12 Applying Geostatistics: Basic Knowledge and Variographic Analysis with Exercises

P.S.N. Murthy, Tanvi Arora[1] and Shakeel Ahmed[1]

Ex-Head, Geostatistics Centre & Gen. Manager, NMDC Ltd., Hyderabad, India
(Presently) Adviser, Essar Ltd., Visakhapatnam, India
[1]National Geophysical Research Institute, Hyderabad-500007, India

INTRODUCTION

Any scientific study is based on a number of information on the system or processes through measurement of parameters defining one or more properties of the physical system. In most cases these parameters change their values in space and sometimes in time also and hence they are called variables. Rainfall, effective recharge, thickness of an aquifer, hydraulic head, transmissivity, permeability, storage coefficient, etc. are the examples in hydrogeology. Although, these parameters are often highly variables but this spatial variability is not purely random and, according to Matheron (1963), if measurements are made at two different locations, the closer the measurement points are to each other, the closer the measured values. These variables are given the name of Regionalized Variables (Re. V).

Although the characterization and estimation of a regionalized variable (Re.V.) can be made on a purely deterministic basis, it is more convenient and usual to introduce geostatistics in a probabilistic framework, bearing in mind that this artifact is only a tool for performing an estimation (Marsily, 1986).

In this article attempt has been made to include a few basic aspects that are often needed by a hydrogeologist to practise geostatistics and for others it will be a refreshing starting from basic definition and formulae of statistics, part of mathematics and algebra that are used.

BASIC STATISTICS AND MATHEMATICS USED

Data speak most clearly when they are organized. Much of statistics deals with the organization, presentation and summary of data.

Frequency Distribution

One of the most common and useful presentation of data sets is the frequency table and the histogram. A frequency table records how often observed values fall within certain intervals or classes. A typical data set is given in Table 1 (Isaaks and Srivastava, 1989).

Table 1: Values of a single variate data set observed on a regular grid.

81	77	103	112	123	19	40	111	114	120
•	•	•	•	•	•	•	•	•	•
82	61	110	121	119	77	52	111	117	124
•	•	•	•	•	•	•	•	•	•
82	74	97	105	112	91	73	115	118	129
•	•	•	•	•	•	•	•	•	•
88	70	103	111	122	64	84	105	113	123
•	•	•	•	•	•	•	•	•	•
89	88	94	110	116	108	73	107	118	127
•	•	•	•	•	•	•	•	•	•
77	82	86	101	109	113	79	102	120	121
•	•	•	•	•	•	•	•	•	•
74	80	85	90	97	101	96	72	128	130
•	•	•	•	•	•	•	•	•	•
75	80	83	87	94	99	95	48	139	145
•	•	•	•	•	•	•	•	•	•
77	84	74	108	121	143	91	52	136	144
•	•	•	•	•	•	•	•	•	•
87	100	47	111	124	109	0	98	134	144
•	•	•	•	•	•	•	•	•	•

The frequency table of the data set from Table 1 is given in Table 2.

Table 2 also shows, in its last column, the cumulative frequency distribution. For many Earth Science applications, such as pollution studies, the cumulative frequency above a lower limit is of more interest. Rather than recording the number of values within certain class, we record the total number of values below certain cut off. The second and third columns in the Table 2 can be represented graphically and they are called histogram and cumulative histogram respectively. The percent frequency and cumulative percent frequency forms are used interchangeably, since one can be obtained from the other. The frequencies or the cumulative frequencies can be expressed in percentage. Incidentally, in the above example, since the total number of values is 100, the frequency and its percent are same. If a curve is plotted taking y-axis as the frequency, it is known as Probability Plot or Probability Distribution Function (PDF). A number of distribution function based on the shape of the curve are available. However, among the most common are the Gaussian/Normal and Log-normal distribution. Some of the estimation tools work better if the distribution of data values is close to a Gaussian or normal distribution.

Table 2: Frequency and cumulative frequency table

Class of values	*Number of observations/ Frequency*	*Cumulative Frequency*
0-10	1	1
10-20	1	2
20-30	0	2
30-40	0	2
40-50	3	5
50-60	2	7
60-70	2	9
70-80	13	22
80-90	16	38
90-100	11	49
100-110	13	62
110-120	17	79
120-130	13	92
130-140	4	96
140-∞	4	100

Summary Statistics

The centre of the distribution can be defined by mean, median and mode. The measure of spread can be defined as variance and standard deviation. The shape of the distribution is described by the coefficient of skewness and coefficient of variation. Taken together, these statistics provide feel of the data.

The mean, m, is the arithmetic average of the data values:

$$m = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{1}$$

The number of data is n and z_1, ... z_n are the data values. The mean of the above presented 100 values is 97.55.

The median, *M*, is the midpoint of the observed values if they are arranged in increasing order. The median can easily be read from a probability plot. The mode is the value that occurs most frequently.

The variance, σ^2, is the average squared difference of the observed values from their mean and is given by:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - m)^2 \tag{2}$$

The standard deviation, σ, is simply the square root of the variance. This shows the spread of the values from the mean. The coefficient of skewness showing the symmetry of the distribution is defined as:

$$\text{Coeff. of Skewness} = \frac{\frac{1}{n}\sum_{i-1}^{n}(z_i - m)^3}{\sigma^3} \tag{3}$$

The coefficient of variation (CV) is a statistics that is often used as an alternative to skewness to describe the shape of the distribution. It is defined as the ratio of the standard deviation to the mean:

$$\text{CV} = \frac{\sigma}{m} \tag{4}$$

Some Remarks

The mean is quite sensitive to erratic extreme values. If the 145 ppm value in the above data set had been 1450 ppm, the mean would change to 110.6 ppm. The median, however, would be unaffected by this change because it depends only how many values are above and below it.

