United States
Department of
Agriculture

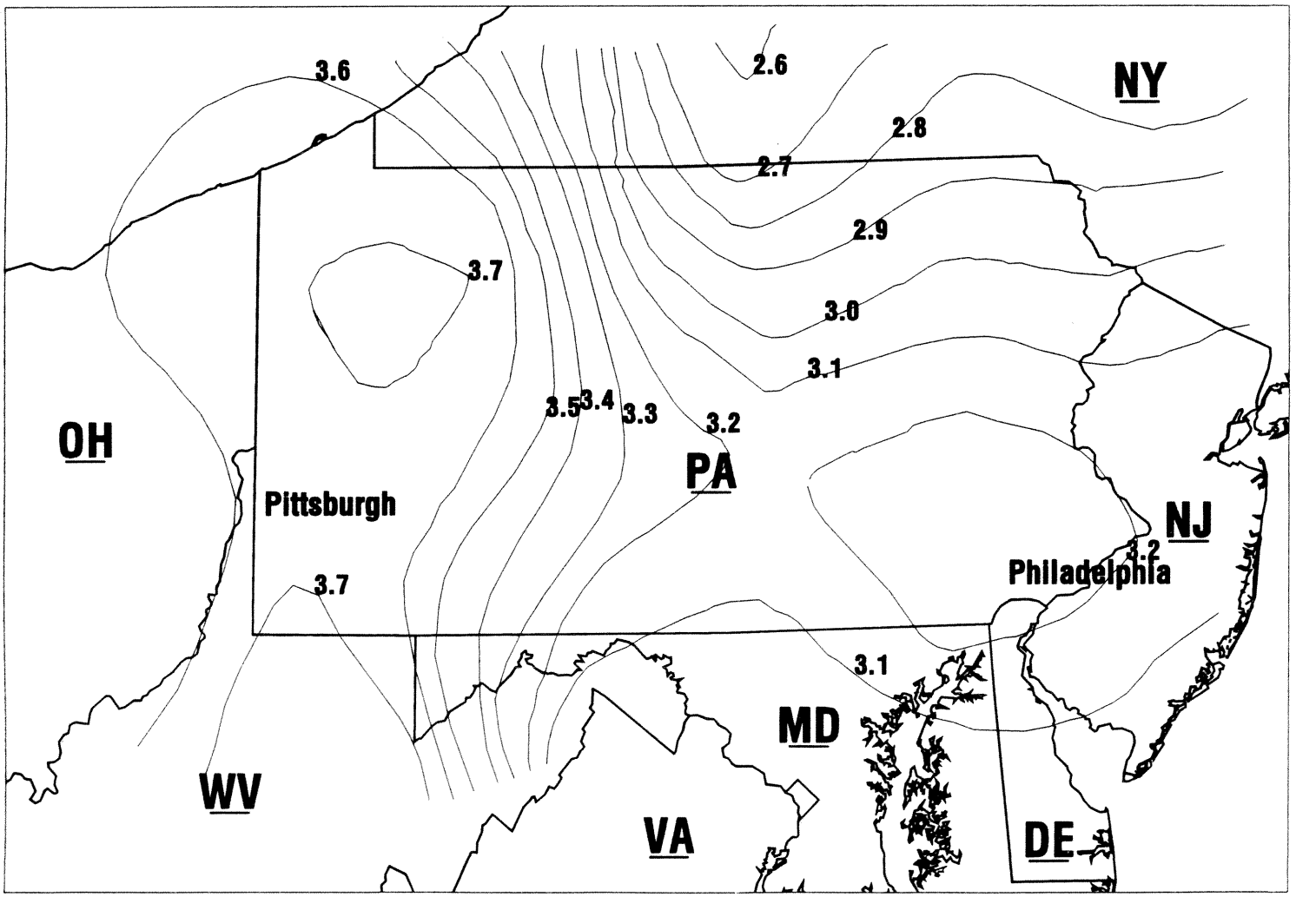
Forest Service

**Southern
Research Station**

Research Paper
SRS-7

Kriging Direct and Indirect Estimates of Sulfate Deposition: A Comparison

**Gregory A. Reams, Manuela M.P. Huso,
Richard J. Vong, and Joseph M. McCollum**



Kriging Direct and Indirect Estimates of Sulfate Deposition: A Comparison

Gregory A. Reams, Manuela M.P. Huso, Richard J. Vong, and Joseph M. McCollum

Abstract

Due to logistical and cost constraints, acidic deposition is rarely measured at forest research or sampling locations. A crucial first step to assessing the effects of acid rain on forests is an accurate estimate of acidic deposition at forest sample sites. We examine two methods (direct and indirect) for estimating sulfate deposition at atmospherically unmonitored forest sites. The direct method only uses directly measured deposition data, while the indirect method additionally incorporates precipitation measurements from a spatially denser network of monitoring sites. Sulfate deposition values were estimated by point kriging using both the direct and indirect methods. By using the supplemental data from the precipitation monitoring network, estimates of sulfate deposition improved substantially, particularly at sites that are relatively isolated to the acid deposition monitoring network. Cross-validation procedures indicate that by using the indirect method, a reduction of approximately 20 to 25 percent in the predicted error sum of squares occurred.

