*Mathematical Geology, Vol. 21, No. 8, 1989*

# Cumulative Semivariogram Models of Regionalized Variables[1]

**Zekâi Şen[2,3]**

*The cumulative semivariogram approach is proposed for modeling regionalized variables in the geological sciences. This semivariogram is defined as the successive summation of half-squared differences which are ranked according to the ascending order of distances extracted from all possible pairs of sample locations within a region. This procedure is useful especially when sampling points are irregularly distributed within the study area. Cumulative semivariograms possess all of the objective properties of classical semivariograms. Classical semivariogram models are evaluated on the basis of the cumulative semivariogram methodology. Model parameter estimation procedures are simplified with the use of arithmetic, semilogarithmic, or double-logarithmic papers. Plots of cumulative semivariogram values vs. corresponding distances may scatter along a straight line on one of these papers, which facilitates model identification as well as parameter estimation. Straight lines are fitted to the cumulative semivariogram scatter diagram by classical linear regression analysis. Finally, applications of the methodology are presented for some groundwater data recorded in the sedimentary basins of the Kingdom of Saudi Arabia.*

**KEY WORDS:** cumulative semivariogram, classical semivariogram, regionalized variable.

## INTRODUCTION

Field measurements of geological variables, such as ore grades, chemical constitutions in groundwater, fracture spacings, porosity, permeability, aquifer thickness, and dip and strike of a structure, are dependent on the relative positions of measurement points within the study area. Measurements of a given variable at a set of points provide some insight into the regional variability. This variability determines the regional behavior as well as the predictability of the variable concerned. In general, the larger the variability, the more heterogeneous is the geological environment. As a result, the number of measurements required to model, simulate, estimate, and predict the regional behavior

[1]Manuscript received 11 July 1988; accepted 30 May 1989.

[2]Faculty of Earth Sciences, Hydrogeology Department, King Abdulaziz University, P.O. Box 1744, Jeddah 21441, Kingdom of Saudi Arabia.

[3]Permanent address: Department of Hydraulics, İnsaat Fakültesi, Istanbul Teknik Üniversitesi, Ayazağa, Istanbul, Turkey.



0882-8121/89/1100-0891$06.00/1 

is expected to be large. Large variability implies also that the degree of dependence might be rather small even for data whose locations are close to each other. A geological interpretation of such a situation may be that either the region was subjected to active geological phenomena (such as tectonics, volcanism, deposition, erosion, recharge, etc.) or to some human activities such as pollution, groundwater abstraction, mining, etc.

However, many types of geological variables are known to be related spatially in that the closer their positions, the greater is their dependence. Spatial dependence especially is pronounced in hydrogeological data due to groundwater flow as a result of the hydrological cycle, which homogenizes the distribution of chemical constituents within the heterogeneous mineral distribution in geological formations.

In order to quantify the degree of variability within spatial data, variance techniques can be used in addition to classical autocorrelation methods (Box and Jenkins, 1970). However, these methods are not helpful to account directly for the regional dependence or for the variability in terms of sample positions. The drawbacks are due to either nonnormal (asymmetric) distribution of data and/or irregularity of sampling positions. However, the semivariogram (SV) technique, developed by Matheron (1963, 1971, 1973) and used by many researchers (Clark, 1977; Cooley, 1979; David, 1977; Myers et al., 1982; Journel, 1985; Aboufirassi and Marino, 1984; Hoeksema and Kitanidis, 1984; Carr et al., 1985) in diverse fields, such as geology, mining, hydrology, earthquake prediction, and groundwater, can be used to characterize spatial variability. The SV is a prerequisite for best linear unbiased prediction of regionalized variables through the use of kriging techniques (Krige, 1982; Journel and Huijbregts, 1978; Davis, 1977).

