Statistical Techniques for Groundwater Data Analysis

VIJAY KUMAR
National Institute of Hydrology, Roorkee

INTRODUCTION

Groundwater is one of the most valuable natural resources, which supports human health, economic development and ecological diversity. Because of its several inherent qualities (e.g., consistent temperature, widespread and continuous availability, limited vulnerability, low development cost, drought reliability, etc.), it has become an immensely important and dependable source of water supplies in all climatic regions including both urban and rural areas of developed and developing countries.

The distribution of water on the earth is highly unbalanced. Nearly 97.41% of water is confined to the world's oceans and are unsuitable for human and livestock consumption. Major portion of the remaining 2.59% is locked up in the glaciers (1.953%) and beneath the surface as groundwater (0.614%), thus leaving only a meager 0.015% in the rivers and lakes for consumption by terrestrial inhabitants. This quantity of water was enough to support civilization. But, in recent years due to population explosion and urbanization, water resources from rivers, ponds and lakes have become inadequate. Therefore, an urgent need has arisen for exploring and managing surface as well as groundwater resources for continuous and dependable water supply for the growing needs of population. It is now well-recognized fact that water is finite and vulnerable resource, and it must be used efficiently and in an ecologically sound manner for present and future generations.

The evaluation, rational development and management of groundwater resources require a thorough knowledge of the subsurface environment and an understanding of the hydrological processes that governs the occurrence, movement and yield of groundwater. To do this various types of groundwater data is collected and stored in various forms. To be able to provide this information the first step is to obtain the information on the temporal and spatial characteristics of groundwater by having a network of observational stations. The basic data collected for different hydro-meteorological phenomenon through this observational network is called the observed or field data. Such observed data have to be processed to ensure its reliability. The processing of groundwater data is necessary to represent the data in more informative and useful form so that the data becomes meaningful to make some preliminary inferences and for further use.

Statistics deals with methods to draw inferences about the properties of a *population* based on sample data from that population. Population refers to a collection of objects. It can be finite or infinite, for example, the collection of all flow data of a river at a given site. Often, the measurements of the entire population are not available and what is generally available is a limited number of observations or a finite *sample*. Based on this sample, properties of the population are determined assuming these to be unbiased estimates of the properties of the population.

A variable whose value at any time is not influenced by the value at earlier time(s) is known as a *random variable*. Such a variable can be discrete which can take on only a finite set of values, such as number of rainy days in a year at a place. It can also be continuous and can take on any value, for example, the water level of river at a gauging site or the magnitude of rainfall at a place.

In many problems, the sample data consist of measurements on a single random variable; the techniques of analysis are called univariate analysis and estimation. Univariate analysis is carried out by using the measurements of the random variable, which is called sample information, to identify the statistical properties of the population from which the sample measurements are likely to have come. After the underlying population has been identified, one can make probabilistic statements about the future occurrences of the random variable, this represents univariate estimation.

Once the data is collected, checking of the data is required. The checking of the data can be done on different levels with increasing complexity. The different levels of data validation are according to the protocols: (a) field validation; (b) data entry validation; (c) primary validation (first checks); (d) secondary validations (integration checks); and (e) tertiary validations (advanced statistics).

STATISTICAL PARAMETERS

The important features of the groundwater data can be captured by a few characteristics of the data. It is known as summary statistics and includes measure of location, measure of spread and measure of shape. There are four principal moments for characterizing probability distributions:

(i) the central tendency or the value around which all other values are clustered,

(ii) the spread of the sample values around mean,

(iii) the asymmetry or skewness of the frequency distribution, and

(iv) the flatness of the frequency distribution.

These characteristics are expressed in terms of the parameters of distributions, the parameters can themselves be expressed in terms of moments.

The basic measure of location of data is some type of average value. Various measures like mode, median and arithmetic average exist.

