# GEOSTATISTICAL ANALYSIS OF MONITORING DATA

**Mark Dowdall**
*Norwegian Radiation Protection Authority, Environmental Protection Unit, Polar Environmental Centre, Tromso, Norway*

**John O'Dea**
*Institute of Technology, Ballinode, Sligo, Ireland*

**Keywords:** anisotropy, cross-validation, data analysis, estimation, geostatistics, interpolation, intrinsic hypothesis, kriging, monitoring, nugget, range of influence, regionalized variables, sampling, semi-variogram, sill, spatial analysis, stationarity, variography

## Contents

## Summary

Continuing demands for improved environmental data analysis techniques and the need for achieving greater efficiency in monitoring programs has led to the increased use of novel data analysis techniques and the application of methods devised for other purposes to the analysis of environmental monitoring data. Although initially developed as a means of ore reserve estimation, the application of geostatistical techniques to the analysis of environmental monitoring data has proved successful in a wide variety of settings. This increased interest in geostatistics as a means of environmental data analysis is reflected in the upsurge of reports in the literature detailing the implementation of geostatistical routines in fields as diverse as atmospheric science, epidemiology, and soil chemistry. The geostatistical approach differs from that of classical statistics in its adoption of the concept of regionalized variables. While classical statistical methods are concerned with independent random variables, geostatistics describes the analysis of variables that are not independent but instead are regionalized or correlated in either space or time. The extent of this correlation is measured using one of the primary tools of geostatistics: semi-variography. The semi-variogram assesses the strength of the correlation between samples as a function of the distance separating them (in time or space) and this information is then used in the second primary tool of geostatistics: the estimation technique known as kriging. Kriging interpolates between sampled points at which the value of the variable in question is known, to produce estimates for the variable value at unsampled locations. The typical output of a geostatistical study is an isopleth map of the kriging estimates for the variable of concern. This article presents an overview of the theory and practice of geostatistics in environmental monitoring data analysis. An assessment of the advantages and possible pitfalls of this technique to the field of environmental monitoring will be made and the latest developments in the area will be presented.

## 1. Introduction

A primary goal of environmental monitoring is the determination of the temporal or spatial distribution of the pollutant or variable of interest in order to assess the levels of the pollutant present and how it varies with location or time. Environmental data sets are often difficult to analyze in relation to these objectives as the data may be clustered, highly skewed in its distribution, or may exhibit features such as global or local trends that make spatial analysis difficult by conventional means. For logistical, economical, or technical reasons the data set may be relatively small or may be "missing" some values for locations that could not be sampled. The implementation of estimation/interpolation procedures may therefore be required to fully describe the occurrence of the pollutant of interest over the area being studied, to estimate values of the pollutant for unsampled locations, and to allow for the production of contour maps. At a more advanced level, such techniques may be necessary for decision making in relation to, for example, remediation decisions based on an estimation of the probability of local areas exceeding some cutoff contaminant level. The explosion of interest in geostatistical techniques as applied to environmental monitoring is indicative of both the need to address the problems inherent in the analysis of environmental data and the recognition that geostatistical techniques have much to offer over and above more conventional spatial analysis methods. This article introduces the underlying concepts in the practice of geostatistics and describes their application to the analysis of environmental monitoring data. Although the number of reported applications of the technique is increasing rapidly, no individual case studies will be presented to ensure the article is unhindered by the specificities of any single environmental parameter. Terms commonly used in geostatistical terminology such as "deposit" and "grade" (which reflect the underlying geological origin of the techniques) have been omitted and replaced with terms commonly encountered in environmental monitoring, such as "variable," "study area," and "sample." The mathematics of the article have been limited to the standard geostatistical equations and no derivations are presented.

