WATER RESOURCES RESEARCH, VOL. 33, NO. 4, PAGES 537–546, APRIL 1997

# A geostatistical approach to contaminant source identification

Mark F. Snodgrass and Peter K. Kitanidis
Department of Civil Engineering, Stanford University, Stanford, California

**Abstract.** A geostatistical approach to contaminant source estimation is presented. The problem is to estimate the release history of a conservative solute given point concentration measurements at some time after the release. A Bayesian framework is followed to derive the best estimate and to quantify the estimation error. The relation between this approach and common regularization and interpolation schemes is discussed. The performance of the method is demonstrated for transport in a simple one-dimensional homogeneous medium, although the approach is directly applicable to transport in two- or three-dimensional domains. The methodology produces a best estimate of the release history and a confidence interval. Conditional realizations of the release history are generated that are useful in visualization and risk assessment. The performance of the method with sparse data and large measurement error is examined. Emphasis is placed on formulating the estimation method in a computationally efficient manner. The method does not require the inversion of matrices whose size depends on the grid size used to resolve the solute release history. The issue of model validation is addressed.

## Introduction

In recent years, hydrogeologists have placed increasing emphasis on modeling contaminant transport in the subsurface. One important aspect of this problem is identifying the source of the contamination. For a variety of legal and regulatory reasons it is often necessary to determine over what time period and at what concentration contaminants were actually released. The release history needs to be inferred from a limited number of point concentration measurements. Over the years a variety of methods have been developed to solve this problem. *Skaggs and Kabala* [1994] used Tikhonov regularization to transform the ill-posed (i.e., algebraically underdetermined) inverse problem into a minimization problem with a unique solution. *Skaggs and Kabala* [1995] applied the method of quasi-reversibility to solve the problem. *Gorelick et al*. [1983] used linear programming to estimate the location and magnitude of steady state tracer injections. *Bagtzoglou et al*. [1992] used particle tracking in reverse time to estimate a probability distribution of the original solute position. *Wagner* [1992] developed a deterministic approach for the combined estimation of model parameters and the solute source characteristics. A brief summary of these methodologies and others is given by *Skaggs and Kabala* [1994].

To date, however, Bayesian analysis has not been applied to the problem of pollutant source characterization as it has been to other aspects of hydrogeology. Hydrogeologists have used Bayesian theory and geostatistical techniques to estimate hydraulic head and conductivity fields for many years [e.g., *Hoeksema and Kitanidis*, 1984; *Dagan*, 1986; *Kitanidis*, 1986, 1995; *Tarantola*, 1987], but they have overlooked the applicability of these techniques to the problem of source identification. There are many advantages to developing a method of contaminant source identification within the geostatistical and Bayesian framework.

Copyright 1997 by the American Geophysical Union.

Paper number 96WR03753.
0043-1397/97/96WR-03753$09.00

There will always be uncertainty in contaminant concentration estimates and release history. It makes sense therefore to treat these quantities as random functions that can be described by their statistical properties. In this framework, estimation uncertainty is recognized and its importance can be determined. Bayesian analysis provides for the quantification of estimation error and aids in the evaluation of the sources of uncertainty. A major advantage of the geostatistical approach is that it allows for the validation and improvement of the chosen model. The proposed method is very general and in fact includes Tikhonov regularization and many common interpolation schemes as special cases. Often in practice, one of these interpolation schemes is adhered to blindly without regard to whether or not it is appropriate for the problem at hand. In contrast, the geostatistical approach allows us to choose the scheme that is most consistent with the data and other information from a virtually infinite list including, but not limited to, the aforementioned ones. Model parameters may be fitted to the data in a systematic fashion. The procedure for checking whether the model is adequate or should be improved is called model validation or model criticism.

This paper presents the methodology and demonstrates its versatility. The first section reviews the algorithm used to find the best estimate following the geostatistical approach, followed by a brief discussion of model validation. The geostatistical approach is then applied to estimating the release history of a conservative solute. Conditional realizations of possible solutions are also generated to show what the actual solution might look like given the available information. The performance of the method in an application with sparse data and large measurement error is demonstrated. The most computationally efficient forms of the equations are presented because in the end, a method is practical only if it is cost effective to employ.

## General Methodology

The objective is to estimate an unknown function, in this case the release history, which is a function of time. For con-

venience the unknown function is represented as a random process because there is uncertainty associated with the function and its true value may never be found. However, we can imagine the set of all possible functions that fit the data and are consistent with additional information. In a Bayesian approach, each of these functions is assigned a probability that it is the solution. The expected value of this set is sought as a best estimate along with its covariance as a measure of the estimation uncertainty.

The standard estimation problem may be expressed in the following form:

$$\mathbf{z} = \mathbf{h}(\mathbf{s}, \mathbf{r}) + \mathbf{v} \tag{1}$$

where $\mathbf{z}$ is an $m \times 1$ vector of observations and $\mathbf{s}$ is an $n \times 1$ "state vector" obtained from the discretization of the unknown function that we wish to estimate. The vector $\mathbf{r}$ contains other parameters needed by the model function $\mathbf{h}(\mathbf{s}, \mathbf{r})$. For example, $\mathbf{r}$ might include unknown aquifer parameters such as the velocity or dispersivity. In practice, these parameters may also need to be estimated, but in this work we will assume they are known perfectly. Thus $\mathbf{h}(\mathbf{s}, \mathbf{r})$ reduces to $\mathbf{h}(\mathbf{s})$. The measurement error is represented by the vector $\mathbf{v}$. Following geostatistical methodology, $\mathbf{s}$ and $\mathbf{v}$ are represented as random vectors. We shall assume that $\mathbf{v}$ has zero mean and known covariance matrix $\mathbf{R}$. Furthermore, we will model $\mathbf{s}$, the unknown, as a random vector with expected value

$$E[\mathbf{s}] = \mathbf{X}\beta \tag{2}$$

where $\mathbf{X}$ is a known $n \times p$ matrix and $\beta$ are $p$ unknown drift coefficients. The covariance of $\mathbf{s}$ is

$$\mathbf{Q}(\theta) = E[(\mathbf{s} - \mathbf{X}\boldsymbol{\beta})(\mathbf{s} - \mathbf{X}\boldsymbol{\beta})^T]$$

where $\mathbf{Q}(\theta)$ is a known function of unknown parameters $\theta$. A simplification occurs if the function $\mathbf{h}(\mathbf{s})$ is linear with respect to $\mathbf{s}$ or if we can linearize the function. In fact, for conservative solute transport, the relation between the solute observations $\mathbf{z}$ and the solute input $\mathbf{s}$ is linear, so that we can rewrite equation (1) as

$$\mathbf{z} = \mathbf{H}\mathbf{s} + \mathbf{v} \tag{3}$$

where $\mathbf{H}$ is a known matrix. We now wish to find the estimate of $\mathbf{s}$ that maximizes its posterior probability. The approach used is detailed by *Kitanidis* [1995] and reviewed below.

