnosing spatial problems which are built into our knowledge base are illustrated in the following list and diagnostic flow chart shown in Figure 1:

- Estimate the model using the Ordinary Least Squares (OLS) model in SpaceStat.

- Examine the diagnostic tests which accompany the OLS regression output for the presence of either a spatial lag or a spatial error process in the residuals. A low p-value (up to 5%) suggests that a significant problem exists. The test statistic for presence of an error process is RS_{λ}, and the test statistic for presence

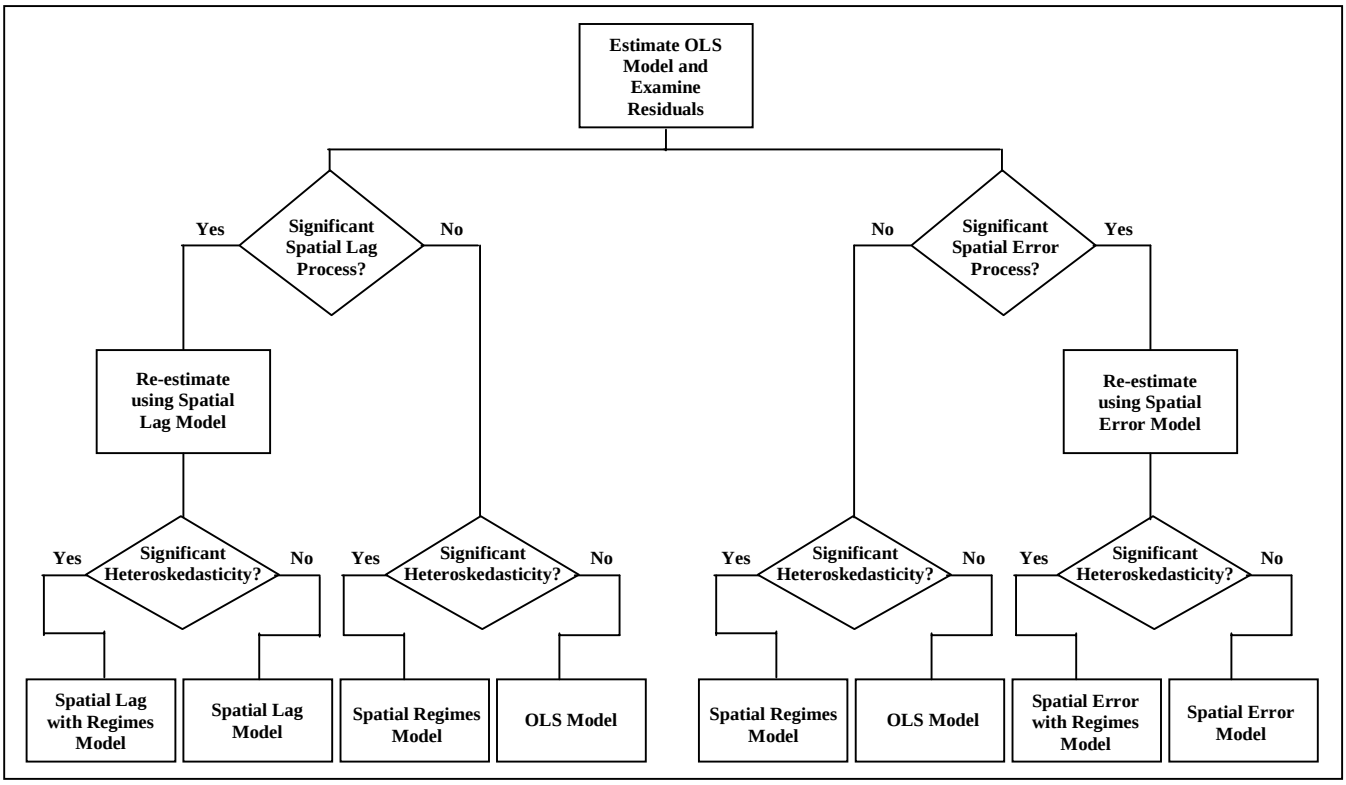

Figure 1. Diagnosing a Spatial Process

of a lag process is RS_ρ. To diagnose the problem, apply the rule: if RS_ρ is significant (low p-value) while RS_λ is not, then spatial lag is most likely the correct error structure. Conversely, if RS_λ is significant while RS_ρ is not, then spatial error is most likely the correct error structure.

- If either of these problems exists, re-estimate the model using the correct error form, and then check for any remaining heteroskedasticity. If present, see next bullets.