Arithmetic Mean: If x_1, x_2 ... x_n represent a sequence of observations, the mean of this sequence is the ratio of the sum of values and the number of values:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

where $\bar{x}$ represents the sample mean; population mean is generally represented by μ. The mean is the average concentration of the sample, and as such, it provides a indication of the central tendency. Caution is warranted when using the mean as an indicator of central tendency when dealing with water quality: it is extremely sensitive to the presence of outliers and censored data. However, the mean also provides information about the total quantity of the contaminant present and, therefore, is included. The median value is the value above which and below which half of the sample population lies. It is not affected by outliers or by censored data. Consequently, the median is a reasonably good indicator of central tendency.

Variance: The measure of spread or dispersion about the mean is given by variance and the standard deviation. Standard deviation is often used instead of variance since its units are the same as the units of the variable being described. A small value of standard deviation indicates that observations are clustered tightly around a central value. On the other hand, a large standard deviation indicates that values are scattered widely about the mean and the tendency for central clustering is weak. It represents the dispersion of data about the mean and is expressed as:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2}$$

The unbiased estimate of population standard deviation (s) from the sample is given as the square root of the variance, i.e.,

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n}} \tag{3}$$

The coefficient of variation C_V is a dimensionless dispersion parameter and is equal to the ratio of the standard deviation and the mean:

$$C_V = s / \overline{x} \tag{4}$$

Coefficient of Skewness: The shape of the distribution is described by the coefficient of skewness and the coefficient of variation. Coefficient of skewness provides information on the symmetry while the coefficient of variation provides information on the length of the tail for certain type of distribution. It is a non-dimensional measure of the asymmetry of the distribution of the data. An unbiased estimate of the coefficient is given by:

$$C_S = \frac{n\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)(n-2)s^3} \tag{5}$$

Symmetrical frequency distributions have very small or negligible sample skewness coefficient C_s, while asymmetrical frequency distributions have either positive or negative coefficients. Often a small value of C_s indicates that the frequency distribution of the sample may be approximated by the normal distribution since $C_s = 0$ for this function.

Standard Errors of Sample Statistics

Because of the short length of most records, the statistics calculated from the sample are only estimates of the true or population values which would be available if very large samples were available. The reliability of the statistics calculated from the sample can be judged from the standard error of the estimate (SEE). According to the statistical theory, the probability that the true or population value of each statistic is within one standard error of estimate of the value calculated from the available data is about 68%.

The standard errors of mean, standard deviation and coefficient of skewness are respectively, given below:

$$S_e(\overline{x}) = s / \sqrt{n} \tag{6}$$

$$S_e(S) = s/\sqrt{2n} \tag{7}$$

$$S_e(C_s) = \sqrt{6n(n-1)/[(n+1)(n+2)(n+3)]} \tag{8}$$

Clearly, the standard error of estimate for each parameter becomes smaller as the length of record used in the analysis becomes even longer.

Test for means (t-Test)

The most common parametric test used to check whether or not two samples are from the same population is the *t*- test. The main assumptions of this test are: (i) the observations are independent, (ii) the observations are drawn from normally distributed populations, and (iii) these populations have the same variance. Hence this test is useful to determine whether the mean of the samples are significantly different from each other. Thus, the t-test indicates whether both the series belong to the same population or not. According to this test, the '*t*' statistic of the samples is determined by:

$$t = \frac{\left|\overline{X_1} - \overline{X_2}\right|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{9}$$

where, $\overline{X_1}$ and $\overline{X_2}$ are the arithmetic mean of the two samples of size *n*1 and *n*2 respectively; *S* is the unknown population standard deviation estimated from the samples variances *s*1 and *s*2 as:

$$S = \frac{(n_1 - 1)s_1 + (n_2 - 1)s_2}{n_1 + n_2 - 2} \tag{10}$$

If the statistic *t* is less than the tabulated value of Student's distribution at some chosen significance level α and *n*1+*n*2-1 degrees of freedom then the hypothesis that 'the means of both the samples are not significantly different' may be accepted at the chosen significance level.

Test for variances (F-test)

F-test is commonly used for testing whether or not the variances of two samples are significantly different. According to this test, the *F* statistic of the samples is determined as:

$$F = \frac{s_1^2}{s_2^2} \tag{11}$$

Where, S_1 and S_2 are sample variances. If the computed *F* is less than the tabulated value of *F* distribution at some chosen significance level α, and *n*1-1 and *n*2-1 degrees of freedom then the hypothesis that 'the variances of both the samples are not significantly different' may be accepted at the chosen significance level.