The estimation procedure is divided into two parts. First the optimal structural parameters are found, and then the unknown function is estimated. The structural parameters $\theta$ are estimated by maximizing the probability of the measurements given $\theta$:

$$p(\mathbf{z}|\theta) \propto |\Sigma|^{-1/2}|\mathbf{X}^T\mathbf{H}^T\Sigma^{-1}\mathbf{H}\mathbf{X}|^{-1/2} \exp\left[-\tfrac{1}{2}\mathbf{z}^T\Xi^{-l}\mathbf{z}\right] \tag{4}$$

where

$$\Sigma = \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R} \tag{5}$$

$$\Xi = \Sigma^{-1} - \Sigma^{-1}\mathbf{H}\mathbf{X}(\mathbf{X}^T\mathbf{H}^T\Sigma^{-1}\mathbf{H}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}^T\Sigma^{-1} \tag{6}$$

Maximizing $p(\mathbf{z}|\theta)$ is equivalent to minimizing

$$L(\theta) = \tfrac{1}{2}\ln|\Sigma| + \tfrac{1}{2}\ln|\mathbf{X}^T\mathbf{H}^T\Sigma^{-1}\mathbf{H}\mathbf{X}| + \tfrac{1}{2}\mathbf{z}^T\Xi^{-1}\mathbf{z} \tag{7}$$

Note that the drift parameters have been eliminated from the analysis by integrating over all $\beta$ [see *Kitanidis*, 1995]. This process removes the bias caused by the unknown coefficients $\beta$. This minimization can be achieved by taking derivatives of $L(\theta)$ with respect to $\theta$ and setting them to zero. Define

$$g_i = \frac{\partial L}{\partial \theta_i} = \frac{1}{2}\,tr\left(\Xi\,\frac{\partial\Sigma}{\partial\theta_i}\right) - \frac{1}{2}\,\mathbf{z}^T\left[\Xi\,\frac{\partial\Sigma}{\partial\theta_i}\,\Xi\right]\mathbf{z} \tag{8}$$

where the $i$th element of $\theta$ is $\theta_i$.

These equations can be solved numerically. The recommended procedure is to apply Gauss-Newton iterations. Form the Fisher information matrix

$$F_{ij} = \frac{1}{2}\,tr\left[\Xi\,\frac{\partial\Sigma}{\partial\theta_i}\,\Xi\,\frac{\partial\Sigma}{\partial\theta_j}\right] \tag{9}$$

and update the previous estimate of $\theta$,

$$\theta_{l+1} = \theta_l - \mathbf{F}^{-1}\mathbf{g} \tag{10}$$

For the examples that will be discussed below, it was necessary to apply the Marquardt modification by replacing $\mathbf{F}^{-1}$ with $(\mathbf{F}^{-1} + \lambda\mathbf{I})$, where $\mathbf{I}$ is the identity matrix and $\lambda$ is a positive parameter. The mean square estimation error of the structural parameters can by approximated using $\mathbf{F}^{-1}$ and the Cramer-Rao inequality.

Once the iterations have converged, form and solve the system

$$\begin{bmatrix} \Sigma & \vdots & \mathbf{H}\mathbf{X} \\ \cdots & & \cdots \\ (\mathbf{H}\mathbf{X})^T & \vdots & 0 \end{bmatrix}\begin{bmatrix} \Lambda^T \\ \cdots \\ \mathbf{M} \end{bmatrix} = \begin{bmatrix} \mathbf{H}\mathbf{Q} \\ \cdots \\ \mathbf{X}^T \end{bmatrix} \tag{11}$$

where $\Lambda$ is a $m \times n$ matrix of coefficients and $\mathbf{M}$ is $p \times n$ matrix of multipliers. The best estimate of the function is

$$\hat{\mathbf{s}} = \boldsymbol{\Lambda}\mathbf{z} \tag{12}$$

and its covariance is

$$\mathbf{V} = -\mathbf{X}\mathbf{M} + \mathbf{Q} - \mathbf{Q}\mathbf{H}^{\mathbf{T}}\Lambda^T \tag{13}$$

It should be noted that the solution does not require the inversion of $\mathbf{Q}$ or any matrix with dimensions which depend on the number of grid points. The number of operations needed to calculate $\hat{\mathbf{s}}$, $\mathbf{V}$ and $\partial L/\partial\theta_i$ is on the order of $n^2m$.

This method is elegant but does not enforce the nonnegativity of concentration. If this constraint is imposed, the computational time increases. However, a solution can be found by working in a transformed space and solving the equations iteratively. This method will be developed below.

Instead of using concentration we will use a transformation of the concentration [*Box and Cox*, 1964]. Define

$$\tilde{\mathbf{s}} = \alpha(\mathbf{s}^{1/\alpha} - 1) \tag{14}$$

and the inverse operation

$$\mathbf{s} = \left(\frac{\tilde{\mathbf{s}} + \alpha}{\alpha}\right)^{\alpha} \tag{15}$$

where $\alpha$ is a positive number. The parameter $\alpha$ may be chosen as small as possible while ensuring that $\tilde{\mathbf{s}} > -\alpha$. The exact value of $\alpha$ used has little impact on the method's overall performance. (A method for the optimization of parameter $\alpha$ is given by *Kitanidis and Shen* [1996].) However, smaller values of $\alpha$ cause less drastic transformations, which tends to help the method converge slightly more rapidly. When in doubt, a very large $\alpha$ value may be used without fear of adversely affecting the solution. Notice that at the limit

$$\lim_{a\to\infty}(\alpha(\mathbf{s}^{1/\alpha} - 1)) = \ln(\mathbf{s})$$

The logarithm transformation could be used; however, it encounters difficulty when the concentration is zero. As a result of this transformation, $\mathbf{s}$ is constrained to be nonnegative and the relationship between $\tilde{\mathbf{s}}$ and $\mathbf{s}$ is one to one. It should be further noted that the claim is not that this transformation is inherently the best, but only that it is a useful mathematical tool for enforcing the nonnegativity constraint, thus solving the problem at hand. Thus equation (1) in the transformed space becomes

$$\mathbf{z} = \tilde{\mathbf{h}}(\tilde{\mathbf{s}}) + \mathbf{v} \tag{16}$$

where $\tilde{\mathbf{h}}(\tilde{\mathbf{s}}) = \mathbf{h}[((\tilde{\mathbf{s}} + \alpha)/\alpha)^{\alpha}]$.

In this case the transfer function $\tilde{\mathbf{h}}(\tilde{\mathbf{s}})$ is not linear with respect to the transformed unknown $\tilde{\mathbf{s}}$. For this reason the solution is arrived at iteratively. The best estimate can be found by following the quasi-linear procedure detailed by *Kitanidis* [1995]. To summarize, the algorithm is as follows.

Make an initial estimate of the unknown $\tilde{\mathbf{s}}_l$. Find the derivative of $\tilde{\mathbf{h}}$ with respect to $\tilde{\mathbf{s}}$ at $\tilde{\mathbf{s}}_l$:

$$\tilde{\mathbf{H}}_l = \left.\frac{\partial \tilde{\mathbf{h}}}{\partial \tilde{\mathbf{s}}}\right|_{\tilde{\mathbf{s}}=\tilde{\mathbf{s}}_l} \tag{17}$$

Define

$$\Sigma = \tilde{\mathbf{H}}_l \mathbf{Q} \tilde{\mathbf{H}}_l^T + \mathbf{R} \tag{18}$$

$$\mathbf{z}_{0l} = \mathbf{z} - \tilde{\mathbf{h}}(\tilde{\mathbf{s}}_l) + \tilde{\mathbf{H}}_l \tilde{\mathbf{s}}_l \tag{19}$$

Solve

$$\begin{bmatrix} \Sigma & \vdots & \tilde{\mathbf{H}}_l \mathbf{X} \\ \cdots & & \cdots \\ (\tilde{\mathbf{H}}_l \mathbf{X})^T & \vdots & 0 \end{bmatrix} \begin{bmatrix} \Lambda^T \\ \cdots \\ \mathbf{M} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{H}}_l \mathbf{Q} \\ \cdots \\ \mathbf{X}^T \end{bmatrix} \tag{20}$$

where $\Lambda$ is a $m \times n$ matrix of coefficients and $\mathbf{M}$ is $p \times n$ matrix of multipliers. Then

$$\tilde{\mathbf{s}}_{l+1} = \Lambda \mathbf{z}_{0l} \tag{21}$$

The structural parameters $\theta$ are found by minimizing

$$L(\theta) = \tfrac{1}{2} \ln |\Sigma| + \tfrac{1}{2} \ln |\mathbf{X}^T \tilde{\mathbf{H}}_l^T \Sigma^{-1} \tilde{\mathbf{H}}_l \mathbf{X}| + \tfrac{1}{2} \mathbf{z}_{0l}^T \Xi^{-1} \mathbf{z}_{0l} \tag{22}$$

where

$$\Xi = \Sigma^{-1} - \Sigma^{-1} \tilde{\mathbf{H}}_l \mathbf{X} (\mathbf{X}^T \tilde{\mathbf{H}}_l^T \Sigma^{-1} \tilde{\mathbf{H}}_l \mathbf{X})^{-1} \mathbf{X}^T \tilde{\mathbf{H}}_l^T \Sigma^{-1} \tag{23}$$

