Spatial Analysis of the Fish Species Richness of Adirondack Lakes:

Applications of Geostatistics and Nonparametric Regression

BU-1186-M

February, 1993

James P. Hobert¹ Biometrics Unit Cornell University

N. S. Altman²

Biometrics Unit

Cornell University

Carl L. Schofield*

Natural Resources Dept.

Cornell University

¹Research supported by National Institute of Environmental Health Sciences Training Grant EHS-5-T32-ES07261-03.

 $^2\mathrm{Research}$ supported by Hatch Grant 151410 NYF.

* James P. Hobert is a graduate student of statistics, Biometrics Unit, Cornell University. N. S. Altman is Assistant Professor, Biometrics Unit, Cornell University. Carl L. Schofield is Senior Research Associate, Natural Resources Department, Cornell University, Ithaca, NY 14853. The authors thank C. McCulloch and G. Casella for comments that substantially improved this article and S. P. Neuman for providing them with useful data. • '• , , ,

Abstract

۰.

Legislation passed in 1990 lowering the allowable sulfur dioxide emission levels in the U. S. should reduce acidity in the Adirondack lakes of New York State. Fish species richness, the number of species existing in a lake, is one indicator of lake conditions which may be affected by acidity.

Fish species richness in Adirondack lakes depends on a number of physical, chemical, and biological factors, including area, depth, acidity, and predation. Data on these and other factors are available for a sample of 1166 lakes. The data are analysed with the goal of quantifying the effects of acid deposition after controlling for the other important factors, which include spatial location, to predict the likely effect of the 1990 legislation on species richness.

Two independent analyses were carried out: kriging with covariates and nonparametric regression within the framework of semi-parametric additive models. Nonparametric regression was found to provide a simpler, interpretable model for these data.

The environmental conclusion based on the nonparametric regression equation is that sulfate deposition reductions in the Adirondack region would certainly not decrease species richness and could be instrumental in its increase.

KEY WORDS: Backfitting algorithm; Fish species richness; Geostatistics; Nonparametric regression.

1. INTRODUCTION

Legislation (National Acid Precipitation Assessment Program 1990) was passed in 1990 decreasing allowable sulfur dioxide (SO_2) emission levels in the United States. A decrease in SO_2 emission levels should result in lower levels of sulfate deposition. This in turn should lead to significant changes in the pH of many lakes in the Adirondack region of New York State (Schofield 1990).

The goal of this analysis was to use the data gathered during the Adirondack Lakes Study (Kretser, Gallagher and Nicolette 1989) to investigate the effect of changes in sulfate deposition on fish species richness, which we take as the number of different species existing in a lake. The available data are measurements of fish species richness and nineteen covariates on each of the 1166 drainage lakes sampled. (The original data set described 1469 lakes but reclaimed lakes, non-drainage lakes and those with missing observations were not included.) The nineteen covariates are: lake area (hectares), lake pH, mean lake depth (meters), lake latitude, lake longitude, lake elevation (meters), lake type (six types classified by drainage), concentrations (micromoles per liter) of calcium, SiO_2 (silica), total aluminum, monomeric aluminum, dissolved organic carbon, chloride, and total phosphorus, an indicator variable for thermal stratification (shallow or deep) and four indicator variables for the presence or absence of bogs, beavers, predacious fish, and outlet dams (which would prevent migration).

An obvious and easily implemented statistical procedure to use in this situation is standard multiple regression analysis. A multiple regression equation was formulated (which did not include latitude or longitude as covariates) but a spatial

1

plot of the residuals suggested a violation of the assumption of independent errors. This regression equation and its problems are the topic of Section 2. Sections 3 and 4 of this article describe two fundamentally different attempts to adequately describe species richness as a function of the covariates with a more realistic model incorporating spatial variation. Although these two analyses are quite different, they are both based on models which are simple modifications of the standard regression model. The first analysis is based on geostatistics methodology (Cressie 1991) and the second on semi-parametric additive models (Hastie and Tibshirani 1990) and nonparametric regression. (See Yakowitz and Szidarovszky (1985) for an enlightening formal comparison of 'riging and (kernel) nonparametric regression.) Section 5 gives environmental and statistical conclusions.