Keywords: Acid deposition, geostatistics, interpolation, monitoring, variograms.

Introduction

In many studies that involve sampling within a given area, sample independence fundamental to much of classic statistical analysis cannot be assumed because the observed variables are spatially correlated. Regionalized variable theory and geostatistical analysis (Matheron 1971) provide a way to use spatial correlation to derive unbiased estimates of the observed variable at unmeasured points. The techniques involved have been applied in a wide range of disciplines, from toxic waste management to population biology, and the unifying theme was spatially correlated variables.

Acidic deposition is rarely measured at forest research or sample locations because monitoring remote forest sites is difficult, time-consuming, and expensive. These logistical problems push the costs over budgets. An accurate estimate of acidic deposition at these sites is critical to assessing the effects of acid rain on forests. In this paper, we examine two methods (direct and indirect) for estimating sulfate deposition and develop contour maps of sulfate deposition for Pennsylvania. These estimates can then be used to estimate sulfate deposition at unmonitored forest research sites within the State.

The two methods presented are not the only procedures capable of producing valid deposition estimates. For example, cokriging could be used. Cokriging is most useful when a highly correlated covariable is sampled intensely. However, available software is limited and cokriging is poorly understood by potential users because the statistical notation is complex (Stein and Corsten 1991). We present the indirect method as an alternative to cokriging. The indirect method can be implemented using the univariate kriging procedures more readily understood by potential users, and user-friendly software for implementing univariate kriging is available from numerous sources.

Estimation Procedures

Geostatistical analysis involves two basic steps: (1) modeling the degree of similarity among measured points as a function of their separating distance, and (2) interpolating values among measured points using the knowledge of their spatial correlation in the estimation procedure. The degree of similarity between points is evaluated by using the semi-variance statistic, $\gamma(h)$, which is defined as half the expected squared difference between values a given distance, h, apart:

$$\gamma(h) = \frac{1}{2}E[z(x_i) - z(x_i + h)]^2 = \frac{1}{2N(h)}\sum_{i=1}^{N(h)}[z(x_i) - z(x_i + h)]^2 , \quad (1)$$

where

$z(x_i)$ = measured sample value at point x_i (x_i can be multidimensional),
$z(x_i + h)$ = sample value at a point a distance of h from x_i, and
$N(h)$ = total number of pairs of points within an interval h of each other.

The plot of $\gamma(h)$ against h is the semi-variogram, and a model of the semi-variance is fit to this set of points.

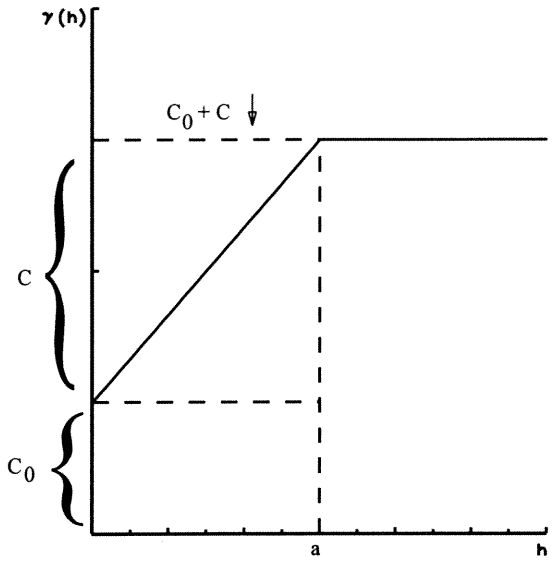


Figure 1—Theoretical semi-variogram defined as a linear model, illustrating (a) the range, (c_o) the nugget variance, and (c_1) the sill. The sill $c_1=c_0+c$, c is the slope term of the linear model.

