

Comment on “A reassessment of the groundwater inverse problem” by D. McLaughlin and L. R. Townley

Peter K. Kitanidis

Civil Engineering, Stanford University, Stanford, California

Introduction

Many methods are available for the solution of the following inverse problem: determine the conductivity or transmissivity of a porous medium from head or pressure observations and other information. *McLaughlin and Townley* [1996] (hereinafter referred to as MT) review a family of methods and present maximum a posteriori estimation and a functional formulation as common characteristics of prevailing methodologies.

The authors deserve due credit for their efforts to unify seemingly disparate approaches. It is hard, however, to squeeze different methods into a narrow mold without knocking them out of shape. My objective is to clarify and supplement some important points made by MT in order to make this work more useful, particularly to readers who may be interested in using the “geostatistical approach.” I will focus on estimation (also known as inference) issues using the first two statistical moments, leaving aside the other important questions of how to solve the forward problem, determine sensitivity matrices, or use other statistics. My key points are as follows:

1. The geostatistical approach does not fit within the narrow definition of maximum a posteriori estimation adopted by MT; an alternative study of the relation between Bayesian estimation and geostatistical analysis is that of *Kitanidis* [1986].
2. The geostatistical approach is well suited to function estimation, a point that was missed by MT.
3. The structural analysis part of inverse modeling, which was neglected by MT is at least as important as the “least squares” or “conditioning” part that is the focus of MT.

Geostatistical Approach Versus MAP

It is essential to start with a synopsis of the geostatistical approach (GA). The objective is to estimate the log conductivity or other unknown function $s(x)$, where x indicates a location. It is immaterial whether the function is discretized into a vector \mathbf{s} . Discretization should be avoided when possible but is often essential for the solution of the direct problem solution through numerical methods. The crucial characteristic of the problem is that there are only n observations, whereas the number of unknowns, m , can increase without an inherent limit and in the nondiscretized case is essentially infinite.

Information about the function is represented through deterministic functions, namely, the mean and covariance functions, of a few parameters. First, The mean function may be expanded as

$$\mu(x) = \sum_{i=1}^p X_i(x)\beta_i \quad (1)$$

where $X_i(x)$ are the p basis functions of the mean and p is a small number. A common case is that of a constant mean, $p = 1$ and $X_1(x) = 1$. The covariance function is usually a function of the separation vector $q(x - x'; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are a few parameters that need to be adjusted from data fitting. The observations, arranged in vector \mathbf{y} , are related to the unknowns and are affected by observation error.

GA is not antithetical to zonation if zonation is based on solid information. For example, if three hydrogeologically distinct zones have been identified with a reasonable degree of certainty, they may be represented easily in the $\mu(x)$, by setting

$$p = 3 \quad X_i(x) = \begin{cases} 1 & x \text{ in zone } i \\ 0 & \text{otherwise} \end{cases}$$

However, it is always required that $p \ll n$.

There are three sets of unknowns: \mathbf{s} , $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$. If all of them are treated as parameters, the textbook maximum a posteriori (MAP) method is to maximize the probability density function of data with respect to these three parameter sets:

$$\text{MAP} := \max_{\mathbf{s}, \boldsymbol{\beta}, \boldsymbol{\theta}} p(\mathbf{y}|\mathbf{s}, \boldsymbol{\beta}, \boldsymbol{\theta}) \quad (2)$$

This is not what is done in GA, which was proposed by *Kitanidis and Vomvoris* [1983] and named geostatistical because of its resemblance with a methodology advanced by *Matheron* [1963] for the solution of some interpolation and averaging problems. Instead, the $\boldsymbol{\theta}$ parameters are found by maximizing the likelihood of the data \mathbf{y} conditional only on $\boldsymbol{\theta}$, and then the best linear unbiased estimate (BLUE) of \mathbf{s} is found conditional on the data.

$$\text{GA} := \begin{cases} \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}) \\ \text{then} \\ \text{BLUE of } \mathbf{s} \end{cases} \quad (3)$$

