A PRACTICAL ESTIMATION TECHNIQUE FOR SPATIAL DISTRIBUTION OF GROUNDWATER CONTAMINANT

Sungkwon Kang
Department of Mathematics
Chosun University, Kwangju 501-759, Korea and
Institute for Scientific Computation
Texas A&M University, College Station, TX 77843-3404, U.S.A.
skang@chonnam.chonnam.ac.kr

Thomas B. Stauffer
AFRL/MLQR, Tyndall AFB, FL 32403, U.S.A.
tom_stuffer@ccmail.aleq.tyndall.af.mil

Kirk Hatfield
Department of Civil Engineering
University of Florida, Gainsville, FL 32611, U.S.A.
khatf@ce.ufl.edu

Abstract: To predict the fate of groundwater contaminants, accurate spatially continuous information is needed. Because most field sampling of groundwater contaminants are not conducted spatially continuous manner, a special estimation technique is required to interpolate/extrapolate concentration distributions at unmeasured locations. A practical three-dimensional estimation method for *in situ* groundwater contaminant concentrations is introduced.

1 INTRODUCTION

Groundwater contamination is an important environmental issue. Researchers have conducted extensive field experiments to analyze geophysical, chemical and biological processes that control the fate and movement of groundwater contaminants [1,3,4,9,11,12]. In order to describe and predict underlying physical, chemical and biological processes affecting chemical fate and transport, accurate spatially continuous information is needed. Because most field sampling

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35359-3_40

S. Chen et al. (eds.), *Control of Distributed Parameter and Stochastic Systems*

of groundwater contaminants are not conducted spatially continuous manner, a special estimation technique is required to interpolate/extrapolate contaminant concentration distributions at unmeasured locations. These interpolations/extrapolations are complicated by uncertainties often associated with an unknown distribution of contaminant fluxes in space and time reflected in a complex velocity field within a heterogeneous aquifer.

Many geostatistical techniques have been developed for estimating geophysical/chemical parameters and groundwater contaminant concentrations (see, for example, [2,5,6,7,10] and references there in). But, application of these estimation methods to groundwater contaminant data often fail to obtain satisfactory results because methods are based on the *geostatistical intrinsic assumptions* [10], and because field data behave irratically or contain "outliers." The geostatistical intrinsic hypothesis (or stationary assumptions) is that spatial correlationship between data points depend only on the separation vector (modulus and direction) and not on the individual sample location. But, the global behavior of contaminant plume follows dynamical process governed by groundwater flow; consequently, the concentration of contaminant strongly depends on sample location. Questions concerning the locations or regions of high pollutant concentrations and contaminant plume dimensions are key issues in enviromental concerns. The first obstacle, the intrinsic geostatistical hypothesis, can be overcome by extracting the *plume macroscopic behavior* from the field data. The concept of macroscopic plume behavior is essentially similar to that of *drift* or *trend* in geostatistics in the sense of nonstationarity. In geostatistics, the general profile of most regionalized variables is assumed to be stationary, and, hence, the slowly varying minor nonstationary components (drift or trend) observed in the field data may be approximated by lower order polynomials. However, the greater portion of a goundwater contaminant plume exhibits nonstationary characteristics. Thus, a major component of the plume should be estimated from dynamical processes and measured in a large region. On the other hand,=0Dthe drift should be estimated in a "small neighborhood" of the point where kriging to be performed. Additional problems arise when concentrations are estimated. For example, conventional semivariograms are too sensitive to obtain correct spatial correlations for data exhibiting a wide variance. These correlations are needed for determining variogram models in a kriging procedure. The log transformation commonly used to compress data variance contains a logical conflict between original data structure and application of kriging algorithm. Consequently, to make spatial interpolations of data exhibiting a large variance, there is a need to develop a new robust estimator. This paper introduces a new robust estimator which is consistent and robust.

The general procedure of our proposed estimation method is following:

(1) Divide the field site into several subregions based on all available information. (2) In each subregion, macroscopic plume behaviors (or deterministic transport components) are estimated from the field data. These estimated values are subtracted from the field data to obtain residuals. (3) Based on complexities of spatial distribution of the residuals, divide each subregion into

several small blocks. (4) In each block, calculate experimental variograms using a robust estimator and determine mathematical variogram models. (5) Perform kriging to estimate residual at each desired location. (6) Finally, combine kriged residual values with the estimated macroscopic transport components.