The variance, since it involves squared differences, is also quite sensitive to the extreme values. The variance of the 100 values is 688. Standard deviation is often used instead of the variance since its units are the same as the units of the variable.

The coefficient of skewness suffers even more than the mean and variance from sensitivity to erratic extreme values. A coefficient of variation greater than one indicates the presence of some erratic high sample values that may have a significant impact on the final estimates. The CV for the above data is 0.269, which reflects the fact that the histogram does not have a long tail of high values.

Bivariate Distribution

In the field of hydrogeology, due to scarcity of data, we always prefer to work on multivariable information. A typical data set for a two variable system is given in Table 3.

In the data set in Table 3, values of two variables are available at all the 100 locations. Values of the first variable are written above the circle (denoting the locations of measurement) and the values of the second variable are written below the circle in Italics. The basic statistics of the two sets of data are given in Table 4 for comparison.

Table 3: Values of a bi-variate data set observed on a regular grid

81 • *15*	77 • *12*	103 • *24*	112 • *27*	123 • *30*	19 • *0*	40 • *2*	111 • *18*	114 • *18*	120 • *18*
82 • *16*	61 • *7*	110 • *34*	121 • *36*	119 • *29*	77 • *7*	52 • *4*	111 • *18*	117 • *18*	124 • *20*
82 • *16*	74 • *9*	97 • *22*	105 • *24*	112 • *25*	91 • *10*	73 • *7*	115 • *19*	118 • *19*	129 • *22*
88 • *21*	70 • *8*	103 • *27*	111 • *27*	122 • *32*	64 • *4*	84 • *10*	105 • *15*	113 • *17*	123 • *19*
89 • *1*	88 • *18*	94 • *20*	110 • *27*	116 • *29*	108 • *19*	73 • *7*	107 • *16*	118 • *19*	127 • *22*
77 • *15*	82 • *16*	86 • *16*	101 • *23*	109 • *24*	113 • *25*	79 • *7*	102 • *15*	120 • *21*	121 • *20*
74 • *14*	80 • *15*	85 • *15*	90 • *16*	97 • *17*	101 • *18*	96 • *14*	72 • *6*	128 • *28*	130 • *25*
75 • *14*	80 • *15*	83 • *15*	87 • *15*	94 • *16*	99 • *17*	95 • *13*	48 • *2*	139 • *40*	145 • *38*
77 • *16*	84 • *17*	74 • *11*	108 • *29*	121 • *37*	143 • *55*	91 • *11*	52 • *3*	136 • *34*	144 • *35*
87 • *22*	100 • *28*	47 • *4*	111 • *32*	124 • *38*	109 • *20*	0 • *0*	98 • *14*	134 • *31*	144 • *34*

Table 4: Various statistics of the two data sets

Statistics	*Symbol*	*First variable V*	*Second variable U*
Population	n	100	100
Mean	m	97.6	19.1
Median	M	100.5	18.0
Mode		111.0	15.0
Standard Deviation	σ	26.2	9.81
Minimum	min	0.0	0.0
Maximum	max	145.0	55.0
Coeff. of Variation	CV	0.27	0.51

REFRESHING PRELIMINARY ALGEBRA AND MATHEMATICS

A few equations and expressions are reproduced below for refreshing and their applications in later part of the theory.

$$(x+y)^2 = x^2 + y^2 + 2xy \tag{5}$$

Similarly $$(ax+by)^2 = a^2x^2 + b^2y^2 + 2abxy \tag{6}$$

Let us substitute x_1 for x and x_2 for y in the equation (5).

$$(x_1+x_2)^2 = x_1^2 + x_2^2 + 2x_1x_2 \tag{7}$$

Let us substitute λ_1, λ_2, x_1, x_2 for a, b, x and y respectively in the equation (6).

Then $$(\lambda_1x_1 + \lambda_2x_2)^2 = \lambda_1^2x_1^2 + 2\lambda_1\lambda_2x_1x_2 + \lambda_2^2x_2^2 \tag{8}$$

$(\lambda_1x_1 + \lambda_2x_2)$ can be written in the form of = $\sum_{i=1}^{2}\lambda_ix_i$ (9)

Then $$(\lambda_1x_1 + \lambda_2x_2)^2 = \lambda_1^2x_1^2 + \lambda_2^2x_2^2 + 2\lambda_1\lambda_2x_1x_2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\lambda_i\lambda_jx_ix_j \tag{10}$$

This is because in the equation (10), when $i = j = 1$ the resulting term is $\lambda_1^2 x_1^2$; when $i = j = 2$ the term is $\lambda_2^2 x_2^2$; and when $i = 1, j = 2$ the term is $\lambda_1\lambda_2x_1x_2$. Similarly when $i = 2$, $j = 1$ the term $\lambda_2\lambda_1x_1x_2$ results.

Therefore, equation (10) can also be written as

$$\left(\sum_i \lambda_ix_i\right)^2 = \sum_{j=i}\lambda_i\lambda_jx_ix_j + \sum_{i\neq j}\lambda_i\lambda_jx_ix_j \tag{11}$$

The above formula can be simplified and written as

$$\left(\sum_i \lambda_ix_i\right)^2 = \sum_i\sum_j \lambda_i\lambda_jx_ix_j \tag{12}$$

This is an important derivation. Generally one is prone to confusion as to how the terms with j have come, when it is not there originally. This formula has got extensive application in statistics and geostatistics.

PROBLEM OF CALCULATING MEAN AND CONCEPT OF VARIOGRAM

Let us summarize again the most common statistics and their forms used in the development of geostatistical methods.