The MAP and GA approaches differ in philosophy, results, and implementation. The MAP approach makes sense only if the total number of unknowns is less than the number of observations. This is clearly not the case in the function estimation case where the number of unknowns is large or even infinite. In the final analysis, MAP is a fitting method, and it is unreasonable to fit, say, a thousand parameters to a hundred observations. That is why when implementing the complete MAP approach the function should be parameterized so that the effective number of parameters in \mathbf{s} is small, for instance, by discretizing the domain into a small number of “zones.” For example, see *Gavalas et al.* [1976], who used 8 or 16 uniform zones with 75 or 80 observations. Incidentally, it is common practice in weighted least squares to use models with few adjustable parameters [e.g., see *Cooley et al.*, 1986].

In contrast to the MAP approach, in the GA approach the parameters to be adjusted are basically $\boldsymbol{\theta}$. Then the conditional

mean and covariance of \mathbf{s} follow. There are no inherent limitations on the size of \mathbf{s} . A rigorous explanation of GA within the context of probability theory is given by *Kitanidis* [1986, 1995]. Intuitively, one may say that the method consists of first fitting $\boldsymbol{\theta}$ and then fitting \mathbf{s} , but from an estimation viewpoint, the first fitting is principal and the second fitting is subordinate.

MT noted that for the linear case (or one iteration of the Gauss-Newton method of successive linearizations) and for given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the MAP and BLUE approaches yield the same result. *Carrera and Glorioso* [1991] have demonstrated that, and I may add that the relation between Bayesian and geostatistical BLUE methods has been discussed extensively and under general conditions by *Kitanidis* [1986]. An earlier but more limited in scope treatment was presented by *Kimeldorf and Wahba* [1970]. It has thus been known that MAP and BLUE produce the same algorithm under the very restrictive conditions mentioned previously, but this should not leave the impression that the two approaches are identical. Philosophical differences aside, the MAP and GA approaches applied in their entirety do yield different results, as was demonstrated by *Kitanidis* [1996a]. In fact, GA works without function discretization, whereas the textbook MAP method works when \mathbf{s} is small in dimension (such as by discretizing the domain into a small number of zones). The two methods are implemented differently, as will be discussed next.

Function Estimation

MT describe the parameterization of *Hoeksema and Kitanidis* [1984] as “pixel” (which could be misinterpreted) and differentiate it from the case where the function is not discretized. However, the distinction between an arbitrarily large number of pixels and a continuous description has no fundamental significance. The Hoeksema and Kitanidis discretization was motivated solely by the need to solve the direct problem without making the assumptions needed for an analytical solution. Other than that, their method is the same as the method of *Kitanidis and Vomvoris* [1993], in which the function is not discretized. MT [p. 1145] also mention the methods of *Gavalas et al.* [1976] and *Hoeksema and Kitanidis* [1984] side by side when they discuss the computational difficulties associated with estimation applied with a pixel-type parametrization; this may leave the erroneous impression that the estimation is implemented in the same way in the two works. (The question of linear versus nonlinear estimation is a separate issue that has no bearing upon this discussion. Iterative BLUE estimation methods that solve nonlinear least squares problems have been discussed [Yeh et al., 1995; Kitanidis, 1995, etc.]

The crucial characteristic of the function-estimation problem is that the number of observations is fixed and usually small whereas the number of unknowns can be quite large and can increase without an inherent limit. It has been noticed in estimation applications [e.g., *Schweppe*, 1973, pp. 96–97] that methods can be formulated in two different ways depending on whether the number of observations exceeds the number of unknowns (as usually is the case in science and engineering) or the other way around (as is the case in function estimation). GA [e.g., *Kitanidis and Vomvoris*, 1983; *Hoeksema and Kitanidis*, 1984; *Dagan*, 1985] naturally chose the latter approach. In contrast, approaches that proceed from the minimization of a penalty functional (application of MAP) [e.g., *Gavalas et al.*, 1976; *Carrera and Neuman*, 1986; *Loaiciga and Marino*, 1987;

etc.] chose to implement the former approach, which makes sense when $m < n$.

