ELSEVIER

PII:S0309-1708(96)00005-X

Advances in Water Resources, Vol. 19, No. 6, pp. 333–342, 1996
Copyright © 1996 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0309-1708/96/$15.00 + 0.00

On the geostatistical approach to the inverse problem

Peter K. Kitanidis

Stanford University, Stanford, California 94305-4020, USA

(Revised received 8 February 1996)

The geostatistical approach to the inverse problem is discussed with emphasis on the importance of structural analysis. Although the geostatistical approach is occasionally misconstrued as mere cokriging, in fact it consists of two steps: estimation of statistical parameters ("structural analysis") followed by estimation of the distributed parameter conditional on the observations ("cokriging" or "weighted least squares"). It is argued that in inverse problems, which are algebraically undetermined, the challenge is not so much to reproduce the data as to select an algorithm with the prospect of giving good estimates where there are no observations. The essence of the geostatistical approach is that instead of adjusting a grid-dependent and potentially large number of block conductivities (or other distributed parameters), a small number of structural parameters are fitted to the data. Once this fitting is accomplished, the estimation of block conductivities ensues in a predetermined fashion without fitting of additional parameters. Also, the methodology is compared with a straightforward maximum *a posteriori* probability estimation method. It is shown that the fundamental differences between the two approaches are: (a) they use different principles to separate the estimation of covariance parameters from the estimation of the spatial variable; (b) the method for covariance parameter estimation in the geostatistical approach produces statistically unbiased estimates of the parameters that are not strongly dependent on the discretization, while the other method is biased and its bias becomes worse by refining the discretization into zones with different conductivity. Copyright © 1996 Elsevier Science Ltd

1 INTRODUCTION

Mathematical solutions to the equations that describe groundwater flow and the transport and transformation of solutes are important tools for predicting the behavior of a hydrogeologic system. However, predictions made by such simulation models cannot be reliable without representative values for the hydrogeologic parameters, the mass sources and sinks, boundary conditions, etc. Most parameters must be inferred from data.

The literature on estimating groundwater parameters is voluminous[25] and the problem is still receiving much attention, as adequate site characterization is recognized as crucial in making good predictions and decisions and as new methods for measurements are developed. The diversity among available methods confounds those who have not devoted substantial effort in studying the problem.

In my judgement, the well-studied and established methods of statistical inference provide the basis for solving estimation problems encountered in hydrogeology, especially challenging inverse problems. For example, methods of nonlinear least squares or nonlinear regression have proven particularly useful.[4,5] However, the estimation of spatial processes involves distinctive challenges that have not been addressed in other fields. Principal among them is that the estimation from some measurements of an arbitrarily large number of parameters, obtained from the discretization of a spatial function, is an ill-posed problem. The established but unfortunate term "ill-posed" simply means that there are more unknowns than equations, resulting in an under determined system of equations. The numerous different ways of possible reformulation (or "parametrization") of this problem into a problem with a well-defined solution accounts for the proliferation of available methods.

Carrera and Glorioso[3] pointed out similarities between two methods for the solution of the inverse problem, particularly the estimation of conductivity or transmissivity from head and other measurements: the

geostatistical approach or GA;[6,10–12,18,21,22,24] and a somewhat related maximum likelihood approach or MAP.[2] This work has two principal objectives:

1. to discuss the essential points of the geostatistical approach, and
2. to highlight through theoretical analysis and examples the key differences between GA and MAP.

2 THE GEOSTATISTICAL APPROACH

One of the methods for the solution of the inverse problem, proposed by Kitanidis and Vomvoris,[18] is based on a parametrization familiar from the theory of random fields that has been used extensively in the stochastic approach to subsurface flow and transport.[7,9] This parametrization has been popularized in geophysics by the classical geostatics of Matheron;[19] for this reason, this method was named "geostatistical". The formal framework within which Kitanidis and Vomvoris[18] approach the inverse problem is that of standard statistical inference[1,8,20,23] and the theory of stochastic differential equations.

The spatially variable parameter to be determined (typically the log-conductivity) is represented as a realization of an appropriately characterized random field. In practice, the characterization is usually limited to specifying a mean function and a covariance function. Such a model is justified as a practical way to represent the structure of the unknown function without making overly strong or restrictive assumptions. Matheron[19] and others have argued that deterministic trend functions are not appropriate for the description of erratic variability, especially small-scale variability or random-walk type variability; information about such variability is more suitably represented through low-order statistics. Geological processes are complex and it is difficult to predict in quantitative terms the variability of parameters such as hydraulic conductivity, porosity, etc. In most geostatistical applications, the simplest mean function (i.e. a constant) is conservatively adopted and the structure of the process is described through the variogram. This approach is compatible with the fundamental principle of inference that dictates that: the simplest and least restrictive empirical model that agrees with what is otherwise known should be fitted to the data; and that complexity should be added to the empirical model only if it leads to a statistically significant improvement in the fit to the data and is not inconsistent with other information.

In this approach, parameter estimation is carried out in two stages. The first stage, known as structural analysis, is the characterization of the random field. For example, in the most basic case, the variogram is selected. The second stage is that of conditioning on the data, which is a well-defined mathematical problem once the structure has been selected: derive estimates (or the conditional distribution) of the unknown parameters given the observations. In the case that the relation between the set of observations and the unknowns can be linearized, the minimum-variance linear unbiased estimates of the parameters are obtained in a straightforward fashion through an algorithm known in geophysics as cokriging. By contrast, the process of structural analysis involves inducing a model from the data. This process may not be fully automated but must be applied in three steps:

1. *Exploratory analysis* of the data leads to a tentative selection of a simple empirical model with a very small number of adjustable parameters to describe the structure of the variable to be estimated;
2. *Parameter estimation*, where for a given model the best possible (in some sense) estimates of the adjustable parameters are obtained through fitting to the data;
3. *Model criticism* (also known as *validation*), where the fitted model is checked by performing some statistical tests that may reveal model inadequacies and may lead to model modifications.

It is nonsensical to introduce an empirical model with too many parameters or with parameters that cannot be identified from the data. An implicit assumption in the approach just described is that reasonably accurate estimates of the parameters can be identified, otherwise, it may be important to recognize the uncertainty in the covariance parameters and their effect on the second step. A methodology for accounting for parameter uncertainty has been described,[14] but it is computationally intensive and may be impractical to use in routine applications. Another reason to avoid using a model with unidentifiable parameters is that it will likely lead to difficulties in the fitting of parameters (step 2) which is a nonlinear optimization procedure. One cannot overemphasize the importance of checking that the results of the optimization make sense and that the estimates are reasonable, stable, and with acceptable variance of sampling error. Finally, model criticism is vital because this is where, by subjecting the empirical model to tests as severe as possible, one develops some basis for trusting the results of the inverse modeling. No methodology or model that precludes the possibility of being tested and found incompatible with the data can be considered scientific.