The purposes of this paper are to point out some practical difficulties as well as subjectivities of classical SV in the case of irregular sampling points, and to present fundamentals of cumulative SV models. The cumulative semivariogram (CSV) technique of modeling regional variability provides a simple procedure for identifying the underlying model and for estimating its parameters. In this work, linear regression is employed for parameter estimation.

## PRACTICAL DIFFICULTIES OF CLASSICAL SEMIVARIOGRAMS

The classical SV, $\tau(h)$, for any distance, $h$, is defined as the half-squared difference of two measurements separated by this distance. As $h$ varies from zero to the maximum possible distance within the study area, the relationship of the half-square difference to the separation distance emerges as a theoretical function which is called the "semivariogram." The sample SV is an estimate of this theoretical function calculated from a finite number, $n$, of samples. The sample SV can be estimated reliably for small distances when the distribution

of sampling points within the region is regular. As the distance increases, the number of data pairs for calculation of SV decreases, which implies less reliable estimation at large distances.

In various disciplines of the geological sciences, the sampling positions are irregularly distributed in the region and, therefore, an unbiased estimate of SV is not possible. Some distances occur more frequently than others and, accordingly, their SV estimates are more reliable than others. Hence, a heterogeneous reliability dominates the sample SV. Consequently, the sample SV may have ups and downs even at small distances. Such a situation gives rise to inconsistencies and/or experimental fluctuations with the classical SV models which are, by definition, nondecreasing functions (*i.e.*, a continuous increase with distance is their main property). In order to give a consistent form to the sample SV, different researchers have used different subjective procedures. These are:

1. Journel and Huijbregts (1978) advised grouping of data into distance classes of equal length in order to construct a sample SV. However, the grouping of data pairs into classes causes a smoothing of the sample SV relative to the underlying theoretical SV. If a number of distances fall within a certain class, the average of half-squared differences within this class is taken as the representative half-squared difference for the midclass point. The effect of outliers is partially damped, but not completely smoothed by the averaging operation.

2. To reduce the variability in the sample SV, Myers et al. (1982) grouped the observed distances between samples into variable length classes. The class size is determined such that a constant number of sample pairs falls in each class. The mean values of distances and half-squared differences were used for the classes as a representative point of sample SV. Even this procedure resulted in an inconsistent pattern of sample SV (Myers et al., 1982) for some choices of the number, $m$, of pairs falling within each class. However, Myers et al. (1982) observed that choosing $m = 1000$ gave a discernible shape. The choice of constant number of pairs is subjective and, in addition, averaging procedures smooth out the variability within the experimental semivariogram. As a result, the sample SV provides a distorted view of the variable in that it does not provide, for instance, greater frequency (shortwave length) variations. However, such short wavelength variations, if they exist, are so small that they can be safely ignored.

The above procedures have two basic common properties; namely, predetermination of a constant number of pairs or distinctive class lengths, and the arithmetic averaging procedure for half-squared differences as well as distances. The former needs a decision which in most cases is subjective, whereas the latter can lead to unrepresentative SV values. In classical statistics, only in the case of symmetrically distributed data, the mean value is the best estimate,

otherwise, the median becomes superior. Moreover, the mean value is sensitive to outliers.

## THE CUMULATIVE SEMIVARIOGRAM

The CSV is a graph which shows the relationship of successive half-squared difference summations to the ranked (ascending order) distances which are extractable from the sample positions within the study area. Ordering distances eliminates the subjectivity in selection of distance classes, whereas the successive summation eliminates the averaging procedure. Additionally, the summation of half-squared differences leads to nondecreasing sample functions. Thus, the CSV does not have the shortcomings of the classical SV. The sample CSV can be obtained by carrying out the following steps:

1. Calculate distances between every possible pair of sample positions. If the number of data positions is $n$, then $m = n\,(n - 1)/2$ half-squared differences and distances, $h_i$, $(i = 1, 2, \cdots, m)$ exist. For instance, $n = 7$ data positions give rise to 21 different half-squared difference and distance values.