In this step, $\mathbf{z}_{0l}$ and $\tilde{\mathbf{H}}_l$ are treated as constants. The minimization can be achieved efficiently by applying Gauss-Newton iterations as described above. Once $\theta$ is found, it is used to find an updated value of $\tilde{\mathbf{s}}_l$. This two-step iterative process is continued until $\tilde{\mathbf{s}}_l$ and $\theta$ converge. The final value of $\tilde{\mathbf{s}}_l$ is the best estimate of the transformed function, and its covariance is given by

$$\mathbf{V} = -\mathbf{X}\mathbf{M} + \mathbf{Q} - \mathbf{Q}\mathbf{H}^{\mathbf{T}} \mathbf{\Lambda}^T \tag{24}$$

The best estimate of the unknown function is

$$\hat{\mathbf{s}} = \left( \frac{\tilde{\mathbf{s}}_l + \alpha}{\alpha} \right)^{\alpha} \tag{25}$$

The use of this transformed variable has the advantage of returning a physically possible solution; however, the computational time is greater depending on the number of iterations needed to converge to a solution. Fortunately, convergence is generally rapid for the problem at hand.

## Model Validation

One of the most important steps in applying the geostatistical approach is to develop a model that describes the structure of the unknown function to be estimated. Like any function, a release history can be represented through a statistical structure. From the measurements we attempt to develop the best possible model of this structure that describes available information. Geostatistics gives us a methodology for developing a model and then checking it. Model development and validation are discussed by *Delfiner* [1976], *Dubrule* [1983], *Kitanidis and Vomvoris* [1983], *Davis* [1987], *Borgman* [1988], and *Kitanidis* [1988]. *Box and Jenkins* [1976] discuss the issue within the context of time series analysis, but the methodology they describe is just as applicable to hydrogeological problems.

The choice of a model is far from arbitrary. It is arrived at through careful and thoughtful analysis of the data and through validation. In estimation, one starts with the simplest model that is consistent with all available information. Two principles guide the selection of a model: "indifference" and "parsimony" [*Box and Jenkins*, 1976]. Indifference here means that if there is no information suggesting a trend, then a constant mean should be adopted. For example, in the problem discussed in this paper, there is no way to know if the release occurred as one catastrophic event, and if so, when that event took place, or if the release was slow and gradual. Given this lack of information the best choice is to assume a constant mean. This accurately reflects the state of our a priori knowledge. Parsimony means that the model should use as few unknown parameters as possible; that is, additional parameters should be introduced into the model only if they improve significantly the ability of the model to reproduce the data.

Once the first tentative model has been selected, it must be validated. This model validation process is discussed in a number of works [e.g., *Anscombe and Tukey*, 1963; *Andrews*, 1971; *Delfiner*, 1976; *Belsley et al.*, 1980; *Box* 1980; *Kitanidis*, 1988, 1991]. Model validation is of paramount importance, as it allows the practitioner to reexamine the model and provides an opportunity for improvements. This process is far superior to blind adherence to one given model, as in previous approaches to the source identification problem.

The purpose of this paper, however, is to demonstrate the applicability of the geostatistical approach to the problem of contaminant source identification. A full discussion of model validation goes beyond the scope of this work.

## Conditional Simulations

The preceding section outlined the methodology to find the best estimate of the process and the error. At times, however, it is useful to have an idea of how the actual process might look, given the available information. This can be achieved by generating realizations of the process that are conditional on all the observations. Viewing a number of conditional realizations can aid in assessing how accurate the best estimate may be. The procedure for generating conditional realizations is discussed by *Gutjahr et al.* [1994] and *Kitanidis* [1995]. To review, the approach is as follows. First, an unconditional realization $\mathbf{s}_{ui}$ is generated. A simple and general approach is to decompose the covariance matrix so that

$$\mathbf{Q} = \mathbf{C}\mathbf{C}^T \tag{26}$$

and then compute the vector

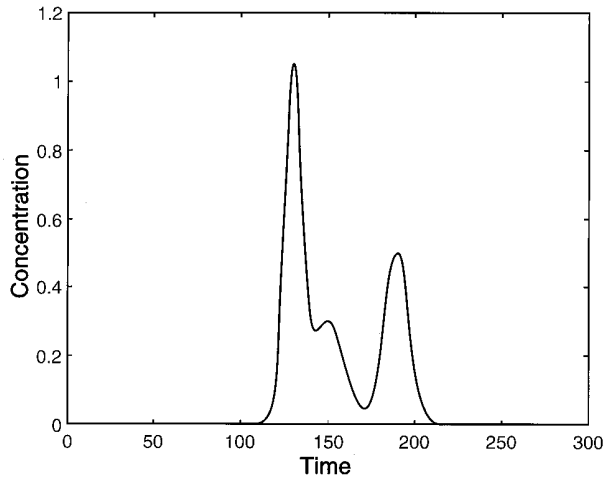


**Figure 1.** The contaminant release history.

$$\mathbf{s}_{ui} = \mathbf{C}\mathbf{u}_i \tag{27}$$

Here $\mathbf{u}_i$ is a vector of independent, identically distributed normal variants. A realization of the error vector $\mathbf{v}_i$ must also be independently generated with zero mean and covariance $\mathbf{R}$. Then the conditional realization $\mathbf{s}_c$ may be found by minimizing

$$(\mathbf{s}_c - \mathbf{s}_{ui})^T\mathbf{G}(\mathbf{s}_c - \mathbf{s}_{ui}) + (\mathbf{z} + \mathbf{v}_i - \mathbf{h}(\mathbf{s}_c))^T\mathbf{R}^{-1}(\mathbf{z} + \mathbf{v}_i - \mathbf{h}(\mathbf{s}_c)) \tag{28}$$

with respect to $\mathbf{s}_c$. Here

$$\mathbf{G} = \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}^{-1} \tag{29}$$

For the transformed case, $\tilde{\mathbf{h}}$ is used, and conditional realizations of the transformed function $\tilde{\mathbf{s}}$ are computed which are then back transformed. Note that this operation may be achieved without explicitly calculating the inverse of $\mathbf{Q}$. Decomposition of $\mathbf{Q}$, however, may take considerable computational effort, which is why in practice spectral decomposition methods, such as fast Fourier transforms, are preferred.

## Simulations and Results

As was stated in the introduction, the motivation behind this paper was to formulate the release history estimation work of *Skaggs and Kabala* [1994, 1995] and others within a geostatistical framework. The example problem we will examine is modeled after the example used by *Skaggs and Kabala* [1994, 1995]. Let us consider the advective and dispersive transport of a conservative solute in a one-dimensional (1-D) homogeneous aquifer, even though the estimation method described in the work applies to transport in any domain. Suppose the conservative solute is injected at the left boundary of the aquifer and at some later time $T$ the concentration is measured at various points in the aquifer. From this data we wish to estimate the amount of solute injected as a function of time. Figure 1 shows the true solution

$$s(t) = \exp\left(-\frac{(t-130)^2}{50}\right) + 0.3\exp\left(-\frac{(t-150)^2}{200}\right) + 0.5\exp\left(-\frac{(t-190)^2}{98}\right) \tag{30}$$

This release history is used to generate the contaminant plume that is measured. Figure 2 shows the plume after 300 time units and the location of the measurements. We will estimate the release history from the data using the two methods described above. First let us outline the specific example problem.

