ORNL/TM-1999/294

SPATIAL MODELING OF LANDSCAPE PATTERN

H. I. Jager, M. Kramer, and W. S. Overton

Environmental Sciences Division

Publication No. 4957

ORNL/TM--1999/294

ENVIRONMENTAL SCIENCES DIVISION

SPATIAL MODELING OF LANDSCAPE PATTERN

H. I. Jager
Environmental Sciences Division

M. Kramer
U.S. Bureau of the Census, Washington, D.C.

W.S. Overton
Oregon State University, Corvallis, Oregon

ESD Publication No. 4957
December 1999

Prepared for
U.S. Environmental Protection Agency
Office Of Research And Development
under IAG DW89934074-2

Prepared by
Environmental Sciences Division
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831
managed by
LOCKHEED MARTIN ENERGY RESEARCH CORPORATION
for the U.S. DEPARTMENT OF ENERGY
under contract number DE-AC05-96OR22464

CONTENTS

List of Figures

Figure **Page**

ACKNOWLEDGEMENTS

This work was supported by the Office of Research and Development, U.S. Environmental Protection Agency under Interagency Agreement DW89934074-2 with the U.S. Department of Energy under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.. The ideas that led to the approach here of "focusing on the pattern rather than the noise" originated from discussions with Dr. Scott Overton at Oregon State University. Dr. Carolyn Hunsaker provided advice and support from her experience with regional analysis of Adirondack lakes. The manuscript was improved based on the insightful reviews and discussions with Drs. Carolyn Hunsaker, Jon Nyquist, Bob O'Neill, Tony King, Marie Josée Fortin, and Mike Sale. This paper does not necessarily reflect the opinions of the U.S. Environmental Protection Agency; no official endorsement should be inferred.

SUMMARY

When the goal is a spatial model of an environmental variable, both spatial autocorrelation and correlations with other environmental variables can be useful predictors. In this paper, we describe a common-sense approach for modeling spatial patterns. Spatial-statistical models typically have a deterministic and a random part, and the problem has always been deciding how to define the partition between the two parts. Here, we provide a practical set of constraints that partitions the spatial model in a useful way. The deterministic component explains large-scale spatial patterns. This "grand" spatial pattern is inferred from environmental predictor variables that are available everywhere in the spatial region of interest (e.g., remotely-sensed data). The random component supplements this large-scale pattern with small-scale, local deviations from the grand pattern that are obtained by kriging. We define some "common-sense" expectations for the spatial autocorrelation structure and use these as constraints to help us build a multivariate model of spatial pattern. This approach developed out of the need for a spatial model to predict the acid neutralizing capacity (ANC) of unsampled lakes in upstate New York. We used the model to produce regional estimates of lake chemistry for all unsampled lakes based on acid deposition, lake elevation, and geologic region, supplemented by a local correction obtained by kriging the ANC residuals measured in neighboring lakes.

1. INTRODUCTION

Spatial patterns and relationships among species and environmental variables that vary together in the natural landscape have always been a key focus of ecological research. Multivariate techniques such as multiple linear regression have traditionally been used to describe relationships between the available predictor variables and the environmental variable of interest (Pauly, 1980 is one of many, many examples). Unfortunately, these models remove the spatial context of the environmental data. More recently, geostatistical models have been borrowed from the field of geology and applied to environmental problems (e.g., Burrough, 1983). These models are also sub-optimal in that the interactions between multiple environmental variables and/or species are neglected. In this paper, we show how geostatistical tools can be used to improve a multiple linear regression model by taking the spatial autocorrelation between sample locations into account and by providing guidance in variable selection.

The geostatistical interpolation technique, kriging, was originally developed by the mining industry to guide exploration for ore deposits (Journel and Huijbregts 1979). In mining applications, kriging is used to predict the spatial distribution of a regionalized variable (RV), most often the ore grade of a mineral. The principle motivation for kriging in geology is that all attributes of the geology are costly to sample, and so one must make do with fewer samples. This is true for the main RV of interest as well as for its potential predictors. As a result, geologists have traditionally modeled spatial patterns in their RVs as a function of spatial location (a polynomial surface is fitted to spatial coordinates to represent average ore grade), rather than as a function of more "meaningful" or causal geologic predictor variables.

While many of the problems facing environmental science are similar to mining problems, most problems are different enough to warrant some modifications to the standard geostatistical tool kit. Geostatistical tools have been borrowed for environmental studies because of the recognition that many environmental properties are autocorrelated in space (LaJaunie 1973; Robertson 1987; Legendre and Troussellier 1988; Legendre *et al*. 1989). This means, simply, that locations that are close together are more similar than those that are far apart. The spatial characterization of toxic waste concentrations in soils (Meyers and Bryan 1984; Simpson 1984) is an example of an environmental application that has a direct parallel in mining. Generally, there are differences between the problems of mining and geology and those of environmental science and ecology. In contrast to geologists, environmental scientists often have some level of relevant and spatially extensive surface information (e.g., on land use, elevation, and climate) available to aid in predicting large-scale spatial patterns.

Another distinction is that environmental landscapes change on shorter time scales than do geologic formations. In many instances these temporal changes that are of interest (Fedorov 1989; e.g., the recent interest in landscape responses to global climatic change). Because natural landscapes are relatively dynamic, spatial models capable of predicting future landscapes based on environmental relationships described in the past are most useful. Models based on environmental predictors of spatial pattern are more robust to temporal changes than models based simply on geographic location. To illustrate this, suppose that smallmouth bass adults prefer nesting sites protected from high velocity and the potential effects of flooding.   A model based on location will be fine for describing patterns in nest density as long as the flow stays the same as it was at the time measurements were taken. However, a model that predicts the spatial pattern in smallmouth nests by using easily-measured variables such as stream bed elevation and in-stream cover that correlate with velocity, should capture the spatial features of the population even if the velocity field shifts in response to changing stream flow.

In this paper, we demonstrate a spatial analysis approach that predicts the spatial distribution of environmental data by combining an estimate of the large-scale pattern obtained from multivariate relationships among environmental variables and local deviations from this pattern estimated by kriging. Because the predicted spatial pattern is supplemented by local deviations obtained by kriging, we refer to our program for model development as "PATTERN+". We will begin by providing a brief statistical background and defining notation used in the paper. This is followed by a review of variations of kriging in the geostatistics literature that are related to our method. In the last half of the paper, we describe the use of PATTERN+ to predict lake ANC for lakes throughout the Adirondack region of upstate New York.

2. METHODS

2.1 Statistical background and definitions

The general form of a spatial statistical model for the regionalized variable Z at location **x** is:

$$Z(x) = M(x) + e(x), \qquad (1)$$

where $M(\boldsymbol{x})$ is a function giving the mean at $\boldsymbol{x}$. The residuals (e) are second-order stationary. Second-order stationarity of ε asserts that the values of $e(\boldsymbol{x})$ are realizations from distribution(s) with a common mean and variance that is independent of location $\boldsymbol{x}$. This implies (1) that the expected value of the residuals is zero, E[e] = 0 and (2) that the covariance function for $\boldsymbol{e}$ describing the covariance between values at two locations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ that are separated by distance $h = |\boldsymbol{x}_i\text{-}\boldsymbol{x}_j|$ is a function only of the distance between locations (and not absolute position $\boldsymbol{x}$). If M($\mathbf{x}$) is known, then the true errors $e(\boldsymbol{x}) = Z(\boldsymbol{x}) - M(\boldsymbol{x})$ are known to satisfy the first condition of stationarity, and if the second condition is also met, we can proceed with kriging the $e(\boldsymbol{x})$.

In practice, neither $M(\boldsymbol{x})$ or $e(\boldsymbol{x})$ is known and we have to estimate $M(\boldsymbol{x})$ by $M^*(\boldsymbol{x})$, forming estimated residuals $R^* = Z(x) - M^*(x)$ which are used in kriging. This is known as residual kriging. For a review of approaches to kriging nonstationary data, including residual kriging, see Jernigan (1986). The model for Z given in Equation (1) views the observed landscape as one possible realization of many alternative landscapes that might have been generated from the underlying statistical process. The "true" process is not identifiable from a single observed landscape. Thus, there is no single spatial-statistical model that is correct (see Myers 1989). Because statistics cannot tell us "the correct model," it seems reasonable to choose a model that meets the pragmatic goals of environmental research. In many cases this translates into predicting values of a RV at unsampled locations on the basis of relevant and accessible auxiliary information.

The semivariogram links the covariance between values of an RV at two locations to the distance separating them. The semivariance increases as the covariance decreases, with a maximum value at the sill. The sill (Figure 1) represents the average variation of two locations far enough apart to be statistically independent.

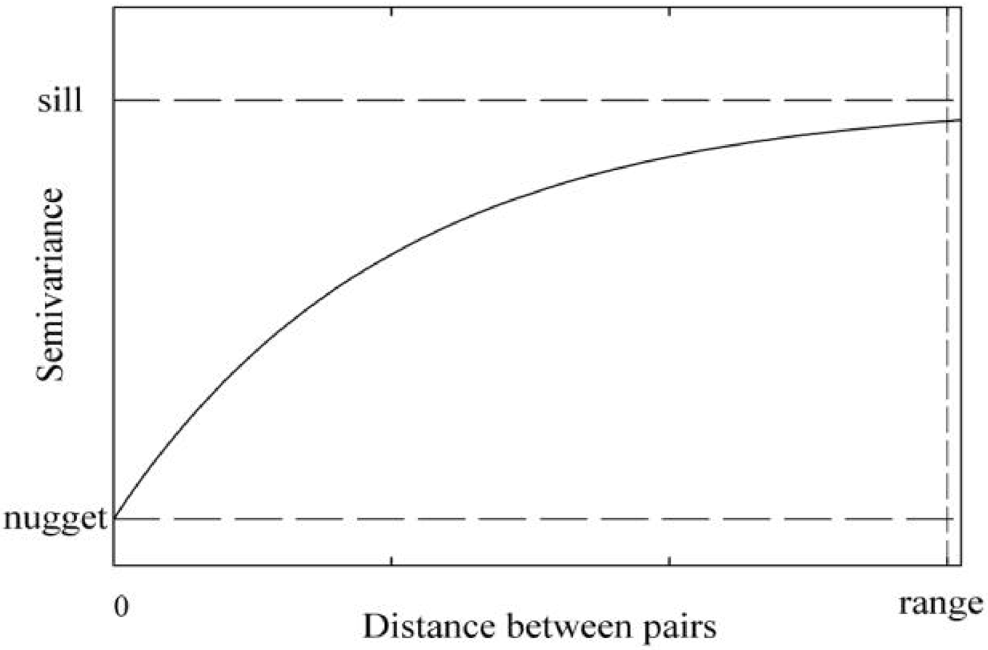

Figure 1 The semivariogram models the average squared difference between pairs of points as a function of the distance between them. Three parameters (nugget. sill. and range) are shown.