2. For each pair of sample positions, find the half-squared differences, $d(h_i)$, between data values. Hence, for each distance, $h_i$, a corresponding half-squared difference may be calculated.

3. Rank the distances in ascending order with their attached half-squared differences, $d[h^{(i)}]$, where superscript $(i)$ indicates the rank. For instance, $d[h^{(1)}]$, is the half-squared difference corresponding to the smallest distance.

4. Successive summation of the ordered half-squared differences yields the sample CSV as

$$\tau_c(h_k) = \sum_{i=1}^{k} d[h^{(i)}] \qquad (k = 1, 2, \cdots, m) \tag{1}$$

where $\tau_c(h_k)$ is the value of the $k^{\text{th}}$ ordered distance CSV value.

The following attributes of the CSV must be kept in mind in any application:

1. The CSV is a nondecreasing function; however, local flat portions, implying constancy of the regionalized variable at certain distances, i.e., the same value has been observed at two locations $h$ apart, may occur.

2. The slope of the theoretical cumulative SV at any distance is an indicator of dependence between pairs of regionalized variables separated by that distance.

3. The sample CSV reflects even small dependencies between data pairs which are not possible to detect with the classical SV due to averaging.

4. The sample CSV is free of subjectivity because no *a priori* selection of distance classes is involved. In fact, real distances are employed in construction of the sample CSV rather than class midpoint distances.

## SAMPLE CLASSICAL AND CUMULATIVE SEMIVARIOGRAMS

The CSV proposed in the previous section is applied to transmissivity, total dissolved solids, and piezometric level data for the Wasia Sandstone aquifer in the eastern part of Saudi Arabia. A complete hydrogeological study of this area has been performed recently (Subyani, 1987).

GAMA3 software, developed for computing the classical SV (Journel and Huijbregts, 1978, p. 224), has been applied to groundwater variables such as transmissivity, piezometric level, and total dissolved solids from the Wasia Sandstone. The resulting sample SV and sample CSV plots (Fig. 1-3) indicate that the half-squared difference points are scattered in such a way that a clear pattern in the sample SV's, which suffer from fluctuations even at small distances is not possible. Comparisons of the sample SVs (Fig. 1-3) with the sample CSVs indicate that the latter are more orderly and have distinctive non-decreasing patterns. A sample CSV often yields more or less a straight line for large distances, which corresponds to the sill concept in the classical SV. Furthermore, the sample CSV starts as a curve before it becomes almost a straight line. The length of the distance domain over which the sample CSV occurs as a curve is a counterpart of the range in the classical SV. Hence, the range is determined straightforward from the sample cumulative SV. The piezometric level sample cumulative SV (Fig. 3) shows an initial range which has zero half-squared differences for about 10 km. Such a portion implies physically that the piezometric level does not change significantly within distances less than 60 km. In fact, the Wasia aquifer has remained free of any tectonic movements, it