CORRELATION

Correlation attempts to measure the strength of relationship between two quantitative variables by means of a single number called Correlation coefficient(r). Correlation coefficient gives the measure of how the two variables X and Y vary together. Two variables are positively correlated if the large values of one variable tend to be associated with large values of the other variable, and similarly the smaller values of each variable. Two variables are negatively correlated if the large values of one variable tend to be associated with the smaller values of the other. The final possibility is that the variables are not related.

Correlation coefficient is the statistic that is most commonly used to summarize the relationship between two variables. It provides a measure of the linear relationship between two variables. It is actually a measure of how much close the observed values come to falling on a straight line. For a sample, It is given as

$$r_{X,Y} = \frac{Cov(x_i, y_i)}{\sigma_x \sigma_y} \tag{12}$$

where Cov(xi,yi) is the sample covariance between X and Y and σ_x and σ_y are the sample standard deviation of X and Y respectively i.e.

$$Cov(x_i, y_i) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n} \tag{13}$$

It can be shown that $-1 \leq r \geq +1$. If $r = +1$, then the scatter plot will be straight line with a positive slope; if $r = -1$, then the scatter plot will be a straight line with a negative slope. A value of $r = 0$ implies a lack of linearity and not necessarily a lack of association. For $|r| < 1$ the scatter plot appears as a cloud of points that becomes fatter and more diffuse as $|r|$ decreases from 1 to 0.

A serial correlation is the association between the successive terms in the same series x_i by lagging them suitably as per the requirement. Auto-correlation, which expresses the correlation between all pairs of observations x_i and x_{i+k} within the time series as a function of their spacing *k*, is defined as

$$\rho_k = \frac{COV(x_i, x_{i+k})}{\sigma_{x_i} . \sigma_{x_{i+k}}} \tag{14}$$

The distance *k* is known as the lag between x_i and x_{i+k}. This equation gives an idea of the dependence or association between the values of the same series. In practical applications the value of ρ_k are estimated in hydrology from observed samples with the estimate r_k as

$$r_k = \frac{\frac{1}{N-k} \sum_{i=1}^{N-k} x_i x_{i+k} - \frac{1}{(N-k)^2} \left(\sum_{i=1}^{N-k} x_i \right) \left(\sum_{i=1}^{N-k} x_{i+k} \right)}{\left[\frac{1}{N-k} \sum_{i=1}^{N-k} x_i^2 - \frac{1}{(N-k)^2} \left(\sum_{i=1}^{N-k} x_i \right)^2 \right]^{0.5} \left[\frac{1}{N-k} \sum_{i=1}^{N-k} x_{i+k}^2 - \frac{1}{(N-k)^2} \left(\sum_{i=1}^{N-k} x_{+ki} \right)^2 \right]^{0.5}} \tag{15}$$

where N is the length of data. For $k = 0$, the correlation coefficient $r_0 = 1$ and for other values of *k*, the value of r lies between ±1. A plot between the lag k against the serial correlation r_k is defined as correlagram. The significance of r_k can be tested at 95% confidence limit. The test value $(r_k)_t$ is computed from

$$(r_k)_t = \frac{-1 \pm 1.645\sqrt{N-k-1}}{N-k} \qquad (16)$$

A negative value of r_1 gives indication of marked high frequency (i.e. short-period) oscillations in the rainfall series. On the other hand, positive values indicate Markov linear type persistence (Mirza et al. 1998).

REGRESSION

It is a widely used approach to describe linear cause and effect relations between two variables. The objective is to predict a dependent variable based on an independent variable. The linear regression equation is:

$$y_i = a + bx_i + \in_i \; i = 1, 2, \ldots n \qquad (17)$$

where, y_i is the ith value of the dependent or regressed variable, x_i is the ith value of the independent or regressor variable. The regression line crosses the y-axis at a point a (the intercept), and has a slope b, and $\in_i$ is the random error term for the ith data point. The variables involved in regression should be chosen carefully and there should be a logical reason behind this choice. A scatter plot of y vs. x should be made to ascertain the dependence structure. Sometimes, a transformation of x, such as a power or log transformation, improves the regression relation.