3 FITTING DATA IN THE GEOSTATISTICAL APPROACH

It has always been stressed[6,10–13,18,21,22] that the geostatistical approach consists of two steps: estimation of structural parameters followed by estimation of the

distributed parameter (e.g. block transmissivities). Despite the emphasis given to that particular point, this approach has often been misunderstood as "mere cokriging". However, as we will argue, cokriging is not data fitting in any useful sense of the word and consequently there can be no such thing as "an inverse method that is mere cokriging".

Consider what cokriging has to do with matching data. Given expressions professing to represent the covariance functions of log-transmissivity and head and the cross-covariance between log-transmissivity and head, one may obtain the best estimates of log-transmissivity and head conditional on observations. Cokriging reproduces the measurements, no matter how absurd the assumed covariances or how senseless the results at points where no observations are available. That the estimates reproduce the observations is a consequence of the so-called "exact interpolator" property. It is a consequence of how cokriging is formulated. It happens automatically and not because there is some basis to hope that the predictions are any good. Of course, the real challenge is not to reproduce the observations but to devise a rational procedure to estimate quantities that we do not know, such as conductivities at points between measurement locations. There are in principle infinite degrees of freedom in choosing a spatial function and consequently any data set can be reproduced through cokriging in a trifling fashion even with the most absurd geostatistical model.

Where the important data matching or optimization occurs is in structural analysis, which is by far the most interesting step in the analysis. There, a few (structural) parameters are estimated or fitted to the data. In the geostatistical approach, all observations are used, rather than just the log-transmissivity observations as done in *ad hoc* methods such as experimental variogram analysis of log-transmissivity data.

To appreciate better why the estimation of structural parameters is the real history matching step, we will review an intuitive interpretation of the structural parameter estimation method used in Kitanidis and Vomvoris[18] and other applications of the approach, skipping details that can be found in Kitanidis.[15] We treat the observations $z_1, \ldots, z_n$ as if they were given in a certain order. Consider that using a starting subset of the data $z_1, \ldots, z_{k-1}$, we can predict through cokriging that we expect the next observation to be $\check{z}_k$. Since cokriging has not used z_k as an observation, there is generally an error $e_k = \check{z}_k - z_k$, which is a true measure of the predictive ability of the model. We consider next an expanded data base $z_1, \ldots, z_k$, compute the predicted value $\check{z}_{k+1}$ and the prediction error $e_{k+1} = \check{z}_{k+1} - z_{k+1}$. We continue with this process until we obtain a set of fitting residuals which depend on the structural parameters. (By the way, such an approach is sometimes termed "cross-validation", but this is a misnomer because the issue is parameter estimation.) A basic principle of inference is that it is eminently reasonable that parameter values that give small residuals are more likely than parameter values that give large residuals, provided of course that the model makes sense and a small number of parameters are fitted to many observations. My point is that the approach recommended in Kitanidis and Vomvoris[18] effectively selects the parameters that give the best fit in a well-defined sense.

Thus, the basic idea in the geostatistical approach is that, instead of attempting to solve the ill-posed problem of fitting a spatial function, the parameters that describe its structure are fitted. Thus, we end up with a well-defined parameter estimation problem with perhaps dozens of observations and only a couple of parameters. For such a case, it is meaningful to attempt history matching, the same way that it makes sense for a biochemist to use the sequence of one hundred or so observations of chemical concentration in a laboratory reactor in order to fit a couple of parameters that describe the rate of biochemical transformation; and the same way that it would be meaningless to fit a model with more parameters than data, just to reproduce the data.

4 LINEAR CASE

4.1 General formulation

Let s represent the variable (such as log-transmissivity) that needs to be estimated over the flow domain. The process has a mean parametrized by $\boldsymbol{\beta}$ and a covariance function parametrized by $\boldsymbol{\theta}$. We discretize the spatial domain so that $\mathbf{s}$ is the m vector of the discretized variable values and

$$E[\mathbf{s}] = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

where $\mathbf{X}$ is a known m by p matrix, $\boldsymbol{\beta}$ is a p vector of unknown drift coefficients, and $E[\cdot]$ denotes expected value. Furthermore, $\mathbf{s}$ has a covariance matrix

$$E[\mathbf{s} - \mathbf{X}\boldsymbol{\beta})(\mathbf{s} - \mathbf{X}\boldsymbol{\beta})^T] = \mathbf{Q}(\boldsymbol{\theta}) \tag{2}$$

that is considered a known function of parameters $\boldsymbol{\theta}$. The exponent ^{T} stands for matrix transpose.

The $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ parameters are treated as unknown constants and are supposed to be few in number, far fewer than the observations. The observations are related to the unknown spatial process and the other parameters through:

$$\mathbf{z} = \mathbf{h}(\mathbf{s}) + \mathbf{v} \tag{3}$$

where $\mathbf{z}$ is the vector of observations. The observation error $\mathbf{v}$ is random with zero mean and covariance matrix $\mathbf{R}$, fixed or (for generality) a known function of $\boldsymbol{\theta}$.

The vector $\mathbf{z}$ is a random vector, because it is a

function of $\mathbf{s}$ and $\mathbf{v}$ that are random vectors. The joint pdf of $\mathbf{z}$ and $\mathbf{s}$ depends on the distribution of $\mathbf{s}$ and $\mathbf{v}$ and also on the function $\mathbf{h}$:

$$p(\mathbf{z},\mathbf{s}) = p(\mathbf{z}|\mathbf{s})p(\mathbf{s}) \tag{4}$$

It is common to model the observational errors as Gaussian:

$$p(\mathbf{z}|\mathbf{s}) \propto |\mathbf{R}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\mathbf{z}-\mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{h}(\mathbf{s}))\right] \tag{5}$$

where here $|\cdot|$ denotes the determinant of a square matrix. The prior probability distribution of $\mathbf{s}$ is considered Gaussian (consistent with most previous work):

$$p(\mathbf{s}) \propto |Q|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{Q}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta})\right] \tag{6}$$

Then,

$$p(\mathbf{z},\mathbf{s}|\boldsymbol{\beta},\boldsymbol{\theta}) \propto |\mathbf{R}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\mathbf{z}-\mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{h}(\mathbf{s}))\right] |\mathbf{Q}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{Q}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta})\right] \tag{7}$$

