INTERVAL ESTIMATES FOR DIATOM INFERRED LAKE pH HISTORIES

GARY W. OEHLERT

UNIVERSITY OF MINNESOTA
SCHOOL OF STATISTICS
TECHNICAL REPORT #469

APRIL 1986

UNIVERSITY OF MINNESOTA
SCHOOL OF STATISTICS
DEPARTMENT OF APPLIED STATISTICS
ST. PAUL, MINNESOTA 55108

Interval Estimates for Diatom Inferred Lake pH Histories

Gary W. Oehlert

Technical Report No. 469

Department of Applied Statistics

University of Minnesota

ABSTRACT

We investigate ordinary least squares and Bayesian methods for constructing interval estimates for historical lake pH's inferred from diatom sediments. The Bayesian method explicitly models several forms of variability, including the sampling and classification variability of the diatom records, estimation variability, and measurement error in observed pH's. The two methods produce similar interval estimates, but the Bayesian model allows design recommendations to be made.

April 23, 1986

Interval Estimates for Diatom Inferred Lake pH Histories

Gary W. Oehlert

Technical Report No. 469

Department of Applied Statistics

University of Minnesota

Introduction

Diatoms are small plants with siliceous cell walls that inhabit most waters. The variety of diatoms is large, and different diatom taxa have evolved to occupy different ecological niches. In particular, many diatom taxa are pH sensitive and prefer, or are most competitive, in a certain pH range. Since this aspect of the ecology of diatoms is fairly well known, it is possible to estimate from a given collection of diatoms the pH of the water from which they were taken. This estimation is most useful for reconstruction of a lake's historical pH record. Because the cell walls of diatoms are siliceous, they are preserved in the sediments that accumulate on the bottom of the lake. If these sediments are undisturbed, then each layer of the sediments will contain diatom remains from a specific time period, and these remains can be used to estimate the pH of the water at that time.

Lake pH histories are of more than scientific interest due to the

national debate over acid deposition and it effects. Current surveys, for example the National Lake Survey sponsored by the U.S. Environmental Protection Agency, can determine the number and distribution of lakes which are currently acidic, but they cannot determine whether those lakes have become acidic recently (say within the past 30 to 50 years) or been acidic for hundreds of years. Historical data on lake pH's are sparse and frequently of poor quality (NRC 1986). Thus the need for inferred pH histories is great.

The pH reconstruction process typically uses two diatom data sets. The first is a calibration set used to determine the pH prediction procedure, and the second is the historical set from a lake where we wish to estimate the pH history. A common protocol would be to sample the present diatom flora at lakes in the area of the target lake and to make simultaneous pH measurements on these lakes. Next, the diatoms are subsampled, each diatom in the subsample is classified taxonomically, and each taxonomic group is classified into one of five pH preference categories (acidobiontic - optimum below pH 5.5, acidophilic - usually below pH 7.0, indifferent, alkaliphilic - usually above pH 7.0, and alkalibiontic - occurs only above pH 7.0). The pH category of each taxon can usually be found in the literature, but it may also be determined from the taxon's distribution in the calibration set. Four to five hundred diatoms comprise a typical subsample. The data have now been reduced to a pH and k, a five-vector of proportions for each lake.

Three techniques are common for estimating pH from the diatom proportions. The first two involve indices computed from the five-vector of proportions and the third is multiple regression. Nygaard (1956) proposed an index alpha defined by

$$\alpha = \frac{5\,k_1 + k_2}{k_4 + 5\,k_5},$$

where k_1 is the proportion of diatoms that are acidobiontic and so on. Renberg and Hellberg (1982) defined a new index called B:

$$B = \frac{k_3 + 5\,k_2 + 40\,k_1}{k_3 + 3.5\,k_4 + 108\,k_5} \quad .$$

The pH prediction procedure assumes that pH is a linear function of log(α) or log(B). The coefficients of the linear relationship are determined by least squares.

Charles (1985a) uses these three techniques on 37 Adirondack lakes with surface pH's ranging from 4.5 to 7.8. One drawback of the index based techniques is that they are undefined for acid lakes with no diatoms in the more alkaline categories. For these lakes, Charles sets the denominator of the index ratios to 0.01. The coefficients of determination for the three regressions are log(α) 0.89, log(B) 0.91, and multiple regression 0.94, and residual standard errors range from 0.28 to 0.38 pH units. Thus all three techniques are producing good estimates of pH.