The term "geostatistics" refers to the application of the theories devised by Matheron in the 1960s based on empirical work conducted by Krige a decade earlier pertaining to the estimation of gold ore reserves in South African mines. The term reflects the fact that the earliest applications of the techniques were confined to the area for which they were developed, although recent years (from the mid 1980's) have seen an upsurge in the number and diversity of the fields in which they are being implemented. Geostatistical methods differ substantially from classical statistical approaches by utilizing the concept of regionalized variables and using tools such as semi-variography to examine the spatial characteristics of these variables. Information provided by the semi-variogram is then used in the estimation procedure known as kriging to provide estimates of variable values at unsampled locations as well as a measure of the reliability of estimates made at individual locations.

## 2. Regionalized Variables

Random variables are variables that vary in a probabilistic manner between individual samples or observations and are considered to be spatially independent whereas, in contrast, a regionalized variable is generally considered to consist of two components:

(a) a random component where an observation of a variable at a point $w_i$, within the larger study area $w$, is a realization of a random variable $Z(w_i)$ at the point $w_i$,

(b) a nonrandom or structured component in which the random variables for two locations $w_i$ and $w_{i+h}$ (separated by a distance $h$) are not considered spatially independent

The concept of regionalized variables is often invoked to support the phenomena, experienced by many practitioners in the environmental field, that the values of a variable measured at two nearby locations are likely to be less dissimilar than variable values at distant locations. Certain stages in the geostatistical process require that the regionalized variable obey specific conditions, these conditions usually being referred to as the "intrinsic hypothesis" or "weak stationarity." A variable obeys the intrinsic hypothesis if:

(a) the expected difference (in the value of the variable) between individual points in the data set is zero,

$$E\left[Z(w_{i+h})-Z(w_i)\right]=0 \quad \text{for all locations } w \tag{1}$$

(b) the variance of the set of differences in pairs of variable values is only a function of the separation distance $h$,

$$\text{Var}[Z(w_{i+h})-Z(w_i)]=2\gamma(h) \quad 2\gamma(h) \text{ being the variogram value} \tag{2}$$

Although the measure of spatial correlation known as the semi-variogram requires fulfillment of the intrinsic hypothesis, other less common measures of spatial correlation, such as the correlogram, require that $E[Z(w)]$ exists and is constant for all $w$ and that the covariance exists and is only a function of the separation distance, $h$ (a condition also known as strong stationarity). The presence of a trend in the data set is indicated when the expected difference does not equal zero, this trend usually being modeled separately before incorporation into the estimation process.

## 3. The Semi-Variogram

The semi-variogram is a geostatistical tool for the calculation of the extent of the spatial correlation exhibited by a regionalized variable and is described by the function:

$$\gamma(h) = \frac{1}{2N_h}\sum_{i=1}^{N_h}[Z(w_{i+h}) - Z(w_i)]^2 \qquad (3)$$

where $N_h$ represents the number of points in the data set separated by $h$.

Thus it can be seen that the semi-variogram for a regionalized variable consists of a plot of half the variance of the set of differences exhibited by pairs of points or samples separated by a vector $h$ as a function of $h$. A typical semi-variogram is depicted in Figure 1. As only a limited number of data points are available to calculate the semi-variogram, an experimental (or empirical) semi-variogram is initially plotted and a theoretical model is fitted to the result. The application of the theoretical model is the modeling of the spatial structure. The maximum semi-variance that the semi-variogram attains is deemed the sill, the separation distance or "lag" at which the sill (or a certain portion of it for an asymptotic model) is reached is known as the "range of influence." Sample points separated by a distance less than the range of influence are spatially correlated, that correlation being described by the theoretical model applied.