Furthermore, the (marginal) probability distribution of $\mathbf{z}$ given only $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is:

$$p(\mathbf{z}|\boldsymbol{\beta},\boldsymbol{\theta}) = \int_s p(\mathbf{z},\mathbf{s}|\boldsymbol{\beta},\boldsymbol{\theta})d\mathbf{s} = |R|^{-\frac{1}{2}}|\mathbf{Q}|^{-\frac{1}{2}}\int_s \times \exp[-\tfrac{1}{2}\{(\mathbf{z}-\mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{h}(\mathbf{s})) + (\mathbf{s}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{Q}^{-1}(\mathbf{s}-\mathbf{X}\boldsymbol{\beta})\}]\, d\mathbf{s} \tag{8}$$

4.2 Reduced case

Because our objective here is to highlight the similarities and differences between the geostatistical approach and the MAP method, nonessential parts will be eliminated to allow us to focus on a case that is analytically manageable. First, we will neglect uncertainty in the drift coefficients so that the drift can be subtracted from the spatial process. This is formally equivalent to setting $\mathbf{X} = 0$. Second, we consider that the relation between the observations and the spatial process is linear:

$$\mathbf{z} = \mathbf{H}\mathbf{s} + \mathbf{v} \tag{9}$$

Note that these simplifications are made only in order to make the main point more visible and that the more general case can be addressed without conceptual difficulty, as has been done elsewhere. In particular, note that the geostatistical approach has been extended for nonlinear systems.[17] Under these conditions the joint pdf of $\mathbf{z}$ and $\mathbf{s}$:

$$p(\mathbf{z},\mathbf{s}|\boldsymbol{\theta}) \propto |\mathbf{R}|^{-\frac{1}{2}}|\mathbf{Q}|^{-\frac{1}{2}} \times \exp\left[-\tfrac{1}{2}\{(\mathbf{z}-\mathbf{H}\mathbf{s})^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{s}) + \mathbf{s}^T\mathbf{Q}^{-1}\mathbf{s}\}\right] \tag{10}$$

and the pdf of the observations $\mathbf{z}$ is

$$p(\mathbf{z}|\boldsymbol{\theta}) \propto |\mathbf{R}|^{-\frac{1}{2}}|\mathbf{Q}|^{-\frac{1}{2}} \times \int_s \exp\left[-\tfrac{1}{2}\{(\mathbf{z}-\mathbf{H}\mathbf{s})^T\mathbf{R}^{-1} \times (\mathbf{z}-\mathbf{H}\mathbf{s}) + \mathbf{s}^T\mathbf{Q}^{-1}\mathbf{s}\}\right]d\mathbf{s} \tag{11}$$

4.3 Maximum *a posteriori* probability (MAP) method

Maximum *a posteriori* probability (MAP) estimation boils down to maximizing eqn (10) with respect to both $\mathbf{s}$ and $\boldsymbol{\theta}$. (It is the same method that Carrera and Neuman[2] call maximum likelihood.) The problem is equivalent to the minimization of the negative logarithm of eqn (10),

$$\tfrac{1}{2}\ln|\mathbf{R}| + \tfrac{1}{2}\ln|\mathbf{Q}| + \tfrac{1}{2}[(\mathbf{z}-\mathbf{H}\mathbf{s})^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{s}) + \mathbf{s}^T\mathbf{Q}^{-1}\mathbf{s}] \tag{12}$$

Setting the derivative with respect to vector $\mathbf{s}$ equal to zero and solving for $\mathbf{s}$ (as function of $\boldsymbol{\theta}$):

$$\mathbf{s} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{z} \tag{13}$$

This is the best estimate of $\mathbf{s}$ (if $\boldsymbol{\theta}$ were given) and has mean square error matrix:

$$\mathbf{V}_s = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1})^{-1} \tag{14}$$

These two equations constitute a linear estimator (essentially the same as with cokriging or Gaussian conditional mean estimation). Since $\mathbf{s}$ is a function of $\boldsymbol{\theta}$, we may substitute into the objective function obtaining what we need to minimize with respect to $\boldsymbol{\theta}$:

$$\tfrac{1}{2}\ln|\mathbf{R}| + \tfrac{1}{2}\ln|\mathbf{Q}| + \tfrac{1}{2}\mathbf{z}^T\{\mathbf{R}^{-1}\mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\}\mathbf{z} \tag{15}$$

or, using a matrix identity,

$$\tfrac{1}{2}\ln|\mathbf{R}| + \tfrac{1}{2}\ln|\mathbf{Q}| + \tfrac{1}{2}\mathbf{z}^T(\mathbf{R} + \mathbf{H}\mathbf{Q}\mathbf{H}^T)^{-1}\mathbf{z} \tag{16}$$

Thus, the estimation problem has been separated into two problems:

1. Structural analysis, or estimation of covariance parameters $\boldsymbol{\theta}$ from the minimization of eqn (16);
2. Linear estimation, or estimation of spatial function $\mathbf{s}$ from eqn (13).

4.4 Geostatistical approach (GA)

In the geostatistical approach,[18] the estimation of structural parameters is based on the maximization of the expression of eqn (11), which after performing the

integration analytically is:

$$p(\mathbf{z}|\boldsymbol{\theta}) \propto |\mathbf{R}|^{-\frac{1}{2}}|\mathbf{Q}|^{-\frac{1}{2}}|\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}+\mathbf{Q}^{-1}|^{-\frac{1}{2}} \times \exp\left[\tfrac{1}{2}\mathbf{z}^T\{\mathbf{R}^{-1}-\mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}+\mathbf{Q}^{-1})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\}\mathbf{z}\right] \tag{17}$$

or (using matrix identities)

$$p(\mathbf{z}|\boldsymbol{\theta}) \propto |\mathbf{R}+\mathbf{HQH}^T|^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}\mathbf{z}^T(\mathbf{R}+\mathbf{HQH}^T)^{-1}\mathbf{z}\right] \tag{18}$$

Thus, structural parameters are estimated by minimizing with respect to $\boldsymbol{\theta}$ the function:

$$\tfrac{1}{2}\ln|\mathbf{R}+\mathbf{HQH}^T|+\tfrac{1}{2}\mathbf{z}^T(\mathbf{R}+\mathbf{HQH}^T)^{-1}\mathbf{z} \tag{19}$$

The same results could have been obtained if, like Kitanidis and Vomvoris,[18] we had used the well-known fact that the linear transformation of a Gaussian vector is a Gaussian vector, so that $\mathbf{z}$ is Gaussian with a mean zero and covariance matrix $\mathbf{R}+\mathbf{HQH}^T$.

