Spatial Data Mining: An Emerging Tool for Policy Makers

by Sanjay Chawla, Shashi Shekhar, Wei Li Wu, and Xinhong Tan

Just as the widespread use of relational databases triggered interest in classic data mining (CDM) techniques, widespread use of spatial databases has increased interest in "mining" useful but implicit spatial patterns among data. Data mining products are being successfully used as decision-making and planning tools in both the public and private sectors. Knowledge extraction from geospatial data was highlighted as a key area of research at a recently concluded National Science Foundation workshop on geographic information systems, and a January 20, 2000, article in the *New York Times* on spatial data mining (SDM) demonstrates that interest in this technology has spread to the public domain.

One of the corollaries of the information age is that society has become inundated with large quantities of data. The sheer size of these data sets often makes it difficult to search for meaningful patterns or relationships among data. Data mining is a technique that allows researchers to overcome this obstacle and discover potentially interesting and useful patterns of information embedded in large databases. A pattern can be a summary statistic such as the average or mean, or a statistical relationship such as a correlation between two events. A well-publicized pattern that has now become part of data mining lore was discovered in the transaction database of a national retailer: People who buy diapers in the afternoon also tend to buy beer. This was an unexpected finding that the company put to profitable use by rearranging store merchandise.

The promise of data mining is the ability to rapidly and automatically search for local and potentially high-utility patterns using computer algorithms. Data mining draws on techniques from machine-learning, database management, and statistics to rapidly search for patterns in the data. Although many data mining techniques were inspired by classic statistical techniques, there is one major difference: In statistics, data are used to test the validity of a hypothesis, while in data mining, patterns and hypotheses are "discovered" by exploring the data. Thus, data mining encompasses a set of techniques to automatically generate hypotheses, followed by their validation and verification via standard statistical tools.

Identifying efficient tools for extracting information from geo-spatial data is important to organizations that own, generate, and manage large geospatial data sets. In this article, we will discuss the differences between CDM and spatial data mining (SDM) techniques, develop a model for incorporating spatial properties into both classic statistical analyses and a data mining framework, apply this model to an example from ecology involving wildlife habitat, and discuss the implications of the SDM model for policy makers.

Classic Data Mining vs. Spatial Data Mining

The difference between CDM and SDM is similar to the difference between classic and spatial statistics. One of the fundamental assumptions of classic statistical analysis is that data samples are independently generated, much like successive tosses of a coin or rolls of a die, where each toss or roll has no relationship to the previous one. When it comes to the analysis of spatial data, however, the assumption of independence is generally false because spatial data tend to be highly self-correlated. For example, people with similar characteristics, occupations, and backgrounds tend to cluster together in the same neighborhoods. The economies of a region tend to be similar. Changes in natural resources, wildlife distribution, and temperature usually vary gradually over an area.

The tendency of like things to cluster in a space is so fundamental that geographers have elevated this phenomenon to the status of the first law of geography: "Everything is related to everything else, but nearby things are more related than distant things." In spatial statistics, an area within statistics devoted to the analysis of spatial data, this tendency is called spatial autocorrelation. Ignoring spatial autocorrelation when analyzing data with spatial characteristics may produce hypotheses or models that are inaccurate or inconsistent with the data set. Thus, CDM algorithms often perform poorly when applied to spatial data sets.

Any data set that has a spatial, locational, or geographic component can be considered a spatial database. Examples of common spatial databases include maps, repositories of remote-sensing images, and the decennial census. Spatial data mining involves the search for patterns embedded in large spatial databases. Although contemporary SDM involves the use of computers, the following well-known examples of what we now call SDM occurred long before the invention of computers:

- In 1855, when Asiatic cholera was sweeping through London, an epidemiologist marked on a map those locations where the disease had struck. The epidemiologist discovered that the locations formed a cluster, at the center of which was a water pump. When government authorities turned off the water pump, the cholera epidemic began to subside.
- In 1909, a group of dentists discovered that the residents of Colorado Springs had unusually healthy teeth. They attributed this occurrence to the high level of natural fluoride in the local drinking water. Now all municipalities in the United States ensure that their drinking water supply is fortified with fluoride.
- In 1919, an investigator discovered (using maps) that all the continents could be fitted together like a giant jigsaw puzzle. Based on this discovery, the theory of Gondwanaland—which states that all the continents once formed a single landmass—was proposed.