A theoretical semi-variogram (fig. 1) has three parameters, the nugget (c_0), sill (c_1), and range (a). In most cases, $\gamma(h)$ increases with increasing h up to some maximum; this maximum is approximately the variance of the variable of interest. The distance at which the semi-variance levels off is the range (a), the separation distance at which sample values appear to be spatially uncorrelated. Although $\gamma(0) = 0$ by definition, some finite positive value occurs for $\gamma(h)$, as h approaches 0. This value is the nugget or intercept (c_0) and embraces the variance that occurs at distances smaller than the sampling intervals. The height above the nugget at which the semi-variance is considered constant is the sill (c_1).

Several variogram models, such as the spherical, Gaussian, and exponential, guarantee a positive definite covariance matrix for any set of lags (Journel and Huijbregts 1978). We selected a spherical model based on its previous success in fitting variogram models to deposition data (Haas 1990). This model was used for all semi-variograms in this analysis:

$$\gamma(h) = c_0 + c_1\left[\frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3\right] \quad ,0<h\leq a$$
$$\gamma(h) = c_0 + c_1 \quad ,h>a \ . \qquad (2)$$

Modeling the semi-variogram is a critical step towards interpolation by kriging—an algorithm for determining the weights to be assigned to each point in the interpolation procedure. If variable Z has been measured at locations x_1, $x_2,\ldots,x_n$, resulting in a set of sample values $z(x_1)$, $z(x_2),\ldots,$ $z(x_n)$, the estimate of the value of Z at some unmeasured point x_0, is a linear combination or weighted average of all the observed variables:

$$\hat{z}(x_0) = \lambda_1 z(x_1) + \lambda_2 z(x_2) + \ldots + \lambda_n z(x_n) \ , \qquad (3)$$

where

λ_i = coefficients or weights associated with each of the observed values.

In kriging, the weights (λ_i) are chosen so the estimator is unbiased and the error associated with the estimate is less than that of any other linear sum (i.e., minimum variance). An unbiased estimate $\hat{z}(x_0)$ must equal $E[z(x_0)]$, therefore, the weights λ_i must sum to 1. The variance is minimized subject to this constraint (i.e., $\Sigma\lambda_i = 1$) by using a Lagrangian multiplier μ. A solution is found when

$$\sum_{j=1}^{n} \lambda_j \gamma(|x_i - x_j|) + \mu = \gamma(|x_i - x_0|) \ . \qquad (4)$$

The allocation of the weights among points is determined by the semi-variogram model. Higher weights are assigned to samples closer to the point interpolated. Kriging is a minimum variance, unbiased estimator, and an exact interpolator because if the point being estimated has been sampled, the weight (λ_i) for all other points will be set to zero and the kriging algorithm will return the sample value.

In determining which semi-variogram model to use, two factors must be remembered: (1) lag distances h must provide a sufficient number of points to produce reliable estimates of $\gamma(h)$, particularly at small lags (in small data sets this requirement often forces the use of fairly large lag intervals); and (2) because only the few nearest points are generally used for kriging estimates, the semi-variogram needs to be accurate only at the shortest lag distances, and the fit of the model at distances greater than half the maximum separation of points in the data set is usually inconsequential (Burgess and Webster 1980).

Direct and Indirect Methods of Deposition Estimation

The obvious and most common method of estimating acid deposition values at unmonitored sites involves

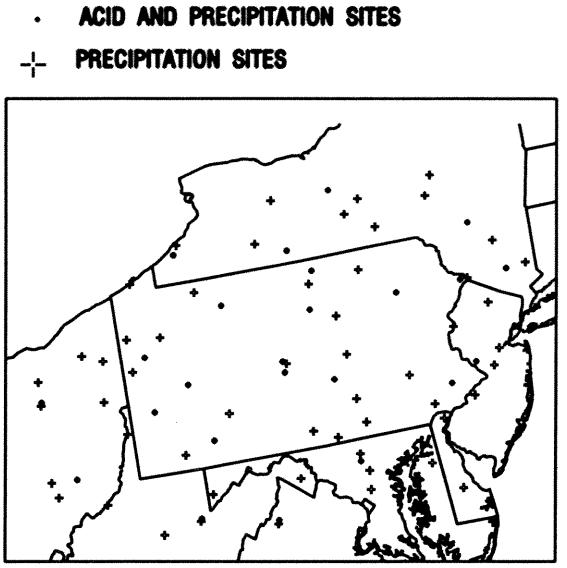


Figure 2—Sample locations of the pollutant monitoring network (24 sites) and the precipitation monitoring network of the National Weather Service (54 locations).

interpolating values between existing deposition monitoring sites. This direct method only uses directly measured deposition data. Several different methods for calculating the weights have been developed including inverse squared distance weighted averaging, Thiessen polygons, least squares polynomials, optimal interpolation, multiquadratic interpolation, and various forms of kriging (Vong and others 1989). Using specific criteria, Tabios and Salas (1985) compared these and other methods and found that kriging and optimal interpolation were better than other techniques, but no technique was uniformly the best.