The semivariogram function is expressed as *g(h) = C(0) - C(h),* where *C(0)* is the sill, *C(h)* is the covariance between values of the RV at the two locations, and h is the distance between the two locations. The minimum distance between locations at which independence occurs is the range. The nugget represents the variation between locations in very close proximity. If there is a nugget effect, the semivariogram function is greater than zero even at very short distances but *g(0)* is defined to be zero.

The empirical semivariogram is constructed as follows: First, the squared differences between all possible pairs of locations in the data are calculated. The pairs of locations are then grouped into distance classes, (e.g., 0–10 km). Next, calculate the average of the squared differences of all pairs for each distance class. Each average (divided by two — hence the term semivariance) is plotted against its distance class. Empirical semivariograms can be constructed separately for pairs of samples oriented in different directions (e.g., N-S or E-W). If the semivariograms differ when constructed in this fashion, the RV is said to be anisotropic; if not, it is isotropic.

The empirical semivariogram is used to determine the appropriate parametric semivariogram model and to estimate its parameters. One can then use this parametric semivariogram model, rearranged as $C(h) = C(0) - g(h)$, to estimate the covariance between any two locations, whether they are part of the original sample or not, if the distance between them is known. A variance-covariance (VC) matrix can be built for all locations of interest. Covariances between all pairs of locations are on the off-diagonal, variances on the diagonal.

2.2 Iterative residual kriging with an external drift

The PATTERN+ model includes both a deterministic component and a random component of the regionalized variable *Z*, as in Equation (1). The deterministic component, which represents changes in the mean of Z in space, is usually referred to as "drift" in geostatistics. Instead, we use the term "explained spatial pattern" (or "pattern") to emphasize the relationship with traditional approaches to modeling spatial organization in ecology and to avoid confusion with time-series terminology. There is not enough information contained in one observed landscape to uniquely identify both a deterministic pattern component and a random component. To estimate one component it is necessary to assume that the other is known. This induces a bias in the resulting parameter estimates that can be reduced by using an iterative procedure and can be eliminated completely by using separate subsets of data for the estimation of each component. Rather than assume one model or the other to be known, we iterate between the estimation of semivariogram and pattern model parameters (Neuman and Jacobson 1984). In contrast to Neuman and Jacobsen, we develop the pattern model using relevant environmental variables rather than spatial coordinates. Similar "drift" models have been described by Delhomme (1978), Stein (1984), Ahmed and DeMarsily (1987), and VerHoef (1993).

The PATTERN+ procedure is outlined in Figure 2. The first phase selects candidate explanatory variables that might be incorporated into an external pattern function $F(\boldsymbol{Y})$. These variables are screened using a stepwise procedure that considers both the predictive capability and pattern characteristics of the candidate variables. The residuals are used to construct directional semivariograms to verify that large-scale spatial pattern in the mean has been removed. If not, additional explanatory variables are added to the pattern model. The second phase of the procedure is a refinement step in which the coefficients of the pattern model and the semivariogram parameters of the residuals are adjusted in turn until convergence is reached for both.

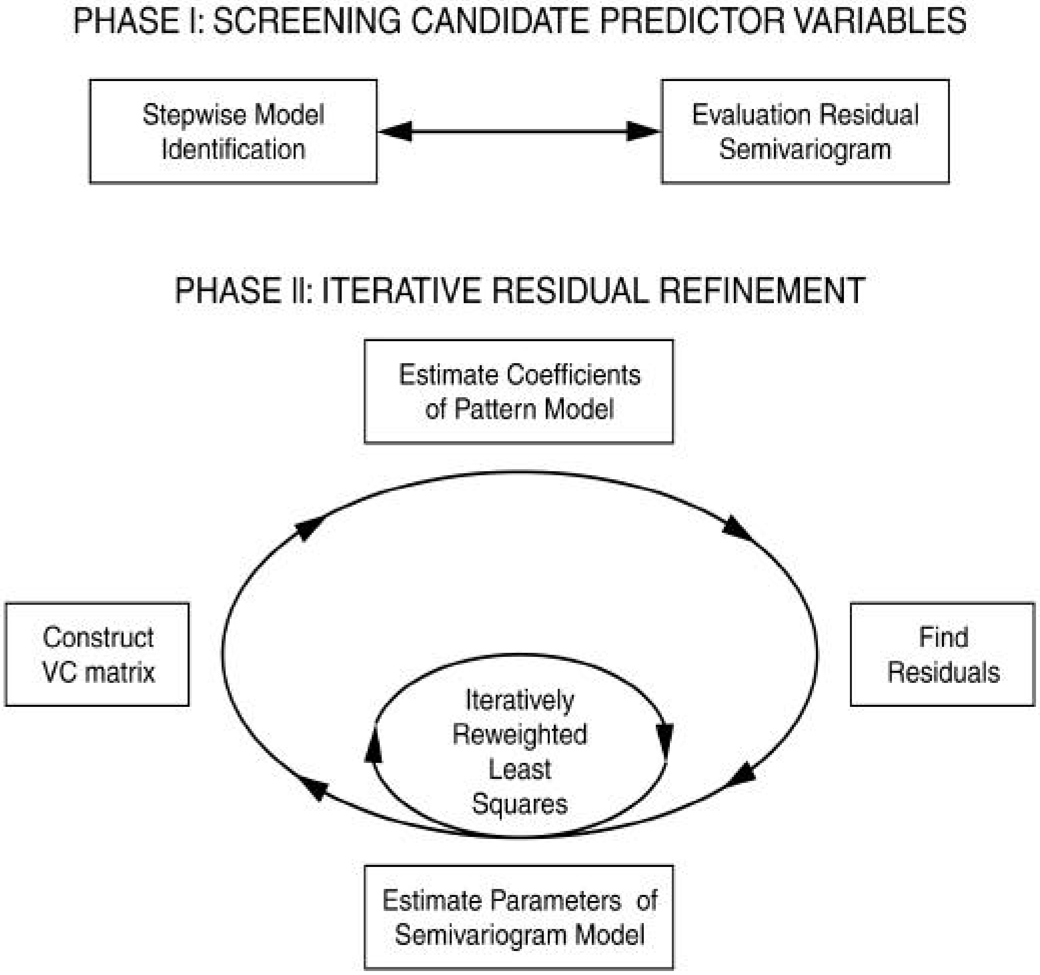
PHASE I: SCREENING CANDIDATE PREDICTOR VARIABLES
Stepwise Model Identification
Evaluation Residual Semivariogram
PHASE II: ITERATIVE RESIDUAL REFINEMENT
Estimate Coefficients of Pattern Model
Find Residuals
Iteratively Reweighted Least Squares
Construct VC matrix
Estimate Parameters of Semivariogram Model

Figure 2 Diagram of the spatial modeling procedure used to develop a pattern model of lake ANC and to characterize the spatial autocorrelation in the residuals.

In addition to modeling differences in mean *Z*, the variances within different subregions or non-spatially defined sub-populations can be equalized at each step of the analysis (Jager and Overton 1993). This can help to meet the homogeneity of variances assumption required by the multivariate pattern model without data transformation and also allows pooled estimation of the semivariogram parameters from scaled residuals.

2.3 Guidelines for modeling environmental pattern

We identified three key features or constraints for developing an environmental pattern model that are not usually imposed on a non-spatial multiple regression model: (1) there is some spatial variation in the mean (pattern), (2) this pattern is spatially smooth relative to the RV itself, and (3) the removal of a proposed pattern model leaves residuals that appear to be pattern-free (according to criteria listed later).

According to the first part of this definition, spatial patterns explained by environmental predictors can be linear in space (e.g., environmental gradients), they can have some other functional form, or they may be discontinuous in space. The important distinction is that average behavior of the RV appears to depend on location in a way that is predictable based on the spatial distribution of environmental features. Important spatial distinctions may exist that are not necessarily monotonic or even continuous. For instance, abrupt changes in geology, elevation, or land use suggest that different patches are best modeled as originating from processes with different means. As an illustration, consider the density of spruce-budworm infestation on trees as an RV of interest. The expected (mean) density is likely to be higher in patches of coniferous forest dominated by spruce trees than it is in patches of forest dominated by hardwoods. In this case, the categorical variable *land use* influences the mean pattern, but the relationship with space is not one that can be described as a simple function of spatial coordinates, particular if the goal is a generally-applicable model that can be applied in different regions.

According to our second criterion, we expect the mean to vary relatively slowly, if not continuously, in space, relegating higher frequency fluctuations to the stochastic portion ε of the spatial model Equation (1). Referring back to the spruce budworm example, the mean should represent intermediate- and large-scale spatial variation in budworm densities but should not reflect microhabitat features such as locations of individual spruce trees.

Our last criterion is to ensure that removal of a proposed pattern function results in residuals with a semivariogram that appears to be pattern-free. Geostatistical analysis can be used to guide the predictor selection process by paying attention to the semivariograms of the main variable, candidate predictor variables, and the residuals of each proposed model. For example, if anisotropy is evident in the semivariogram of the RV (a difference between the empirical semivariance curves for pairs of locations oriented in different directions), then candidate environmental predictors with the same directional anisotropy can most likely help in explaining the pattern portion of the RV. Pattern-free errors meet the stationarity assumptions of geostatistics.

Taken to its logical extreme, these guidelines suggest that the ideal spatial model will have "white" residuals (no spatial autocorrelation). In practice, we believe that the complexity and the non-decomposability of scales of influence on an ecological RV will usually leave some amount of unexplained local spatial autocorrelation that is most conveniently treated in a statistical manner.

