Chapter 5
Geostatistics in Groundwater Modelling

Bhabesh C. Sarkar

1 Introduction

Geostatistics is a collection of statistical techniques for the analysis of spatial data. Geostatistics has been described by several authors (Matheron 1971; David 1977, 1986; Isaak and Srivastava 1989; Kitanidis 1997). In recent years, these tools have developed from research topics into basic techniques in the design and, such as mining, geology and hydrology, among others. The aim of this chapter is to present application of geostatistical tools in groundwater modelling and mapping. A typical spatial data set, such as groundwater levels, monthly precipitations, or transmissivities, is composed of scattered readings in space, denoted by $z(x)$, where x represents the measurement location. Having such information, geostatistics provides many techniques to solve a variety of hydrogeological resources problems, such as: (i) Estimation of z at an unmeasured location: interpolation and mapping of z; (ii) Estimation of one variable based on measurements of other variables: co-estimation of piezometric head and transmissivity; (iii) Estimation of the gradient of z at an arbitrary site: estimation of groundwater flow velocity based on observed heads; (iv) Estimation of the integral of Z over a defined block: estimation of contamination volume based on point measurements; and (v) Design of sampling and monitoring networks, such as groundwater quality monitoring. Many of the groundwater related variables are spatial functions presenting complex variations that cannot be effectively described by simple deterministic functions, such as polynomials. Such phenomena are subject of geostatistics that are named as regionalized variables. Annual point precipitation is an example of a regionalized variable. Transmissivity also displays spatial variations due to complex processes governing the transport, deposition and compression of materials in sedimentary deposits.

B. C. Sarkar (✉)
Department of Applied Geology, Indian Institute of Technology (ISM), Dhanbad, India
e-mail: bhabeshsarkar2005@gmail.com

P. K. Sikdar (ed.), *Groundwater Development and Management*,
https://doi.org/10.1007/978-3-319-75115-3_5

Another example of a regionalized variable is the concentration of a chemical compound in groundwater that varies in both space and time.

Geostatistics goes beyond the interpolation problem by considering the studied phenomenon at unsampled locations as a set of correlated random variables. Let $Z(x)$ be the value of the variable of interest at a certain location x. This value is unknown (e.g. temperature, rainfall, piezometric level, geological facies, etc.). Although there exists a value at location x that could be measured, geostatistics considers this value as random since it was not measured, or has not been measured yet. However, the randomness of $Z(x)$ is not complete, but defined by a cumulative distribution function (CDF). Typically, if the value of Z is known at locations close to x (or in the neighbourhood of x) one can constrain the CDF of $Z(x)$ by this neighbourhood. If a high spatial continuity is assumed, $Z(x)$ can only have values similar to the ones found in the neighbourhood. Conversely, in the absence of spatial continuity $Z(x)$ can take any value. The spatial continuity of the random variables is described by a model of spatial continuity that can be either a parametric function in the case of variogram-based geostatistics, or have a non-parametric form when using other methods such as multiple-point simulation or pseudo-genetic techniques. By applying a single spatial model on an entire domain, one makes the assumption that Z is a stationary process. It means that the same statistical properties are applicable on the entire domain.

The variations of these processes can be so complicated that estimating their values are difficult, even if measurements from nearby locations are available. Geostatistics recognizes these difficulties and provides statistical tools for: (i) calculating the most accurate (according to well defined criteria) predictions, based on measurements and other relevant information, (ii) quantifying the accuracy of these predictions, and (iii) selecting the parameters to be measured, and where and when to measure them, if there is an opportunity to collect more data. Considering that spatial data represents only an incomplete picture of the natural phenomenon of interest, it is logical to use statistical techniques to process such information. Geostatistics has adopted the procedure and some of the most practical and yet powerful applicable tools of probability theory (Rouhani 1986).

2 Statistical Modelling

The possible outcome of a random selection of a sample is expressed by its probability distribution that may or may not be known. In the case of a *discrete distribution*, which can only assume integer values, the distribution would associate to each possible value X, a probability $P(X)$. The individual value of $P(X)$ will be positive and the sum of all possible $P(X)$ will be equal to 1. The function $f(x)$ is a mathematical model that provides the probability that the random variable X would take on any specified value x, i.e. $f(x) = P(X = x)$. This function, $f(x)$ is called the probability distribution of the random variable X and describes how the probability values are distributed over the possible values, x of a random variable X. In the case of a *continuous distribution*, to each possible value x, a density of probability $f(x)$ is associated so that probability of a value lying between x and $x + dx$ is $f(x)\,dx$, where dx is infinitesimal. This serves as a mathematical model for describing the uncertainty of an outcome for a continuous variable. The probability of x lying between lower limit, (a) and upper limit, (b) is expressed as:

$$\mathrm{Prob}(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

The individual probability density value will be positive and the sum of all such values extending from –∝ to +∝ will be 1. The probability of X being smaller than or equal to a given value x is called the cumulative probability distribution function F (x):

$$\mathrm{Prob}(X \leq x) \int_{-a}^{b} f(x)dx = F(x); F(-\propto) = 0; \mathrm{and} F(+\propto) = 1$$

The following holds true for the cumulative distribution function, $F(x)$:

(i) $0 \leq F(x) \leq 1$ for all x;
(ii) $F(x)$ is non-decreasing.

The usual practice to determine the characteristics of an aquifer is to collect drill hole samples, analyse the properties of those samples and infer the characteristics of the aquifer from the properties. *If one uses classical statistics to represent the properties of sample values*, an assumption is made that the values are realisations of '*a random variable*'. The relative positions of the samples are ignored and it is assumed that all sample values in aquifer have an equal probability of being selected. The fact that two samples taken close to each other is more likely to have similar values than if taken far apart is also not taken into consideration.

Sample spacing remains wide in the initial stages of groundwater exploration that provide broad knowledge of an aquifer. It is in this early stage of exploration, quality of the aquifer is examined by estimating mean (average) value, 'm' of the aquifer. For this purpose, 'n' samples of same support (size, shape and orientation) are taken at points X_{i}. The sample values are used to estimate 'm' of the population mean, μ and the confidence limits of the mean. The estimator for this purpose would vary according to the probability distribution of sample values. In classical statistical analysis, since it is assumed that all sample values are independent (i.e. random), the location X_{i} of the sample is ignored. The parameters estimated from a classical statistical model refer to variables such as thickness, permeability, porosity, etc. Theoretical models of probability distributions which are commonly encountered in aquifers to represent sample value frequency distribution are either Normal (Gaussian) or Lognormal. Various other distributions are known but the assumption of either normality or lognormality can be made for most aquifers and the use of more complex distributions is not justified.

