CONF-9109345--1

DE92 007840

EXPLANATORY MODELS for ECOLOGICAL RESPONSE SURFACES

HENRIETTE I. JAGER
Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6036

W. SCOTT OVERTON
Department of Statistics, Oregon State University, Corvallis, OR 97331

ABSTRACT

Patterns of ecological variables over landscapes are appropriately considered to be response surfaces, functionally determined by spatially patterned landscape or environmental variables, and by atmospheric influences and inputs. In this perspective, spatial models may provide a good description for such patterns, but must be considered surrogates for causal models of response. Enhanced understanding follows modeling the response by a mechanistic explanatory model, replacing spatial variables by topographic, landscape, geologic, or other environmental variables. Such an analysis is provided by multiple regression of the response on the chosen suite of explanatory variables.

It is usual in regression analysis to analyze the residuals and to modify the analysis based on the properties of the residuals. With spatially distributed data, residuals must be analyzed in the spatial context, with attention to spatial autocorrelation and non-stationarity. The geostatistical techniques of semivariogram analysis and kriging, when applied to the residuals, provide evidence of response pattern not accounted for by the regression and of spatial structure in the errors. This evidence can be used to improve the regression analysis and to guide the search for additional explanatory variables.

The approach is illustrated by an example from the National Lake Survey. This example involves features of the sampling design in the analysis, and takes into account the identity and location of the finite population of lakes sampled. Implications for national monitoring programs are identified.

Conference, Integrating Geographic Information Systems and Environmental Modelling, Boulder, Co., September 15-19, 1991. This research was supported by Cooperative agreement CR816721 between the U.S. Environmental Protection Agency and Oregon State University and under Interagency Agreement DW89934074-2 with the U.S. Department of Energy under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.. This is Environmental Sciences Division publication number ____.

INTRODUCTION

It is often the spatial patterns in environmental and ecological variables that arouse interest and demand explanation. For environmental response variables, the causal influences of interacting environmental factors produce the patterns of interest. Ecological response variables by definition involve living organisms and are at least one step removed from spatial patterns in the physical environmental. The spatial organization of ecological variables, such as species abundances, is often viewed as a collection of individual species responses to variation in the physical environment (Gleason, 1926), although competition and other ecosystem interactions may also influence spatial arrangement.

The response of ecological variables to spatial environmental gradients can be direct, or it can be biologically integrated through interactive responses of the ecosystem. Physical factors such as soil and geologic and physiographic structure provide the base physical environment and substrate. Climatic factors such as precipitation, temperature, constituents of atmospheric deposition, and solar radiation represent important exogenous driving variables. Complex ecosystems provide yet another dimension of the environment of particular ecological processes, and also interact with the physical environment, modifying it and ameliorating its effects. The influence of these factors is seldom simple and predictive models must be tailored to the ecological problem at hand. Quite often, simple explanatory variables suffice for complex relations. For example, the environmental cue that notifies temperate populations of spring's arrival is widely modeled as the "degree-day," a cumulative measure of temperature that ignores temperatures below a given threshold. This simple measure mimics the way in which many organisms physiologically integrate ambient temperatures, as evidenced by the success of the model.

Other variables operate as complex environmental gradients (Whitaker, 1970) that reflect a suite of primitive gradients that covary in a predictable way in nature. Populations apparently respond to these complex gradients, which act as surrogates for their causal constituents. For example, elevation itself does not cause changes in species composition, but temperature, soils, and rainfall are covarying environmental factors that influence species composition along an elevational gradient. Atmospheric pressure and density so closely follow elevation that their effects can almost be considered to be caused by elevation. The issue of whether variables in a model are causal is complicated by the number of links (indirect effects) and the scale of interest. To illustrate, a mobile population is likely to cue on any convenient signal that a potential habitat falls into its elevation range (e.g., the presence of familiar prey species or nest building materials). Such a proximate cue is likely to be an integrated index of elevation rather than a direct measurement of environmental variables. Similarly, explanatory models may use variables that integrate complex patterns of causal factors, rather than the base causal factors themselves. Elevation is useful because we are familiar with the environmental factors associated with elevation, and because elevation is an integrative factor that carries certain causal complexes with a high degree of reliability.

Understanding the spatial organization of ecological systems is a fundamental part of ecosystem

study. While discovering the causal relationships of this organization is an important goal, our purpose of spatial description on a regional scale is best met by use of explanatory variables that are somewhat removed from the mechanistic causal level. Regional level understanding is best obtained from explanatory variables that reflect spatial gradients at the regional scale and from categorical variables that describe the discrete constituents of (statistical) populations, such as lakes. The scale on which we are concerned with spatial pattern is quite different from the scale of study of ecosystems; our scale is more the scale that Whitaker (1970) addressed in his treatment of gradients (p. 34) than in his treatment of ecological processes (p. 6).

In this paper, we use a regression model to predict lake acid neutralizing capacity (ANC) based on environmental predictor variables over a large region. These predictions are used to produce model-based population estimates. Two key features of our modelling approach are that it honors the spatial context and the design of the sample data. The spatial context of the data are brought into the analysis of model residuals through the interpretation of residual maps and semivariograms. The sampling design is taken into account by including stratification variables from the design in the model. This ensures that the model applies to a real population of lakes (the target population), rather than whatever hypothetical population the sample is a random sample of.

OVERVIEW

Environmental Predictors of Spatial Pattern

Faced with a spatially distributed environmental response variable, our goal is to construct a response surface for spatial pattern that consists of a regression model involving appropriate explanatory variables. There are several qualities of interest to us in identifying a model. First, we seek robust model relations that describe the regional-scale manifestations of environmental processes. This implies that we do not consider spatial processes to contribute to the causal determination of pattern in the response variables at the scale being analysed here. Second, the explanatory variables of interest are spatially extensive (known at each location of interest). When used in the manner proposed, these extensive data can provide enhanced resolution of spatial patterns, as well as enhanced population inferences.