- If neither of these problems exists (spatial error or lag) then check for the presence of significant heteroskedasticity. If present, this suggests that spatial regimes are present in the data. Often mapping the residuals can aid in the discovery of exactly what these regimes might be (i.e. urban-rural, city center-suburbs, presence or absence of some other spatial phenomenon).

- If spatial error is present with heteroskedasticity, then estimate a spatial error model with regimes. If spatial lag is present with heteroskedasticity, then estimate a spatial lag model with regimes. If neither spatial error or spatial lag is present but heteroskedasticity is, then estimate a spatial regimes model.

- When none of these problems are evident in the model's residuals, then no further spatial modeling is necessary.

Importance of Spatial Modeling : Some Examples

In a well-known article, Jaffe (1989) came to the dismal conclusion that "there is only weak evidence that spillovers are facilitated by geographic coincidence of universities and research labs within the state". This conclusion was decisively refuted by Anselin, Varga, and Acs (1997), who used a spatial econometric approach to carefully model spatial interaction. Anselin et. al. find a positive and significant relationship between university research and innovative activity, both directly, and indirectly through spillovers on private sector R&D. This example illustrates the importance of careful spatial modeling in drawing correct conclusions and inferences from the data.

We illustrate this again using another example, which applies the diagnostic methodology from our knowledge base outlined above. First we estimate the determinants of crime patterns in the Columbus, Ohio region, using neighborhoods as the units of analysis, and an OLS model. The results are presented in Table 1, below.

Table 1: Explaining Crime Rates in the Columbus, Ohio Region						
	OLS Model Single Regime		Spatial Lag Model With Two Spatial Regimes:			
			Inside City Center		Outside Center	
Variable	coeff	p-val	coeff	p-val	coeff	p-val
INCOME	-1.597	0.000	-1.746	0.041	-0.707	0.021
HOUSEVALUE	-0.274	0.011	-0.045	0.808	-0.209	0.024
Diagnostic Test	Conclusion	(p-val)	Conclusion	(p-val)		
Heterosk. Test	Present	(0.001)	Not present	(0.911)		
RS_λ Error test	Not present	(0.147)	Not present	(0.612)		
RS_ρ Lag test	Present	(0.030)	Not present	(0.420)		
Regimes test	NA		Regimes Present	(0.023)		
Weights matrix	5 nearest neighbors		5 nearest neighbors			
Rsquared	0.552		0.718			

Beneath the coefficient estimates are the diagnostic tests for heteroskedasticity and spatial error or lag processes. These suggest that heteroskedasticity is present, and/or a spatial lag process. We re-estimated the model using the spatial lag specification, and tested again for heteroskedasticity, which was still significant (this step skipped in Table 1). This led to our final specification, with spatial lag combined with spatial regimes (Table 1). The binary regimes are defined as "whether in city center, or not." The diagnostics presented for this second model show that there is no remaining heteroskedasticity, no further lag or error structure modeling needed, and that the spatial regimes are significantly different.

Upon examination of the goodness-of-fit (R-squared), one can see that the model fit has improved dramatically with the spatial modeling. Also, comparison of estimated coefficients across the OLS versus spatial model reveals the extent of bias imparted by the OLS model.

So far, we have presented a brief overview of spatial regression and the typical steps involved in diagnosing spatial problems. Generally, in order to arrive at the correct spatial model, substantial *a priori* knowledge of spatial statistics is necessary. A system that minimizes this cognitive load by hiding the complexities and which helps the user in developing appropriate models would greatly facilitate the use of spatial models in decision making. The following section discusses the architecture of such a spatial decision support system.

Architecture of Our SDSS

A unique feature of our SDSS is the inclusion of a knowledge base which interacts with both the modeling stage and the GIS (Figure 2). To illustrate this using the Columbus crime data example above, our diagnostics told us that some sort of spatial regimes were likely present, because of the continuing heteroskedasticity even after a spatial lag model was specified. Linkage between the spatial econometrics module (the model base) and the GIS (core component) is crucial at this point, in order to discover the 'offending' pattern of spatial regimes that need inclusion in the model.

To facilitate this, we build a dynamic link between SpaceStat (model base) and ArcView (GIS core component), which is contained within the 'model base interface' in Figure 2. This link allows the user to capture and then dynamically plot the residuals from model estimation, and to simultaneously test them for significant patterns of spatial association. The visual display on the map can help the analyst discover the pattern in any unmodeled (remaining) spatial heterogeneity. The model base interface would then allow the analyst to construct new spatial variables for inclusion in the model, based on the information displayed on the map about remaining spatial heterogeneity. Selected areas could be used to create spatial regime indicator variables, for example, which can be passed back to SpaceStat for inclusion in the spatial regression model.