2. MULTIPLE REGRESSION

2.1 MODEL AND EQUATION

· · ·

The response variable Y_i was defined by $\ln(\text{species richness}_i + 1)$. The first step in constructing a regression model was choosing a reasonable set of covariates from the nineteen available. The SAS procedure RSQUARE (SAS Institute 1985), Mallows's C_p statistic, and previous studies of this type (Schofield 1990) were used to determine a reasonable subset of the covariates. The parameters for a model with these covariates and all first order interaction terms were estimated. Each term in the model was kept or discarded based on a hypothesis test that the true value of its parameter was zero. The result is the following regression model

$$Y_{i} = \beta_{0} + \beta_{1} PRED_{i} + \beta_{2} PH_{i} + \beta_{3} AREA_{i} + \beta_{4} SILI_{i} + \beta_{5} DEPTH_{i} + \beta_{6} PH_{i}^{2} (2.1)$$
$$+ \beta_{7} AREA_{i}^{2} + \beta_{8} (PH*PRED)_{i} + \beta_{0} (PH^{2}*PRED)_{i} + \beta_{10} (AREA*PRED)_{i} + \epsilon_{i}.$$

The covariates in (2.1) are described in Table 1. The last three covariates are interactions and the ϵ_i are assumed to be iid N(0, σ^2), i=1,2,...,1166.

Least squares estimation of the parameters in (2.1) resulted in the following equation for the conditional expectation, $E(Y_i|x)$,

$$E(Y_{i}|x) = -0.778 + 0.615pH_{i} + 9.60*10^{-3}AREA_{i} + 1.90*10^{-3}SILI_{i} + 2.81*10^{-2}DEPTH_{i} - 3.94*10^{-2}pH_{i}^{2} - 4.41*10^{-5}AREA_{i}^{2}$$

for those lakes possessing predator species, and

(2.2)

$$E(Y_{i}|\mathbf{x}) = -7.736 + 2.435pH_{i} + 1.435*10^{-2}AREA_{i} + 1.90*10^{-3}SILI_{i} + 2.81*10^{-2}DEPTH_{i} - 0.163pH_{i}^{2} - 4.41*10^{-5}AREA_{i}^{2}$$

for those lakes possessing no predator species.

The R-square for this regression equation was 0.6958 and the mean square error (MSE) was 0.2125. Standard diagnostic plots did not suggest the violation of any of the assumptions of model (2.1) aside from that described below.

2.2 RESIDUALS

The fact that longitude and latitude were not covariates in model (2.1) is suspicious since it is intuitively clear that spatial location is an important piece of information for explaining species richness (because it is a proxy for important, but unavailable, covariates). This in mind, it is not surprising that the residuals from (2.2) are spatially correlated. Figure 1 shows a smoothed three-dimensional plot of residuals versus position of lakes. It is quite evident from this plot that there is a strong northwest to southeast trend among the residuals.

The trend in Figure 1 suggests that the assumption of independent errors of model (2.1) is violated. If the true errors in (2.1) are not independent, prediction intervals based on (2.2) will not necessarily have their prescribed confidence levels and it is difficult to determine what the true confidence levels are.

Figure 1 demonstrated that spatial location is an important covariate (which agreed with our intuition), but lake latitude, lake longitude, their squares and the product of the two were deemed insignificant covariates during the model building stage. The problem is that by using the multiple regression model (2.1), we were assuming that we could describe the effects of spatial location on species richness with <u>linear</u> functions of lake latitude, lake longitude, their squares or some simple function of the two. This was apparently a poor assumption in this case and in order to correctly model the effects of spatial location, a more flexible model must be used. The development and application of such a model is the topic of the remainder of this article.