In this paper we compare two interval estimates of pH based on the multiple regression model. The first is the standard prediction interval for multiple regression. The second is a Bayesian posterior prediction interval which explicitly models the sources of variability in the pH prediction process. Both procedures yield reasonable prediction intervals. However, the Bayesian procedure allows us to see how the length of the prediction intervals varies as a function of model parameters. This allows statisticians to advise diatomists as to which aspects of the pH reconstruction process are introducing the most variability, so that extra effort can be put to bear in the most useful areas.

Prediction Models

We will use the following notation. Let y be an observed pH measurement and let σ be the standard error of this measurement. Suppose that there are M taxa of diatoms, that in a given sample the true proportions of the M taxa are $\mathbf{p} = (p_1, p_2, \ldots p_M)'$, that the subsampled counts are $\mathbf{n} = (n_1, n_2, \ldots n_M)'$, and that the total count is N. Let **C** be the 5 by M classification matrix of zeros and ones assigning each diatom taxon to one of the pH preference categories, and let $\mathbf{k} = \mathbf{Cn}/N$ be the 5-vector of proportions in the pH preference categories. We observe neither **n** nor C, because there may have been taxonomic classification errors or misassignments of taxa to pH categories. Instead, we observe $\mathbf{n}^* = \mathbf{Tn}$, where **T** is an unobserved MxM matrix of ones and zeros classifying each observed taxon, C^*, the 5 x M pH classification matrix reported by the investigator, and $\mathbf{k}^* = \mathbf{C}^*\mathbf{n}^*/N$. We assume that **T** is invertible. There are L lakes in the calibration set; **K** is the L x 5 matrix whose rows are the **k** proportions from the L lakes, $\mathbf{K}^*$ is the observed version of **K**, and **y** is the vector of length L of observed lake pH's. Let **z** (5 by 1) be a vector of coefficients, and let $\mathbf{p}^* = \mathbf{Tp}$ be the permuted version of **p**.

The regression approach

The regression approach is a standard least squares prediction interval. We assume that a pH measurement in a lake with diatoms distributed as $\mathbf{k}^*$ has expected value $\mathbf{k}^{*\prime}\mathbf{z}$, with a normally distributed error independent of all other lakes. The least squares estimate of **z** is

$$\hat{\mathbf{z}} = (\mathbf{K}^{*\prime}\mathbf{K}^*)^{-1}\mathbf{K}^{*\prime}\mathbf{y} \quad .$$

To predict the pH associated with diatoms $\mathbf{k}^*_{L+1}$ we form an interval with center $\mathbf{k}^{*\prime}_{L+1}\hat{\mathbf{z}}$ and width

$$t(1 - \alpha/2, L-5)s(1 + k^*_{L+1}{}'(K^{*\prime}K^*)^{-1}k^*_{L+1})^{1/2} ,$$

where $t(1 - \alpha/2, L-5)$ is the $1 - \alpha/2$ point of a t distribution with L-5 degrees of freedom and s is the root mean square error of the regression. See Weisberg (1985) p. 229.

It is worth noting that even though we are in an "errors in variables" situation (the carriers of our regression are measured with error), we do not need to adopt a structural or functional approach. This is because we are interested in predicting unobserved pH measurements rather than estimating model coefficients. See Madansky (1959).

There is a great temptation to compute a confidence interval for the mean of a lake's pH measurements rather than a prediction interval for an unobserved measurement, because the confidence interval will be much shorter than the prediction interval. This temptation must be resisted, since it is based on the premise that all variability about the regression line is "measurement" error that would average out over many measurements. In fact, true pH measurement error has a variance of about 0.03, much less than the error mean square observed in the example below, and the remaining components of the error mean square need not average out with more measurements. Thus it is safer to stay with prediction rather than confidence intervals.

The Bayesian approach

The Bayesian approach computes the posterior prediction interval for an unobserved pH value given the observed pH and $\mathbf{n}^*$ values at the calibration lakes and the observed pH classification matrix $\mathbf{C}^*$. This interval is based on the posterior predictive distribution:

$$f(y_{L+1} | y_1, n^*_1, y_2, n^*_2, \ldots, n^*_L, n^*_{L+1}, C^*).$$

(We will always use the symbol f to denote a density or probability function, and we will rely on context and the arguments to determine for which random variables f is a density, on which random variables they are conditioned and so on.) In the Bayesian model, $\mathbf{z}$, $\mathbf{p}$, $\mathbf{C}$, $\mathbf{T}$, and σ are unobserved random variables with prior distributions, and $\mathbf{n}^*$ and y (which we observe at each of the lakes in the calibration set) and $\mathbf{C}^*$ have likelihoods conditional on the unobserved parameters. We must specify the prior distributions and likelihoods to compute the posterior via Bayes Theorem.

