Journal of the Korean Society of Remote Sensing, Vol. 11, No.3, 1995

Spatial Estimation of Point Observed Environmental Variables: A Case Study for Producing Rainfall Acidity Map

Kyu-Sung Lee

Inha University, Department of GeoInformatic Engineering

點觀測 환경 인자의 空間 推定 - 남한 지역의 降雨 酸度 분포도 작성

이 규 성

인하대학교 지리정보공학과

Abstract

The representation of point-observed environmental variables in Geographic Information Systems (GIS) has often been inadequate to meet the need of regional-scale ecological and environmental applications. To create a map of continuous surface that would represent more reliable spatial variations for these applications, I present three spatial estimation methods. Using a secondary variable of the proximity to coast line together with rainfall acidity data collected at the 63 acid rain monitoring stations in Korea, average rainfall acidity map was created using co-kriging. For comparison, two other commonly used interpolation methods (inverse distance weighting and kriging) were also applied to rainfall acidity data without reference to the secondary variable.

These estimation methods were evaluated by both visual assessments of the output maps and the quantitative comparison of error measures that were obtained from cross validation. The co-kriging method produced a rainfall acidity map that showed noticeable improvement in reproducing the inherent spatial pattern as well as provided lower statistical error as compared to the methods using only the primary variable.

Received 13 November 1995 ; Accepted 8 December 1995.

요 약

기후, 토양, 대기, 지질, 지하수 등과 관련된 측정자료는 지역적 규모의 생태 및 환경 목적의 지리정보시스템(GIS)에서 자주 요구되는 공간자료이다. 이와 같은 환경 인자는 자료의 특성상 한정된 지점에서의 점관측자료(point observations)에 의존하여 전체 대상 지역의 지리적 분포를 추정하는 補間法(spatial interpolation)을 이용하여 수치지도의 형태로 변환되고 있으나, 그 추정의 정확도와 관련하여 다른 GIS 공간자료와의 연계분석시 많은 주의가 요망되고 있다. 전국 63개소에서 측정된 강우산도 자료를 이용하여 보다 정확도가 높은 연속면(continuous surface)을 나타내는 디지탈지도를 제작하기 위하여 세 가지 공간추정방법을 적용하였다. 미측정지점에서의 강우산도를 추정하기 위하여 강우산도와 상관관계가 높은 서남해안으로부터의 거리를 보조변수로 사용하여 Co-kriging 방법을 적용하였고, 위의 추정 결과와의 비교 목적으로 보조변수를 사용하지 않는 거리반비례평균법(Inverse Distance Weighting)과 Kriging을 이용하였다. 세 가지 공간보간법에 의하여 추정된 연속면 지도를 비교한 결과, 보조 변수를 이용한 Co-kriging 방법에 의한 수치지도가 강우산도의 미세한 분포 양상을 나타내는 데 적합하게 판정되었다. 또한 실지 관측치와 추정치와의 차이를 분석하는 역검정방법(cross validation)을 이용하여 추정오차를 구한 결과, Co-kriging에 의한 추정치가 최소의 오차를 보여 주였다. Co-kriging이 현재의 GIS 사용자들에게 다소 익숙지 않은 공간추정방법이지만, 여러 종류의 점관측 환경인자의 공간추정에 매우 적합한 방법이라 할 수 있다.

INTRODUCTION

The validity of information derived from a geographic information system(GIS) is crucially dependent on the quality of data in each of the map layers. Each map layer will have some inherent errors regarding the positional and thematic accuracies as a result of the map generation procedure. These errors will be compounded when data are manipulated in multiple-map operations. Of particular concern are errors generated when data derived from course sampling are applied at a fine scale. This is often the case with data collected from sparsely distributed point observations, such as data from climate variables, soil survey, geological survey, and many other environmental variables.

Although these variables are among the critical variables in GIS analysis for a variety of environmental and ecological applications, it has been difficult to define proper methods and procedures for generating suitable data layers for GIS databases. Traditionally, point observation data are transformed to map by tedious hand- drawing of contours, requiring visual interpolation among the sparse point data from measurement locations. Today the process is often automated in a two-step procedure. The first step is to predict the unknown values of a variable of interest for every grid point using simple local averaging procedure or more

sophisticated interpolation algorithms(Willmott et al., 1985; Harcum and Loftis, 1987). The second step is to construct contour lines on the interpolated grid surface. In this paper, I will focus on the first step of estimating data at the unsampled grid points.

