AN ESTIMATION METHOD FOR GROUNDWATER ELEVATION*

CHOON KYUNG CHO AND SUNGKWON KANG**

Abstract. An estimation method for groundwater level elevations is introduced. Using geostatistical techniques and anisotropies, experimental variograms show significant improved correlations compared with those from conventional techniques. The estimation method is applied to a field experimental data set.

Keywords. geostatistics, kriging, groundwater level, optimization

1991 AMS Subject Classification: 86A32

1. Introduction

Groundwater and groundwater contamination have been important issues in environmental concerns. Since movement of groundwater contaminants follows groundwater flow direction, estimation of groundwater level elevation is one of key factors for determining the fate of the contaminants and their spatial distributions.

In this paper, we introduce a geostatistical method and some of new concepts such as averaging process and ratio for estimating groundwater level elevation. A typical set of groundwater data such as measured piezometric heads consists of observations at a number of irregularly spaced locations. The problem we are considering in this paper is to estimate the value at an unmeasured location in the same aquifer from the given set of observations.

Geostatistics [6,9] can be considered as a collection of techniques for the solution of estimation problems involving spatial variables. In recent years, geostatistical methods have been applied to a variety of problems in geohydrology [1]. Among

*The authors thank Dr. Stauffer for providing field data. **The research of this author was supported in part by the KOSEF 951-0104-024-2.

them, the punctual kriging and the universal kriging have been used in soil and water sciences to estimate a spatially distributed random variable at unsampled locations. In this paper, groundwater level elevations at unmeasured locations are estimated using a modified punctual kriging method.

The contents of this paper are following. Regionalized variable, experimental variograms and mathematical models needed in kriging procedure are introduced in Section 2, and the punctual kriging is introduced in Section 3. In Section 4, the 26 water level elevation samples collected at an experimantal site are chosen for kriging. An experimental variogram is obtained by using the robust R_p estimator ($p = 1$) and a mathematical model to fit the experimental variogram is obtained by using the cubic polynomials. The new concepts such as the averaging process and the ratio are introduced to obtain a good experimental variogram.

2. Semivariogram, mathematical models and drift

In this section, we introduce regionalized variable, experimental variograms, mathematical models, and drift needed in kriging procedure.

A mineralized phenomenon can be characterized by the spatial distribution of a certain number of measurable quantities called "regionalized variable." Specifically, a *regionalized variable* is a numerical function with a spatial distribution which varies from one place to another with apparent continuity, but the changes of which cannot be represented by any workable function [11]. Typical regionalized variables are functions describing natural phenomena that have geographic distributions such as the elevation of the ground surface, changes in grade with in an ore body. A regionalized variable is *stationary* if the statistics of the random variable $Z(X_i)$ and $Z(X_j)$ at locations X_i and X_j are equal. That is, the statistics determined for $Z(X)$ are equal to those for $Z(X + \mathrm{h})$ for every h. Specifically, the first-order stationary regionalized variable satisfies $E[Z(X)] = E[Z(X + \mathrm{h})]$ and the second-order stationary regionalized variable satisfies $E[Z(X_1), Z(X_2)] = E[Z(X_1+\mathrm{h}), Z(X_2+\mathrm{h})]$, where $E[Z(X)]$ is the expectation of the regionalized variable $Z(X)$ at the location

X: We will say that the regionalized variable satisfies the *intrinsic hypothesis* if for all h the first and second moments of the difference $Z(X+\text{h})-Z(X)$ depend only on the distance between the two points X+h and X, and not on their individidaul locations [9], i.e.,

$$E\big[Z(X+\text{h})\big] = m(\text{h}),$$

$$E\big[\{Z(X+\text{h}) - m(\text{h})\}^2\big] = 2\gamma(\text{h}),$$

where the function $\gamma(\text{h})$ is the *semivariogram* or *intrinsic function* to be explained in the following.