MT seem to advance the continuous equivalent of the approach used in MAP. Their preoccupation with the penalty functional in the form that involves the inverse of the covariance of \mathbf{s} [MT, equation (49)], is puzzling. The inverse of the covariance (whether a generalized function or a matrix) is generally difficult to find and is not needed in function estimation using stochastic methods! For example, see *Kitanidis and Vomvoris* [1983], who solve the function estimation problem without ever having to estimate the inverse of the covariance function of \mathbf{s} and without discretizing. Additionally, MT’s emphasis on representing the function a posteriori through a series (infinite in the nondiscretized case, arbitrarily large in the discretized case) is odd considering that much better alternatives are around.

The seminal work of *Gavalas et al.* [1976] used a singular value decomposition of the prior covariance in order to reduce the effective number of parameters. Their approach, which is related to the overall approach suggested by MT, is not a panacea considering that (1) a straightforward decomposition has computational cost proportional to m^3 and (2) it is still an approximation when the approach in GA is exact.

GA effectively utilizes a series with a much smaller number of terms (equal to number of observations plus number of unknown drift coefficients). This issue has been discussed by *Cressie* [1993, p. 303] and *Kitanidis* [1996b] for kriging but applies to all BLUE, whether or not a function is discretized, and extends to nonlinear cases that involve successive linearizations. *Dagan* [1985] demonstrated this point clearly for the linear given-mean case. We can start with an intuitive argument: the “cokriging” weights and Lagrange multipliers are functions of space and can serve to define the set of basis functions. At each location, these weights and multipliers are found from the solution of a linear system of $n + p$ equations, but it was recognized early on that [*Hoeksema and Kitanidis*, 1984, p. 1008] “. . .only the right sides [of the system] change for each new estimation location. This allows great computational efficiency, since the matrix of left side constants needs to be factored only once.”

Next, to make the same point more formally and explicitly, let us focus on the best estimate. Let \mathbf{y} be the n by 1 vector of observations, \mathbf{s} the m by 1 the vector of unknowns, and $\hat{\mathbf{s}}$ the vector of its estimates. The prior statistics of \mathbf{y} and \mathbf{s} are

$$E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}, \quad E[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T] = \mathbf{Q}_{yy}$$

$$E[\mathbf{s}] = \mathbf{X}_0\boldsymbol{\beta}, \quad E[(\mathbf{s} - \mathbf{X}_0\boldsymbol{\beta})(\mathbf{s} - \mathbf{X}_0\boldsymbol{\beta})^T] = \mathbf{Q}_{00}$$

$$E[(\mathbf{s} - \mathbf{X}_0\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T] = \mathbf{Q}_{0y} \quad (4)$$

(We start with \mathbf{X}_0 and \mathbf{Q}_{00} and then compute \mathbf{X} , \mathbf{Q}_{0y} , and \mathbf{Q}_{yy} by solving a direct problem.) Using equation (7) of *Kitanidis* [1986, p. 501], the best estimate is

$$\begin{aligned} \hat{\mathbf{s}} &= [\mathbf{Q}_{0y}\mathbf{Q}_{yy}^{-1} + (\mathbf{X}_0 - \mathbf{Q}_{0y}\mathbf{Q}_{yy}^{-1}\mathbf{X})(\mathbf{X}^T\mathbf{Q}_{yy}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}_{yy}^{-1}] \mathbf{y} \\ &= \mathbf{X}_0\mathbf{b} + \mathbf{Q}_{0y}\boldsymbol{\xi} \end{aligned} \quad (5)$$

where we defined

$$\mathbf{b} = (\mathbf{X}^T\mathbf{Q}_{yy}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}_{yy}^{-1}\mathbf{y}$$

$$\boldsymbol{\xi} = (\mathbf{Q}_{yy}^{-1} - \mathbf{Q}_{yy}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{Q}_{yy}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}_{yy}^{-1}) \mathbf{y} \quad (6)$$