One of the major constraints to kriging (and many interpolation techniques) is the paucity of acid deposition data in any given area. This scarcity of data makes it difficult to adequately determine which model describes the spatial variability in the data, and subsequently, to accurately estimate deposition at unmonitored sites. Developing a means of estimating deposition that would also use precipitation measurements would be advantageous for two reasons. First, these measurements are more easily obtained than deposition measurements. Second, the sampling network for precipitation measurements is much denser than that for deposition measurements. First proposed by Granat (1988) and discussed by Vong and others (1989), the indirect method uses both measurements. This method is based on the observation that the acid concentration field over a region tends to be much smoother (less variable) than the precipitation field, and that concentration and precipitation are apparently not correlated (Granat 1988). Because deposition is calculated as the product of volume weighted mean concentration (VWMC) of an ion and precipitation, the variability in deposition values over a region should be closely associated with precipitation. The indirect approach, then, is to estimate concentration at sites where only precipitation has been measured, then calculate deposition as the product of the estimated VWMC and the known precipitation. The resulting data set represents a denser sample, facilitating the use of interpolation techniques such as kriging. In this paper we estimate deposition values at 24 National Acid Deposition Program (NADP) monitoring sites in and around Pennsylvania using both methods and compare the results.

Methods

Volume weighted mean concentration of sulfate and mean annual precipitation values averaged over 3 years (1985-87) for 24 sites in and around Pennsylvania were obtained from the NADP data bank and from the Pennsylvania State Environmental Resources Research Institute (Lynch and others 1987). Additional mean annual precipitation values for the same 3 years were obtained from the National Weather Service (NWS) for 54 other sites in the same region (fig. 2). The longitude and latitude coordinates of each point were converted to Alber's conic, equal area projection coordinates (Pearson 1990) to accurately calculate distances between sample points.

Sulfate deposition values were estimated by point kriging using both the direct and indirect methods. The validity of each technique was evaluated by cross-validation. To cross-validate, each sample point from the data set was successively dropped and its value was estimated from the remaining (n-1) points. The predictive ability of the two interpolation techniques can then be evaluated by comparing the estimated with the observed value using the Predicted Error Sum of Squares (PRESS) statistic (Green 1983):

$$PRESS = \sum_{i=1}^{n} [z(x_i) - \hat{z}(x_i)]^2 . \qquad (5)$$

The percent improvement (PI) using the indirect method versus the direct method can be estimated:

$$PI = (1 - \frac{PRESS_{indirect}}{PRESS_{direct}}) \times 100 . \qquad (6)$$

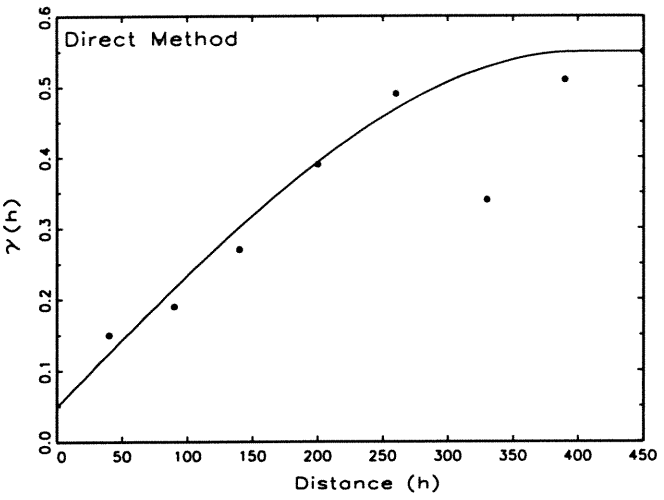


Figure 3—Variogram for the direct method. Model is based on the 24-site pollutant monitoring network.

The direct versus indirect estimation methods were evaluated by the following two criteria: (1) comparison of the mean and variance of the observed and interpolated values, and (2) the sum of squared errors between observed and interpolated values (PRESS statistic).