3. A Case Study: Lake Chemistry in the Adirondack Region

Data from the U.S. Environmental Protection Agency Eastern Lake Survey data for upstate New York (Linthurst *et al.* 1986) were used to demonstrate the PATTERN+ approach for lake acid neutralizing capacity (ANC) expressed in μeq/L. Our goals were to obtain (1) a model for the mean of lake ANC as a function of environmental predictors and (2) a model of the spatial autocorrelation structure of the residuals after removing the spatially varying mean. The final estimate of lake ANC for any unsampled lake will be the sum of two estimates: (1) the local mean (based on values of environmental predictors at the lake) and (2) the residual (based on local residuals measured at neighboring sample lakes).

3.1 Protocols for the Adirondack Lake case study

In the spatial analysis of Adirondack lakes, several environmental variables were considered as candidates for the pattern model. The variables that we considered as potential influences on mean lake ANC were: (1) lake elevation (m), (2) pH of precipitation, (3) precipitation (cm per year), and (4) watershed slope (%). Interactions among these four variables were also considered. Local values of these variables are available in digitized form for each of the lakes in the lake population of interest, and each variable is expected, based on previous studies (e.g., Hunsaker *et al.* 1987), to have an influence on lake ANC. The spatial distribution of the RV lake ANC and two candidate pattern variables, precipitation pH and lake elevation, are shown in Figures 3 through 5. Three types of information were included in the selection of pattern variables: (1) the predictive capability of the candidate variables was assessed using stepwise regression analysis, (2) pattern effects of the candidates were compared with those exhibited by lake ANC based on directional semivariogram analysis, and (3) scaled-residual semivariograms were assessed after each step of the stepwise regression analysis for removal of pattern effects.

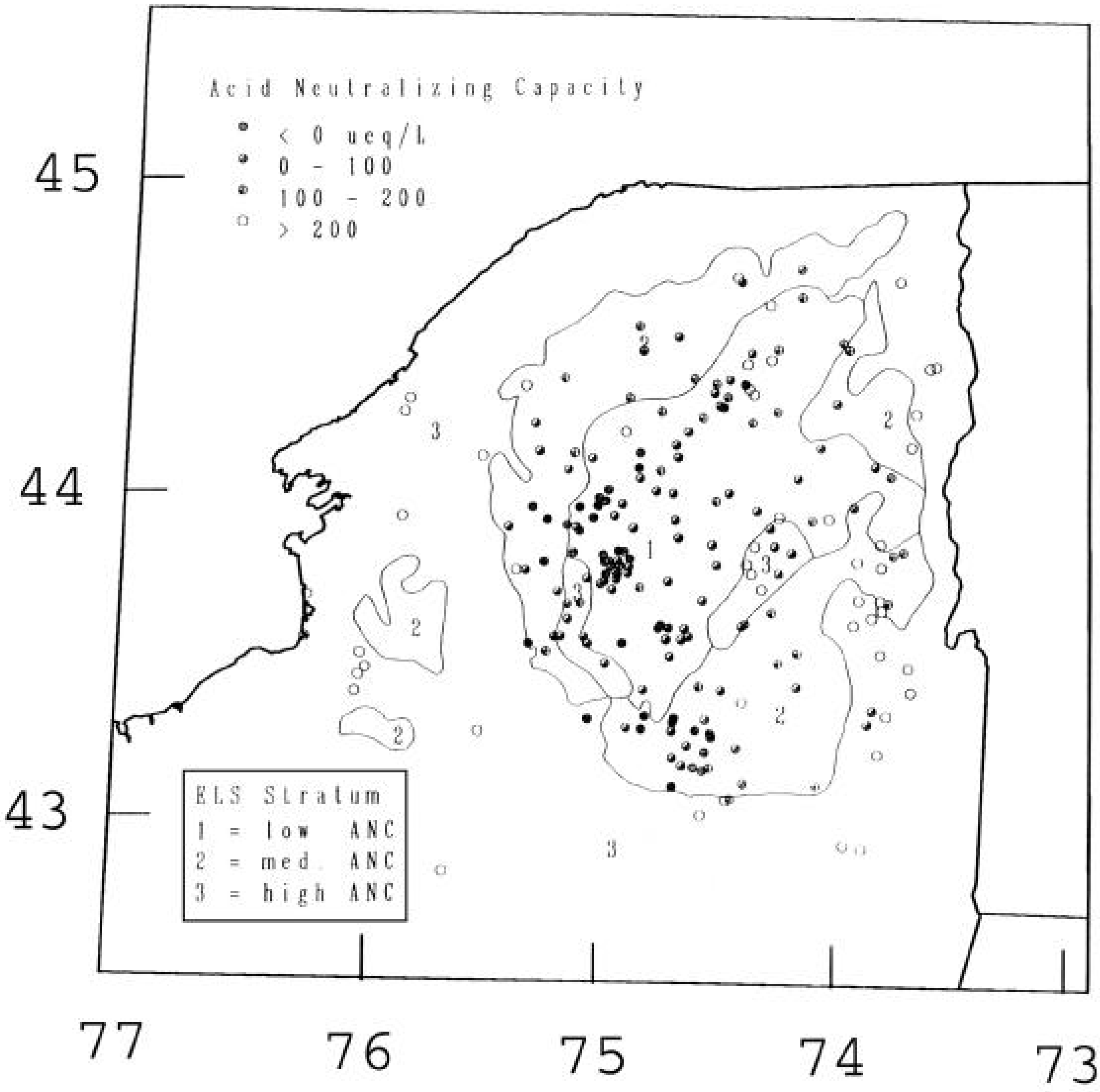

Figure 3 Map of upstate New York showing the acid neutralizing capacity (ANC) of sampled lakes. The outlined regions represent areas of low, medium, and high *a priori* ANC based on geology.

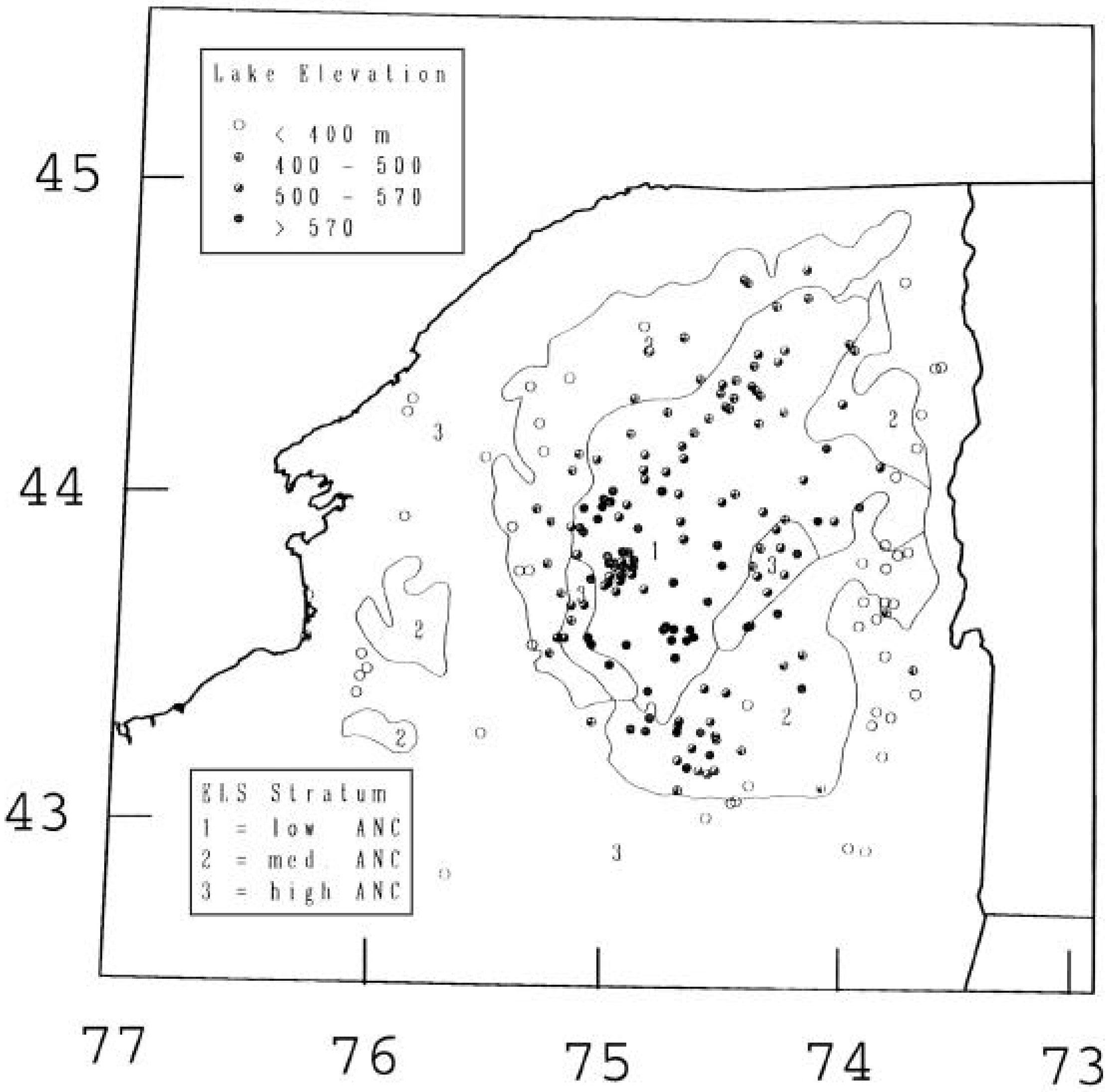

Figure 4 Map of upstate New York showing the elevation of sampled lakes. The outlined regions represent areas of low, medium, and high *a priori* ANC based on geology.