Expectation

If we can calculate the expected value of a distribution, it is known as Expectation and is denoted by *E*. It is also known as the first moment of the variable. Practically, one calculates the expected value of a variable using arithmetic mean keeping in mind that the number of values available can be approximated by ∝. However, we write an expected value of *Z* as follows:

$$m_z = E[z] \quad \forall z \tag{13}$$

Variance

The variance of the variable *Z* can be written in the following form:

$$\sigma_z^2 = E[(z_i - m_z)^2] \tag{14}$$

It is important to note that the variance in the general case is the variance of many realization available at a single location. However, the variance given by the equation (14) is the spatial variance of the variable and it does not depend on the location *i*.

Covariance

Similarly an expression for covariance of a variable *Z* between two points *i* and *j* in space can be given as:

$$C_{ij} = E[(z_i - m_x)(z_j - m_x)] \tag{15}$$

By definition the expected value m_z should not change by adding or removing one or a few values of the variable from the data. However, since we simply calculate an arithmetic mean, it will not be possible to ensure this. This means that a true mean or a true expected value can never be calculated from the data. Thus we can neither calculate a correct variance nor a correct covariance. We have to then introduce another hypothesis known as Intrinsic.

If we work on another variable *y* defined as:

$$y = z_i - z_j \tag{16}$$

i.e. the first order difference of the primary variable *Z*, then the expected value of the variable *y* can be shown to be 0 provided the variable *Z* has an unknown but constant mean m_z. Since the expected value of *Y* is 0, we need not calculate it from the data. If we calculate the variance of this new variable *Y*, we find the following expression.

$$\text{Var}\ (y) = E[(z_i - z_j)^2] \tag{17}$$

Though it is the variance of the new variable *Y*, it is a new function for the variable *Z* that does not depend on the location of points nor is affected by the value of m_z. This new function is called variogram of the variable *Z* between two points in space *i* and *j* and is denoted by

$$2\gamma_{ij} = E[(z_i - z_j)^2] \tag{18}$$

The variogram depends only on the separation vector between the two points *i* and *j* and not on the location of *i* and *j*. Here *Z* is called Intrinsic and more precisely Intrinsic Random Function (IRF) of order zero because its first order difference has become stationary. Similarly if $(k + 1)^{th}$ order difference of a variable becomes stationary, the variable is called IRF-k, i.e. Intrinsic Random Function of order k. Most of the hydrogeological variables could be categorized as IRF-0 from the practical point of view. We will show in the subsequent text that the complete estimation theory can be developed using variograms and thus there is no need to calculate the mean.

STRUCTURAL ANALYSIS AND COMPUTATION OF COVARIANCE AND VARIOGRAM

As both the covariance and the variogram are the function of separation vector between two points i and j, values of separation vectors e.g., h_1, h_2 etc. are decided first such that

$$h = \left| x_i - x_j \right| \tag{19}$$

Depending upon the value of h, the data are grouped into pairs and some function as defined below are averaged to obtain a covariance (C_{ij}) or a variogram (γ_{ij}).

$$C^z(h) = \frac{1}{N_h}\sum_{i=1}^{N_h}\{z(x_i) - m_z\}\{z(x_i + h) - m_z\} \tag{20}$$

$$\gamma^z(h) = \frac{1}{2N_h}\sum_{i=1}^{N_h}\{z(x_i) - z(x_i + h)\}^2 \tag{21}$$

N_{h} is the number of pairs for a given h. It is usual to write $C(h)$ or C_{ij} or $C(x_{\mathrm{i}}, x_{\mathrm{j}})$ etc. and correspondingly for the variogram also. If we consider equation (18), the expression in the equation (21) should be called semi-variogram. However, in practice only semi-variograms are used; hence for convenience we call it simply a variogram. It is now clear that in practice, we cannot calculate a true mean. Thus we prefer to work on variogram rather than covariance.

Properties of Covariance and Variogram

A relation between a variogram and a covariance can be established as follows if both the functions exist (i.e., the mean exist or the variable is stationary).

$$\gamma(h) = C(0) - C(h) \tag{22}$$

Here $C(0)$ is equal to the variance of the variable. Thus in case of stationary variables one function can be calculated from the other.

Behaviour of the Variogram for Large Separation Distance, *h*

As the distance h increases, $\gamma(h)$ increases. However, after a large distance, $\gamma(h)$ stabilizes around a constant value except sometimes when it increases continuously. In case the variogram increases continuously with h, it is known as unbounded variogram and then no covariance exist. But if it stabilizes around a constant value (known as sill), it is called bounded variogram and both covariance and variogram exist [cf. equation (22)]. Unbounded variogram also shows the nonstationary nature of the variable.

Behaviour Close to Origin

Theoretically, $\gamma(h) = 0$ if $h = 0$ regardless of the type of variogram. However, often variograms exhibit a jump at the origin. This apparent jump is called nugget effect.

Anisotropy in the Variogram

Variogram can be calculated in different directions by taking the separation vector in particular direction. If variograms in different directions are different then the parameter is anisotropic but if variograms calculated in all possible directions are more or less same, the parameter can be considered as isotropic and so the variograms.

Parameters of the Variogram

A variogram function can be defined essentially by sill and range: sill is the constant value on the y-axis around which a variogram stabilizes after a large distance and range is the value at x-axis at which the variogram becomes constant or nearly constant. The sill value is usually very close to the variance of the variable. In addition, the sudden apparent jump near the origin that occurs in some cases is known as nugget effect. Also, the shape of the variogram between origin and the point of stabilization is different in different variables, which purely depends on its nature of variability. Depending on this shape, variograms are categorized into different type of variogram viz., *Linear, Spherical, Exponential, Gaussian, Cubic* etc.