We make the following general assumptions. First, $\mathbf{T}$, $\mathbf{C}$, $\mathbf{z}$, σ, and $\mathbf{p}_i$, i=1,2, ... , L+1 are <u>a priori</u> independent. Second, y_i and $\mathbf{n}_i^*$ are independent of each other and of y_j and $\mathbf{n}_j^*$ given $\mathbf{T}$, $\mathbf{C}$, $\mathbf{z}$, σ, $\mathbf{p}_i^*$, and $\mathbf{p}_i^*$. Third, $\mathbf{C}^*$ is independent of $\mathbf{n}_i^*$, y_i, $\mathbf{p}_i$, $\mathbf{T}$, $\mathbf{z}$, and σ. Under these assumptions, we may show that the posterior predictive distribution is proportional to

$$E_{\mathbf{p}_i^*\text{'s},\ \mathbf{C},\ \mathbf{T}\ \mid\ \mathbf{n}_i^*\text{'s},\ \mathbf{C}^*} \int_0^{\infty}\int_{-\infty}^{\infty} \prod_{i=1}^{L+1} f(y_i \mid \mathbf{C}, \mathbf{z}, \sigma^2, \mathbf{p}_i^*, \mathbf{T})\, f(\mathbf{z})\, f(\sigma^2)\ d\mathbf{z}d\sigma^2,$$

where the outer expectation is with respect to the conditional distribution of $\mathbf{C}$, $\mathbf{T}$, and the $\mathbf{p}_i^*$'s.

We assume that the distribution of a pH measurement y_i given $\mathbf{C}$, $\mathbf{z}$, $\mathbf{p}_i$, and σ, is normal with mean $\mathbf{z}'\mathbf{C}\mathbf{p}_i$ and standard deviation σ. Equivalently, the mean may be written $\mathbf{z}'\mathbf{C}\mathbf{T}^{-1}\mathbf{p}_i^*$. We will use the same classification matrices $\mathbf{C}$ and $\mathbf{T}$ and standard error σ at all lakes. This assumption is reasonable if one investigator or team is responsible for all data collection, though it may be suspect if the data are merged from several sources. We will use a multivariate normal prior for $\mathbf{z}$ with mean μ_z and variance $\sum_z$. This should be an informative prior, since we know roughly what the coefficients in the regression should be. The normal shape is chosen mostly for convenience, because it

allows us to compute the inner integral in closed form.

The conditional distribution of $\mathbf{n}$ given $\mathbf{p}$ is multinomial with mean $N\mathbf{p}$. Thus, the conditional distribution of $\mathbf{n}^*$ given $\mathbf{p}^*$ is multinomial with mean $N\mathbf{p}^*$. Our prior for $\mathbf{p}$ will be Dirichlet, with a constant shape parameter γ. Ordinarily, γ will be small reflecting limited prior knowledge of the diatom proportions. The p_i and p_j are a posteriori independent given T and the n_k^*'s, so their joint posterior distribution is a product of Dirichlets with parameters of the form $T^{-1}n_i^* + \gamma$. Thus, p_i^* has a Dirichlet posterior distribution with parameter $n_i^* + \gamma$, and is independent of the other p_j^*'s.

The matrices C and T are more difficult to model, but there is some simplification since they only enter through the product CT^{-1}. This simplification arises because we only need to know the taxonomic classification (T) up to pH category.

Common or physically large diatoms are more likely to be correctly classified at the first step (T) than are small or rare diatoms. We shall model only the frequency dependence and not the size dependence, and assume that each diatom is classified independently of the others. Suppose that j identical diatoms (from L+1 lakes) out of a total sample of size I (from the same L+1 lakes) must be classified. We assume that the diatom is classified into a taxon of the correct pH preference group with probability $\rho_l + (\rho_h - \rho_l)*\min(j/I/\varepsilon, 1.0)$; if misclassified, the diatom goes into some taxon chosen so that the pH preference category is uniform over the four incorrect categories. We assume a uniform prior for the pH categories, so that the probability that the diatom is truly from a taxon in pH category i given that the investigator classified the diatom in a taxon from pH category j is equal to the probability that the diatom will be classified into a taxon

from pH category j given that it is from a taxon in pH category i. This specifies the posterior distribution of T given the n_i^*'s.