Automated interpolation schemes are generally based either on a simple local averaging process(i.e., inverse distance weighting) or on rather sophisticated interpolation methods that use geostatistical theory. In either case, the estimation at the unsampled locations is essentially based on a weighted linear combination of the known sample data:

$$\hat{U}_0 = \sum_{i=1}^{n} w_i U_i \qquad \text{.....(1)}$$

where $\hat{U}_0$ is the estimate of a point variable at the unsampled location o, U_i is the actual measurements of the variable, and w_i is the weight for U_i. In case of inverse distance weighting method, the weight w_i is determined to be inversely proportional to its distance from the location of $\hat{U}_0$. For geostatistics-based interpolation schemes such as kriging, the weights are calculated using a spatial autocorrelation function(semivariogram) that defines the spatial self-dependence of a variable by the lag-distance. More detailed descriptions of spatial interpolations including inverse distance weighting and kriging can be found in Cressie (1991a, 1991b), Isaaks and Srivastava (1989), and Burrough (1986).

For the case of environmental variables where the distances between measurement points are sparsely distributed, the interpolated maps created by these procedures may not be adequate to represent the site specific variations. It is similar situation for the acid rain monitoring in Korea where the average distance between the monitoring stations is about 40 km. Furthermore, it is important to produce a map where the estimates of the unsampled area have appropriate degree of accuracy. The objective of this study is to define a proper methodology to create map of point observed environmental variables, which will have more reliable estimations over the unsampled locations.

STUDY AREA AND DATA USED

As a case study, I attempted to create rainfall acidity maps over the southern part of Korean Peninsula. During the last few years, there has been a growing concern about acid rain over the country. Consequences of rapid industrialization accomplished during the last few decades in Korea have brought several environmental problems that are increasingly getting public awareness. Acid rain is one of the environmental elements that are great concern to public and its impacts on natural ecosystems as well as human life are often recognized as one of critical

environmental problems in this country.

Starting from 1991, the Forestry Research Institute (FRI), a government research organization, initiated the acid rain monitoring program to study a long term effects of acid rain on forest ecosystem. Considering that forest lands occupy two third of the total land in Korea, the program would contribute substantial information that are useful for maintaining forest productivity as well as analyzing overall environmental conditions in Korea. Currently there are total 65 acid rain monitoring stations and they are evenly distributed all over the country (Figure 1). For this study, I used rainfall acidity data that were collected during the three months from March to May in 1992 and the average value was calculated for each of the 63 monitoring stations within the main land. The site related data (longitude, latitude, elevation, proximity to coast line, etc...) were collected and introduced into a GIS database. The longitude and latitude of the monitoring stations were converted to a plane rectangular coordinate system that includes the whole area as a single coordinate zone.

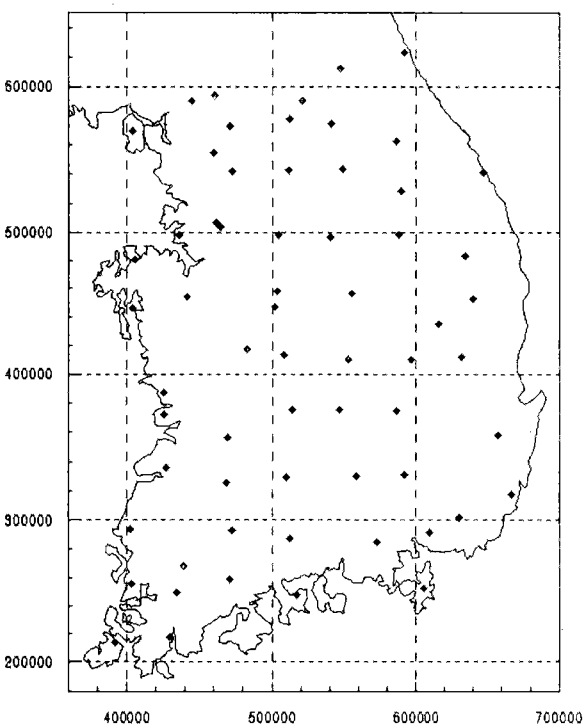

Figure 1. Location of the 63 acid rain monitoring stations used for this study.

SPATIAL ESTIMATIONS

Typical interpolation methods use known values of one variable to predict the same variable at unknown locations. For instance, temperature data collected at the weather stations are used to predict temperature values at unknown grid points. However, considering that there are several other variables (such as elevation for the temperature) that may be correlated to the primary variable of interest, it would be useful to include these secondary variables in the estimation. By incorporating secondary variables during the estimation process, we might expect advantages of both presenting site-specific variation and improving the estimation accuracy.