Modelling a Theoretical Variogram

The variogram calculated from the field data is an erratic curve and known as an experimental variogram. It is not possible to use this variogram in the estimation purpose due to various reasons discussed later. Therefore, the curve of the experimental variogram is approximated by another theoretical curve with a defined mathematical expression. This smooth curve fitted to the experimental variogram is known as theoretical variogram. This fitting or modelling is performed in several ways mostly visual or using some form of difference between the two variograms but on a trial and error basis. Sometimes an automatic modelling has also been proposed but is not proved to be very useful. Mathematical functions for the common variogram type are given below. In all the expressions, c is the sill, a the range and h is the separation vector.

Model in h^2 including Linear Model

$$\gamma(h) = ch^{\lambda} \qquad \lambda<2 \qquad (23)$$

If $\lambda = 1$ the variogram model is called Linear.

Spherical model

$$c\left[\frac{3}{2}\left(\frac{h}{a}\right)-\frac{1}{2}\left(\frac{h}{a}\right)^{3}\right] \qquad h \le a \qquad (24)$$

$$\mathrm{c} \qquad h > a \qquad (25)$$

Exponential Model

$$c\left[1-\exp\left\{-\frac{h}{a}\right\}\right] \qquad (26)$$

Gaussian Model

$$c\left[1-\exp\left\{-\left(\frac{h^2}{a^2}\right)\right\}\right] \qquad (27)$$

Cubic Model

$$c\left[7\left(\frac{h}{a}\right)-8.75\left(\frac{h}{a}\right)^{3}+3.5\left(\frac{h}{a}\right)^{s}-0.75\left(\frac{h}{a}\right)^{7}\right] h \le a \qquad (28)$$

$$c \qquad h > a \qquad (29)$$

Difference between the approach of statistics and geostatistics

Statistics	Geostatistics
Statistics considers all the numerical values *Z*(*x*), *Z*(*x*') as independent realization of the same numerical function *Z*.	Geostatistics considers all the numerical values *z*(*x*), *z*(*x*') as the particular realizations of random variables *Z*(*x*) and *Z*(*x*')
In other words it does not take into account the spatial auto-correlation between two neighbouring values *Z*(*x*) and *Z*(*x* + *h*)	In other words it takes into account the spatial auto-correlation between two neighbouring values *z*(*x*) and *z*(*x* + *h*).
Statistics does not take the regionalization into consideration. In other words, the location of the samples is ignored.	Geostatistics takes the regionalization into consideration. Even when populations can have same parameters like mean, variance, skewness and kurtosis still they can differ considerably. Because the regionalizations can be different.
The variability with the distance and the anisotropies are not taken into consideration	Geostatistics also takes into consideration, the anisotropies, the correlation with the distance. In other words Geostatistics takes into account the holistic view.

EXERCISES IN LINEAR GEOSTATISTICS

1. Probabilities

Question 1.1

The variance of a random variable x is, by definition, $V(x) = E[x - \mu]^2$ where $\mu = E(x) = \int xf(x)dx$.

Show that $V(x) = E(x^2) - \{E(x)\}^2$

Answer

$$\begin{aligned}\mathrm{Var}(X) &= \int (x-\mu)^2 f(x)dx \\ &= \int x^2 f(x)dx - 2\mu\int xf(x)dx + \mu^2 \int f(x)dx \\ &= E(x^2) - 2\ \mu\mu + \mu^2\ .1 \\ &= E(x^2) - \mu^2 \\ &= E(x^2) - \{E(x)\}^2\end{aligned}$$

Question 1.2

Let x be a random variable and "a" a constant.

(a) Show that $E(ax) = a\ E(x)$.

(b) Show that $\mathrm{Var}(ax) = a^2\ \mathrm{Var}(x)$.

Answer

(a) $E(ax) = \int (ax)f(x)dx = a\int xf(x)dx = a\mu = aE(x)$

(b) $$\begin{aligned}\mathrm{Var}(ax) &= E[(ax)^2] - E[(ax)]^2 \\ &= a^2E(x^2) - a^2\mu^2 \\ &= a^2V(x)\end{aligned}$$

Question 1.3

Let x, y and z be three random variables such that $z = x + y$.
Show that
$V(z) = V(x) + V(y) + 2\mathrm{Cov}(x, y)$

Answer

$$\begin{aligned}E(x + y) &= \int\int (x+y)f(x,y)dx.dy \\ &= \int\int xf(x,y)dx.dy + \int\int yf(x,y)dx.dy \\ &= \int\int x[f(x,y)dy]dx + \int\int f(x,y)dx.y.dy \\ &= \int\int x.f(x)dx + \int y.f(y).dy\end{aligned}$$

$$E[(x + y) = E(x) + E(y)$$

Also $V(x + y) = E[(x + y)^2] - [E(x + y)]^2$

$$= E(x^2 + 2.x.y + y^2) - (\mu_x + \mu_y)^2$$
$$= E(x^2) + 2.E(x.y) + E(y^2) - \mu_x^2 - 2.\mu_x.\mu_y - \mu_y^2$$
$$= (E(x^2) - \mu_x^2) + (E(y^2) - \mu_y^2) + 2[E(x.y) - \mu_x.\mu_y]$$
$$= V(x) + V(y) + 2\text{Cov}(x, y)$$

In fact

$$\text{Cov}\ (x, y) = E[(x - \mu_x)(y - \mu_y)]$$
$$= E(x.y - \mu_x.y - \mu_y.x + \mu_x.\mu_y)$$
$$= E(x.y) - \mu_x.E(y) - \mu_y.E(x) + \mu_x\mu_y$$
$$= E(x.y) - \mu_x.\mu_y - \mu_x.\mu_y + \mu_x.\mu_y = E(x.y) - \mu_x.\mu_y)$$
$$\text{Cov}\ (x^2) = (\text{Cov}(x.x)) = E(x, x) - m^2 = E(x^2) - m^2$$
$$E(x^2) = \text{Cov}(x, x) + m^2$$

2. Properties of Variogram and Covariance Models

Question 2.1

Let $Z(x)$ be a stationary random function. Show that the co-variance that is equal to

$C(h) = E\{[Z(x + h) - m)[Z(x) - m]\}$ has the following properties:

(i) $C(0) \geq 0$

(ii) $C(h) = C(- h)$

(iii) $|C(h)| \leq C(0)$

Answer

(i) $C(h) = E\{[Z(x+h) - m][Z(x) - m]\}$

$C(0) = E\{[Z(x) - m]^2\} = V(x) \geq 0$

(ii) $C(-h) = E\{[Z(x-h) - m][Z(x) - m]\}$ Put $x - h = y$

$= E\{[Z(y) - m][Z(y+h) - m]\}$

$= C(h)$

Because the random variable is stationary.