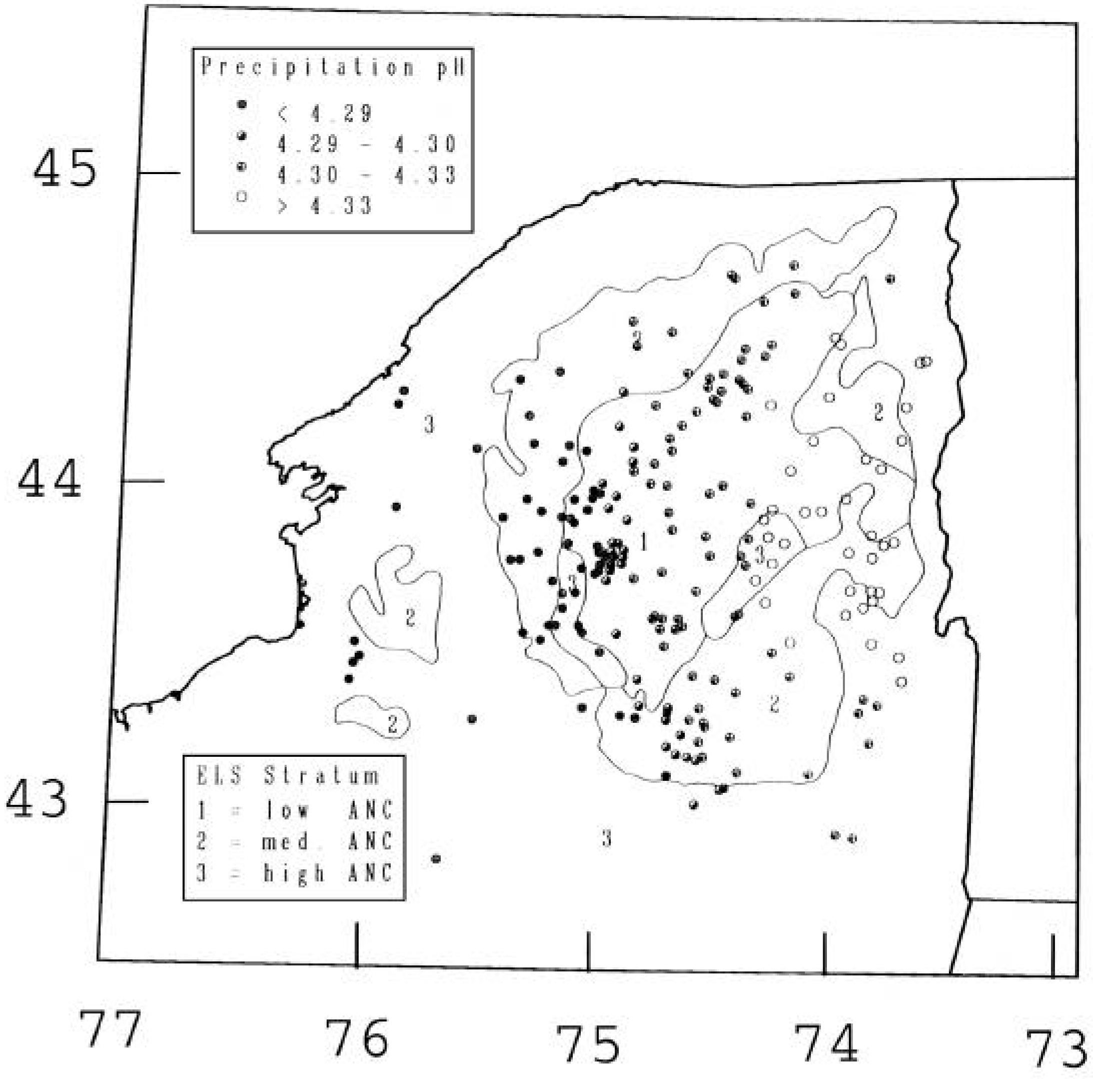

Figure 5 Map of upstate New York showing the pH of precipitation input to sampled lakes. The outlined regions represent areas of low, medium, and high *a priori* ANC based on geology.

As a part of the stepwise ordinary least squares (OLS) analysis, three subpopulations — the three ANC subregions described by Omernick and Powers (1983) and Landers *et al.* (1988) — were treated as qualitative indicator variables (Figures 3 through 5). These three categories were selected based on geology and other information that was available prior to sampling of the lakes. The general pattern is one of concentric rings: the inner Adirondack region representing the low-ANC subregion, surrounded by medium-ANC subregions, surrounded by a high-ANC subregion in the outer ring. This example problem involves both categorical and continuous influences on the spatial pattern of mean ANC.

From the total of 155 sampled lakes, 113 lakes were included in the OLS analysis. We removed two outliers from the regression analysis based on extreme values of ANC. The remaining 153 lakes were divided into four random subsets of approximately the same size. Three of these (subsets 1-3) were used in estimating the parameters of the pattern model, and three (2-4) were used in the estimation of semivariogram parameters. Ideally, we could have used completely different subsets in each step of the estimation procedure, but given the limitations of sample size, we decided to allow one-half of the sample lakes to be included in both steps. We strongly recommend that separate subsets of data be used to develop the two component models, and the PATTERN+ program allows the user to specify a dummy variable that splits the data. This provision ensures that the estimates from the two components are independent and do not induce bias in one-another.

Our assumption in defining a pattern function is that it produces residuals that share a common underlying distribution or, at least, come from distributions sharing a common mean and variance. In some cases, removing the mean from each subpopulation leaves residuals that differ widely in their variability, one subpopulation showing little variability and another showing a wider range of values. In this case, it may be possible to fit a "pseudostationary" semivariogram model to encompass both groups by allowing each group to have a different sill (variance) and fitting a semivariogram to data after standardizing (Journel and Huijbregts 1979). In this case study, we used the transformation $Z = \log_{10}(\text{ANC} + 150)$, suggested by Hunsaker *et al.* (1987), to equalize variances. An alternative approach that we recommend uses untransformed ANC and scales the residuals to avoid heteroscedascicity (Jager and Overton 1993).

3.2 Detecting the presence of a spatial pattern

Although there is no true test for deciding whether the underlying process that gave rise to the landscape had a significant spatial pattern (Myers 1989), several indications have been offered for detecting spatial pattern (referred to as "drift" in geostatistics literature; see Starks and Fang 1982). All of these rely on information provided by the empirical semivariogram, which is based on a single realization of the RV. One indication that there is a spatial gradient in the mean is given when the empirical semivariogram does not level off to a constant sill value but continues to climb above the variance (estimated by the sample variance). Geostatisticians are quick to point out that there is no requirement for semivariogram models to have a sill. While this is true in theory, we are not interested in providing the model with the greatest possible amount of flexibility. Instead we have a subjective problem of modeling a real spatial process in nature with more degrees of freedom than we know what to do with. Why not welcome the constraint that the residual semivariances reach an independently estimated sill? We propose that the apparent absence of a finite sill, or even a sill much higher than the sample variance, provides us with a concrete guideline or constraint for model development. The absence of a finite sill implies large negative correlations between locations that are far apart. VerHoef (1993) points out that negative spatial dependence appears to be rare in nature. Common sense suggests that small negative correlations might occur at random, but it defies logic to expect the strength of the correlation to increase *ad infinitum*.

A second clue pointing toward expainable spatial pattern is the presence of a parabolic shape in the empirical semivariogram. We know that a monotonic linear or higher order spatial gradient causes a parabolic shape. We assume here that the converse holds and use a parabolic shape to indicate the presence of a gradient. If there is a monotonic pattern in the mean, then the pattern must have a direction. This will cause the appearance of anisotropy among empirical semivariograms constructed separately for pairs of points oriented in different directions. Those pairs of points oriented in the direction of fastest change will diverge in the values of the RV, causing the experimental semivariogram to rise above the theoretical sill. At the same time, pairs perpendicular to this direction will have values that are similar to each other, causing the experimental semivariogram to remain low. These features are mainly related to spatial patterns that are gradient-like (i.e., monotonic in space over relatively large distances in the region of interest).

Empirical semivariograms were produced for the Adirondack lakes. Comparison of the directional semivariograms (separate analysis for pairs of lakes oriented in different directions) clearly suggests the presence of a linear gradient. We hypothesize that spatial patterns in lake ANC cause (1) a parabolic shape of the semivariogram and (2) anisotropy, with larger semivariances in the direction of the pattern.

3.3 Iterative residual refinement (Phase II)

We initiated the iterative residual procedure using the final pattern model provided by the screening process. Our first step was to fit a semivariogram model to the residuals of the final pattern model. The general form of the semivariogram is given in Equation (2). Function d(h) varies from one to zero as distance (h) runs from 0 to the range. The shape of the function d(h) depends on the particular model form chosen.

$$g(h) = n + (s - n)(1 - d(h)) \quad (2)$$

We compared three commonly used semivariogram models: the exponential, spherical, and circular model. Three semivariogram parameters— the nugget (*n*), the sill (*s*), and the range (*r*)— were estimated for each model. Iteratively reweighted nonlinear least squares was used to estimate parameters *n*, *s*, and *r*. The algorithm used for nonlinear least squares is a modification of the Levenberg-Marquardt algorithm. The weights used in this algorithm (Cressie 1985) are related to the variance of the semivariogram estimator for each distance class. The weights emphasize shorter distance classes and those supported by more pairs of points. The function in Equation (3) was minimized for all distance classes by nonlinear least squares.

$$\frac{N(h)^{\frac{1}{2}}}{g(h;n,s,r)} \left(g_{ex}(h) - g(h;n,s,r)\right)^{2} \quad (3)$$

Here, $g_{ex}(h)$ is the empirical semivariance calculated from sample pairs in distance class *h* and *g(h;n,s,r)* is the semivariance predicted by the parametric model for distance class *h*. The model, *d(h)* in Equation (2), resulting in the smallest least-squares error between the fitted model and empirical semivariogram values was chosen. *N(h)* is the number of sample pairs in distance class *h*. The selected semivariogram model and its parameter values were then used to construct a VC matrix for the residuals.

In the second step, the pattern model was refitted using generalized least squares and the estimated VC matrix to account for spatial autocorrelation among the residuals. This produced new estimates of the regression coefficients in the pattern model. After several iterations, convergence was achieved for both the pattern model coefficients and the semivariogram parameters. Although we are minimizing the potential bias in these estimates through our use of an iterative procedure, we made an effort to reduce the possibility of bias even further by using different subsets of lakes in each step. We were unable to afford dividing the lakes into two completely different subsets because we needed to maintain adequate sample sizes for both. As a compromise, we divided the sample lakes into four random groups and used three in each step, leaving us with an overlap of one-half of the lakes used in both.