3. SPATIAL ANALYSIS I: VARIOGRAMS AND KRIGING

3.1 MODEL

•

The assumption of iid errors in model (2.1) was unreasonable because the residuals of equation (2.2) were spatially correlated. One way to improve on (2.1) is to abandon the iid errors assumption and model the correlation structure of the residuals. The modified model is

$$Y_{i} = \beta_{0} + \beta_{1} PRED_{i} + \beta_{2} PH_{i} + \beta_{3} AREA_{i} + \beta_{4} SILI_{i} + \beta_{5} DEPTH_{i} + \beta_{6} PH_{i}^{2} \quad (3.1)$$
$$+ \beta_{7} AREA_{i}^{2} + \beta_{8} (PH*PRED)_{i} + \beta_{9} (PH^{2}*PRED)_{i} + \beta_{10} (AREA*PRED)_{i} + Z(\mathbf{z}_{i})$$

which with obvious notation becomes

$$\mathbf{Y}_{i} = \mathbf{x}_{i}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{s}_{i}) \tag{3.2}$$

where \mathbf{s}_{i} , i=1,2,...,1166, are vectors representing the location of the lakes in \Re^{3} . The correlation structure of the <u>stochastic process</u> { $Z(\mathbf{s}_{i}) : i=1,2,...,1166$ } is characterized by the variogram (Cressie 1989)

$$2\gamma(d) = Var(Z(s_i) - Z(s_i))$$
 for i, j=1,2,...,1166 (3.3)

where d is the Euclidian distance between lakes at \mathbf{s}_{i} and \mathbf{s}_{j} . The variogram in (3.3) is unknown and is therefore a parameter as well as a univariate function. It is also isotropic, depending only on the distance between lakes and not on direction. It is further assumed that the Z's are normal and $E(Z(\mathbf{s}_{i}))=0$ which along with (3.3) makes the errors a zero-mean intrinsically stationary process.

3.2 ESTIMATION

If we assume that there exists a covariance function , $C(\cdot)$, on the error process defined as

$$C(d) = Cov(Z(s_i), Z(s_j))$$
 for i, j=1,2,...,1166 (3.4)

then

$$2\gamma(d) = 2C(0) - 2C(d).$$
 (3.5)

If the variogram function were known, (3.5) could be used to calculate the covariance matrix of the residual process and β could in turn be estimated using generalized least squares. Neuman and Jacobson (1984) suggest using an estimate of the variogram, $\hat{\gamma}(\cdot)$, to calculate $\hat{\beta}$ in this fashion and then iterating between $\hat{\gamma}(\cdot)$ and $\hat{\beta}$ until convergence. The first estimate of the variogram is calculated using the residuals from model (2.2), i.e. $\hat{\gamma}_1 \equiv 0$, $\hat{\beta}_1 = \hat{\beta}$ of (2.2).

A valid variogram must be a negative-definite function (Feller 1971, Cressie 1990). Estimates calculated using, say, the "classic" method-of-moments estimator are not necessarily negative-definite. Kriging (Cressie 1989) and the estimation technique of Neuman and Jacobson both require a negative-definite variogram estimate. It is therefore convenient to fit a member of a parametric family of variograms to the initial variogram estimate in order to guarantee negative-definiteness.

One goal of this analysis was to construct prediction intervals for species richness of a prespecified set of lakes under different sulfate reduction scenarios. Standard error calculations (required for prediction intervals) for the iterative method of Neuman and Jacobson are difficult to derive.

To avoid this problem, the Neuman and Jacobson technique was applied to half of the data (selected at random) to get an estimate of the variogram of the error process. Within the algorithm, variogram estimates were all calculated using the method-of-moments estimator and exponential variogram models were fitted to those estimates using the weighted least squares method given in Cressie (1985). The convergence criterion was that no element of $\hat{\beta}$ change by more than one half of one percent between two iterations. Convergence required four iterations and the resulting variogram was

$$2\widehat{\gamma}(d) = 0.115 + 0.321 \left\{ 1 - \exp\left(\frac{-d}{1.01}\right) \right\} \text{ for } d \ge 0.$$
 (3.6)

Figure 2 shows this variogram and the method-of-moments estimate from which it was fit.

