Glossary

A

advanced geospatial methods
Methods that include an explicit spatial correlation model. These methods may also include spatial trend and statistical error components. Advanced methods are also known as geostatistical methods.

Akaike information criterion
A method for determining the quality of an estimated statistical model and for selecting the best model from a set of candidate models. The selected model is the one for which the predicted difference between the model and the truth is minimized.

anisotropy
The degree of spatial correlation is dependent on direction, typically assessed using directional variograms, covariograms, or correlograms.

autocorrelation
A quantifiable relationship between sampled observations as they relate to one another in space and/or time. This relationship is expressed as a function of sampling distance or temporal interval.

B

bandwidth parameter
A parameter in some regression models that helps fit the model to the sampled data.

Bayesian information criterion
A method for selecting the best model from a set of candidate models; these candidate models may have different numbers of parameters. For a model with k free parameters and a set of data {x} containing n points, with P(x|k) the probability of the observed data and L the maximum likelihood function, the BIC may be expressed as: $-2\ln p(x|k) \approx BIC = -2\ln L + k\ln(n)$, where the model with the lowest BIC is predicted to be the best one. The BIC is derived assuming that the data distribution is exponential (Schwarz 1978; Zhao 2013).

bins
For plotting data in a histogram, groups of data within a range of values are plotted on the x-axis, the height of each bar plot corresponds to the number of values in that group. For a variogram, grouping of data point pairs at similar lag distances. Each lag bin ideally should contain 20–30 sample pairs.

breakline
A defined line with X, Y, and Z values at each vertex.

C

convex hull
Boundaries of data location.

coregionalization
Scale dependence in multivariate variogram modeling.

correlogram
A plot of the correlation between pairs of data points versus lag.

covariance
A quantitative measure of how two sets of observations, on average, vary together around their respective means.

covariance function
An expression of the extent to which two variables change together.

covariogram
A plot of covariance versus lag.

cross-validation
A geospatial model assessment method implemented by eliminating one observed value at a time from the data set, using the model to calculate a predicted value at that point, and then comparing the predicted value with the observed value.

cross-variogram
A spatial correlation model used when there are secondary data. The empirical variogram of the correlation between two variables.

D

detrend
Remove the trend component of the data set, usually based on other information or secondary data.

drift
In geostatistics literature, spatial trend is called drift.

E

edge effects
In models of physical systems, the manner in which the boundaries of the study area are handled can have an impact on the calculations. If the influences across a boundary are not incorporated correctly the interpolation results will have errors at the edges of the modeled area.

empirical variogram
Also called the experimental variogram or the sample variogram. A plot of half of the mean of the squares of the differences in measured values grouped by lag function of distance.

exact interpolator
An interpolation method that produces values exactly equal to observed values at all measurement locations.

F

fast radial basis functions
See radial basis functions (RBFs). A reference to the fast interpolation enabled by the use of RBFs.

first statistical moment
An estimate of the central tendency of a sampled population.

G

gamma
The semivariance value used to plot the y-axis of a variogram.

Gaussian anamorphosis
A method to transform a raw variable to a Gaussian (normally distributed) variable.

genetic algorithm
A mathematical procedure used in optimization to find possible solutions. Genetic algorithm is a heuristic optimization technique that mimics the concept of natural evolution and mechanisms such as crossover, mutations, and survival of the fittest. The genetic “code” for a given possible solution is composed of various model decision variables expressed in binary digits. A “population” of different solutions with different genetic codes are allowed to interact as a way to explore the universe of solutions to find an “optimal” solution.

geospatial analysis
Process of compiling and analyzing data related in time or space.

geospatial data
Data that are referenced in 2-D (x and y) or 3-D (x, y and z) space and/or time (t); where x, y and z represent spatial coordinates (e.g. latitude, longitude and depth, respectively) and t represents a specific time of sampling.

geospatial methods
Spatial or temporal analytical methods used to estimate values (such as concentrations) at unsampled locations or times. These methods require data with information about sampling locations and times. Some methods can also generate measures of uncertainty associated with the estimates.

geospatial support
Geometrical size, shape and spatial orientation of the units or regions associated with the measurements.

geostatistical methods
A category of geospatial methods that make statistical assumptions about the sampled population and use statistical metrics to predict estimated values or uncertainty in space or time.

global data variation
Systemic changes in data over relatively large temporal or spatial scales.