### Example Problem

In the 1-D case we have an analytical solution

$$c(x, T) = \int_0^T s(\tau) f(x, T-\tau)\, d\tau \tag{31}$$

where $c(x, T)$ is the concentration at distance $x$ from the source and time $T$. The source is a function of time and is expressed by $s(\tau)$. The transfer function $f(x, t-\tau)$ applies the appropriate weight to the source function [*Skaggs and Kabala*, 1994]:

$$f(x, t-\tau) = \frac{x}{2\sqrt{\pi D(t-\tau)^3}}\exp\left[-\frac{(x - v(t-\tau))^2}{4D(t-\tau)}\right] \tag{32}$$

The velocity is $v$, and the dispersion coefficient is $D$. Let $x_i$, $i = 1, \cdots, m$ be the $m$ points at which the measurements are taken (in this example, $m = 25$. Let us discretize the time domain into $n$ temporal points $t_i$, $i = 1, \cdots, n$, with a time step $\Delta t = T/n$. In this example, $\Delta t = 1$ and $n = T = 300$. Then we can set up the system as follows:

$$\mathbf{z} = \begin{bmatrix} c(x_1, T) \\ c(x_2, T) \\ \vdots \\ c(x_m, T) \end{bmatrix} + \mathbf{v},\ \mathbf{s} = \begin{bmatrix} s(t_1) \\ s(t_2) \\ \vdots \\ s(t_n) \end{bmatrix} \tag{33}$$

$$\mathbf{H} = \Delta t \begin{bmatrix} f(x_1, T-t_1) & \cdots & f(x_1, T-t_n) \\ f(x_2, T-t_1) & \cdots & f(x_2, T-t_n) \\ \vdots & \ddots & \vdots \\ f(x_m, T-t_1) & \cdots & f(x_m, T-t_n) \end{bmatrix} \tag{34}$$

Again, $\mathbf{z}$ is the vector of known observations and $\mathbf{s}$ is the unknown we wish to estimate. A realization of the estimation error $\mathbf{v}$ is generated to have zero mean and covariance matrix $\mathbf{R}(m \times m)$. The problem will be solved using the two methods developed in the preceding section. For this problem a constant, but unknown, mean is assumed. Thus

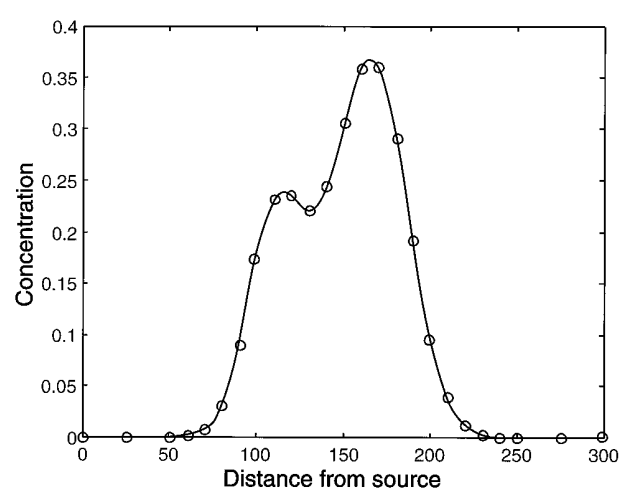


**Figure 2.** Contaminant plume after 30 time units, with measurement locations denoted by circles.

$$\mathbf{X} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{35}$$

and $\beta$ is the mean of the function, an unknown scalar. The assumed form of the mean function and the form of the covariance matrix are subject to validation as discussed above.

**Solution 1: Unconstrained Case**

In the first case the solution is not constrained to be nonnegative. The covariance of the measurement errors used was

$$\mathbf{R} = \sigma_R^2 \mathbf{I} \tag{36}$$

where $\sigma_R^2$ was assigned the value $1 \times 10^{-12}$. The covariance matrix $\mathbf{Q}$ is assumed to have Gaussian form.

$$Q(t_i, t_j|\theta) = \sigma^2 \exp\left(-\frac{(t_i - t_j)^2}{l^2}\right) \tag{37}$$

where $(t_i - t_j)$ is the separation distance (time units) and

$$\theta = \begin{bmatrix} \sigma^2 \\ l \end{bmatrix}.$$

This covariance function is used for illustration, though it may not be best, as will be shown later. The variance $\sigma^2$ and the characteristic time scale $l$ are estimated as described above. The best estimates are

$$\begin{aligned} \sigma^2 &= 0.046 \pm 0.0096 \\ l &= 13.24 \pm 1.2 \end{aligned} \tag{38}$$

The mean square errors of estimation are estimated using the Cramer-Rao inequality. Once these parameters are estimated, we solve for the best estimate. Figure 3 shows the best estimate and the approximate 95% confidence interval given by the covariance matrix $\mathbf{V}$, that is the bounds are given by

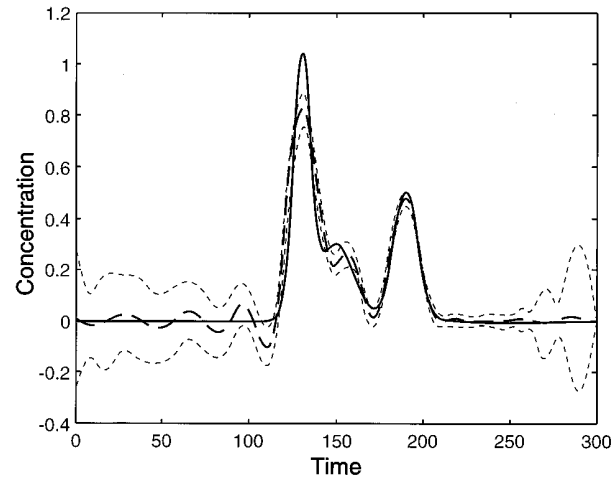


**Figure 3.** Unconstrained case with Gaussian covariance: the true solution (solid line), best estimate (dashed line), and approximate 95% confidence interval (dotted lines).

this problem and to a achieve a more accurate estimate, we use the second method described.

**Solution 2: Constrained Case**

The solution is constrained to be nonnegative by introducing a new variable:

$$\tilde{\mathbf{s}} = \alpha(\mathbf{s}^{1/\alpha} - 1) \tag{39}$$

and an initial guess of $\tilde{\mathbf{s}}$ is made (any reasonable guess will do). Define

$$\tilde{\mathbf{h}}(\tilde{\mathbf{s}}_0) = \int_0^t \left(\frac{\tilde{\mathbf{s}}_0 + \alpha}{\alpha}\right)^{\alpha} f(x, t - \tau)\, d\tau \tag{40}$$

and find the derivative of the transfer function at the initial guess, $\tilde{s}_0$.

$$\tilde{\mathbf{H}} = \Delta t \cdot \begin{bmatrix} \left(\frac{\tilde{\mathbf{s}}_0(t_1) + \alpha}{\alpha}\right)^{\alpha-1} f(x_1, T - t_1) & \cdots & \left(\frac{\tilde{\mathbf{s}}_0(t_n) + \alpha}{\alpha}\right)^{\alpha-1} f(x_1, T - t_n) \\ \left(\frac{\tilde{\mathbf{s}}_0(t_1) + \alpha}{\alpha}\right)^{\alpha-1} f(x_2, T - t_1) & \cdots & \left(\frac{\tilde{\mathbf{s}}_0(t_n) + \alpha}{\alpha}\right)^{\alpha-1} f(x_2, T - t_n) \\ \vdots & \ddots & \vdots \\ \left(\frac{\tilde{\mathbf{s}}_0(t_1) + \alpha}{\alpha}\right)^{\alpha-1} f(x_m, T - t_1) & \cdots & \left(\frac{\tilde{\mathbf{s}}_0(t_n) + \alpha}{\alpha}\right)^{\alpha-1} f(x_n, T - t_n) \end{bmatrix} \tag{41}$$

$$s_i \pm 2\sqrt{V_{ii}}$$

where $V_{ii}$ are the diagonal elements of $\mathbf{V}$. Five realizations conditioned on the data are shown in Figure 4. The uncertainty associated with the estimate can be seen by examining Figure 3. The confidence interval and the scatter of the realizations show the degree of error in the estimate. Also note that the solution oscillates to negative values of concentration. To solve

The best estimate of the function is found iteratively by applying the quasi-linear method [*Kitanidis*, 1995].