(iii) It is to be shown that $-C(0) \leq C(h) \leq C(0)$

to show that $C(h) \leq C(0)$

$$0 \leq E\{[Z(x+h) - Z(x)]^2\} = E\{[(Z(x+h) - m) - (Z(x) - m)]^2\}$$

The expectation of a square is always ≥ 0

$$= E[Z(x+h) - m]^2 + E[Z(x) - m]^2 - 2E[Z(x+h) - m] \times E[Z(x) - m]\}$$
$$= 2C(0) - 2C(h) \geq 0 \Rightarrow C(h) \leq C(0)$$

to show that $-C(h) \geq C(0)$ again:

$$E\{[Z(x+h) - m) + (Z(x) - m)]^2\} \geq 0 \Rightarrow E[Z(x+h) - m]^2$$

$$\Rightarrow E\{[Z(x+h)-m]^2 + E[Z(x)-m]^2 +$$
$$2E[Z(x+h)-m]\times E[Z(x+h)-m]\} \geq 0$$
$$\Rightarrow 2C(0)+2C(h)\geq 0 \Rightarrow -C(0)\leq C(h)$$

Hence: $-C(0)\leq C(h)\leq C(0)\Rightarrow |C(h)|\leq C(0)$

Question 2.2

Let $Z(x)$ be a stationary random function. Show that the variogram that is equal to $\gamma(h)=\frac{1}{2}E[Z(x+h)-Z(x)]^2$ has the following properties:

$\gamma(0)=0$

$\gamma(-h)=\gamma(h)\geq 0$

Answer

(i) $\gamma(0)=\frac{1}{2}E[Z(x)-Z(x)]=0$

$\gamma(h)$ is the expectation of a square, hence it has to be positive.

$\gamma(-h)=\frac{1}{2}E[Z(x-h)-Z(x)]^2$

Let $y = x - h$ $\gamma(x,h)=\gamma(y,h)$, because it is intrinsic

$$\gamma(-h)=\gamma(y,h)=\frac{1}{2}E[Z(y)-Z(y+h)]^2=\gamma(h)\Rightarrow\gamma(h)=\gamma(-h)$$

Question 2.3

Let $Z(x)$ be a stationary random function. Show that

$\gamma(h)=C(0)-C(h)$

Answer

$$\begin{aligned}\gamma(h) &= \frac{1}{2}E\{[Z(x+h)-m]-[Z(x)-m]\}^2\\ &= \frac{1}{2}\{E[Z(x+h)-m]^2+E[Z(x)-m]^2\\ &\qquad -2E[(Z(x+h)-m)(Z(x)-m)]\}\\ &= \frac{1}{2}[2C(0)-2C(h)]\\ \gamma(h) &= C(0)-C(h)\end{aligned}$$

Question 2.4

Let $Z^*(x)$ be a stationary random function.

$Z^*(x)=\sum_{i=1}^{n}\lambda_i Z(x_i)$ where λ_i are constants. Show that

$$V(Z^*(x))=\sum_i\sum_j\lambda_i\lambda_j \text{ Cov }(Z(x_i),Z(x_j))$$

$$= -\sum_i \sum_j \lambda_i \lambda_j \gamma(Z(x_i), Z(x_j)) \text{ where } \sum \lambda_i = 0$$

Answer '

We know that $\text{Cov}(Z(x_i), Z(x_j)) = C(0) - \gamma(Z(x_i), Z(x_j))$

Hence $V(Z^*(x)) = \sum_i \sum_j \lambda_i \lambda_j [C(0) - \gamma(Z(x_i), Z(x_j))]$

$$= C(0) \sum_i \sum_j \lambda_i \lambda_j - \sum_i \sum_j \lambda_i \lambda_j \gamma(Z(x_i), Z(x_j))$$

$$= C(0)(\sum (\lambda_i)^2 - \sum_i \sum_j \lambda_i \lambda_j \gamma(Z(x_i), Z(x_j))$$

But $\sum \lambda_i = 0$ by the hypothesis.

Therefore: $V(Z^*(x)) = -\sum_i \sum_j \lambda_i \lambda_j \gamma(Z(x_i), Z(x_j))$ If $\sum \lambda_i = 0$

3. Standard Models

Spherical Model

$$\gamma(h) = \begin{cases} C\left(\frac{3}{2} \times \frac{h}{a} - \frac{1}{2} \times \frac{h^3}{(a)^3}\right) & \text{when } |h| < a \\ C & \text{when } |h| \geq a \end{cases}$$

Question 3.1

Trace the curve of $\gamma(h)$ when sill $c = 2$ and range $a = 100$ m.