In addition, we account for the relevant sampling design features of the data on which the analyses are based, preserving the informative ties with the well defined populations being sampled. The case study presented here involves a survey of lakes in which the sample was drawn from a list (sampling frame) of lakes in upstate New York that were represented on 1:250,000 scale maps. The Eastern Lake Survey (ELS) was a synoptic survey with a single index sample taken during the a period of low within-lake variability (fall) from each lake. The target population of lakes was defined by eliminating frame lakes that met one or more of the following criteria: (1) the point identified as a lake on the 1:250,000-scale map appeared not to be a lake on more detailed maps; (2) proximity of the lake to intense urban, industrial, or agricultural land use; or (3) lake surface area smaller than 4 hectares. In the context of the sampling design, we

demonstrate the utility of frame data (data available for all lakes in the list frame) in generating improved population estimates, as well as in generating better description of spatial pattern. Our approach is consistent with the model-based approach to population description common to finite sampling, as reported by Royall and Cumberland (1981), for example. The model that we describe here exploits the more-extensive information in the explanatory variables to improve inferences relative to the response variable.

The approach described here can be used at many scales in environmental science, but we will concentrate on regional level studies. This implies that we are interested in multivariate relationships that explain spatial patterns over a relatively large region. In two non-stationary kriging techniques, universal kriging and kriging with intrinsic random functions, the pattern model is usually local and parameter estimates for the drift or pattern are generally not of interest. In contrast, residual kriging proceeds from a global model of predicted pattern. While kriged residuals were not actually added to our regression estimates, the iterative residual kriging framework was useful to us in our spatial analysis of residuals.

This case study involves a response variable, acid neutralizing capacity (ANC), for the finite population of lakes. ANC is measured by titration and is used here in $\mu eq\ L^{-1}$. This variable measures the ability of the lake to buffer acidic inputs. Based on sample data from the ELS (Linthurst, et al., 1986), and frame data on the entire population, we predicted lake ANC for population lakes that were not included in the sample. The predictions combine the large-scale spatial patterns inherent in the suite of explanatory variables to provide a multivariate pattern model for ANC, expressed as a spatial response surface. This pattern can be supplemented by local deviations that are obtained by kriging interpolation. In this paper, we focus on pattern estimation rather than on the kriging portion.

The proposed methods have important implications for investigators conducting regional surveys of environmental or ecological resources. There are emerging analytic methods that will utilize both the predicted mean surface and the spatial model of the residuals. Facsimile populations constructed to have exactly the location of real population units and patterns of mean and variability closely similar to the real population are proving very useful in assessment of sampling designs, both before and after conduct of the sample. Those aspects of the present effort directed toward improving the model for residuals will be useful in constructing realistic facsimile populations.

In the spatial analysis of residuals described here, our regression or response surface model can be viewed, from the perspective of geostatistics, as a model of the deterministic portion of a stochastic process or the "drift". (Our use of the term "spatial pattern" is synonymous, at times, with the geostatistical term "drift", but it is also used more generally to refer to observed spatial structure.) This case study is different from usual geostatistical analysis in several respects. The sample is a probability sample of an explicit stratified population of lakes, with the sampling intensity varying among strata. This fact causes us to approach several parts of the analysis differently then if the sample were unstratified. Additionally, predictions are made for the specific locations at which population units occur, rather than for all points in the domain. The

spatial distribution of the frame lakes (the full set of "lakes" represented on the maps used in sample selection) is therefore automatically considered, with no assumption that lake density is spatially uniform. Finally, the use of environmental predictor variables to model spatial pattern departs from the common use of geostatistical models to represent the drift or pattern in that pattern will not be expressed solely as a function of location. This is because ecologists, unlike geologists, generally have access to spatial data that can help to explain large-scale patterns of interest.

The spatial analysis of residuals used here evolved from the iterative residual analysis procedure outlined by Neuman and Jacobsen (1984). Details of the methods used can be found in Jager et al. (in preparation). Beginning with a regression analysis that assumes uncorrelated and homogeneous errors, the program alternately estimates semivariogram model parameters from the resulting residuals, constructs the variance-covariance (VC) matrix of the errors, and reestimates the regression coefficients by weighted least squares (as defined in Graybill, 1983; p.177), until the estimates stabilize. We have modified the earlier analyses by accounting for heterogeneity of variance of the residuals among strata prior to spatial analysis of residuals.

Punctual kriging maps were produced using SURFER (Golden Software) and point semivariograms were used in the spatial analysis of residuals. Our interpretation of this data is that the volumetric support of the ANC measurement is 6.2 L based on the size of the sample bottle used. We did not consider the measurement as a lake average because "each lake is represented by an index chemistry, rather than, for example, mean chemistry or some other integration over time and space." (Linthurst et al., 1986; p. 3).

Residual Analysis — General Considerations

Regression analysis and geostatistics share the same basic model,

$$\mathbf{Z} = \mathbf{Xb} + \mathbf{R} \qquad [1]$$

but while regression focuses on the model and parameter estimates **b**, geostatistics focuses on the correlation structure of the residuals, **R**. These are opposite sides of the same coin. In regression, unknown properties of the residuals are a hindrance to analysis. In geostatistics, the presence of a deterministic spatial pattern (drift) is a hindrance to obtaining a valid semivariogram model. The usual ordinary least squares regression procedure (OLS) assumes that the (residual) errors are homogeneous and uncorrelated. Both heterogeneous and correlated errors can be taken into account by using generalized least squares (GLS), a procedure that explicitly accounts for the variance-covariance structure of the residuals. In the presence of heterogeneous and correlated errors, estimates by OLS will be less efficient than estimates by GLS, but still unbiased. Loss of efficiency is modest unless correlation is high, although the efficiency lost due to heterogeneity of errors may be more serious. Thus analysis by OLS can be used to generate the residuals that provide the initial basis for residual analysis. Iteration and reanalysis by GLS is not expected to generate a greatly different set of residuals, and the theoretical gain in

efficiency from GLS will not be fully achieved because the variance-covariance (VC) matrix must be estimated from the observed residuals.