The covariance of the measurement errors used was

$$\mathbf{R} = \sigma_R^2 \mathbf{I} \tag{42}$$

where $\sigma_R^2$ was assigned the value $\sigma_R^2 = 1 \times 10^{-12}$. The results in Figure 5 were generated using a Gaussian covariance matrix and $\alpha = 4$. The covariance parameters were found to be

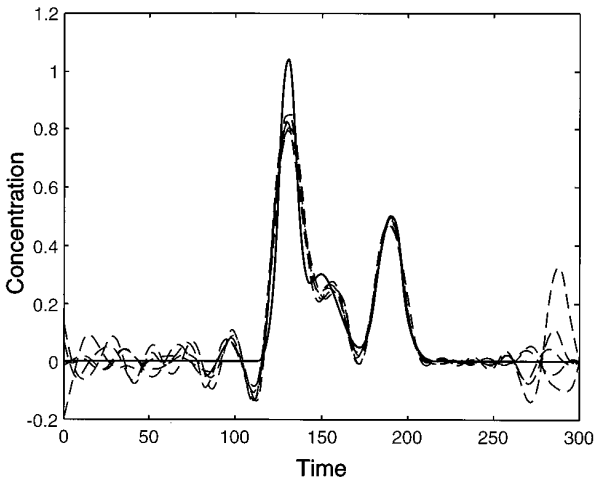


**Figure 4.** Unconstrained case with Gaussian covariance: the true solution (solid line) and five equally likely realizations (dashed lines).

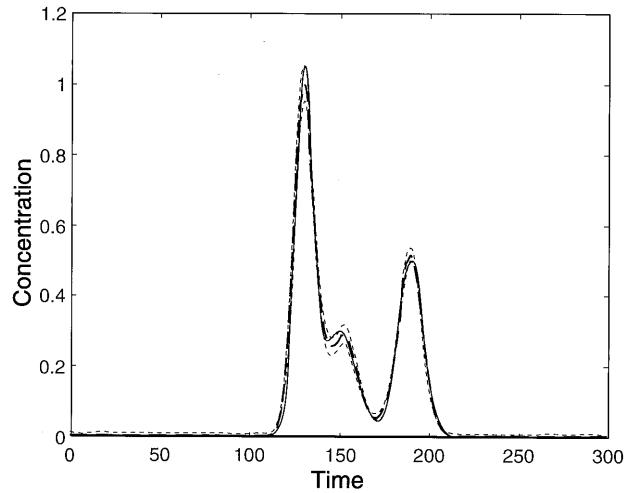


**Figure 5.** Constrained case with Gaussian covariance: the true solution (solid line), best estimate (dashed line), and approximate 95% confidence interval (dotted lines).

$$\sigma^2 = 1.82 \pm 2.86$$
$$l = 17.5 \pm 6.4 \tag{43}$$

Figure 6 shows five realizations. The accuracy of the estimate has increased significantly, as can be seen by the small error bound and the low degree of spread in the realizations, because we introduced in the analysis the information that concentration is nonnegative. Note here that the confidence intervals is approximated in the transformed space and the bounds are then calculated using the inverse transformation operator.

### Correspondence Between the Geostatistical Approach and Other Estimation Schemes

An important aspect of the geostatistical method applied here is its generality. By using certain generalized covariance functions, many common regularization and interpolation schemes may be duplicated. This equivalence between Bayesian theory and common interpolation schemes is discussed by among others, *Kimeldorf and Wahba* [1970], *Matheron* [1981], *Micchelli* [1986], and *Salkauskas* [1982]. Presented below are the results using two such common schemes.

**Tikhonov regularization.** Tikhonov regularization is a commonly used method to transform an ill-posed inverse problem to a well-posed minimization problem. *Skaggs and Kabala* [1994] use this method to estimate the release history. Tikhonov regularization is generally formulated as minimizing the function

$$V(\kappa) = \left\| y(w) - \int_a^b K(\lambda, w)s(\lambda)\, d\lambda \right\|^2 + \kappa^2 \|Ls\|^2 \tag{44}$$

where $s$ is the function to be estimated, $K$ is a known function of spatial or temporal variables $\lambda$ and $w$, and $\kappa$ is the regularization parameter. The regularization operator is generally of the form

$$\|Ls\|^2 = \int_a^b \left( \frac{d^n s}{d\lambda^n} \right)^2 d\lambda \tag{45}$$

The first term on the right-hand side of (44) forces the unknown function to reproduce the data. The second term enforces smoothness or another desired property of the unknown function. This formulation, although reasonable, has three problems. First, there is no obvious and simple method to find the optimal regularization parameter $\kappa$. Second, as often applied, the only measure of the estimation error is given by the scalar quantity $V(\kappa)$. This quantity tells little about the uncertainty associated with the estimate; i.e., it does not specify a confidence interval (although, as was pointed out by a reviewer of this paper, methods have been developed to compute bounds that account for the uncertainty introduced by measurement error [e.g., *Weese*, 1992; *Provencher*, 1982]). The third problem is that it may be difficult to formulate the problem in a computationally efficient manner. The geostatistical approach solves all of these problems.

If $n = 2$, which is the most popular case, then Tikhonov regularization in one dimension is equivalent to using the geostatistical approach for a first-order intrinsic function (i.e., linear trend)

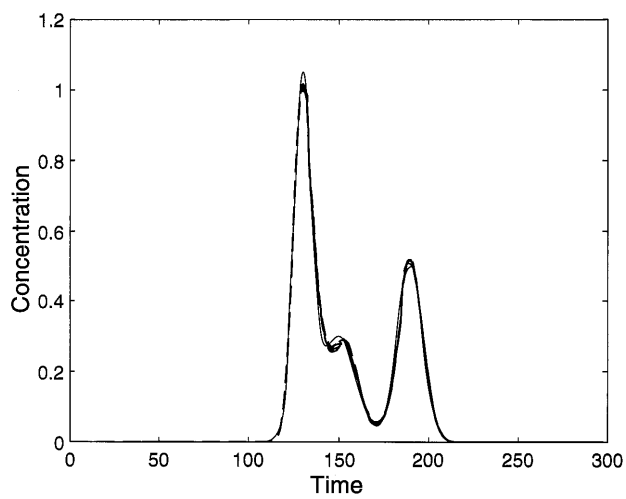


**Figure 6.** Constrained case with Gaussian covariance: the true solution (solid line) and five equally likely realizations (dashed lines).

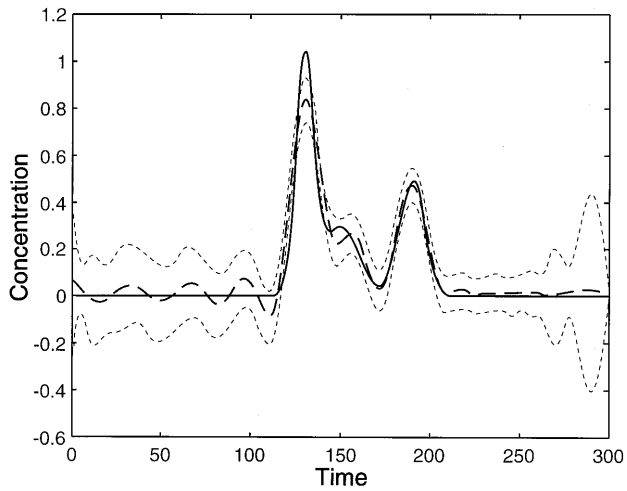


**Figure 7.** Unconstrained case with cubic spline generalized covariance: the true solution (solid line), best estimate (dashed line), and approximate 95% confidence interval (dotted lines).