At the second stage, the taxonomic groups must be assigned to pH categories. Here we will assume that a taxonomic group is correctly categorized with probability β, and miscategorized into one of the two adjacent categories with probability $1-\beta$, independently of the other taxonomic groups. (For extreme pH categories, the probability of correct classification will be taken to be $1-(1-\beta)/2$.)

The variance of lake surface water pH's measurements is about 0.03, but in some lakes there can be additional variability due to a nonuniform sedimentation rate across the lake bottom (Charles 1985b). Our model does not explicitly include this source of variation, but we can allow for it by putting prior probability on larger σ's. With this in mind, we take a uniform prior from 0.03 to 0.07 for σ^2.

Computation of the posterior predictive distribution is done with a combination of analytic, numerical, and Monte Carlo methods. The integral over z may be done in closed form to obtain that the posterior predictive is proportional to

$$E_{p_1^*, \ldots, p_{L+1}^*, C, T \mid n_1^*, \ldots, n_{L+1}^*, C^*} \int_{0.03}^{0.07} \sigma^{-L-1} \left| \frac{K'K}{\sigma^2} + \sum_z^{-1} \right|^{-1/2} \times$$

$$\exp \frac{1}{2} \left| \left(\frac{K'y}{\sigma^2} + \sum_z^{-1}\mu_z\right)' \left(\frac{K'K}{\sigma^2} + \sum_z^{-1}\right)^{-1} \left(\frac{K'y}{\sigma^2} + \sum_z^{-1}\mu_z\right) - \frac{y'y}{\sigma^2} \right| d\sigma^2 .$$

Next numerically integrate with respect to σ^2, and finally do a Monte Carlo computation of the expectation with respect to C, T, and the p^*'s. (To do the

numerical integration over σ^2, we use 24 point Gauss quadrature.)

Example

We illustrate the use of the methods described here on the data used in Charles (1985a). This data consists of a calibration set of 37 lakes, each with pH measurements (y values) and observed diatom counts $\mathbf{n}^*$ from surface sediments; Big Moose Lake, for which there are diatom counts ($\mathbf{n}^*$) going down through about 30cm of sediment; and a pH category matrix $\mathbf{C}^*$ (Don Charles, personal communication). We have taxonomic data on many layers of sediment from Big Moose Lake, but we will only show the results for the surface sediment (depth 0.0-0.5 cm) representing 1979 conditions. There are 270 different diatom taxa present in these lake sediments.

One problem in this data is that there is no pH tolerance information for some of the taxa. We have chosen to impute a pH tolerance category for these taxa, rather than delete the taxa all together. To impute the tolerance, we take the mean tolerance category of all taxa of the same genus for which we have pH tolerance information. Some genera, e.g. Cyclotella, Eunotia, Frustulia, and Tabellaria, are strongly clustered in one tolerance category, most have some species in each of two or three tolerance categories, and only a few, e.g. Cymbella, Fragilaria, and Navicula have species in four tolerance categories. To account for the fact that these categories are imputed, we use a β value of one third for the imputed tolerance categories rather than the larger β value used for cases where the value is not imputed.

We begin using the least squares prediction technique. There were no diatoms found in the most alkaline pH category, so the regression is done with the first 4 categories only. Figure 1 shows a plot of observed pH versus

predicted pH. A 95% prediction interval for a 1979 pH measurement in Big Moose Lake is 4.70 ± 0.75, based on the surface sediments. Standard regression diagnostics do not indicate any problems with assumptions: a normal probability plot of the residuals is acceptably linear and the largest Cook's distance is less than 0.3.

To use the Bayes technique, we must completely specify the priors. Our prior for **z** is multivariate normal with mean (4, 5, 6, 7, 8)' and variance 3.33 times the identity matrix. This is a broad prior centered at a reasonable guess for the regression parameters. The prior for **p** is Dirichlet with common parameter γ. We will take γ to be 0.01.