4. RESULTS

4.1 Screening candidate environmental variables (Phase I)

The results of our stepwise OLS procedure in Table I are given for the final model for all three subregions combined. Elevation (ELEV), the pH of precipitation (pH), the subregion indicator variable (S3) identifying lakes in the high-ANC subregion, and a pH*S3 interaction term were included in the final model. The latter term was needed to allow for a different slope for pH in the high-ANC subregion. The final OLS model explained 73% of the variation in $\log_{10}(\mathrm{ANC} + 150)$. The multivariate pattern model for low- and medium-ANC subregions is given by Equation (4). The model for high-ANC subregions is given by Equation (5).

$$\mathrm{log10(ANC + 150) = -8.59 - 0.0012\ ELEV + 2.67\ pH} \qquad (4)$$

$$\mathrm{log10(ANC + 150) = 6.17 - 0.0012\ ELEV - 0.72\ pH} \qquad (5)$$

Table 1. Ordinary least squares results for the final model obtained by a stepwise procedure: $E[Z] = \beta_0 + \beta_1$ Elev $+ \beta_2$ pH $+ \beta_3$ S3 $+ \beta_4$ pH*S3.

Source	Degrees of freedom	Sum of squares	Mean square error	Statistics
Model	4	6.9840	1.7460	F = 71.65
Error	108	2.6318	0.0241	P < 0.0001
Total	112	9.6158		$R^2 = 0.73$

Characteristics of the directional semivariograms for lake ANC (Figure 6) suggest that a kriging model that assumes stationary spatial autocorrelation will not perform as well as a model that incorporates the deterministic spatial pattern. Each point has as its y-axis value half of the average squared difference between pairs of lakes separated by the distance shown on the x-axis and oriented in the direction indicated by the legend. The semivariance of pairs of lakes oriented in the E to W and NE to SW (157.5°,0.0°,22.5°) directions has a parabolic shape near the origin and continues to rise beyond the theoretical sill. In contrast, the semivariogram constructed from lakes oriented in the N to S direction remains well-below the sill. These features give a fairly clear indication of an E to W gradient in mean lake ANC. The anisotropy caused by the gradient influences a wide angle (between about -45° and 22.5°).

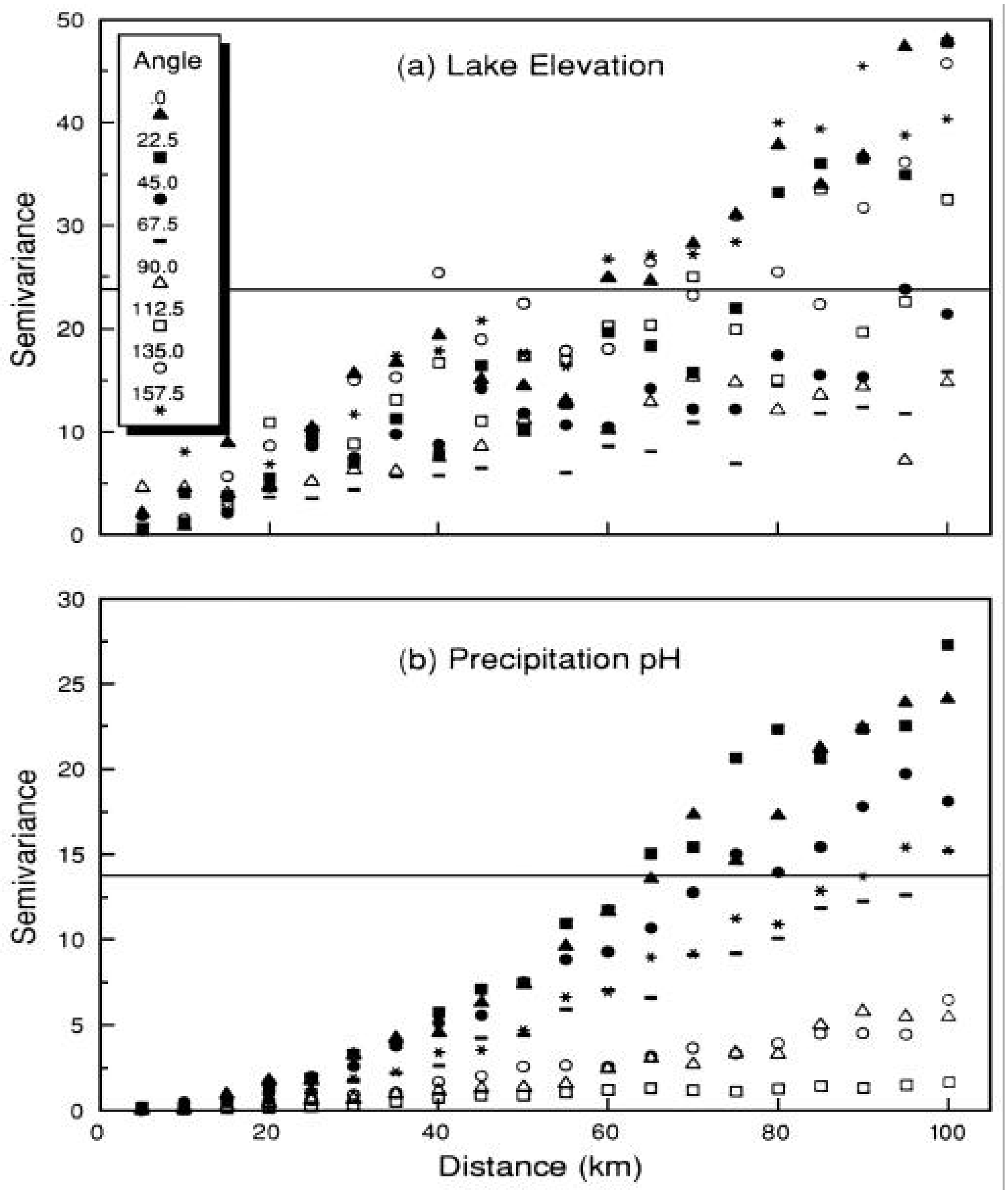

Figure 6 Directional experimental semivariograms for the regionalized variable of interest. $\log_{10}$(ANC + 150). Symbols identify the orientation between pairs of lakes (East = 0°). The horizontal line identifies the theoretical sill (overall variance).

Compare the directional semivariograms of lake ANC (Figure 6) with the directional semivariograms of candidate predictor variables lake elevation (Figure 7a) and precipitation pH (Figure 7b). Among the candidate pattern variables, precipitation pH shows the most clear-cut pattern effects. Note that for the gradient direction, the semivariogram has a parabolic shape and continues to rise above the theoretical sill, whereas there is almost no variation in the 112.5° direction. There is a clear ordering of directions, with a peak in the sill index at 22.5° and a trough at 112.5°. The amount of precipitation showed similar patterns to the pH of precipitation. Elevation shows gradient features in the E-W direction (0°), but without a parabolic shape.

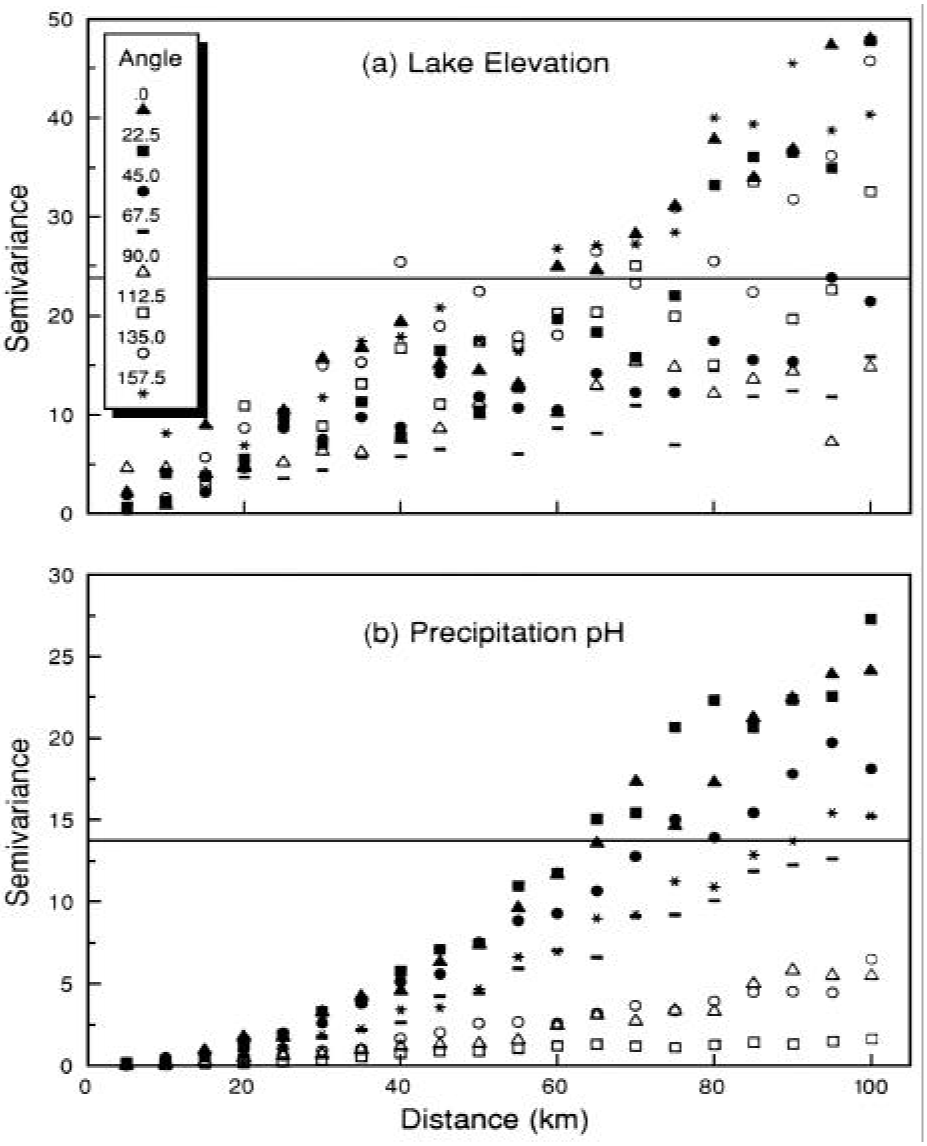

Figure 7 Directional experimental semivariograms for two candidate variables used to predict spatial pattern in lake ANC: (a) lake elevation and (b) precipitation pH. Symbols identify the orientation between pairs of lakes (East = 0°). The horizontal line identifies the theoretical sill (overall variance).

The results of evaluating the directional semivariograms are summarized in a diagram of a sill index (defined as the maximum value attained by the semivariogram for each direction) plotted against direction (Figure 8). In this case, a sinusoidal curve emerges with a 90° period, where the gradient direction (the direction of fastest change in ANC) is indicated by the peak of the sill index.