The purpose of this paper are to provide a systematic methodology for estimating *in situ* groundwater contaminant concentrations, to introduce the $\mathcal{R}_p$-*estimator* for producing correlations between data points characterized by large variance, and to address some of mathematical problems related to estimation of goundwater contaminants. The method can be used generally to estimate space and time dependent geophysical, chemical and biological parameters; thus it may be useful to those developing numerical models for capturing the main feature of the groundwater contaminant distributions.

2 ESTIMATION OF MACROSCOPIC PLUME BEHAVIOR

As the first step for estimating the global plume behavior represented in the field data, the field site is divided into several subregions. Adjacent subregions may be overlapped to obtain spatially continuous information. The size of each subregion strongly depends on the geological structure of the field site and the global characteristics of data distributions. Each region showing distinctive distribution behavior is contained in a separated subregion. All available field data are visualized and analyzed. Also, any information related to the field site such as geological aquifer history are incorporated.

The macroscopic behavior is a spatially continuous large scale behavior describing the main profile of the plume movement. This step is difficult because the parameters in the transport equations, such as, dispersion/diffusion and seepage velocity share a highly nonlinear interdependence in space and time. Also, the measured contaminant concentration themselves add uncertainties due to spatial variabilities and unequal analytical confidence intervals. The criteria on how to choose basis functions and some specific approximation functions to estimate the macroscopic plume behavior are proposed. The approximating functions are chosen based on the global characteristics of the solute transport process in porous media. Nonlinear optimization techniques are needed for estimating parameters appeared in the approximating basis functions.

The basis functions are chosen using the following criteria:

(1) "Simple" functions are preferred because they are easily evaluated. At the same time, a "low order" approximation should capture the main profile of the plume movement. Here, the order of the approximation refers to the number of basis functions needed for the approximation.

(2) The approximation functions must capture the global plume behavior outside sampling network. In many practical applications, sampling networks often don't cover the entire extent of the contaminant plume; however, extrapolation of plume movement outside the sampling network is often desired.

(3) Basis functions should be "robust" in the sense that they are not very sensitive to unevenly spaced data points.

According to the selection criteria described above, specific basis functions are chosen based on the available field data and the process to be described (e.g., transport of a tracer or the spatial distribution of aquifer permeability). Focusing on the problem at hand, advective, dispersive/diffusive solute transport in porous media depends on space, time, and solute concentration. If, in addition, recharge, chemical, biological, and other reactive processes are considered, then the solute transport may be approximated [8], in Cartesian coordinates, using

$$\frac{\partial C}{\partial t} = \nabla \cdot (D \cdot \nabla C) - \nabla \cdot (VC) + f \tag{2.1}$$

with appropriate initial and boundary conditions, where $C = C(x, y, z; t)$ is the concentration of the solute, i.e., the mass of solute per unit volume of fluid, $D = D(x, y, z; t; C)$ is the dispersion tensor, $V = V(x, y, z; t; C)$ is the pore water velocity vector, $f = f(x, y, z; t; C)$ is a "forcing function" related to recharge, chemical, and biological activities. Note that the coefficients D and V depend on space, time and concentration itself.

As a simple case, assume that the porous medium is homogeneous, isotropic, saturated, the flow is steady-state, and that there is no external source. Then the transport equation (2.1) can be simplified as

$$\frac{\partial C}{\partial t} = \left[D_x \frac{\partial^2 C}{\partial x^2} + D_y \frac{\partial^2 C}{\partial y^2} + D_z \frac{\partial^2 C}{\partial z^2} \right] - \left[\bar{v}_x \frac{\partial C}{\partial x} + \bar{v}_y \frac{\partial C}{\partial y} + \bar{v}_z \frac{\partial C}{\partial z} \right], \tag{2.2}$$

where D_x, D_y and D_z are dispersion coefficients in the x, y and z-directions, $\bar{v}_x$, $\bar{v}_y$ and $\bar{v}_z$ are the average linear pore water velocities in each coordinate direction defined by $\bar{v}_x = v_x/\phi$, $\bar{v}_y = v_y/\phi$, $\bar{v}_z = v_z/\phi$, in which v_x, v_y, and v_z are specific discharge components, and ϕ is the porosity of the medium. If a contaminant is released instantaneously at the origin $(x, y, z) = (0, 0, 0)$, the mass distribution of the contaminant at time t is given by

$$C(x, y, z; t) = \frac{M}{8(\pi t)^{3/2} \phi \sqrt{D_x D_y D_z}} \exp \left(-\frac{\bar{X}^2}{4D_x t} - \frac{\bar{Y}^2}{4D_y t} - \frac{\bar{Z}^2}{4D_z t} \right), \tag{2.3}$$