Answer

The expression for $\gamma(h)$ when sill $c = 2$ and range $a = 100$ m will be

$$\gamma(h) = \begin{cases} \left(3 \times \frac{h}{100} - \frac{h^3}{(100)^3}\right) & \text{when } |h| < 100 \\ 2 & \text{when } |h| \geq 100 \end{cases}$$

The slope at the origin is equal to $\gamma'(h)$ when $h = 0$ which is equal to 3/100.

$$\gamma'(h) = \frac{d\gamma(h)}{dh} = \frac{3}{100} - \frac{3h^2}{100^3}$$

We will calculate $\gamma(h)$ when h takes values 0, 20, 40, 60, 80, 100, 120. Elaborate calculations are shown for two cases. Rest can be calculated in similar manner.

$$\begin{aligned}
\gamma(0) &= 3(0/100) - (0/100)^3 = 0 \\
\gamma(20) &= 3(20/100) - (20/100)^3 = 3 \times 0.2 - 0.008 = 0.59 \\
\gamma(40) &= 3 \times 0.4 - 0.064 = 1.14 \\
\gamma(60) &= 3 \times 0.6 - 0.216 = 1.58 \\
\gamma(80) &= 3 \times 0.8 - 0.512 = 1.9 \\
\gamma(100) &= 2 \\
\gamma(120) &= 2
\end{aligned}$$

Question 3.2

Trace the variogram which is a sum of two spherical variograms, the first variogram having sill = 1, range 50 m, and second variogram having a sill 1 and range 100 m. (Note the curvature at h = 50 m and at h = 100 m).

Answer

It may be remembered that variograms are additive. Sum of two spherical variograms with $c_1 = 1$, $a_1 = 50$, and $c_2 = 1$, $a_2 = 100$

$$\gamma(h) = \gamma_1(h) + \gamma_2(h) = \frac{3}{2} \times \frac{h}{50} - \frac{1}{2} \times \frac{h^3}{50^3} + \frac{3}{2} \times \frac{h}{100} - \frac{1}{2} \times \frac{h^3}{100^3}$$

When $h < a_1$: $\gamma_1(h) + \gamma_2(h) = \frac{3}{2} \cdot \frac{3h}{100} - \frac{1}{2}\left(\frac{8h^3}{100^3} + \frac{h^3}{100^3}\right) = \frac{1}{2}\left[\frac{9h}{100} - \frac{9h^3}{100^3}\right]$

$$= \frac{9}{2}\left[\frac{h}{100} - \frac{h^3}{100^3}\right] = \frac{9h}{200}\left(1 - \frac{h^2}{100^2}\right)$$

When $a_1 \le h \le a_2 ..\text{or}.. h \in [a_1, a_2] : \gamma_1(h) + \gamma_2(h)$

$$= 1 + \gamma_2(h) = 1 + \frac{3}{2} \cdot \frac{h}{100} - \frac{1}{2} \cdot \frac{h^3}{100^3}$$

When $h > a_2$: $\gamma_1(h) + \gamma_2(h) = C_1 + C_2 = 2$

$\gamma'(h) = \frac{d\gamma(h)}{dh} = \frac{1}{2}\left[\frac{9}{100} - \frac{3h^2}{100^3}\right]$ Slope near the origin is when h = 0

$$\begin{aligned}
\gamma'(0) &= 9/200 \\
\gamma(0) &= \frac{9}{2}\left(0 - (0)^3\right) = 0 \\
\gamma(10) &= \frac{9}{2}\left(0.1 - (0.1)^3\right) = 0.445
\end{aligned}$$

$$\gamma(20) = \frac{9}{2}\left(0.2-(0.2)^3\right)=0.86$$

$$\gamma(30) = \frac{9}{2}\left(0.3-(0.3)^3\right)=1.23$$

$$\gamma(40) = \frac{9}{2}\left(0.4-(0.4)^3\right)=1.51$$

$$\gamma(50) = \frac{9}{2}\left(0.5-(0.5)^3\right)=1.69$$

$$\gamma(60) = 1+\frac{3}{2}\times 0.6-\frac{1}{2}\times\left(0.6\right)^3 =1.79$$

$$\gamma(70) = 1+\frac{3}{2}\times 0.7-\frac{1}{2}\left(0.7\right)^3 =1.88$$

$$\gamma(80) = 1+\frac{3}{2}\times 0.8-\frac{1}{2}(0.8)^3 =1.944$$

$$\gamma(90) = 1+\frac{3}{2}\times 0.9-(0.9)^3 =1.99$$

$$\gamma(100) = 1 + 1 = 2$$

$$\gamma(110) = 1 + 1 = 2$$

Question 3.3

Calculate the slope of the spherical variogram near the origin. Show that the tangent at the origin cuts the straight line $Y = c$ at $h = (2/3)a$.

Answer

The tangent at the origin cuts the straight line $Y = c$ at $h = (2/3)a$. The slope of the spherical variogram near origin is obtained by differentiating variogram function and substituting $h = 0$.

$$\gamma(h)=c\left(\frac{3}{2}\cdot\frac{h}{a}-\frac{1}{2}\cdot\frac{h^3}{a^3}\right)$$

$$\gamma'(h)=\frac{d\gamma(h)}{dh}=c\left(\frac{3}{2a}-\frac{3h^2}{2a^3}\right)$$

Substituting $h = 0$, $\gamma'(0)=\frac{3c}{2a}$ which is the slope near origin.

The equation of the tangent near the origin is $Y=\frac{3c}{2a}\cdot h$

When this tangent cuts the line $Y = c$ at the point of intersection, the values of (x, y) are common for both the lines, which can be obtained by equating them.

$$Y = c = \frac{3c}{2a}h \qquad \therefore h = \frac{2a}{3}$$

Question 3.4

Show that the spherical variogram has the same behaviour at the origin as the linear variogram with the following equation:

$$\gamma(h) = \frac{3c}{2a}|h|$$

Answer

At the origin, $\frac{h^3}{a^3}$ term is so small that it can be neglected for practical purposes. Hence the equation of the variogram can be considered as

$$\gamma(h) \cong c\left(\frac{3}{2}\cdot\frac{|h|}{a}\right) = \frac{3c}{2a}\cdot|h|$$

But this is nothing but the linear variogram. Hence it is proved that spherical variogram has the same behaviour at the origin as the linear variogram with the equation $\gamma(h) = \frac{3c}{2a}|h|$.

