

## TITLES RECENTLY PUBLISHED BY IAHS

- Land Subsidence. Proceedings of the Venice Symposium, March 1984  
 Publ.no.151 (1986), price \$45
- Experiences in the Development and Application of Mathematical Models in Hydrology and Water Resources in Latin America (invited in Spanish). Proceedings of the HYDROMATH Tequi-igilpa (Honolulua) Symposium, September 1983  
 Publ.no.152 (1985), price \$30
- Scientific Basis for Water Resources Management. Proceedings of the Jerusalem Symposium, September 1985  
 Publ.no.153 (1985), price \$42
- Hydrology in the Service of Man, volumes 1-4. Proceedings of the IAHS/IAHS Cambridge Symposium, September 1985  
 Publ.no.154 (1985), price \$40 the set
- Proceedings of the symposia held during the Second IAHS Assembly, Budapest, July 1986:
- Modelling Seasonally-Induced Processes  
 Publ.no.155 (1986), Price \$40
- Conjunctive Water Use  
 Publ.no.156 (1986), price \$48
- Monitoring to Detect Changes in Water Quality Series  
 Publ.no.157 (1986), price \$40
- Integrated Design of Hydrological Networks  
 Publ.no.158 (1986), price \$40
- Drainage Basin Sediment Delivery. Proceedings of the Albuquerque Symposium, August 1986  
 Publ.no.159 (1986), price \$45
- Hydrologic Applications of Space Technology. Proceedings of the Cocoa Beach Workshop, August 1985  
 Publ.no.160 (1986), price \$45
- Kart Water Resources. Proceedings of the Ankara Symposium, July 1985  
 Publ.no.161 (1986), price \$45
- Atmospheric Forecasts, Management and Effects. Proceedings of the Davos Symposium, September 1986  
 Publ.no.162 (1987), price \$50
- Developments in the Analysis of Groundwater Flow Systems. Report prepared by a Working Group of the IAHS International Commission on Groundwater  
 Publ.no.163 (1986), price \$35
- Water for the Future: Hydrology in Perspective. Proceedings of the Rome Symposium, April 1987  
 Publ.no.164 (1987), price \$50
- Erosion and Sedimentation in the Pacific Area. Proceedings of the Corvallis Symposium, August 1987  
 Publ.no.165 (1987), price \$55
- Proceedings of the symposia held during the IUGG Assembly, Vancouver, August 1987
- Large Scale Effects of Seasonal Snow Cover  
 Publ.no.166 (1987), price \$42
- Forest Hydrology and Watershed Management  
 Publ.no.167 (1987), price \$35
- The Influence of Climate Change and Climatic Variability on the Hydrologic Regime and Water Resources  
 Publ.no.168 (1987), price \$55
- Irrigation and Water Allocation  
 Publ.no.169 (1987), price \$32
- The Physical Basis of Ice Sheet Modelling  
 Publ.no.170 (1987), price \$40
- Hydrology 2000. Report of the IAHS Hydrology 2000 Working Group  
 Publ.no.171 (1987), price \$22
- Side Effects of Water Resources Management. Report prepared by an IHP-III Working Group  
 Publ.no.172 (1988), price \$40
- Office of the Treasurer IAHS  
 2000 Florida Avenue NW  
 Washington, DC 20009  
 USA
- Bureau des Publications  
 de l'UICGI  
 140 Rue de Grenelle  
 75700 Paris, France
- IAHS Press  
 Institute of Hydrology  
 Wallingford, Oxfordshire  
 OX10 8BB UK
- A copy of a catalogue of IAHS publications and information on individual membership of IAHS may be obtained free of charge from any of these addresses.

## CONSEQUENCES OF SPATIAL VARIABILITY IN AQUIFER PROPERTIES AND DATA LIMITATIONS FOR GROUNDWATER MODELLING PRACTICE

By

Adrian Peck  
 Steven Gorelick  
 Christian de Marsily  
 Stephen Foster  
 Vladimir Kovalevsky

### Contributors

Jean Bahr	Dan Rosbjerg
Bryson Bates	Nicholas Sitar
James Butler	Lloyd Townley
Jeffrey Cawfield	Dirk Van Rooy
Jaime Gómez-Hernández	Brian Wagner
Peter Kitanidis	



A contribution to the International  
 Hydrological Programme of UNESCO  
 (Project 2.4d)

Report prepared by a Working Group  
 of the International Commission on  
 Groundwater (ICGW) of the International  
 Association of Hydrological Sciences



IAHS Publication No. 175

PLEASE SEND ORDERS AND/OR ENQUIRIES TO:

# THE CONCEPT OF PREDICTIVE PROBABILITY AND A SIMPLE TEST FOR GEOSTATISTICAL MODEL VALIDATION

Peter K. Kitanidis

Civil Engineering  
Stanford University  
Stanford, CA 94305

## Abstract

The concept of predictive probability is discussed as a unified way to develop validation tests. The predictive probability for the model used in linear geostatistics is derived and is used to derive one validation test.