First, we will be concerned with one of the basic statistical measures of geostatistics, the semivariance, which is used to express the rate of change of a regionalized variable along a specific orientation. Let $Z(X)$ be a function whose difference $Z(X+\text{h})-Z(X)$ have the first and second moments depending only on the distance h between locations X and $X+\text{h}$. The *semivariogram* or *intrinsic function* is denoted by $\gamma(\text{h})$ and defined by

$$\begin{aligned}\gamma(\text{h}) &= \frac{1}{2} var\big[Z(X+\text{h}) - Z(X)\big] \\ &= \frac{1}{2} E\big[Z(X+\text{h}) - Z(X) - E[Z(X+\text{h}) - Z(X)]\big]^2.\end{aligned}$$

Semivariograms can be obtained by several estimators such as the classical estimator, the Cressie-Hawkins robust estimator, and the R_p-estimators. The *classical estimator* [3] is defined by the arithmetic mean of the squared difference $\big[Z(X_i) - Z(X_i+\text{h})\big]$, that is,

$$\gamma_c(\text{h}) = \frac{1}{2N(\text{h})} \sum_{i=1}^{N(\text{h})} |Z(X_i) - Z(X_i+\text{h})|^2,$$

where $N(\text{h})$ is the number of data pairs separated by the vector h. Note that the classical estimator γ_c is not robust since the influence of outliers on the semivariogram γ_c increases as $|Z(X) - Z(X+\text{h})|^2$ increases. On the other hand, the *Cressie-Hawkins robust estimator* [3] defined by

$$\gamma_{ch}(\text{h}) = \frac{1}{2} \frac{\left[\frac{1}{N(\text{h})} \sum_{i=1}^{N(\text{h})} |Z(X_i) - Z(X_i+\text{h})|^{\frac{1}{2}}\right]^4}{0.457 + 0.494/N(\text{h})}$$

is resonably robust, but not robust enough so that they do not produce correct correlation between data points showing erratic behaviors which are commonly observed in field data. The R_p-*estimator* [7] is given by

$$R_p(\mathrm{h}) = \left[\frac{1}{N(\mathrm{h})}\sum_{i=1}^{N(\mathrm{h})} |Z(X_i) - Z(X_i + \mathrm{h})|^p\right]^{\frac{1}{p}}$$

For $0 < p \leq 1$ the estimator is robust; whereas for $p > 1$ the estimator becomes sensitive to apparent outliers. As p approaches 0, robust effects are increased and the estimator R_p reduces effects of outlier for $0 < p \leq 1$.

The semivariogram is known only at discrete points. The discrete experimental semivariogram should be modelled by a continuous function, a mathematical model described below, that can be evaluated for any desired distance $h = ||\mathrm{h}||$ because the experimental variogram cannot provide variance for every lag h. The *spherical model*

$$\gamma(h) = \begin{cases} c_0 + \beta(\frac{3h}{2\alpha} - \frac{h^3}{2\alpha^3}), & h \leq \alpha, \\ c_0 + \beta, & h > \alpha, \end{cases}$$

and the *exponential model*

$$\gamma(h) = c_0 + \beta(1 - exp(-\frac{h}{\alpha}))$$

are commonly used [3,4], where $h = ||\mathrm{h}||$ is the radius of the vector h, and the parameters c_0 is the nugget effect, $c_0 + \beta$ is the sill value, and α is the distance h at which variogram reaches the sill value. Also, the *third-order polynomial model* is given by

$$\gamma(h) = a + bh + ch^2 + dh^3.$$

The choice of mathematical model depends strongly on data distribution represented by the experimental semivariogram.

We now consider drifts. Kriging is based on the geostatistical assumption which will be described in the next section, i.e., the expected value of the sample is stationary regardless of sample location. But, practically, if the locations of the samples are different, then the corresponding values are different. Let $Z(X)$ be a regionalized variable. Then the *drift* $m(X)$ is defined by

$$m(X) = E[Z(X)].$$

That is, the drift at a point X is the expected value of the regionalized variable Z at point X [4]. Let $Z(X)$ be a regionalized variable with drift $m(X)$. Then, the *residual* $Y(X)$ is defined by

$$Y(X) = Z(X) - m(X).$$

Then it is easy to see that the residual $Y(X)$ from a regionalized variable satisfies that

$$E[Y(X)] = 0.$$

The form of the drift $m(X)$ depends on spatial data distribution. Often $E[Z(X)] = m(X)$ is expressed as a linear combination of polynomials in the spatial coordinates. Also, $m(X)$ may be a linear combination of any functions $\{f_\ell : \ell = 1 \text{ to } k\}$, the coefficients a_ℓ of which are unknown, so that the drift $m(X) = \sum_{\ell=1}^{k} a_\ell f_\ell(X)$ remains unknown.