Incorporating the secondary variables is beneficial only when they correlate with the primary variable to be estimated. To find the secondary variables that are correlated with rainfall acidity, several geographical and topographical factors were analyzed.

Proximity to the coast line of western and southern part of the country showed the highest correlation with the acidity data among the possible secondary variable analyzed (Figure 2). Once the coordinates of each of the 63 monitoring stations were digitized, its proximity from coast line was obtained by a simple map overlay. Topographic variables (elevation and aspect) did not show any significant correlations with the acidity data. Since the 63 monitoring stations were located at relatively low elevation and flat areas, the range of variations for these variables is relatively narrow. Therefore, it seems reasonable to observe the poor correlation between topography and acidity measurements.

A simple way to incorporate secondary variable is a regression model in which the environmental variable can be estimated from a set of independent variables. The regression model can be derived from the acidity data and the corresponding location variables including longitudinal coordinate, latitudinal coordinate, and proximity to coast line. Several regression models, however, derived from the data set did not show adequate evidence to explain the total variation and, therefore, were not considered for this study.

Co-kriging method, that is relatively unknown to general GIS user community, was used to incorporate a secondary variable to create rainfall acidity map. In addition to the two-variable estimation method, rainfall acidity maps were generated using two other commonly used single-variable interpolation methods for comparison. Unlike co-kriging, the two other interpolation methods (inverse distance weighting and ordinary kriging) use only a single variable itself to be estimated without incorporating the secondary variable.

Co-Kriging Estimation

Like linear interpolation (Equation 1), co-kriging is a weighted linear estimation. However, unlike kriging, which is a linear combination of only one primary variable, co- kriging is a linear

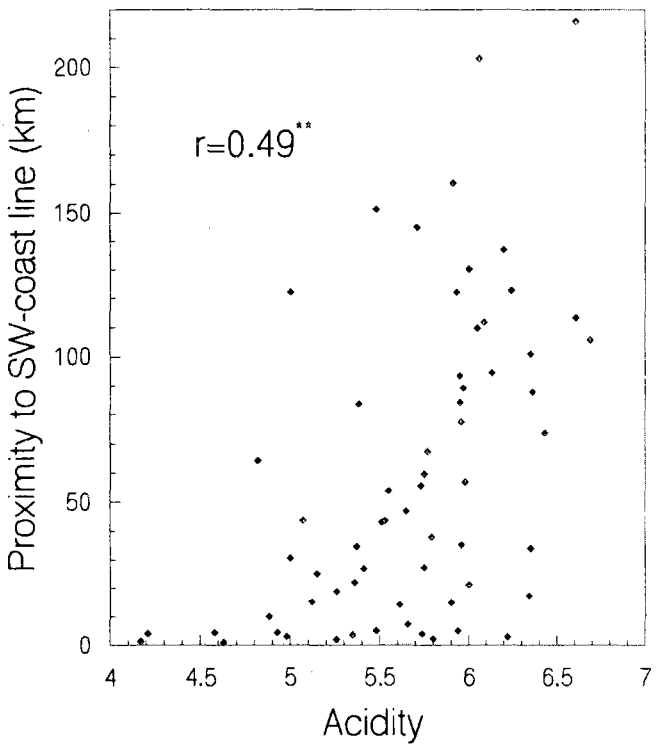

Figure 2. The relationship between the average rainfall acidity and proximity to coast line (based on the sample data collected at the 63 monitoring stations).

combination of both the primary variable to be estimated and a secondary variable that is correlated to the primary variable. Co-kriging estimate is defined by

$$\hat{U}_0 = \sum_{i=1}^{m} a_i U_i + \sum_{j=1}^{n} b_j V_j \qquad \text{.....(2)}$$

where $\hat{U}_0$ is an estimate at location o, U_i is a measured point data from the primary variable, V_j is a sample from the secondary variable, and a_i and b_j are the corresponding weights for U_i and V_j. The numbers (m and n) of neighboring samples are not usually the same for U and V unless the two variables are only measured at the same locations.

Co-kriging can be divided into two major steps. The first step is to define the spatial self-dependence for each of the two variables and the spatial inter-dependence between the two variables. The spatial dependence is defined by modeling the semivariogram and cross-

semivariogram from the sample data. Once the semivariogram and cross-semivariogram functions have been found, the co-kriging weights a_i and b_j are calculated from a system of equations designed to minimize the variance of the estimation errors. Minimization of the error variance is one of features that makes kriging and co-kriging more robust than other ordinary interpolation methods. Detailed descriptions of co-kriging and the system of equations are given by Myers (1983) and Isaaks and Srivastava (1989).