Over the years, information databases have grown so large that it has become both useful and necessary to automate the search for potentially meaningful patterns. For example, an interesting problem in crime analysis is the detection, explanation, and prediction of "hot spots"—locations in a community or city that experience outbreaks of increased criminal activity. The classic statistical approach to detection of such spots is for an expert to use a GIS to correlate different map-layers of attribute data available for that city. The promise of data mining is that it allows the analysis to be recast as a search problem in a database. By using high-speed computers and smart algorithms, it is possible to search the database for clusters of data that may characterize *potential* hot spots. Thus, the domain expert who earlier searched for hot spots with the aid of a GIS is now involved in setting up the correct problem, and then interpreting the output from a data mining algorithm to determine which of the hot spots are worthy of further analysis using standard statistical techniques. Data mining is a tool for generating candidate hypotheses from data on which no a priori information is available.

Spatial data mining holds the promise of discovering patterns within existing spatial databases with minimal human intervention. As such, SDM can be a powerful aid in policy decision making. The following areas in which SDM is already playing an important role were showcased in a January 20, 2000, *New York Times* article:

- Monitoring lending patterns of institutions. Consumer advocacy groups are using SDM to map the lending practices of banks and other lending institutions. By relating the location of banks with the demographics of surrounding neighborhoods, SDM techniques can provide more accurate information about whether poor neighborhoods are being denied fair access to credit.
- Crime mapping and hot-spot analysis. Techniques from SDM can be used to detect local patterns in crime databases and examine related databases to search for an explanation. For example, a sudden spurt in crime in a given neighborhood may be the result of an ex-convict moving into the neighborhood.
- Protecting the environment. Spatial data mining is being used to design optimal habitat environments for birds on the endangered species list. For example, by using SDM to identify factors that influence a bird's choice of nesting location, conservation managers can ensure that these factors are preserved.

Developing a Spatial Regression Model

In the previous section, we gave a general overview of SDM and examples of some potential applications of SDM techniques. In this section, we attempt to develop a new model for SDM by incorporating the concept of spatial autocorrelation—the idea that "Everything is related to everything else, but nearby things are more related than distant things." The value of such a model can be illustrated by considering the case of standard regression analysis.

Regression analysis is a standard technique in statistics that is used to quantify the relationship between a dependent variable Y and an independent variable (or variables) X. For example, Y might be the number of crime incidents in different Twin Cities neighborhoods, and X might be the average house value in each of these neighborhoods. The goal of an investigator concerned with identifying crime hot-spots in the Twin Cities might be to build a model that can predict the number of crime incidents likely to occur in a given neighborhood based on the average house value in that neighborhood.

The classic statistical approach to building such a model consists of two steps. First, the investigator describes the relationship between Y and X using a linear regression equation such as the one shown in Figure 1 (a). In this equation, ε represents the residual error. In the classic approach, the value of ε is assumed to be generated from identical and independent distributions (for example, a standard bell-curve distribution). This assumption is based on the presupposition that an error associated with one data-sample observation is not dependent on or related to errors associated with other data-sample observations.