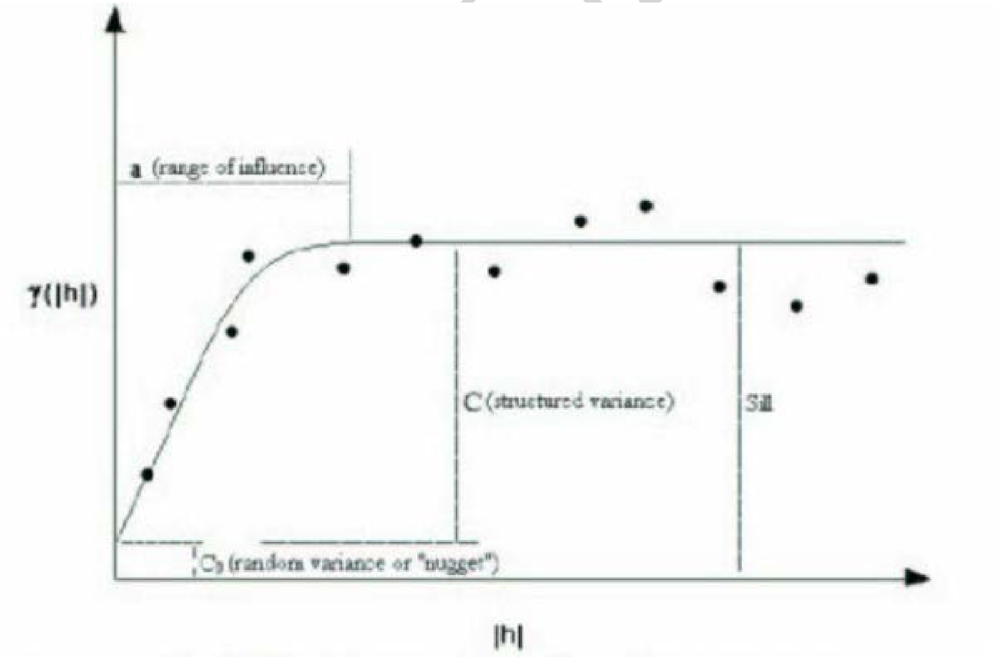

Figure 1. Schematic representation of a typical semi-variogram

The sill can be often be divided into two components, a random or "nugget" component of the spatial structure and a nonrandom or structured component. The nugget consists of two components, the aforementioned random component and a variance introduced through errors in the sampling and analytical measurement of the variable and the inherent randomness in the data. Although the former can be reduced through better sampling and analytical procedures, the latter cannot be mitigated against. The nugget component must be accounted for in subsequent stages of the analysis procedure such as kriging. Semi-variograms are usually constructed for a number of different directions in order to establish whether the spatial structure is isotropic, or exhibits the same properties irrespective of direction. If the range of influence varies between directional semi-variograms then the anisotropy is geometric and the range of influence exhibits an ellipsoid shape as opposed to the circular shape exhibited under isotropic conditions. A second type of anisotropy is zonal anisotropy, which exists when the sill value varies with direction. The primary purpose of the semi-variogram is the calculation of the range of influence. This distance describes the extent of the spatial correlation of the data and is used in the assignment of weights to known samples in order to estimate the value of the variable at unsampled locations.

## 4. Theoretical Semi-Variogram Models

The purpose of fitting a theoretical model to the empirical (or "raw") semi-variogram is to describe the spatial structure for the entire study area. Model fitting is a rather subjective procedure, although it is possible to employ least-square techniques to judge the fit of the proposed model. Testing of the appropriateness of a model may be achieved using a cross-validation kriging procedure in which a series of known points are estimated using the various models, the models producing the best quality estimates being the ones deemed most suitable. In dealing with complex semi-variograms it may be necessary to use a nested model structure, using separate models to describe various stages of the spatial structure. A number of model types are commonly encountered in the analysis of environmental data. They may be conveniently divided into bounded and nonbounded groups. Bounded models are those in which the semi-variance reaches a definite sill; nonbounded models are those where the semi-variance does not.