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}$$

with the generalized covariance [*Kimeldorf and Wahba*, 1970]

$$Q(\theta) = \theta|h|^3 \tag{46}$$

where $h$ is the separation distance in units of time. This covariance matrix also corresponds to cubic spline interpolation.

For the unconstrained case, using this covariance matrix and $\sigma_R^2 = 1 \times 10^{-12}$, the optimal covariance parameter was

$$\theta = 7.82 \times 10^{-6} \pm 2.02 \times 10^{-8} \tag{47}$$

The best estimate and the approximate 95% confidence interval are shown in Figure 7, and five conditional realizations are shown in Figure 8.

For the constrained case, with $\sigma_R^2 = 1 \times 10^{-12}$, and $\alpha = 6$, the following result was obtained:

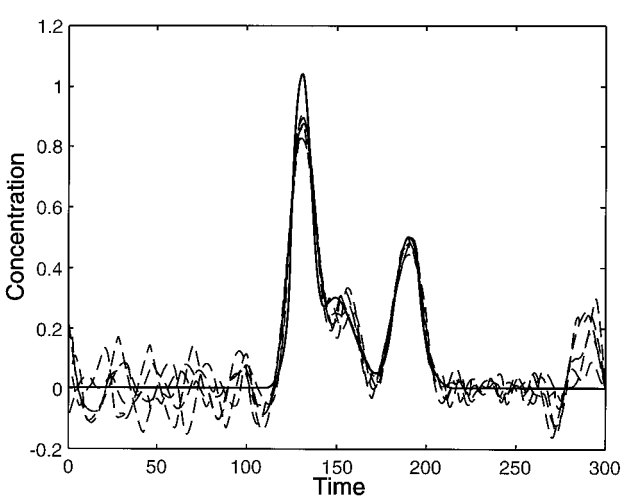


**Figure 8.** Unconstrained case with cubic spline generalized covariance: the true solution (solid line) and five equally likely realizations (dashed lines).

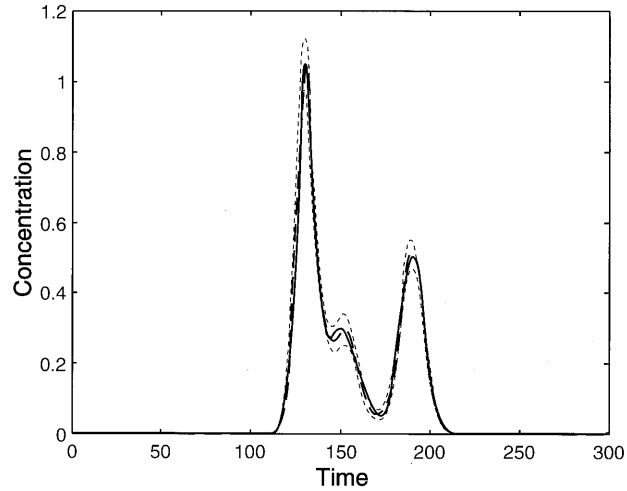


**Figure 9.** Constrained case with cubic spline generalized covariance: the true solution (solid line), best estimate (dashed line), and approximate 95% confidence interval (dotted lines).

$$\theta = 7.07 \times 10^{-5} \pm 1.33 \times 10^{-7} \tag{48}$$

The best estimate and the approximate 95% confidence interval are shown in Figure 9 and five conditional realizations are shown in Figure 10.

**Thin plate splines.** Another popular interpolation method is the thin plate spline. This method is equivalent to the geostatistical approach with the generalized covariance [*Micchelli*, 1986]

$$Q(\theta) = \begin{cases} \theta|h|^2 \ln|h| & h \neq 0 \\ 0 & h = 0 \end{cases} \tag{49}$$

Using this covariance matrix and $\sigma_R^2 = 1 \times 10^{-12}$ for the unconstrained case, the following result was obtained:

$$\theta = 1.24 \times 10^{-4} \pm 2.1 \times 10^{-6} \tag{50}$$

The best estimate and the approximate 95% confidence interval are shown in Figure 11. Five conditional realizations are shown in Figure 12.

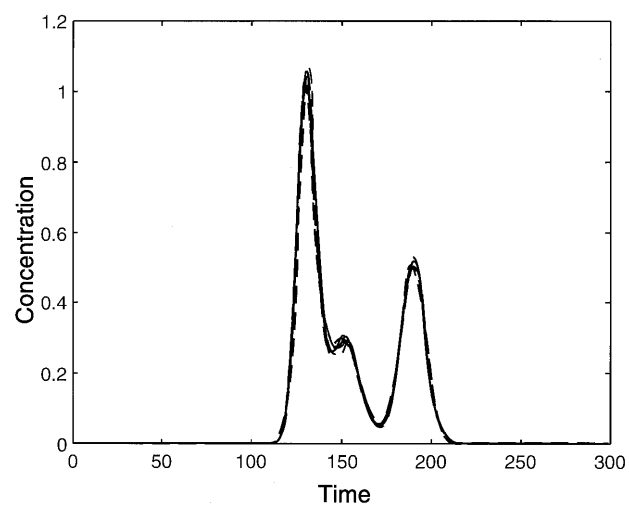


**Figure 10.** Constrained case with cubic spline generalized covariance: the true solution (solid line) and five equally likely realizations (dashed lines).

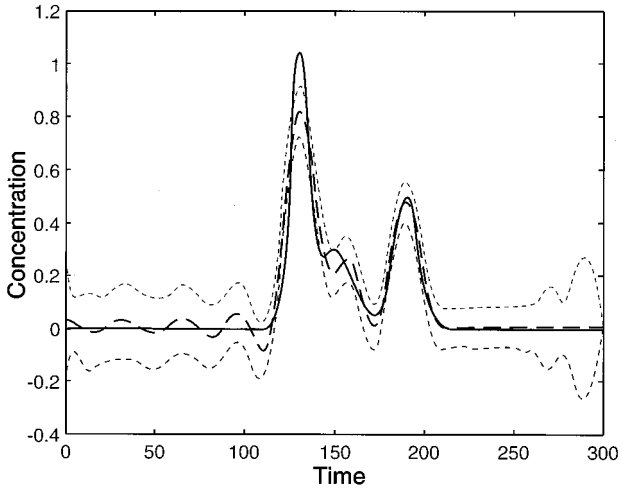


**Figure 11.** Unconstrained case with thin plate spline generalized covariance: the true solution (solid line), best estimate (dashed line), and approximate 95% confidence interval (dotted lines).

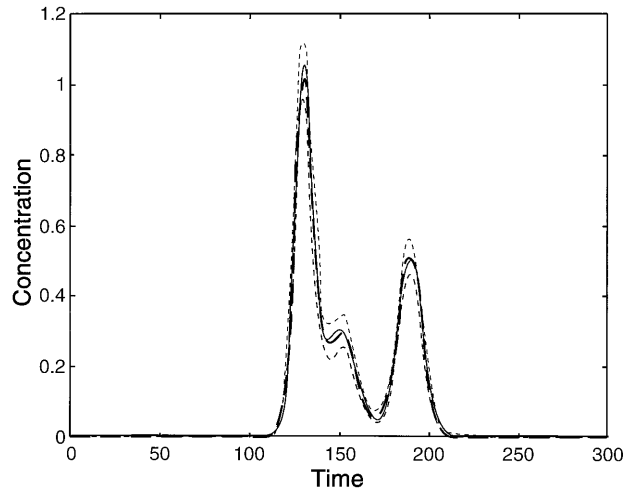


**Figure 13.** Constrained case with thin plate spline generalized covariance: the true solution (solid line), best estimate (dashed lines), and approximate 95% confidence interval (dotted line).