Analysis by either OLS or GLS can eliminate spatial pattern (drift) and better provide for effective kriging. If the residuals satisfy weak stationarity assumptions, then the empirical semivariogram will describe the VC structure of the sampled lakes. The residuals will not fully satisfy the kriging assumptions if a deterministic trend or spatial pattern remains. The presence of a gradient-like spatial pattern can be noted from an experimental semivariogram constructed from the residuals (the residual semivariogram). Several features of such a semivariogram confounded with drift are anisotropy (the correlation depends on direction, as well as distance), a sill that rises above the sample variance, and a parabolic shape near the origin (Starks and Fang, 1982). Alternating least squares and spatial analysis of residuals thus seems a natural approach to investigating spatial pattern.

ADIRONDACK CASE STUDY

The U.S. Environmental Protection Agency conducted a synoptic survey of lakes in the eastern U.S. during the fall of 1984 (Linthurst, et al., 1986; Blick, et al., 1987). A probability sample of lakes was drawn from the 1:250,000 scale map population. The sample was stratified by previously drawn contours that identified regions in which surface waters would be expected to have low, medium, or high alkalinity on the basis of geology, soils, and other information. Figure 1 shows the region of upstate NY that we are concerned with, the sample lakes, and the boundaries of three strata. Many chemical and some physical measurements were made on each of the lakes included in the sample, including ANC. The base properties of the population/sample in Subregion I-A (containing the Adirondacks) that is addressed by our analysis are provided in Table 1. Strata are alkalinity classes. $\hat{N}$ is the estimated size of the target population, as specified by the survey, and n is the number of target lakes in the sample.

Stratum	Frame Size	Sample Size	Weight	n	$\hat{N}$	SE($\hat{N}$)
1	711	75	9.633	57	549.08	33.08
2	542	65	8.338	51	425.54	26.13
3	431	68	6.719	47	315.79	22.14
Total	1684	208		155	1290.11	

Table 1 -- Characteristics of the Sampling Design for Subregion I-A of the Eastern Lake Survey (Linthurst, et al., p 36).

The sample weights vary among the strata; weights can be thought of as the number of population lakes represented by each sample lake. Although the differences in weights in Table 1 are not large, the strata should still be analyzed separately. The differences between the sample size of 208 and 155, and between the population size of 1684 and the estimated 1290, are due to non-target lakes contained in the frame population. $\hat{N}$ is the estimated size of the target population, and all design-based estimates apply to this population. In our exercise of predicting ANC for the non-sample lakes, we are unable to identify the lakes in the frame population that are non-target; thus our projections are for the full frame population of 1684 lakes. This is a fundamental limitation to this model-based method; any subpopulation must be identifiable on the frame in order to make inferences in this manner about that subpopulation. However, the spatial patterns and distribution of the frame lakes still provide assistance in making inferences about the spatial patterns of ANC among lakes, even though those inferences cannot be restricted to the target population.

Data from probability samples sometimes pose analytic problems by virtue of design features that invalidate conventional statistical analyses. In this example, there is only one general restriction and that is associated with the strata, and with the differential sampling rate among the strata. It is necessary to conduct the regression analyses by stratum, and to maintain the stratum distinctions if the fitted models are different. If the regression relations are equal for different strata then those strata can be pooled for analysis and prediction. In certain cases, regressions weighted by the sampling weights are appropriate. Likewise, the sample weights should be taken into account when kriging and an extension of the method proposed here can be used for kriging from a variable-probability sample.

Regression Analysis

Several environmental and topographic variables were considered as possible predictors of ANC. These were elevation (m), pH of precipitation, precipitation (cm per year), and watershed slope (%). Elevation and slope were obtained for the lake locations (both those sampled and predicted) from the TOPOCOM digitized elevation maps from 1:250,000 scale maps. The pH of precipitation and precipitation amount were obtained from relatively large-scale maps (Olsen and Slavich, 1986) converted to GIS coverage. In addition, the stratum to which each lake belongs was included as an indicator variable. All interactions between these variables were evaluated for inclusion in the model. Other potential explanatory variables, such as lake type, lake depth, and lake size were not available simply because they were measured only on the sample lakes. However, inclusion of these variables in the regression analysis provides evidence of their potential value, in case the frame data were to be made available.

Logarithmic transformation of ANC was made to make the distribution more symmetric. The new variable was defined as, LANC = $\log_{10}$(ANC + 150), to also account for possible negative values of ANC. LANC was regressed on the suite of explanatory variables. The regression equation for ANC for lakes belonging to the two lower ANC strata (strata nos. 1 and 2 in Table 1) is shown in Equation 2. Equation 3 gives the equation for lakes in the high ANC stratum no. 3. These were obtained simultaneously by using dummy variables to indicate stratum

membership for all stratum-specific parameters ($R^2 = 0.67$).

$$\text{LANC} = -9.08 - 0.0012 \text{ Elevation} + 2.78 \text{ pH} \quad [2]$$
$$\text{LANC} = 6.77 - 0.0012 \text{ Elevation} - 0.85 \text{ pH} \quad [3]$$

These prediction equations produce a spatial pattern for LANC by virtue of the spatial pattern of the explanatory variables, elevation and pH. Part of the pattern in ANC has been "explained" by the association with these two spatially patterned explanatory variables. Figure 2 illustrates the spatial pattern apparent in the observed data, and Figure 3 the smoother patterns in the fitted regression. Figure 4 presents the spatial patterns observed in the predicted values generated on the frame population; visual enhancement derives from the great amount of information present in the explanatory variables known on the frame. The known spatial patterns in these explanatory variables are interpreted via the regression equations to generate spatial patterns in LANC (the response surface). The surfaces in these figures were generated by the kriging routine in SURFER.

Residual Analysis — Case Study

Given that we have generated a response surface reflecting the pattern predicted by the explanatory variables, we now wish to explore spatial patterns in the residuals, both in terms of means of residuals, and in terms of their variance/covariance structure. The spatial properties of residuals are used here to identify a regression model that leaves no evidence of unexplained pattern and to improve the precision of the regression analysis. In some cases, this structure can also be used to improve the response surface predictions by adding kriged residuals.