This visual interaction is crucial in building good models. The diagnostics can only tell us that a problem exists; without visualization, it is often impossible to determine how to correct the problem. It is at present possible for the analyst to interact with SpaceStat as described above, albeit in a rather clumsy and unfriendly fashion. Our SDSS will facilitate this sort of user interaction, among other things.

A few SDSS have been discussed in the literature (Densham, 1991; Moon, 1992; NCGIA, 1992;

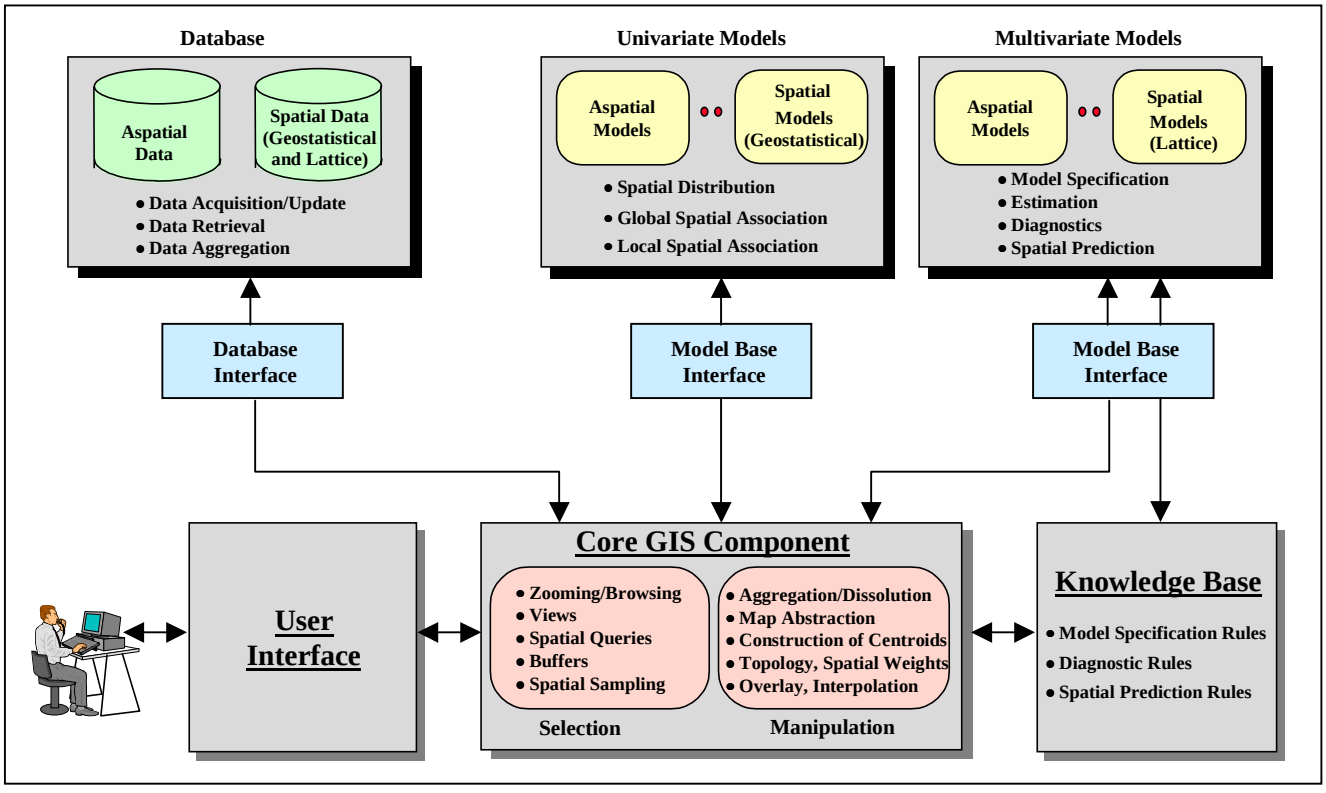

Figure 2. Architecture of the Spatial Decision Support System

Sugumaran, 1998) to solve a variety of problems related to natural resource management, urban planning, land use planning, environment management, water quality etc. However, these systems have a very narrow focus and are application domain specific; hence they do not have a broad range of application. Researchers are currently focusing on designing distributed, adaptive decision support systems (Ferrand, 1996; Chuang et al., 1997), which are configurable, based on the problem at hand.