Parameter Estimation

The regression coefficients (a and b) are estimated by minimizing the sum of squares of deviations of y_i from the regression line. For a point x_i, the corresponding $\hat{y}_i$ given by the regression equation will be:

$$\hat{y}_i = a + bx_i \qquad (18)$$

The residual error at this point is $e_i = y_i \, \hat{y}_i$ It provides a measure of how well the least-squares line conforms to the raw data. If the line passes exactly through each sample point, the error e_i would be zero. The sum of square of errors is:

$$S_{se} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (19)$$

Minimizing S_{se} leads to the following values of parameters:

$$b = S_{xy}/S_{xx}$$

and

$$a = \bar{y} - b\,\bar{x} \qquad (20)$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (21)$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Goodness of Regression

The goodness of regression is measured by the variability of the dependent variable that is explained by the regression relation. Some important indicators of the goodness of regression are:

Mean square error (mse): $$s^2 = \frac{S_{se}}{n-2} \tag{22}$$

Standard error of regression $$s = (\mathrm{mse})^{0.5} \tag{23}$$

Correlation coefficient $$r = S_{xy} / (S_{xx}\, S_{yy})^{0.5} \tag{24}$$

Coefficient of determination $$R^2 = 1 - \frac{S_{se}}{S_{yy}} \tag{25}$$

where

$$S_{se} = \sum \left(y_i - \hat{y}_i \right)^2 = \text{Error sum of squares}$$

$$S_{yy} = \sum \left(y_i - \bar{y} \right)^2 = \text{Total sum of squares}$$

The coefficient of determination R^2 (or sometimes r2) is one of the measure of how well the least squares equation performs as a predictor of y. The coefficient of determination represents the fraction of variance that is explained by regression. Essentially, R^2 tells us how much better we can do in predicting y by using the model and computing ˆy than by just using the mean ¯y as a predictor. R^2 takes on values between 0 and 1. The closer this ratio is to unity, the 'better' is the regression relation. S_{yy} measures the deviations of the observations from their mean and S_{se} measures the deviations of observations from their predicted values.

Inferences on Regression Coefficients

The variances of coefficients *a* and *b* are needed to determine their confidence bands. From eq. (20),

$$b = \sum (x_i - \bar{x})(y_i - \bar{y}) / S_{xx} = \sum y_i (x_i - \bar{x}) / S_{xx} = S_{xy}/S_{xx} \tag{26}$$

So,

$$\mathrm{var}(b) = \sum (x_i - \bar{x})^2 \,\mathrm{var}(y_i) / S_{xx}^2 = S_{xx} s^2 / S_{xx}^2$$

Hence $\sigma_b^2 = s^2 / S_{xx}$

or $\sigma_b = s/\surd S_{xx}$

Thus, the standard error of *b*, $S_b = s / \surd S_{xx}$. (27)

The variance of coefficient *a*

$$\mathrm{var}(a) = \mathrm{var}(\bar{y} - b\bar{x}) = \mathrm{var}(\bar{y}) - \bar{x}^2\ \mathrm{var}(b)$$
$$= s^2 / n + s^2 \bar{x} / \mathrm{S_{xx}}$$

So, the standard error of *a*

$$S_a = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \tag{28}$$

Test of hypothesis concerning *a* and *b*

Hypothesis $\mathrm{H_0}$: $a = a_0$ versus $\mathrm{H_a}$: $a \neq a_0$ is tested by computing $t = (\mathrm{a} - \mathrm{a_0})/\mathrm{S_a}$ which has a *t* distribution with *n-2* degrees of freedom. $\mathrm{H_0}$ is rejected if $|t| > t_{(1-\alpha/2),\,(n-2)}$.

Hypothesis $\mathrm{H_0}$: $b = b_0$ versus $\mathrm{H_b}$: $b \neq b_0$ is tested by computing $t = (\mathrm{b} - \mathrm{b_0})/\mathrm{S_b}$. $\mathrm{H_0}$ is rejected if $|t| > t_{(1-\alpha/2),(n-2)}$.

Hypothesis $\mathrm{H_0}$: $b = 0$ is tested by computing $t = (b - 0)/S_b$. $\mathrm{H_0}$ is rejected if $|t| > t_{(1-\alpha/2),\,(n-2)}$ and in this case the regression equation is able to explain a significant amount of variation in *y*.

Confidence Intervals

The confidence interval at the α% significance level indicates that in repeated applications of the technique, the frequency with which the confidence interval would contain the true parameter value is (100 - α)%. A typical value of α is 0.05 which corresponds to (1-0.05)*100% = 95% confidence limits. These intervals are defined if the true relationship between the variables is linear and the residuals e_i are independent, normally distributed random variables with constant variance.

If the model is correct, then a/S_a and b/S_b should follow *t* distribution with *(n-2)* degrees of freedom. Hence, for coefficient *a*, the lower and upper limits are:

$$(\mathrm{l_a},\ \mathrm{u_a}) = \{\mathrm{a} - \mathrm{t}_{(1-\alpha/2),(n-2)}\ \mathrm{S_a},\ \ \mathrm{a} + \mathrm{t}_{(1-\alpha/2),(n-2)}\ \mathrm{S_a}\} \tag{29}$$

For coefficient *b*, the lower and upper limits are:

$$(\mathrm{l_b},\ \mathrm{u_b}) = \{\mathrm{b} - \mathrm{t}_{(1-\alpha/2),(n-2)}\ \mathrm{S_b}\ ,\ \mathrm{b} + \mathrm{t}_{(1-\alpha/2),(n-2)}\ \mathrm{S_b}\} \tag{30}$$

where $t_{(1-\alpha/2),\,(n-2)}$ represents Student's *t* values corresponding to the probability of exceedance α/2 and *(n-2)* degrees of freedom.

Confidence Intervals on Regression Line

These depend on the variance of $\hat{\bar{y}}_k$ which is the predicted mean value of $\hat{y}_k$ for a given x_k:

$$\hat{\bar{y}}_k = a + bx_k \tag{31}$$

Then,

$$\operatorname{var}(\hat{\bar{y}}_k) = \operatorname{var}(a) + x_k^2 \operatorname{var}(b) + 2x_k \operatorname{cov}(a,b)$$
$$= s^2\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}\right] \qquad (32)$$

Hence, the standard error of $\hat{\bar{y}}_k$ would be

$$S_{\hat{\bar{y}}_k} = s\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}\right]^{1/2} \qquad (33)$$

So, the lower and upper confidence limits on the regression line are:

$$(L,U) = [\hat{\bar{y}}_k - S_{\hat{\bar{y}}_k} t_{(1-\alpha/2),(n-2)},\ \hat{\bar{y}}_k + S_{\hat{\bar{y}}_k} t_{(1-\alpha/2),(n-2)}] \qquad (34)$$

Confidence intervals on an individual predicted value of y are:

$$S'_{\hat{y}_k} = S\left[1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]^{1/2} \qquad (35)$$

Many a times a dependent variable is dependent on several other quantities. The equation in this case is of the form of

$$Y = a + b*X_1 + c*X2 + \text{...........} \qquad (36)$$

where,

Y = Dependent variable

X_1, X_2 = Independent variables

a, b, c = Coefficients of the polynomial regression

The regression coefficients in the above equation are also determined by the least square procedure. This type of regression is known as multiple regression. The above equation can be written in the matrix form as below

$$[\,Y\,] = [\,X\,][\,B\,] \qquad (37)$$

where,

[Y] = column matrix of dependent variable of order N*1

[B] = column matrix of coefficients to be calculated M*1

[X] = square matrix of independent variables of order n*m

N = no. of data points

M = no. of coefficients to be calculated

The general solution of the above matrix is given by

$$[\,B\,] = [\,X^T X\,]^{-1}[\,X^T Y\,] \qquad (38)$$

TREND ANALYSIS

Statistical analysis of water table data is also carried out to determine any underlying trends in the data i.e. to determine whether the groundwater is rising of falling in a particular aquifer. Trend indicates a long term growth or decline in the time series of groundwater owing to man's activities. The man's activities are related to artificial discharge and artificial recharge. Artificial discharge usually refers to pumping whereas artificial recharge refers to introduction of water into ground by wells, pits, excavations or by irrigation. The trends present in the data may be in general approximated by the polynomial equation of the form

$$Y = a + b*t \tag{39}$$

where,

Y = observed data; t = time; a, b = coefficients of the regression.