Equation (5) shows that the best estimate is expanded exactly on a basis consisting of the p columns of \mathbf{X}_0 and the n columns of \mathbf{Q}_{0y} , no matter how large the dimension of the unknown vector. One may organize the computations as follows: solve one system of $n + p$ linear equations

$$\begin{bmatrix} \mathbf{Q}_{yy} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \xi \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} \quad (7)$$

and then perform $n + p$ multiplications for each sampling location of s (equation (5)). Note the following points: (1) The cost of obtaining the best estimate in GA is proportional to m whereas in the “other” approach where a system of m equations is solved is proportional to m^3 . (2) The issue of sampling a function is separate from the discretization that may be needed to derive the cross covariance \mathbf{Q}_{0y} and the measurement error covariance matrix \mathbf{Q}_{yy} . (3) The “analytical result” is contained trivially in this solution: if one considers that \mathbf{s} is a scalar, the value sampled at generic location x , equation (5) becomes

$$s(x) = \sum_{i=1}^p X_i(x)b_i + \sum_{i=1}^n q_{sy}(x, i)\xi_i \quad (8)$$

where $X_i(x)$ is from equation (1) and $q_{sy}(x, i)$ is a function of x that gives the cross covariance between $s(x)$ and observation y_i . This result is analytical and exact. Of course, even an analytically derived function needs to be sampled on a grid before it can be plotted, and the vectorized equation (5) is better suited for that purpose. Finally, (d) In the nonlinear case, the same approach is applied except that the $q_{sy}(x, i)$ functions need to be recomputed at every iteration along with other quantities affected by the linearization, as was essentially done by *Yeh et al.* [1995] and *Kitanidis* [1995].

The generation of a conditional realization, as discussed by *Kitanidis* [1995], involves the generation of an unconditional realization (which can be accomplished efficiently using methods outside of the realm of inverse problems) followed by the solution of a problem similar to the determination of the best estimate. The computation of variances of estimation or of the complete covariance is organized in GA along similar lines. I thus wonder what are the advantages of computing the covariance through, say, equation (D12) of MT compared with computing the covariance through an equivalent cokriging formulation, such as equation (24) of *Kitanidis* [1995], which serves the same purpose and is easier to compute.

For the application of GA, it is required to have a nonsingular matrix of coefficients in the “cokriging system”; this matrix in the special case considered by MT is the same as the covariance matrix of data. However, I sensed from the discussion in MT [pp. 1131, 1149] a confusion between the real issue of ill-posedness with the nonissue of an ill-conditioned measurement covariance matrix. An ill-conditioned covariance matrix, if we rule out a defective formulation such as using inappropriate functions as covariance functions or badly chosen units that represent observations of different quantities, indicates redundant observations and can be easily fixed. Of course, this embarrassment of riches (redundant data) is the exact opposite of the root cause of the inherent ill-posedness of function estimation (insufficiency of data). In applications, there is always plenty of “observation error,” so there are no redundant observations and the issue of an ill-conditioned matrix of coefficients should not even arise. Thus I see no

problem in applying the geostatistical approach to estimate a function.

Prior Statistics and Structural Analysis

The focus of MT is on the case that the mean and covariance functions $\mu(x)$ and $q(x, x')$ are given. They interpret these functions as prior statistics, which strictly speaking means independent of the observations and based only on other information. They add [MT, p. 1150], “One might argue that it is not essential, or even desirable, to estimate regularization parameters from field data.” (Regularization parameters in our context means the parameters that control $\mu(x)$ and especially $q(x, x')$.)

However, if this view were adopted and consistently applied, one would have to choose the $\mu(x)$ and $q(x, x')$ functions without even peeking at the data, which I would argue is inexpedient, or even impossible. Usually, $\mu(x)$ is practically unknown a priori so that a large prior variance would have to be introduced in order to avoid biasing the results towards an unreliable $\mu(x)$. Consider for example the similar problem of interpolation: the first step in practice is to examine the data for guidance on what variogram, spline, or weight function to use. Of course, I am taking for granted the availability of a reasonable number of observations and of an acceptable conceptual model, because otherwise it is pointless to talk about formal approaches to interpolation or inverse problems.