Figure 1. Classic and Spatial Regression Models

$$X \longrightarrow \boxed{Y = \beta X + \varepsilon} \longrightarrow Y$$

(a) The classic regression model to determine the relationship between variable *Y* and variable *X*.

$$X \longrightarrow \boxed{Y = \rho W y + \beta X + \varepsilon} \longrightarrow Y$$

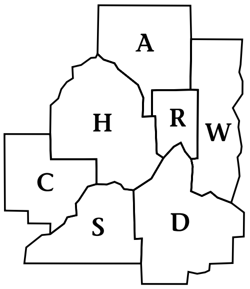
(b) The classic regression model, modified to account for spatial autocorrelation in the dependent variable by inclusion of a "correcting" term.

In the second step, the investigator uses the linear regression equation to calculate the parameter β based on the available data for X and Y. Solving for β in this equation is similar to calculating the slope and intercept of a straight line that represents, in graphic terms, the relationship between the dependent and independent variable. Theoretically, the resulting value for β could then be used as a predictive instrument. In the case of crime incidence in the Twin Cities, for example, β could be used to predict the incidence of crime for a given neighborhood based on average house values for the neighborhood.

Although useful, the classic statistical approach ignores spatial characteristics of the data used to generate the model—specifically, the spatial relationships that might exist among the different locations where the data were collected. This explains why, when classic linear regression is used to model phenomenon where the data samples have a spatial component, the value of the error term ε is *not* always distributed identically and independently across the data set. Instead, the value of ε often varies systematically over space. This is because the independent and dependent variables themselves are spatially related in ways that differentially affect the values of these variables, and in turn the value of ε. In short, because the classic approach ignores the first law of geography and assumes that there is no spatial autocorrelation present among the data, it cannot adequately account for such relationships.

The spatial regression approach attempts to overcome this problem by introducing into the regression equation an additional "correction" term. An investigator using such an approach to construct a model would describe the relationship between X and Y using a spatial regression equation such as the one shown in Figure 1 (b). In this equation, the correction term ρWy attempts to capture the spatial characteristics of the data. The spatial information is encoded using what is called a contiguity matrix, abbreviated W. For example, Figure 2 (a) shows a map of the seven counties that constitute the

Figure 2. Map with Corresponding Contiguity Matrix



A = Anoka County
C = Carver County
D = Dakota County
H = Hennepin County
R = Ramsey County
S = Scott County
W = Washington County

(a) A map of the seven counties in the Twin Cities metropolitan region.

$$\begin{array}{c|ccccccc}
 & A & C & D & H & R & S & W \\
\hline
A & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
C & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
D & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
H & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\
R & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\
S & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
W & 1 & 0 & 1 & 0 & 1 & 0 & 0
\end{array}$$

(b) The contiguity matrix (abbreviated *W*) for the map shown in *(a)*. A non-zero entry in the matrix indicates that the corresponding spatial entities on the map (in this case, counties) are neighbors.

Twin Cities metropolitan area. The contiguity matrix (*W*) for this map is shown in Figure 2 (b). Counties whose borders touch each other are indicated by a 1 in the matrix, and counties whose borders do not touch are indicated by a 0. The contiguity matrix essentially quantifies the first law of geography—that nearby things are more related than distant things — because counties that touch each other are more "nearby" than those that do not touch. The investigator would use the available data to solve for ρ and β, where ρ signifies the degree to which the values in the dependent variable are spatially autocorrelated.

Another shortcoming of the classic approach involves the way that the outcomes of various models are interpreted. Consider the example shown in Figure 3. Here the goal is to predict the actual location of bird nests marked by an *A* in Figure 3 (a). Figures 3 (b) and 3 (c) show the results of two models used to predict the location of nests, indicated by a *P*. Although the predictions of the model used in Figure 3 (c) are clearly closer to the actual location of the nests than are the predictions of the model used in Figure 3 (b), the classic approach would fail to distinguish between the predictions of these two models.

Having argued that there are significant drawbacks to using a classic approach to spatial data analysis, and that the introduction of a spatial regression equation can overcome these problems and lead to more accurate predictions, we now demonstrate the value of a spatial regression analysis framework using the example of ecological habitat prediction.