Exponential Model

$\gamma\,(h) = C[1 - \exp(-|h|/a)]$

Question 3.5

Trace the variogram curve when sill $c = 2$ and range $a = 30$ m. Make a comparison with those of the questions 3.1 and 3.2.

Answer

$$\gamma'(h) = \frac{d\gamma(h)}{dh} = c(-e^{-\frac{|h|}{a}})\cdot\left(-\frac{1}{a}\right) = \frac{c}{a}\left(e^{-\frac{h}{a}}\right)$$

We can get the slope near the origin by differentiating the Gamma function and substituting $h = 0$ in the differential.

$$\gamma'(0) \;=\; \frac{c}{a}\left(e^{-\frac{0}{a}}\right) = \frac{c}{a}\left(e^{-0}\right) = \frac{c}{a} = \frac{2}{30}$$

$$\gamma(0) \;=\; 2\left(1 - e^{-\frac{0}{30}}\right) = 0$$

$$\gamma(20) = 2\left(1-e^{-\frac{20}{30}}\right)=0.97$$

$$\gamma(40) = 2\left(1-e^{-\frac{40}{30}}\right)=1.47$$

$$\gamma(60) = 2\left(1-e^{-\frac{60}{30}}\right)=1.73$$

$$\gamma(80) = 2\left(1-e^{-\frac{80}{30}}\right)=1.86$$

$$\gamma(100) = 2\left(1-e^{-\frac{100}{30}}\right)=1.93$$

Note: The exponential variogram, in comparison to spherical variogram, rises very quickly in the beginning. On the contrary it increases very slowly and reaches the sill value very slowly. Since it is an exponential function, theoretically it reaches the sill value at infinity only. Therefore a practical range is defined below.

Question 3.6

Calculate the practical range of the model, i.e. the value of h, for which $\gamma(h)$ attains 95% of its sill value.

Answer

$$\gamma(h)=95\% \times c=0.95c=c\left(1-e^{-\frac{h}{a}}\right)=0.95=\left(1-e^{-\frac{h}{a}}\right)$$

$$e^{-\frac{h}{a}}=1-0.95=0.05$$

$$h=-a\times \log(0.05) \Leftrightarrow h \cong 3a.$$

Note: One important point to be noted is that in spherical model, a represents the sill value at which the $\gamma(h)$ reaches the maximum value. Whereas in exponential model, theoretically, the sill value is reached when $h = \infty$. That is why we are forced to define a working sill value of 95% value.

Question 3.7

Calculate the slope of the variogram near the origin. Show that the tangent near the origin cuts the line $Y = c$ at $h = a$.

Answer

We can get the slope near the origin by differentiating the Gamma function and substituting $h = 0$ in the differential.

$$\gamma'(h) = \frac{d\gamma(h)}{dh} = c(-e^{-\frac{|h|}{a}}) \cdot \left(-\frac{1}{a}\right) = \frac{c}{a}\left(e^{-\frac{h}{a}}\right)$$

The slope near the origin = $\gamma'(0) = \frac{c}{a}\left(e^{-\frac{0}{a}}\right) = \frac{c}{a}\left(e^{-0}\right) = \frac{c}{a}$

The equation of the tangent near the origin is $Y = \frac{c}{a}h$

Since it cuts the line $Y = c$, both can be equated near point of intersection. In other words:

$$\frac{c}{a}h = c \quad \text{or} \quad h = a\,.$$

Gaussian Model

$$\gamma(h) = c\left(1 - e^{-\frac{h^2}{a^2}}\right) = c[1 - \mathrm{Exp}(-h^2/a^2)]$$

Question 3.8

Trace the curve of the Gaussian variogram, when Sill $c = 2$ and the Range $a = 50$. Make a comparison of the curves in questions 3·5 and 3·1.

Answer

We can get the slope near the origin by differentiating the Gamma function and substituting $h = 0$ in the differential.

$$\gamma(h) = c\left(1 - e^{-\frac{h^2}{a^2}}\right)$$

$$\gamma' h) = \frac{d\gamma(h)}{dh} = c\left[-e^{-\frac{h^2}{a^2}} \times \left(-\frac{2h}{a^2}\right)\right] = c\left(\frac{2h}{a^2}\right) \times e^{-\frac{h^2}{a^2}}$$

$$\gamma'(0) = c\left(\frac{2 \times 0}{a^2}\right) \cdot e^{-\frac{0}{a^2}} = 0$$

$$\gamma(0) = 2\left(1 - \mathrm{Exp}\,(-0^2/50^2)\right) = 0 \qquad \gamma(10) = 2\left(1 - \mathrm{Exp}\,(-1/25)\right) = 0.08$$

$$\gamma(20) = 2(1 - \mathrm{Exp}\,(-4/25)) = 0.30 \qquad \gamma(30) = 2(1 - \mathrm{Exp}\,(-9/25)) = 0.60$$

$$\gamma(40) = 2(1 - \mathrm{Exp}\,(-16/25)) = 0.95 \qquad \gamma(50) = 2(1 - \mathrm{Exp}\,(-25/25)) = 1.26$$

$$\gamma(60) = 2(1 - \mathrm{Exp}\,(-36/25)) = 1.53 \qquad \gamma(70) = 2(1 - \mathrm{Exp}\,(-49/25)) = 1.72$$

$$\gamma(80) = 2(1 - \mathrm{Exp}\,(-64/25)) = 1.85 \qquad \gamma(90) = 2(1 - \mathrm{Exp}\,(-81/25)) = 1.92$$

$$\gamma(100) = 2(1 - \mathrm{Exp}\,(-100/25)) = 1.96$$

Comparison

The Gaussian scheme presents a parabolic behaviour near the origin with a horizontal tangent. Hence it rises very slowly at the beginning, in comparison to either Exponential or Spherical Schemes. The Gaussian scheme also shows a concavity whereas the other two schemes are convex. Lastly, for the values of h nearer to the range, the Spherical and Gaussian schemes appear same.