3. Kriging

Kriging is an interpolation method to estimate values at unmeasured locations. It uses information from the semivariogram to find an optimal set of weights. The kriging is based on the *geostatistical intrinsic assumption*:

1) $E[Z(X)] = E[Z(X + \mathrm{h})]$.

2) For any vector h the increment $[Z(X) - Z(X + \mathrm{h})]$ has a finite variance which does not depend on X, i.e.,

$$Var[Z(X) - Z(X + \mathrm{h})] = E[Z(X) - Z(X + \mathrm{h})]^2.$$

For unmeasured location X_0, assume that the value at $Z^*(X_0)$ can be obtained as a linear combination of the selected sample values, that is,

$$Z^*(X_0) = \sum_{i=1}^{n} w_i Z(X_i), \tag{3.1}$$

where $Z^*(X_0)$ is the estimated value of Z at X_0 and w_i's are weights to be determined. Optimization of the statistic $Z^*(X)$ will be performed by imposing the following two constraints

$$E\big[Z^*(X_0) - Z(X_0)\big] = 0, \tag{3.2}$$

$$E\big[\{Z^*(X_0) - Z(X_0)\}^2\big] \quad \text{is a minimum with respect to} \quad w_i, \tag{3.3}$$

where $Z(X_0)$ is the value of the random function Z at X_0. These restrictions imply that the difference $Z^*(X_0) - Z(X_0)$ is unbiased and the variance of this difference is a minimum. Using (3.1) and (3.2), it can be written

$$E\left[\left[\sum_{i=1}^{n} w_i Z(X_i)\right] - Z(X_0)\right] = 0. \tag{3.4}$$

Taking the expectation of each value and equating it to the mean, m, which is assumed to be constant, yield

$$\sum_{i=1}^{n} w_i E\big[Z(X_i)\big] - E\big[Z(X_0)\big] = \sum_{i=1}^{n} w_i m - m = 0 \tag{3.5}$$

resulting in

$$\sum_{i=1}^{n} w_i = 1. \tag{3.6}$$

We assume that the mean value, m, is constant, but unknown. For *punctual kriging* which is most commonly used in practice, the minimization of (3.3) is carried out over $(w_1, w_2, \cdots w_n)$ subject to $\sum_{i=1}^{n} w_i = 1$, i.e., minimize

$$E\left[\sum_{i=1}^{n} Z(X_i) - Z(X_0)\right]^2 - 2\lambda\left(\sum_{i=1}^{n} w_i - 1\right) \tag{3.7}$$

with respect to $w_1, w_2, \cdots, w_n$ and λ (the parameter λ is a *Lagrange multiplier* that ensures $\sum_{i=1}^{n} w_i = 1$). Now the condition $\sum_{i=1}^{n} w_i = 1$ implies that

$$\begin{aligned}&\left[\sum_{i=1}^{n} w_i Z(X_0) - Z(X_0)\right]^2 \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \big(Z(X_i) - Z(X_j)\big)^2/2 + 2\sum_{i=1}^{n} w_i \big(Z(X_0) - Z(X_i)\big)^2/2.\end{aligned} \tag{3.8}$$

So that (3.7) becomes

$$-\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(X_i - X_j) + 2\sum_{i=1}^{n} w_i \gamma(X_0 - X_i) - 2\lambda\Big(\sum_{i=1}^{n} w_i - 1\Big). \qquad (3.9)$$