Using this covariance matrix, $\sigma_R^2 = 1 \times 10^{-12}$, and $\alpha = 6$ for the constrained case, the following result was obtained.

$$\theta = 1.15 \times 10^{-4} \pm 5.9 \times 10^{-7} \tag{51}$$

The best estimate and the approximate 95% confidence interval are shown in Figure 13. Five conditional realizations are shown in Figure 14. The best estimate agrees remarkably well with the true solution.

### Results With Large Measurement Error and Sparse Data

In practical applications the available data is usually sparse and there may be a high degree of measurement uncertainty. To illustrate how the method performs under these conditions we consider the following example. Figure 15 shows the position and values of the new measurements. Notice that the measurements vary considerably from the actual plume. Here the number of measurements is reduced to 15 and the variance of the measurements is increased to $\sigma_R^2 = 1 \times 10^{-3}$. Figure 16 shows the results. For illustration purposes a Gaussian covariance matrix and a constant mean are used with $\alpha = 8$. The optimal covariance parameters were calculated to be

$$\sigma^2 = 0.33 \pm 0.2$$

$$l = 17.4 \pm 8.3$$

Results are presented here for only the constrained case. The model results reflect the greater degree of uncertainty. The size of the 95% confidence interval has increased, and the best estimate is flatter and smoother. This is because the extreme peaks and valleys in actual release history cannot be resolved with the current degree of measurement error.

## Conclusions

This paper has demonstrated the application of Bayesian analysis and the geostatistical approach to the problem of release history estimation. The method improves on previous

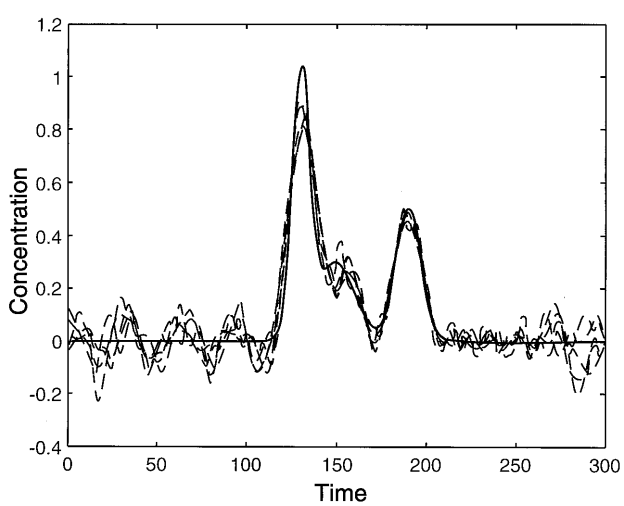


**Figure 12.** Unconstrained case with thin plate spline generalized covariance: the true solution (solid line) and five equally likely realizations (dashed lines).

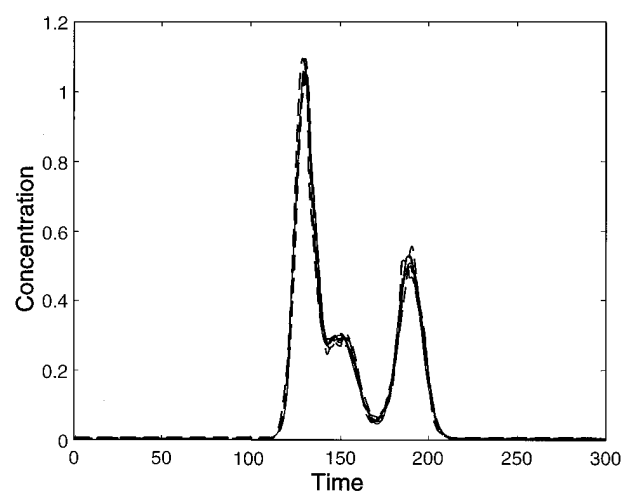


**Figure 14.** Constrained case with thin plate spline generalized covariance: the true solution (solid line) and five equally likely realizations (dashed lines).

solutions to the problem because it is more general and makes no blind assumptions about the nature and structure of the unknown source function. Many common solution methods such as least squares, splines, and Tikhonov regularization make implicit assumptions about the statistical structure of the unknown process. Through the geostatistical approach, only assumptions that are supported by the data are made. The issues of model development and validation were discussed here. By following the two guiding principles of geostatistics (indifference and parsimony) to arrive at tentative models, and then by criticizing and refining the model through the validation process, meaningful estimates may be obtained.

The method provides a measure of the error associated with the best estimate of the release history. This is extremely important in deciding how much faith to place on the estimate. It can also help identify sources of uncertainty, thus aiding in deciding if and where new measurements should be taken. Measurements are often costly. If more data are needed, the new sampling points should be located in the position where they will do the most good. The error bounds generated with the best estimate can aid in this decision. For example, if there is large uncertainty at time $T$ in the best estimate of the release history, a new measurement can be taken at position $x = Tv$, where $v$ is the groundwater velocity. This position corresponds to the location of the center of mass of a pulse injection at time $T$ before the measurement time.

The results for the unconstrained case (Figures 3 and 4) compare well with the results achieved by *Skaggs and Kabala* [1995] using the method of quasi-reversibility. The method presented here has the advantage that it provides a measure of estimation uncertainty and a method for optimizing the parameters used in the model. The geostatistical method is computationally efficient. The number of operations needed to compute the best estimate is, to leading order, $mn^2$ times the number of iterations needed to find the structural parameters, where $m$ is the number of measurements and $n$ is the number of grid points over which the release history is discretized. The number of iterations is usually small; in this example, only five were needed. In contrast, methods used elsewhere require operations of the order of $n^3$. Typically, $m \ll n$; therefore the developed method is much more efficient then other approaches.

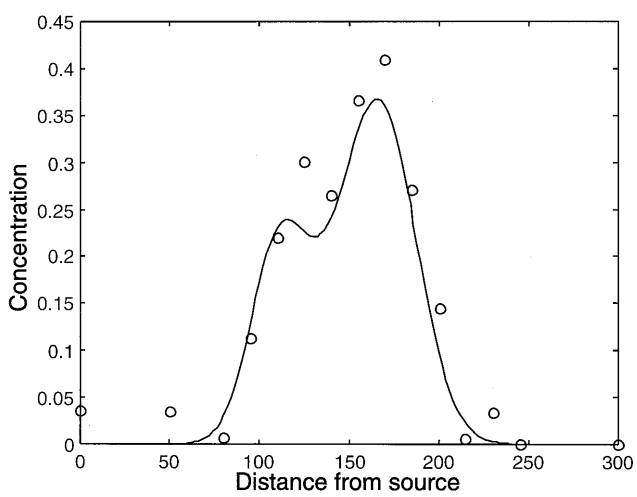


**Figure 15.** Contaminant plume in practical application with sparse data and large measurement error, with measurement locations denoted by circles.

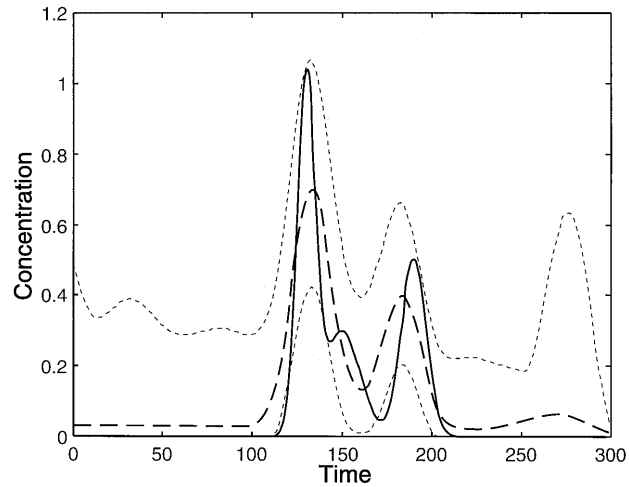


**Figure 16.** Constrained case with Gaussian covariance and large measurement error: the true solution (solid line), best estimate (dashed line), and approximate 95% confidence interval (dotted lines).