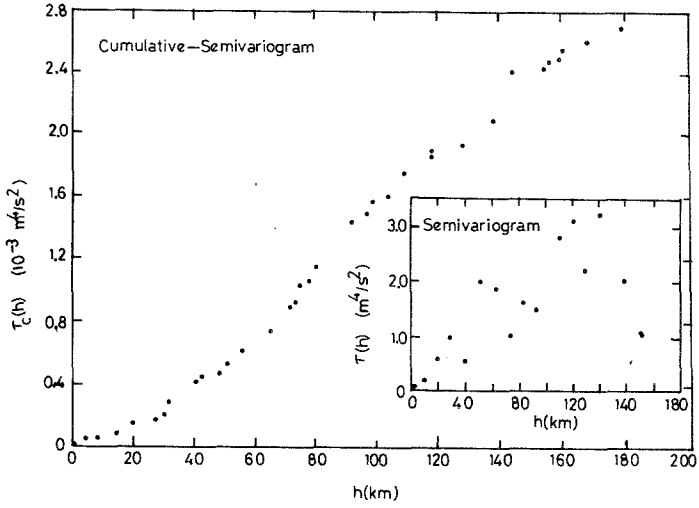


**Fig. 1.** Sample cumulative semivariogram for Wasia Sandstone transmissivity.

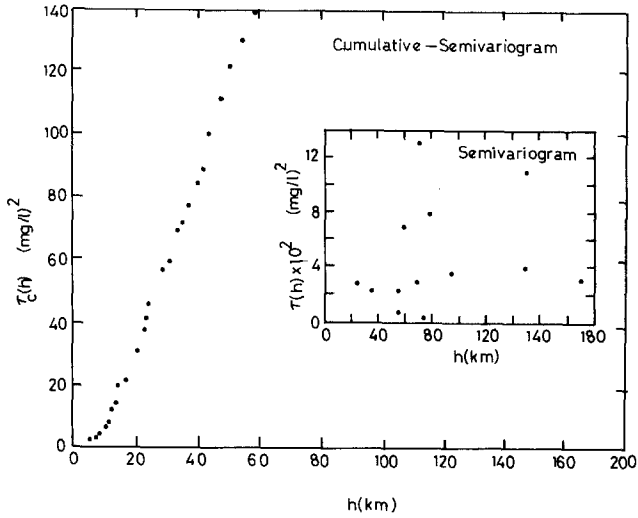


**Fig. 2.** Sample cumulative semivariogram for Wasia Sandstone dissolved solids.

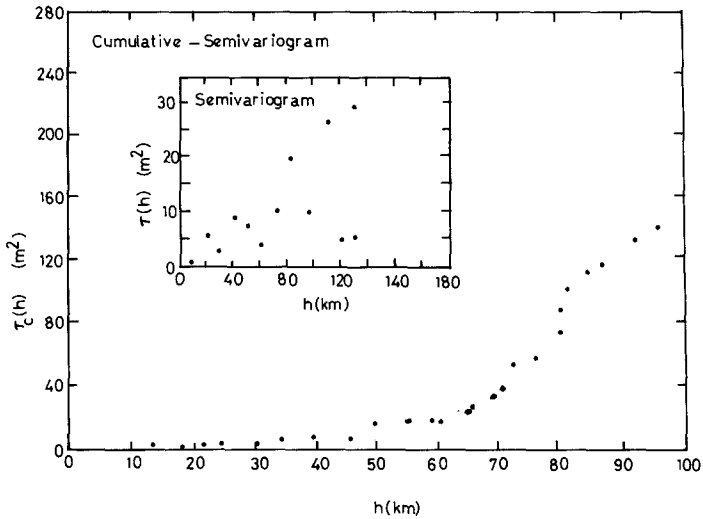


**Fig. 3.** Sample cumulative semivariogram for Wasia Sandstone piezometric level.

is extensive, and the recharge is negligible, but it is discharged by local well groups which are at large distances from each other (Powers et al., 1966).

## THEORETICAL CUMULATIVE SEMIVARIOGRAM MODELS

In order to be able to apply kriging estimation techniques to a regionalized variable, a functional relationship must be established between the distance and the measure of regional dependence which, herein, is the CSV. These models must be nondecreasing functions. Although numerous functions have this property, in practice restricting them to a few simple ones is desirable.

By considering basic definitions of both the classical and cumulative SVs, they may be related through an integration as

$$\tau_c(h) = \int_0^h \tau(u)\, du \tag{2}$$

or through differentiation as

$$\tau(h) = \frac{d\tau_c(u)}{du}\bigg|_{u=h} \tag{3}$$

Therefore, a CSV counterpart may be found for any given classical SV using Eq. (2). Furthermore, Eq. (3) indicates that the theoretical classical SV value at any distance is equal to the slope of the theoretical CSV at the same distance. In the following, models which have been used previously for SVs by many researchers will be assessed from the CSV point of view.