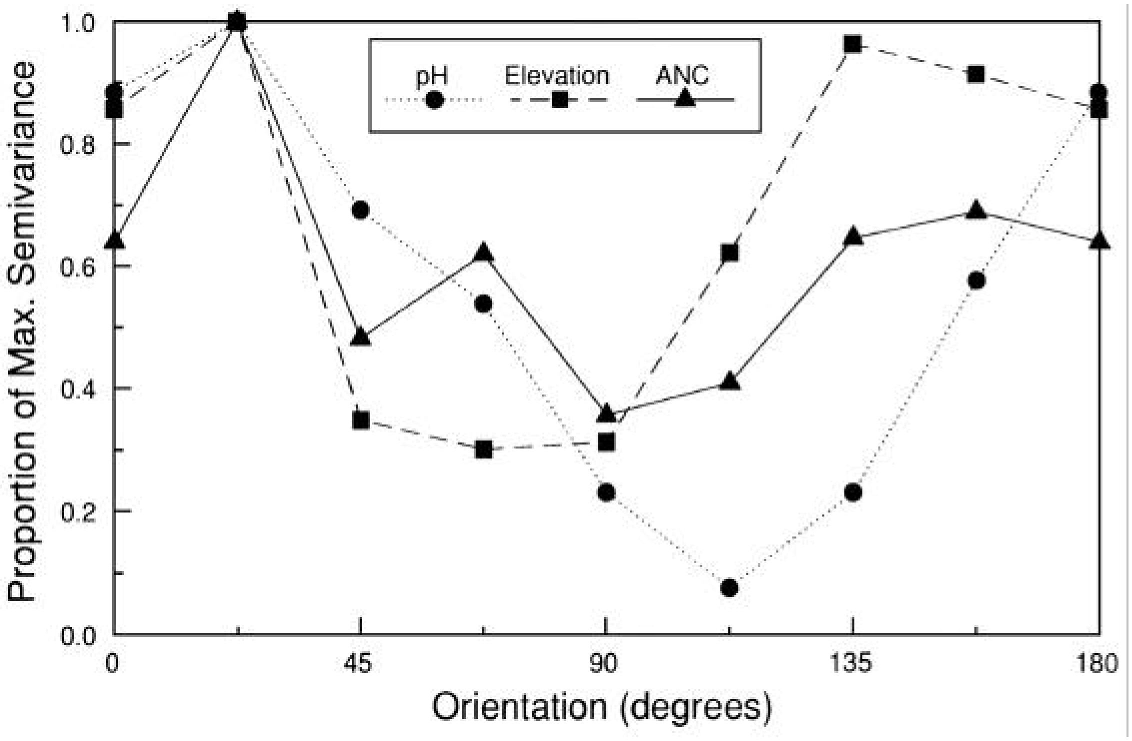

Figure 8 Sinusoidal pattern of anisotropy shown by the empirical sill (scaled maximum semivariance) as a function of the orientation between lakes. Curves are shown for lake $\log_{10}$(ANC + 150) and two pattern variables: precipitation pH and lake elevation.

In this case study, fitting a pattern model removed most of the spatial autocorrelation structure in the data. This can be seen by comparing the empirical semivariogram for the residuals from the final GLS residuals (after iterative residual analysis; Figure 9a) with the original semivariogram for lake ANC (prior to the removal of explainable pattern in the mean; Figure 9b).

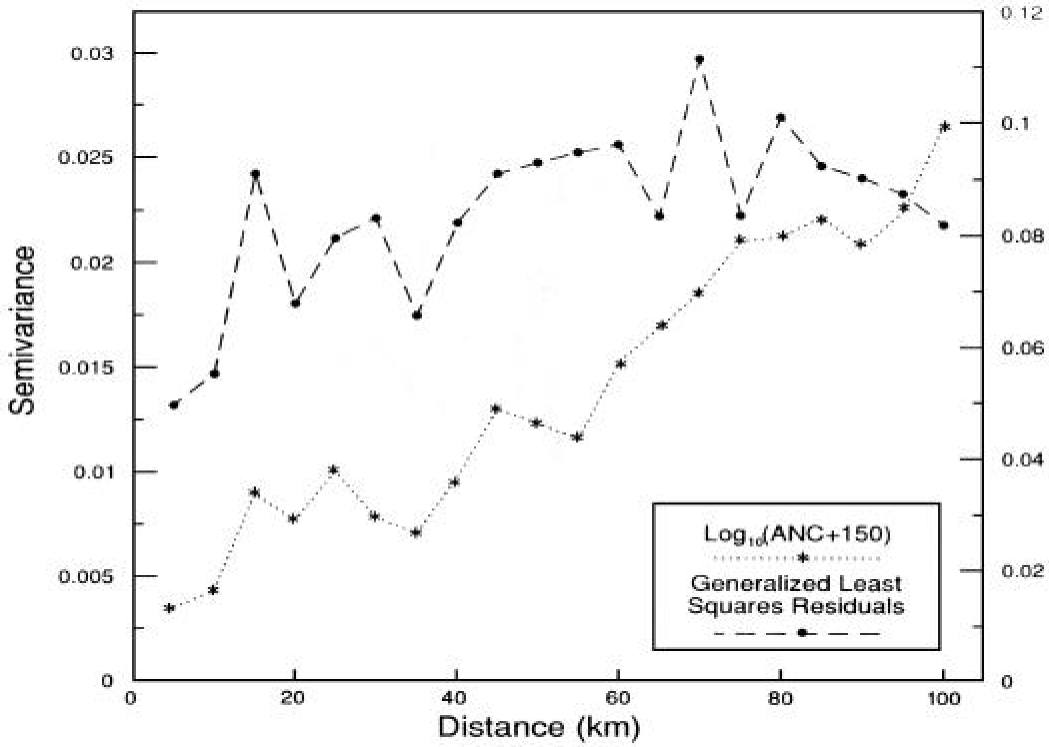

Figure 9 Omnidirectional experimental semivariogram for log$_{10}$(ANC + 150) (right axis) and for residuals from the pattern model developed by ordinary and generalized least squares (left axis) following the iterative residual kriging procedure.

4.2 Iterative residual refinement (Phase II)

The iterative procedure converged to give stable values for both the pattern model coefficients and the semivariogram parameters. The semivariogram model with the smallest total (weighted) deviation from the empirical semivariogram of the residuals follows the exponential model with nugget = 0.0116, sill = 0.0248, and range = 19.63 km. The form of this model is given by Equation (6). The fit of this model to the empirical semivariogram of the GLS residuals is shown in Figure 10.

$$g(h;n,s,r) = n + (s - n)\left(1 - \exp\left(\frac{-h}{r}\right)\right), \quad 0 < h < r \qquad (6)$$

The refined coefficients obtained for the pattern model by GLS are similar to those obtained initially by OLS. The mean square error is higher (0.0298), and as a result, the r^2 (0.67) is somewhat reduced because an estimate of spatial autocorrelation among lakes is incorporated (Table 2). This is expected because the OLS model assumes that lakes are not spatially correlated, thus underestimating the true residual variance. The influence of rainfall pH is slightly increased for all subregions, with compensations in the intercept.

Table 2. Generalized least squares results for the final model obtained by the iterative residual kriging procedure.

Source	Degrees of freedom	Sum of squares	Mean square error	Statistics
Model	4	6.3950	1.5987	F = 53.61
Error	108	3.2208	0.0298	P < 0.0001
Total	112	9.6158		$R^2 = 0.67$

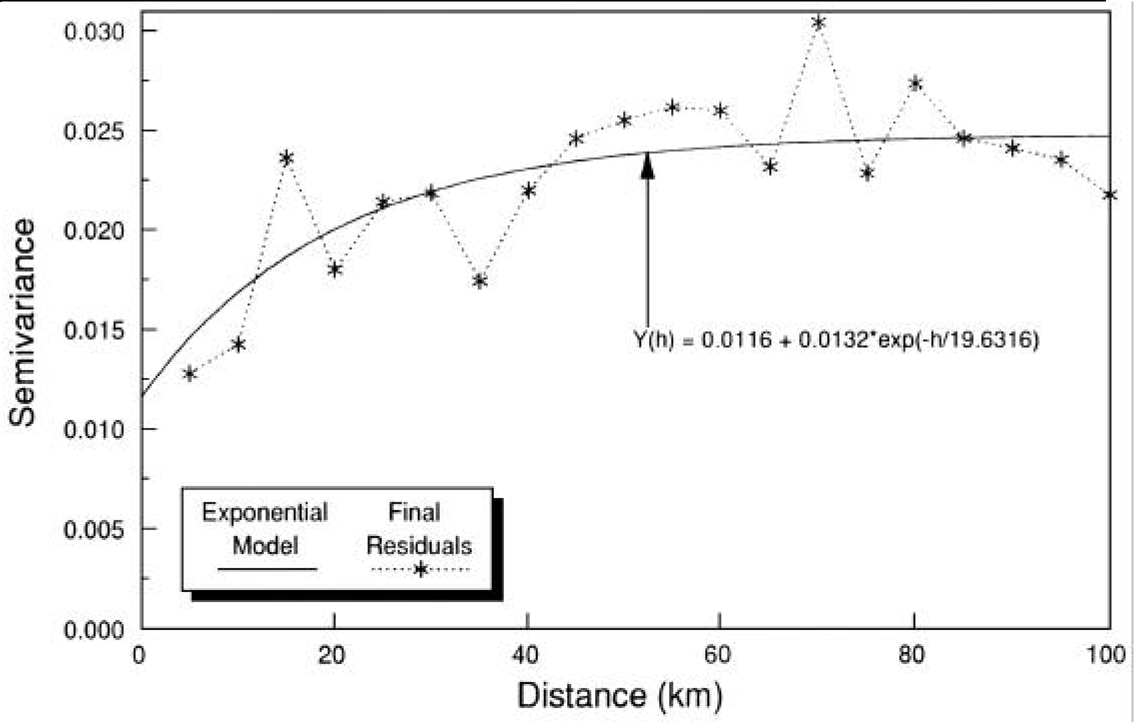

Figure 10 Comparison of the final exponential model fit to the final generalized least-squares residuals.