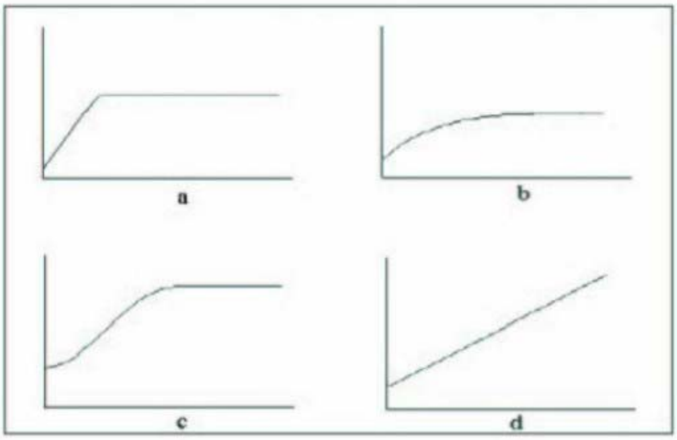

Figure 2. Theoretical semi-variogram models: (a) spherical, (b) exponential, (c) gaussian, (d) linear

A commonly used model in environmental geostatistics, the spherical model, is indicative of a high degree of spatial continuity. It is linear near the origin before flattening out as it approaches the sill at $a$ (the range of influence). It is described by the equation:

$$\gamma(h) = C_0 + C\{1.5h/a - 0.5(h/a)^3\}\text{, when } h < a \qquad (4)$$

$$\gamma(h) = C_0 + C\text{, when } h \geq a \qquad (5)$$

where $a$ is the range, $h$ is the lag distance, $C_{\mathrm{O}}$ is the nugget, $C$ is the sill minus $C_{\mathrm{O}}$, and $\gamma(h)$ is the semi-variogram value.

A useful rule of thumb is that the tangent at the origin reaches the maximum semi-variance at a distance of approximately $2/3\,a$.

The exponential model approaches the sill asymptotically and is described by the equation:

$$\gamma(h) = C_0 + C\{1 - \exp(-h/a)\} \qquad (6)$$

Due to the asymptotic nature of the ascension, calculation of the range is usually accomplished by determining the lag distance at which 95% of the sill value is attained.

Gaussian models are used to model extremely continuous variables. Resembling a spherical model except for the parabolic nature of the ascension at short lags and an asymptotic approach to the sill, it is described by:

$$\gamma(h) = 1 - \exp(-3h^2/a^2) \qquad (7)$$

The linear model is a simple unbounded model described by:

$$\gamma(h) = C_0 + mh \qquad (8)$$

where $m$ is the slope.

In practice, many data sets are based on irregularly spaced data for which it is not possible to obtain pairs of points separated by exact distances. In these cases it is usual to prescribe lag tolerances which are often the lag distance plus or minus a certain fraction of the lag distance. The reason that simple interpolation between points on the empirical semi-variogram is not used relates to the constraint that solutions of the kriging equations, encountered further on in the estimation process, are required to be both positive and unique. This constraint limits the models that may be applied to the empirical semi-variogram to those which are known as positive definite. Although a full description of the condition of positive definiteness is beyond the scope of this article, the constraint is necessary to ensure the mathematical stability of the kriging process.

Although the semi-variogram is the most common measure of spatial correlation encountered, others do exist and are usually based on the variation in some parameter between pairs of samples as a function of the distance separating the points. Time series analysts often use a correlogram which utilizes a normalized covariance of the data and/or a madogram which employs the mean absolute difference between the sample points. Modifications of the madogram may utilize the median of the differences or the median squared difference.

## 5. Semi-Variogram Modeling for Environmental Data

As construction of a semi-variogram is vital for any geostatistical study, it is worthwhile considering a number of points that are relevant to the production of the semi-variogram for environmental data sets. The subject of perhaps the most attention is the sampling scheme adopted. Data sets can be typically divided into two groups on this matter: those designed specifically with a geostatistical approach in mind and those in which geostatistical analysis methods were decided upon after the sampling scheme had been designed and samples taken. The former data set will exhibit a scheme that may incorporate two specific sampling stages, the first to construct the experimental semi-variogram and the second (which may or may not include the data of the first stage) designed for the optimal implementation of the estimation procedure based on the information provided by the semi-variogram. The latter data set, in which geostatistics has been adopted *a posteriori*, may feature any of the combinations of random, systematic or stratified sampling schemes commonly encountered in environmental monitoring.