Geostatistics takes the view that $\mu(x)$ and $q(x, x')$ are consistent both with prior information and with the data. The process of selecting these functions is known as structural analysis. One aspect of structural analysis is the estimation of the structural parameters. *Kitanidis and Vomvoris* [1983], *Hoeksema and Kitanidis* [1984], and others assume no prior information about the parameters β of the mean functions and estimate the covariance parameters θ through the method of restricted maximum likelihood (RML). RML in the context of inverse problems can be interpreted intuitively as a “cross-validation” method, in which the inverse method is trained on the available data to improve its predictive performance through adjustment of its parameters θ (see discussion by *Kitanidis* [1991]).

Structural analysis, which is so underrated by MT, is crucial in function estimation. Inverse problems are basically data fitting problems and MT present equations (48) or (49) in MT as general fitting criteria. However, “fitting” has different significance in algebraically overdetermined problems (e.g., fit 2 parameters to 100 observations) and in algebraically underdetermined problems (e.g., fit 1000 parameters to 40 observations). Function-estimation problems are underdetermined: if one can find a function that fits the data, one can find other functions that do the same. Different \mathbf{C}_v , \mathbf{C}_a , and $\bar{\mathbf{a}}$ in (48) may lead to substantially different best estimates and particularly confidence intervals. Note that in underdetermined problems “a straightforward minimization without considering the uncertainty and reliability of the estimated parameters will result in meaningless parameter estimates” [*Yeh and Yoon*, 1981, p. 665]. My point is that the optimization that leads to equation (48), i.e., structural analysis, is even more interesting than the optimization that follows it.

MT raise the issue of estimation accuracy of the structural parameters and quote other sources that suggest that large sample sizes are needed to obtain reasonable estimates of covariance parameters. There have been conflicting statements

in the literature, but take into consideration that θ estimation may take place in different situations and with separate objectives. One situation is where one constructs realizations according to a certain probabilistic model and then evaluates how many observations are needed in order to identify the actual model and the actual parameters. Obviously, complex models with many parameters are hard to identify from data, particularly (1) when a generalized covariance function (GCF) is mistaken for an ordinary covariance function, (2) when RML is used to identify covariance parameters when it is intended for GCF (for a practical discussion of the differences, see Kitanidis, 1993]), or (3) when one judges the accuracy of the results with excessively high standards. This mathematical exercise is in any case interesting and germane to some problems of unconditional probabilities but is irrelevant to inverse problems. Formation parameters were likely not generated according to a specific probabilistic model, and the success of our estimation is not based on the premise that we have identified this model and the "actual" parameters.

Instead, the functions $\mu(x)$ and $q(x, x')$ are empirical models that are convenient means to quantify information about the structure. The parameters that need to be estimated are the parameters of the generalized covariance function that must be parameterized parsimoniously so that distinct parameters correspond to distinct and significant structural features in the data. What are the compelling reasons for introducing into an empirical model adjustable parameters that cannot be identified with acceptable precision? See *Box and Jenkins* [1976], who discuss the building of empirical models and the role of parsimony, and *Wahba* [1990], who criticizes the use of models with redundant parameters in interpolation through splines. Instead, one should use simple but comprehensive stochastic models with few parameters. In the uncommon case that a stochastic model has been ordained and final predictions are sensitive to uncertain parameters, Bayesian methods can be used to incorporate the effect of parameter uncertainty on predictions [e.g., Kitanidis, 1986].

Furthermore, the menu of GCF is broader than the menu of ordinary covariance functions. In my view, the most useful models in inverse modeling are generalized covariance functions. Certainly the models most useful in spline and Tikhonov regularization methods correspond to generalized covariance functions (such as the thin plate spline in two dimensions, or the cubic spline in one dimension). In contouring, the linear and modified linear [Hardy, 1990] variograms appear to be the most widely used models. These models have only one or two parameters that need to be estimated from data.

