In Environmental Modeling with GIS
M. Goodchild
B. Parks
L. Steyaert
Oxford U. Press
1993

VI Spatial Statistics

DONALD E. MYERS

Spatial statistics and geographic information systems are natural partners, especially for the analysis and modeling of environmental data. The variety of problems that are addressed by spatial statistics and ways in which geographic information systems can aid in the manipulation of environmental data are amply illustrated in the chapters presented in this section. Environmental data sets have two characteristics that set them apart from many other kinds. First they are nearly always multivariate; that is, there is more than one variate or analyte of interest, and these are correlated in some sense. Second, each data value is associated with a location either by specific coordinates or by association with some area or volume. This positional association is also normally manifested in another way, namely through some form of spatial correlation. At least three perspectives on the way in which these two characteristics are utilized are presented in this collection. Subsequent sections of this introduction will identify where these occur in the various chapters and provide an overview.

SPATIAL STATISTICAL MODELS

The objective in constructing an environmental model may vary considerably, but a model is most often used to explain or predict. Cressie and Ver Hoef (Chapter 40) point out that environmental models not only need to be spatially based but should incorporate time and often need to incorporate the relationship between an organism and its environment. Spatial statistics seems to be a new field, but in fact it has its roots in classical statistics and in particular in the work of Fisher, Yates, and Whittle. Several examples of models are given, especially in the context of sampling design. In addition to an overall discussion of the relevance of spatial statistics to environmental modeling, several examples are given to illustrate this relevance. The first example uses a random function model and pertains to acid rain, that is, wet deposition of hydrogen ions in the Eastern U.S. resulting from the dispersal into the atmosphere of various sulfurous and nitrous pollutants produced by industry and transportation. This is linked to the identification and prediction of ecological effects. The methods used in this example are a part of what is known as geostatistics. A more complete overview and description of geostatistics is provided in Cressie's Chapter 41. An example using data for the percent vertical cover in a dolomite glade illustrates the use of a spatial lattice model. Spatial point pattern methods are illustrated using data from a census of longleaf pines in southern Georgia. In this model, time is incorporated into the birth, death, and growth process.

Geostatistics is based on the use of a random function model wherein the data are viewed as a nonrandom sample from one realization of the random function. It had its origins in applications to mining, hydrology, meteorology, and forestry, although the developments in the latter two areas proceeded in slightly different ways. Cressie describes the importance of the spatial correlation function, which is most often given in the form of a variogram. Difficulties pertaining to and methods for estimating the variogram are described. The random function model leads easily to a regression form for the estimator, known as kriging, for spatial prediction or estimation. An extension useful in the presence of a nonstationarity, known as universal kriging, includes the thin plate spline as a special case. It is shown that nonpoint data are easily accommodated by the model and the model is easily adaptable to simulation. The final section is devoted to multivariate geostatistics wherein both spatial and intervariable correlation are incorporated in the model. Matheron and coworkers derived a linear predictor for one variate utilizing the spatial correlation for that variate and the intercorrelation with other variates. The work of Myers, which provides the general setting, shows that the single-variate case is subsumed in the general case and provides the natural extension of the single-variate case.

Geostatistical methods explicitly incorporate the position coordinates into the random function model. Anselin (Chapter 46) models spatial correlation in a manner that does not explicitly utilize coordinates. A heuristic description might be that spatial correlation means that values at

locations close together are more correlated than values for locations far apart. In geostatistics this concept is used to derive an estimator for the values at unsampled locations. Anselin uses the same concept to characterize clusters or patterns. In the geostatistical model, spatial correlation is quantified by the spatial correlation function, that is, the variogram or covariance. In the discrete space autoregressive model, spatial correlation is quantified by contiguity or weight matrices. For the latter, the data locations are considered to be a subset of a regular or irregular grid or lattice. An example is given using data from the Global Change Database (Chapter 36) from an area around the border between the Central African Republic, Sudan, and Zaire. Four variables, GREEN (greenness vegetation index), TEMP (temperature), ELEV (modal elevation), and PREC (precipitation), are used. One part of the analysis is concerned with examination of the clustering of GREEN versus the clustering of the other variables, and the second part of the analysis considers GREEN as the dependent variable in a spatial regression using the remaining as explanatory variables. Several statistics useful for testing for spatial correlation are described, and there is a comparison between the use of least squares and maximum likelihood for determining the coefficients in the spatial regression equation. Anselin has incorporated these methods into a software package called SPACESTAT.