For the co-kriging estimate, I initially calculated semivariances for each of the acidity data and the proximity data. Semivariance is calculated by

$$\gamma(h) = \frac{1}{2m} \sum_{i=1}^{m} (U_i - U_{i+h})^2 \qquad \text{.....(3)}$$

where m is the number of pairs separated by the lag-distance h, U_i is the value of a rainfall acidity or proximity to coast line at location i, and U_{i+h} is the value at the distance h away from i. While there were only 63 samples for rainfall acidity data, proximity to coast line could be obtained anywhere on the country. To calculate the semivariance and cross-semivariance, proximity values were systematically obtained at a spacing of 10x10 km. Semivariograms representing the change in semivariance of each variable with increasing lag-distance h, were produced (Figure 3).

The idea of using a semivariogram to describe the spatial dependence of a variable with itself can be extended to situations with two variables. Instead of considering the change in only one variable, cross-semivariances consider the simultaneous change in two variables with increasing lag distance. The cross- semivariances between rainfall acidity and proximity were calculated as follows:

$$\gamma_{uv}(h) = \frac{1}{2m} \sum_{i=1}^{m} (U_i - U_{i+h})(V_i - V_{i+h}) \qquad \text{.....(4)}$$

where m is the number of pairs compared, U_i and V_i are the sample values of rainfall acidity and proximity respectively, and U_{i+h} and V_{i+h} are the sample values of rainfall acidity and proximity separated by the lag-distance h.

Since we are dealing with two-dimensional space rather than a one-dimensional function, the semivariograms are also two dimensional. To examine the directional variability of semivariances, I calculated four semivariograms from each of four major directions. Using the four semivariograms, the anisotropy ratio and angles were determined. Each of the sample semivariograms and cross-semivariogram was then fitted to an appropriate mathematical function by nonlinear least square minimization procedure. As can be seen in Figure 3, the best