where M is the mass of contaminant introduced at the point source, $\bar{X} = x - \bar{v}_x t$, $\bar{Y} = y - \bar{v}_y t$ and $\bar{Z} = z - \bar{v}_z t$ [8]. The averaged pore water velocities $\bar{v}_x$, $\bar{v}_y$ and $\bar{v}_z$ contribute movement of the center of mass of the contaminant plume (the propagation process). D_x, D_y, and D_z contribute to the longitudinal and transverse spreading of the plume around the plume centroid (the dispersion/diffusion process). The solution (2.3) of the ideal equation (2.2) is a simple representation of these two processes throughout two parameter sets $V = (\bar{v}_x, \bar{v}_y, \bar{v}_z)$ and $D = (D_x, D_y, D_z)$.

To approximate the gross distribution of contaminant concentrations in space, we propose the following linear combination of exponential functions.

$$F(x, y, z; a, b, c) = \sum_{i=1}^{m} c_i \exp \left(- \left(\frac{x - a_i^x}{b_i^x} \right)^2 - \left(\frac{y - a_i^y}{b_i^y} \right)^2 - \left(\frac{z - a_i^z}{b_i^z} \right)^2 \right), \tag{2.4}$$

where m is the number of basis functions, a_i^x, a_i^y, a_i^z, b_i^x, b_i^y, b_i^z, and c_i, $1 \leq i \leq m$, are parameters to be determined. In the same context as transport equations (2.1) and (2.2), the parameter set $a = \{(a_i^x, a_i^y, a_i^z) : 1 \leq i \leq m\}$ represents the propagation or advection process, the set $b = \{(b_i^x, b_i^y, b_i^z) : 1 \leq i \leq m\}$ represents the dispersion/diffusion process, and the set $c = \{c_i : 1 \leq i \leq m\}$ is related to the magnitude of the source load.

3 $\mathcal{R}_P$-ESTIMATOR

The residuals are obtained by subtracting the macroscopic transport portion from field data. The experimental semivariogram or variogram is used to describe the pattern of spatial correlation displayed by the residuals. A mathematical model is fitted to this experimental variogram, and this model is used in kriging to estimate the residuals at unmeasured locations. Some of mathematical models commonly used in practice can be found in [10]. In this section, we introduce the "$\mathcal{R}_p$-*estimator*," where $\mathcal{R}$ stands for "robust" and $p > 0$ indicates the order of robustness. For $0 < p \leq 1$, the estimator is robust; whereas for $p > 1$ the estimator becomes sensitive to apparent outliers. The estimator satisfies the following properties:

(1) It is consistent such that spatial correlations among the original data are preserved under linear transformation. Thus, the original data structure, estimation of correlations, and kriging procedures are consistent. (2) It is robust. It reduces outlier effects on estimated correlations between data points. (3) It is systematic. Depending on the distribution of the data, the order p of robustness can be adjusted systematically.

Let $Z(\mathbf{x})$ be a regionalized function on a domain Ω in three dimiensional space, and $Z(\mathbf{x}_i)$ be the realization of the function $Z(\mathbf{x})$ at $\mathbf{x}_i = (x_i, y_i, z_i) \in \Omega$, $i = 1, 2, \cdots, n$. Let $p > 0$ be a positive real number. For any vector $\mathbf{h} = (h_x, h_y, h_z)$, we define the $\mathcal{R}_p$-*estimator* as

$$\mathcal{R}_p(\mathbf{h}) = \left[\frac{1}{n(\mathbf{h})} \sum_{i=1}^{n(\mathbf{h})} |Z(\mathbf{x}_i) - Z(\mathbf{x}_i + \mathbf{h})|^p \right]^{\frac{1}{p}}, \tag{3.1}$$

where $n(\mathbf{h})$ is the number of data pairs separated by the vector $\mathbf{h}$.

For any positive integer k, and for any positive real number $p > 0$, the following inequalities hold:

$$\begin{aligned} a_1^p + a_2^p + \cdots + a_k^p &\leq (a_1 + a_2 + \cdots + a_k)^p, \quad p \geq 1, \\ a_1^p + a_2^p + \cdots + a_k^p &\geq (a_1 + a_2 + \cdots + a_k)^p, \quad 0 < p \leq 1, \end{aligned} \tag{3.2}$$

where $a_i \geq 0$, $i = 1, 2, \cdots, k$. Thus, the function

$$(a_1, a_2, \cdots, a_k) \mapsto \frac{a_1^p + a_2^p + \cdots + a_k^p}{k} \tag{3.3}$$

is a convex function for $p \geq 1$ and a concave function for $0 < p \leq 1$. As $p > 0$ approaches 0, robust effects are increased and the estimator $\mathcal{R}_p$ defined by equation (3.1) reduces effects of outliers for $0 < p \leq 1$.