One of the limitations of existing SDSS is the lack of adequate support for spatial modeling. Also, these systems do not provide mechanisms for applying different spatial models in problem solving and performing sensitivity analysis. This type of "what-if" analysis is essential in unstructured problem solving and decision making. In this research, we focus on designing a SDSS which not only facilitates specifying appropriate spatial models for the problem at hand, but also provides knowledge-based support for sensitivity analysis. We extend the models suggested by Murphy (1995), and Mennecke (1997) for "GIS as a decision support tool" in generating a model for our proposed SDSS environment. Conceptually, the proposed SDSS is comprised of the following components: a) database, b) model-base, c) knowledge-base, d) user interface, and e) core GIS. The architecture of the SDSS is shown in Figure 2, and its components are briefly described in the following paragraphs.

Database: The database component provides access to spatial and non-spatial data stored within the organization. It encompasses traditional databases as well as GIS databases that contain spatial and temporal data. Managerial decision making requires easy access to large volumes of internal and external data, as well as data of different types (quantitative, qualitative, spatial, temporal etc.). The database interface component facilitates accessing different types of data both internal and external to the organization.

Model-base: The model-base component provides access to a large number of models necessary for analyzing and solving unstructured problems. The model base supports both aspatial and spatial models. It also facilitates development and testing of new models. There are two broad categories of models supported by the model-base, namely, univariate models, and multivariate models. The model base interface component facilitates the model management activity, through interaction with the knowledge base.

Knowledge-base: The knowledge base consists of a number of rules that help in model selection and execution. It enables the user to select a particular type of

model to use in problem solving and also perform sensitivity analysis. It interacts with both the model base and the core GIS via the model base interface. The knowledge base may also contain organizational policies, procedures, business rules and constraints that would influence the types of models to be used in problem solving. Prior results of unstructured problem solving are also stored here.

Core GIS Component: The core GIS component is capable of assembling, storing, manipulating, and displaying geographically referenced information, i.e., data identified according to their locations. Geographical information consists of both textual data ("attribute" or "aspatial" data) as well as spatial data (data which includes cartographic coordinates). Thus, the core GIS not only provides users with tools for managing and linking attribute and spatial data, but also advanced modeling functions, designing and planning, and imaging capabilities.

User Interface: The interface component provides the user interface for the SDSS. It provides a graphical interface for the user to interact with the system during a decision support session. The interface can be customized according to the tastes and preferences of individual users. The dialog component also provides different presentation modes as well as different reporting capabilities.

Implementation of SDSS

We are in the process of implementing an SDSS environment and demonstrating its usefulness by applying it to various domains. The components of the SDSS are implemented using available commercial products to the extent possible. Existing GIS and spatial modeling and analysis tools have various limitations, and more importantly, do not interface well. Hence, a major emphasis of this project is to develop an SDSS environment that integrates some of the well known GIS and analysis tools in order to harness the power of each of these products. In the proposed SDSS environment, the core GIS component will be the ArcView GIS product. The model-base is created using the SpaceStat software (Anselin, 1995). As mentioned earlier, the model-base may contain univariate models and multivariate models. In this implementation, we primarily focus on the multivariate models. While a rudimentary interface exists between ArcView GIS and SpaceStat, we intend to develop a full-blown interface between these two products that would greatly improve interoperability.

We also intend to develop a comprehensive database interface module that would enable the user to access data from both internal and external data sources and retrieve the necessary data for problem solving. The knowledge-base is implemented using an expert system shell called Jess (Java Expert System Shell), from Sandia National Laboratories (Friedman-Hill, 1999). Model selection rules and sensitivity analysis rules can be easily implemented in Jess. The user interface is being implemented in Visual Basic. This GUI would enable the user to interact with the system and develop various spatial models and apply them to the problem that is currently being solved, and visually examine the results. This also helps the user in performing sensitivity analysis. The next two paragraphs provide some basic information on the functionalities of ArcView and SpaceStat products.

ArcView: ArcView software is the GIS component in our system. ArcView is used to map geo-referenced data and to analyze spatial relationships among data points using SQL (structured query language). Many data formats are readily imported into ArcView, and we facilitate this within our system. The geo-referenced data may be recorded at different spatial levels - i.e. points (cities, street addresses, zip code centroids), lines (roads, rivers, railroads) or polygons (zip code boundaries, census tracts, county boundaries). These can be simultaneously layered for complex multi-layered queries. Queries are useful for defining subsets of the data, creating categorical variables, etc. for subsequent multivariate analysis. Also imbedded in ArcView as an add-on is a multivariate spatial modeling component, which is useful in the analysis of geo-statistical (grid) data.