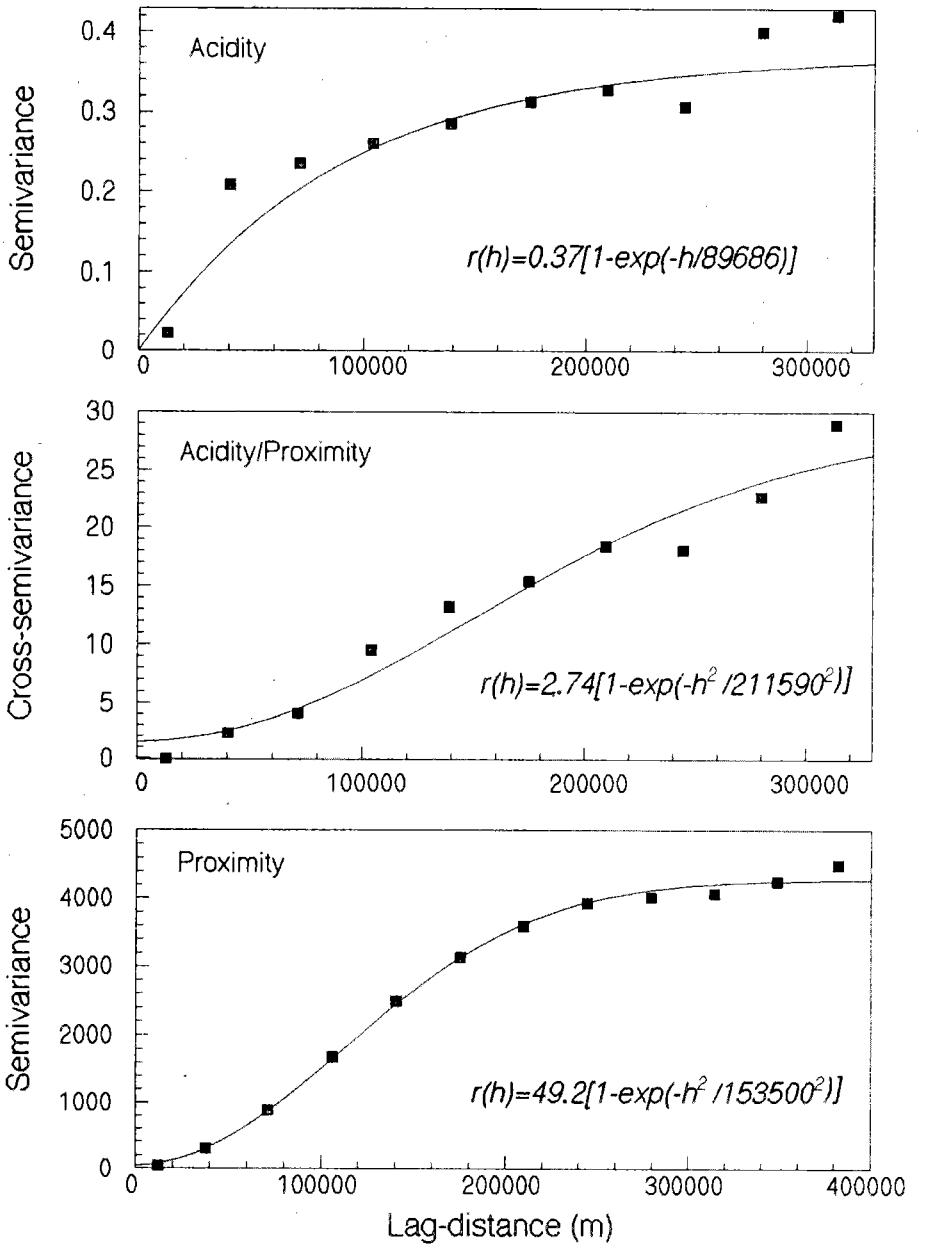

Figure 3. Semivariogram and cross-semivariogram models used for the co-kriging and the kriging.

fitting models for both the semivariograms and cross-semivariogram were found using exponential function.

Once the semivariogram and cross-semivariogram models were obtained, the rainfall acidity value at an unknown grid point could be estimated by calculating the co- kriging weights in equation (2) from the following system of matrix equation:

$$W = C^{-1} D \qquad \text{.....(5)}$$

where W is a vector of the co-kriging weights a_i and b_i, C is a matrix showing the covariances of the acidity and proximity between the neighboring samples surrounding the point to be estimated, and D is a vector of covariances between the unknown point to be estimated and the nearby sample points. The components in system C and D were computed from the semivariogram and cross-semivariogram models obtained previously in Figure 3. Rainfall acidity estimates were made for each of 2x2 km^2 grid cells over the country of 163x235 grids. Most co-kriging computations described above were performed using a geostatistics package developed by the Environmental Protection Agency (Yates and Yates, 1990).

Inverse Distance Weighting and Kriging

To compare the outcome of co-kriging estimation method, additional rainfall acidity maps were created by applying more commonly used interpolation methods that did not use the secondary variable. The first method used was inverse distance squared weighting to interpolate the 63 sample data over the space of the same 163x235 grid cells. In this method, the prediction at an unknown location was made by solving the weights in equation (1):

$$w_i = \frac{1/d_i^2}{\sum_{i=1}^{n} 1/d_i^2} \qquad \text{.....(6)}$$

where d_i is the distance between the sample point of U_i and the location being estimated.

The kriging estimation followed essentially the same steps as co-kriging method except that the kriging uses only rainfall acidity data without the help of the secondary variable of proximity data. In the kriging method, only one semivariogram model (Figure 3) derived from the acidity data was used to compute the covariances in the system of matrix C and D in equation (5).

EVALUATIONS OF DIFFERENT METHODS

Although the ideal evaluation would be to compare the predicted values with the true values, it is unrealistic to acquire a comprehensive data set that measures rainfall acidity for every grid points over the entire area. The three estimation methods were evaluated by both visual interpretations of the outcome acidity maps and quantitative assessment of statistical error measures.

A simple and effective evaluation was carried out by visually comparing the rainfall acidity maps produced from the above estimation methods (Figure 4). The two acidity maps created by inverse distance weighting and kriging, which did not incorporate the secondary variable, are very similar in which they show a plain monotonic change from one monitoring station to another. These two maps show a spatial pattern of rainfall acidity that tend to decrease toward eastern part of the country.

The acidity map created by co-kriging is not much different from the two single-variable estimation methods. Although the overall patterns of rainfall acidity are similar for all three maps, the map created by co-kriging method shows subtle directional details that toward north-eastern part of the country. In addition, it does not show the distinct positional pattern corresponding to the location of the 63 monitoring stations, which are noticeable on the two other maps. The north-eastern directional pattern displayed at the co-kriging map is not surprising given that the secondary variable used for this estimation is the proximity to south-western coast line. Since the proximity values are also monotonically changed to one direction without site variation, the acidity map estimated by incorporating proximity data does not appear to represent the site specific variations. If elevation was the secondary variable that showed high correlation with rainfall acidity, the co-kriging acidity map would have shown site specific variations influenced by elevation.