References

- Box, G. E. P., and G. M. Jenkins, *Time Series Analysis*, 575 pp., Holden-Day, Merrifield, Va., 1976.
- Carrera, J., and L. Glorioso, On geostatistical formulations of the groundwater flow inverse problem, *Adv. Water Resour.*, 14(5), 273–283, 1991.
- Carrera, J., and S. P. Neuman, Estimation of aquifer parameters under transient and steady state conditions, 1, Maximum likelihood method incorporating prior information, *Water Resour. Res.*, 22(2), 199–210, 1986.
- Cooley, R. L., L. F. Konikow, and R. L. Naff, Nonlinear-regression groundwater flow modeling of a deep regional aquifer system, *Water Resour. Res.*, 22(13), 1759–1778, 1986.
- Cressie, N. A. C., *Statistics for Spatial Data*, 900 pp., John Wiley, New York, 1993.
- Dagan, G., Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem, *Water Resour. Res.*, 21(1), 65–73, 1985.
- Gavalas, G. R., P. C. Shah, and J. H. Seinfeld, Reservoir history matching by Bayesian estimation, *Soc. Pet. Eng. J.*, 16, 337–350, 1976.
- Hardy, R. L., Theory and applications of the multiquadric-biharmonic method: 20 years of discovery, *Comput. Math. Appl.*, 19(8/9), 163–208, 1990.
- Hoeksema, R. J., and P. K. Kitanidis, An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling, *Water Resour. Res.*, 20(7), 1003–1020, 1984.
- Kimeldorf, G., and G. Wahba, A correspondence between Bayesian estimation of stochastic processes and smoothing by splines, *Ann. Math. Stat.*, 41, 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., Orthonormal residuals in geostatistics: Model criticism and parameter estimation, *Math. Geol.*, 23(5), 741–758, 1991.
- Kitanidis, P. K., Generalized covariance functions in estimation, *Math. Geol.*, 25(5), 525–540, 1993.
- Kitanidis, P. K., Quasilinear geostatistical theory for inverting, *Water Resour. Res.*, 31(10), 2411–2419, 1995.
- Kitanidis, P. K., On the geostatistical approach to the inverse problem, *Adv. Water Resour.*, 19(6), 333–342, 1996a.
- Kitanidis, P. K., Analytical expressions of conditional mean, covariance, and sample functions in geostatistics, *J. Stoch. Hydrol. Hydraul.*, 10(4), 279–294, 1996b.
- Kitanidis, P. K., and E. G. Vomvoris, A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, *Water Resour. Res.*, 19(3), 677–690, 1983.
- Loaiciga, H. A., and M. A. Marino, The inverse problem for confined aquifer flow: Identification and estimation with extensions, *Water Resour. Res.*, 23(1), 92–104, 1987.
- Matheron, G., Principles of geostatistics, *Econ. Geol.*, 58, 1246–1266, 1963.
- McLaughlin, D., and L. R. Townley, A reassessment of the groundwater inverse problem, *Water Resour. Res.*, 32(5), 1131–1161, 1996.
- Schweppe, F. C., *Uncertain Dynamic Systems*, 563 pp., Prentice-Hall, Englewood Cliffs, N. J., 1973.
- Wahba, G., Comment on a paper by Cressie, *Am. Stat.*, 44(3), 255–256, 1990.
- Yeh, T.-C. J., A. L. Gutjahr, and M. Jin, An iterative cokriging-like technique for groundwater flow modeling, *Ground Water*, 33(1), 33–41, 1995.
- Yeh, W. W.-G., and Y. S. Yoon, Aquifer parameter identifiability with optimum dimension in parameterization, *Water Resour. Res.*, 17(3), 664–672, 1981.

P. K. Kitanidis, Civil Engineering, Stanford University, Stanford, CA 94305-4020. (e-mail: pkk@ce.stanford.edu)

(Received August 19, 1996; revised October 24, 1996; accepted April 4, 1997.)