Introducing the additional information that concentration is a nonnegative quantity improves the performance of the method. More computational time is required because more iterations are needed. However the method is still efficient. The results compare well with those obtained by *Skaggs and Kabala* [1994] using Tikhonov regularization. The method presented here offers the added benefit of automatically supplying a measure of the estimation uncertainty and a means of optimizing the structural parameters. Additionally, the best form of the mean and covariance can be determined.

The generality of the method and its correspondence with other methods have been discussed. Noting this connection adds insight into the estimation procedure. Various types of splines have been used in interpolation and estimation to produce estimates with a desired degree of smoothness. The cubic spline and the thin plate spline discussed here produce very smooth estimates (see Figures 7 and 9) that in this case agree remarkably with the actual function. They require only one structural parameter. If this sort of result is desired, then the corresponding generalized covariance function may be used in the geostatistical approach.

One limitation of the method is that the location of the potential sources must be known. However, the method is robust enough to handle complex cases with multiple potential sources and spatially varying velocity field and dispersion coefficients. Uncertainty in the measurement error can also be included in the analysis. It has been demonstrated that the method performs well even with sparse and uncertain data. If the measurement error is unknown, then $\sigma_R^2$ can be estimated along with the other structural parameters.

The methodology was presented here with attention to computational efficiency. All calculation needed in the estimation process can be completed without inverting matrices whose size depends on the discretization of the unknown parameter. If a quick solution is desired, we recommend that the first method (unconstrained case) be used. A solution can be found with a minimal of computational effort. If a more accurate solution is desired, then it may be worth it to apply the second

method (constrained case). A number of generalized covariance matrices can be used to enforce certain desired properties on the estimate.

**Acknowledgments.** Funding for this research was provided by the U.S. Department of Navy, Naval Facilities Engineering Service Center, Port Hueneme, California, under grant R-815738-01 through the U.S. Environmental Protection Agency–supported Western Region Hazardous Substance Research Center. The contents of this study do not necessarily represent the views of these organizations.

## References

Andrews, D. F., Significance tests based on residuals, *Biometrica*, *58*, 139–148, 1971.

Anscombe, F. J., and J. W. Tukey, The examination and analysis of residuals, *Technometrics*, *5*, 141–160, 1963.

Bagtzoglou, A. C., D. E. Dougherty, and A. F. B. Tompson, Application of particle methods to reliable identification of groundwater pollution sources, *Water Resour. Manage.*, *6*, 15–23, 1992.

Belsley, D. A., A. Kuhn, and R. E. Welsch, *Regression Diagnostics*, Wiley-Interscience, New York, 1980.

Borgman, L. E., New advances in methodology for statistical tests useful in geostatistical studies, *Math. Geol.*, *20*, 383–403, 1988.

Box, G. E. P., Sampling and Bayes' Inference in Scientific Modeling and Robustness, *J. R. Stat. Soc.*, *Ser. A*, *143*, 383–430, 1980.

Box, G. E. P., and D. R. Cox, An analysis of transformations, *J. R. Stat. Soc.*, *Ser. B*, *26*, 353–360, 1964.

Box, G. E. P., and G. M. Jenkins, *Time Series Analysis*, Holden-Day, Merrifield, Va., 1976.

Dagan, G., Statistical theory of groundwater flow and transport: Pore to laboratory, laboratory to formation, and formation to region scale, *Water Resour. Res.*, *22*(9), 120–134, 1986.

Davis, B. M., Uses and abuses of cross-validation in geostatistics, *Math. Geol.*, *19*, 241–248, 1987.

Delfiner, P., Linear estimation of nonstationary spatial phenomena, in *Advanced Geostatistics in the Mining Industry*, edited by M. Guarascio, M. David, and C. Huijbregts, pp. 49–68, D. Reidel, Norwell, Mass., 1976.

Dubrule, O., Cross validation of kriging in a unique neighborhood, *Math. Geol.*, *15*, 687–699, 1983.

Gorelick, S. M., B. Evans, and I. Remson, Identifying sources of groundwater pollution: An optimization approach, *Water Resour. Res.*, *19*(3), 779–790, 1983.

Gutjahr, A., B. Bulard, S. Hatch, and L. Hughson, Joint conditional simulations and the spectral approach for flow modeling, *Stoch. Hydrol. Hydraul.*, *8*(1), 79–108, 1994.

Hoeksema, R. J., and P. K. Kitanidis, Comparison of Gaussian conditional mean and kriging estimation in the geostatistical solution of the inverse problem, *Water Resour. Res.*, *21*(6), 825–836, 1985.

Kimeldorf, G. S., and G. Wahba, A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *Ann. Math. Stat.*, *41*(2), 495–502, 1970.

Kitanidis, P. K., Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resour. Res.*, *22*(4), 499–507, 1986.

Kitanidis, P. K., The concept of predictive probability and a simple test for geostatistical model validation, in *Consequences of Spatial Variability in Aquifer Properties and Data Limitations for Groundwater Modelling Practice*, *IAHS Publ.*, *175*, 178–189, 1988.

Kitanidis, P. K., Orthonormal residuals in geostatistics: Model criticism and parameter estimation, *Math. Geol.*, *23*, 741–758, 1991.

Kitanidis, P. K., Quasi-linear geostatistical theory for inversing, *Water Resour. Res.*, *31*(10), 2411–2419, 1995.

Kitanidis, P. K., and K.-F. Shen, Geostatistical estimation of chemical concentration, *Adv. Water Resour.*, *19*(6), 369–378, 1996.

Kitanidis, P. K., and E. G. Vomvoris, A geostatistical approach to the inverse problem, *Water Resour. Res.*, *19*(3), 677–690, 1983.

Matheron, G., *Splines and Kriging: Their Formal Equivalence*, *Geol. Contrib. 8*, edited by D. F. Merriam, Dep. of Geol., Syracuse Univ., Syracuse, N. Y., 1981.

Micchelli, C. A., Interpolation of scattered data: Distance matrices and conditionally positive definite functions, *Constr. Approx.*, *2*, 11–22, 1986.

Provencher, S. W., A constrained regularization method for inverting data represented by linear algebraic or integral equations, *Comput. Phys. Commun.*, *27*, 213–227, 1982.

Salkauskas, K., Some relationships between surface splines and kriging, in *Multivariate Approximation Theory II*, edited by W. Schempp and K. Zeller, pp. 313–325, Birkhäuser, Basel, Switzerland, 1982.

Skaggs, T. H., and Z. J. Kabala, Recovering the release history of a groundwater contaminant, *Water Resour. Res.*, *30*(1), 71–79, 1994.

Skaggs, T. H., and Z. J. Kabala, Recovering the history of a groundwater contaminant plume: Method of quasi-reversibility, *Water Resour. Res.*, *31*(11), 2669–2673, 1995.

Tarantola, A., *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier, New York, 1987.

Wagner, B. J., Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling, *J. Hydrol.*, *125*, 275–300, 1992.

Weese, J., A reliable and fast method for the solution of Fredholm integral equations of the first kind based on Tikhonov regularization, *Comput. Phys. Commun.*, *69*, 99–111, 1992.

P. K. Kitanidis and M. F. Snodgrass, Department of Civil Engineering, Terman Engineering Center, Stanford University, Stanford, CA 94305-4020. (e-mail: msnod@leland.stanford.edu; pkk@ce.stanford.edu)

(Received August 1, 1996; revised December 1, 1996; accepted December 6, 1996.)