## 1. Introduction

### Deduction Versus Induction

"Deduction, based on the temporary pretense that the current model is true, is attractive because it involves the statistician in 'exact' estimation calculations which he alone controls. By contrast, induction testing on the idea that the current model may not be true is messy and the statistician is much less in control. His role is now to present analyses in such a form, both numerical and graphical, as will accurately portray the current situation to the investigators' mind and appropriately stimulate his colleagues' imagination, leading to the next step. Although this inductive jump is the only creative part of the cycle and hence scientifically the most important, the statistician's role in it may appear inexact and indirect."

G.E.P. Box [1980, page 425]

The emphasis of statistical methods in the water sciences has been on parameter estimation and prediction, i.e., deduction, with little attention paid to model validation.

i.e., inductive testing. Indeed, at the earlier stages of stochastic hydrology, it was common for a model to be justified on little more than its ability to preserve certain characteristics of the historical record, such as mean values, variances, skewness, and correlation coefficients. In groundwater modeling, a groundwater flow model would be considered calibrated when it reproduced some measurements (usually pressure head). Parameter estimation methods would then focus on the purely algorithmic problem of how to minimize some fitting criterion.

Of course, the approach of "preserving observed features of the system" is perfectly reasonable if the underlying model is accepted as true and the data are reliable and relevant. However, one may wonder how the predictive capability of a model may be improved by preserving measurement error or spatial variability at scales much smaller than the model could reasonably reproduce. For example, a regional groundwater flow model, discretized on a scale of kilometers cannot account for random measurement error or for variability at a scale of, say, meters, even though such variability is evident in the data. Do we achieve anything by adjusting parameters in the regional model to reproduce this type of variability? Also, a model with many parameters may reproduce every available measurement, particularly if there are few measurements and many parameters, as is often the case in hydrogeology. What does this tell about the predictive capability of the model?

There have, of course, been exceptions. A case in point is geostatistical analysis, which has developed somewhat separately. If anything, the geostatistical literature, which is practical and pragmatic, has gone to the other extreme by deemphasizing the algorithmic problem of parameter estimation and relying heavily on "validation" in model development. Also, stochastic hydrology now makes wide use of ARIMA techniques [Box and Jenkins, 1976], which emphasize parsimonious modeling and validation rather than moment preservation.

The need to shift some of the emphasis from modeling and parameter estimation to validation becomes gradually accepted in the hydrologic community. For example, according to Klemes [1986, page 16]:

"...Simulation (with models which cannot be directly tested) has been a safe game for the modeller but a precarious one for the user. The missing corrective potential of comparing simulation results with the actuality has encouraged proliferation of both deterministic and stochastic simulation models and exaggerated claims of their performance, which has led to their uninhibited application to problems far beyond their capabilities."

This work deals with model validation for spatial processes, particularly in conjunction with linear geostatistics, i.e., linear minimum-variance unbiased estimation methods. These methods are well known for their application to simple interpolation problems, such as kriging to estimate from measurements values on a regular grid for purposes of contour mapping. A better kept secret is that the applicability of estimation methods is much wider than interpolation or curve fitting. In the author's opinion, these methods provide the environment in which data, our description of the physics of hydrologic processes, and probability theory come together to improve understanding of complex hydrologic processes or systems.

### Model Validation

Model is defined as the assembly of all assumption which allow to make predictions. For example, in groundwater modeling they include assumptions about

- (i) The physics of the process, such as continuity equation, Darcy's law, two or three-dimensional flow.

- (ii) The input or boundary conditions. They include leakage, pumping, fixed head or discharge, and effect of streams.
- (iii) The spatial variability of parameters. For example, the transmissivity may be assumed constant in zones or varying in a way which can be represented using geostatistical methods.
- (iv) The structure of uncertainty or error. It may be reasonable to hypothesize that all head measurements are contaminated with measurement errors which are uncorrelated random variables with zero mean and constant variance. Processes that are neglected or simplified, such as representing three-dimensional flow by a two-dimensional one, introduce an error whose statistical structure may be part of the model.