Let S represent the set of data points not used in calculating (3.6). Kriging (Cressie 1989) was then used to predict the values of the error process for the data in S, call these $\hat{Z}(\mathbf{s}_j)$, $\mathbf{j} \in S$. The data in S were then transformed by subtracting these predictions from the corresponding responses $(\mathbf{y}_i - \hat{Z}(\mathbf{s}_j))$. We then assumed that the transformed data (in S) followed the original iid error model (2.1). This assumption allowed parameter estimation via standard least squares. Figure 3 is a smoothed three-dimensional plot of the resulting residuals. A trend similar to but less severe than that in Figure 1 is apparent.

The fact that this method failed to remove the spatial correlation of the residuals (in S) led us to suspect that the assumption of an isotropic variogram was invalid. To check the validity of the isotropic assumption, the longitude-latitude plane was divided into four directional classes and separate sample variograms of the initial residuals (from model (2.2)) were calculated for each. A pair of residuals was assigned to the first directional class if the angle (polar coordinates) of the vector through the two points was in the interval $(112.5^{\circ}, 157.5^{\circ})$. The second class contained angles in

the interval (67.5°, 112.5°). Similarly, the third and fourth classes possessed angles in the intervals (22.5°, 67.5°) and (157.5°, 202.5°), respectively. Exponential variograms were fit to each of the resulting variogram estimates. Although no formal comparison was made, these variograms appear to be quite different. Figure 4 shows all four exponential models. The isotropic assumption does not appear to be reasonable. This implies that the above analysis could be improved by abandoning the isotropic assumption and using an anisotropic parametric variogram model throughout. As parametric anisotropic variogram models and the fitting of such models to variogram estimates are not well developed in the geostatistics literature this topic was not pursued further.

4. SPATIAL ANALYSIS II: NONPARAMETRIC REGRESSION

4.1 MODEL

Modeling the correlation structure of the errors via a stochastic process is one way to compensate for the absence of spatial location in the list of covariates. Another way is to allow nonlinear covariate effects. A nonparametric term was added to model (2.2) so that species richness could be modeled as a (not necessarily linear) function of spatial location. The updated model is a semi-parametric additive model (Hastie and Tibshirani 1990)

$$Y_{i} = \beta_{0} + \beta_{1} PRED_{i} + \beta_{2} PH_{i} + \beta_{3} AREA_{i} + \beta_{4} SILI_{i} + \beta_{5} DEPTH_{i} + \beta_{6} PH_{i}^{2} + \beta_{7} AREA_{i}^{2} + \beta_{8} (PH*PRED)_{i} + \beta_{9} (PH^{2}*PRED)_{i} + \beta_{10} (AREA*PRED)_{i} + g(LATI_{i}, LONG_{i}) + \epsilon_{i}$$

$$(4.1)$$

which with obvious notation becomes

$$Y_{i} = x_{i}\beta + g(LATI_{i}, LONG_{i}) + \epsilon_{i}$$
(4.2)

where $\epsilon_i \sim \text{iid N}(0,\sigma^2)$, LATI and LONG represent the lake latitude and lake longitude in decimal degrees and the function g is assumed to be some smooth bivariate function of lake latitude and lake longitude.

4.2 ESTIMATION

Parameter estimation for additive models can be accomplished using the backfitting algorithm (Hastie and Tibshirani 1990). Bivariate locally-weighted quadratic regression (Cleveland 1988), which is a straightforward extension of univariate LOESS (Cleveland 1979), was used to estimate the non-linear component at each step. A span of 127 was selected using cross-validation on the residuals from the original least squares estimate.

The fact that bivariate locally-weighted quadratic regression (BLQR) is a linear estimator affords two things: i) the backfitting algorithm is guaranteed to converge for semi-parametric models when linear estimators are used and ii) computer intensive iteration is not necessary as the final estimates of β and g can be solved for explicitly (Hastie and Tibshirani 1990). Put $H_1 = X(X^tX)^{-1}X^t$ and let H_2 be the hat matrix of the BLQR technique. Then

$$\widehat{\mathbf{g}}^{(\infty)} = [\mathbf{I} + \mathbf{H}_2(\mathbf{I} - \mathbf{H}_1\mathbf{H}_2)^{-1} - (\mathbf{I} - \mathbf{H}_2\mathbf{H}_1)^{-1}]\mathbf{Y}$$
(4.3)

and

$$\widehat{\beta}^{(\infty)} = (X^{t}X)^{-1}X^{t}((I - H_{2}H_{1})^{-1} - H_{2}(I - H_{1}H_{2})^{-1})Y.$$
(4.4)