Residual analysis begins with the residuals produced by the GLS regression models, [2] and [3]. Directional semivariograms for both LANC and the residuals were examined for anisotropy and other indications of drift (Figure 5a and b). The residual semivariograms do not differ markedly and are therefore more isotropic. We interpret this as verification that the regression model has removed much of the gradient-like pattern in LANC. Note also that the relative nugget is larger in the residual semivariograms. We interpret this to mean that the model has partially explained the apparent spatial autocorrelation in LANC.

As the three strata are spatially defined, it is natural to examine the magnitudes of residual variance for evidence of homogeneity, by stratum, as in Table 2.

Given evidence of heterogeneity of variance among strata, it is easy to eliminate these differences simply by dividing the residuals from the various strata by their respective standard deviations. Use of standardized residuals is preferred so that all strata can be analyzed together, spatially and otherwise, for further pattern. Post-stratification based on the observed pattern is possible, with the limitation that scaling requires these classes (clusters) to be identified on the frame (classes must be identified in terms of frame variables).

Further increase in the nugget after scaling leaves very little residual spatial autocorrelation

Stratum	n	Residual Mean Square Error	Root-Mean Square Error
1	57	0.0115	0.1043
2	51	0.0205	0.1431
3	47	0.0607	0.2464
Pooled	155	0.02938	0.1714

Table 2 -- Summary of LANC residual error for each stratum

(Figure 6). Attempts to fit a parametric model for residual semivariance by maximum likelihood estimation revealed no significant difference between the nugget and the sill parameter. In addition, the predictions made using the fitted semivariogram in cross-validation showed virtually no correlation with the measured values. This leads us to wonder whether a "perfect" explanatory model implies uncorrelated errors? It would be surprising if there was not a remnant of small-scale autocorrelation reflecting high-resolution spatial processes below the resolution of the regionally defined explanatory variables. In our example, such processes could affect two lakes very close together. The issue is clouded by the simple fact that correlation among residuals, due to the regression estimation process, will generate spatial autocorrelation when the explanatory variables are spatially patterned. Thus, we would not expect total elimination of autocorrelation in the residual surface, even if the model were perfect, having uncorrelated errors. In this case study, we conclude that there is a virtual absence of autocorrelation following the pattern fit, which suggests that there is no evidence of remaining spatial pattern in these data.

Patterns in the residuals exposed by a spatial analysis of residuals contribute to the general understanding of the population. The VC matrix can be thought of as a partitioned matrix with blocks of locations (sample units) that share the same variance on the diagonal. Recall that differences in variance among strata have been eliminated by scaling, so that all remaining variation will be on subpopulations that are not identified by the strata. If spatial autocorrelation were present, then the off-diagonal blocks would define the covariances between sample units from different subpopulations (in this case, strata).

IMPLICATIONS FOR NATIONAL MONITORING PROGRAMS

Model-Based Estimation - Case Study

Population statistics for the population of Adirondack lakes obtained using our pattern model are given in Table 3, in contrast to statistics obtained from the design-based methods employed in the Eastern Lake Survey. Although there is a potential inherent increase in precision from use

of the model-based method, the inability to identify non-target lakes in the frame population prevents us from taking advantage of this potential in generating population statistics. The model-based results of this table should be considered only as indicating the potential of the method.

The methods of projection to obtain Table 3 are straightforward. For model-based analyses, a predicted value is made for each lake in the frame but not in the sample; these are combined with the sample values and the full set of 1649 lake estimates are simply analysed for the population characteristics. The backtransformed model-based estimates for ANC in Table 3 are generated by backtransforming the predicted and observed values of LANC: $ANC = 10^{LANC} - 150$. The two alternative model-based estimates are discussed below. Design-based estimates are produced by expansion of the sample values to population estimates. This is performed via standard probability sample estimation formulae, such as provided by Overton, this volume. In this framework, LANC and ANC are simply treated as two variables on the sample. In Table 3, $\hat{N}$ is the estimated population size, $\hat{T}_y$ estimates the total ANC over all lakes projected to be in the population, and $\hat{N}_{50\ \mu eq/L}$ is the estimated number of lakes with ANC <= 50 μeq L^{-1}.

	LANC		ANC (μeq · L^{-1})			
				Model-Based		
Statistic	Design-Based	Model-Based	Design-Based	Back-Transform	Bias-Corrected	Non-Transformed
$\hat{T}_y$	3,146	3,882	278,325	199,716	227,664	319,698
Mean	2.439	2.354	215.74	121.11	138.18	193.87
Std. Dev.	0.286	0.246	413.34	194.29	206.98	315.55
Median	2.399	2.315	100.54	56.44	69.18	75.99
$\hat{N}$	1290.1	1649	1290.1	1649	1649	1649
$\hat{N}_{50\mu eq/L}$	487.6	774	487.6	774	691	664

Table 3 -- Comparison of Model- and Design-Based Population Estimates for Adirondack Lakes (Region I-A)

Several features of the predictions are notable. First, $\hat{N}$ is different for the two estimation methods: while the sample provides an estimate of the size of the target population for the design-based method, the target lakes are not identified on the frame used in producing model-based estimates. Comparing the estimates of $\hat{N}$ and $\hat{N}_{50}$ suggests that a large proportion of the non-target lakes have low ANC, but we have no way to confirm this from the survey data. Secondly, the discrepancy in sample sizes (1649 vs. 1684) is accounted for by 35 lakes from the

high ANC stratum in the far west portion of the state that were included in the original Lake Survey frame but that are not included in our analysis because data on the predictor variables are lacking. Excluding sample lakes west of 76° from the design-based estimates has little effect on the gap between the design-based and model-based estimates. Thirdly, for LANC, there is reasonably good agreement between the estimates that are less sensitive to $\hat{N}$ - the population size (the mean, standard deviation, and median), and poor agreement for those parameters influenced by population size ($\hat{N}_{50}$ and $\hat{T}_y$). Agreement of $\hat{T}_y$ would be greatly improved if it were possible to remove the non-target lakes from the model-based estimates. Finally, agreement is poor for all estimates of the parameters for ANC. The model-based estimates (generated from the backtransformed predictions of LANC) appear to be badly biased. This bias can be reduced by adding a bias-correction term involving the estimated prediction variances, s^2, from the regression: $E[ANC] = \exp\{k \cdot LANC + \frac{1}{2}k^2s^2\} - 150$, where $k = \ln(10)$. This correction does narrow the gap between the model-based and design-based estimates, but not by much. Alternatively, the analysis can be conducted without the log transformation. Preliminary results of a re-analysis without transforming ANC showed better agreement with the design-based estimates, especially for mean ANC. Resolution of the remaining differences hinges on identification of the non-target lakes.