We also need to specify the ρ_l, ρ_h, ε, and β parameters in the distribution of the classification matrix. In our example, we set ρ_l equal to 0.6, ρ_h equal to 0.97, ε equal to 0.006, and β equal to 0.9. The ρ_l, ρ_h, and β values were chosen to be somewhat larger than lower bounds estimated by a diatom expert (Charles, personal communication). The ε value corresponds to approximately 96 individuals present for maximal probability of correct taxonomic classification. The average number of taxonomically misclassified diatoms is 1317 (out of approximately 17000 diatoms counted).

The calculation of the posterior distribution of pH involves an integration which we do by Monte Carlo. In our example, we use 4000 samples, but even 4000 is probably too few for truly stable estimation. This is because the likelihood varies greatly from sample to sample, and the final average tends to be dominated by just a few of 4000 samples. Some form of importance sampling would improve the efficiency here, but I have not implemented it.

Using the prior parameters given above, the posterior predictive distri-

bution for pH is approximately normal with a mean of 4.83 and a variance of 0.126. A central 95% posterior interval is 4.16 to 5.55, showing a slight asymmetry. This interval is slightly narrower and about 0.1 higher than the least squares prediction interval. Figure 2 compares the predictive distributions for Big Moose Lake pH computed via least squares and Bayesian techniques.

Design considerations

The Bayesian model is conceptually difficult and expensive to implement, but it does have the advantage of having explicit, meaningful parameters that describe various aspects of the pH reconstruction system. We may investigate changes in the properties of system observables as functions of the parameters by simulating the system with different parameters. Thus, to the extent that our model approximates the way the diatom identification is performed, we can indicate to diatomists which changes in parameters will most improve system performance. The prototypical question is whether to count more diatoms under the current system, or to count the same or fewer number of diatoms but put more effort into correctly classifying them.

One measure of system performance is the residual mean square in the least squares regression of pH on the abundance of diatoms in the five tolerance categories. This mean square is the controlling factor in how wide our interval estimates of pH will be, so it is sensible to see how this varies as parameters change.

We simulate the system in the following way. Use the 37 calibration lakes in Charles' Adirondack study, and choose **n** for each lake to be multinomial **p**, where **p** is chosen from the posterior distribution for **p** used in our

Bayesian method above. Use Charles (1985a) linear regression coefficients from the Adirondack study to compute the true lake pH from $\mathbf{p}$. Next, assign the taxa to pH categories based on the parameters and the assumption that the true pH categories of the taxa are the observed categories. Finally, compute the observed proportions of diatoms in the different tolerance categories for each lake, and compute the residual mean square for the regression. Here, we will repeat this process 100 times to get an estimate of the expected residual mean square for given parameter values. Note that this residual mean square does not include a component for σ^2.

We study the effects of ρ_1, ρ_h, ε, β, and N by looking at a quarter fraction of a 2^5 factorial design. Design points and observed geometric mean MSE's are given in Table 1. (The distributions of the MSE's are approximately log normal. The geometric means have a relative standard error of about eight percent.) The observed mean square error from Charles' data was 0.107. This includes a component from σ^2 which we expect to be about 0.05, so the quantity corresponding to the simulations should be about 0.06. This value of 0.06 is well within the distributions of all eight parameter settings.

Analysis of the results of this experiment leads to the conclusion that residual mean square error depends primarily on β, ε, and ρ_h, and only slightly on ρ_1 and N. Two lines of thought explain why this should be so. First, most of the error is the result of misclassification, not multinomial sampling, so N should have little effect on MSE. Second, variability in the proportions lying in each tolerance category is driven mostly by errors in abundant taxa, and these errors are controlled by β and ρ_h, and to a lesser extent ε. Quantitatively, doubling the sample size from 450 to 900 reduces mean square error an amount approximately equal to that obtained when changing

β from 0.9 to 0.915, ε from 0.002 to 0.0036, or ρ_h from 0.95 to .965.

The following design principle seems to be established. To the extent that it is possible, more effort should be spent to correctly classify the abundant taxa, both taxonomic and pH tolerance classification. Small increases in classification accuracy can offset the effect of a decreased sample size, so increased classification accuracy is usually desirable even if fewer diatoms must be counted to maintain a constant level of effort.

Recent developments

One criticism of the Bayesian model used here is that it treats all diatoms uniformly. Certainly some taxa are more distinctive and easier to identify than others, but this fact is ignored. If there were data indicating which taxa were consistently identified correctly, and listing sets of taxa that tend to be confused, the model for T given above could be extended to include this information and presumably give better posterior distributions for pH. Fortunately, such data are forthcoming.