Applying a Spatial Regression Model to Ecological Habitat Prediction

The availability of accurate spatial habitat models is an important tool for wildlife management, as well as protection of critical habitat and endangered species. Because the underlying process governing the interaction between wildlife and environmental factors is complex, statistical techniques are often used to gain insight into the process on the basis of data collected during fieldwork. Writing in the 1997 issue of *Ecological Modelling,* Uygar Ozesmi and William Mitsch developed a spatial habitat model for predicting the nesting locations of the wetland-breeding red-winged blackbird (*Agelaius phoeniceus L.*) in the Great Lakes region. We will use their model, and the accompanying data they used to generate it, to illustrate the benefits of using spatial regression analysis techniques.

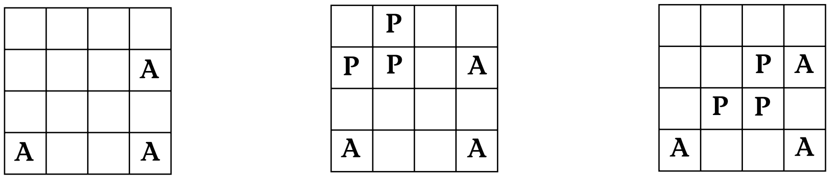
The goal of the Ozesmi and Mitsch study was to create a model able to predict nest locations (dependent variable) based on several explanatory (independent) variables, including distance to open water, water depth, and dominant vegetation durability. Data were collected in 1995 and 1996 from two wetland sites—Darr and Stubble—located on the shores of Lake Erie in Ohio. The data collected at the Darr site were used to build a classic regression model. This model was then tested using the data collected at the Stubble site in order to evaluate the model's predictive power.

To create the model, a uniform grid was imposed on the Darr wetland, and in each cell the values of several structural and environmental factors (independent variables) were recorded. These included water depth, dominant vegetation durability, and distance to open water. For each cell, it was also noted whether or not a red-winged blackbird nest was present (dependent variable). The geometry of the Darr wetland, the locations of the nests, and the spatial distribution of the independent variables are shown in Figure 4.

When the data are mapped, it is apparent that both the dependent and independent variables show a moderate to high degree of spatial autocorrelation. For example, Figure 5 shows hypothetical random spatial distributions for an independent variable and for bird nests

Figure 3. Classic and Spatial Model Interpretation

A = actual nest in pixel P = predicted nest in pixel



			A
A			A

(a) The actual locations of nests in a grid, with locations marked by an *A*.