2.1 The Normal Distribution Theory

This distribution is characterised by a symmetrical bell-shape and its probability density function (p.d.f.), $f(X)$ is expressed (Davis 1986) as:

$$\text{p.d.f.}, f(X) = \left[\left(1/S\sqrt{2\pi}\right)\right]\exp\left[-\left((1/2)\left(X_i - \bar{X}\right)^2\right)/S^2\right] \text{ for } -\infty \leq X \leq \infty$$

where $\bar{X}$ is the sample mean which is an estimate of the population mean μ, and S is the sample standard deviation, an estimate of the population standard deviation σ. The distribution can be standardised by expressing $\left[\left(X_i - \bar{X}\right)/S\right]$ equal to Z:

$$f(Z) = \left[1/\sqrt{2\pi}\right]\exp\left[-\tfrac{1}{2}Z^2\right]$$

This standard normal distribution has a zero mean and unit standard deviation, i.e. N(0,1). The cumulative probability density function (c.d.f.), F(X) of a normal distribution has the expression:

$$\text{c.d.f.}, F(X) = \left[1/\sqrt{2\pi}\right]\int_{-\alpha}^{x} \exp\left[-\left((1/2)\left(X_i - \bar{X}\right)^2\right)/S^2\right]dx$$

2.2 Fitting a Normal Distribution

To check the assumption of normality, or in other words, to fit a normal distribution to an experimental histogram, a convenient graphical method known as the probability-paper method can be used. Cumulative frequency distribution of the values are calculated and plotted in an arithmetic-probability paper against the upper limits of the class values. From the definition of arithmetic-probability scale, the cumulative distribution of a normally distributed variable will plot as straight line on arithmetic-probability paper. If the points obtained by this approach can be considered or closely approximated as distributed along a straight line, the assumption of normality can be accepted, and the theory of normal distribution to estimate the mean, variance and confidence limits of mean can then be applied.

Other methods to test the fit of a normal distribution include: (i) measures of degree of skewness and kurtosis, and (ii) χ^2 (Chi-squared) goodness of fit test. For a normal variate, the degree of skewness is zero and that of kurtosis is 3, and the calculated value of χ^2 must be less than or equal to the table value of χ^2 at '$\propto$' level of significance and '*f*' degrees of freedom.

2.3 Estimation of Mean, Variance and Confidence Limits

The sample mean and sample variance for a normal distribution are estimated as follows:

$$\text{Sample mean}, \bar{X} = [1/n]\sum_{i=1}^{n} X_i$$
$$\text{Sample variance}, S^2 = [1/(n-1)]\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

where $S = \sqrt{}S^2$ which is an estimate of the population standard deviation. The mean value, '*m*' of the aquifer is estimated by:

$$m = \bar{X}; \text{with variance}, V = S^2/n$$

If m_{p} be confidence limits of the true mean '*m*' such that the probability of '*m*' being less than m_{p} is *p*, then $m_{1\text{-p}}$ is the confidence limit such that the probability that '*m*' is larger than $m_{1\text{-p}}$ is $1 - p$. The probability that '*m*' falls between m_{p} and $m_{1\text{-p}}$ is $1 - 2p$ confidence limits of the mean. The following equations can be used to calculate m_{p} and $m_{1\text{-p}}$ for the mean value, '*m*' of an aquifer:

$$\text{Lower limit}, m_{\mathrm{p}} = m - t_{1-\mathrm{p}}\left(S/\sqrt{}n\right); \text{and}$$
$$\text{Upper limit}, m_{1\text{-p}} = m + t_{1-\mathrm{p}}\left(S/\sqrt{}n\right)$$

where $t_{1\text{-p}}$ is the value of student's *t*-variate for $f = n - 1$ degrees of freedom, such that the probability that '*t*' is smaller than '$t_{1\text{-p}}$' is 1 – p.

2.4 Measures of Skewness, Kurtosis and Chi-squared goodness of Fit

Degrees of skewness and kurtosis of a sample distribution are given by the equations:

$$\text{Skewness}, Sk = [1/(n-1)]\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^3/S^3$$
$$\text{Kurtosis}, Ku = [1/(n-1)]\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^4/S^4$$

Once the optimum solution for '*m*' has been determined, it is desirable to check for the goodness of fit of a normal distribution to the sample distribution. Chi-squared $(\chi)^2$ test provides a robust technique for the fit. The test statistics is given by:

$$\chi^2\text{Calculated} = \sum_{i=1}^{n} (O_i - E_i)^2/E_i$$

where O_i = observed frequency in group *i* and E_i = expected frequency in group *i*.

2.5 The Lognormal Distribution Theory

In many aquifers, where the distribution of the sample values is asymmetrical, either positively or negatively skewed, it has been observed that this skewed distribution can be represented either by a 2-parameter or a 3-parameter lognormal distribution. If $\log_e(X_\mathrm{i})$ has a normal distribution, we call it a 2-parameter lognormal distribution, and if $\log_e(X_\mathrm{i} + C)$ has a normal distribution, we call it a 3-parameter lognormal distribution (where *C* is the additive constant). The value of the additive constant, *C* is:

(i) Positive for a positively skewed distribution, i.e. a distribution showing an excess of low values with tail towards high values; and
(ii) Negative for a negatively skewed distribution, i.e. a distribution showing an excess of high values with tail towards low values.

The p.d.f. of a lognormal distribution is given by the expression:

$$f(X) = \left[1/\left(x\beta\sqrt{}2\pi\right)\right]\exp\left[-1/2\left(\frac{(\ln x - \alpha)}{\beta}\right)^2\right]$$

where α = logarithmic mean, i.e. log mean and β^2 = logarithmic variance, i.e. log variance.

The probability distribution of a 3-parameter lognormal variate, X_i is defined by:

- the additive constant, *C*;
- the logarithmic mean of $(X_\mathrm{i} + C)$
- the logarithmic variance of $(X_\mathrm{i} + C)$.

2.5.1 Fitting a Lognormal Distribution

For '*n*' samples with values X_i ($i = 1, 2, ..., n$), the cumulative frequency distribution of a 2-parameter lognormal variate plots as a straight line on logarithmic probability paper. If the variate is 3-parameter lognormal, the cumulative curve shows either an excess of low values for positively skewed distribution and or an excess of high values for negatively skewed distribution. In such cases, plot of (X_i + C) will be a straight line on logarithmic probability paper conforming to a lognormal distribution.

2.5.2 Estimation of Additive Constant (C)

If a large number of samples are available, the cumulative distribution may be plotted on a log-probability paper. Different values of 'C' can then be tried until the plot of (X_i + C) is reasonably assumed to be a straight line. Alternatively, the value of 'C' can be estimated using the following approximation:

$$C = \frac{M_\mathrm{e}^2 - F_1 F_2}{F_1 + F_2 - 2M_\mathrm{e}}$$

where M_e is the sample value corresponding to 50% cumulative frequency (i.e. the median of the observed distribution) and F_1 and F_2 are sample values corresponding to 'p' and '$1 - p$' percent cumulative frequencies respectively. In theory, any value of 'p' can be used but a value between 5% and 20% gives best results.

2.5.3 Estimation of Logarithmic Mean and Logarithmic Variance

$$\begin{aligned}\text{Let}, y_i &= \log_\mathrm{e}(X_i + C)\\ \log_\mathrm{e}\text{ mean}, \alpha \text{ or } \bar{Y} &= [1/n]\sum_{i=1}^{n} y_i\\ \log_\mathrm{e}\text{ variance}, \beta^2 \text{ or } v(y) &= [1/(n-1)]\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\end{aligned}$$

2.5.4 Estimation of Average for a Deposit

$$\begin{aligned} m^* &= e^{\bar{\mathrm{y}}+\mathrm{v}(\mathrm{y})}\\ &= e^{\left(\alpha+\left(\beta^2\right)\right)/2}\\ &= e^{\alpha}.e^{\left(\beta^2/2\right)}\\ \text{Average value,} \qquad m &= (m^* - c)\\ \text{Variance,} \qquad S^2 &= m^2[\exp(v) - 1]\end{aligned}$$

2.5.5 Estimation of Central 90% Confidence Limits

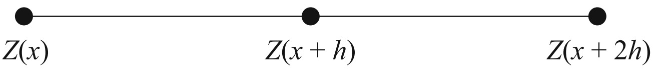
The lower and upper limits for the estimation of Central 90% confidence interval of the mean of a lognormal population can be obtained by using factors $\psi_{0.05(\mathrm{v,n})}$ and $\psi_{0.95(\mathrm{v,n})}$:

$$\begin{aligned}\text{Lower limit} &= \left(\psi_{0.05(\mathrm{v,n})} m^*\right) - C; \text{ and}\\ \text{Upper limit} &= \left(\psi_{0.95(\mathrm{v,n})} m^*\right) - C.\end{aligned}$$

3 Geostatistical Modelling

Classical statistics produce an error of estimation stated by confidence limits but ignores the spatial relations within a set of sample values. These limitations point to the need for an estimation technique that is capable of producing estimates with minimum variance. Such estimates are achieved with the use of geostatistics based on the *'Theory of Regionalised Variables'*, i.e. a variable that is related to its position in space and has a constant support.

The underlying assumption of geostatistics is that the values of samples located near or inside a block of ground are most closely related to the value of the block. This assumption holds true if a relation exists among the sample values as a function of distance and orientation. The function that measures the spatial variability among the sample values, is known as the semi-variogram function, $\gamma(h)$. Comparisons are made between each sample of a data set with the remaining ones at a constantly increasing distance, known as the lag interval.



Thus, a semi-variogram function numerically quantifies the spatial correlation of aquifer parameters (e.g. thickness, bedrock elevation, porosity, permeability etc.). If $Z(x_\mathrm{i})$ be the value of a sample taken at position x_i and $Z(x_\mathrm{i} + h)$ be the value at 'h' distance away from x_i position, the mathematical formulation of a semi-variogram function, $\gamma(h)$ is given by the expression:

$$\gamma(h) = (1/2N)\sum_{i=1}^{N}(Z(x_i) - Z(x_i + h))^2$$

where N is number of sample value pairs, $Z(X_\mathrm{i})$ is the value of Regionalized Variable at location X_i and $Z\ (X_\mathrm{i} + h)$ is the value of Regionalized Variable at a distance 'h' away from X_i.

Spatial variance changes from arrangement to arrangement. The function $2\gamma(h)$ is called the variogram function. It is the semi-variogram function $\gamma(h)$ that is used rather than variogram function $2\gamma(h)$ because the relation between semi-variogram and covariogram (i.e. plot of covariance between $Z(x_\mathrm{i})$ and $Z(x_\mathrm{i} + h)$ with constantly increasing values of 'h') is straight forward:

$$2\gamma(h) = E\left[2\gamma(h)^*\right]$$

where E is the Expected Value which is the probability weighted sum of all possible occurrences of regionalized variable; and $2\gamma(h)$* is the experimental variogram function based on sample values;

$$\begin{aligned}
\text{or,}\quad & 2\gamma(h)\\
&= E\left[2\gamma(h)^*\right] = E\left[(Z(x_i) - Z(x_i + h))^2\right]\\
&= E\left[(Z(x_i) - m + m - Z(x_i + h))^2\right] \text{ where } m \text{ is the sample mean}\\
&= E\left[((Z(x_i) - m) - (Z(x_i + h) - m))^2\right]\\
&= E\left[(Z(x_i) - m)^2 + (Z(x_i + h) - m)^2 - 2(Z(x_i) - m)(Z(x_i + h) - m)\right]\\
&= E\left[(Z(x_i) - m)^2\right]\\
&\quad + E\left[(Z(x_i + h) - m)^2 - 2E[(Z(x_i) - m)(Z(x_i + h) - m)] = 2 \text{ variance}\right.\\
&\quad -2 \text{ covariance } (h)
\end{aligned}$$

Hence the fundamental relation: $\gamma(h) = \sigma^2 - CV(h)$.

Graphical representation of a semi-varigram is given in Fig. 5.1.

An experimental semi-variogram permits the interpretation of several characteristics of the aquifer as follows:

(i) *The Continuity (C):* The continuity is reflected by the rate of growth of $\gamma(h)$ for constantly increasing values of 'h'.

(ii) *The Nugget Effect (C_o):* This is the name given to the semi-variogram value, $\gamma(h)$ at $h \to 0$. It expresses the local homogeneity (or lack thereof) of aquifer. The nugget effect represents an inherent variability of a data set which could be due to both the spatial distribution of the values together with any error encountered in sampling.

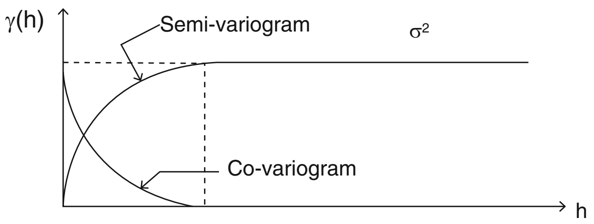


Fig. 5.1 Relation between semi-variogram and co-variogram

(iii) *The Sill Variance (C_o + C):* The value where a semi-variogram function $\gamma(h)$ plateaus is called the sill variance. For all practical purposes, the sill variance is equal to the statistical variance of all sample values used to compute an experimental semi-variogram.

(iv) *The Range (a):* The distance at which a semi-variogram levels off at its plateau value is called the range (or zone) of influence of semi-variogram. This replaces the conventional geological concept of an area of influence. Beyond this distance of separation, values of sample pairs do not correlate with one another and become independent of each other.

(v) *The Directional Anisotropy:* This denotes whether or not the aquifer has greater continuity in a particular direction compared to other directions. This characteristic is analysed by comparing the respective ranges of influences semi-variograms computed along different directions. Where the semi-variograms in different directions are very similar, it is said to be isotropic.

In practice, since sampling grids are rarely uniform, semi-variograms are computed with a tolerance on distance (i.e., $h \pm dh$) and a tolerance on direction (i.e. $\alpha \pm d\alpha$) to accommodate sample pairs not falling on the grid. The tolerances on distance and direction should be kept as low as possible in order to avoid any directional overlapping.

4 Semi-Variogram Models

There are several mathematical models of semi-variogram. However, three most commonly encountered models in aquifer modelling (Fig. 5.2) are:

4.1 Spherical Model

This model is encountered most commonly in aquifer where sample values become independent once a given distance of influence (i.e. the Range) '*a*' is reached. The equations are given by:

$$\begin{array}{ll}
\gamma(h) = Co + C\left[3/2(h/a) - ½\left(h^3/a^3\right)\right] & \forall h < a; \\
\gamma(h) = Co + C & \forall h \geq a; \\
\gamma(h) = Co & \forall h \text{ tends to } 0; \\
\gamma(h) = 0 & \forall h = 0.
\end{array}$$

This model is common in most aquifers and said to describe transition phenomena as it is the one which occurs when one has geostatistical spatial structures independent of each other beyond the range but, within it, sample values are highly correlated.

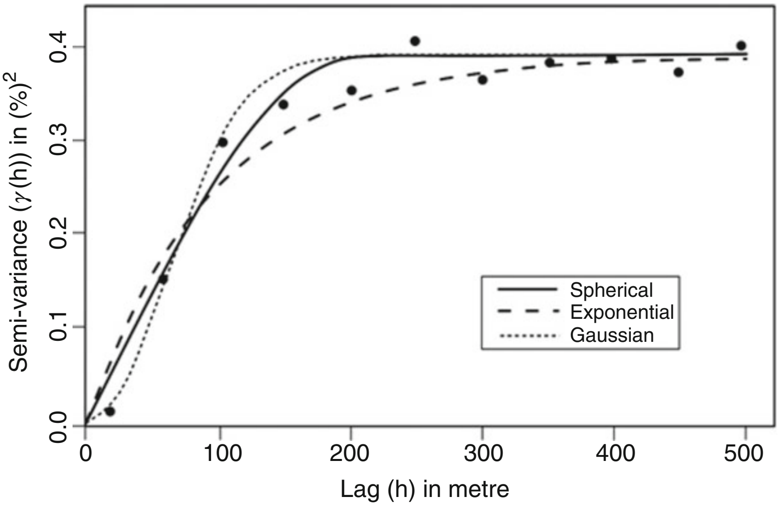


Fig. 5.2 Common semi-variogram models

4.2 Exponential Model

This model is not encountered too often in aquifers since its infinite range is associated with a too continuous process. The equation is: $\gamma(h) = C\ [1 - e^{-h/a}$]. The slope of the tangent at the origin is C/a. For practical purposes, the range can be taken as $3a$. The tangent at the origin intersects the sill at a point where 'h' equals 'a'.

4.3 Gaussian Model

This model is characterised by two parameters C and a. The curve is parabolic near the origin and the tangent at the origin is horizontal, which indicates low variability for short distances. Excellent continuity is observed which is rarely found in geological environments. Practical range is $\sqrt{3}a$. The equation is: $\gamma(h) = C\left[1 - e^{\left(-\mathrm{h}^2/\mathrm{a}^2\right)}\right]$.

5 Practice of Semi-VARIOGRAM Modelling

The behaviour at the origin for both nugget effect and slope plays a crucial role in fitting of a model to an experimental semi-variogram. While the slope can be assessed from the first three or four semi-variogram values, the nugget effect can be estimated by extrapolating back to the γ(h) axis. The choice of nugget effect is extremely important since it has a very marked effect on kriging weights and in turn on kriging variance. Three methods for semi-variogram model fitting include:

5.1 Hand Fit Method

The sill ($Co + C$) is set at the value where experimental semi-variogram stabilizes. In theory, this should coincide with the statistical variance. Estimate of nugget effect is achieved by joining the first three or four semi-variogram values and projecting this line to the γ(h) axis. By projecting the same line until it intercepts the sill provides 2/3rd the range. Using the estimates of Co, C and 'a', calculate a few points and examine if the model curve fits the experimental semi-variogram. Although this method is straight forward, and simple to practice, there is an element of subjectivity involved in the estimation of model parameters.

5.2 Non-linear Least Squares Fit Method

Like any curve fitting technique, this method uses the principle of polynomial fit by least squares to fit a model with sum of the deviations squared of the estimated values from the real values being minimum. Unfortunately, polynomials obtained by least squares do not guarantee the positive definite function (otherwise semi-variance could turn out to be negative).

5.3 Point Kriging Cross-Validation Method

Point kriging cross-validation (PKCV) is a technique referred to as a procedure for checking the validity of a mathematical model fitted to an experimental semi-variogram that controls the kriging estimation (Davis and Borgman 1979).

The principle underlying the technique is as follows:

> '……… a sample point is chosen in turn on the sample grid that has a real value. The real value is temporarily deleted from the data set and the sample value is kriged using the neighbouring sample values confined within its radius of search. The error between the estimated value and the real value is calculated. The kriging process is then repeated for rest of the known data points'. A crude semi-variogram model is initially fitted by visual inspection to the experimental semi-variogram. Estimates of the initial sets of semi-variogram parameters (viz., Co, C and 'a') are made from the initial model and cross-validated through point kriging empirically. The error statistics such as mean error, mean variance of errors and mean kriging variance are then computed. The model parametes are varied and adjusted until: (i) a ratio of mean variance of the errors (estimation variance) to mean kriging variance approximating to unity (in practice, a value of 1 ± 0.05 has been

observed to be the acceptable limits); (ii) a mean difference between sample values and estimated values close to zero; and (iii) an adequate graphical fit to the experimental semi-variogram are achieved. For a good estimate, most of the individual errors should also be close to zero. A model approximated or fitted by this approach eliminates subjectivity.

6 Geostatistical Estimation – Kriging

Kriging is an optimal spatial interpolation technique. In general terms, a kriging system calculates an estimated value, G^* of a real value, G by using a linear combination of weights, a_i of the selected surrounding 'n' values such that:

$$G^* = \sum_{i=1}^{n} a_i g_i, \text{ where } \sum_{i=1}^{n} a_i = 1 \text{ and } g_i \text{ are the sample values.}$$

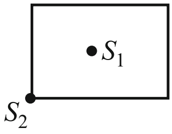
If G* is the estimate of a block average grade G by applying straight average method, i.e.

$$G^* = 1/\mathrm{n} \sum_{i=1}^{n} g_i$$

then equal weight is given to all the sample values, and the error of estimation of G is:

$$\sigma_E^2(S \text{ to } V) = E\left[(G^* - G)^2\right] = -\overline{\gamma}(S, S) - \overline{\gamma}(v, v) + 2\overline{\gamma}(S, V)$$

In many cases, however, we know that to assign equal weight to all selected surrounding samples may not provide the best possible estimate. Consider the case of a block valued by a centre sample and a corner sample as configured below:



Clearly, the centre sample should be given a greater weight than the corner sample. Say, we give weight a_1 to S_1 and a_2 to S_2. The new grade estimate would be:

$$G^* = a_1 g_1 + a_2 g_2$$

The weights of selected surrounding sample values are so chosen that:

- G^* is an unbiased estimate of G, i.e. $E\,[(G^* - G)] = 0$; and
- Variance of estimation of G by G^*, i.e. $E\,[(G^* - G)^2]$ is minimum.

By definition, Kriging is known as Best (because of minimum estimation variance) Linear (because of weighted arithmetic average) Unbiased (since the weights sum to unity) Estimator – BLUE.

7 Practice of Kriging

Once the model semi-variogram parameters characterizing all information about the expected sample variability are defined, the subsequent step involves estimation of grid cell values together with their associated variances through kriging. At this stage, a geological domain is considered within an aquifer which is further divided into smaller grids equalling the size of a geocellular block. Decision on the choice of a geocellular block size is generally influenced by several factors such as measurement points, hydrogeological framework, precision of measurement data, desired use of grid cell, and capability of manipulating a huge number of grid cells. The arrays of geocellular blocks are kriged producing kriged estimate and kriging variance for each of them and an overall average. The following input parameters are found to be adequate for geocellular block kriging:

- a minimum of four measurement points (because of the necessity to define a surface) and a maximum of 16 measurement points (because of reasonable computational time and cost) with at least one measurement point in each quadrant to krig a geocellular block; and
- the radius of search for measurement points around a geocellular block centre to be within the semi-variogram range of influence.

The individual values are averaged to produce a mean kriged estimate and a mean kriging variance in order to provide global estimates. The 95% geostatistical confidence limits are calculated as:

$$m \pm 1.96\sqrt{\sigma_k^2},$$

where m = mean kriged estimate and σ_{k}^2 = mean kriging variance.

8 Applications

Geostatistics can be used in a variety of groundwater modelling studies, such as: (i) mapping of spatial variables; (ii) simulation of hydrogeological fields; (iii) co-estimation of hydrological fields using physical relationships, such as co-mapping of piezometric head and transmissivity using groundwater flow equations; (iv) sampling and monitoring designs; and (v) groundwater resource management under uncertainty. One of the earliest applications of geostatistics in groundwater was in the area of mapping spatial variables, such as transmissivity maps, piezometric surfaces, and precipitation fields. Kitanidis (1997) in his book has dealt with geostatistics and applications to hydrogeology. In fact, the power of the methods described becomes most useful when utilizing measurements of different types, combining these with deterministic flow and transport models, and incorporating geological information to achieve the best characterization possible.

Geostatistical mapping also yields the accuracy map that indicates the areas of high and low precision. Simulation of hydrogeological fields is another application of geostatistics. Simulation usually means the generation of spatial data, such that their mean and their covariance are the same as the original data. There are various useful applications for simulated data. For example, by generating different spatial rainfall patterns, one can determine the statistical distribution of runoff. Co-estimation allows the user to utilize the information in one variable in the estimation of another. In some instances, a variable that is sampled at a lower cost can be used to improve the accuracy of another variable which is costly to measure. If there are known physical relationships between variables, they can be used to further improve our estimation. The estimation variance is a measure for the accuracy of estimated fields. This measure can help us to design sampling activities based on the maximization of gained information. In some instances, such as groundwater quality monitoring, the estimated magnitude of the variable of interest is as important as its accuracy. So the sampling may be designed not only for improving the precision of the estimated field, but also for targeting those areas which exhibit critical estimated values. Water resources management problems usually include many variables that exhibit uncertainty. Ignoring the stochastic nature of these problems may yield non-optimal solution. Geostatistics provides the framework to quantify these uncertainties and incorporate them in our decisions.

9 A Case Study

A study has been aimed at modelling of spatial phenomena of groundwater distribution during pre-monsoon and post-monsoon periods in respect of the year 2014 using geostatistical techniques with reference to rainwater harvested groundwater level inside the IIT(ISM) campus. The groundwater level data of 44 recharge bore wells located within the ISM campus area were collected on a monthly basis during the year 2014. Pre-monsoon (May to June), monsoon (July to September) and post-monsoon (October to December) measurements of groundwater level data of the recharge bore wells have been utilized for spatial modelling of the groundwater fluctuation employing the theory and applications of ‘Regionalised Variable’. The modelling study reveals the spatial variability of the fluctuation and estimates the rise in the groundwater level employing Ordinary Kriging.

Statistical analyses of groundwater level data for these periods were carried out to compute the distribution parameters. Geostatistical methods utilize an understanding of the inter-relations of measurement (sample) values and provide a basis for quantifying the geological concepts of (i) an inherent variability; (ii) a change in the continuity of inter-dependence of measurement (sample) values according to the spatial variability; and (iii) a range of influence of the inter-dependence of measurement (sample) values. Based on these quantifications, geostatistics produces an estimated map with minimum variance, and provides an error of estimation both on a local and a global scale. The underlying assumption of geostatistics is that the values of samples located nearby are most closely related to one other than the distant ones. This assumption holds true if a relation exists among the sample values as a function of distance and orientation. The function that measures the spatial variability among the sample values, is known as the semi-variogram function, γ(h). Comparisons are made between each sample of a data set with the remaining ones at a constantly increasing distance, known as the lag interval.

Geostatistical analysis was initiated with computation of experimental semi-variogram and fitting appropriate mathematical model to it that characterizes the spatial variability of the groundwater level. A semi-variogram model exhibit various spatial characteristics, viz. nugget effect (C_0), continuity (C), sill ($C_0 + C$), range of influence (a) and directional anisotropy. Semi-variogram constitute the major tool in geostatistics to express the spatial dependence among neighbouring values measured in pairs. Most commonly used mathematical models of semi-variogram include spherical, exponential, gaussian, and pure nugget effect. The behaviour at the origin for both nugget effect and slope plays a crucial role in fitting of a model to an experimental semi-variogram. While the slope can be assessed from the first three or four semi-variogram values, the nugget effect can be estimated by extrapolating back to the γ (h) axis. The choice of nugget effect is extremely important since it has a very marked effect on kriging weights and in turn on kriging variance. The appropriateness and rationality of a semi-variogram model fit was carried out employing point kriging cross-validation technique.

3D omni-directional experimental semi-variograms for pre- and post-monsoon groundwater levels and that of the fluctuations have been carried out using GEXSYS software (Sarkar 1988). The spatial variability analyses revealed experimental semi-variograms with moderately low nugget effect and increasing tendency of semi-variogram values with constantly increasing distances levelling off at respective range of influences. Cross-validated models as obtained employing point kriging cross-validation technique for Pre-monsoon, Post-monsoon, and Fluctuation are given in Figs 5.3, 5.4 and 5.5 and semi-variogram model parameters obtained through point kriging cross-validation are given in Table 5.1.

Prior to the grid cell kriging, the grid size of the study area was decided by taking into account the various parameters i.e. area, fluctuation of groundwater and the best fitted grid cell which can cover the maximum extent near to the boundary of the ISM. A grid cell size of 25 m × 25 m dimensions was selected on the basis of appropriate fitting of the cells in the periphery of the boundaries. Having delineated the cells of the dimension of 25 m × 25 m ordinary kriging was performed cell by cell to provide kriged estimate and kriged standard deviation. The plots show the spatial distribution of groundwater levels generated in the study area. The plots of pre-monsoon (Fig. 5.6) and post-monsoon (Fig. 5.7) groundwater levels display the spatial

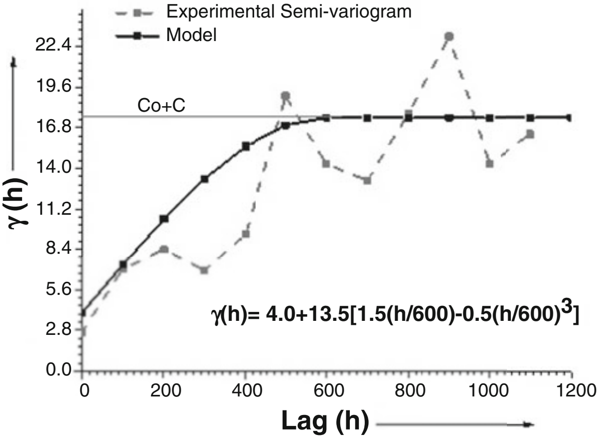


Fig. 5.3 Semi-variogram model of pre-monsoon period

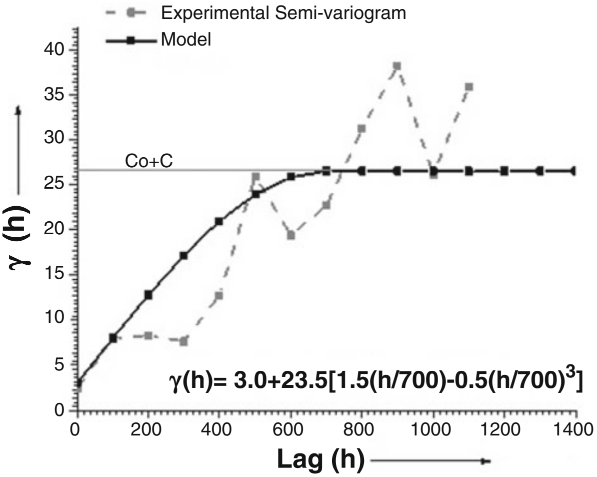


Fig. 5.4 Semi-variogram model of post-monsoon period

distribution maps of kriged estimate and kriged standard deviation of groundwater levels along with that of the fluctuation (Fig. 5.8) in the study area. Groundwater flow maps of pre- and post-monsoon periods have been developed that provide the direction of flow of groundwater (Fig. 5.9).

Statistical analyses of pre-monsoon and post-monsoon groundwater levels provided a negatively skewed characteristic while that in respect of the fluctuations between pre- and post-monsoon provided a positively skewed characteristic. Estimated mean and standard deviation values corresponding to each of these periods and that of the fluctuation are (240.55 m; 4.08 m), (242.44 m; 3.40 m) and (1.09 m; 1.44 m) respectively.

Spatial variability analyses of pre-monsoon and post-monsoon groundwater levels and that of the fluctuation between pre- and post-monsoon periods revealed a spherical function fit. Pre-monsoon period exhibited a nugget effect of 3.0 m^2, a

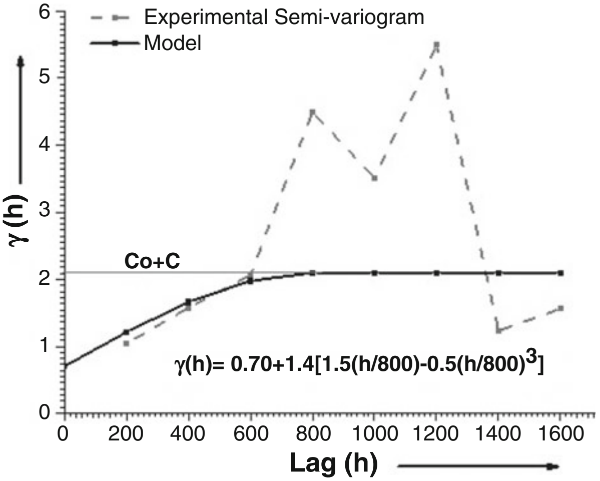


Fig. 5.5 Semi-variogram model of fluctuation

Table 5.1 Semi-variogram model parameters obtained through Point Kriging Cross-validation

Semi-variogram parameters	Pre-monsoon period	Post-monsoon period	Fluctuation
Nugget effect, C_0	3.0	4.0	0.7
Continuity, C	23.5	13.5	1.40
Sill, $C_0 + C$	26.5	17.5	2.1
Range (a)	700	600	800
Ratio of KV:EV	0.98	1.04	1.03

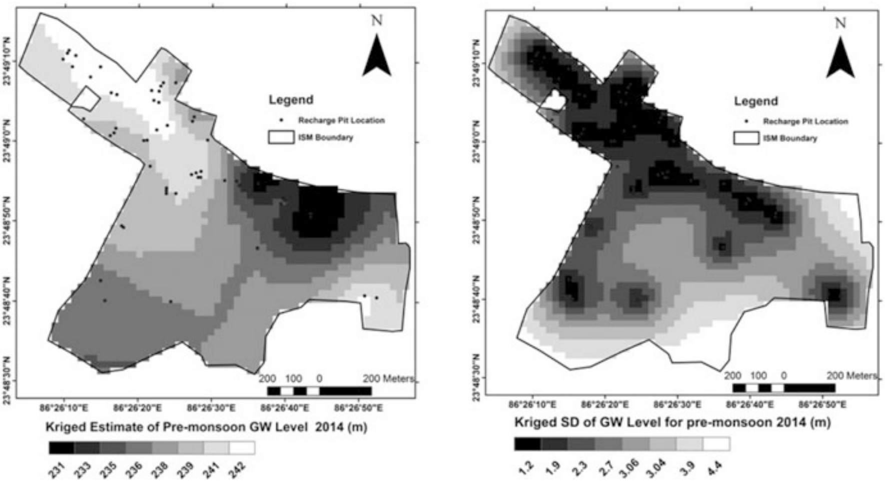


Fig. 5.6 Spatial distribution of kriged estimate and kriged standard deviation of groundwater levels in respect of pre-monsoon period

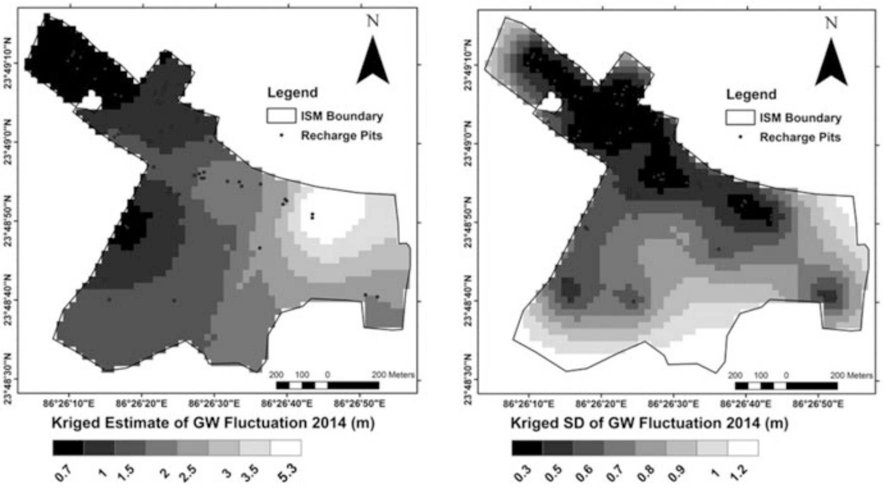


Fig. 5.7 Spatial distribution of kriged estimate and kriged standard deviation of groundwater levels in respect of post-monsoon period

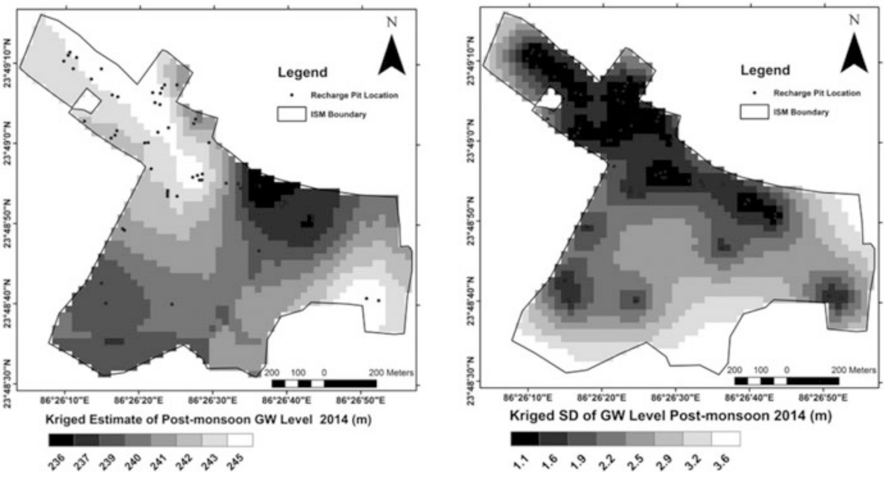


Fig. 5.8 Spatial distribution of kriged estimate and kriged standard deviation of groundwater levels in respect of fluctuation

continuity of 23.5 m^2 and a range of 700 m; post-monsoon period exhibited a nugget effect of 4.2 m^2, a continuity of 13.5 m^2 and a range of 600 m; fluctuation between pre- and post-monsoon periods displayed a nugget effect of 0.7 m^2, a continuity of 1.40 m^2 and a range of 800 m. Model fitting exercise has been carried out employing point kriging cross-validation technique yielding a ratio of estimation variance to

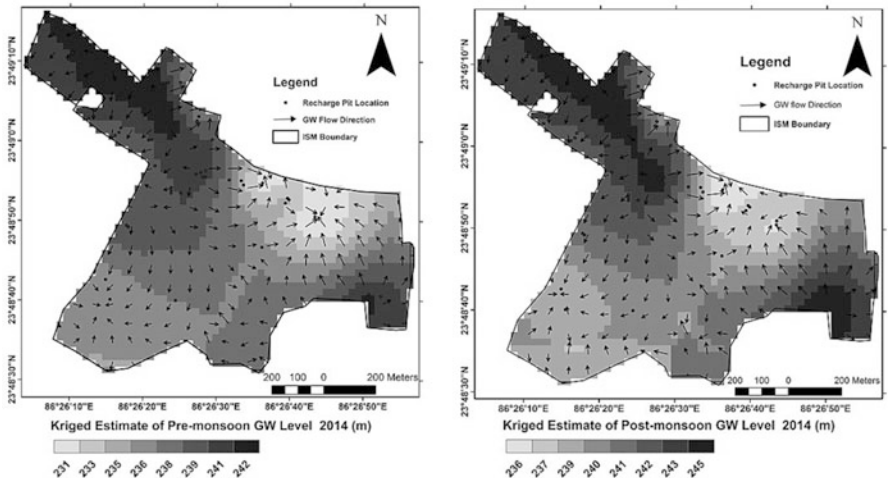


Fig. 5.9 Groundwater flow maps of pre- and post-monsoon periods

kriging variance as 0.98, 1.04 and 1.03 respectively with adequate graphical fit to experimental semi-variograms. Following are the geostatistical model equations of groundwater levels:

$$\text{Pre-monsoon} \; : \gamma(h) = 3.0 + 23.5\left[1.5(h/700) - 0.5(h/700)^3\right]$$

$$\text{Post-monsoon} \; : \gamma(h) = 4.0 + 13.5\left[1.5(h/600) - 0.5(h/600)^3\right]$$

$$\text{Fluctuation} \; : \gamma(h) = 0.7 + 1.40\left[1.5(h/800) - 0.5(h/800)^3\right]$$

Geostatistical estimation was initiated with gridding the IIT (ISM) campus area into cells 25 m × 25 m with each cell defined in space in terms of northing and easting. Block kriging has been carried out for each of these cells which provided kriged estimate and associated kriged standard deviation in respect of pre-monsoon and post-monsoon groundwater levels and that of the fluctuation between pre- and post-monsoon periods. Assessment of the goodness of fit (R^2) of kriged estimates was carried out through a correlation plot of the measured (true) values versus the kriged estimated values of groundwater level and assuming the best fit straight line as the regression model. The *R* values of pre- and post-monsoon and that of fluctuation are 0.94, 0.88 and 0.88 respectively and have been found to be significant through *t*-test of significance on *R*. The calculated values of '*t*' on *R* are 18.40, 12.80 and 11.90 as compared to the critical value of 1.67, thereby indicating a significant correlation.