The PIRLA project (Paleoecological Investigation of Recent Lake Acidification, Charles and Whitehead 1985) is a multidisciplinary study of lake acidification. Part of this project includes diatom reconstructions of historical lake pH's conducted by several research groups. To improve the quality of the data obtained, the diatomists have instituted a "truth-in-counting" system describing the way each diatomist classifies diatoms. In this system, each diatomist gives each taxon a code. The codes are of the form (1) the name used for the taxon, (2) a four point rating giving the degree of confidence in the classification, and (3) a list of taxa which might be confused with the current taxon. The four point rating scale for confidence is subjective and

ranges from "I believe I use this taxon consistently ... and that other PIRLA investigators would agree with me" to "there are almost certainly inconsistencies in the identification of this taxon in my data set".

Data such as this will allow the Bayesian model to include the probability of each individual taxon being misclassified, and if misclassified, know into which taxa it is most likely to go. This could provide a great improvement over the ad hoc model of misclassification currently in use.

Summary

Historical lake pH measurements are rare, and when available are often of poor quality. This means that inference about trends in lake acidity must usually be made with indirect methods such as diatom pH reconstructions. As in all scientific work, an interval estimate is preferred over a point estimate.

Ordinary least squares regression provides a sensible prediction interval for the unobserved pH measurements, but care must be taken when deriving confidence intervals for mean lake pH to prevent an overstatement of accuracy. Bayesian posterior prediction intervals can agree closely with the least squares intervals, but their computational cost makes them less attractive.

The major advantage of the Bayesian approach is the explicit modeling of the pH reconstruction process. To the extent that our modeling is accurate, it allows us to simulate the pH reconstruction process and make recommendations about ways to most effectively improve the precision of the technique. The current model implies that accurate classification of the most abundant diatoms is paramount, even if this implies that fewer diatoms will be counted.

Finally, better and more extensive data are on the horizon, so that the modeling done in the Bayesian technique will be a more accurate reflection of reality.

Acknowledgements

I would like to thank Don Charles for introducing me to the problem of diatom pH reconstruction, answering many, many questions, and providing me with the raw data for the Adirondack lakes. Thanks also to Rebecca Bormann for carefully entering the Adirondack lake data into the computer.

REFERENCES

Charles, Donald F. (1985a), "Relationships between surface sediment diatom assemblages and lakewater characteristics in Adirondack lakes," Ecology, 66, 994-1011.

Charles, D. F. (1985b), "Spatial variability of diatom assemblages in Big Moose Lake surface sediments," Eight North American Diatom Symposium, Murray, Kentucky.

Charles, D. F. and Whitehead, D. R. (1985), "The PIRLA Project: Paleoecological Investigation of Recent Lake Acidification," IVth International Symposium on Paleolimnology, Ossiach, Austria.

Madansky, Albert (1959), "The Fitting of Straight Lines When Both Variables are Subject to Error," Journal of the American Statistical Association, 54, 173-205.

National Research Council (1986), Acid Deposition: Long-Term Trends, National Academy Press, Washington, D. C.

Nygaard, G. (1956), "Ancient and recent flora of diatoms and Chrysophyceae in Lake Grobso," in Studies on the Humic, Acid Lake Bribso, K. Berg and I. C. Peterson (eds.), pp. 32-94, Fol. Limnol. Scand. 8:1-273.

Renberg, I. and Hellberg, T. (1982), "The pH history of lakes in southwestern Sweden, as calculated from the subfossil diatom flora of the sediments," Ambio, 11, 30-33.

Weisberg, S. (1985), Applied Linear Regression, 2nd ed., Wiley, New York.