(i) and (ii) are often considered the object of physics-based mathematical modeling while (iii) and (iv) are the often considered the object of statistics. While the misconception that the one precludes the other still persists, the idea that the most fruitful approach may be the combination of the two gradually gains wider acceptance.

The objective of every validation test is the uncovering of inconsistencies between the assumed model and the data. (It is assumed, of course, that the model itself is free of logical inconsistencies.) Discrepancies, if their presence is ascertained, must be resolved by modifying either the model or some of the data. For example, the initially contemplated model of a regional aquifer turned out to be incompatible with the data in the case examined by Hoeksema and Kitanidis [1984], prompting a revision which resulted in introducing previously neglected leakage.

A fundamental and widely accepted principle is that the data used for validation must not have been taken into account in the process of model calibration or even model selection. Split-sample validation procedures divide the data base into two segments; the one used in model development and calibration and the other only in model validation. The model is judged acceptable when it reproduces that data almost equally well in both segments. Split-sample tests appear to simulate the actual prediction process in a realistic way and are the ones practitioners respect the most. However, in the case of small samples, the verdict may be overly dependant on the way the data were divided. Residual-testing is another commonly used method. [see Beisley et al., 1980].

## 2. A Brief Review of Validation Methods Used in the Estimation of Spatial Processes

Geostatistical estimation methods have long recognized the importance of testing the validity of the fitted model before it is accepted and used for predictions. Such tests become particularly important in complicated cases, such as those with variable drift or the joint analysis of logtransmissivity and head. In such cases, simple graphical methods of model calibration and validation, such as plotting of raw or experimental variograms, are not adequate. In practice, the model is "validated" by focusing on specific cases and ascertaining that the observed behavior is consistent with the expected one.

In applied geostatistics, an important problem is to test the validity of the (usually fitted) variogram or generalized covariance function. The following method is often used. Assume that we have a sample consisting of  $n$  point measurements  $y_1, \dots, y_n$ . Drop measurement  $y_i$ . Then, using the remaining measurements and the variogram, estimate through kriging the value of  $y$  at the location of the dropped measurement and its estimation variance,  $\hat{y}_i$  and  $\sigma_i^2$ , respectively. If the model is correct, the actual standardized error (or "residual")

$$e_i = \frac{y_i - \hat{y}_i}{\sigma_i} \tag{1}$$

is the sampled value of a random variable with an average value of 0 and a mean square value of 1. Since we have at our disposal  $n$  such values, each obtained by dropping one measurement at a time, we can test whether this is the case. Thus, calculate

$$S_1 = \frac{1}{n} \sum_{i=1}^n e_i \tag{2}$$

$$S_2 = \frac{1}{n} \sum_{i=1}^n e_i^2 \tag{3}$$

The first number must be near zero while the second must be near one. If this is not the case, the modeler may reject the model.

In the application of this test the key question is: How large must the difference between the actual values (calculated from Equations 2 and 3) and the "theoretical values" (0 and 1, respectively) be before it becomes a good reason to reject a particular model? Or, to put it another way, when is this difference significant rather than the result of random sampling error? Before answering this question in an objective way, one must derive a more complete probabilistic description of the statistics  $S_1$  and  $S_2$ , used for testing. It is desirable to obtain the sampling distribution (or at least its variance) given that the right model is used. Unfortunately, the distributions of  $S_1$  and  $S_2$  are not known a priori and this difficulty limits the usefulness of these statistics despite their intuitive appeal. The sampling distributions of  $S_1$  and  $S_2$  can be derived for Gaussian data [see Kitanidis, 1988] but the procedure is computationally intensive and is seldom used in applications.

An intuitively appealing way to compare different models is the following [DeLfiner, 1976; Kafritsas and Bras, 1981]. For each measurement location use kriging to predict the value using other measurements. This is done for all models. The model which gives the smallest absolute kriging error is given a grade of 1; the second best is given a grade of 2; and so on. The grade of each model may then be averaged over all points kriged and the model with the lowest average grade may be picked as the one which performs the best among the compared models.

In addition to looking at overall statistics, one may study the empirical frequency curve or histogram of the standardized residuals. Outliers, excessive skewness, or bimodal distributions may be detected by visual inspection. In current geostatistical practice the  $i$ th observation is considered an outlier when  $e_i$  is larger than 3. Another useful approach is plotting on normal probability paper. Of course, such criteria rely on a Gaussian assumption.

A limitation of the standardized residuals is that they are not independent so that one cannot apply the usual goodness-of-fit tests. Kitanidis and Vomvoris [1983] recommended using normalized rather than standardized residuals. These are linear combinations of the measurements which, assuming that the true model is used, do not depend on the drift coefficients, have zero mean and unit variance. Thus, their average is approximately normally distributed, with zero mean and variance  $1/(n-p)$  and their sum of squares is approximately chi-squared distributed with  $n-p$  degrees of freedom, where  $n$  is number of measurements and  $p$  is number of drift coefficients. Standard checks for verifying normality or detecting outliers may be applied.

This paper has two objectives: First, to review the rudiments of a general approach for development of validation tests. Second, to apply this approach to justify theoretically and to generalize tests previously used in Kitanidis and Vomvoris [1983]. Hoeksema and

### 3. The General Problems of Parameter Estimation, Validation and Prediction

Assume that there are  $n$  measurements arranged, for the sake of notational convenience, in an  $n \times 1$  vector,  $y$ . For example, these may be measurements of head or logtransmissivity at points, of weighted averages over given areas, or of gradients of these functions. Following the notation of Kiandis [1986], let  $y_0$  be the  $m \times 1$  vector of unknown quantities, i.e., quantities to be predicted from the measurements. Let  $M$  stand for the model, i.e., the assembly of all assumptions which allow us to make predictions on the basis of "structural" parameters which are not perfectly known a priori but are to be estimated from the data. Examples of parameters are the mean, variance, and integral scale (correlation length) of a stationary process with exponential variogram. These parameters are incidental, i.e., they are of interest only to the extent that they affect the inference about the unknown  $y_0$ . Given the model and its parameters  $\theta_M$ , one can calculate the joint probability density function (pdf) of  $y$  and  $y_0$ , denoted by  $(y, y_0 | M, \theta_M)$ . Indeed, one may implicitly define the model as the set of assumptions which allow the calculation of this pdf.

#### Parameter Estimation

Let  $p^*(\theta_M | M)$  be the prior pdf of model parameters assuming that model  $M$  is valid. This pdf summarizes all available information about the parameters which is "prior" to measurements  $y$ . The pdf which combines prior and data information is the posterior,  $p^*(\theta_M | M, y)$ , which can be calculated through application of Bayes' theorem

$$p^*(\theta_M | M, y) = c p(y | M, \theta_M) p^*(\theta_M | M) \quad (5)$$

where  $p(y | M, \theta_M)$  is the likelihood of the data given  $\theta_M$ , and  $c$  is a positive constant such that  $p^*(\theta_M | M)$  is a proper pdf.

#### Prediction

The joint pdf of the unknowns,  $y_0$ , and the structural parameters,  $\theta_M$ , given the model, and data may be factored into the product of a conditional times a marginal, always given  $M$  and  $y$ :

$$p(y_0, \theta_M | M, y) = p(y_0 | M, \theta_M, y) p^*(\theta_M | M) \quad (6)$$

The pdf of the unknowns is the marginal

$$\begin{aligned} p(y_0 | M, y) &= \int \theta_M p(y_0 | M, \theta_M, y) p^*(\theta_M | M) d\theta_M \\ &= \int \theta_M \frac{p(y_0 | M, \theta_M)}{p(y | M, \theta_M)} p^*(\theta_M | M) d\theta_M \end{aligned} \quad (7)$$

where multiple integrals are represented in vector notation

$$\int f(x) dx = \int x_1 \int x_2 \dots \int x_n f(x_1, x_2, \dots, x_n) dx_1 \dots dx_n \quad (8)$$

It is important to point out that the essence of the integration in Equation (7) is that uncertainty in structural parameters is taken into account. Equation (5) through (7) are Model Validation

In the problems of parameter estimation and prediction, the model is assumed given, a fact which has been emphasized by making all distributions conditional on  $M$ . In model validation, however, it is the validity of the model itself which is questioned. The means to perform our critique is to test whether the actual data was likely to be generated from the model and the parameters which are plausible given prior information. The joint pdf of  $y$  and  $\theta_M$  may be factored into a conditional pdf times a marginal pdf,

$$p(y, \theta_M | M) = p(y | M, \theta_M) p^*(\theta_M | M) \quad (9)$$

Integration over all possible values of the parameters we obtain the pdf of the measurements given only the model,

$$p(y | M) = \int p(y | M, \theta_M) p^*(\theta_M | M) d\theta_M \quad (10)$$

The expression of Eq. (10) is known as the predictive probability density function [Box, 1980]. On the basis of the predictive distribution, one could devise tests to evaluate the likelihood of the actual data,  $y$ , given the model and prior information about the parameters.

#### The General Linear Model

From the perspective of probability theory Equations (5) through (10) represent the complete solutions to the problems of parameter estimation, prediction, model validation, and selection. In practice, of course, these equations are only a theoretically sound foundation on which one can build practical methods to solve actual problems. The difficulties associated with the direct application of these equations are:

- a) These equations generally cannot be integrated analytically and the numerical quadrature of multiple integrals may be prohibitively expensive, especially when there are hundreds of measurements and unknowns and several structural parameters; and
- b) In some of the most important problems, especially in groundwater modeling, the vectors of measurements and unknowns involve quantities which are related through partial differential equations. A case in point is head and logpermeability which are related through continuity and Darcy's law. Evaluating  $p(y_0 | M, \theta_M)$  means solving nonlinear stochastic differential equations. In practice, only a few lower moments of this distribution can be calculated.

The most important special case, which is the subject of linear geostatistics, is known as the general linear model. In summary, the measurements are modeled as a realization of a vector stochastic process whose mean is proportional to some parameters and whose covariance model has a given form but depends on some parameters which must be determined from data. That is,

$$y = X\beta + \epsilon \quad (11)$$

where  $X$  is a known  $n \times p$  matrix which may depend on the spatial coordinates of the

measurement locations;  $\beta$  is a  $p \times 1$  vector of structural parameters to be referred as "drift coefficients"; and  $\epsilon$  is zero-mean random vector with known covariance matrix  $Q_{yy}$ . Note that usually  $Q_{yy}$  is a known function of the covariance structural parameters  $\theta$  which must be estimated from data. However, in the next section we will concentrate on the case of known covariance functions. That is, the covariance is treated as part of the model whose validity is tested. This case can be treated analytically and the results are useful in more complicated cases.

Furthermore, it is assumed that  $y$  given  $\beta$  is normally distributed

$$p(y|M, \beta) = (2\pi)^{-n/2} |Q_{yy}|^{-1/2} \exp\{-1/2(y - X\beta)^T Q_{yy}^{-1}(y - X\beta)\} \quad (12)$$

where  $| \cdot |$  indicates matrix determinant. Also  $p(\beta|M)$  is assumed Gaussian with mean  $b$  and covariance matrix  $V$ . Note that while Equation (11) can be found at the foundation of linear geostatistics, the applicability of the Gaussian assumption is not universally accepted. However, Kitanidis [1986] has argued that linear geostatistics is most appropriate for cases which can be treated as Gaussian. Thus, we will consider the Gaussian assumption as part of the linear model.

The parameter estimation and prediction problems for the general linear model have been addressed in Kitanidis [1986], with numerical examples and comparisons to be reported in a forthcoming paper. The present work deals with model validation starting with an analytical derivation of the predictive distribution and proceeding to the development of some simple tests.

#### 4. Model Validation for Given Covariance Function Parameters

The predictive distribution  $p(y|M)$  can be calculated from (10) and (12):

$$p(y|M) = \int p(y|M, \beta) p(\beta|M) d\beta$$

$$= \int (2\pi)^{-n/2} |Q_{yy}|^{-1/2} \exp\left\{\frac{1}{2}(y - X\beta)^T Q_{yy}^{-1}(y - X\beta)\right\} \\ \cdot (2\pi)^{-r/2} |P|^{-1/2} \exp\left\{-\frac{1}{2}(\beta - b)^T P^{-1}(\beta - b)\right\} d\beta \quad (13)$$

In this equation, as well as throughout this work,  $Q_{yy}$  is assumed invertible. This usually can be achieved through preprocessing of the data and removal of redundant measurements, if they are present. We have also defined

$$r = \text{rank}(V) \quad (14a)$$

$$P = L^T(LVL^T)^{-1}L \quad (14b)$$

where  $L$  is any  $r \times p$  matrix of rank  $r$  so that  $LVL^T$  is invertible. Note that if  $r = p$ ,  $V$  is invertible. One can then set  $L$  equal to the  $r \times p$  identity matrix and  $P$  is the inverse of  $V$ . However, if  $r < p$ , one cannot write the pdf of  $\beta$  in the usual sense. One can only write the pdf of  $L\beta$  which is the vector of  $r$  combinations of the parameters whose covariance matrix,  $LVL^T$ , is invertible. This yields the result shown on Equation (13). The integration of (13) may be achieved analytically by taking advantage of the properties of the Gaussian distribution.

$$p(y|M) = (2\pi)^{-\frac{n+r}{2}} |Q_{yy}|^{-1/2} |P|^{-1/2} |P^{-1}|^{-1/2} \\ \exp\left\{\frac{1}{2}(y^T Q_{yy}^{-1} y - b^T P^{-1} b + b^T P^{-1} y)\right\} \quad (15)$$

where we have defined  $P^{-1}$ ,  $b^{-1}$ , and  $r^{-1}$  through the equations

$$P^{-1} = X^T Q_{yy}^{-1} X + P^{-1} \quad (16)$$

$$b^{-1} = X^T Q_{yy}^{-1} y + P^{-1} b \quad (17)$$

$$r^{-1} = \text{rank}(P^{-1}) \quad (18)$$

Equation (15) gives the predictive pdf of the data given the model and prior information about the drift coefficients. Note that this expression does not depend on fitted values of the drift coefficients. Consequently, it can be used to devise validation tests which avoid double-accounting of the data. (That is, first fitting the model to the data and then attempting to validate the model by comparing the fitted model to the same data.) For the common case of  $P^{-1}$  invertible ( $r^{-1} = p$ ), one can verify that

$$p(y|M) = \frac{\exp\{-\frac{n+r}{2} \cdot b^{-1}\}}{2} |Q_{yy}|^{-1/2} |P|^{-1/2} |P^{-1}|^{-1/2} \\ \exp\left\{\frac{1}{2}(y - Xb)^T (Q_{yy}^{-1} - Q_{yy}^{-1} X X^T Q_{yy}^{-1} X + P^{-1}) (y - Xb) + c\right\} \quad (19)$$

where  $c$  does not depend on data

$$c = b^{-1} X^T Q_{yy}^{-1} X (X^T Q_{yy}^{-1} X + P^{-1})^{-1} X^T Q_{yy}^{-1} X \cdot X^T Q_{yy}^{-1} X \\ - P^{-1} X^T Q_{yy}^{-1} X + P^{-1} P^{-1} + P^{-1} b \quad (20)$$

With the exception of Bayesian geostatistics [Kitanidis, 1986], the assumption of interest is that the drift coefficients are a priori completely unknown. In this special case,  $b^{-1} = 0$  and  $P^{-1} = \text{log} \delta^2 \rightarrow \infty$  and Equation (19) reduces to:

$$p(y|M) \propto \exp\left\{\frac{1}{2} y^T Q_{yy}^{-1} Q_{yy}^{-1} X (X^T Q_{yy}^{-1} X)^{-1} X^T Q_{yy}^{-1} y\right\} \quad (21)$$

Note that a rather peculiar property of this distribution is that

$$p(y + Xb|M) = p(y|M) \quad (22)$$

where  $b$  is any arbitrary  $p$ -dimensional vector. One may easily verify this relation by substituting  $y$  by  $y + Xb$  and noting that  $b$  drops out. That is, in the case of a priori "contrast" of the data. For example, consider the case that we measure logtransmissivity at various locations and the model is that logtransmissivity is a realization of an intrinsic random field with given variogram. In this case,  $p=1$ ,  $X^T = [1, \dots, 1]$  and  $b$  stands for the logtransmissivity mean. The expression of Equation (21) depends only on the

value of authorized increments of the data, i.e., linear function of the data  $\sum_{i=1}^n \lambda_i y_i$  which

do not depend on the value of the mean because  $\sum_{i=1}^n \lambda_i = 0$ . Since this predictive

distribution has no way of distinguishing between  $y$  and  $y \pm X\theta$ , it cannot be used to calculate the probability of: say,  $y_i > 0$ . By the same token,  $p(Y|M)$  cannot be used to check the consistency of a single actual observation  $y_i$  with the model. This is a direct consequence of assuming complete ignorance about the mean of  $y$ . However, since the covariance matrix is assumed known, one can calculate the probability of linear combination of data which do not depend on the values of the unknown drift coefficients. For example, one can check whether  $y_1 - y_2$  seems to be consistent with the model.

The predictive distribution can be used to develop a host of useful validation tests. We will focus on only one case which is similar to the  $S_p$  test (Equation 3) which is often used in applied geostatistics.

**A General Test**

A useful test of the overall goodness of fit of the assumed model can be developed as follows. It can be proved that

$$g(Y) = y^T Q_y^{-1} y - b^T P b^{-1} + b^T P b \tag{23}$$

is always nonnegative. A plot of  $p(Y|M)$  vs  $g(Y)$  is shown in Figure 1. Now, let  $Y$  be the vector of actual measurements which, according to the model, is a sample of random vector  $y$ . If the model is correct, it is unlikely that  $g(Y)$  would be at the tails of the distribution. With this in mind, calculate the probability.

$$\sigma_T = \text{Pr}\{g(Y) > g(Y)\} \tag{24}$$

It may be useful to remind that  $y$  is a vector of random variables, while  $Y$  is the vector of the numerical values of the actual data.  $g(Y)$  is a quadratic function of Gaussian random variables. Consequently, it follows a chi-squared distribution.

In the special case of most interest in geostatistics, for a priori perfectly unknown drift parameters,  $g(Y)$  follows the well-known (central) chi-squared distribution with  $n-p$  degrees of freedom. Thus, performing the test is facilitated because

$$\sigma_T = \text{Pr}\{\chi^2(n-p) > g(Y)\} \tag{25}$$

can be calculated from widely available tables.

If  $g(Y)$  is very large so that  $\sigma_T$  is small, such as  $\sigma_T = 0.01$ , then "the measurements are at the tail of the predictive distribution" and this can be interpreted as evidence against the assumption that  $Y$  was generated from  $p(Y|M)$ . Similarly,  $g(Y)$  near zero is unlikely given the model, so that if  $\alpha$  is near 1 one may conclude that the model is not consistent with the data. This point can be illustrated with the following example. Assume that we analyze logtransmissivities of a given aquifer and we enter the model that they are a realization of an intrinsic field with linear variogram. In this case,  $p = 1$  and  $X$  is an  $n \times 1$  vector of 1's and  $(Q_y^{-1})_{ij} = \theta d_{ij}$  where  $\theta$  is minus the slope of the variogram and  $d_{ij}$  is the distance between measurements. Assume, however, that when the measurements are obtained they are the same everywhere so that  $g(Y) = 0$ . It is obvious that the data, which indicate lack of spatial variability, discredit the model which presumes that measurements vary from point to point. More generally, whenever the degree of spatial variability is overestimated by the model,  $g(Y)$  is near zero.

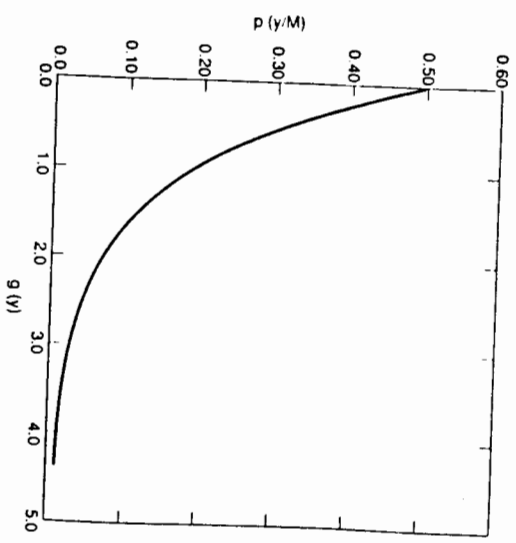


Figure 1. Predictive probability function.

## 5. Example

Consider the 29 measurements of Kastriss and Bras (1981, page 92) also analyzed in Kitanidis (1983). We consider the following models:

Model 1: Intrinsic random field with linear variogram with nugget. In this case,  $p = 1$ ,  $X$  is a column vector of 1's, and  $Q_{Yj}$  is given by

$$Q_{Yj} = \theta_1 \cdot 1 + \theta_2 D \quad (26)$$

where 1 is the identity matrix and  $D$  is the symmetric matrix with  $D_{ij}$  the distance between the locations of measurements  $i$  and  $j$ . In this case,  $\theta_1$  is the nugget and  $\theta_2$  is the negative slope of the variogram. It is given that

$$\theta_1 = 0$$

$$\theta_2 = -0.0015$$

We estimate  $g(Y) = 46.1$

$$\sigma_T = \text{Prob}\{x^2(28) > 46.1\} = 0.017$$

The small value of  $\sigma_T$  indicates that the data are an unlikely sample given the model. The test thus shows that the data are not very consistent with this model.

Model 2: Same as model one except for the numerical values of  $\theta_1$  and  $\theta_2$ . In this model

$$\theta_1 = 0.00427$$

$$\theta_2 = -0.00116$$

We calculate  $g(Y) = 28$

$$\sigma_T = \text{Prob}\{x^2(28) > 28\} = 0.46$$

Model 3: Intrinsic model of order 1 with linear generalized covariance function. In this case,  $p = 3$ ; the first column of  $X$  consists of ones, the second of the first spatial coordinates of the data, and the second of the second spatial coordinates of the data. The covariance matrix has the same form as before and we are given

$$\theta_1 = 0.00286$$

$$\theta_2 = -0.00174$$

The data yield  $g(Y) = 24.7$  with tail probability

$$\sigma_T = \text{Prob}\{x^2(26) > 24.7\} = 0.46$$

Thus, models 2 and 3 pass the overall validation test.

## 6. Discussion

Geostatistics is no exception in that one should always start with the simplest model which is consistent with past experience and the objectives of modelling. Model selection

is preceded by an exploratory analysis in which data are organized and presented in ways which bring out important features of variability and interdependence. Last but perhaps most important, it is always guided by common sense and the understanding of the physical processes which are involved.

Nevertheless, before using the model to make predictions, the model's consistency with the data must be double-checked. This is where "model validation" comes in. Unfortunately, there is no single test which can be used to prove or disprove a model (assuming that the data are trusted) or to identify "peculiar" measurements. Instead, one proceeds cautiously by developing tests which compare the behavior of the actual data with the behavior expected assuming that the model is correct. Each one may check against one possible model inadequacy, such as whether the mean square error calculated using the model is a good measure of the actual square error of estimation. Of particular interest are tests which check the model's performance under conditions as similar as possible to those under which the model will be asked to perform.

There is no complete set of tests which check against all eventualities. However, the concept of predictive probability presents a unified way to develop such tests. The idea is to calculate the probability density function of the data given only the model and prior information about the parameters. Next, to evaluate the pdf of a statistic of interest (function of the data). Then, one may judge whether the actually measured statistic is likely to have been generated from this pdf. One may develop formal statistical tests on the basis of which the model is rejected if a statistic falls outside of a given range of values. Instead, most engineers may prefer a less rigid approach in which the results of each test may strengthen or weaken our confidence in the model. At the end of the exercise, the modeller should have a better understanding of the strengths and weaknesses of his model, which in itself is not a minor achievement. To paraphrase Box: "Put your models to the test - then use judgement." After all, no hydrogeological model is a perfect representation of reality.

In this paper, we discussed the concept of predictive probability which was derived for the model used in linear geostatistics (kriging, universal kriging, cokriging, etc.) when the covariance is known. Then, we used the predictive probability to derive the pdf of a statistic which can be used to check against overestimation or underestimation of the mean square error. Unlike the  $S_1$  statistic (Equation 3), which is widely used for the same purpose, its distribution is known a priori. Consequently, one can judge whether the sampled value is too small or too large for the model to be consistent with data.

## ACKNOWLEDGMENTS

This work has been supported by the National Science Foundation under Grant No. ECE-8517598 and the Minnesota Water Resources Center under Grant USD1 14-08-0001-G-1025-02. The comments of Dr. David Freyberg are appreciated.

## REFERENCES

- Belsey, D. A., Kuh, E. & Welch, R. E. (1980) *Regression Diagnostics*. New York: Wiley-Interscience, 292 pp.
- Box, G. E. P. & Jenkins, G. M. (1976) *Time Series Analysis*. Holden-Day.
- Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *J. Royal Statist. Society, Series A* 143(4): 383-430.
- Delinger, P. (1976) Linear estimation of nonstationary spatial phenomena. *Advanced Geostatistics in the Mining Industry*. Edited by M. Guarascio, M. David, and C. Huijbregts. Hingham, MA: D. Reidel, 49-68.
- Hoeksema, R. J. & Kitanidis, P. K. (1984) An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modelling. *Water Resour. Res.* 20(7): 1003-1020.
- Hoeksema, R. J. & Kitanidis, P. K. (1985) Analysis of the spatial structure of properties of selected aquifers. *Water Resour. Res.* 21(4): 563-572.
- Karlsruhe, J. & Bras, R. L. (1981) The practice of kriging. R. M. Parsons Lab. TR No. 263, MIT.
- Kitanidis, P. K. (1983) Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resour. Res.* 19(4): 909-921.
- Kitanidis, P. K. (1985) Minimum-variance unbiased quadratic estimation of covariances of regionalized variables. *J. Int. Assoc. of Math. Geology*, 17(2): 195-208.
- Kitanidis, P. K. (1986) Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resour. Res.* 22(4): 499-507.
- Kitanidis, P. K. (1988) *Geostatistics in Groundwater Modelling*. Notes to be published.
- Kitanidis, P. K. & Vomvoris, E. G. (1983) A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations. *Water Resour. Res.* 19(3): 677-690.
- Klemes, V. (1986) Operational testing of hydrological simulation models. *Hydrological Sciences J.* 31(1): 13-24.