After the structural parameters have been obtained, the pdf of $\mathbf{s}$ given $\mathbf{z}$ can be obtained easily, given that they are jointly Gaussian, as done, for example, in Dagan.[6] The result is eqns (13) and (14).

As the analysis has demonstrated, the estimation consists of two stages: structural analysis and linear estimation (cokriging or Gaussian conditional mean). The second-stage formulae are identical in GA and MAP methods. However, the objective functions to be minimized for estimation of structural parameters are different, as can be seen by comparing eqns (16) and (19).

4.5 Illustrative example

Consider a generic linear case where the covariance is known except for a multiplicative constant $\boldsymbol{\theta}$ and the observation error is extremely small so that $\mathbf{R}$ can be neglected in comparison to $\mathbf{Q}$:

$$\mathbf{Q}=\boldsymbol{\theta}\mathbf{Q}_0, \qquad \mathbf{R}=c\mathbf{R}_0 \tag{20}$$

where $\mathbf{Q}_0$ is a known $m\times m$ matrix of rank m, $\mathbf{R}_0$ is a known $n\times n$ matrix of rank n, $\mathbf{H}$ is an $n\times m$ matrix of rank n, and the scalar $c\simeq 0$. The numerical values of $\mathbf{Q}_0$, $\mathbf{R}_0$, and $\mathbf{H}$ do not matter. We will be concerned here with cases where m is larger than n, because a domain can be discretized into an extremely large number of zones or nodes, whereas measurements are limited in number. It is reiterated that m is the number of parameters (zones in the MAP approach) that determine the discrete number of $\mathbf{s}$ variables. Then,

$$|\mathbf{Q}|=\theta^m|\mathbf{Q}_0| \tag{21}$$

$$\mathbf{R}+\mathbf{HQH}^T\simeq\theta\mathbf{HQ}_0\mathbf{H}^T,\ |\mathbf{R}+\mathbf{HQH}^T|\simeq\theta^n|\mathbf{HQ}_0\mathbf{H}^T| \tag{22}$$

Next, we will apply the two methods, the obtained estimators, and evaluate their properties.

MAP

$$\min_{\theta}\left[\frac{1}{2}|\mathbf{Q}_0|+\frac{m}{2}\ln\hat{\theta}+\frac{1}{2}\mathbf{z}^T\frac{(\mathbf{HQ}_0\mathbf{H}^T)^{-1}}{\hat{\theta}}\mathbf{z}\right] \tag{23}$$

Set the derivative with respect to $\hat{\theta}$ equal to zero:

$$\frac{m}{2}\frac{1}{\hat{\theta}}-\frac{1}{2}\mathbf{z}^T\frac{(\mathbf{HQ}_0\mathbf{H}^T)^{-1}}{\hat{\theta}^2}\mathbf{z}=0 \tag{24}$$

to obtain the MAP estimate:

$$\hat{\theta}_{MAP}=\frac{1}{m}\mathbf{z}^T(\mathbf{HQ}_0\mathbf{H}^T)^{-1}\mathbf{z} \tag{25}$$

GA

$$\min_{\theta}\left[\frac{1}{2}|\mathbf{HQ}_0\mathbf{H}^T|+\frac{n}{2}\ln\hat{\theta}+\frac{1}{2}\mathbf{z}^T\frac{(\mathbf{HQ}_0\mathbf{H}^T)^{-1}}{\hat{\theta}}\mathbf{z}\right] \tag{26}$$

Take derivative, then set equal to 0:

$$\frac{n}{2}\frac{1}{\hat{\theta}}-\frac{1}{2}\mathbf{z}^T\frac{(\mathbf{HQ}_0\mathbf{H}^T)^{-1}}{\hat{\theta}^2}\mathbf{z}=0 \tag{27}$$

$$\hat{\theta}_{GA}=\frac{1}{n}\mathbf{z}^T(\mathbf{HQ}_0\mathbf{H}^T)^{-1}\mathbf{z} \tag{28}$$

4.6 Comparison

Clearly, these two estimators are different. To test for bias, compute the expected value of each estimator:

$$\begin{aligned}E[\hat{\theta}_{MAP}]&=\frac{1}{m}E[\mathbf{z}^T(\mathbf{HQ}_0\mathbf{H}^T)^{-1}\mathbf{z}]\\&=\frac{1}{m}E[Tr((\mathbf{HQ}_0\mathbf{H}^T)^{-1}\mathbf{zz}^T)]\\&=\frac{1}{m}Tr((\mathbf{HQ}_0\mathbf{H}^T)^{-1}E[\mathbf{zz}^T])\\&=\frac{1}{m}Tr(\mathbf{I}_{n\times n})\theta=\frac{n}{m}\theta\end{aligned} \tag{29}$$

Following exactly the same steps:

$$E[\hat{\theta}_{GA}]=\frac{1}{n}E[\mathbf{z}^T(\mathbf{HQ}_0\mathbf{H}^T)^{-1}\mathbf{z}]=\frac{n}{n}\theta=\theta \tag{30}$$

Thus, while the geostatistical approach yielded an unbiased estimator, the MAP approach yielded a biased estimator, with bias that depends critically on the discretization (into zones of constant conductivity, for example). The finer the grid, the worse off the bias and, for a very fine discretization, the MAP method ends up with one hundred percent relative bias!

To appreciate the practical significance of the bias in the covariance parameter, consider how it affects the computed mean square error, given for both cases by

eqn (14). For the example considered here:

$$\mathbf{V}_s = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1})^{-1}$$
$$= [\mathbf{Q}_0 - \mathbf{Q}_0\mathbf{H}^T(\mathbf{H}\mathbf{Q}_0\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{Q}_0]\theta \tag{31}$$

The mean square estimation matrix $\mathbf{V}_s$ is thus proportional to the value of θ. If a greatly biased estimate of θ is used, the computed mean square estimation matrix is not representative of the actual error. If the domain is discretized into a very large number of zones, the MAP method will miscalculate the mean square estimation error to be zero.

5 ASYMPTOTIC BIAS ANALYSIS

The structural parameters in both methods are obtained from a nonlinear optimization scheme. Theoretical bias analysis in nonlinear estimation in its most general form is quite complex. However, we may analyze the asymptotic case, i.e. that the sample is effectively "large" and that the MAP and GA estimates $\hat{\boldsymbol{\theta}}_{MAP}$ and $\hat{\boldsymbol{\theta}}_{GA}$ are "close" to the true parameters $\boldsymbol{\theta}$. Our objective here is to demonstrate that while the GA parameter estimation method is unbiased, the same cannot be said about the MAP parameter estimation method.