The reliability of the estimate is often assessed by the mean squared error (MSE) which can account for both bias and stability of the error distribution:

$$MSE = \frac{1}{n} \sum_{i=1}^{63} (U_i - \hat{U}_i)^2 \qquad \text{.....(7)}$$

Since the three estimation methods (co-kriging, kriging, and inverse distance weighting) are exact interpolator, in which they produce the exact sample value for each sample location, their mean squared errors are all zero. To compensate for this limitation, I used a technique called cross validation to indirectly measure the estimation error using the sample data set. In cross validation, one of the 63 samples was temporarily eliminated from the data set and the

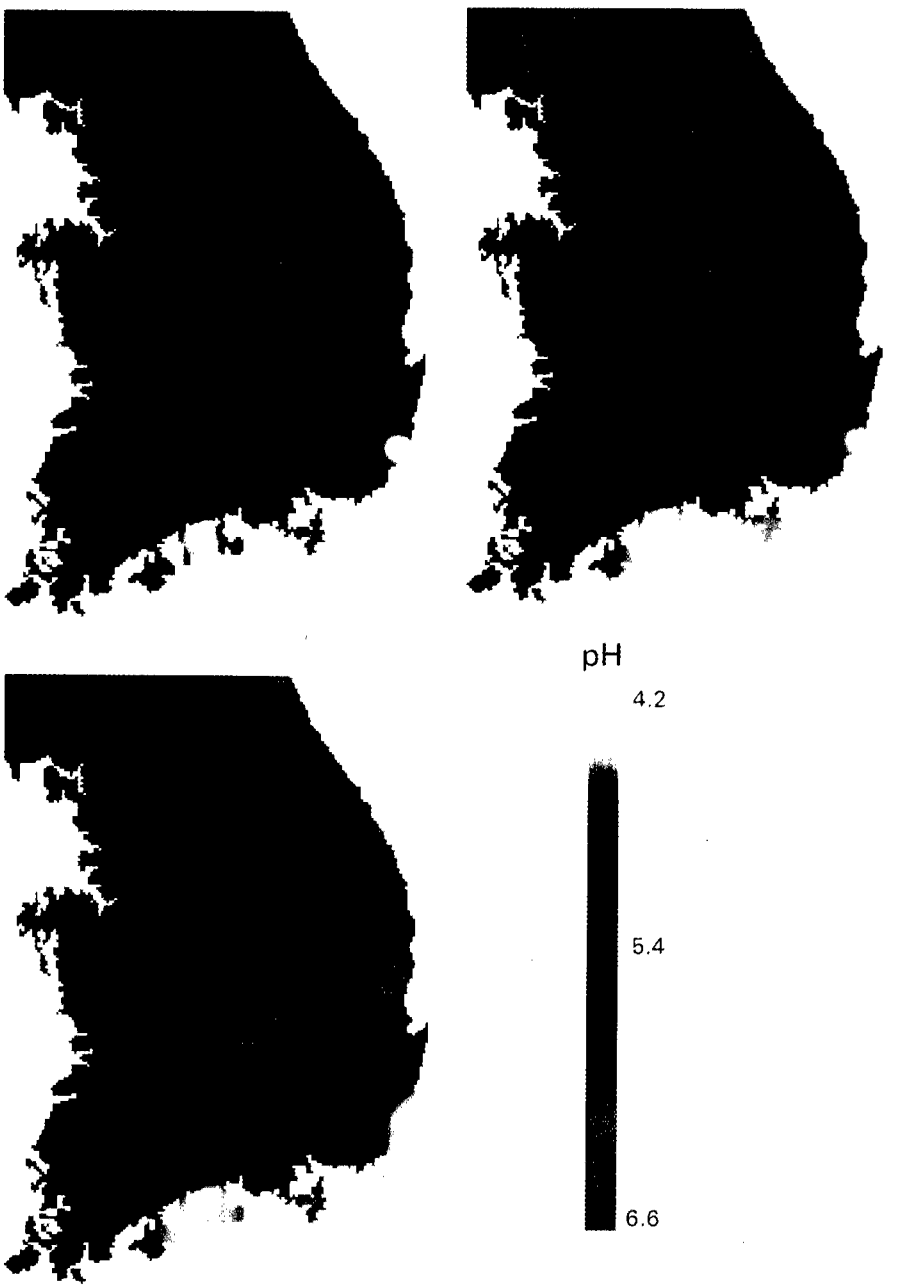

Figure 4. Maps of the estimated rainfall acidity values by three different interpolation methods: (a) inverse distance squared weighting, (b) kriging, and (c) co-kriging.

estimation was carried out for the location of the eliminated sample point. The estimated value was then compared to the true value at the sample point. This exercise was repeated for each of the remaining 62 samples. At the above equation, U_i is the measured sample value at the station i and $\hat{U}_i$ is the estimated value at the same location obtained by cross validation.