The semi-variance for sulfate deposition at the 24 monitored sites was calculated and a model for the semi-variogram was determined (fig. 3):

$$\begin{aligned} (h) &= 0.05 + 0.5[\frac{3}{2}(\frac{h}{400}) - \frac{1}{2}(\frac{h}{400})^3],\ 0<h\leq 400 \\ (h) &= 0.55\ . \qquad\qquad h>400 \end{aligned} \tag{7}$$

Choosing this semi-variogram model was based on the five criteria outlined by Myers and others (1982) for semi-variogram selection:

$$(a)\ \frac{1}{n}\sum_{i=1}^{n}[z(x_i) - \hat{z}(x_i)]\ \textit{should be}\ 0,$$

$$(b)\ \frac{1}{n}\sum_{i=1}^{n}[z(x_i) - \hat{z}(x_i)]^2\ \textit{should be small,}$$

$$(c)\ \frac{1}{n}\sum_{i=1}^{n}[\frac{z(x_i) - \hat{z}(x_i)}{\sigma_i}]^2\ \textit{should be}\ 1,$$

(d) sample correlation of $z(x),\hat{z}(x)$ *near* 1, *and*

(e) sample correlation of $\hat{z}(x),[z(x) - \hat{z}(x)] / \sigma_i$ *near* 0 .

Each point was then estimated by ordinary point kriging using the cross-validation procedure described, and the squared difference between the predicted and observed value was calculated.

The sample correlation between concentration and precipitation was calculated and was not found significant ($p = 0.43$, $n = 24$). The semi-variogram model for sulfate concentration was determined using the same five criteria (fig. 4):

$$\begin{aligned} \gamma(h) &= 1.524 + 18.423[\frac{3}{2}(\frac{h}{400}) \\ &\quad - \frac{1}{2}(\frac{h}{400})^3],\ 0<h\leq 400 \\ \gamma(h) &= 19.947\ .\ h>400 \end{aligned} \tag{8}$$

From this, kriged estimates of sulfate concentration at each of the 54 NWS sites were made (Fig. 5). Estimates of sulfate concentration were multiplied by mean annual precipitation values at the 54 NWS sites to estimate deposition. The semi-variogram for sulfate deposition was calculated from the resulting data set of 78 (24 NADP sites and 54 NWS sites) deposition values, with each of the 78 sites treated as an observed value (Fig. 6):

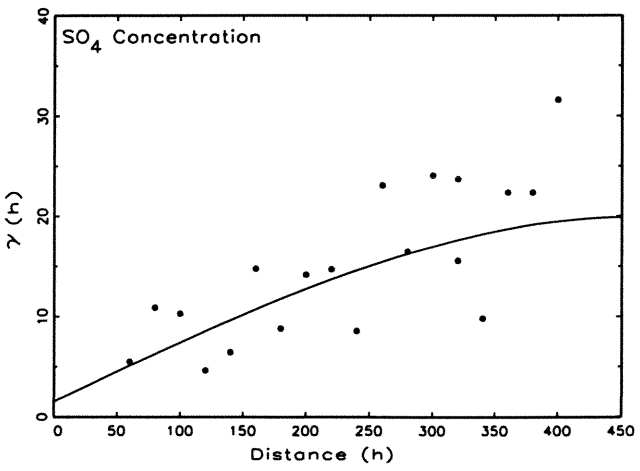


Figure 4—Variogram for sulfate concentration. Model is based on the 24-site pollutant monitoring network.