The MAP method estimates structural parameters from the minimization with respect to $\boldsymbol{\theta}$ of:

$$L_{MAP}(\boldsymbol{\theta}) = \tfrac{1}{2}\ln|\mathbf{R}| + \tfrac{1}{2}\ln|\mathbf{Q}| + \tfrac{1}{2}\mathbf{z}^T(\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{z} \tag{32}$$

$$\frac{dL_{MAP}}{d\theta_i} = \tfrac{1}{2}Tr[\mathbf{R}^{-1}\mathbf{R}_i] + \tfrac{1}{2}Tr[\mathbf{Q}^{-1}\mathbf{Q}_i] - \tfrac{1}{2}Tr[(\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{H}\mathbf{Q}_i\mathbf{H}^T + \mathbf{R}_i) \times (\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{z}\mathbf{z}^T] \tag{32a}$$

where $\mathbf{R}_i = \frac{\partial \mathbf{R}}{\partial \theta_i}$ and $\mathbf{Q}_i = \frac{\partial \mathbf{Q}}{\partial \theta_i}$. Taking expected values and assuming that $E[\mathbf{z}\mathbf{z}^T] = \mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R}$

$$E\left[\frac{dL_{MAP}}{d\theta_i}\right] = \frac{1}{2}Tr[\mathbf{R}^{-1}\mathbf{R}_i] + \frac{1}{2}Tr[\mathbf{Q}^{-1}\mathbf{Q}_i] - \frac{1}{2}Tr[(\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{H}\mathbf{Q}_i\mathbf{H}^T + \mathbf{R}_i)] \tag{33}$$

From identity

$$|\mathbf{R}|\,|\mathbf{Q}|\,|\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1}| = |\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R}| \tag{34}$$

by taking logarithms

$$\ln|\mathbf{R}| + \ln|\mathbf{Q}| + \ln|\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1}| = \ln|\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R}| \tag{35}$$

and differentiating with respect to θ_i

$$Tr[\mathbf{R}^{-1}\mathbf{R}_i] + Tr[\mathbf{Q}^{-1}\mathbf{Q}_i] - Tr[(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1}) \times (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{R}_i\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1}\mathbf{Q}_i\mathbf{Q}^{-1})] = Tr[(\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{H}\mathbf{Q}_i\mathbf{H}^T + \mathbf{R}_i)] \tag{36}$$

we obtain

$$E\left[\frac{dL_{MAP}}{d\theta_i}\right] = \frac{1}{2}Tr[(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1}) \times (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{R}_i\mathbf{R}^{-1}\mathbf{H} + \mathbf{Q}^{-1}\mathbf{Q}_i\mathbf{Q}^{-1})] \tag{37}$$

The GA method, on the other hand, minimizes:

$$L_{GA}(\boldsymbol{\theta}) = \tfrac{1}{2}\ln|\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R}| + \tfrac{1}{2}\mathbf{z}^T(\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{z} \tag{38}$$

$$\frac{dL_{GA}}{d\theta_i} = \tfrac{1}{2}Tr[(\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1} \times (\mathbf{H}\mathbf{Q}_i\mathbf{H}^T + \mathbf{R}_i)] - \tfrac{1}{2}Tr[(\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1} \times (\mathbf{H}\mathbf{Q}_i\mathbf{H}^T + \mathbf{R}_i)(\mathbf{H}\mathbf{Q}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{z}\mathbf{z}^T] \tag{39}$$

Taking expected values

$$E\left[\frac{dL_{GA}}{d\theta_i}\right] = 0 \tag{40}$$

Rao (p. 367)[20] has proved that asymptotically:

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = -\mathbf{F}^{-1}\frac{dL}{d\boldsymbol{\theta}} \tag{41}$$

where $\mathbf{F}$ is the Fisher information matrix (computed for the true value of the parameters). Intuitively, one may view this as a Gauss–Newton iteration starting with the true value $\boldsymbol{\theta}$; because of the assumed proximity of $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}$, one iteration should be sufficient to lead to the final estimate. Then, taking expected values,

$$E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] = -\mathbf{F}^{-1}E\left[\frac{dL}{d\boldsymbol{\theta}}\right] \tag{42}$$

which expresses the important property that for asymptotic unbiasedness, the expected value of the gradient of the log-likelihood function needs to be zero. As already demonstrated, this is not the case for the MAP method,

$$E[\hat{\boldsymbol{\theta}}_{MAP} - \boldsymbol{\theta}] = -\mathbf{F}^{-1}E\left[\frac{dL_{MAP}}{d\boldsymbol{\theta}}\right] \tag{43}$$

Carrying out the same procedure for the GA estimator:

$$E[\hat{\boldsymbol{\theta}}_{GA} - \boldsymbol{\theta}] = -\mathbf{F}^{-1}E\left[\frac{dL_{GA}}{d\boldsymbol{\theta}}\right] = \mathbf{0} \tag{44}$$

Thus, asymptotically the MAP estimator is biased, whereas the GA estimator is unbiased.

6 EXAMPLE FROM INVERSE MODELING

6.1 Case study

We will compare the GA and MAP methods for an inverse problem that is similar to the one in Kitanidis.[17] Consider steady one-dimensional flow without sources or sinks:

$$\frac{d}{dx}\left(K(x)\frac{d\phi}{dx}\right) = 0 \tag{45}$$

where ϕ is the hydraulic head and K is the hydraulic conductivity. It is given as boundary or auxiliary conditions that $\phi(0) = 1$ and the discharge is $q = 5.93$. The downgradient boundary head, $\phi(1)$, is not given. The objective is to estimate the log-conductivity $Y = \ln K$ and the head ϕ, and to evaluate the uncertainty associated with the estimation over a grid that covers the domain of interest with spacing $\Delta x = \frac{1}{100}$.

The log-conductivity is represented as a realization of a stationary function with exponential covariance function:

$$C_Y(x_i, x_j) = v \exp\left(-\frac{|x_i - x_j|}{l}\right) \tag{46}$$

where v and l are structural parameters to be estimated. The measurement errors are taken to be independent and identically distributed with variance $\sigma_R^2 = 10^{-6}$, i.e. the measurement error covariance matrix was taken to be equal to:

$$\mathbf{R} = \sigma_R^2 \mathbf{I} \tag{47}$$

where $\mathbf{I}$ is the identity matrix. Before proceeding, let us review the methods to be applied. Here, $\mathbf{s}$ is the n by 1 log-conductivity vector. $\mathbf{X}$ is an n by 1 vector of 1s, β is the process mean, and the covariance parameters $\boldsymbol{\theta}$ are the sill and length parameters of the exponential covariance function.