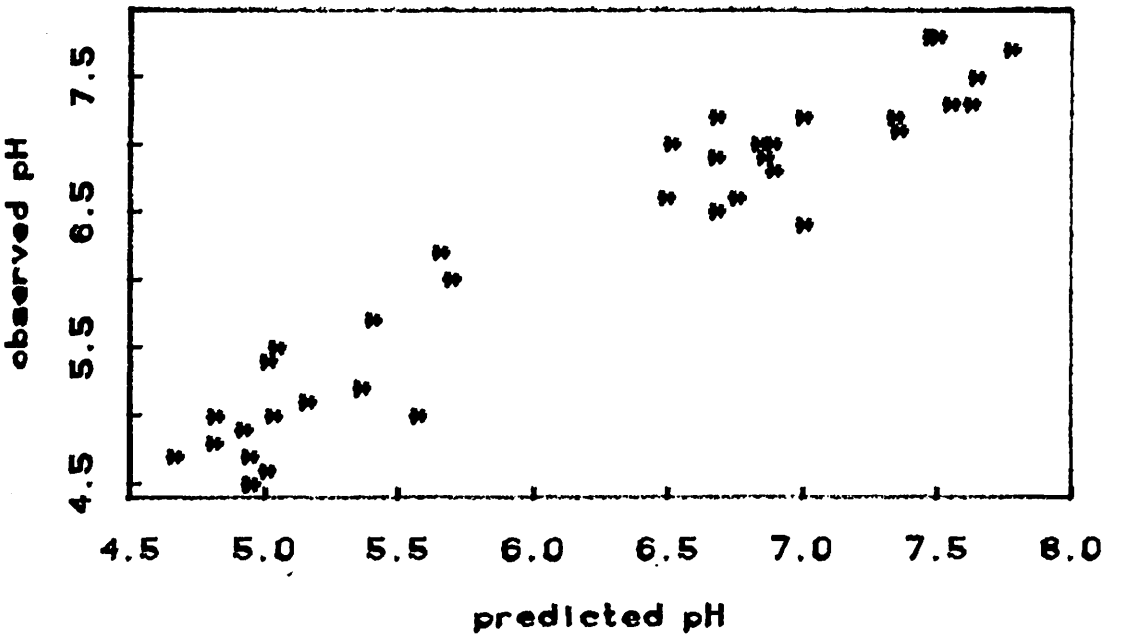
Table 1

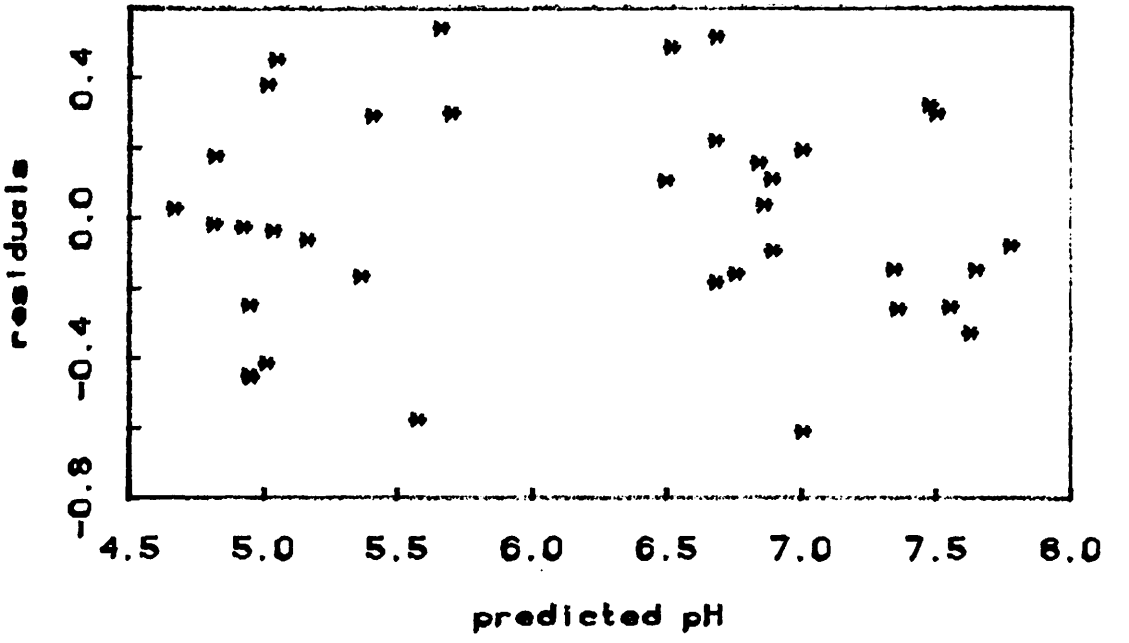
Geometric mean MSE for simulated regressions