4.3 SEMI-PARAMETRIC REGRESSION EQUATION

• •

Equations (4.3) and (4.4) were used to calculate the final estimates of β and g. The result was the following semi-parametric regression equation

$$EY_{i} = 0.650 + 0.152 pH_{i} + 9.656*10^{-3}AREA_{i} + 1.649*10^{-3}SILI_{i}$$

+ 3.349*10⁻²DEPTH_i - 4.690*10⁻³pH_i² - 4.441*10⁻⁵AREA_i²
+ $\hat{g}(LATI_{i},LONG_{i})$

for those lakes possessing predator species, and

(4.5)

$$\begin{split} \mathrm{EY}_{\mathrm{i}} &= -5.913 + 1.871 \mathrm{pH}_{\mathrm{i}} + 1.440 * 10^{-2} \mathrm{AREA}_{\mathrm{i}} + 1.649 * 10^{-3} \mathrm{SILI}_{\mathrm{i}} \\ &+ 3.349 * 10^{-2} \mathrm{DEPTH}_{\mathrm{i}} - 0.122 \mathrm{pH}_{\mathrm{i}}^{2} - 4.441 * 10^{-5} \mathrm{AREA}_{\mathrm{i}}^{2} \\ &+ \widehat{\mathrm{g}}(\mathrm{LATI}_{\mathrm{i}}, \mathrm{LONG}_{\mathrm{i}}) \end{split}$$

for those lakes not possessing predator species.

The R-square for this regression was 0.7560 and the MSE was 0.1826. Note that these regression coefficients are nearly identical to those in (2.2). Standard regression diagnostic plots did not suggest the violation of the additive model assumption that $\epsilon_i \sim \text{iid N}(0,\sigma^2)$. Figure 5 is a contour plot showing \hat{g} . This plot shows that \hat{g} is attempting to model the spatial correlation shown in Figure 1.

An estimate of the error degrees of freedom is required in the calculation of prediction intervals. It is not at all clear how many degrees of freedom should be associated with the function \hat{g} . We take as the definition of degrees of freedom of a

linear smoother the trace of its hat matrix (Hastie and Tibshirani 1990). The trace of the final BLQR hat matrix (see (4.3)) was 75 while that of the final multiple linear regression hat matrix (see (4.4)) was 11.

Figure 6 shows a smoothed three-dimensional plot of new residuals (from (4.5)) versus position of lakes. A comparison of this plot with the plot in Figure 1 suggests that the spatial correlation of the residuals has been eliminated. The semi-parametric equation (4.5) is therefore a satisfactory equation on which to base the desired prediction intervals.

4.4 PREDICTION INTERVALS AND BOX PLOT

The legislation discussed in the introduction will most likely lead to a 50% sulfate deposition reduction in the Adirondack region over the next couple of years. A model relating sulfate deposition to lake pH (Schofield 1990) was used to <u>predict</u> the pH of 20 prespecified lakes for the 50% sulfate deposition reduction scenario. These predictions were then used in conjunction with (4.5) to form prediction intervals (PI's) for the species richness of those twenty lakes. (Prediction intervals were also computed for 4 other scenarios (Hobert 1992).) Table 2 shows these PI's, for seven lakes, along with corresponding PI's for a "no change" scenario and PI's based on equation (2.2). (Acidity for the "no change" scenario was predicted using the Schofield model.) Some environmental implications of the results in Table 2 are discussed in Section 5.

Pointwise standard error calculations for the PIs in Table 2 are straightforward since all estimates are linear (conditional on the bandwidth).