For environmental monitoring, it is advantageous to have a thorough knowledge of the source, movement, and fate of the pollutant of interest to allow for optimal design of sampling schemes for semi-variogram calculation. Initial exploratory sampling and investigation of the study area provide valuable information prior to the actual sampling of the site. The literature provides many discussions of the "best" sampling schemes for construction of the semi-variogram, but adoption of any one should be made only after consideration of a number of factors including cost, practicability, and the variable of interest. As the goal of the sampling is the description of the omni-directional and directional semi-variograms, the basic sampling design is often a radial design, samples being taken at intervals along a number of transects with orientations of 45° relative to one another, allowing for calculation of the directional semi-variograms. The number of samples to be taken along each transect is a function of the range of influence. The distance between the samples must be small enough to allow for the ascending portion of the semi-variogram to be described accurately and a logarithmic spacing is often adopted. As the range of influence is unknown, a literature survey may provide information pertaining to the approximate range as calculated in other studies. Alternatively, varying sample spacing along each transect may be employed, the sample spacing increasing with increasing distance from the center of the confluence of the transects. The maximum distance between samples does not need to be greater than approximately one half of the maximum lag distance allowed by the boundaries of the proposed study area. The low number of pairs of samples separated by greater than this distance presents the possibility that the semi-variance values calculated for these lags are less reliable than for those at lag distances where a greater number of pairs have been used. The positioning of the transects within the study area is of some importance, being typically centered on the perceived pollution source, or, in the absence of a source, at the center of the study area. Some consideration should be given to factors such as the mode of transport of the pollutant or variable (water movements, colloidal transport, etc.) and the pollution source (aerial deposition, point source, etc.). The transects do not have to reach the site boundaries and should not be placed subject to where it is assumed high or anomalous variable values will occur, as the purpose of this sampling stage is not the location of regions of elevated variable values. Transects have been found to yield noisy semi-variograms due to the inherently noisy nature of environmental variables and combinations of transect sampling and regular grids have been suggested as possible remedies.

Systematic grid sampling for semi-variogram construction is often encountered where the samples are placed at regular intervals either over the entire survey area or within a defined part. This approach offers some advantages, as lag distances are regularly spaced and the initial sampling plan may be incorporated easily into secondary sampling for estimation purposes. If deviations from the regular grid are made for any reason, it is important to note the polar coordinates for the relevant samples to allow these samples to be incorporated into the semi-variogram calculations (for an irregular sampling scheme it is necessary to obtain coordinates for all samples). Irregular sampling patterns may be implemented where imposition of a structured grid is impossible. This approach can be used to avoid locations where noise may be introduced into the semi-variogram as a result of secondary pollution sources. If an irregular sampling pattern is used, the definition of lag and direction tolerances is necessary in order to obtain enough points for reliable calculation of the semi-variance. Aside from sample spacing for semi-variogram calculation, another consideration of some importance is the concept of the sample "support." The sample support describes the physical attributes of the sample. For soil sampling, the volume and shape of a soil core, or the mass of a soil sample, are used to describe the support. The support should be chosen carefully, bearing in mind its relation to the perceived pollution source. A typical example is sampling to construct a semi-variogram for the aerial deposition of a pollutant from a stack. All samples in this case should be taken from the same depth or otherwise a variance will be introduced as a result of changes in the sample support.

Sampling schemes for empirical semi-variogram calculation can influence the appearance of the semi-variogram in a number of ways. Variations in the nature of the support may superimpose noise onto the semi-variogram structure, making an assessment of the sill more difficult. The appearance of a totally random spatial structure (typified by a semi-variogram where the nugget value, $C_0$, is equal to the sill) may indicate that the sample spacing has been too large to capture the spatial structure at shorter lag distances. This situation is easily rectified by resampling at shorter distances to delineate the short-range structure. Although the appearance of a noisy semi-variogram can make model fitting more difficult, there are methods by which the semi-variogram may be smoothed to make the underlying structure somewhat clearer. Investigation of the number of pairs that constitute the semi-variance for each lag as well as the individual contribution of each pair to the overall semi-variance provides information as to possible reasons for a noisy structure. Low numbers of pairs (< ~40–60) can be corrected for by increasing the lag tolerance, and individual sample pairs may be removed to observe their effects on the semi-variogram value. Directional semi-variograms tend to exhibit more noise than omni-directional semi-variograms due to reduced pair numbers for each lag so it is usually prudent to model the omni-directional semi-variogram prior to the directional.