The coefficient *b* (slope) represents the rate of increase or decrease of the variable.

SPATIAL INTERPOLATION

The sustainable management of groundwater resources needs quantitative information on its behaviour in space and time. As groundwater use has increased, issues associated with the quality of groundwater resources have likewise grown in importance. The groundwater data sets typically contain many variables measured at several spatially scattered locations. Knowledge of spatial variability of groundwater data is essential for making reliable groundwater interpretations and for making accurate predictions of groundwater data at any particular location in the aquifer. With the availability of distributed hydrological models, which can handle large volume of data, the spatial information of hydrological data at a grid pattern is not only useful but is necessary to get reliable results. Various methods, both simple and complicated, are available and are in use to get the information about any parameter on a pre-specified grid when the parameter is measured at random points in the field. Spatial interpolation is the process of using points with known values to estimate values at other points. GIS can provide effective spatial analysis capabilities required to use various data in modeling studies. The various methods of interpolation are discussed below.

Regression Model

A global interpolation method, regression model relates a dependent variable to a number of independent variables in a linear equation (an interpolator), which can then be used for prediction or estimation. The trend surface analysis, a type of regression model, approximates points with known values with a polynomial equation.

The equation or the interpolator can then be used to estimate values at other points. A linear or first-order trend surface uses the equation:

$$z_{xy} = b_0 + b_1 x + b_2 y \tag{40}$$

where, the attribute value z is a function of x and y coordinates. The b coefficients are estimated from the known points. Higher-order trend surface models are required to approximate more complex surfaces. A cubic or a third-order model, for example, includes hills and valleys. A cubic trend surface is based on the equation:

$$z_{xy} = b_0 + b_1x + b_2y + b_3x^2 + b_4y^2 + b_5xy + b_6x^3 + b_7y^3 + b_8x^2y + b_9xy^2 \tag{41}$$

A third-order trend surface required estimation of 10 coefficients (i.e., b_i), compared to three coefficients for a first-order and six coefficients for second-order surface. A higher order trend surface model therefore required more computation than a lower-order model does. In GIS (ILWIS software) this method is known as trend surface method. This method calculates pixel values by fitting one surface through all point values in the map. The surface may be of the first order up to the sixth order. A trend surface may give a general impression of the data. Surface fitting is performed by a least squares fit.

Thiessen Polygon

Thiessen polygons assume that any point within a polygon is closer to the polygon's known point than any other known points. Thiessen polygons do not use an interpolator but require initial triangulation for connecting known points. Each known point is connected to its nearest neighbors. Thiessen polygons are easily constructed by connecting lines drawn perpendicular to the sides of each triangle at their midpoints. In GIS (ILWIS software) this method is known as nearest point method. This method assigns to pixels the value, identifier or class name of the nearest point, according to Euclidean distance. This method is also called Nearest Neighbour.

Inverse Distance Weighted Interpolation

Inverse distance weighted (IDW) interpolation is an exact method that enforces that the estimated value of a point is influenced more by nearby known points than those farther away. The general equation for the IDW method is:

$$Z_0 = \frac{\sum_{i=1}^{n} Z_i \frac{1}{d_i^k}}{\sum_{i=1}^{n} \frac{1}{d_i^k}} \tag{42}$$

where Z_0 is the estimated value at point 0, Z_i is the Z value at known point *i*, d_i is the distance between point *i* and point 0, n is the number of known points used in estimation, and *k* is the specified power. The power *k* controls the degree of local influence. A power of 1.0 means a constant rate of change in value between points (linear interpolation). A power of 2.0 or higher suggests that the rate of change in values is higher near a known point and levels off away from it. The degree of local influence also depends on the number of known points used in estimation. An important characteristic of IDW interpolation is that all predicted values are within the range of maximum and minimum values of the known points. In GIS (ILWIS software) this method is known as Moving Average method.

Kriging

Kriging is an alternative to many other point interpolation techniques. Unlike straightforward methods discussed above, Kriging is based on a statistical method. Kriging is the only interpolation method available in ILWIS that gives you an interpolated map and output error map with the standard errors of the estimates. Kriging is such a technique, which takes into consideration the spatial structure of the parameter and so scores over the other methods. Kriging is based on the Theory of Regionalized Variables. When a variable is distributed in space, it is said to be "regionalized". All the parameters generally used in groundwater hydrology, such as transmissivity, hydraulic conductivity, piezometric heads, vertical recharge etc. can be called regionalized variables.

Semivariogram

Kriging uses the semivariogram to measure the spatial correlated components, a component that is also called spatial dependence or spatial autocorrelation. The semivariogram is half of the arithmetic mean of the squared difference between two experimental measures, ($Z(x_i)$ and $Z(x_i+h)$), at any two points separated by the vector h. Before values of any parameter can be estimated with kriging, it is necessary to identify the spatial correlation structure from the semivariogram, which shows the relationship between semivariance and the distance between sample pairs.

$$\gamma^*(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2 \qquad (43)$$

where $\gamma^*(h)$ = estimated value of the semivariance for lag h; N(h) is the number of experimental pairs separated by vector h; $z(x_i)$ and $z(x_i + h)$ = values of variable z at x_i and x_i+h, respectively; x_i and x_i+h = position in two dimensions. A plot of $\gamma^*(h)$ versus the corresponding value of h, also called the semivariogram, is thus a function of the vector h, and may depend on both the magnitude and the direction of h. A sample plot of semivariogram is shown in Fig. 1.

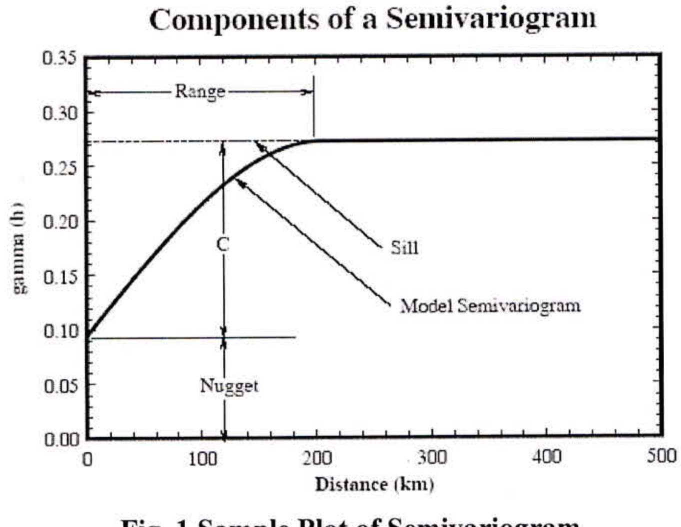

Fig. 1 Sample Plot of Semivariogram

The distance at which the variogram becomes constant is called the range, *a*. It is considered that any data value Z(x) will be correlated with any other value falling within a radius, *a* and thus range corresponds to the zone of influence of the RV. The value of the semivariogram at a distance equal to the range is called the sill. Semivariograms may also increase continuously without showing a definite range and sill. The value of the semivariogram at extremely small separation distance is called the nugget effect.

Structural Analysis

The observed data is used to calculate the experimental semivariogram. A mathematical function used to approximately represent this semivariogram is known as the theoretical semivariogram. Some of the theoretical semivariogram models are (Fig. 2).

Spherical model: -
$$\gamma(h)=\begin{cases} C_0[1-\delta(h)]+C\left[\dfrac{3}{2}\dfrac{h}{a}-\dfrac{1}{2}\dfrac{h^3}{a^3}\right] & h\le a \\ C_0+C & h>a \end{cases} \tag{44}$$

Exponential model: -
$$\gamma(h)=C_0[1-\delta(h)]+C\left[1-\exp\left(-\frac{h}{a}\right)\right] \tag{45}$$

Gaussian model: -
$$\gamma(h)=C_0[1-\delta(h)]+C\left[1-\exp\left(-\frac{h^2}{a^2}\right)\right] \tag{46}$$

Linear model: -
$$\gamma(h)=C_0[1-\delta(h)]+bh \tag{47}$$

where, δ(h) is the Kronecker delta = $\begin{cases} 1 & h=0 \\ 0 & h\neq 0 \end{cases}$, C_0 is the Nugget effect, C_0+C is the Sill, *a* is the Range and *b* is the slope.