6.2 Method GA

Maximize with respect to $\boldsymbol{\theta}$ the expression:

$$p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{r}) = \int_{\mathbf{s}} p(\mathbf{z}|\beta, \boldsymbol{\theta}) d\beta \propto |\mathbf{R}|^{-\frac{1}{2}} |\mathbf{Q}|^{-\frac{1}{2}} |\mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X}|^{-\frac{1}{2}} \int_{\mathbf{s}} \times I(\mathbf{s}) d\mathbf{s}) \tag{48}$$

where the integrand is:

$$I(\mathbf{s}) = \exp[-\tfrac{1}{2}(\mathbf{z} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{z} - \mathbf{h}(\mathbf{s})) + \mathbf{s}^T\mathbf{G}\mathbf{s}\}] \tag{49}$$

$$\mathbf{G} = \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}^{-1} \tag{50}$$

The estimate of $\mathbf{s}$ is simply the value that maximizes function I (s).[17]

6.3 Method MAP

The maximum *a posterior* method can be applied in different ways; here we will present the most straightforward implementation of the method (which may differ in details from other implementations). The objective is to maximize with respect to $\mathbf{s}$, β, and $\boldsymbol{\theta}$ the following expression:

$$p(\mathbf{z}, \mathbf{s}|\beta, \boldsymbol{\theta}) \propto |\mathbf{R}|^{-\frac{1}{2}}|\mathbf{Q}|^{-\frac{1}{2}} \times \exp[-\tfrac{1}{2}(\mathbf{z} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{z} - \mathbf{h}(\mathbf{s}))] \times \exp[-\tfrac{1}{2}(\mathbf{s} - \mathbf{X}\beta)^T\mathbf{Q}^{-1}(\mathbf{s} - \mathbf{X}\beta)] \tag{51}$$

We will minimize the negative logarithm of eqn (51):

$$L_{MAP} = \tfrac{1}{2}\ln|\mathbf{R}| + \tfrac{1}{2}\ln|\mathbf{Q}| + \tfrac{1}{2}(\mathbf{z} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{z} - \mathbf{h}(\mathbf{s})) + \tfrac{1}{2}(\mathbf{s} - \mathbf{X}\beta)^T\mathbf{Q}^{-1}(\mathbf{s} - \mathbf{X}\beta) \tag{52}$$

which means that the following conditions must be met:

$$\frac{\partial L_{\mathrm{MAP}}}{\partial \beta} = -(\mathbf{s} - \mathbf{X}\beta)^T\mathbf{Q}^{-1}\mathbf{X} = \mathbf{0} \tag{53}$$

$$\frac{\partial L_{\mathrm{MAP}}}{\partial \mathbf{s}} = -(\mathbf{z} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}\mathbf{H} + (\mathbf{s} - \mathbf{X}\beta)^T\mathbf{Q}^{-1} = \mathbf{0} \tag{54}$$

$$\frac{\partial L_{\mathrm{MAP}}}{\partial \theta_i} = \tfrac{1}{2}Tr[\mathbf{R}^{-1}\mathbf{R}_i] + \tfrac{1}{2}Tr[\mathbf{Q}^{-1}\mathbf{Q}_i] - \tfrac{1}{2}(\mathbf{z} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}\mathbf{R}_i\mathbf{R}^{-1}(\mathbf{z} - \mathbf{h}(\mathbf{s})) - \frac{1}{2}(\mathbf{s} - \mathbf{X}\beta)^T\mathbf{Q}^{-1}\mathbf{Q}_i\mathbf{Q}^{-1}(\mathbf{s} - \mathbf{X}\beta) = \mathbf{0} \tag{55}$$

The minimization is achieved using Gauss–Newton iterations.

6.4 Results

Application of the two methods for six conductivity and 12 head observations yielded the following parameter estimates

	GA	MAP
v	2.58	0.786
l	0.649	1.42
$\ln(\det(\mathbf{Q}))$	−251	−446
$\ln(\det(\mathbf{V}))$	−398	−561

The MAP method estimated a smaller variance but a larger correlation length than GA. These estimates produce different $\mathbf{Q}$ and $\mathbf{V}$ matrices (where $\mathbf{Q}$ is the prior and $\mathbf{V}$ is the posterior covariance matrix of $\mathbf{s}$). What is of

more interest in such inverse problems, however, is how good are the best estimates of s and their mean square error (MSE) of estimation. To measure how close the best estimate is to the actual, we use the actual mean square error which is 0.0562 for GA and 0.568 for MAP. That is, the two methods produced practically equally accurate best estimates. But the computed mean square error and the confidence intervals in the MAP method are too low. As Figs 1 and 2 show, the actual values were outside the 95% confidence bounds in five cases for GA but in 37 values for MAP (out of a total of 100). Thus, the MAP method does not evaluate appropriately the reliability of its estimates.

Note that both methods yield estimates that reproduce all observations equally well as shown in Figs 1 and 2, as well as in Figs 3 and 4 that show that the head is predicted with excellent precision despite errors in the log conductivity.

Additional numerical experiments (not reported here) support the contention that by increasing $\frac{m}{n}$, the ratio of pixels to observations, the MSE computed by MAP tends to zero, unlike the MSE computed by GA. For example, for a subset of the original data consisting of two conductivity and five head observations, the results are shown in Figs 5 and 6. As Fig. 6 demonstrates, 76 out of the 100 values estimated using MAP are outside the 95% confidence interval, indicating that MAP underestimates seriously the mean square error. The actual mean square error in this case is 0.347 for GA and 0.415 for MAP.

Finally, in order to illustrate the importance of structural analysis, consider that the structural parameters are not optimized but are arbitrarily set at $v = 3$ and $l = 0$. That is, assume that we perform only what in the linear case is known as "cokriging" and more generally as "weighted linear or nonlinear least squares". The GA and MAP methods yield the same best estimate of s (see Figs 7 and 8). It is clear that despite the reproduction of

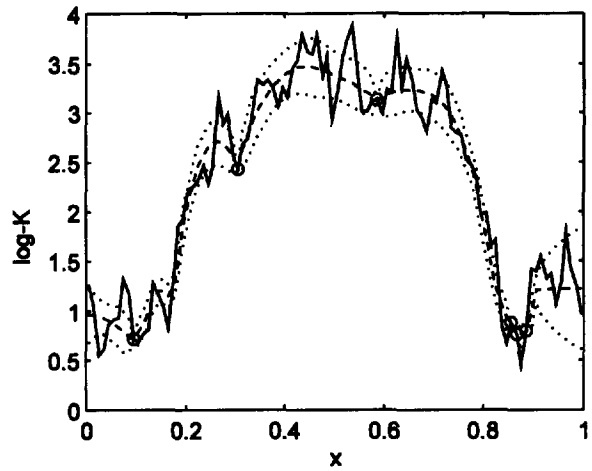

Fig. 2. Application of MAP. Log conductivity: actual (solid line), estimated (dashed line), 95% confidence interval (dotted lines), and observations (open circles).

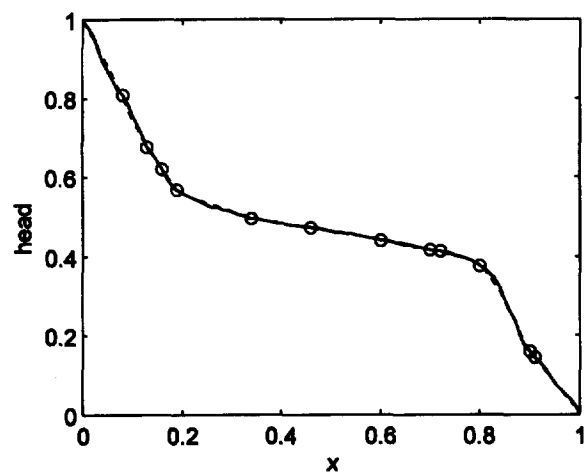

Fig. 3. Application of GA. Head: actual (solid line), computed for best estimate of log conductivity (dashed line), and observations (open circles).

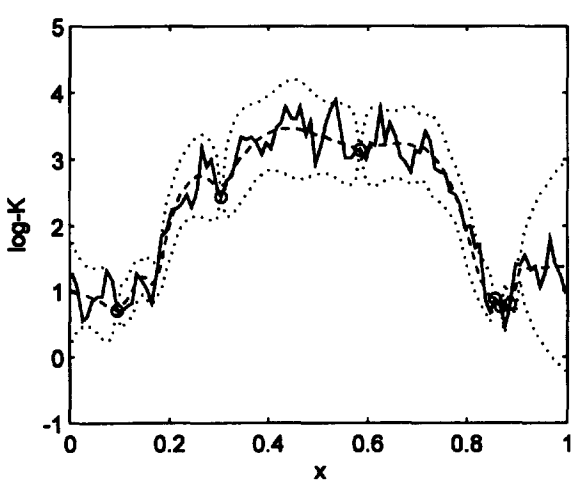

Fig. 1. Application of GA. Log conductivity: actual (solid line), estimated (dashed line), 95% confidence interval (dotted lines), and observations (open circles).

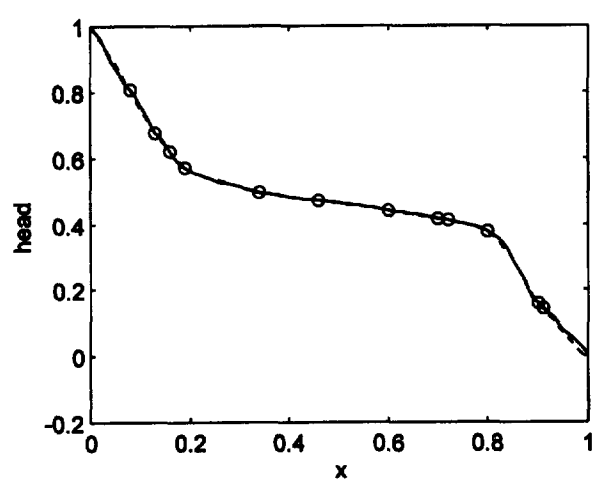

Fig. 4. Application of MAP. Head: actual (solid line), computed for best estimate of log conductivity (dashed line), and observations (open circles).

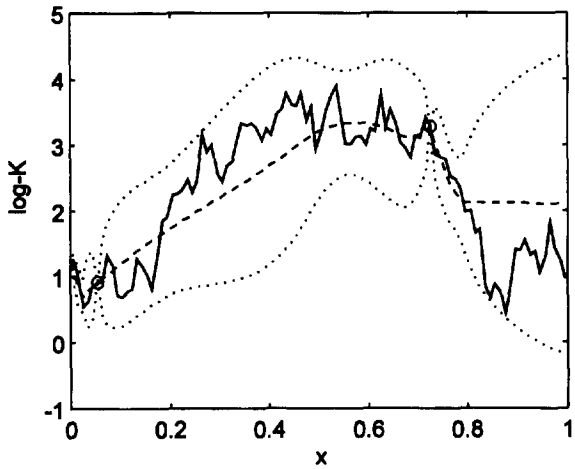

Fig. 5. Application of GA. Log conductivity: actual (solid line), estimated (dashed line), 95% confidence interval (dotted lines), and observations (open circles).

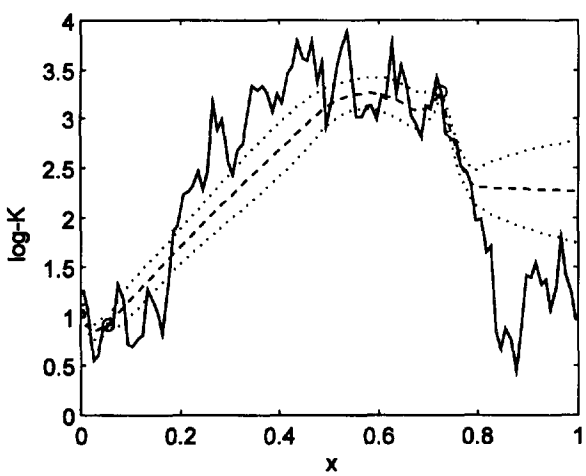

Fig. 6. Application of MAP. Log conductivity: actual (solid line), estimated (dashed line), 95% confidence interval (dotted lines), and observations (open circles).

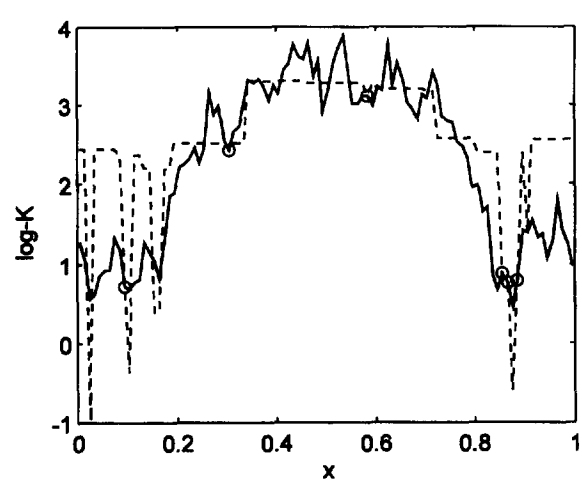

Fig. 7. Using suboptimal structural parameters. Log conductivity: actual (solid line), estimate (dashed line), and observations (open circles).

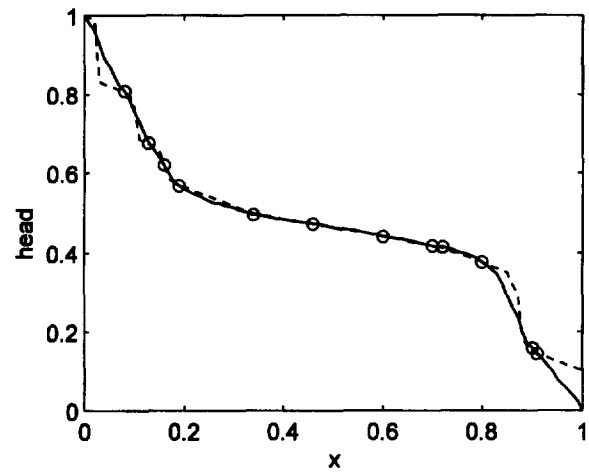

Fig. 8. Using suboptimal structural parameters. Head: actual (solid line), from best estimate of log conductivity (dashed line), and observations (open circles).

the data, this approach gives suboptimal results at locations where there are no observations.

7 CONCLUDING REMARKS

This paper made some points on the principles and application of the geostatistical approach to the inverse problem. In particular, the significance of the structural analysis part was emphasized. It is during the structural analysis that the important data fitting decisions are made. Furthermore, it was demonstrated that the MAP method of estimation of structural parameters differs in principle and in practice from the estimation method in Kitanidis and Vomvoris[18] and that the former is biased, unless the domain is descretized into only a few zones, while the latter is unbiased even if every node or element is treated as a separate zone.

ACKNOWLEDGMENTS

Funding for this work was provided by the office of Research and Development, U.S. Environmental Protection Agency, under agreement R-815738-01 through the Western Region Hazardous Substance Research Centre. The content of this study does not necessarily represent the views of the agency.

REFERENCES

1. Box, G. E. P. & Jenkins, G. M., *Time Series Analysis*, Holden-Day, San Francisco, 1976.
2. Carrera, J. & Neuman, S. P., Estimation of aquifer parameters under transient and steady state conditions, 1. Maximum likelihood method incorporating prior information. *Water Resour. Res.*, **22**(2) (1986) 199–210.
3. Carrera, J. & Glorioso, L., On geostatistical formulations

of the groundwater flow inverse problem. *Adv. Water Resour.*, **14**(5) (1991) 273–283.
4. Cooley, R. L., Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 1. Theory. *Water Resour. Res.*, **18**(4) (1982) 965–967.
5. Cooley, R. L., Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 2. Applications. *Water Resour. Res.*, **19**(3) (1983) 662–676.
6. Dagan, G., Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem. *Water Resour. Res.*, **21**(1) (1985) 65–72.
7. Dagan, G., *Flow and Transport in Porous Media*, Springer, Berlin, 1989, 465 pp.
8. Draper, N & Smith, H., *Applied Regression Analysis*, 2nd edition, Wiley Interscience, New York, 1981.
9. Gelhar, L. W., *Stochastic Subsurface Hydrology*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
10. Hoeksema, R. J. & P. K. Kitanidis, An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling. *Water Resour. Res.*, **20**(7) (1984) 1003–1020.
11. Hoeksema, R. J. & P. K. Kitanidis, Comparison of Gaussian conditional mean and kriging estimation in the geostatistical solution of the inverse problem. *Water Resour. Res.*, **21**(6) (1985) 825–836.
12. Hoeksema, R. J. & P. K. Kitanidis, Prediction of transmissivities, heads, and seepage velocities using mathematical models and geostatistics. *Adv. Water Resour.*, **12**(2) (1989) 90–102.
13. Hoeksema, R. J. & Clapp, R. B., Calibration of groundwater flow models using Monte Carlo simulations and geostatistics, in *ModelCARE 90: Calibration and Reliability in Groundwater Modelling* (pp. 33–42) IAHS Publ. No. 195, 1990.
14. Kitanidis, P. K., Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resour. Res.*, **22**(4) (1986) 499–507.
15. Kitanidis, P. K., Orthonormal residuals in Geostatistics: Model criticism and parameter estimation. *Mathematical Geology*, **23**(5) (1991) 741–58.
16. Kitanidis, P. K., Geostatistics. In *Handbook of Hydrology*, ed. D. R. Maidment, pp. 20.1–20.39, McGraw-Hill, New York, 1993.
17. Kitanidis, P. K., Quasilinear geostatistical theory for inversing. *Water Resour. Res.*, **31**(10) (1995) 2411–2419.
18. Kitanidis, P. K. & E. G. Vomvoris, A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations. *Water Resour. Res.*, **19**(3) (1983) 677–690.
19. Matheron, G., *The Theory of Regionalized Variables and its Applications*, 212 pp., Ecole de Mines, Fontainbleau, France, 1971.
20. Rao, C. R., *Linear Statistical Inference and Its Applications*, 2nd edition, Wiley, New York, 1973.
21. Rubin, Y. & Dagan, G., Stochastic identification of transmissivity and effective recharge in steady groundwater flow, 1. Theory. *Water Resour. Res.*, **23**(7) (1987a) 1185–1192.
22. Rubin, Y & Dagan, G., Stochastic identification of transmissivity and effective recharge in steady groundwater flow, 2. Case study. *Water Resour. Res.*, **23**(7) (1987b) 1193–1200.
23. Schweppe, F. C., *Uncertain Dynamic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
24. Wagner, B. J. & Gorelick, S. M., Reliable aquifer remediation in the presence of spatially variable hydraulic conductivity: From data to design. *Water Resour. Res.*, **25**(10) (1989) 2211–2225.
25. Yeh, W. W. G., Review of parameter identification procedures in groundwater hydrology: The inverse problem. *Water Resour. Res.*, **22**(1) (1986) 95–108.