Moreover, it is easy to see that, for any positive constant $c > 0$,

$$\left[\frac{1}{n(\mathbf{h})}\sum_{i=1}^{n(\mathbf{h})}|cZ(\mathbf{x}_i) - cZ(\mathbf{x}_i+\mathbf{h})|^p\right]^{\frac{1}{p}} = c\left[\frac{1}{n(\mathbf{h})}\sum_{i=1}^{n(\mathbf{h})}|Z(\mathbf{x}_i) - Z(\mathbf{x}_i+\mathbf{h})|^p\right]^{\frac{1}{p}}, \tag{3.4}$$

and, hence, the estimator $\mathcal{R}_p$ preserves any scaling factor. Therefore, if the original data set has a large range of values, then the range can be scaled by multiplying by a fixed positive constant without destroying any correlation structure found within the original data. The Cressie-Hawkins robust estimator [6]

$$\gamma_{ch}(\mathbf{h}) = \frac{1}{2}\frac{\left[\frac{1}{n(\mathbf{h})}\sum_{i=1}^{n(\mathbf{h})}|Z(\mathbf{x}_i) - Z(\mathbf{x}_i+\mathbf{h})|^{\frac{1}{2}}\right]^4}{[0.457 + 0.494/n(\mathbf{h})]} \tag{3.5}$$

which is commonly used in practice, the squared median of the absolute deviations estimator [7] $\gamma_{smad}(\mathbf{h}) = 2.198 \times [median\,|Z(\mathbf{x}_i) - Z(\mathbf{x}_i+\mathbf{h})|]^2$, and the conventional semivariogram [10]

$$\gamma(\mathbf{h}) = \frac{1}{2\,n(\mathbf{h})}\sum_{i=1}^{n(\mathbf{h})}|Z(\mathbf{x}_i) - Z(\mathbf{x}_i+\mathbf{h})|^2 \tag{3.6}$$

are essentially similar to the $\mathcal{R}_p$-estimator with $p = 1/2$, $p = 1$, and $p = 2$, respectively. However, the semivariogram γ is not robust. The influence of outliers on the semivariogram γ increases by the square $|Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})|^2$ as the difference $|Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})|$ increases. The Cressie-Hawkins estimator γ_{ch} and the squared median of the absolute deviations estimator γ_{smad} are not robust enough so that they do not produce correct correlation between data points showing erratic behaviors which are commonly observed in field data. Moreover, they do not preserve scaling factors and are not systematic.

4 KRIGING

Kriging is to estimate variables at unmeasured locations. It uses the mathematical model variograms fit to experimental variograms. Many kriging methods are available. Among them, the universal kriging and the punctual kriging are simple and can be easily implemented. In this section, the punctual kriging is explained when the experimental variograms are obtained by the $\mathcal{R}_p$-estimator. The application of the estimator to the universal kriging can be done in a similar way.

For each i, $i = 1, 2, \cdots, m$, let $Z(\mathbf{x}_i)$ be a given value at location $\mathbf{x}_i = (x_i, y_i, z_i)$ that is selected for kriging. For a given location $\mathbf{x}_o = (x_o, y_o, z_o)$, assume that the value $Z(\mathbf{x}_o)$ at $\mathbf{x}_o$ can be approximated by a linear sum of known values $Z(\mathbf{x}_i)$, $i = 1, \cdots, m$. Let

$$Z(\mathbf{x}_o) = \sum_{i=1}^{m}\omega_i Z(\mathbf{x}_i), \tag{4.1}$$

where $\omega_i \geq 0$, $i = 1, \cdots, m$, are weights to be determined by the following kriging system:

$$\begin{array}{rcll} \sum_{j=1}^{m} \omega_j \mathcal{R}_p(h_{ij}) + \lambda & = & \mathcal{R}_p(h_{io}), & 1 \leq i \leq m, \\ \sum_{i=1}^{m} \omega_i & = & 1, & \end{array} \tag{4.2}$$

where $\mathcal{R}_p(h_{ij})$ is the correlation value estimated by the $\mathcal{R}_p$-estimator at lag h_{ij}, the subscript p is the order of robustness, h_{ij} is the "correlation lag" between two points $\mathbf{x}_i$ and $\mathbf{x}_j$, λ is the Lagrange multiplier, and $\sum_{i=1}^{m} \omega_i = 1$ is the optimality condition. Note that this kriging system (4.2) is "optimal" in the sense that the method produces the exact (original) value at the sampled location. However, the kriging system is optimal only inside the sampling network (convex) domain. Thus, the estimation procedure for points outside the sampling domain must consider macroscopic properties such as trend, drift, etc., of the original data structure together with the kriging system because the optimality condition in equation (4.2) is no longer valid outside the (convex) domain. The optimality constraint described above is independent of the choice of variogram; $\mathcal{R}_p$-estimator, semivariogram γ in equation (3.6), or any other estimators.