This discussion of the results of Table 3 suggests the need for a well-defined strategy for the use of models in population estimation. If model-based estimation is based on transformation, followed by back-transformation of predictions, then the resultant substantial biases must be dealt with in some manner. Bias correction may not be satisfactory; it generally seems preferable not to transform and to accept the loss of efficiency in return for consistency and near-unbiasedness. Loss of efficiency can be reduced by weighted regression, without incurring bias in predictions, but simple unweighted regressions of the natural variable may still be preferable. However, if the transformed variable represents an alternate measure of the population that is in some sense more satisfying from a scientific standpoint, then perhaps the criteria should be consistency and near-unbiasedness for the parameters of this alternate population, rather than for the original one. In Table 3, one should ask if LANC or ANC is more relevant.

Model-Based Estimation

Model-based methods for predicting survey variables on the unsampled population units can be valuable for regional surveys of environmental resources. A multivariate regression model such as the regression model described here can be used to predict unknown population values from relationships developed from the survey sample. The uncertainty of population estimates can be reduced considerably through the use of information contained in explanatory variables and imparted through a model. Additional information about the location of each resource unit has general potential of further reducing the uncertainty through use of interpolation models. The pattern in the fitted surface is thus generated by the known patterns in the explanatory variables, so that these known patterns have been used to strengthen spatial inferences from the sample. The situation found in this case study would seem to be highly desirable, from the point of view of ecological response surfaces, with virtually all apparent spatial pattern accounted for by the explanatory variables.

The spatial analysis of residuals used in our model-building procedure (evaluation of residual maps and spatial autocorrelation) was useful to us as a model diagnostic tool. The role of geostatistics was restricted to exploration of the spatial pattern of residuals. Model diagnostics such as these increase in value when they provide feedback to ecologists in a meaningful context (e.g., geographic maps) because it allows them to bring external knowledge to bear on the problem.

Application of this model approach to developing regional analysis of spatially distributed environmental data depends on the general availability of extensive data for explanatory variables. In effect, these extensive data must be available for the entire frame of population units. Digitized GIS coverage of spatially extensive environmental data are very valuable for a regression-based approach, but must be in a specific form suitable to the needs of any particular study. Identification of needs and possibilities, and provision of specific explanatory variables as part of the frame materials, would seem to be an integral part of any comprehensive monitoring program.

The validity of model-based estimates depends on the appropriateness of the assumptions made. In the Adirondack case study we have already raised one issue, whether data require transformation to meet model assumptions, and this is probably just the tip of the iceberg. Basing the estimates on a probability sample ensures that an assumption-free (design-based) methodology is always available. Then model-based methods can be used to enhance inferences without the general validity of those inferences depending on the model assumption (see Overton, this volume).

This case study illustrated several important points for future environmental surveys. First, we demonstrated the potential of model-based estimation for the description of spatial pattern in environmental variables. Second, we learned the importance of adequate investment in frame materials as the inadequacies in ELS frame data prevented us from producing population estimates for the target population of lakes. Finally, the value of analyzing residuals in a spatial context was underscored by the model-building procedure that we chose.

REFERENCES

Blick, D.J., Messer, J.J., Landers, D.H., and Overton, W.S. (1987) Statistical Basis for the Design and Interpretation of the National Stream Survey, Phase I: Lakes and Streams. *Lake and Reservoir Management*, 3: 470-475.

Gleason, H.A. (1926) The Individualistic Concept of the Plant Association. *Bulletin of the Tory Botanical Club* 53: 7-26.

Graybill, F.A. (1983) *Matrices with Applications in Statistics*. The Wadsworth Statistics/Probability Series, Wadsworth Publishing Co., Inc., Belmont, CA.

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., Kramer, M., and Overton, W.S. Multivariate Modeling of Environmental Patterns in Space (in preparation).

Linthurst, R.A., D.H. Landers, J.M. Eilers, D.F. Brakken, W.S. Overton, E.P.Meier, and R.E. Crowe (1986) Characteristics of Lakes in the Eastern United States. Vol I, Population Description and Physico-Chemical Relationships. *EPA/600/4-86/007a.*

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

Overton, W.S. (1989) Calibration Methodology for the Double Sample of the National Lake Survey Phase II Sample. *Tech. Report 130*, Department of Statistics, Oregon State University, Corvallis OR 97331.

Royall R.M. and W.G.Cumberland (1981) An Empirical Study of the Ratio Estimator and Estimators of its Variance. *J. Amer. Statis. Assoc.* 76: 66-77.

Starks, T.H. and Fang, J.H. (1982) The Effect of Drift on the Experimental Semivariogram. *Math. Geol.* 14(4): 309-319.

Whitaker, R.H. (1970) *Communities and Ecosystems.* The McMillian Co., New York, 158 pps.