### Linear Model

This model postulates a linear relationship between the cumulative half-squared difference and the distance as

$$\tau_c(h) = \alpha + \beta h \tag{4}$$

in which $\alpha$ and $\beta$ are the model parameters (Fig. 4a). The sample CSV of the regionalized variable that abides by this model will appear as a straight line on arithmetic paper. In fact, $\alpha$ is the intercept on the CSV axis and $\beta$ is the slope of this straight line. This slope corresponds to the sill value in the classical SV which represents a pure nugget effect (Şen, 1979). Furthermore, $\beta$ represents exactly the variance of the underlying random field. Hence, the smaller the slope of the straight line, the smaller the random fluctuation in the regionalized variable. If the slope is equal to zero, theoretically, this indicates a complete deterministic uniform variation in the regionalized variable. The sample CSV scatter diagram and the fitted regression line to pH values measured at 71 sam-

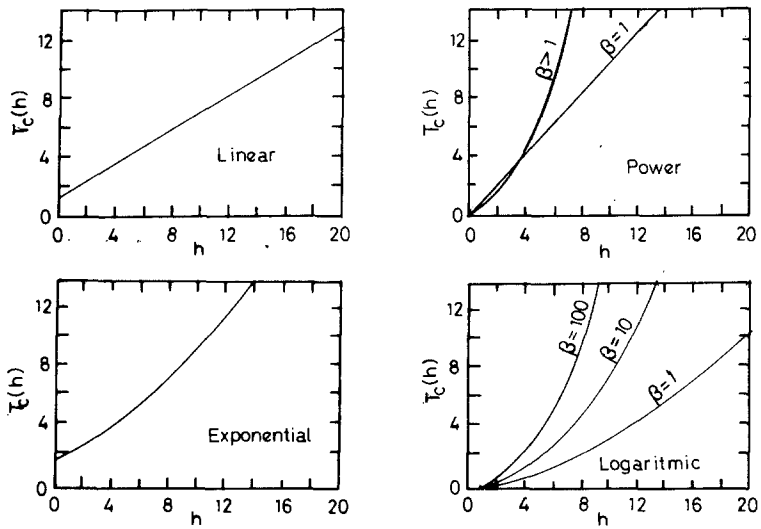


**Fig. 4.** Schematic cumulative semivariogram models.

ple locations within the Umm Er Radhuma Limestone aquifer in the Eastern Provinces, Saudi Arabia (Fig. 5) has the form

$$\tau_c(h) = -0.213 + 1.144h$$

from which the parameter estimates are $\alpha = -0.213$ and $\beta = 1.144$. The hydrochemical data were presented by Şen and Al-Dakheel (1985) for major anions and cations.

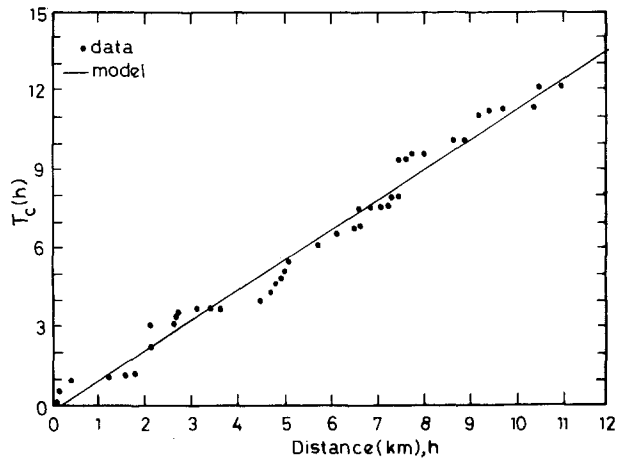


**Fig. 5.** Sample cumulative semivariogram for pH values in Umm Er Radhuma.

## Power Model

This is a two-parameter model which yields a set of different shapes for the theoretical CSV (Fig. 4b). The mathematical expression for this model is