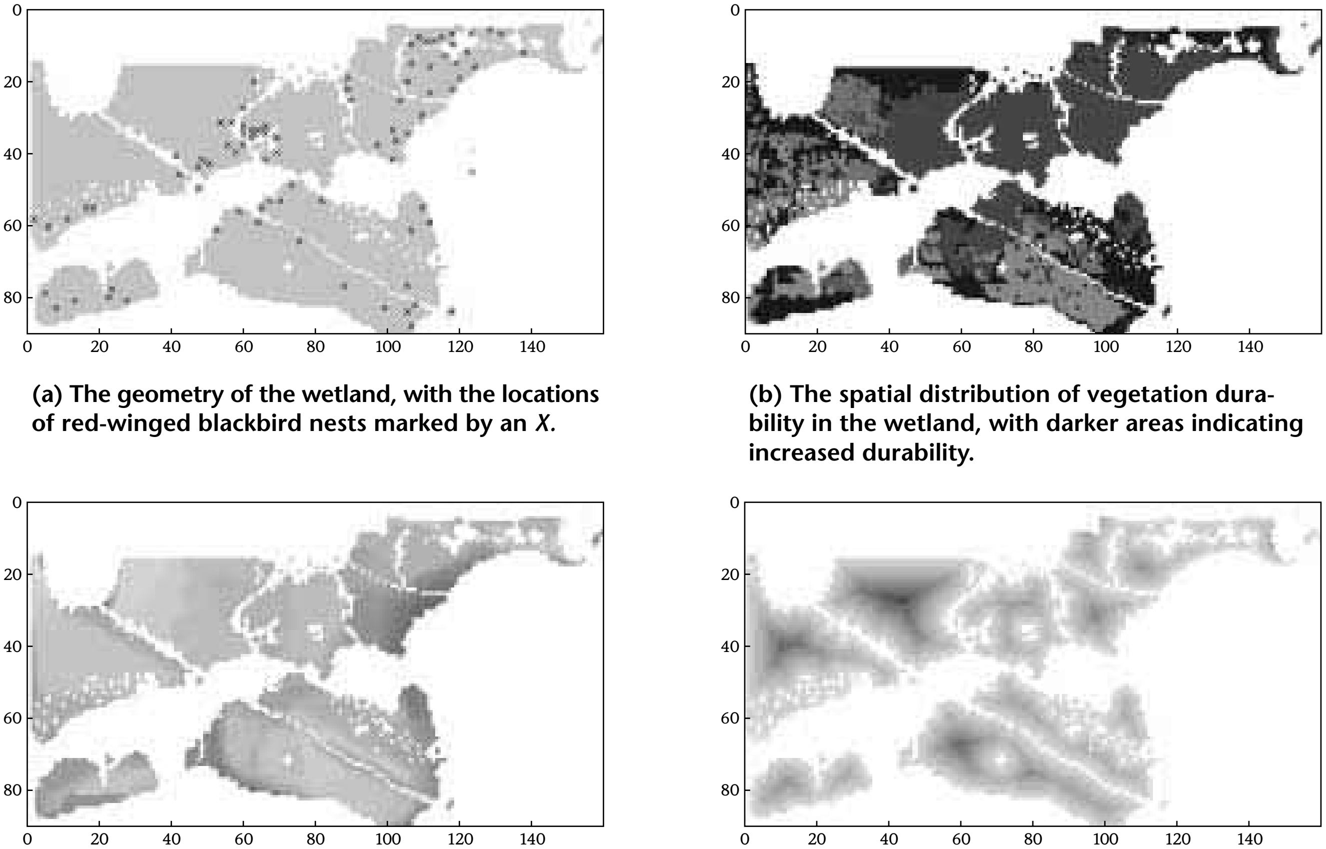
	P		
P	P		A
A			A

(b) The locations of nests predicted by Model 1, with predicted locations marked by a *P*.

		P	A
	P	P	
A			A

(c) The locations of nests predicted by Model 2, with predicted locations marked by a *P*. Although the predictions produced by Model 2 are spatially more accurate than those produced by Model 1, classic measures of classification accuracy are unable to recognize this distinction.

Figure 4. Learning Data Set from the Darr Wetland, 1995



(a) The geometry of the wetland, with the locations of red-winged blackbird nests marked by an *X*.

(b) The spatial distribution of vegetation durability in the wetland, with darker areas indicating increased durability.

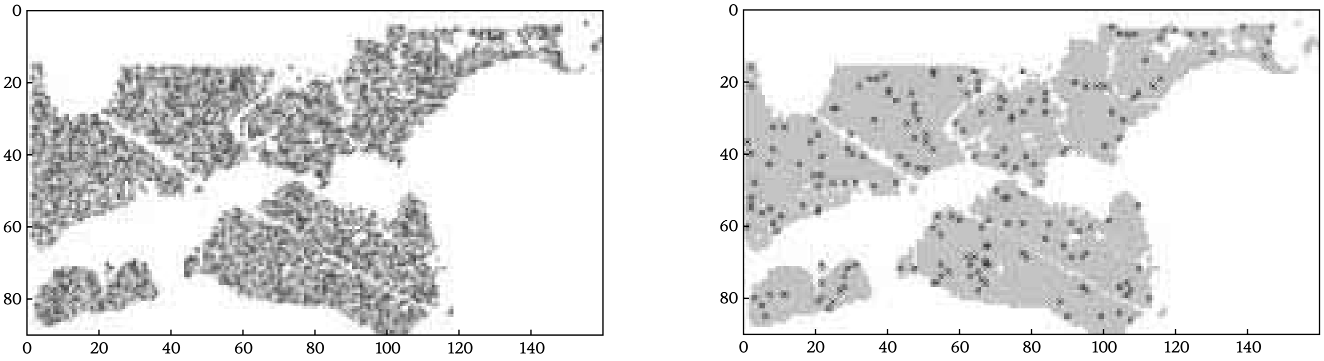
(c) The spatial distribution of water depth in the wetland, with darker areas indicating proximity to water of greater depth.