The cross validation was applied to all three different methods. Table 1 shows the summary statistics for the estimation errors of three methods. Co-kriging method shows the lowest estimation error among the three methods while inverse distance weighting and kriging have slightly higher error. Although the level of estimation error was not significant among the three methods, co-kriging method that incorporates the secondary variable shows lower estimation error as compared to the two single variable methods. As can be seen from Figure 5 that shows the size of residuals by sample value, the two single-variable estimation methods show larger residuals at low and high sample values than the co-kriging estimate. It is worthwhile to mention that cross-validation is a way of comparing the three methods and may not be used as an

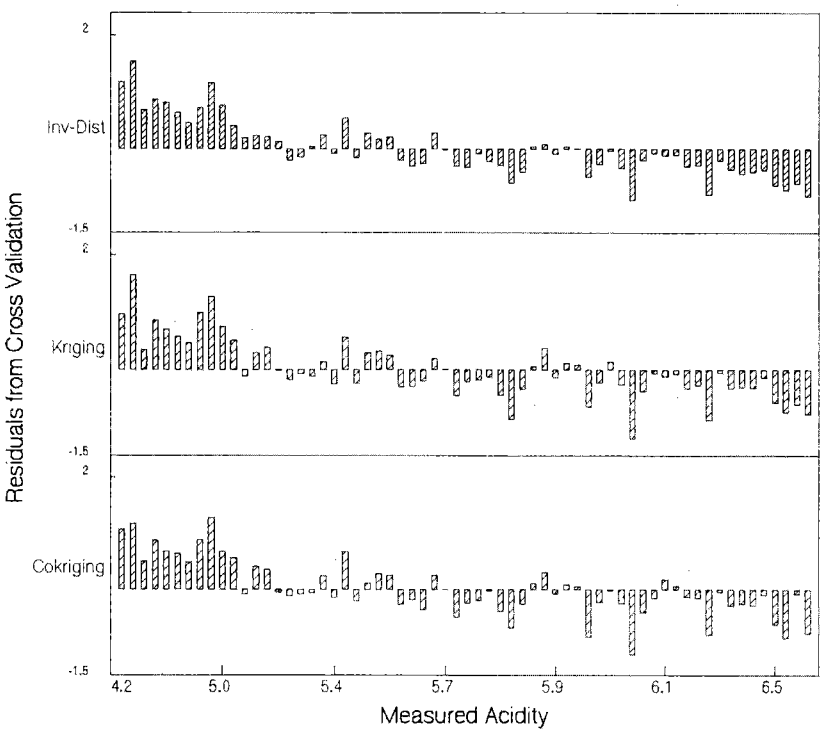

Figure 5. Distribution residuals obtained from cross-validation. Note that the sizes of residuals are decreased at the co-kriging estimate as compared to inverse distance weighting and kriging.

Table 1. Summary statistics of estimation errors computed using the 63 sample data set. These statistics are based on the residuals produced by cross validation.

	Co-Kriging	Inverse-distance	Kriging
SSE	14.967	16.098	17.064
RMSE	0.487	0.502	0.516
MAE	0.368	0.392	0.399

absolute measure to advocate one method over others. True benefit of co-kriging method can be found from its theoretical strength that minimizes the error variance during the estimation process.

DISCUSSIONS

The comparison of the three estimation methods was carried out using different approaches. The error estimates were computed based on the sample data from the 63 acid rain monitoring stations. The stations were limited in number, restricted to relatively level areas, and sited at low elevation area. Without an independent set of true-measured data that can be used to test the estimated values, selecting one best estimation method could be misleading. Yet, some general results are worth pointing out. Throughout the comparison it was clear that co-kriging method using the secondary variable produced rainfall acidity map that shows the north-eastern spatial pattern that is correlated to the acidity data. More importantly, the co-kriging estimates were more reliable in terms of low estimation errors derived from cross-validation.