$$\tau_c(h) = \alpha h^{\beta} \tag{5}$$

in which $\alpha$ is the scale parameter and $\beta$ is the shape parameter. Because $0 < \beta < 2$ for a theoretical SV from a power family (Journel and Huijbregts, 1978, p. 165), parameter $\beta$ for the theoretical CSV in Eq. (5) is restricted to the range $1 < \beta < 3$. The derivative of Eq. (5) yields also a power form for the classical SV. Obviously, use of a double logarithmic paper facilitates parameter estimation. Sulfate concentrations in the Umm Er Radhuma aquifer groundwater show on double logarithmic paper a more or less straight line pattern (Fig. 6). The mathematical expression of this straight line by the regression technique can be found as

$$\log \tau_c(h) = 0.46 + 0.841 \log h$$

hence, parameter estimates are $\log \alpha = 0.46$ or $\alpha = 2.88$ and $\beta = 0.84$. The original form of this model prior to transformation can be written as $\tau_c(h) = 2.88h^{0.84}$.

## Exponential Model

The general form of this model is

$$\tau_c(h) = \alpha e^{\beta h} \tag{6}$$

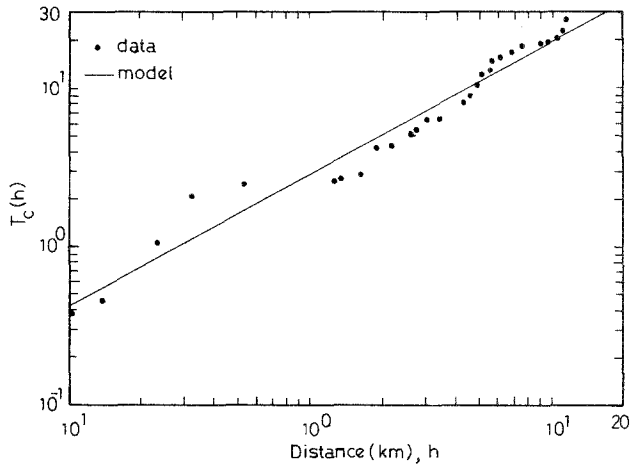


**Fig. 6.** Sample cumulative semivariogram for sulfate in Umm Er Radhuma.

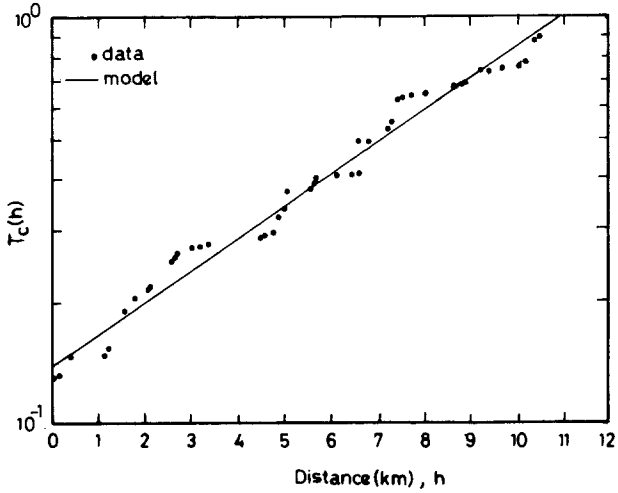


**Fig. 7.** Sample cumulative semivariogram for bicarbonate in Umm Er Radhuma.

where $\alpha$ and $\beta$ are scale and shape parameters, respectively. The main difference of this model from the others is that it has a nonzero value for zero distance, i.e., it has a nugget effect. Forms of different CSVs resulting from Eq. (6) are shown (Fig. 4c). The sample CSV can be checked for concordance with this model by plotting log $\tau_c(h)$ vs. $h$ on semilogarithmic paper. If the sample points appear as a straight line, the exponential model is the generating mechanism of the regional variability within the regionalized variable. The slope of this line directly yields an estimate of $\beta$, whereas the intercept on the $\tau_c(h)$ axis leads to an estimate of $\alpha$. This model does not have a unique classical SV which has appeared in the geostatistical literature. The sample CSV for bicarbonate concentrations in the Umm Er Radhuma aquifer appears as a straight line on semilogarithmic paper (Fig. 7). The appearance of this straight line implies that the convenient model for bicarbonate concentrations for this aquifer is of exponential type. The regression line of this scatter diagram is