(d) The spatial distribution of distance to open water in the wetland, with darker areas indicating increased distance.

in the Darr wetland, as might be expected to occur if there were no spatial autocorrelation present among the data. If there were no spatial autocorrelation present, then the values recorded for variables at one location would have no influence on the values recorded in the vicinity of that location, and thus a random (independent and identical) distribution such as the ones pictured here would be expected. However, as one can see from Figures 4 (b), 4 (c), and 4 (d), the distribution of independent variables reveals a gradual variation in values across space, indicating moderate to high spatial correlation among the data. As a result, the random distribution in Figures 5 (a) and 5 (b) are quite different from the actual distribution of nests in the Darr wetland shown in Figure 4 (a).

The goal of our experiment was to evaluate the effects of including the spatial autocorrelation term ρWy in the regression model used in the Ozesmi

Figure 5. Hypothetical Spatial Distributions of an Independent Variable and Bird Nests in the Darr Wetland



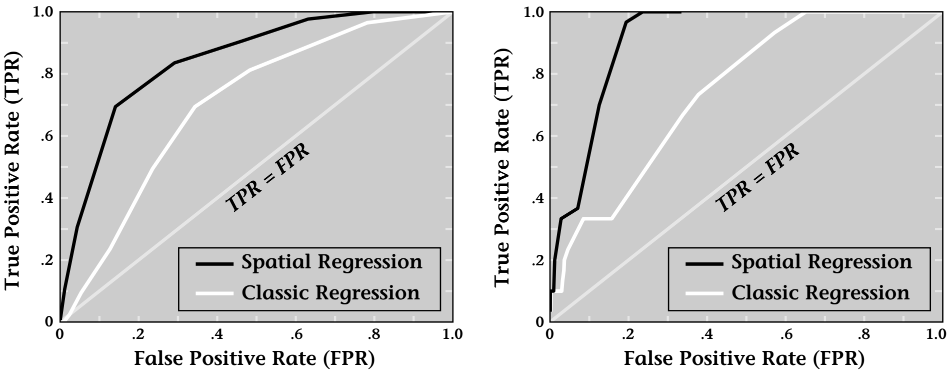
(a) An independent and identical distribution for an independent variable, consistent with the random distribution assumptions underlying classic regression analysis.

(b) A random distribution of bird nests, consistent with the random distribution assumptions underlying classic regression analysis.

and Mitsch study. The 1995 Darr wetland data were used as a learning set to construct both a classic and a spatial regression model. The accuracy and predictive power of the model was then tested using the 1995 Stubble wetland data as a test set. Finally, the predictive ability of the classic and spatial regression models was compared on the basis of receiver operating characteristic (ROC) curves. ROC curves were first used in World War II to distinguish between the radar signatures of friendly and enemy ships. We used them to compare how accurately the two models predicted the location and non-location of red-winged blackbird nests. ROC curves plot the relationship between the true positive rate (TPR) and the false positive rate (FPR) for a predictive model. The TPR represents the proportion of correctly identified nest locations, and the FPR represents the proportion of correctly identified no-nest locations.