Conventional inference techniques are based on an assumption that the sample is selected from a hypothetical population and the inferences pertain to the parameters or characteristics of this population. More realistically the sample is commonly selected from a finite universe. Overton (Chapter 47) describes the use of probability sampling, which was incorporated into the National Surface Water Survey (NSWS) and is an integral part of the current EPA initiative, EMAP. As an example consider the universe of lakes in the NSWS. A probability sample is a subset of this universe selected in such a manner that for each element of the sample the probability of its having been selected is known and this probability is positive. A representation of the universe, called a frame, is used to select the sample. For example, a frame might simply be a list of the elements of the universe or in a spatial context it could be a map. The Horwitz–Thompson theorem provides assurance that certain population parameter estimators based on probability samples are unbiased and determines the variances of these estimators. It is shown that in the case of model-based inferences the usual estimators must be replaced by weighted estimators in order to maintain consistency. In the design of EMAP, the use of probability samples is extended to spatial problems by first overlaying a triangular point grid on the map and then perturbing the grid in a random manner. This provides a sampling grid with the same configuration but randomly positioned. Any small region of fixed area is equally likely to contain a sample grid point.

APPLICATIONS

While most GIS include some routines or programs for the analysis of data, their principal thrust as yet is in the graphical manipulation and presentation of the data. In particular, GIS do not as yet really incorporate any spatial statistical components. Conversely the standard statistical packages include neither spatial statistical routines nor the capabilities of a GIS. The remaining chapters in this section on spatial statistics illustrate the use of a GIS as an aid to spatial statistical analyses or the use of spatial statistics to complement the use of a GIS. These are important because of the lack of an adequate interchange in the literature pertaining to these two fields. Only one of the papers explicitly uses a GIS in the analysis. In general all the papers begin with a geostatistical perspective, and that significantly affects the way in which the connection with a GIS is presented. Englund (Chapter 43) is concerned with the use of simulation, whereas the other chapters pertain to estimation and modeling.

The random function model in geostatistics can be thought of as a collection of realizations together with a probability assignment, although the data are a sample from only one realization. The kriging estimator is a smoother, and the variability exhibited by the data together with the estimated values is less than that of the data. Simulation of additional realizations is a method for reproducing the variability and hence is useful for designing sampling plans and for planning in general. The simulations can be conditioned to the data by kriging. There are several different algorithms commonly used for simulation; Englund has used the sequential Gaussian method. Reproducing or characterizing this variability is useful in determining the reliability of maps produced by kriging. Englund illustrates this by considering two "layers" in a GIS that have been produced by sampling different variates, followed by kriging. In this example neither variate is of interest alone, but certain characteristics of the intersection of the layers are of interest (for example, areas where both variates have high values). If these layers are produced only from the data and interpolated by kriging or some other method, there is still the question of the reliability of the resulting intersection map. Englund shows how to use simulation to quantify and characterize the reliability.

Ver Hoef (Chapter 45) considers three different methods to predict spatial-cover abundance for a glade in the Missouri Ozarks. The first approach is a classical regression with cover abundance as the response variable and shade as the explanatory variable. In the second method, abundance at one location is predicted using only the values for abundance at nearby locations and incorporating the spatial correlation (kriging). The third method is a combination of the other two. Residuals from the regression are used in kriging; that is, the method attempts to separate the dependence on the explanatory variable and

the spatial dependence of abundance on itself. The three methods are compared by using cross-validation. Sequentially, each data value for abundance is deleted and estimated using only the remaining data. Then the mean-square estimation error is computed, and this statistic is used to discriminate between the methods. Note that the use of the term universal kriging, in this chapter, is not quite the same as the usage in the geostatistical literature. What is called universal kriging in this chapter has been called kriging with external drift in the geostatistical literature; it is related to but not the same as co-kriging. The form of universal kriging used herein is found to be superior to either regression or kriging alone when predicting cover abundance.