The GLS pattern model for lakes belonging to the low- and medium-ANC subregions and for lakes in the high-ANC subregion are given by Equation (7) and

Equation (8), respectively. The values predicted for each lake represent our estimates of the mean (M^*) at each lake location and collectively describe the large-scale spatial pattern that we were able to explain using these environmental predictors.

$$\log10(\mathrm{ANC} + 150) = -9.08 - 0.0012\ \mathrm{ELEV} + 2.78\ \mathrm{pH} \qquad (7)$$

$$\log10(\mathrm{ANC} + 150) = 6.77 - 0.0012\ \mathrm{ELEV} - 0.85\ \mathrm{pH} \qquad (8)$$

4.3 Estimates of total lake ANC

Our final estimates of lake ANC were made for the population of lakes that appear on 1:250,000-scale maps of upstate New York (Figure 11). This total ANC estimate (Z^*) is the sum of the pattern model estimate of mean ANC (M^*) and the estimate for the residual (R^*) obtained by kriging interpolation. The transformation that we used to convert back from log units to ANC units (μeq/L) predicts the median for lake ANC at each site.

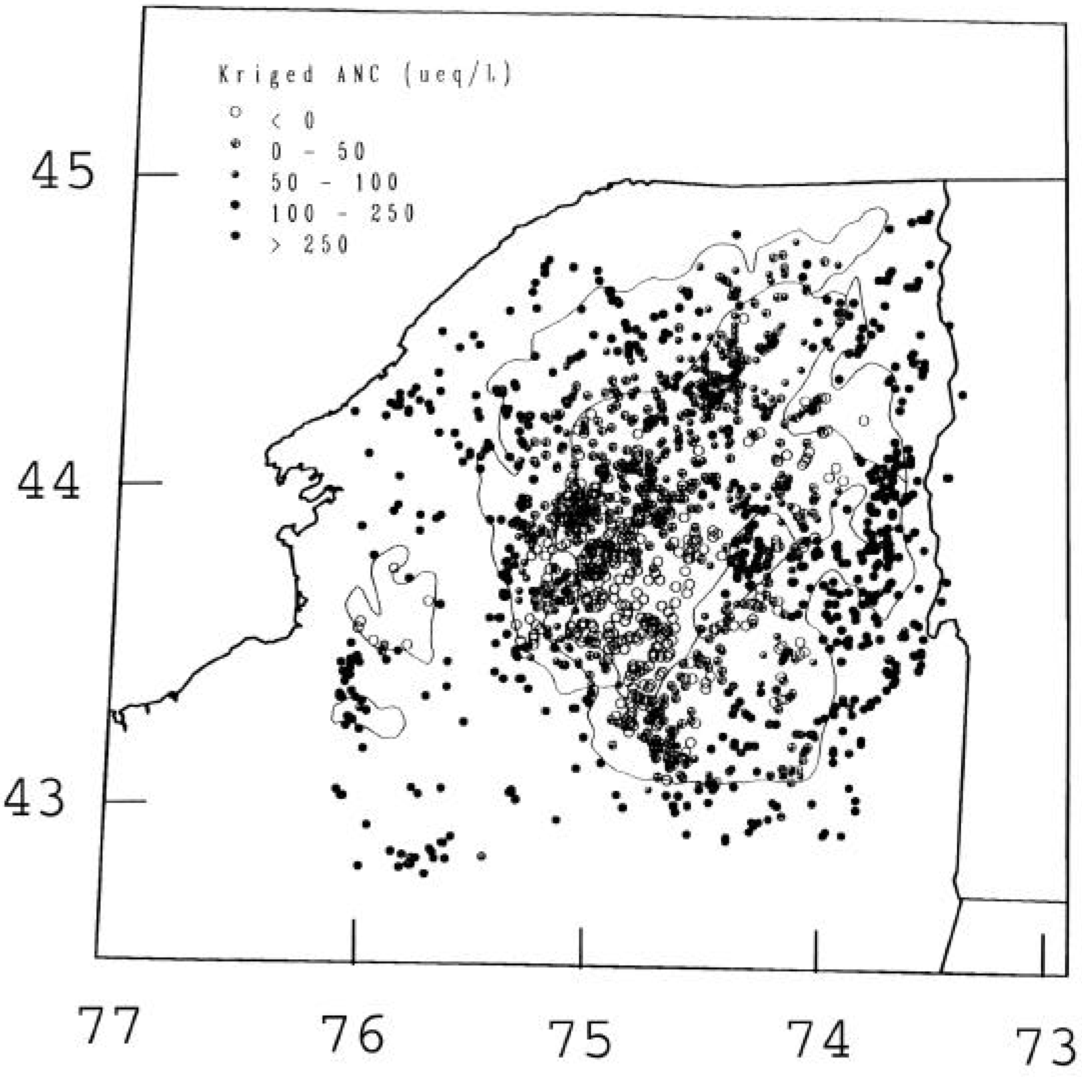

Figure 11 Map of total ANC estimates for the 1:250,000-scale population of lakes in upstate New York State.

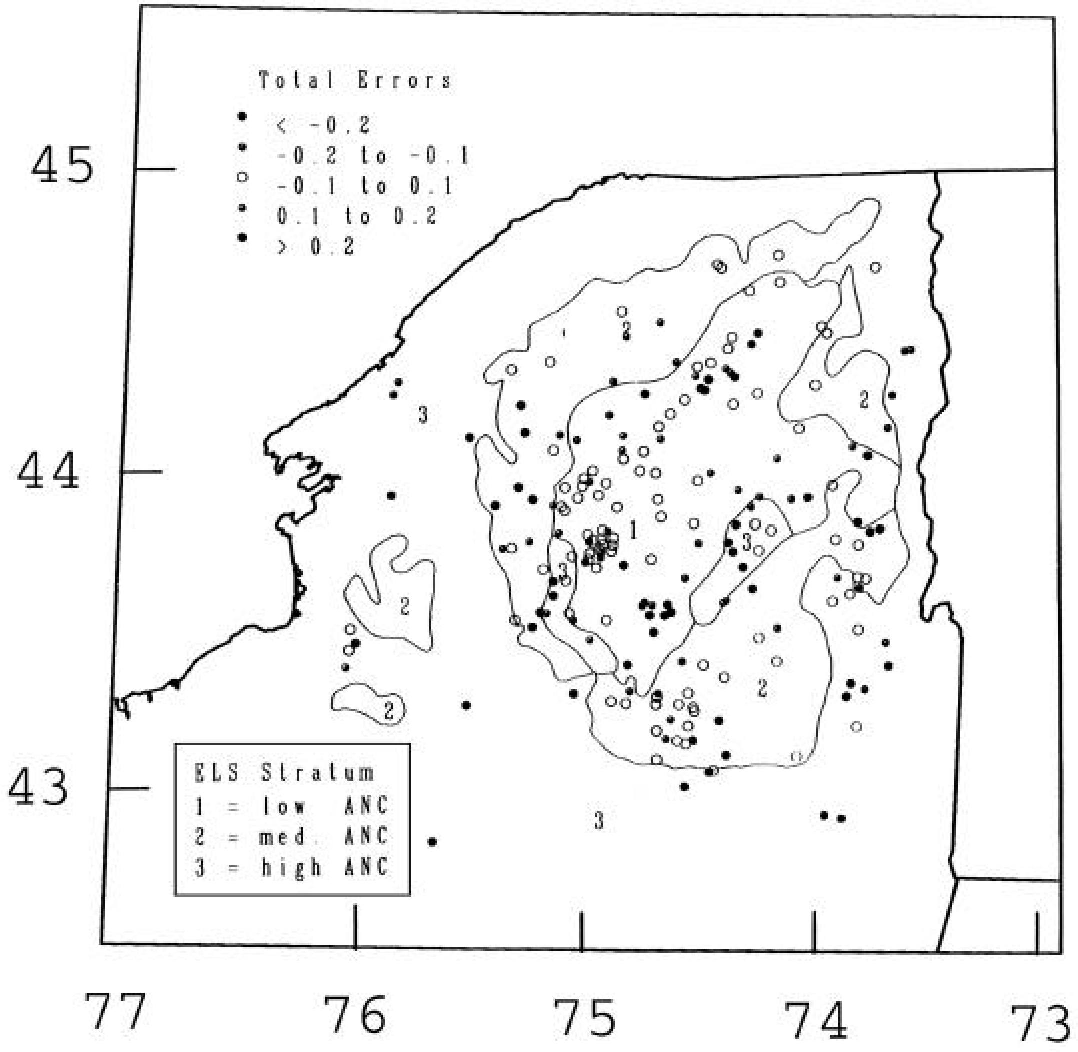

Figure 12 Map of total errors for ANC estimates at sampled lakes.

Total errors *(Z^* - Z)* were calculated for the sample lakes from Equation (1), based on estimated $\log_{10}$(ANC + 150), ($Z^* = M^* + R^*$), and measured $\log_{10}$(ANC + 150), *(Z = M + R)*, where *M* represents the true mean and *R* represents true residual error, both unknown. The total error shown in Figure 12 for sampled lakes includes both the mean prediction error *(M - M^*)* and the kriging error *(R^* - R)* estimated by cross-validation. These errors have units of $\log_{10}$(µeq/L). The expression for the total estimation variance is given by Equation (9). The first term describes the variance in predicting the mean, the second term describes the kriging variance, and the third term represents the covariance between the two parts of the estimator. The variance of predicting the mean is greater for more extreme values of the predictor variables but also depends on how correlated those variables are. The kriging variance depends on the data configuration (the proximity of sample locations to the lake of interest) and is bounded above by the mean square error. The highest errors would therefore be expected in predicting lakes that occur at extreme elevations at one end or the other of the E to NE gradient in precipitation pH and far from any sampled lakes. Note that the total variance can be reduced, through the covariance, by sampling densely at the extremes of the predictor variables.

$$E\left[\left(Z^{*}-Z\right)^{2}\right]=E\left[\left(M^{*}-M\right)^{2}\right]+E\left[\left(R^{*}-R\right)^{2}\right]+2E\left[\left(M^{*}-M\right)\left(R^{*}-R\right)\right] \quad (9)$$

5. Discussion

5.1 Screening candidate environmental variables (Phase I)

Elevation was by far the most important predictor of lake ANC of those available to us. The negative relationship between elevation and water chemistry is discussed elsewhere (e.g., Hunsaker *et al.* 1987; Whitehead *et al.* 1989; Jager *et al.* 1990). The relationship between lake $\log_{10}$(ANC + 150) and elevation appears to be remarkably constant across the three ANC subregions, suggesting that the relationship is linear over a wide range of elevations. The stepwise procedure selected identical models for the low- and medium-ANC subregions, indicating that the relationships between lake ANC and both elevation and the pH of precipitation are similar, even if the values fall in different ranges of the scale. For the high-ANC subregion, there was a different intercept and the relationship with pH of precipitation became negative, although the slope was not significantly different from zero. The pattern model suggests that for lakes in the low- and medium-ANC subregions, a lower pH of rainfall will lower lake ANC, as expected in a titration model. However, the positive association between the pH of rainfall and lake ANC apparently decreases for highly buffered watersheds represented in the high ANC subregion. This is consistent with the theory that these watersheds can still respond to acidic inputs by exporting base cations.