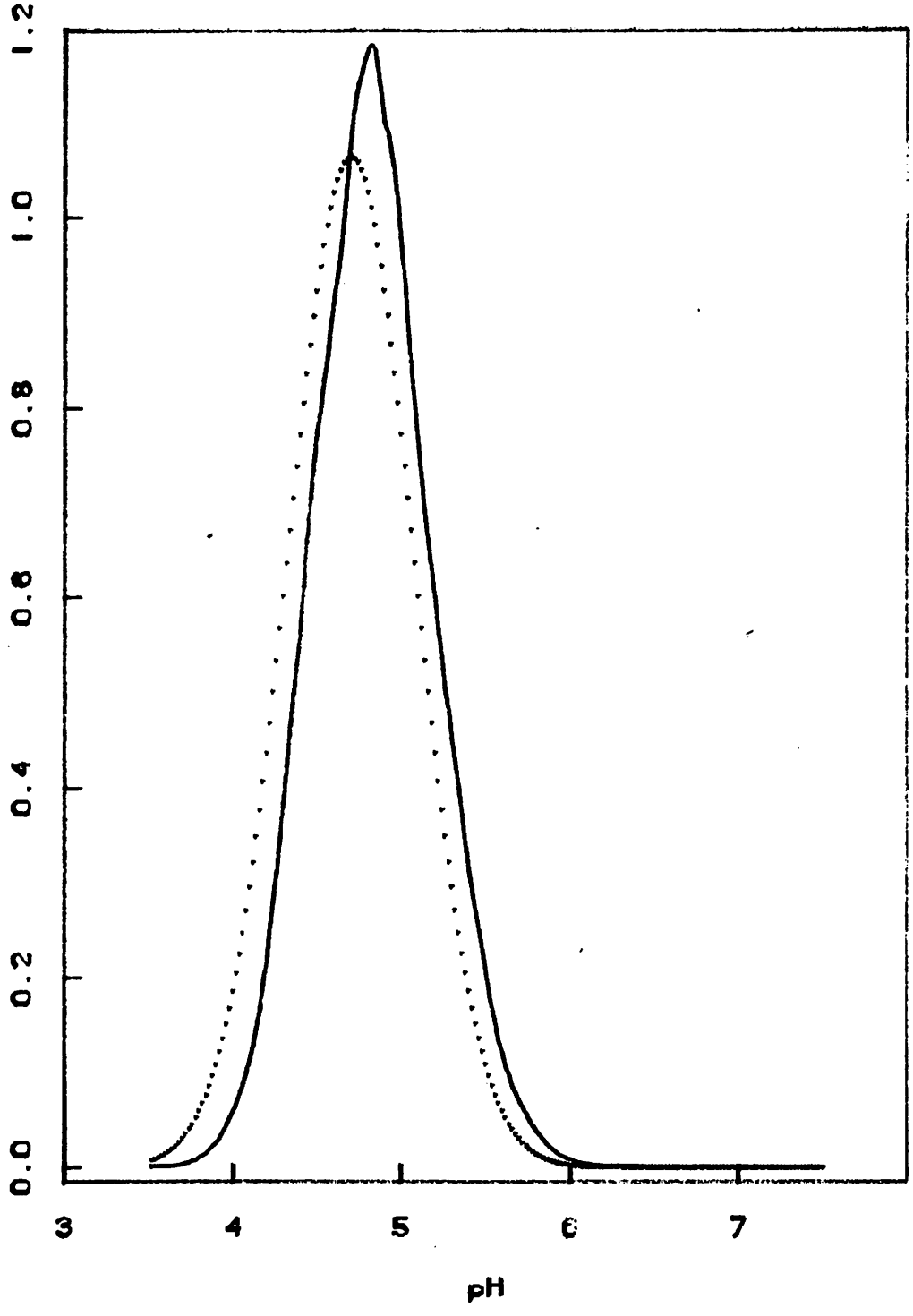
ρ_l	ρ_h	ε	β	N	geometric mean
0.5	0.95	0.006	0.90	450	0.064
0.5	0.95	0.002	0.90	900	0.041
0.5	0.98	0.006	0.95	450	0.033
0.5	0.98	0.002	0.95	900	0.026
0.7	0.95	0.006	0.95	900	0.037
0.7	0.95	0.002	0.95	450	0.031
0.7	0.98	0.006	0.90	900	0.041
0.7	0.98	0.002	0.90	450	0.037

Figure 1. Top: observed versus least squares predicted pH for 37 Adirondack lakes. Bottom: residuals versus least squares predicted pH for 37 Adirondack lakes.

Figure 2. Predictive densities for least squares (dotted) and Bayes (solid) methods.


observed pH
4.5
5.5
6.5
7.5
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
predicted pH


residuals
-0.8
-0.4
0.0
0.4
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
predicted pH


1.2
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
pH