$$\log \tau_c(h) = -0.86 + 0.079\, h$$

and, correspondingly, the model parameter estimates are $\log\alpha = -0.86$ or $\alpha = 0.14$ and $\beta = 0.079$. Hence the original form of the model can be written as

$$\tau_c(h) = 0.14e^{0.079h}$$

## Logarithmic Model

The mathematical expression of this model can be written as

$$\tau_c(h) = \begin{cases} \alpha + \beta \log h & \text{for } h > 1 \\ 0 & \text{for } h < 1 \end{cases} \tag{7}$$

in which $\alpha$ and $\beta$ are two model parameters. This model differs from the exponential one in that it has an intercept on the distance axis similar to the sample CSV for piezometric level (Fig. 3). Different forms of the logarithmic models are presented (Fig. 4d). The model can be depicted from a sample CSV plotted on semilogarithmic paper as $\tau_c(h)$ vs. $\log h$. If the sample points appear as a straight line, the validity of the logarithmic model is confirmed. The slope of this straight line is equal to $\beta$, and the cumulative half-squared difference corresponding to $h = 1$ yields the estimate of $\alpha$. Such a model is similar to what is referred to in the classical SV terminology as the De Wijsian model (DeWijs, 1972).

Other models for the cumulative SV can be constructed from classical cumulative SV models through Eq. (2). For instance, the exponential model of the classical SV, which is

$$\tau(h) = \alpha[1 - \exp(-\beta h)] \tag{8}$$

corresponds to a CSV model, which is

$$\tau_c(h) = \alpha\left[h + \frac{1}{\beta}\exp(-\beta h)\right] \tag{9}$$

in which $\alpha$ and $\beta$ are model parameters. A close inspection of Eq. (9) indicates that for large distances, $(1/\beta)\exp(-\beta h) \approx 0$; consequently, at large distances this model appears as a straight line (on arithmetic paper) whose slope is an estimate of $\alpha$. In addition, this model has an intercept value, $\tau_c(0)$, which is equal to $\alpha/\beta$. Provided that $\alpha$ is known from the slope at large distances, this ratio yields the estimate of $\beta$. These $\alpha$ and $\beta$ values are the parameters of the classical exponential SV model. This last example shows that the CSV method may help to estimate the parameters of the classical SV by simple graphical procedures.

Similarly, the Gaussian classical SV corresponds to the CSV model

$$\tau_c(h) = \alpha[h - \sqrt{2\pi/\beta}\,\phi\,(h, \beta)] \tag{10}$$

where $\phi\,(.,.)$ is the area under the normal probability density function (with zero mean and variance $1/\beta$) from 0 to $h$. Obviously, $\alpha$ can be estimated as the slope of this straight line.

Last but not least, a combination of the aforementioned models can appear in practical situations as mixture models.

## CONCLUSIONS

Principles of cumulative semivariograms have been explained and some CSV models have been presented. The relationship between cumulative and classical SVs are depicted. Sample CSV calculations do not involve any subjectivity, nor do they require any averaging procedure which may lead to inconsistent SVs. The CSV is a nondecreasing function of distance.

The following advantages make the CSV attractive in practical applications.

1. The CSV model may be used for irregularly distributed sample positions within the study region.
2. The CSV method is straightforward in applications without any subjective manipulations.
3. The underlying model for any regionalized variable can be detected by plotting the cumulative half-squared differences vs. distances on arithmetic, semilogarithmic, or double-logarithmic paper. Appearance of sample CSV points on any of these papers as a straight line confirms the type of model. Such an opportunity is missing in the sample classical SV calculations.
4. Model parameter estimates are obtained from the slope and intercept values of the fitted straight line.
5. Any classical SV model has a theoretical CSV counterpart which can be obtained through an integration operation.