Spatial distribution maps of kriged estimate values in respect of pre-monsoon and post-monsoon groundwater levels exhibit a distinct high in the north-western zone and in the south-eastern periphery of the campus associated with geomorphic high. The groundwater levels gradually decrease towards east-central periphery and towards south-western zone with a distinct low observed in the east-central periphery. The kriged standard deviation maps exhibit a relatively high error in the north-western, south-western, southern and eastern peripheries of the campus reflecting a reduced reliability of kriged estimate due to presence of few recharge bore wells (sample location) while in rest of the part, the error reduces towards the areas with high density of recharge bore wells as prominently observed in the north-western zone trending towards east-central part and also in the south-eastern part and north-western part. As regards the spatial distribution map of fluctuation is concerned, it is observed that the kriged estimated fluctuation is maximum in the east-central periphery owing to the presence of geomorphic depression. There is a gradual decrease in the fluctuation towards the north-western side and also towards the western side of the campus. The associated kriged standard deviation has the same distribution pattern as with that of the pre- and post-monsoon kriged standard deviation maps for the same reason of presence of few recharge bore wells. The fluctuation map at the north-western side has the least rise of 0.70 m in groundwater level which gradually increases in the eastern side to 5.3 m. It may be stated that areas showing higher density of recharge bore wells are associated with lower kriged standard deviation which gradually increases towards areas with lesser density of recharge bore wells as evident from the spatial distribution maps of kriged standard deviation. Groundwater flow maps exhibit a similar pattern, i.e. the direction of flow of groundwater is from northwest, west, southwest and southeast sides to the east owing to the geomorphic low.

Attempt made to estimate the replenishable groundwater resource within the campus area using the norms of Groundwater Resource Estimation Committee (GEC) 1997 led to calculation of total replenishable volume of water or dynamic groundwater resource using mean kriged rise of groundwater level of 2.77 m as:

$$\text{Volume of water recharge} = (\text{Area} \times \text{fluctuation} \times \text{specific yield}) + \text{Draft}$$

where specific yield considered for hard rock as per GEC, 1997 is 0.03. Hence, volume of water recharge = ($60 \times 56 \times 25 \times 25 \times 2.77 \times 0.03$ m^3) + Draft. The first term of the equation 174,510 m^3 × 1000 = 174,510,000 litres and the Draft, which is related to the consumption of groundwater for the year 2014 is 19,54,000 × 12 × 30 = 703,440,000.00 litres. Dynamic groundwater resource thus calculated is 877,950,000 litres.

Spatial variability phenomena of groundwater level within the campus of IIT (ISM) Dhanbad for pre-monsoon and post-monsoon periods and that of the fluctuation between pre- and post-monsoon have been analysed and modelled. The spatial variability analyses revealed experimental semi-variograms with moderately low nugget effect and increasing tendency of semi-variogram values with constantly increasing distances and levelling off at respective range of influences. Point Kriging Cross-validation technique has been used for fitting a mathematical model to experimental semi-variograms. This is followed by construction of block grid cells of 25 m × 25 m for which kriged estimate and kriging standard deviation values have been arrived at employing ordinary kriging to estimate the rise in the rainwater harvested groundwater level during the year 2014. The modelling study led to generation of kriged estimate and kriged standard deviation spatial distribution maps in respect of pre-monsoon, post-monsoon and the fluctuation for the year 2014. The study revealed a mean rise of 2.77 m in the groundwater level owing to the rainwater harvesting. The rise in the groundwater level during the study period has led to an estimate of groundwater resource to 174,510,000 litres as compared to the consumption of 703,440,000 litres. The study estimated that about 80% of total volume of groundwater available is consumed and thereby maintaining a balance of about 20%. This figure of groundwater resource balance is expected to improve over the years with continued monitoring study of the fluctuating trend of the groundwater level with implementation of rainwater harvesting and artificial recharge in the campus. Similar rainwater harvesting study can be of use in other areas for assessing spatial and temporal phenomena leading to the usefulness of geostatistical modelling for sustainable development and management of groundwater resource.

10 Conclusion

Geostatistics is a branch of statistics focusing on spatial or spatio-temporal datasets. Developed originally to predict probability distributions of ore grades for mining operations, it is currently applied in diverse disciplines including petroleum geology, hydrogeology, hydrology, meteorology, oceanography, geochemistry, geography, forestry, environmental control, landscape ecology, soil science and agriculture (especially, in precision farming). Geostatistics is intimately related to interpolation methods, but extends far beyond simple interpolation problems in aquifer modelling. Geostatistical techniques rely on statistical model that is based on random function (or random variable) theory to model the uncertainty associated with spatial estimation. Empirical semi-variogram is used in geostatistics as a first estimate needed for spatial interpolation by kriging. Kriging is a group of geostatistical techniques to interpolate the value of a random field (e.g., the elevation of the bedrock as a function of the spatial location) at an unobserved location from observations of its value at nearby locations.

References

David, M. (1977). Geostatistical Ore Reserve Estimation. Elsevier, The Netherlands. 364 p.

Davis, J.C. (1986). Statistics and Data Analysis in Geology. Wiley, New York. 646 p.

Davis, B.M. and Borgman, L.E. (1979). A test of hypothesis concerning a proposed model for the underlying variogram. Proc. 16th APCOM, pp 163-181.

Isaak, M. and Srivastava, R.M. (1989). An Introduction to Applied Geostatistics. Oxford University Press, New York. 561 p.

Kitanidis, P.K. (1997). Introduction to Geostatistics: Applications in Hydrogeology. Cambridge University Press. 249 p.

Matheron, G. (1971). The Theory of Regionalized Variables and its Application. Paris School of Mines, Cah. Cent. Morphologie Math., 5. Fontainebleau.

Rouhani, S. (1986). Comparative Study of Ground Water Mapping Techniques. *J Ground Water*, **24(2):** 207-216.

Sarkar, B.C. (1988). An Integrated System for Geology-controlled Geostatistical Evaluation. Ph.D. thesis. Imperial College of Science and Technology, London. 218 p.