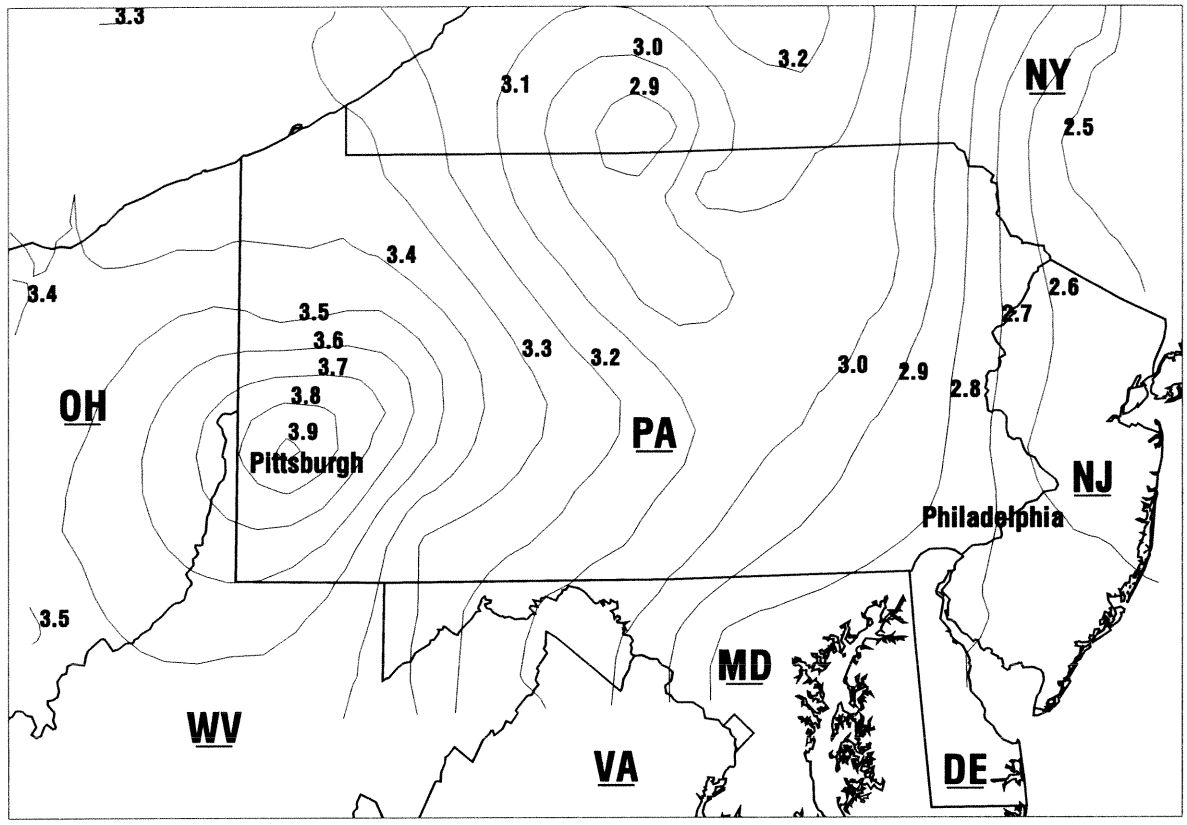


Figure 5—Kriged estimates of sulfate concentration. Response surface is based on the 24-site pollutant monitoring network.

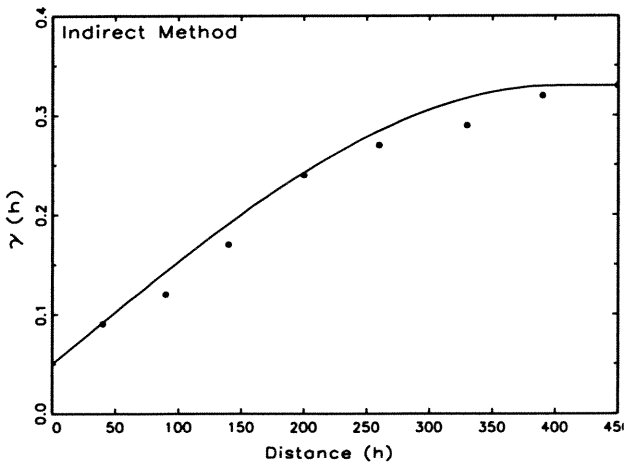


Figure 6—Variogram for the indirect method. Model is based on a network of 78 locations.

$$\gamma(h) = 0.05 + 0.28[\frac{3}{2}(\frac{h}{400}) - \frac{1}{2}(\frac{h}{400})^3], \quad 0<h\leq 400$$
$$\gamma(h) = 0.33 \ . \ h>400 \tag{9}$$

This semi-variogram model was used to point krige deposition estimates for each of the 24 measured sites. The cross-validation procedure and the squared difference between the predicted and observed values were calculated.

Table 1—Comparison of the mean and standard deviation of the observed and kriged estimates of sulfate deposition and the predicted error sum of squares of the kriged estimates using both the direct and indirect approach

Method	Semi-variogram	Observed		Estimated		PRESS	PRESS PA[b]
		Mean	Sd[a]	Mean	Sd[a]		
Direct	Equation (2)	3.251	0.608	3.244	0.422	4.353	1.870
Indirect	Equation (4)	3.251	.608	3.373	.423	3.331	1.426
Percent improvement, equation (1)						23.5	23.7

[a] Sd is abbreviation for standard deviation.
[b] PA is standard abbreviation for Pennsylvania.

Table 2—Comparison of the mean and standard deviation of the observed and kriged estimates of sulfate deposition and the prediction error sum of squares of the kriged estimates using both the direct and indirect approach

Method	Semi-variogram	Observed		Estimated		PRESS	PRESS PA[b]
		Mean	Sd[a]	Mean	Sd[a]		
Direct	Equation (4)	3.251	0.608	3.253	0.408	4.324	1.767
Indirect	Equation (2)	3.251	.608	3.353	.421	3.351	1.477
Percent improvement, equation (1)						23.0	21.0

[a] Sd is abbreviation for standard deviation.
[b] PA is standard abbreviation for Pennsylvania.