By differentiating (3.9) with respect to $w_1, w_2, \cdots, w_n, \lambda$, we have the following optimality conditions

$$-2\sum_{j=1}^{n} w_j \gamma(X_i - X_j) + 2\gamma(X_0 - X_i) - 2\lambda = 0, \quad i = 1, 2, \cdots, n,$$

$$\sum_{i=1}^{n} w_i = 1.$$

Thus, the punctual kriging system is represented by the following matrix form

$$AW = b,$$

where

$$A = \begin{bmatrix} \gamma(X_1 - X_1) & \gamma(X_1 - X_2) & \cdots & \gamma(X_1 - X_n) & 1 \\ \gamma(X_2 - X_1) & \gamma(X_2 - X_2) & \cdots & \gamma(X_2 - X_n) & 1 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \gamma(X_n - X_1) & \gamma(X_n - X_2) & \cdots & \gamma(X_n - X_n) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix},$$

$$W = \big[w_1, w_2, \cdots, w_n, \lambda\big]^T,$$

$$b = \big[\gamma(X_0 - X_1), \gamma(X_0 - X_2), \cdots, \gamma(X_0 - X_n), 1\big]^T.$$

Note that A is a $(n+1) \times (n+1)$ nonsingular symmetric matrix. The optimal weights $w_1, w_2, \cdots, w_n$ can be obtained from $W = A^{-1}b$. The resulting estimation variance of punctual kriging is

$$\sigma^2 = \sum_{i=1}^{n} w_i \gamma(X_0 - X_i) + \lambda - \gamma(X_0 - X_0).$$

4. Application

In this section, we apply the kriging method to a sample data set collected during a field experiment [2,8]. The aquifer at the test site consists of a shallow alluvial terrace deposit averaging approximately $11m$ in thickness. The aquifer is composed of poorly-sorted to well-sorted sandy gravel and gravelly sand with minor amounts of silt and clay. Sediments are generally unconsolidated, and occur as irregular horizontal or nearly horizontal lenses and layers. Marine sediments belongs to the Eutaw Formation and consisting of clays, silts, and fine-grained sands form an aquitard beneath this alluvial aquifer. Table 1 shows the water level elevations of 26 samples in meter scale. Figure 1 shows the elevations in each direction. The spatial range of the 26 samples were $-86 \leq x \leq 103.20, -50.16 \leq y \leq 263.96$ in meter scale, and water level elevation values were ranged from $64.13m$ to $65.29m$.

We need to analyze whether the 26 samples satisfy the geostatistical assumption described in Section 3 before applying kriging. It is easy to observe from Figure 1 that the water level elevations have approximately random distribution in x-direction. On the other hand, the water level elevations in y-direction show an apparent correlation between y coordinate and the heights. Thus, the samples do not satisfy geostatistical assumption and, hence, we cannot apply punctual kriging directly to our samples. Based on the water level distribution profile in Figure 1, a cubic polynomial $a + by + cy^2 + dy^3$ was chosen for finding the drift in y-direction. Parameters a, b, c, d in the cubic polynomial were estimated by solving the following minimization problem :

$$\text{Minimize} \quad \sum_{i=1}^{n} [a + by_i + cy_i{}^2 + dy_i{}^3 - Z(X_i)]^2$$

using the Conjugate Gradient Method (see, for example, [5]). The residuals were obtained by subtracting the drift from the original data. Figure 2 shows the estimated drift in the y-direction, (a), the distribution of the residuals in each direction, (b) and (c), and the fitted mathematical model, (d). It was observed that the variance of the residuals was much less than that of the original data (approximately 60% reduction). Punctual kriging was applied for the residuals to estimated expected residual values at unmeasured locations. Experimental variograms needed in kriging were estimated from the residuals. The variograms were obtained by the four estimators introduced in Section 2; the classical estimator $\gamma_c(h)$, the Cressie-Hawkins estimator $\gamma_{ch}(h)$, and the R_p estimator with $p = 1,\ 1/2$. However, they did not show any apparent correlation between lag h and the corresponding variogram value due to the variance of data distribution. To derive the correlation we used the concepts of averaging process and ratio. Consideration of *averaging process* is to due to statistical aspect resided in data distribution. Within a certain distance, say d, from a lag value h the values of all semivariances are averaged. The averaged value represents the semivariance within the range. The semivariogram obtained by considering averaging process showed more continuous information than that from discrete experimental one. In our applications, the lag average d was chosen as $d = 7$. We observed that the averaged experimental variograms obtained by the four estimators showed improved correlation between variograms and lag h's compared to those of without any averaging process. But, there was a difficulty in fitting a mathematical model to the averaged experimental variogram due to anisotropy effects on data distributions.