DISCLAIMER

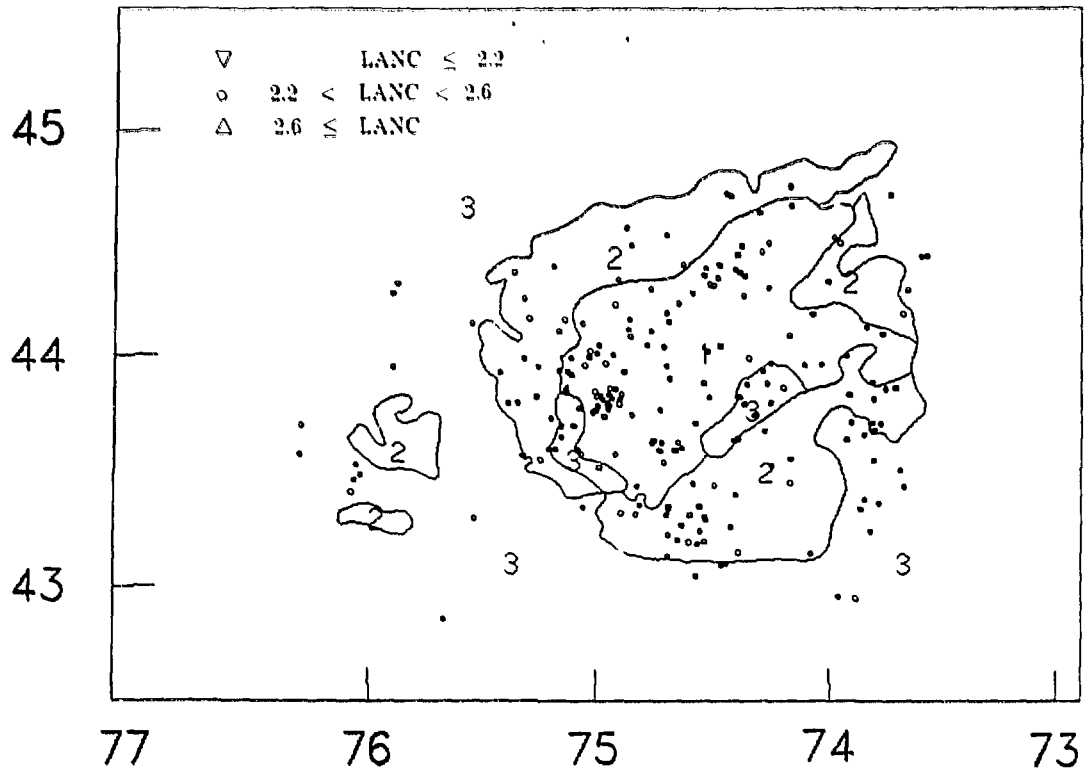

Figure 1 -- Map of sample lakes and stratum boundaries

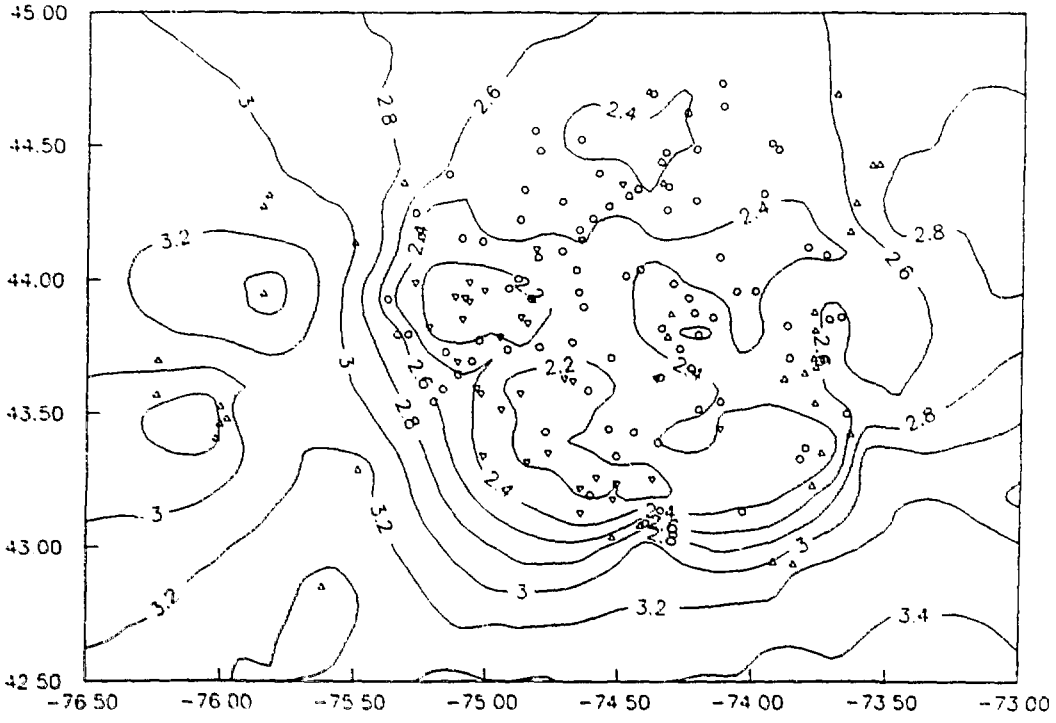

Figure 2 -- Map of observed LANC values on the sample