Results and Discussion

Table 1 compares the cross-validation results. The mean of the estimates using the direct method more closely approximates the true mean of the data and the standard errors of the estimates from the two different methods are essentially equal. However, the difference in the predicted error sum of squares indicates that while for most observations the two methods are practically equivalent, for a few extreme or unusual cases the indirect method more closely predicts the true deposition value.

The difference between the two methods does not lie simply in the choice of variogram models, because the difference in predicted error sum of squares remains when the semi-variogram model derived from the indirect approach is applied in the direct method and vice versa (table 2). The difference seems to lie in how the indirect method utilizes additional data, shifting the weighting scheme for points neighboring certain extreme or unusual data (precipitation) values. Because the indirect method uses a denser network of deposition values to make estimates, it provides more accuracy, particularly at unique or relatively isolated sites. By using the denser precipitation network, the number of sample points within a certain distance of any given point will either increase or remain unchanged.

If the assumption that sample points close together tend to be more alike than points farther apart holds (i.e., the data are spatially correlated), as the semi-variograms indicate, increasing the number of points at a close range can improve the estimates of deposition.

Contour maps drawn using the two methods show a similar pattern of deposition in the western and central part of the region (fig. 7). Insufficient data are available to predict values for the far northeast corner using the direct method;

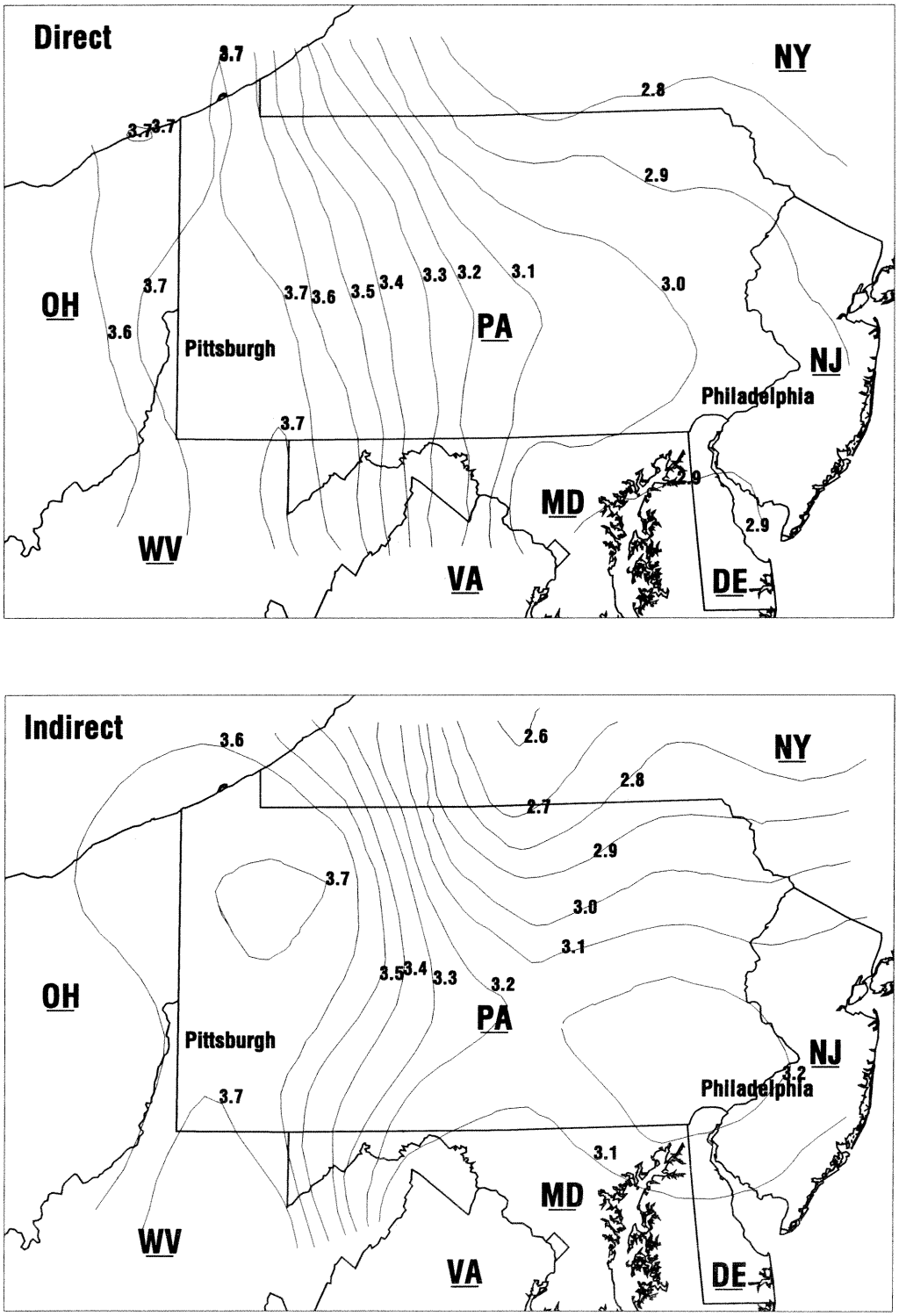


Figure 7—(A) sulfate deposition estimates using the direct method, and (B) sulfate deposition estimates using the indirect method.