The two candidate variables offer an interesting contrast in pattern. The spatial characteristics of elevation and rainfall pH are quite different. Elevation is much more fine-grained than rainfall pH (compare Figures 4 and 5). The pH of precipitation was obtained from large-scale maps (Wampler and Olsen 1987; Olsen and Slavich 1986) and has the appearance of a plane tilted downward in the E to NE direction. Figure 7 shows that both variables exhibit anisotropy and other pattern influences in the directional semivariograms. The pattern of precipitation pH is consistent with that expected in the presence of a strong spatial gradient over the region. This is not surprising because (1) such a gradient is known to exist, radiating northeast from the industrial regions in the midwest and (2) small-scale fluctuations have been lost because the data originated from meteorological monitoring data collected and interpolated over a large region. Although there is patchiness in elevation over the region in upstate New York considered here (e.g., the high-peaks region of the Adirondacks), the pattern of anisotropy probably reflects the local orientation of mountain ranges in the N to S direction, causing smaller average elevational differences between lakes oriented in that direction (see Figure 4).

Comparing the candidate predictors with the variogram for ANC in Figure 6, there are similarities with both elevation and precipitation pH, all of which show directional pattern effects. The ANC semivariogram shows a murkier gradient pattern than precipitation pH, similar to the weak pattern in elevation. The gradient direction could be anywhere between -45 and 22.5°, indicating a broader, less-unidirectional gradient effect. In this respect, lake ANC is intermediate between the sharp peak in precipitation pH and the broad plateau in elevation (Figure 8).

The fact that the residual semivariograms show a decrease in the sill as more variation is explained by the deterministic pattern model. There is also a transition from semivariograms with apparent pattern effects to semivariograms that look more convex and reach a distinct sill (Figure 9). This suggests that most of the large-scale gradient effects in spatial pattern have been accounted for by the model, leaving only small local correlations among lake residuals to be estimated by kriging interpolation.

5.2 Iterative residual refinement (Phase II)

The PATTERN+ model developed here demonstrates the ability to borrow from the field of geostatistics in the development of a predictive model for spatially extensive environmental fields. In this example, the environmental field of interest is lake ANC. The pattern model alone is of some interest. Using this procedure avoids the assumption that no spatial autocorrelation exists among lakes and leads to different estimates of the regression parameters. By itself, the pattern model can give very good predictive capability if the mean square error is small enough. In other cases, strong predictive pattern variables may not be available to explain spatial changes in the mean. If the residuals from such a weak pattern model are large and show considerable spatial autocorrelation, the local kriging estimates can represent an important contribution to the PATTERN+ predictions.

5.3 Future developments

This case study suggested three improvements to the procedure that we used. First, it is preferable to use predictor variables in the pattern model that are measured with greater precision than the environmental variable of interest -- in our case the deposition data were based on data from a fairly coarse monitoring grid. VerHoef et al. (1993) describe an interesting alternative to the semivariogram that is invariant to aggregation that might prove relevant. Second, it is advisable to put effort into separate estimation of variances and scaling of subpopulations that differ greatly in variance and that can be identified a priori. When transformation of the variables is the alternative, back-transformation creates more statistical problems than it solves. Finally, simulation studies are needed to evaluation the ideal partitioning of data for use in parameter estimation for both components of the PATTERN+ model. In particular, we would like to know how much data overlap can be tolerated without causing undue bias in the estimated parameters.

REFERENCES

Ahmed, S. and DeMarsily, G.: 1987, 'Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity', *Water Resour. Res.* **23(9)**, 1717-1737.

Cressie, N.: 1985, 'Fitting variogram models by weighted least squares', *Math. Geol.* **17(5)**, 563-586.

Delhomme, J. P.: 1978, 'Kriging in the hydrosciences', *Adv. Water Resour.* **1(5)**, 251-266.

Fedorov, V. V.: 1989, 'Kriging and other estimators of spatial field characteristics (with special reference to environmental studies)', *Atmos. Environ.* **23(1)**, 175-184.

Hunsaker, C. T., Olson, R. J., and Carpenter, D. E.: 1987, 'Adirondack lake system acidity: differences between headwater and nonheadwater lakes', *The Norwegian National Committee for Hydrology, Vol I.: Acidification and Water Pathways*, pp. 291-201.

Jager, H. I., Sale, M. J., and Schmoyer, R. L.: 1990, 'Cokriging to assess regional stream quality in the Southern Blue Ridge Province', *Water Resour. Res.* **26(7)**, 1401-1412.

Jager, H.I., and W.S. Overton. 1993. 'Explanatory models for ecological response surfaces', *In Environmental Modeling with GIS*, pp. 422--431. Edited by M.F. Goodchild, B.O. Parks, and L.T. Steyaert, Oxford University Press, New York.

Jernigan, R. W.: 1986, 'A primer on kriging', *Statistics 1986*, U.S. Environmental Protection Agency, Washington, D.C., 83 pp.

Journel, A. G., and Huijbregts, Ch. J.: 1978, *Mining Geostatistics*. Academic Press, Inc., Orlando, 600 pp.

Landers, D. H., Overton, W. S., Linthurst, R. A. , and Brakke, D. E.: 1988, 'Eastern Lake Survey, regional estimates of lake chemistry', *Environ. Sci. Technol.* **22(2)**, 128-172.

LaJaunie, C.: 1973, 'A geostatistical approach to air pollution modelling',in Verly, G. *et al.* (eds.), *Geostatistics for Natural Resources Characterization, Part II*, D. Reidel Publishing Co., Dordrecht, Holland, pp. 349-364.

Legendre, P. and Troussellier, M.: 1988, 'Aquatic heterotrophic bacteria: Modeling in the presence of spatial autocorrelation',*Limnol. Oceanogr.* **33(5)**, 1055-1067.

Legendre, P., Troussellier, M., Jarry, V., and Fortin, M. J.: 1989, 'Design for simultaneous sampling of ecological variables: from concepts to numerical solutions', *Oikos* **55**, 30-42.

Linthurst, R. A., Landers, D. H., Eilers, J. M., Brakke, D. F., Overton, W. S., Meier, E. P., and Crowe, R. E.: 1986, 'Characteristics of lakes in the Eastern United States Volume I: population descriptions and physico-chemical relationships', *EPA600/4-86/007a*, U.S. Environmental Protection Agency, Las Vegas.

Meyers, J. C. and Bryan, R. C.: 1984, 'Geostatistics applied to toxic waste . . . a case study',in Verly, G. *et al.* (eds.), *Geostatistics for Natural Resources Characterization, Part II,* D. Reidel Publishing Co., Dordrecht, Holland. pp. 893-901.

Myers, D. E.: 1989,'To be or not to be . . . stationary? That is the question', *Math. Geol.* **21(3)**, 347-361.

Neuman, S. P. and Jacobson, E. A.: 1984, 'Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels', *Math. Geol.* **16(5)**, 499-521.

Olsen, A. R. and Slavich, A. L.: 1986, 'Acid precipitation in North America: 1984 annual data summary from Acid Deposition System Data Base', *EPA 600/4-86-033,*U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Omernick, J. M., and Powers, C. F.: 1983, 'Total ANC of surface waters-a national map', *Ann. Assoc. Amer. Geogr.* **73(1)**, 133-136.

Robertson, G. P.: 1987, 'Geostatistics in ecology: interpolating with known variance', *Ecology* **68(3)**, 744-748.

Simpson, J. C.: 1984, 'Estimation of spatial patterns and inventories of environmental contaminants using kriging',in *Symposium on environmental applications of chemometrics*, August 26-31, Philadelphia. PNL-SA-12160, Pacific Northwest Laboratory, Richland, Washington.

Starks, T. H, and Fang, J. H.: 1982, 'The effect of drift on the experimental semivariogram', *Math. Geol.* **14(4)**, 309-319.

Stein, M. L.: 1984, 'Estimation of spatial variability Part I: nonparametric variogram estimation', *Technical Report No. 73*, Department of Statistics, Stanford University, Stanford. 183 pp.

VerHoef, J. M.. 1993. Universal kriging for ecological data. *In Environmental Modeling with GIS*, pp. 447--453. Edited by M.F. Goodchild, B.O. Parks, and L.T. Steyaert, Oxford University Press, New York.

VerHoef, J. M., N. A. C. Cressie, and D. C. Glenn-Lewin. 1993. 'Spatial models for spatial statistics: some unification'. *Journal of Vegetation Science* 4: 441-452.

Wampler, S. J. and Olsen, A. R.: 1987, 'Spatial estimation of annual wet acid deposition using supplemental precipitation data', in *Tenth Conference on Probability and Statistics,* October 4-5, 1987, Edmonton, Alberta, Canada. American Meteorological Society, Boston.

Whitehead, D. R., Charles, D. F., Jackson, S. T., Smol, J. P., and Engstrom, D. R.: 1989, 'The developmental history of Adirondack (N.Y.) lakes', *J. of Paleolimnology* **2**, 185-206.

This report has been reproduced from the best available copy.

Reports are available to the public from the following source.

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone 703-605-6000 (1-800-553-6847)
TDD 703-487-4639
Fax 703-605-6900
E-mail orders@ntis.fedworld.gov
Web site http://www.ntis.gov/ordering.htm

Reports are available to U.S. Department of Energy (DOE) employees, DOE contractors, Energy Technology Data Exchange (ETDE) representatives, and International Nuclear Information System (INIS) representatives from the following source.

Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831
Telephone 423-576-8401
Fax 423-576-5728
E-mail reports@adonis.osti.gov
Web site http://www.osti.gov/products/sources.html

Reports produced after January 1, 1996, are generally available via the DOE Information Bridge.
Web site http://www.doe.gov/bridge

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof.