11

Geostatistics Applied to Hydrology Problems and Tools

D. E. Myers

University of Arizona, USA

11.1 INTRODUCTION

To understand the importance of geostatistics for the assessment, monitoring and remediation of groundwater pollution it is necessary to consider the importance of geostatistics in hydrology and more generally the transition from a purely deterministic approach for such problems to a stochastic one. The simplest form of Darcy's equation for subsurface flow, valid in an isotropic medium with no source or sink is given by

$$\partial/\partial x_i(K\partial\phi/\partial x_i) = 0;\ i = 1,\ldots,k \tag{11.1}$$

where K is hydraulic conductivity, ϕ is hydraulic head and k is the dimension of the flow domain. In its original form, K is assumed to be a constant. With current computing capability it is also possible to allow K to be a smooth function. A stochastic version of the system is obtained by letting $Y = \ln K$ be a Gaussian process. Even with this assumption the parameters of the process must be estimated and K is usually only known at a finite number of locations. As the emphasis shifted from resource development to groundwater pollution problems (Smith, 1981; Freeze *et al.*, 1989; Neuman, 1993; Kitanidis, 1995), the spatial variability of hydrogeological parameters became more important. Heterogeneity in the subsurface induces fluctuations in flow velocities and in turn dispersion of solutes is enhanced. Whether deterministic or stochastic

Statistics for the Environment 3: Pollution Assessment and Control.
Edited by V. Barnett and K. Feridun Turkman.

models are used, geostatistics is useful for characterizing and modelling the spatial variability of the relevant parameters such as hydraulic conductivity; porosity, permeability and dispersivity.

In trying to apply deterministic models there were at least two perceived obstacles:

- an inadequate understanding of the processes
- inefficient methods for solving the equations

The differences encountered in different environments; porous media vs. fractured rock, the unsaturated zone vs. the saturated zone, gave impetus to the growth of an interest in the use of stochastic models in the 1970s. Previously the use of statistics was limited to empirical statistical models and inappropriate where the physical and chemical processes of subsurface hydrology were well defined. However there are at least three practical problems associated with the use of stochastic models (Kitandis, 1995):

- estimation and inference of parameters, such as hydraulic conductivity from head observations
- determination of the effect of uncertainty in parameters vs. the uncertainty in the head or flow
- compensating for a change in scale.

These problems are all exacerbated by heterogeneities in the subsurface.

11.1.1 Heterogeneity

One of the serious problems that occurs in the use of deterministic models is the incorporation of the heterogeneity that occurs in the subsurface, both in the geology and in the hydrological parameters, particularly hydraulic conductivity. This induces perturbations in the velocities of the groundwater and hence large-scale dispersion of contaminants, this dispersion is not predictable using deterministic models. This heterogeneity exits at multiple scales; in the laboratory, in the field and finally regionally. One solution is to use 'effective' values, whcih are essentially spatial averages. Most often these must be empirically determined for each application which are then used for the parameters in the flow and advection equations. It will usually be necessary to use zonal models and these will be scale dependent. Zonal models create problems in the face of non-stationarities. Anderson (1995) and others have proposed the use of site-specific hydrogeological facies models. A hydrogeological facies is a 'homogeneous but anisotropic unit that is hydrogeologically meaningful for the purposes of field experiments and modelling'. Some work has been done on the use of variograms to characterize facies, Anderson (1995) has suggested the possibility of each facies type having a distinctive variogram. However it is not clear that the family of variograms is sufficiently rich for this purpose.

11.1.2 Stochastic modelling

There are two general methods for incorporating geostatistics into the use of stochastic models; one involves the computation and use of 'effective' parameter values and the other uses simulation. Darcy's equation (isotropic medium, steady state with no sources or sinks), relates hydraulic head (the dependent variable) to hydraulic conductivity (the independent variable). In a zonal model, hydraulic conductivity is treated as a constant (perhaps an 'effective' value) for each of several zones and the equation is solved separately for each. In a continuum model, hydraulic conductivity is allowed to vary and is often modelled as a log-Gaussian process but the parameters of this statistical model are generally unknown and must be modelled from data, moreover the 'scale' of the data is a crucial aspect of the problem. By the use of Monte Carlo methods (conditional or unconditional simulation, to be discussed later), multiple realizations of the hydraulic conductivity field are generated and Darcy's equation is solved for each one separately. The statistical parameters of hydraulic conductivity have a crucial impact on the variability of hydraulic head.

Even if hydrological parameters are assumed constant, or at least constant over a zone, the values must generally be estimated from data and most often from sparse data. This means there will be some uncertainty associated with the estimated values; this uncertainty is generally scale dependent. If the parameters are not constant but rather are modelled by a stochastic process then there will be uncertainty associated with the choice of the model as well as with the values of the parameters in the model. Particularly in this case, the data is 'spatial' (often three-dimensional) and usually exhibits some degree of spatial correlation and spatial variability. This is one of the reasons for the importance of geostatistics in hydrology and groundwater pollution problems.

Hydraulic head and hydraulic conductivity are not the only variables of interest; concentrations of contaminants, rates of spread of a plume, direction of spread of a plume, boundaries of a plume, as well as other hydrological parameters such as porosity, permeability, transmissivity and specific capacity are all relevant. Some are related by state equations and others are not.

11.2 THE TOOLS

11.2.1 Univariate

Before considering the 'tools' it is necessary to consider some of the characteristics of the data. First of all, each data value is associated with a point or a volume in space. In geostatistical terminology the latter is the 'support' of the data (essentially the same as the support of a measure). In some cases a value associated with a volume represents an average over the volume but in others it is only related to the measurement process. Let $\boldsymbol{x}$ denote a point in k-dimensional

space, $v(\boldsymbol{x})$ a 'volume' or area centered at $\boldsymbol{x}$ and $Z(\boldsymbol{x})$ a random function representing the variable of interest such as hydraulic conductivity; then

$$Z_{v(\boldsymbol{x})} = 1/v \int_v Z(\boldsymbol{x})\,d\boldsymbol{x} \tag{11.2}$$

is a linear functional with support v. A change in the 'Support' will most likely result in a change in the statistical characteristics and most of the state equations do not explicitly incorporate the support. Some parameters can not be directly measured but instead must be inferred from other measurements, there may be different techniques for use in the laboratory than in the field. It is usually convenient to decompose $Z(\boldsymbol{x})$ as follows:

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + Y(\boldsymbol{x}) + \varepsilon(\boldsymbol{x}) \tag{11.3}$$

where $m(\boldsymbol{x})$ is the mean function (usually modelled as a linear combination of known linearly independent functions such as monomials in the position coordinates), $Y(\boldsymbol{x})$ is a zero-mean random component satisfying some form of stationarity and characterized by a spatial structure function such as the variogram or (auto)covariance function. Finally $\varepsilon(x)$ is the noise or error term and is considered to be spatially uncorrelated and uncorrelated with $Y(\boldsymbol{x})$. $m(\boldsymbol{x})$ can be interpreted as the regional variation and $Y(\boldsymbol{x})$ the local variation. $Y(\boldsymbol{x})$ is intrinsically stationary if

$$E[\,Y(\boldsymbol{x}+\boldsymbol{h}) - Y(\boldsymbol{x})] = 0 \text{ for all } \boldsymbol{x}, \boldsymbol{h} \tag{11.4}$$

$$\gamma(\boldsymbol{h}) = 0.5\ \mathrm{Var}[Y(\boldsymbol{x}+\boldsymbol{h}) - Y(\boldsymbol{x})] \tag{11.5}$$

exists for all $\boldsymbol{x}$, h and depends only $\boldsymbol{h}$.

In the case where $Y(\boldsymbol{x})$ is second-order stationary, its covariance $C(\boldsymbol{h})$ is related to the variogram $\gamma\ (\boldsymbol{h})$ by

$$\gamma\,(\boldsymbol{h}) = C(0) - C(\boldsymbol{h}) \tag{11.6}$$

The modelling of the variogram or covariance is the most difficult problem in geostatistics. Whereas covariance functions must satisfy a positive-definiteness condition, variograms must satisfy a negative conditional definiteness condition both to ensure that the estimation variance is non-negative and also to ensure that the system of equations has a unique solution, see Myers (1988b).

The estimation or prediction of a value at a non-data location using data at irregularly spaced locations is the simplest problem. Journel (1986), Cressie (1986, 1988), Myers (1991c, 1994a, b) provide overviews on the analysis of spatial data. Neuman (1984) presents an application to $\log_{10} T\,(\boldsymbol{x})$ where $T\,(\boldsymbol{x})$ is

transmissivity; the data are from the Avra Valley aquifer near Tucson, Arizona. If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ denote data locations and $\boldsymbol{x}_0$ a non-data location then in the case where $m\,(\boldsymbol{x})$ is constant and $\varepsilon\,(\boldsymbol{x}) \equiv 0$, $Z(\boldsymbol{x}_0)$ is estimated by

$$Z^*(\boldsymbol{x}_0) = \sum_{i=1,\ldots,n} \lambda_i Z(\boldsymbol{x}_i) \tag{11.7}$$

where the λ_is are obtained as the solution of the system

$$\sum_{i=1,\ldots,n} \lambda_i \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) + \mu = \gamma(\boldsymbol{x}_0 - \boldsymbol{x}_j); \quad j = 1, \ldots, n \tag{11.8}$$

$$\sum_{i=1,\ldots,n} \lambda_i = 1.$$

These are known as the 'ordinary kriging' equations and $Z^*(\boldsymbol{x}_0)$ will be a BLUE where the estimation variance is computed using the variogram, the minimized value is called the kriging variance. Both the estimator and the system of equations are easily modified to allow a non-constant $m(\boldsymbol{x})$, non-zero $\varepsilon(\boldsymbol{x})$ and the use of non-point data as well as for the estimation of non-point values. In the latter case the right-hand side terms in equation (11.7) are replaced by

$$\gamma(V, \boldsymbol{x}_j) = (1/V) \int_v \gamma(\boldsymbol{x} - \boldsymbol{x}_j)\,d\boldsymbol{x}. \tag{11.9}$$

In this latter case the system in (11.8) is known as the 'block kriging' equations. Two kinds of non-linear transformations are commonly used in geostatistics; logarithmic and indicator. The logarithmic is particularly useful in the case of a multivariate log-Gaussian distribution. The indicator transform is given by

$$I(\boldsymbol{x}; \boldsymbol{z}) = 1 \quad \text{if} \quad Z(\boldsymbol{x}) \leqslant \boldsymbol{z} \quad \text{and} \quad 0 \text{ otherwise} \tag{11.10}$$

Usually several different cut-off levels are used. The indicator transform has several advantages, the transformed data is less sensitive to outliers and in addition to point or areal average estimation one can estimate the proportion of a region where the variable is above (or below) a specified cut-off. This is especially useful for determining the need for remediation. Let V be a volume and

$$\Phi(V;\boldsymbol{z}) = (1/V) \int_v I(\boldsymbol{x}; \boldsymbol{z})\,d\boldsymbol{x} \tag{11.11}$$

Then $\Phi(V;\ \boldsymbol{z})$ is the proportion of V where $Z(\boldsymbol{x}) \leqslant \boldsymbol{z}$. Since $E\{I(\boldsymbol{x};\boldsymbol{z})\} =$

$P(Z(\boldsymbol{x}) \leqslant \boldsymbol{x}\}$ under a stationary assumption $E\{I(\boldsymbol{x}; \boldsymbol{z})\}$ can be thought of a 'global' distribution whereas the estimate of $I(\boldsymbol{x}_0; z)$ conditional on data $I(\boldsymbol{x}_i; \boldsymbol{z})$, $i = 1, \ldots, n$ is a local conditional distribution. Both $\Phi(V; \boldsymbol{z})$ and $I(\boldsymbol{x}_0; \boldsymbol{z})$ can be estimated in the same way as $Z(\boldsymbol{x}_0)$ except that the variograms must be modelled on transformed data. An advantage of the indicator transform over the logarithmic transform is that in general it is not necessary to re-transform. Various authors have used indicator variograms to model facies. Journel (1983, 1984) and his students are responsible for much of the work on the application of the indicator transform in geostatistics.

11.2.2 Multivariate

As noted above there are multiple hydrogeological parameters, all of which are spatially variable. These are usually intervariably correlated as well. In some instances data on a correlated variable is used to enhance or replace data on another variable, in other cases the objective is to estimate or predict several variables using data on all variables. In both instances data may not be available on all variables at all locations (sometimes referred to as the 'undersampled' problem). Neuman (1984) gives an example using log specific capacity and log hydraulic conductivity. The general form of the co-kriging estimator is given in Myers (1982, 1984, 1988a, 1992a, b). Let the various variables of interest be $Z_1(\boldsymbol{x}), \ldots, Z_m(\boldsymbol{x})$ and $\boldsymbol{Z}(\boldsymbol{x}) = [Z_1(\boldsymbol{x}), \ldots, Z_m(\boldsymbol{x})]$ then the analogue of 11.7 is

$$\boldsymbol{Z}^*(\boldsymbol{x}_0) = \sum_{i=1,\ldots,n} \boldsymbol{Z}(\boldsymbol{x}_i)\,\Gamma_i \tag{11.12}$$

where the Γ_is are $m \times m$ matrices obtained as the solution of the system

$$\sum_{i=1,\ldots,n} \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j)\,\Gamma_i + \mu = \gamma(\boldsymbol{x}_0 - \boldsymbol{x}_j); \quad j = 1, \ldots, n \tag{11.13}$$

$$\sum_{i=1,\ldots,n} \Gamma_i = I$$

In this case the spatial structure function is matrix valued and must satisfy a matrix form of negative conditional definiteness. The diagonal entries will be variograms or covariances, the off-diagonal entries will be one of two different cross-variograms (one symmetric and the other not) or cross-covariances. Ben-Jemaa *et al.* (1994) use co-kriging for the design of a groundwater monitoring network. Istok *et al.* (1993) use co-kriging for direct estimation of contaminate concentrations as does Myers (1989).

11.3 SIMULATION

Whatever the algorithm, interpolation smoothes the data and the interpolated surface shows less spatial variability than the data. Moreover interpolation provides only one scenario for a given data set but it is often useful to generate alternative scenarios, i.e, the data is viewed as a non-random sample from one realization of the random function $\boldsymbol{Z}(\boldsymbol{x})$ and it can be useful to generate additional realizations of $\boldsymbol{Z}(\boldsymbol{x})$. When a random variable is simulated, the distribution is preserved and hence any parameters. In the case of a random function one will generally not know the full distribution function and only the parameters or univariate characteristics of $\boldsymbol{Z}(\boldsymbol{x})$ can be preserved. Usually this means the spatial structure function (variogram or covariance), mean and univariate distribution. Simulation is useful for many kinds of application, if log hydraulic conductivity is simulated then Darcy's equation can be solved for each realization and the solutions averaged. Simulation also provides a tool for comparing dispersivities measured in the field with those calculated from the flow and dispersion equations.

11.3.1 Turning Bands Algorithm

Although not used quite as much as before it is useful to first consider the Turning Bands Algorithm which was first introduced by Journel (1974). Let $W_1(\boldsymbol{x})$ be a random function defined in 1-space with isotropic covariance function $C_1(r)$ and $\mathrm{d}P(s)$ a uniform probability measure on the unit sphere in k-dimensional space with inner product $\langle \boldsymbol{x}, s \rangle$. Each point s on the sphere can be identified with a 'direction', i.e., a unit vector from the centre of the sphere. Then

$$W(\boldsymbol{x}) = \int W_1(\boldsymbol{x})\,\mathrm{d}P(s) \tag{11.14}$$

The covariance of $W(\boldsymbol{x})$ and that of $W_1(\boldsymbol{x})$ are related by an integral equation which is particularly easy to solve for $k = 3$, i.e.,

$$\mathrm{d}/\mathrm{d}r[rC(r)] = C_1(r) \tag{11.15}$$

In practice one begins with $C(r)$, the covariance of $W(\boldsymbol{x})$, finds $C_1(r)$ and uses this to generate multiple copies of realizations of $W_1(\boldsymbol{x})$. The simulated realization of $W(\boldsymbol{x})$ is obtained as a discrete approximation to (11.14);

$$W_s(\boldsymbol{x}) = a \sum_{j=1,\ldots,q} W_1(\boldsymbol{x}) \tag{11.16}$$

The constant a is to normalize the variance and usually taken to be $(1/q)^{0.5}$. The

directions $s_1, \ldots, s_q$ should be 'equally' spaced; in the case of $k = 3$ geometry intervenes and there are essentially only 15 possible equally spaced directions. The $W_{1,s_i}(\boldsymbol{x})$s are usually generated using a Moving Average (box–Jenkins) algorithm. For the case of $k = 2$ there are two significant differences; one is that there is no maximum number of equally spaced directions and two, the integral equation is not as easily solved. This led to the use of spectral covariances instead, particularly by Gelhar's group at MIT. The method is described in Mantoglou and Wilson (1982). Of course even in one dimension one cannot really simulate an entire realization, i.e., at every point on the line. Instead one simulates at a finite number of points which leads to the idea of bands 'turning in space'. Because the simulated values are obtained as a linear combination of uncorrelated random numbers, the asymptotic univariate distribution will be Gaussian and hence it is best to transform to Gaussian before simulation. The realization can be 'conditioned' to the data by simulating a zero-mean realization and adding to the kriged surface.

11.3.2 L-U or Cholesky decomposition

Since in practice one does not simulate an entire realization but rather only its values at a finite number of locations (usually on a regular grid), simulation of the random function reduces to the simulation of a finite number of correlated random variables and for that one can use the L-U or Cholesky decomposition of the covariance function. Let $y_1, \ldots, y_T$ be the points on the grid, $\boldsymbol{C}$ the covariance matrix of the $Z(y_1), \ldots, Z(y_T)$ and $\boldsymbol{C}^U$ the upper triangular factor. For $w_1, \ldots, w_T$ uncorrelated, mean-zero Gaussian random numbers, $\boldsymbol{C}^U\boldsymbol{W}^T$ is a vector of correlated simulated values whose covariance matrix is $\boldsymbol{C}$, $\boldsymbol{W}$ being the column vector of the w's. Conditioning by simple kriging can be incorporated into the simulation stage by augmenting the covariance matrix for the data locations and between the data locations and the grid points. Again it is best to transform the data to a Gaussian distribution prior to simulation. Both the Turning Bands and the L-U Decomposition algorithms have the disadvantage of severe memory limitations, i.e., a limit on the size of the grid. For L-U decomposition, the practical limit is on the order of a 30×30 grid unless perhaps parallel processing is used.

11.3.3 Sequential Gaussian simulation

One way to avoid the limitations of the two prior algorithms is to exploit the connection between 'simple' kriging and the conditional distribution for multiple jointly Gaussian distributed random variables. Suppose that $Z(\boldsymbol{x})$ is second-order stationary with constant mean m and covariance $C(\boldsymbol{h})$, then (11.7) can be written in a slightly different form with a simpler system of

equations

$$Z^*(\boldsymbol{x}_0) = m + \sum_{i=1,\ldots,n} \lambda_i (Z(\boldsymbol{x}_i) - m) \tag{11.17}$$

where the λ_is are obtained as the solution of the system

$$\sum_{i=1,\ldots,n} \lambda_i C(\boldsymbol{x}_i - \boldsymbol{x}_j) = C(\boldsymbol{x}_0 - \boldsymbol{x}_j); \quad j = 1, \ldots, n \tag{11.18}$$

It is seen that (11.17) is the conditional expectation under the second-order stationary, joint Gaussian assumption. Moreover the minimized estimation is just the conditional variance. Given the data locations and a grid point, use equations (11.17) and (11.18) to obtain an estimated value and a conditional variance. Then this conditional Gaussian distribution is sampled to obtain a simulated value at that grid point. This simulated value is added to the 'data' set and the process repeated for a different grid point. The grid points are visited sequentially, each time a simulated value is obtained by sampling from a univariate conditional Gaussian distribution. Because of the sequential nature of the simulation, the memory limitations are much reduced.

11.3.4 Simulated annealing

Simulated annealing is well known as an optimization method but when combined with the proper cost function and appropriate domain for the cost function it can be used to simulate a random function. For a given set of grid points and a given univariate distribution, generate a simulated value at each grid point by sampling the distribution as many times as there are grid points. In general this will not result in a realization with the desired spatial correlation but it can be 'improved'. The cost function will be a mean square difference between the desired model variogram $\gamma(\boldsymbol{h})$ and the sample variogram $\gamma^*(\boldsymbol{h})$ at any given stage in the 'annealing'. Select two grid points, compute the objective function

$$\sum_{j=1,\ldots,r} [\gamma^*(h_j) - \gamma(h_j)]^2 \tag{11.19}$$

where $h_1, \ldots, h_r$ are user chosen lag distances. Next select two grid points at random, interchange the assigned simulated values and recompute the cost function. If less then retain the interchange and proceed to a new pair of grid points. If not, then the temperature function is used to determine whether the interchange is retained. That is, 'simulated annealing' is used to minimize the cost function given in equation (11.19). The practical aspects of using this algorithm are discussed in Deutsch and Cockerham (1994); Datta-Gupta *et al.* (1995) apply the algorithm to a large porosity data set.

Both the L-U Decomposition and Sequential Gaussian algorithms are easily extended to the joint simulation of correlated random functions. For the extension of the L-U Decomposition algorithm see Myers (1989). For an application of multivariate simulation to flow modeling see Gutjahr *et al.* (1994).

11.4 PARAMETERS

Hydraulic conductivity is one of the principal, perhaps the most important, hydrologic parameter for the flow and transport equations. It can be directly measured, i.e., inferred, in the field by borehole flow meters or by using slug tests. Several problems occur; threshold values for flow meters limit the range of possible values of hydraulic conductivity. The support of the data is not easily determined and change of support problems are non-trivial as noted later. Direct measurements can also be made by testing samples in the laboratory which again leads to change of support problems. Hydraulic conductivity can also be measured or inferred indirectly from grain size distributions or by facies mapping as described in Rehfeldt *et al.* (1992). Smith (1981) uses a proxy variable, porosity, to infer hydraulic conductivity.

The common assumption of a log-Gaussian distribution and stationarity are discussed in relationship to the Borden aquifer study by Woodbury and Sudicky (1992). Even though this parameter is not spatially constant, one may consider 'effective' values for a given volume or region. In general this value is not a simple spatial average. Desbarats (1992a, b, 1994) has derived an appropriate weighting under different conditions and related these to arithmetic and geometric means. In particular a radial weighted average is appropriate under some conditions:

$$K_v^{\varpi} = (1/W)\int K^{\varpi}(\boldsymbol{x})/r^2(\boldsymbol{x})\,\mathrm{d}V \qquad (11.20)$$

$$W = \int 1/r^2(\boldsymbol{x})\,\mathrm{d}V.$$

Anderson (1995) and Gomez-Hernandez and Gorelick (1989) have pointed out the non-uniqueness of effective values.

For solute transport, dispersivity is a crucial parameter but there are problems in reconciling direct measurements with those computed from other parameters using state equations. In particular there are differences in the computed values of the Transverse and longitudinal values when using simulated hydraulic conductivity fields. These can be compared with direct measurements inferred from tracer tests, see for example, Rehfeldt *et al.* (1992).

Using state equations for flow and dispersion one has a choice between using inverse modelling for parameter estimation and using interpolation coupled with a lesser number of direct measurements. Doctor and Nelson (1981) examined this problem in the context of repository characterization. Hoeksema and Kitanidis (1984) use geostatistical methods to estimate hydraulic conductivity and transmissivity. Aboufrassi and Marino (1984) use kriging and co-kriging with transmissivity values for input to flow models.

11.5 CONTAMINANT CONCENTRATIONS

In addition to the modelling of subsurface flow and transport, geostatistics can be used directly with contaminant concentration data or can be used to combine hydrological information with concentration data. Different contaminants may disperse at different rates, influenced by subsurface heterogeneities. The Borden study provided an opportunity to characterize a site both in terms of hydrology and in terms of contaminant concentrations. Woodbury and Sudicky (1991) modelled variograms for the hydrological and transport parameters, Myers (1989b) used multivariate variograms and co-kriging to characterize the spatial distribution of multiple contaminants at the Borden site.

Istok *et al.* (1993) provide a case study in characterizing a polluted soil site, the deposition resulting from the use of pesticides and pesticide by-products. Multivariate geostatistical techniques can be useful both because of the expense of collecting samples (more information is collected at each sample location) but also because of the expense of analysing the samples which is influenced more by the number of samples than the number of variates.

Remediation of contaminated groundwater will usually require pumping of the polluted water, treatment and subsequent replacement. The pumping will cause the plume to move and it is important to predict and control this movement to avoid undesirable dispersion. Gorelick (1995) used conditional simulation to optimize both well placement and pumping regimes for the removal of a contaminant plume. This requires *a priori* modelling of the variogram or covariance(s).

There can be interest in the concentrations and dispersions of solutes in groundwater for reasons other than for environmental monitoring or remediation. Detection and characterization of geochemical patterns either in the soil or in the groundwater are useful for detecting and identifying ore deposits. Kane *et al.* (1982), Myers *et al.* (1982) and Myers (1988c) demonstrate the use of variograms, kriging as well as inverse distance weighting, to analyse patterns in the groundwater that could be indicative of a uranium deposit. This data was collected as a part of the NURE (National Uranium Resource Evaluation) project conducted by the US DOE in the mid-1970s and early 1980s.

Perhaps the simplest form of aquifer monitoring only involves monitoring hydraulic head which in turn can be obtained from Darcy,s equation if the hydraulic conductivity is known. Both conditional simulation and interpolation methods such as kriging allow interpolation of hydraulic conductivity from

a small number of locations to a dense grid which can then be used as the input for a flow model. This is illustrated in Aboufirassi and Marino (1983) who used universal kriging, i.e., the mean function $m(\mathbf{x})$ in equation (11.3) is not constant, and compared those results with results from the use of Trend Surface Analysis. As shown by Marcotte and David (1988), Trend Surface Analysis is the same as universal kriging but using a pure nugget (no spatial correlation) except at the data locations. Ben-Jemaa *et al.* (1994) used co-kriging to allow the use of auxiliary variables as proxies for the principal variable of interest.

11.6 SAMPLING DESIGN

There are two aspects of sample design that occur in the application of geostatistics; first there is the problem of collecting sufficient data to estimate and model a spatial structure function as well as to detect and model any non-stationarities, secondly, after the spatial structure function and mean function have been modelled they will be used either in kriging or in simulation (or both). Note that the decomposition given in (11.3) is not uniquely determined by a finite amount of data as shown in Myers (1989c). Both for the purpose of modelling variograms for hydrological parameters and for contaminant concentrations it will usually be necessary to drill wells, which is expensive; a sample design will depend not only on the number of wells but also their locations. Warrick and Myers (1987) obtained an algorithm for optimizing the frequency distributions of the squared differences that appear in the sample variogram in terms of the sample location pattern. The sampling design for variogram estimation must also be adequate for the detection and characterization of any anisotropies in the variogram, i.e., directional dependence.

Once the spatial structure function is modelled, it can be used to determine a sampling plan for data collection for interpolation; in general a plan that is optimal for estimation and modelling of variograms is not optimal for kriging even though in many cases the same data set is used for both. There are at least two techniques for optimizing a sample plan, one involves using the kriging variance(s) to construct an objective function which is then minimized. This function will usually not be convex and there may be problems with local minimums that are not global. This method depends on the fact that the kriging variance does not depend on the data as such but only on the variogram/covariance model and the sample location pattern.

As an alternative, the variogram or covariance can be used to generate many realizations (conditioned to the data). For each realization, one then searches for an sampling plan optimized with respect to Type I and Type II errors (for example in the case of environmental characterization). Easley *et al.* (1991) used frequency domain simulation of hydraulic conductivity to generate multiple scenarios. Statistically optimal plans may not always be feasible because of various practical constraints on sample locations.

11.7 MODEL-BASED VS. DESIGN-BASED ANALYSES

Geostatistics is a model-based set of techniques and methods yet the model is driven by and is responsive to the data. Obviously the results are dependent on the validity of the model(s) used, i.e, on the underlying statistical assumptions. Although the sample variogram is known to be relatively non-robust with respect to skewed distributions, the kriging estimators are relatively robust with respect to changes in the variogram (type and/or parameters). The kriging estimators are not so robust with respect to skewed distributions, hence one advantage of the indicator transform. Random sampling, i.e., random sample location selection, is of very little importance in geostatistics because the model assumes that there are no replications in the data. Moreover it is assumed that the regional and local variability is on a much more significant scale than are any measurement or instrument errors, hence removal of the error term is not a primary objective as is common in ordinary regression applications.

Although there seem to be few if any applications of the Horwitz–Thompson estimator in hydrology and groundwater pollution problems, there have been attempts to relate this estimator to geostatistical methods. de Gruijter and ter Braak (1990) compare results using the two different approaches. Cordy and Thompson (1995) show how to use the sample variogram to specify the reliability of the Horwitz–Thompson estimator.

11.8 LINGERING STATISTICAL PROBLEMS

Even the best statistical tools are dependent on both the quantity and the quality of the data. In groundwater pollution studies and modelling, data can be both expensive and difficult to obtain. Asymptotic results are seldom useful because sample sizes are nearly always small. The laboratory analyses can be quite expensive and in the case of some organic pollutants there can be problems with non-detects; this will cause serious problems both in modelling distributions and also for spatial structure functions. The use of the indicator transform may partially ameliorate this but will not completely avoid it.

A Gaussian distribution (whether univariate or multivariate) is not necessary to derive the kriging equations (e.g., equations (11.8), (11.13) and (11.18), as the kriging estimators are only best *linear* estimators. Since the data are not viewed as replications, in the context of the random function model, one must be careful about using the usual statistical tests for Gaussian distributions since these in general will assume that spatial distributions are the same as ensemble distributions. The same kind of problem occurs in examining the validity of the model for $m(\mathbf{x})$ in (11.3) as well as the heteroscedastic assumption implicit in the use of the variogram or covariance. The data can be used to examine the reasonableness of the assumptions and model but to allow true statistical testing .