The PI's above tell us nothing of the predicted <u>overall</u> reaction of richness to a given decrease in sulfate deposition. Figure 7 attempts to provide a picture of this

"overall reaction" by concentrating on the differences (for all the lakes) in predicted richness for a 50% decrease in sulfate deposition and a 0% decrease (no change) in sulfate deposition. Heuristically, Figure 7 shows that a 50% decrease in sulfate deposition would increase the richness of many of the lakes and not effect the rest. The reason for this is that sulfate deposition reductions effect lakes with pH between 4.5 and 6.5 much more than lakes with pH outside that range (Schofield 1990).

5. DISCUSSION

5.1 STATISTICAL ISSUES

A researcher using the nonparametric regression method must make two (subjective) decisions: which data smoother to use and which bandwidth selection technique to employ. Correspondingly, a researcher using the geostatistical method must pick a parametric variogram family for use in kriging and she must also decide on some convergence criterion for the algorithm of Neuman and Jacobson. Conditional on these two choices, two people using the nonparametric regression technique will necessarily get the same answer. This is not true for the geostatistical method, however, due to the researcher specific choices required during the calculation of the sample variogram. The nonparametric regression method is conceptually simpler and provides a picture of the spatial effects which is more instructive than a picture of the correlation structure of the residuals.

This leads to the conclusion that even if the use of an anisotropic parametric variogram model guaranteed the success of the geostatistical method, the nonparametric regression method would still be preferred.

Based on a formal comparison of kriging and (kernel) nonparametric regression

(without covariates), Yakowitz and Szidarovszky (1985) concluded that nonparametric regression is as efficient as kriging when the assumptions required for kriging to provide consistent error estimates are true (one of these being that the true variogram is known) and that nonparametric regression is either as efficient or more so when those assumptions are not true. The extent to which nonparametric regression is more efficient in the later case depends on the severity of the violations of the assumptions.

5.2 ENVIRONMENTAL ISSUES

Figure 7 suggests that a 50% reduction in sulfate deposition would have a positive effect on (would increase) the species richness of lakes with low current pH's (4 to 6) and have essentially no effect on the richness of those lakes with current pH's outside this range. Although the prediction intervals in Table 2 are quite wide, the fact that all of the predictions of species richness under the 50% reduction scenario are greater than the corresponding current species richness and Figure 7 suggest strongly that a 50% reduction in sulfate deposition in the Adirondack region would certainly not be detrimental to the species richness and could be quite instrumental in raising it.

13

REFERENCES

• • • • •

- Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. J. Amer. Statist. Assoc. 74, 829-836.
- Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. J. Amer. Statist. Assoc. 83, 596-610.
- Cressie, N. (1985). Fitting Variogram Models by Weighted Least Squares. Journal of the International Association for Mathematical Geology. 17, 563-586.

(1989). Geostatistics. The American Statistician. 43, 197-202.

____ (1991). Statistics for Spatial Data. John Wiley and Sons, New York.

- Feller, W. (1971). An Introduction to Probability Theory and its Applications. John Wiley and Sons, New York.
- Hastie, T. J. and Tibshirani, R. J. (1990). Generalized Additive Models. Chapman and Hall, London.
- Hobert, J. P. (1992). Spatial Analysis of the Fish Species Richness of Adirondack
 Lakes: Applications of Geostatistics and Nonparametric Regression.
 Unpublished M.S. thesis, Biometrics Unit, Cornell University, Ithaca, NY, 75 p.
- Kretser, W., Gallagher, J. and Nicolette, J. (1989). Adirondack Lakes Study. 1984 -1987: An Evaluation of Fish Communities and Water Chemistry. Prepared for Adirondack Lakes Survey Corporation, Ray Brook, New York.
- National Acid Precipitation Assessment Program (NAPAP) (1991). 1990 Integrated Assessment Report. The NAPAP Office of the Director, Washington, D.C.