H

h-scatterplot
A plot of the measured values for pairs of data points. Also known as an h-scattergram or lagged scatterplot.

Hermite polynomials
A set of mutually orthogonal polynomials Hn(x) generated by Hermite's equation. Hermite polynomials are either all-even or all-odd, for instance: H0 = 1, H1 = x, H2 = x2 -1, H3 = x3 -3x, H4 = x4 – 6×2 +3.

homoscedasticity
The equality of variance among sets of data (Unified Guidance).

I

independent and identically distributed (i.i.d)
In traditional statistics, the assumption that each sampling location, regardless of how close they are sampled to one another, is independent, or not spatially or temporally correlated.

K

kernel
The kernel is a weighting function used in locally weighted regression methods.

kriging variance
A calculated value for the degree of confidence in the estimated values at unsampled locations. The kriging variance is calculated using the sampled values within the pre-defined search neighborhood.

kurtosis
A measure of whether the data are peaked or flat near the mean. High kurtosis would show a distinct peak near the mean and drop off rapidly to heavy tails (NIST/SEMATECH 2012). Thus in a populations with high kurtosis, the variance results chiefly from a small number of points with very large deviations. In a population with low kurtosis, the variance results from a larger number of points with small deviations.

L

lag
A parameter of a variogram. When sampling on a regular grid, it is the distance between samples. If the distance between samples is irregular, then the lag may be calculated as the average of the distances between the sampling locations. A sampled interval in time that is used to express temporal relationships between sample observations.

linear model of coregionalization
In kriging and simulation methods a model of spatial correlation that involves more than one variable. Multivariate

spatial autocorrelation.

M

mean
The arithmetic average of a sample set that estimates the middle of a statistical distribution (Unified Guidance).

mean prediction error
The mean of the differences between the predicted values and the true values in a distribution.

mean squared error
The average of the squares of the errors.

more complex geospatial methods
Regression methods with no spatial correlation model. These methods include spatial trend and statistical error components.

N

nugget
An extrapolation of the sample variogram to a lag of zero. The nugget represents measurement errors or spatial variation at distances smaller than the sampling interval. In a variogram, the fact that the y-value is non-zero at an x-value of 0 is sometimes called the nugget effect.

P

prediction standard error
Square root of the prediction variance.

prediction variance
Provides a measure of uncertainty and shows where additional data collection would be most useful. Also called the standard error.

proxy data
Quantitative data such as field data that that can be used to supplement the core laboratory data being examined, data such as membrane interface probe data supplementing VOC data.

R

radial basis function
A basis function that depends only on the distance to the origin.

range
The distance after which the variogram values remain at or close to the sill value.

raster data
The result of spatial interpolation with a computer program that converts discrete point data (for example, monitoring well water level elevations or contaminant concentration) to a continuous grid of predictions with at least one value associated with each grid cell.

realization
In geostatistical modeling, a realization is one model of spatial values prepared using specific values drawn from the statistical distribution of possible values determined by a site-specific spatial correlation analysis. Multiple realizations can be prepared from a given spatial correlation analysis. Also, an independent response produced by a mathematical model.

root mean square error (RMSE)
The difference between predicted values and actual values is the error (or residual). To calculate RMSE, square each error, take the average, then take the square root. (Institute for Statistics Education 2016c)

root mean square standardized error
The root mean square error divided by the error variance.