SpaceStat: SpaceStat software is the multivariate spatial modeling component in our system, which is useful in the analysis of lattice data (Anselin, 1995). It links dynamically to ArcView, and can export data from layers, with latitude and longitude coordinates, for subsequent analysis in SpaceStat. SpaceStat can create a full distance matrix between all pairs of observations in the dataset, which is the basis for a variety of possible spatial weights. SpaceStat contains advanced spatial econometrics with extensive specification tests built in to assist researchers in spatial modeling. Models that can be estimated include linear regression, 2SLS, spatial error, spatial lag, and spatial regimes models, among others - using either Maximum Likelihood or Method of Moments estimators. A full range of sample statistics and data manipulation tools are also included. Individual variables from a SpaceStat dataset can be dynamically explored in ArcView in a variety of ways.

Summary

Traditional DSSs are limited in supporting spatial data/models and researchers are investigating ways to incorporate GIS components into DSS. Spatial Decision Support Systems are increasingly being applied to real world problem solving and decision making, but most are limited to the analysis and modeling of geo-statistical (grid) data. Lattice data are more commonly encountered in business applications, and we focus on developing

SDSS for this type of spatial modeling. In this paper, we have highlighted the importance of spatial modeling which can be applied to lattice-type data, and provided a quick overview of the topic. We have also presented an architecture for a Spatial Decision Support System, that facilitates interactive spatial modeling, problem solving, and decision making in the lattice-data context. A prototype of the environment is being developed using ArcView, SpaceStat and Jess. The interface modules are being implemented in Visual Basic.

References

Anselin, A. and Bera, L. (1998). "Spatial Dependence in Linear Regression Models With an Introduction to Spatial Econometrics", in Giles, D. and A. Ullah (eds.), *Handbook of Applied Economic Statistics* (New York: Marcel Dekker).

Anselin, L. and S. Bao (1997). "Exploratory Spatial Data Analysis Linking SpacsStat and ArcView" in M. Fischer and A. Getis (eds.) *Recent Developments in Spatial Analysis*, Springer-Verlag: Berlin, Chapter 3.

Anselin, L. and A. Varga, and Z. Acs, (1997) "Local Geographic Spillovers Between University Research and High Technology Innovations", *Jl. of Urban Economics*, v 42, pp 422-448.

Anselin, L. (1995). SpaceStat, A Software Program for the Analysis of Data.
URL: http://www.spacestat.com

Anselin, L., (1988) *Spatial Econometrics: Methods and Models*, Kluwer, Dordrecht.

Chuang, T., Yadav, S. B. (1997). An Agent-Based Architecture of an Adaptive Decision Support System. Proc. of the third AMCIS, Indianapolis, Indiana, August 15-17.

Densham, P. J. (1991). Spatial Decision Support Systems. In: Maguire, D.J., Goodchild, M.F., and Rhind, D.W., eds. Geographical Information Systems: Principles and Applications, Vol. 1, Longman, 403-412.

ESRI (2000), Environmental Systems Research Institute, Inc., *ModelBuilder for ArcView Spatial Analyst*, http://www.esri.com.

Ferrand, N. (1996). Modelling and Supporting Multi-Actor Spatial Planning Using Multi-Agents Systems. Proc. of the Third NCGIA Conference on GIS and Environmental Modelling, Santa Fe, Jan.

Friedman-Hill, E. (1999). Jess: The Expert System Shell, Sandia National Laboratories, Livermore, CA. (http://herzberg.ca.sandia.gov/jess)

Goodchild, M.F., Haining, R., Wise, S. (1992). Integrating GIS and Spatial Data Analysis: Problems and Possibilities. Intl. Jl. of Geographic Information Systems. 6(5): 407-423.

Jaffe, A. (1989). "Real Effects of Academic Research", *American Economic Review*, v 79, pp 957-970.

Mennecke, B. E. (1997). Understanding the Role of Geographic Information Technologies in Business: Applications and Research Directions. Jl. of Geographic Information and Decision Analysis, Vol. 1, No. 1, pp. 44-68.

Moon, G. (1992). Capabilities Needed in Spatial Decision Support Systems. GIS/LIS '92, Vol. 2: 594-600.

Murphy, L, (1995). "Geographic Information Systems: Are they Decision Support Systems?" Proceedings of the Twenty Eighth Annual Hawaii International Conference on System Sciences, Maui, Hawaii, pp. 131-140.

NCGIA (1992). A Research Agenda for Geographic Information Analysis. Technical Report 92-7. National Center for Geographic Information Analysis.

Sugumaran, V. (1998). "A Distributed Intelligent Agent-Based Spatial Decision Support System," Proceedings Of AMCIS'1998, Aug. 14-16, Baltimore, MD, pp. 403-405.