1/

- Neuman, S. P. and Jacobson, E. A. (1984). Analysis of Nonintrinsic Spatial Variability by Residual Kriging with Application to Regional Groundwater Levels. Journal of the International Association for Mathematical Geology. 16, 499-521.
- SAS Institute Inc. SAS User's Guide: Statistics, Version 5 Edition. Cary, NC: SAS Institute Inc., 1985. 956 pp.
- Schofield, C. L. (1990). Fish Species Distribution Patterns in Selected Adirondack Drainage Systems. In: Adirondack Lakes Survey: An Interpretive Analysis of Fish Communities and Water Chemistry, 1984-87. Adirondack Lakes Survey Corporation, Ray Brook, New York.
- Yakowitz, S. J. and Szidarovszky, F. (1985). A Comparison of Kriging with Nonparametric Regression Methods. Journal of Multivariate Analysis. 16, 21-53.

Table 1: Description of the Covariates in Model (2.1).

• •



Table 2: For each of the seven lakes, RICHNESS is the current species richness and ESTIMATE is \hat{Y} based on model (4.10). OBSERVED pH is the current pH of the lake. The last three columns provide information concerning the aforementioned two scenarios. The first row under these columns represents the "no change" scenario and the second represents the 50% reduction scenario. The third row is based on the original multiple regression equation (2.2). NEW pH is the predicted pH for the corresponding change in sulfate deposition, PREDICTION is the predicted richness for that pH and 95% PI is a 95% prediction interval for that estimate.

	<u>0</u>]	BSERV	ED	<u>NEW</u>	•	
<u>LAKE</u>	RICHNESS	pH	ESTIMATE	рH	PREDICTION	<u>95% P.I.</u>
Gull	0	5.02	1.101	5.54	1.852	(0.212 , 5.714)
				5.84	2.302	(0.403 , 6.773)
				5.84	3.600	(0.860 , 10.382)
Colden	0	5.07	0.984	5.34	1.336	(0,4.526)
				5.91	2.113	(0.317 , 6.362)
				5.91	3.212	(0.702 , 9.421)
Silver	0	4.29	0.026	4.32	0.052	(0,1.474)
Dollar				4.51	0.224	(0, 1.876)
				4.51	0.043	(0, 1.580)

	OBSERVED		NEW				
LAKE	RICHNESS	<u>рН</u>	ESTIMATE	<u>pH</u>	PREDICTION	<u>95% P.I.</u>	
Pigeon	0	4.72	0.306	4.71	0.297	(0 , 2.039)	
				5.13	0.722	(0 , 3.034)	
				5.13	1.191	(0 , 4.416)	
							1
Little	4	5.66	5.860	5.23	4.421	(1.248 , 12.073)	
Simon				5.73	6.098	(1.947 , 16.093)	
			•	5.73	6.600	(2.044 , 17.957)	
Rose	6	5.08	2.337	4.98	2.127	(0.306 , 6.489)	
	•			5.50	3.266	(0.781 , 9.217)	
				5.50	2.244	(0.312 , 7.021)	
Negro	8	5.98	4.439	5.83	4.095	(1.159, 11.021)	
•				6.31	5.160	(1.610 , 13.538)	
				6.31	6.963	(2.210 , 18.757)	
	•						
				•			
		-					



Figure 1: Smoothed three-dimensional contour plot of the 1166 residuals from the regression equation without a spatial component (2.2) versus position of the lakes corresponding to the residuals. The view is from southeast (bottom right) to northwest (top left).



Figure 2: Sample and fitted variograms of residuals from the final iteration. The solid line represents the sample variogram.



Figure 3: Smoothed three-dimensional contour plot of the 583 residuals from the regression equation with kriging estimate of covariance versus position of the lakes corresponding to the residuals. The view is from southeast (bottom right) to northwest (top left).



Figure 4: Directional exponential variograms. The solid line corresponds to class 1. The large dashes and the small dashes correspond to classes 2 and 4, respectively.



Figure 5:Smoothed three-dimensional contour plot of \hat{g} , the spatial smooth,from regression equation (4.5) versus position of the lakescorresponding to each value in \hat{g} . The view is from southeast(bottom right) to northwest (top left).

•



• •

- !





Difference

Initial calculated pH

Figure 7: Boxplots of the differences between predicted species richness for a 50% sulfate reduction and predicted species richness for a 0% sulfate reduction for each of the 1166 lakes. The data are grouped by predicted pH for a 0% sulfate reduction. (These are the pH's in the "no change" scenario of Table 2).