therefore, the deposition maps differ radically in the east. Deposition estimates are greater and the deposition field less smooth using the denser network of data available through the indirect method.

Conclusion

The indirect method of estimating acid deposition was first proposed several years ago (Granat 1988) but has never been formally tested with actual data. This study shows that using supplemental data from a precipitation monitoring network to improve estimates of acid deposition can be effective, particularly at sites relatively isolated from the acid deposition monitoring network. Although the data used to supplement the deposition data in the indirect approach are not precise measurements, the increase in sampling density they provide appears to outweigh the uncertainty involved in using them.

Literature Cited

Burgess, T.M.; Webster, R. 1980. Optimal interpolation and isarithmic mapping of soil properties. I: The semi-variogram and punctual kriging. Journal of Soil Science. 31: 315-331.

Granat, L. 1988. A comparative study of precipitation chemistry at inland, coastal, and island sites in the Bothnian Bay area. Stockholm, Sweden: University of Stockholm, Department of Meteorology Report CM-73, ISSN 0280-445X. 54 p.

Green, Edwin J. 1983. Evaluating the predictive abilities of regressions with PRESS. Forest Science. 29: 712-714.

Haas, T.C. 1990. Lognormal and moving window methods of estimating acid deposition. Journal of the American Statistical Association. 85: 950-963.

Journel, A.G.; Huijbregts, C.J. 1978. Mining geostatistics, London: Academic Press. 167-177.

Lynch, J.A.; Corbett, E.S.; Grimm, J.W. 1987. Atmospheric deposition: spatial and temporal variation in Pennsylvania - 1987. Publication No. ER8806A. University Park, PA: Environmental Resources Research Institute. Pennsylvania State University. 5-10.

Matheron, G. 1971. The theory of regionalized variables and its applications (Cahiers du Centre de Morphologie Mathematique, No. 5), Fontainebleau, France: Cetre de Morphologie Mathematique.

Myers, D.E.; Begovitch, C.L.; Butz, T.R.; Kane V.E. 1982. Variogram models for regional groundwater geochemical data. Journal of the International Association for Mathematical Geology. 14: 629-644.

Pearson, Frederick, II. 1990. Map projections: theory and practice. Boca Raton, FL: CRC Press. 372 p.

Stein, A.; Corsten, L.C.A. 1991. Universal kriging and cokriging as a regression procedure. Biometrics. 47: 575-587.

Tabios, G.Q.; Salas, J.D. 1985. A comparative analysis of techniques for spatial interpolation of precipitation. Water Resources Bulletin. 21(3): 365-380.

Vong, R.; Cline, S.; Reams, G. [and others]. 1989. Regional analysis of wet deposition for effects research. EPA/3-89/030. Corvallis, OR: U.S. Environmental Protection Agency, Environmental Research Laboratory. 42 p.

Reams, Gregory A.; Huso, Manuela M.P.; Vong, Richard J.; McCollum, Joseph M. 1997. Kriging Direct and Indirect Estimates of Sulfate Deposition: A Comparison. Res. Pap. SRS-7. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 8 p.

Due to logistical and cost constraints, acidic deposition is rarely measured at forest research or sampling locations. A crucial first step to assessing the effects of acid rain on forests is an accurate estimate of acidic deposition at forest sample sites. We examine two methods (direct and indirect) for estimating sulfate deposition at atmospherically unmonitored forest sites. The direct method only uses directly measured deposition data, while the indirect method additionally incorporates precipitation measurements from a spatially denser network of monitoring sites. Sulfate deposition values were estimated by point kriging using both the direct and indirect methods. By using the supplemental data from the precipitation monitoring network, estimates of sulfate deposition improved substantially, particularly at sites that are relatively isolated to the acid deposition monitoring network. Cross-validation procedures indicate that by using the indirect method, a reduction of approximately 20 to 25 percent in the predicted error sum of squares occurred.

Keywords: Acid deposition, geostatistics, interpolation, monitoring, variograms.