The results of these comparisons are shown in Figure 6. For either the classic or spatial model, the higher the curve is above the straight line TPR = FPR, the better the model is at generating accurate predictions. The spatial regression model clearly demonstrates substantial and systematic improvement over the classic model. This is true for both the 1995 Darr (learning) data set and the 1995 Stubble (test) data set.

Figure 6. Receiver Operating Characteristic (ROC) Curves for Classic and Spatial Regression Models



(a) Comparison of the models' performance using the 1995 Darr wetland (learning) data set.

(b) Comparison of the models' performance using the 1995 Stubble wetland (test) data set.

Applications of Spatial Data Mining for Policy Makers

We believe SDM can be a useful tool for researchers and policy makers, both in Minnesota and throughout the nation. As we have demonstrated, a wildlife habitat model based on a spatial regression analysis more accurately describes the relationships between various wetland features and the presence of wildlife nesting sites than does a classic approach to this problem. Such a model could be used by the Minnesota Department of Natural Resources and other environmental agencies to improve conservation efforts for species such as the red-winged blackbird. An SDM approach could also serve as the basis for similar models in other areas of conservation ecology, such as fisheries management or wildlife population control efforts.

Spatial data mining could be applied to other environmental questions as well. For example, one of the greatest challenges currently facing the Metropolitan Council is the issue of how to balance the growth in the Twin Cities metropolitan area with the goal of protecting the remaining natural and agricultural areas in the region. Spatial data mining tools can play a crucial role because they can quickly generate "what if" scenarios regarding the effects of growth and development based on the available data.

Applications of SDM are not limited to environmental issues, of course. For example, SDM techniques could be used by Twin Cities consumer groups to determine whether people of color in the metropolitan region have fair access to credit at local lending institutions. Local law enforcement agencies could use SDM to predict crime hot spots based on existing crime data for the Twin Cities. The Minnesota Department of Transportation could use SDM to plan future highway expansion or light-rail transit construction based on the vast amounts of data collected by traffic sensors embedded in local highways and on-ramps. The potential range of applications is nearly limitless.

In addition, because it can be used with very large data sets, SDM also has the advantage of allowing policy makers and researchers to generate hypotheses across data sets. For example, SDM could be used to combine data from remote sensing, cartographic maps, traffic sensors, and the census in order to generate hypotheses that take into account each of these sources of data.

Spatial data mining is a relatively new term for an approach to data analysis that can be useful in many arenas, including econometrics, environmental management, regional science, geographic analysis, epidemiology, and remote sensing. The potential of SDM lies in its ability to rapidly generate interesting and potentially useful hypotheses that researchers can then verify, modify, and refine using standard statistical techniques. This technique is not a substitute for statistical analysis, but rather a tool that researchers can use to analyze large data sets that have a strong spatial component. Like spatial statistics, which has attained a distinct identity within the field of statistics, we believe SDM can carve out its own niche within the framework of CDM.

■ Sanjay Chawla was a postdoctoral associate in the Department of Computer Science at the University of Minnesota at the time this research was conducted. He is now with Vignette Corporation. His research interests include spatial database management and data mining. Shashi Shekhar is an associate professor in the Department of Computer Science at the University of Minnesota, and an active member of the Army High Performance Computing Research Center as well as the Center for Transportation Studies, both at the University of Minnesota. His research interests include databases, geographic information systems, and intelligent transportation systems. Wei Li Wu is a doctoral student in computer science at the University of Minnesota. She is working on her dissertation in spatial data mining. Xinhong Tan was a masters student in computer science at the University of Minnesota at the time this research was conducted. She is now a software engineer at Parametric Technology Corporation in the Twin Cities.

This study was supported by an interactive research grant from CURA and the Office of the Vice President for Research, University of Minnesota. Interactive research grants have been created to encourage University faculty to carry out research projects that involve significant issues of public policy for the state and that include interaction with community groups, agencies, or organizations in Minnesota. These grants are available to regular faculty members at the University of Minnesota and are awarded annually on a competitive basis.