A similar approach is used by Jager and Overton (Chapter 42) in their study of spatial patterns for acid neutralizing capacity (ANC) for lakes in the Adirondacks of New York. The data are taken from the National Lake Survey and constitute a probability sample. The lakes are stratified according to alkalinity levels (low, medium, and high). Using elevation and pH as explanatory variables, two regression equations are obtained for LANC, where LANC is the base 10 logarithm of (ANC + 150); one regression is used for the low and medium alkalinity levels and one for the high. In each case LANC is regressed on precipitation pH and elevation. Using the regression residuals, sample variograms for LANC were computed and modeled. The authors then infer spatial patterns for LANC using spatial patterns of the explanatory variables and the regression equations.

Rhodes and Myers (Chapter 44) consider a different application of geostatistics to lake survey data. The Eastern Lake Survey–Phase I provided data on a number of variates thought to be relevant to predicting the acidification of lakes as related to effects on marine life in the lakes. Sampled lakes had been selected as a probability sample with one water sample (6.2 liters) taken from each lake irrespective of surface area (lakes smaller than 5 hectares were excluded from the sample). In a geostatistical analysis each data value is either associated with a point or is the spatial average value over an area or volume. Intralake variability is important both in the estimation and modeling of the variogram as well as in the subsequent kriging step. The limitation of one sample location per lake prevents direct estimation of this intralake variability and has a significant impact on the set of sample locations in a kriging neighborhood. Short-range variability will contribute to a nugget effect in the sample variogram when there is a lack of sample pairs for short distances. The magnitude of the nugget can at least be used as a proxy for estimating the intralake variability. The GIS package GRASS 3.1 was used to select pseudo sample locations in some of the larger lakes. This was done by using the GIS to overlay a grid on a map of the lakes and thus to identify grid points within a given lake. To simulate variability within the lakes, pseudo sample locations within a given lake were assigned a value obtained from the lake sample value by adding a random multiple of the square root of the nugget of the variogram. The effect of incorporating these additional sample locations, and thus simulating intralake variability, is evaluated in several ways.

THE FUTURE

While there is clearly interest in merging spatial statistics into GIS, there is little consensus on what techniques or routines should be included or how they would be used. Some interfaces have already been built, including one between the statistical package S+ and the GIS GRASS. Although S+ is not always viewed as a spatial statistics package per se, it is possible to incorporate spatial statistical functions, as well as to take advantage of other graphical features. It is to be hoped that such connections will encourage statistics departments to implement and teach GIS software. Papers on spatial statistics are appearing more often in the statistical literature and at statistics meetings.

One of the reasons for the delay in merging these tools is that GIS is as yet relatively unknown in the statistical community at large and in much of the spatial statistics community in particular. This is in part attributable to the development of GIS in the context of geography and particularly an emphasis on vector-based GIS and hence on vector data. Many spatial statistical methods would more naturally relate to raster data. One aspect of the way statistics functions in an academic setting is pertinent to these problems. Younger statistics faculty need to establish themselves in order to obtain promotion and tenure. This frequently inveighs against significant collaboration in other disciplines such as working in the interface between spatial statistics and GIS. Similarly, an established statistician whose research has been limited to mathematical statistics does not have much incentive to become involved in interdisciplinary work. These trends deserve some attention in the statistical community.

A second possible reason is that spatial statistics, particularly geostatistics and probability sampling, are relatively unknown in the GIS community. This is in part attributable to the origins of GIS, which were principally in geography and closely related fields rather than those giving rise to geostatistics. This is reflected in the initial emphasis on vector-based GIS. As the technology and the software begin to blur the distinction between vector- and raster-based systems, the base for applications will increase. Environmental modeling needs will accelerate this trend.