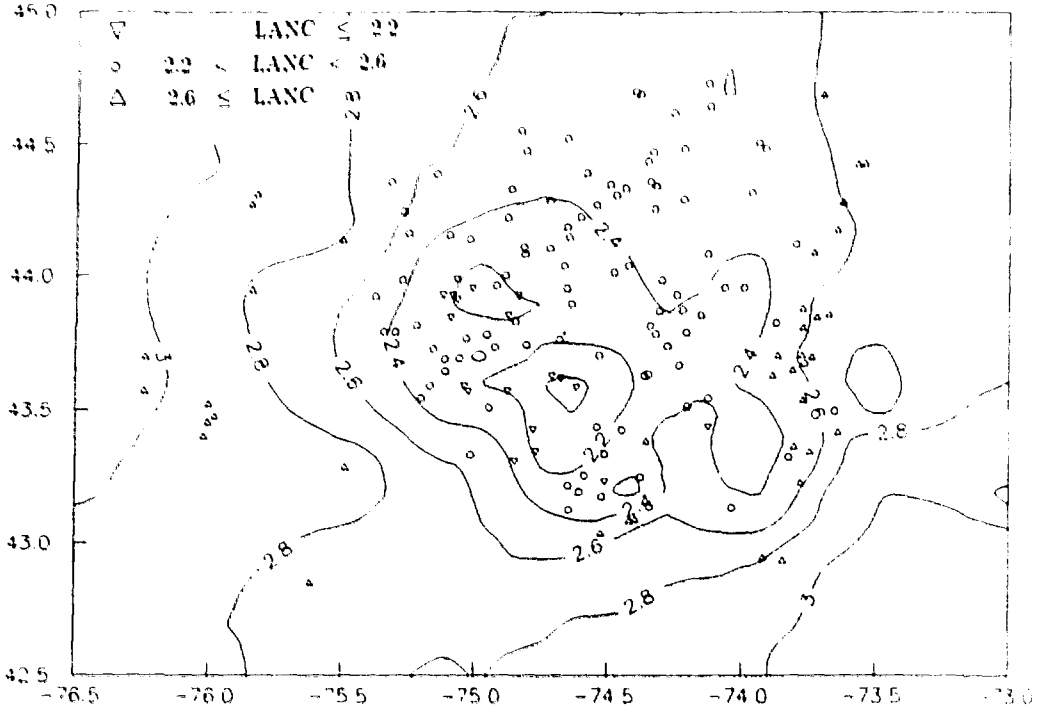

Figure 3 -- Map of predicted LANC values on the sample

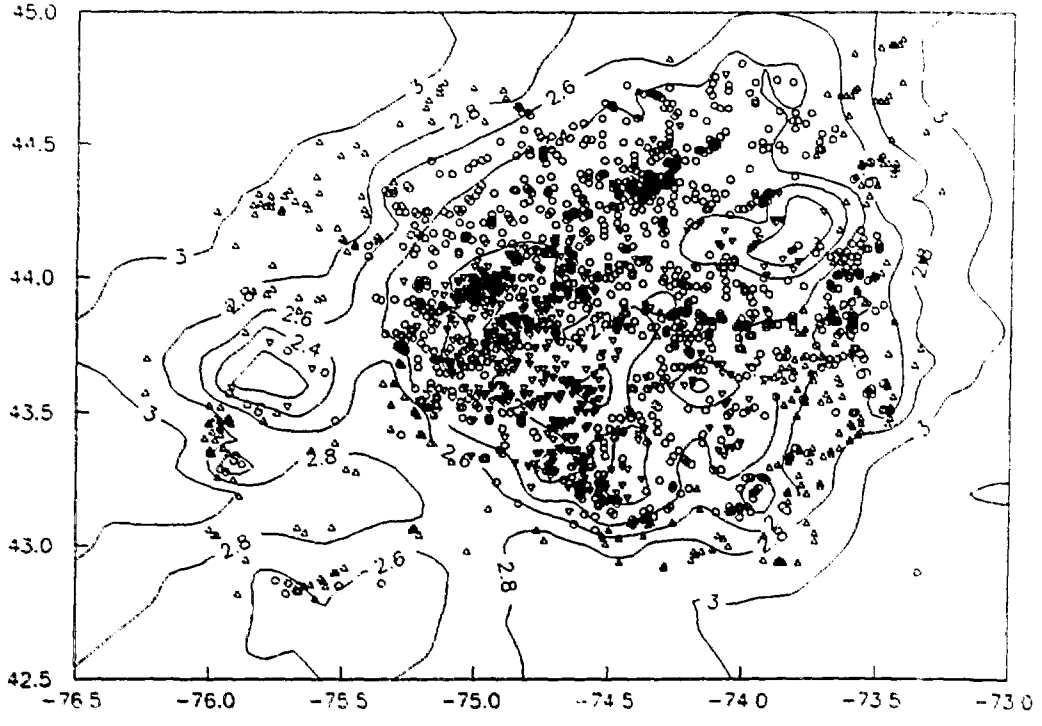

Figure 4 -- Map of predicted LANC values on the frame population