-
-
-

**Bibliography**

AI-Geostats. <www.ai-geostats.org [This website is a complete resource for geostatistical material, software, conferences, and texts.]

Clarke I. (1979). *Practical Geostatistics*, 129 pp. London: Elsevier. [This book is an introduction to geostatistics with emphasis on its application to ore reserve estimation.]

Cressie N. (1993). *Statistics for Spatial Data*, 928 pp. New York: Wiley. [This useful text covers many aspects of spatial statistics and their application.]

David M. (1977). *Geostatistical Ore Reserve Estimation*, 364 pp. Amsterdam: Elsevier. [This authorative text is on geostatistics from a mining perspective.]

Englund E. and Sparks A. (1988). *GEO-EAS: Geostatistical Environmental Assessment Software*. Las Vegas: US Environmental Protection Agency Report. 600/4-88/033. [This semi-variography and kriging software has become the benchmark.]

Gilbert R.O. (1987). *Statistical Methods for Environmental Monitoring*, 336 pp. New York: Van Nostrand Reinhold. [This book is an excellent introduction to statistical methods, including geostatistics, for environmental monitoring.]

Isaaks E.H. and Srivastava R.M. (1989). *Applied Geostatistics*, 561 pp. Oxford: Oxford University Press. [This excellent text covers all aspects of spatial data analysis.]

Journel A.J. and Huijbregts C.J. (1978). *Mining Geostatistics*, 600 pp. London: Academia Press. [This is a definitive text on geostatistics.]

Krige D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Minining Society of South Africa* **52**(6), 119–139. [This paper describes the origin of geostatistics.]

Matheron G., 1971, The theory of regionalised variables and its applications, Les Cahiers duCentre de Morphologie Mathematique, Vol. 5, Ecole des Mines de Paris, Fontainebleau, 212p [This text is definitive on the theory of regionalized variables.]

Pannatier Y. (1996). *Variowin 2.2: Software for Spatial Data Analysis in 2D*. New York: Springer-Verlag [This is a definitive software and text for semi-variography.]

Myers D.E. (1982). Matrix formulation of co-kriging. *Mathematical Geology* **14**, 248–257. [This paper provides a mathematical description of co-kriging.]

Rendu J.M. (1979). Normal and lognormal estimation. *Mathematical Geology* **11**(4), 407–422. [This paper is a mathematically detailed description of kriging with lognormal distributions.]

Webster R. and Burgess T.M. (1980). Optimal interpolation and isarithmic mapping of soil properties. III. Changing drift and universal kriging. *Journal of Soil Science* **31**, 505–524. [This is a useful article on universal kriging.]

**Biographical Sketches**

**Mark Dowdall** completed his PhD at the Institute of Technology, Sligo, Ireland, in 2000 and is currently engaged in radioecological research in the Environmental Protection Unit of the Norwegian Radiation Protection Authority. His research interests include radioecology and advanced statistical methods as applied to the analysis of environmental/radioecological data.

J**ohn O’Dea** has spent his career in physics education both as a lecturer of undergraduate physics in Ireland and a trainer of physics teachers in Africa. He now teaches Environmental Physics in the Institute of Technology in Sligo, Ireland. Since coming to Sligo he has developed an interest in environmental radiation and has published a book, *Exposure: Living with Radiation in Ireland*, and a number of scientific papers on environmental radioactivity in cooperation with Mark Dowdall. Other publications have included topics such as air pollution, instrumentation, and science education.