To account for anisotropy effects on correlation, we considered a *ratio* between two directions x and y. This ratio is a realization of the advection and dispersion/diffusion processes in conjuction with the aquifer material. To get the ratio of given samples, we chose the data range which can represent the main advection and dispersion/diffusion profiles of the data distributions in the considered region. Consider the distributions of water level elevation whose values are from $64.2m$ to $65.2m$ in Figure 1. The reason of choosing these water level elevation values was that samples within this range represent main profile of given 26 samples. The samples with this range were distributed from approximately $x = -50m$ to $x = 50m$. The corresponding y ranges were from $y = -50m$ to $y = 250m$. The ratio was obtained by comparing these ranges. We get the ratio approximately $x : y = 100 : 300 = 1 : 3$. From this ratio, any two points separated by approximately $1m$ distance in x-direction, $3m$ distance in y-direction, are considered to have the same correlation in the average sense. Figure 2 (d) shows the improvement of an averaged experimental variogram by accounting for the ratio $x : y = 1 : 3$ and the fitted cubic polynomial model, $a + bh + ch^2 + dh^3$, corresponding to the variogram obtained by the Rp estimator with p=1. We estimated residuals at lattice points separated by $5m$ in the range of $-70 \leq x \leq 50,\ 0 \leq y \leq 190$ by punctual kriging. Finally, we obtained the estimated water level elevations at those points by summing the estimated residuals and the drifts. Figures 3 and 4 show the contour and the three-dimensional plots of the estimated water level elevations, respectively.

REFERENCES

1. ASCE Task Committe on Geostatistical Techniques in Geohydrology of the Ground Water Hydrology Committee of the ASCE Hydraulics Division, *Review of Geostatistics in Geohydrology. I. Basic concepts; II. Applications*, Hydraulic Engineering **116**, **No. 5** (1990), 612-658.
2. J. M. Boggs, L. M. Beard, W. R. Waldrop, T. B. Stauffer, W. G. MacIntyre and C. P. Antworth, *Transport of tritium and four organic compounds during a natural gradient experiment (MADE-2)*, EPRI Report TR-101998, Electric Power Research Institute, Palo Alto, CA 94304, 1993.
3. Noel A. C. Cressie, *Statistics for Spatial Data*, John Wiley and Sons, Inc., New York, 1991.
4. J. C. Davis, *Statistics and Data Analysis in Geology*, John Wiley and Sons, New York, 1986.
5. R. Fletcher, *Practical Methods of Optimization*, 2nd Ed., John Wiley and Sons, Chichester, 1993.
6. A. G. Journel and J. C. Huijbregts, *Mining Geostatistics*, Academic Press, San Diego, 1993.
7. S. Kang, T. B. Stauffer and K. Hatfield, *A practical three-dimensional estimation method for in situ groundwater contaminants, Part I: Methodology, Part II: Application*, Preprint, 1994.
8. W. G. MacIntyre, J. M. Boggs, C. P. Antworth and T. B. Stauffer, *Degradation kinetics of aromatic organic solutes introduced into a heterogeneous aquifer*, Water Resources Research 29(12) (1993), 4045-4051.
9. G. Matheron, *The theory of regionalized variable and its applications*, Les Cahiers du Centre de Morphologie Mathématique, Fasc. 5, CG Fontainebleau, 1971.
10. K. B. Mital, *Optimization Methods in Operations research and systems analysis*, Wiley Eastern Limited, 1976.
11. R. A. Olea, *Optimum Mapping Techniques Using Regionalized Variable Theory*, Kansas Geological Survey Series on Spatial Analysis **No. 2** (1975).

DEPARTMENT OF MATHEMATICS CHOSUN UNIVERSITY, KWANGJU 501–759, KOREA
E-mail address: sungkang@chollian.dacom.co.kr