Various theoretical CSV models are fitted to the sample CSV by simple least squares. A weighted or generalized least-squares approach would probably be preferable because the sample CSV values are correlated and do not have equal variance. Future researches should be directed toward how to implement a weighted or generalized least-squares approach; in particular, what should the weights be, and how strong are the correlations between neighboring CSV values?

## REFERENCES

Aboufirassi, M., and Marino, M. A., 1984, A Geostatistically Based Approach to the Identification of Aquifer Transmissivities in Yolo Basin, California: Math. Geol., v. 16, p. 125–137.

Box, C. E. P., and Jenkins, G. M., 1970, Time Series Analysis, Forecasting and Control: Golden Day, San Francisco, 498 p.

Carr, J. R., Bailey, R. E. and Deng, E. D., 1985, Use of Indicator Variograms for Enhanced Spatial Analysis: Math. Geol., v. 17, p. 797–812.

Clark, I., 1977, Regularization of semivariogram: Comput. and Geosci., v. 3.

Cooley, R. L., 1979, A Method of Estimating Parameters and Assessing Reliability for Models of Steady State Groundwater Flow, 2, Applications of Statistical Analysis: Water Resour. Res., v. 15, p. 603–617.

David, M., 1977, Geostatistical Ore Reserve Estimation: Elsevier, New York, 340 p.

DeWijs, H. J., 1972, Method of Successive Differences Applied to Mine Sampling: Trans. Inst. of Min. Metal., Sect. A, Min. Industry, n. 81, p. 78–81.

Hoeksema, R. J., and Kitandis, P. K., 1984, An Application of the Geostatistical Approach to the Inverse Problem in Two-Dimensional Groundwater Modeling: Water Resour. Res., v. 20, p. 1003–1020.

Journel, A. J., and Huijbregts, C. J., 1978, Mining Geostatistics: Academic Press, New York, 600 p.

Journel, A. J., 1985, The Deterministic Side of Geostatistics: Math. Geol., v. 17, p. 1–15.

Krige, D. G., 1982, Geostatistical Case Studies of the Advantages of Log-normal, De Wijsian Kriging with Mean for a Base Metal Mine and a Gold Mine: Math. Geol., v. 14, p. 547–555.

Matheron, G., 1963, Principles of Geostatistics: Economic Geology, n. 58, p. 1246–1266.

Matheron, G., 1971, Random Functions and Their Application in Geology: In Merriam, D. F. (ed.), Geostatistics, a Colloquium: Plenum, New York, p. 79–88.

Matheron, G., 1973, The Intrinsic Random Functions and Their Applications: Advances in Applied Probability, n. 5, p. 439–468.

Myers, D. E., Begovich, C. L., Butz, T. R., and Kane, V. E., 1982, Variogram Models for Regional Groundwater Geochemical Data: Math. Geol., v. 14, p. 629–644.

Powers, R. W., Ramirez, L. F., Redmond, C. D., Elberg, E. L., 1966, Geology of the Arabian Peninsula. Sedimentary Geology of Saudi Arabia. U.S. Geol. Survey, Prof Paper 560-D, 1–47, New York.

Şen, Z., 1979. Regional Drought and Flood Frequency Analysis. Theoretical Consideration: J. Hydrol., v. 46, p. 265–279.

Şen, Z., and Al-Dakheel, A. R., 1985, Hydrochemical facies evaluation in Umm Er Radhuma limestone-Eastern Saudi Arabia: Ground Water, v. 24, p. 626–635.

Subyani, A. M., 1987, Hydrogeology of the Wasia Aquifer and Its Geostatistical Modelling: Unpublished M.Sc. thesis, Faculty of Earth Sciences, King Abdulaziz University, 170 p.