With regard to the scaling factor, for each $p > 0$,

$$\mathcal{R}_p(ch) = c\mathcal{R}_p(h) \tag{4.3}$$

for any $c > 0$ and lag h. Thus, the scaling factor $c > 0$ of the original data set is preserved in the correlation estimation step. Moreover, for any constant $c > 0$, the following two kriging systems:

$$\begin{array}{rcll} \sum_{j=1}^{m} \omega_j (c\mathcal{R}_p(h_{ij})) + \lambda & = & c\mathcal{R}_p(h_{io}), & 1 \leq i \leq m, \\ \sum_{i=1}^{m} \omega_i & = & 1, & \end{array} \tag{4.4}$$

and

$$\begin{array}{rcll} \sum_{j=1}^{m} \omega_j \mathcal{R}_p(h_{ij}) + \lambda/c & = & \mathcal{R}_p(h_{io}), & 1 \leq i \leq m, \\ \sum_{i=1}^{m} \omega_i & = & 1 & \end{array} \tag{4.5}$$

are equivalent.

5 ACKNOWLEDGMENTS

This paper has not been subject to USAF review, however, and accordingly does not necessarily reflect the reviews of the USAF. The research of the first author was supported in part by the KOSEF 951-0104-024-2 and the BSRI-97-1426.

References

[1] Adams, E. E. and L. W. Gelhar (1992), Field study of dispersion in a heterogeneous aquifer, 2. Spatial moments analysis, *Water Resour. Res.*, 28(12): 3293-3307.

[2] ASCE Task Committee on Geostatistical Techniques in Geohydrology of the Ground Water Hydrology, Committee of the ASCE Hydraulics Division (1990), Review of geostatistics in geohydrology. I: Basic concepts; II. Applications, *J. Hydraulic Engineering*, 116(5): 612-632; 633-658.

[3] Boggs, J. M., L. M. Beard, W. R. Waldrop, T. B. Stauffer, W. G. MacIntyre and C. P. Antworth (1993), 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.

[4] Boggs, J. M., S. C. Young, L. M. Beard, L. W. Gelhar, K. R. Rehfeldt and E. E. Adams (1992), Field study of dispersion in a heterogeneous aquifer, 1. Overview and site description, *Water Resour. Res.*, 28(12): 3281-3291.

[5] Cooper, R. M. and J. D. Istok (1988), Geostatistics applied to groundwater contamination. I: Methodology; II: Application; III: Global estimates, *J. Environmental Engineering*, 114(2): 270-286; 287-299; 114(4): 915-928.

[6] Cressie, N. and D. M. Hawkins (1980), Robust estimation of the variogram: I, *Mathematical Geology*, 12(2): 115-125.

[7] Dowd, P. A. (1984), The variogram and kriging: Robust and resistant estimators, Geostatistics for Natural Resourses Characterization, Part I, *NATO ASI Ser., Ser. C*, 12: 91-107.

[8] Freeze, R. A. and J. A. Cherry (1979), *Groundwater*, Prentice-Hall, Englewood Cliffs, NJ.

[9] Garabedian, S. P., D. R. Leblanc, L. W. Gelhar, and M. A. Celia (1991), Large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachusetts, 2. Analysis of spatial moments for a nonreactive tracer, *Water Resour. Res.*, 27(5): 911-924.

[10] Journel, A. G. and J. C. Huijbregts (1991), *Mining Geostatistics, 5th ed.*,=20 Academic Press, San Diego.

[11] Leblanc, D. R., S. P. Garabedian, K. M. Hess, L. W. Gelhar, R. D. Quardri, K. G. Stollenwerk, and W. W. Wood (1991), Large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachussetts, 1. Experimental design and observed tracer movement, *Water Resour. Res.*, 27(5): 895-910.

[12] Mackay, D. M., D. L. Freyberg, P. V. Roberts, and J. A. Cherry (1986), A natural gradient experiment on solute transport in a sand aquifer, 1. Approach and overview of plume movement, *Water Resour. Res.*, 22(13): 2017-2079.