S

sample extent
The observation domain or area of characterization. It can be defined by the spatial boundary of a site or the duration of sampling.

sample interval
The sampling distance or frequency data are collected. A sampling interval can be regular (for example, equal-spaced sampling grid or time intervals) or irregular (for example, nested sampling intervals or time steps).

sample support
The larger mass, length, area, or time represented by a smaller sample or group of composite samples.

sampling optimization
Improving the spacing, timing and number of sample observations to adequately and defensibly characterize a site. Sampling optimization is geared towards optimizing the cost, time and manpower associated with field sampling efforts.

search neighborhood
The search neighborhood defines the area over which data points are considered when interpolating a value at a new location.

second statistical moment
One measure of the spread or dispersion of sampled observations around the mean.

secondary correlated data
Both qualitative information such as the location of a physical barrier or soil type, and quantitative data such as concentrations of another contaminant or mineral, or quantitative data (proxy data) collected by a different method from the primary data being modeled.

secondary stationarity
Assumes the mean and covariance of a population are consistent over space and time.

semivariance
Also known as the standard error. One-half the squared difference in values between pairs of sampling points.

semivariogram
Value calculated based on the absolute difference between the sample observations separated by the lag.

sill value
A variogram is a plot of the squared differences between measured values as a function of distance between sampling locations. The sill value is the point at which the variance of the differences increases until the spatial autocorrelation is no longer present.

simple geospatial methods
Methods that are computationally simple, so that large data sets can be efficiently mapped. These methods have no spatial correlation or statistical error model and only require that the data are related in space or time, or both.

skewness
A measure of asymmetry of a data set (Unified Guidance).

smoothing interpolators
Another term for inexact interpolators, because they produce smoother surfaces with fewer discontinuities that better reflect the spatial correlation of the broader data set.

spatial correlation
A relationship between a measured factor such a mineral concentration and its location.

spatial trend
The differences in values between points across a sampled area. The spatial trend represents local average of the data as a function of the location. Large-scale variation.

spatial variogram
A measure of how the data are related by distance, a spatial variogram is a plot of the differences between measured values as a function of distance between sampling locations (see variogram).

splines
The result of a function which creates a contour between sampled data points.

stationarity
The assumption that the statistical characteristics of a population do not change over time or space. That is the statistical properties of a distribution, such as the mean and variance, are translation invariant.

stationary
A distribution whose population characteristics do not change over time or space (Unified Guidance).

statistical significance (GSMC-1 glossary)
Statistical difference exceeding a test limit large enough to account for data variability and chance (Unified Guidance). A fixed number equal to alpha (α), the false positive rate, indicating the probability of mistakenly rejecting the stated null hypothesis (H_0) in favor of the alternative hypothesis (H_A). Or, the p-value sufficiently low such that the analyst will reject the null hypothesis (H_0).

T

temporal variogram
A measure of how the data are related by time, a temporal variogram is a plot of the differences between measured values as a function of time between sampling events.

V

validation
A geospatial model assessment method that is implemented by dividing the observed data set randomly into two data sets and then using each set to calculate predicted values for the other set.

variance
The square of the standard deviation (USEPA 1989); a measure of how far numbers are separated in a data set. A small variance indicates that numbers in the data set are clustered close to the mean.

variogram
A plot of the variance (one-half the mean squared difference) of paired sample measurements as a function of the distance (and optionally the direction) between samples. Typically, all possible sample pairs are examined, distance and directions. Variograms provide a means of quantifying the commonly observed relationship that samples close together will tend to have more similar values than samples far apart (USEPA 1989). A graphical tool used in geospatial analysis.

variogram cloud
A plot of the semivariance of all pairs of data points (y-axis) as a function of the lag (x-axis).

Z

z-score
The observed value of the z statistic (Stark 2013).

z-statistic
A test statistic whose distribution under the null hypothesis has an expected value of zero and can be approximated by the normal distribution (Stark 2013).