The study was also sponsored in part by the United States Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory Cooperative Agreement No. DAAH04-95-2-0003 and Contract No. DAAH04-95-C-0008. The contents of this study do not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

Project Awards

To keep our readers up-to-date about CURA projects, each issue of the *CURA Reporter* features a few capsule descriptions of new projects underway. The projects highlighted in this issue are made possible through CURA's Communiversity Personnel Grants. The grants are awarded twice each year to grassroots organizations in the community. Each grant supports the extra personnel needed by local organizations, usually by providing an advanced graduate student who works directly with the organization receiving the award. The grants are competitive, and organizations working with people of color are favored. The projects described here are only a sampling of projects that will receive CURA support during the coming year.

■ **Empowering Latino Youth.** La Escuelita is a youth-serving organization that focuses on the development of Latino youth, specifically in the areas of academic enrichment, service-learning, recreation, and cultural empowerment. A graduate student will research literature on youth development and Latino youth-serving organizations, and write a report that includes a framework for evaluating such organizations.

■ **Resources for American Indian Women.** The Minnesota Indian Women's Resource Center in the Phillips neighborhood of Minneapolis provides services to American Indian women and their families in the Twin Cities metro region and on several area reservations. A graduate student will assist the center's staff in gathering resource material and developing new curricula in several areas, including culturally-based chemical dependency treatment, parenting, domestic violence, and sexual assault.

■ **Legal Advocacy for Battered Women.** Based in St. Paul, the Battered Women's Legal Advocacy Project works to eliminate oppression of and violence against women. Project staff will work with a graduate student to assess the status of battered female criminal defendants in Hennepin County and southwestern Minnesota, and to identify how many were in need of legal representation in the year prior to the project start date. The research project will document the nature of the representation provided, the defense strategies employed at trial, the length of sentences, and the adequacy of representation.

■ **Services for American Indian Children at Risk.** The Indian Child Welfare Law Center is a nonprofit agency that provides culturally appropriate legal services to American Indian families in the child protection system. The center serves as a community development resource for education, advocacy, and public policy, and helps connect American Indian children at risk with an existing network of service people and programs. The organization will have a graduate student update and reorganize its information database in order to assist the center in setting goals for the future.

■ **Preventing Child Abuse and Neglect in the Latino Community.** Unidas Para Los Ninos (United for Children) is a coalition of individuals and organizations formed to prevent child abuse and neglect in the growing metro area Latino community. A graduate student will research culturally appropriate resources on child abuse and neglect for use by social service programs and health practitioners working with Spanish-speaking families in the area.

■ **Real Estate Foreclosures.** This project will focus on real estate home loan foreclosures by subprime lenders throughout the Twin Cities metro area during the 1990s. A graduate student will work with Minnesota ACORN (Association of Community Organizations for Reform Now)—an advocate for low- and moderate-income families in the Twin Cities—to determine whether there is a correlation between foreclosures and non-conventional mortgages, and whether the hardest hit areas are neighborhood minority communities with a stable rate of home ownership.

■ **Racial and Low-Income Housing Impact Statement.** The Metropolitan Interfaith Council on Affordable Housing (MICAH) mobilizes congregations and people of all faiths to advocate for public policies that increase the supply of affordable housing in the Twin Cities metro area and promote fair housing for all residents. A graduate student will research and review existing literature on affordable housing, identify particular housing policies and practices in the Twin Cities area, and work with MICAH staff to prepare a racial and low-income impact statement regarding metro area housing policies.

■ **Business Plan for Social Service Organization.** Damiano is a community-based organization that provides social service programs to low-income families in the central hillside neighborhood of downtown Duluth. Using a local architectural firm's feasibility study as a foundation, the organization will work with a graduate student to develop a business plan for development of rental space in the Damiano Center, where the organization is housed.