Question 3.9

Show that the behaviour of the variogram is parabolic near the origin. Model in $|h|^{\alpha}$

$$\gamma(h) = |h|^{\alpha} \quad 0 \leq \alpha < 2$$

Answer

The behaviour of the variogram is parabolic near the origin.
Behaviour of e^{x} at the origin is:

$$e^{x} = 1 + x + \frac{x^{2}}{2!} + \ldots\ldots\ldots\ldots\ldots\ldots + \frac{x^{n}}{n!} \quad (\text{for } x \approx 0)$$

Hence $\gamma(h) \approx c(1 - (1 - \frac{h^{2}}{a^{2}} + \frac{h^{4}}{2!a^{4}} + \ldots\ldots\ldots)) \approx c\frac{h^{2}}{a^{2}}\ (\text{for } h \rightarrow 0)$: This is because as $h \rightarrow 0$, the higher powers of h can be neglected. This represents the equation of a parabola of axis (0, y).

Hence the variogram has a parabolic behaviour near the origin. A few clarifications below:

Origin = h = 0 Behaviour near the origin means $h \approx 0$.

Model in $|h|^{\alpha}$

Question 3.10

Trace the curves for different values of α (for example α = ½, 1, 3/2).

Answer

The curves for different values of α (for example α = ½, 1, 3/2).

α = 0.5

h	0	1	2	3	5	10	15	20	25
$\gamma(h)$	0	1	1.41	1.73	2.24	3.16	3.87	4.47	5

α = 1

h	0	1	2	3	5	10	15	20	25
$\gamma(h)$	0	1	2	3	5	10	15	20	25

α = 1.5

h	0	1	2	3	5	10	15	20	25
$\gamma(h)$	0	1	2.83	5.2	11.18	31.62	58.1	89.4	125

Clarification: It may be clarified that α can not take values greater than or equal to 2. It can be 1.9 or 1.999, but cannot be equal to 2 or more.

4. Calculation of Experimental Variograms

Regular Grids in One Direction

Question 4.1

```
             1        3        5        7       9        8       6        4        2
Case I       I--------I--------I--------I------- I--------I-------I--------I--------I
             5        1        9        2       3        7       8        4        6
Case II      I--------I--------I--------I------- I--------I-------I--------I--------I
```

The above figures represent the value of $Z(x)$ at each point. Calculate the experimental variograms in both the cases.

Answer

The theoretical formula for the variogram is

$$\gamma(h)=\frac{1}{2}E\left[Z(x+h)-Z(x)\right]^2.$$

But to calculate practically $\gamma(h)=\frac{1}{2N}\sum\left(Z(x+h)-Z(x)\right),^2$ where N represents the number of pairs and h represents the lag. The variograms for the cases are calculated using the formula and presented in Table 5. The distance lag and the number of pairs for both the cases are same since the measurement points are same.

Table 5: Variograms of two data sets having same mean and variance

h	*γ(h)–Case I*	*γ(h)–Case II*	*Number of pairs*
1	1.81	10.44	8
2	6.43	8.29	7
3	11.92	4.58	6
4	14.80	5.40	5
5	10.50	11.75	4
6	5.83	4.50	3
7	2.50	6.50	2
8	0.50	0.50	1

It may be noted that though both data sets are numeral from 1 to 9 with the same mean and variance but have different variograms.

Question 4.2

```
10 12   36   72   91   38   31    50    69    07   21    18    71    08   40  82
I --- I --- I --- I --- I --- I --- I --- I --- I --- I --- I --- I --- I --- I --- I--- I
```

Calculate the experimental Variogram for the above distribution.

Answer

h	γ(*h*)	*Number of pairs*	*h*	γ(*h*)	*Number of pairs*
1	647.7	15	9	707.8	7
2	962.7	14	10	834.6	6
3	896.2	13	11	543.4	5
4	615.7	12	12	481.6	4
5	810.7	11	13	484.0	3
6	1116.5	10	14	1450.0	2
7	861.5	9	15	2592.0	1
8	565.8	8			

Question 4.3

2.4 3.0 - 1.8 1.5 3.0 - 2.5

I-------- I-------- I-------- I-------- I-------- I-------- I-------- I

The numbers represent the tenors in percentage. Two values are missing. Calculate the experimental variogram.

Answer

Note: While calculating the variograms with missing values, the pairs with missing values is omitted and it is not taken into consideration while calculating the number of pairs.

h	0	1	2	3	4	5	6	7
γ(*h*)	0	0.45	0.52	0.60	0.22	0.18	0.13	0.01
Number of pairs	6	3	3	3	3	1	1	1

CONCLUSIONS

Geostatistical methods have found applications in many fields of Earth Sciences and much more in hydrogeology. In this article, efforts have been made to introduce the subject in a possible simple way by defining the basic statistics utilized to develop the method, the basic mathematical derivations required as well as a few critical comparisons between classical statistical methods and that of geostatistical methods. Variography, the most important component of the geostatistical methods, has been dealt with in details and finally a number of examples of computing the variograms are presented for demonstration purpose.

REFERENCES

Isaaks, H.E. and Srivastava, R.M., 1989. An Introduction to Applied Geostatistics, Oxford Universtity Press, USA.

Marsily, G.De., 1986. Quantitative Hydrogeology: Groundwater Hydrology for Engineers, Academic Press.

Matheron, G., 1963. Principles of Geostatistics, *Econ. Geol.* **58:** 1246-1266.