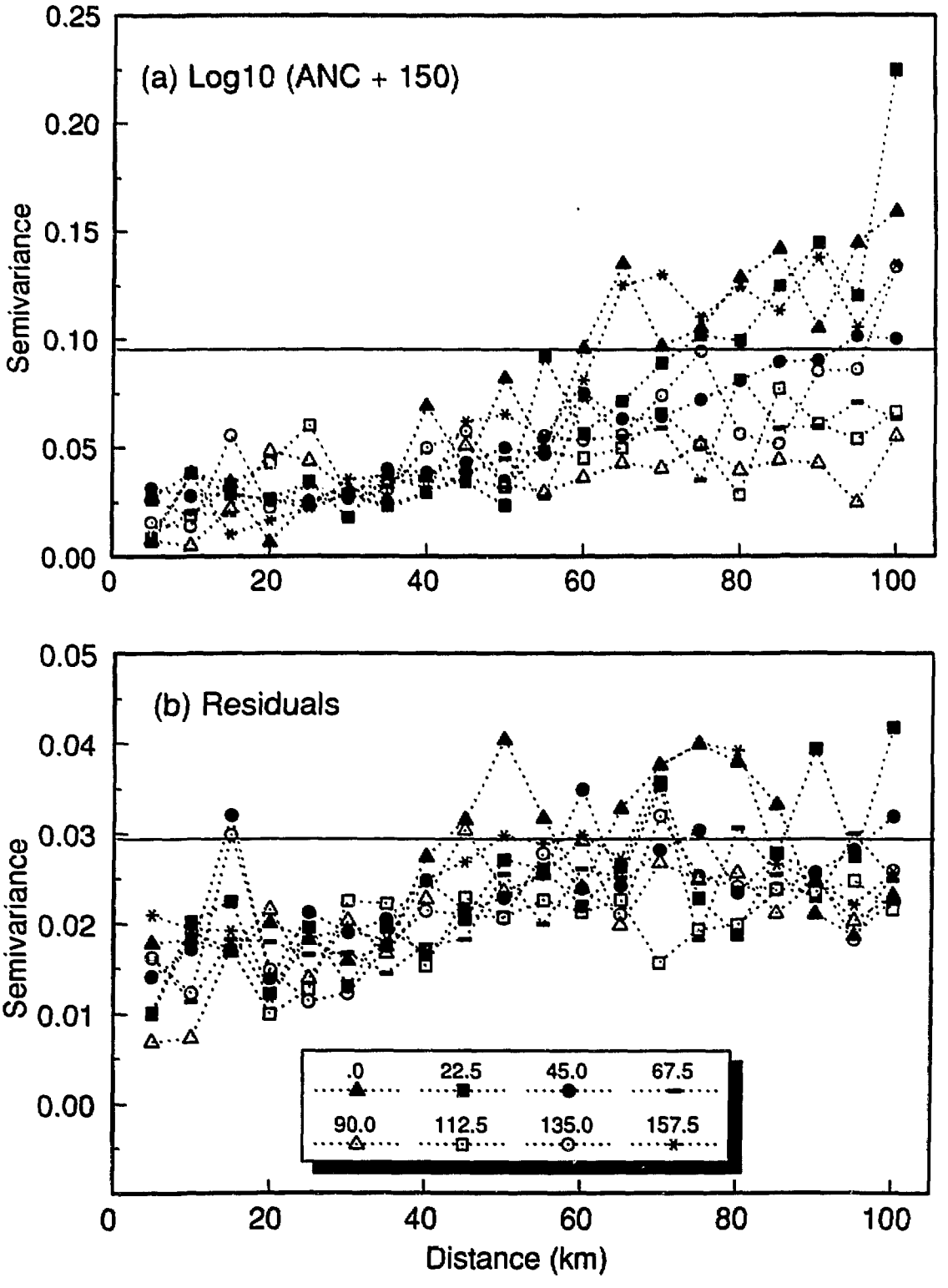

Figure 5 -- Directional semivariograms for (a) LANC and (b) scaled residuals.

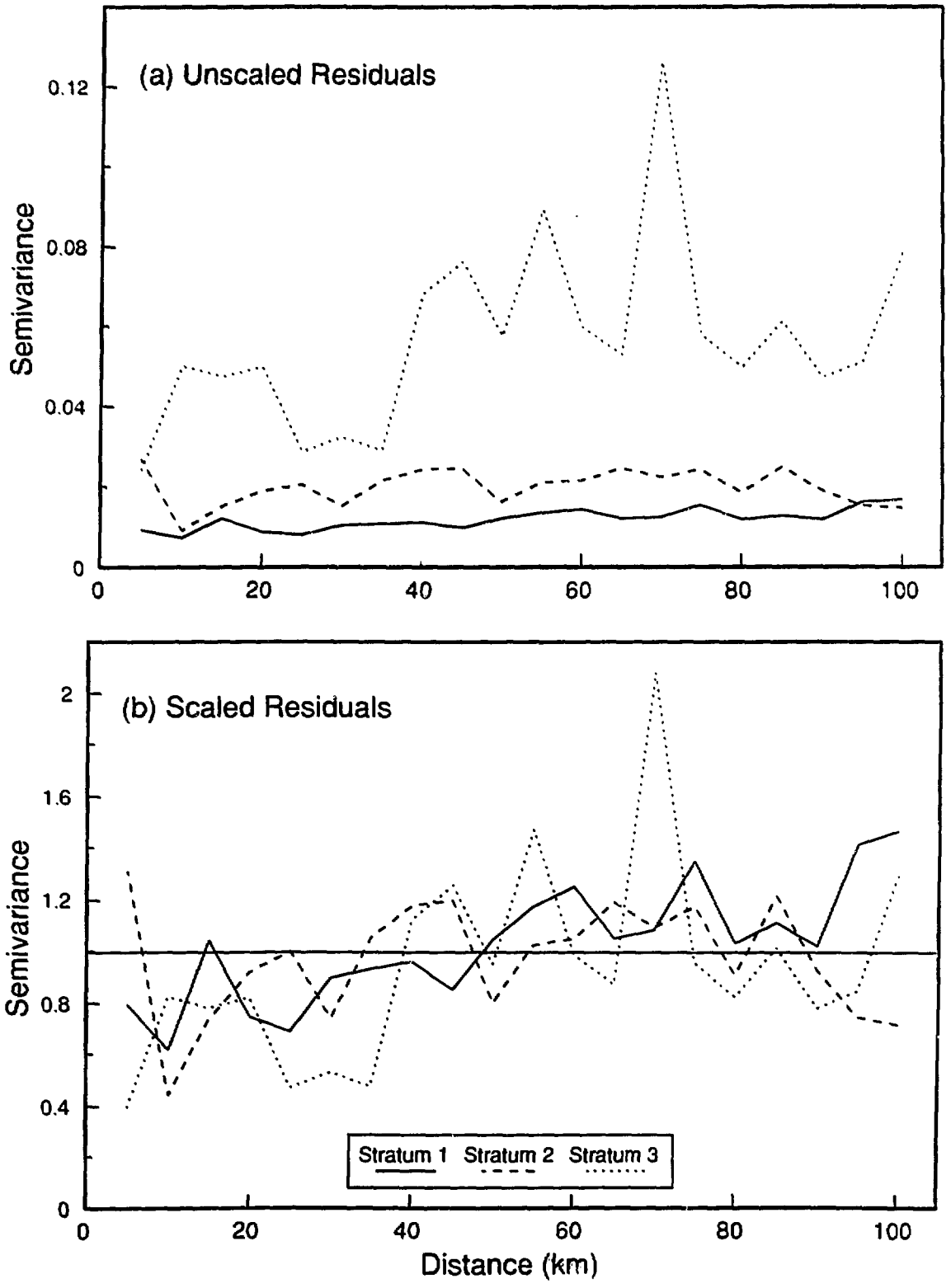

Figure 6 -- Semivariograms of (a) unscaled and (b) scaled residuals.