Probably, the most profound factor that can distinguish co-kriging method from ordinary inverse-distance weighting method can be found from the strong theoretical basis. Co-kriging method uses the spatial inter-dependence of the two variables that are highly correlated and interpolates using only those neighboring samples that are within the range of influence. The geostatistics-based methods are often considered better (or optimal) since the interpolation weights, w_i in Equation (1), are determined in a way that minimizes the variances of estimation errors in the kriging and co-kriging methods (Burrough, 1986). Minimization of the error variance is one of features that makes kriging and co-kriging more robust than other ordinary interpolation methods.

The search strategy to find the neighboring samples is another important step in spatial interpolation and deserves some comment. There is no clear rule to decide the search radius and the number of the neighboring samples to be used for local averaging interpolations, such as inverse distance method. With co-kriging and kriging, the search radius can be defined more

effectively using the range of influence determined from semivariograms. The numbers of neighboring samples to be used for predicting a rainfall acidity value are not the same for the acidity and proximity variables. While there are only 63 points of rainfall acidity data, the proximity data exist for every point over the entire area. In cases such as this, where the sample number of the primary variable to be estimated is small but the secondary variable can be collected more frequently, co-kriging should be a particularly effective estimation method. Environmental data are often collected by relatively small number of point observations. To overcome the limitation using only the small number of point observed data itself, other secondary variables that is highly correlated with the environmental variable of interest can be incorporated into the estimation procedure.

Geostatistics-based spatial interpolation techniques, particularly kriging, are used increasingly for spatial estimation. However, it is rare to find examples of spatial interpolation using co-kriging method. This might be due to several factors including the lack of software to carry out the intensive computations and, more importantly, the lack of understanding in the theory behind co-kriging. As Cressie (1991) pointed out, co- kriging should be considered for the analysis of the multivariate spatial data. Finally, there might be several other variables that can affect the site-specific variations of the rainfall acidity in Korea. Prevailing wind pattern, land use/land cover, and proximity to urban/industrial area would be other important factors that should be considered in creating a better rainfall acidity map.

SUMMARY AND CONCLUSIONS

Point observed environmental variables are often required in ecological and environmental applications of geographic information systems. To create a more reliable rainfall acidity map that would better represent the spatial variations over the country, three spatial estimations were applied. Using sample data collected from the 63 acid rain monitoring stations and its proximity to coast line, average rainfall acidity during the spring season of 1992 was estimated by co-kriging method. For comparison, I also used two estimation methods based only on the sample data without introducing the secondary variable.

Comparison between different estimation methods was carried out using both qualitative and quantitative evaluations of the outcome of each method. By incorporating the secondary variable of proximity to coast line, the estimated rainfall acidity map using co-kriging showed improvement over the estimations from inverse distance weighting and kriging methods. Although co-kriging method is unfamiliar and relatively complex to implement, it has great potentials to numerous situations of creating spatial map layer for point observed environmental variables. As we are dealing with a variety of environmental data that are measured at sparsely

distributed points and have to covert the point measured data into a map layer in GIS environment, it became more crucial to better represent both thematic and locational accuracies. In cases where there are a limited number of samples of the primary variable but other spatial variables are highly correlated with the primary variable, co-kriging method will be a very appropriate method to be considered.

Acknowledgment

The author acknowledges the generous support by 1994-1995 Inha University Research Fund for the research reported in this paper.

REFERENCES

Burrough, P.A., 1986. Principles of Geographical Information Systems for Land Resources Assessment. Oxford University Press, New York. 194 pgs.

Cressie, N. 1991a. Geostatistical analysis of spatial data. In "Spatial Statistics and Digital Image Analysis" by National Research Council, National Academy Press, Washington D.C., pgs. 87-108.

Cressie, N., 1991b. Statistics for Spatial Data. John Wiley & Sons, Inc., New York. 900 pgs.

Harcum, J.B. and J.C. Loftis, 1987. Spatial interpolation of penman evapotranspiration. Transactions of the American Society of Agricultural Engineers. 30 (1):129-136.

Isaaks, E.H. and R.M. Srivastava, 1989. Applied Geostatistics. Oxford University Press, New York. 561 pgs.

Myers, D.E., 1983. Estimation of linear combinations and co-kriging. Journal of Mathematical Geology, 15 (5):633-637.

Willmott, C.J., C.M. Rowe, and W.D. Philpot, 1985. Small-scale climate maps: a sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. The American Cartographer, 12(1):5-16.

Yates, S.R. and M.V. Yates, 1990. Geostatistics for Waste Management: a user's manual for the GEOPACK geostatistical software systems. U.S. Environmental Protection Agency, EPA/600/8-90/004.