As has been seen in all of the above, estimation and modelling of the spatial structure function is a crucial step. The variogram must satisfy two conditions:

(i) $-\sum_{i=1}\sum_{j=1} \lambda_i \lambda_j \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) > 0$

for any $\boldsymbol{x}_i, \ldots, \boldsymbol{x}_n$ and any $\lambda_1, \ldots, \lambda_n$ (not all zero) such that

$$\sum_{i=1}^{\Delta} \lambda_i = 0$$

(ii) $\lim_{|\boldsymbol{h}| \to \infty} \gamma(\boldsymbol{h})/|\boldsymbol{h}|^2 = 0.$

The failure of the second condition is an indication of non-stationarity. Since it is not easy to verify these conditions directly, the practical solution is to use a family of known valid models. These all depend on at least two parameters; moreover, positive linear combinations of valid models are also valid hence relatively complicated models can be constructed from this small family. Although several alternatives to the sample variogram have been and are being used, modelling the variogram is still partly an art and relies on experience.

There are additional problems when the model is both spatial and temporal as would be common for the movement of plumes in the subsurface. Most commonly used variogram models are isotropic and hence independent of dimension; a geometric anisotropy can be incorporated essentially as a change in distance. This approach does not work in the case of spatio-temporal applications. Some authors constructed spatio-temporal variogram models as a positive linear combination of a spatial and a temporal variogram; Myers and Journel (1990), Rouhani and Myers (1990) and Myers (1992a) show that this can lead to non-invertible kriging matrices.

In the multivariate case, the spatial structure function is matrix valued and must satisfy a condition similar to (i) above where the λs are vectors. Whereas the diagonal elements must satisfy conditions (i) and (ii), the condition on the off-diagonal elements is only given in terms of the associated diagonal elements. The Cauchy–Schwartz condition is necessary but not sufficient when $m > 2$. The problem is further complicated by at least two different forms of a cross-variogram. The most common one is given by

$$\gamma_{st}(\boldsymbol{h}) = 0.5\ \mathrm{Cov}\{Z_s(\boldsymbol{x}+\boldsymbol{h}) - Z_s(\boldsymbol{x}), Z_t(\boldsymbol{x}+\boldsymbol{h}) - Z_t(\boldsymbol{x})\}. \qquad (11.21)$$

However, unlike the cross-covariance, $\gamma_{st}(\boldsymbol{h})$ is symmetric, i.e., $\gamma_{st}(\boldsymbol{h}) = \gamma_{st}(-\boldsymbol{h})$. Estimation of this cross-variogram requires that data be available on $Z_s(\boldsymbol{x})$, $Z_t(\boldsymbol{x})$ at the same locations. The pseudo-cross variogram

$$\hat{\gamma}_{st}(\boldsymbol{h}) = 0.5\mathrm{Var}\{Z_s(\boldsymbol{x}+\boldsymbol{h}) - Z_t(\boldsymbol{x})\} \qquad (11.22)$$

presented in Myers (1991) is not symmetric and does not require data at common locations but in general is harder to model. It also essentially requires second-order stationarity rather than the weaker intrinsic stationarity which is sufficient for equation (11.21). The cross-variogram given in (11.21) or a symmetric cross-covariance can be used in the Linear Coregionalization Model (LCM), this is a linear combination of valid variograms (or covariances) where the coefficients are positive-definite matrices. The validity of the model then is reduced to verifying that the coefficient matrices are positive definite. The LCM is based on the idea that each of the random functions $Z_s(\boldsymbol{x})$, $s = 1, \ldots, m$ can be written as a linear combination of uncorrelated second-order stationary random functions $Y_1(\boldsymbol{x}), \ldots, Y_p(\boldsymbol{x})$. Writing $\boldsymbol{Y}(\boldsymbol{x}) = [Y_1(\boldsymbol{x}), \ldots, Y_p(\boldsymbol{x})]$ this implies that

$$\boldsymbol{Z}(\boldsymbol{x}) = [Z_1(\boldsymbol{x}), \ldots, Z_m(\boldsymbol{x})] = [Y_1(\boldsymbol{x}), \ldots, Y_p(\boldsymbol{x})]A = \boldsymbol{Y}(\boldsymbol{x})A \qquad (11.23)$$

and hence the matrix variogram of $\boldsymbol{Z}$ is given in terms of the matrix variogram of $\boldsymbol{Y}$, which is diagonal. In hydrological applications or transport applications where there are state equations, the cross-covariances can sometimes be derived from those equations.

Several different simulation algorithms were described above, the choice between these is usually made on the basis of memory requirements, and speed of computation rather than on intrinsic characteristics of the simulated random functions. In particular little is known about the extent to which these are equivalent or even what equivalence means or should mean. This problem is discussed in greater detail in Myers (1994c).