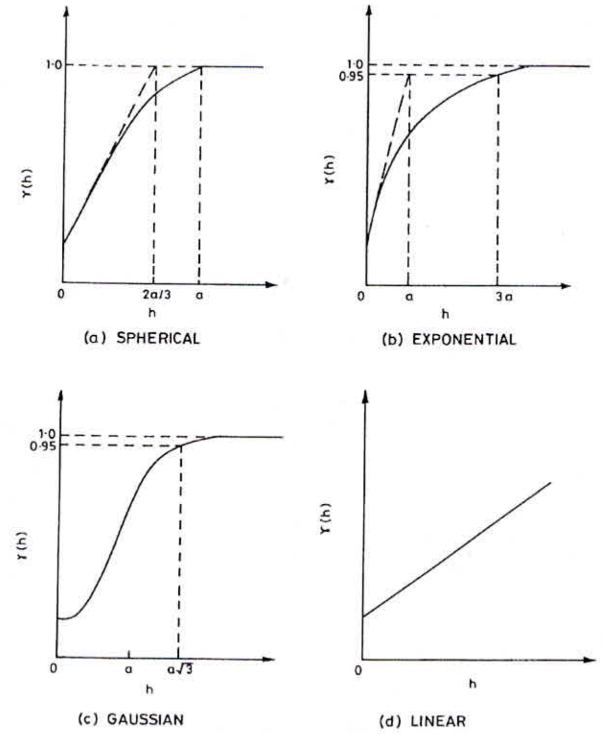

Fig. 2 Theoretical models of semivariogram

Kriging

Kriging is a geostatistical method for spatial interpolation. It is a technique of making optimal, unbiased estimates of regionalized variables at unsampled locations using the structural properties of the semivariogram and the initial set of data values. Consider a situation in which a property is measured at a number of points, x_i , within a region to give values of $z(x_i)$, i=1,2,3,...,N. (x_i is the coordinate of the observation point in 1, 2 or 3-dimensional space). From these observations, the value of the property at any place x_0 can be estimated as

$$z^*(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i) \qquad \text{i=1,2,3,.......,N} \tag{48}$$

where,

$z^*(x_0)$ = estimated value at x_0

λ_i = weights chosen so as to satisfy suitable statistical conditions

$z(x_i)$ = observed values at points x_i, i= 1,2,3,.......,N

N = sample size

In kriging, the weights λ_i are calculated so that $Z^*(x_0)$ is unbiased and optimal i.e.

$$E\{Z^*(x_0) - Z(x_0)\} = 0 \tag{49}$$

$$\text{Var}\{Z^*(x_0) - Z(x_0)\} = \text{minimum} \tag{50}$$

The best linear unbiased estimate of $Z(x_o)$ is obtained by using Lagrangian techniques to minimise Eq. (50) and then optimising the solution of the resulting system of equations when constrained by the nonbiased condition of Eq. (20). The following system of equations, known as kriging system, results from the optimization:

$$\begin{cases} \sum_{j=1}^{N} \lambda_j \gamma(x_i, x_j) + \mu = \gamma(x_i, x_0) & i = 1,2,3,......., N \\ \sum_{j=1}^{N} \lambda_j = 1 \end{cases} \tag{51}$$

where,

μ = Lagrange multiplier

$\gamma(x_i, x_j)$ = semivariogram between two points x_i and x_j

Solution of the above set provides the values of λ_i, which can be used with Eq. 19 for estimation. The minimum estimation variance, or kriging variance, is written as:

$$\sigma_k^2(x_0) = \sum_{i=1}^{N} \lambda_i \gamma(x_i, x_0) + \mu \tag{52}$$