Interpolation Methods and Model Prediction

Sampling measurements made at discrete points, such as measurements of contaminant concentrations, can be used to build a model for the whole site. Different methods are available to make models for contaminant concentrations at all points within the site. Simple, more complex, and advanced interpolation methods can support these models, provided that appropriate data requirements are met.

▼*Read more*

Sample points and monitoring points are individual (discrete) points collected at a particular time and at a particular place (point). Because all points in time and space cannot be sampled, values for nearby unsampled points are inferred based on the data collected. Geospatial methods offer systematic approaches to fill in the gaps in between sampling locations. Before interpolating or making predictions, an appropriate geospatial method must be selected based on the characteristics of the available data and the project requirements. See Work Flow and Flow Charts for guidance on selecting methods. Many different interpolation approaches can be used, ranging from simple methods that are easy to apply, to more complex or advanced methods that require significant effort to estimate the parameters used by the method. Ideally, the interpolation approach would be both easy-to-implement and accurate. Practitioners often must compromise, however, to find an approach that is both useable and realistic.

Categories of Geospatial Interpolation Methods

The major geospatial methods can be grouped into three categories depending on their complexity: simple, more complex, and advanced. These methods are presented in Table 3.

Table 3. Organizing geospatial interpolation methods

	Simple	**More Complex**	**Advanced**
Description	Mechanistic methods with no statistical error model	Regression methods with no spatial correlation model.	Extension of regression methods including a spatial correlation model. Also known as geostatistical methods.
Data Requirements	Works best with larger data sets	Works best when other predictor variables are available.	Works best when sufficient data are available to estimate correlation model.
Statistical Assumptions	None	Regression residuals are spatially uncorrelated and normally distributed (after transformation).	Residuals after trend removal are stationary and normally distributed (after transformation).
Provide Prediction Uncertainty	None	Yes. Prediction standard error or variance provides measure of uncertainty. True uncertainty is higher when the residuals are spatially correlated.	Yes. Prediction standard error or variance provides measure of uncertainty. True uncertainty is higher because standard error does not include uncertainty from estimation of model parameters.
Example Methods	Inverse distance weighting (IDW), natural neighbors, Thiessen polygons	Parametric and nonparametric regression; local linear regression; splines and kernel methods	Kriging, conditional simulation*
*Conditional simulation provides a set of possible interpolated values that can be used to estimate the probability distribution of predictions at each location			

Simple methods provide a quick, simple approach to modeling spatial data. First, simple methods are conceptually simpler because they require almost no assumptions be made about the data or the variable to be mapped, only that the sampled

data relate to one another in space or time, or both. Second, these methods are computationally simpler, so that large data sets can be efficiently mapped. Because the simple methods do not impose much structure on the data, they generally require more data in order to produce a good interpolation. Third, these methods are statistically simpler because they do not include a statistical error component that can be used to estimate the error in the predictions at unsampled locations.

In contrast, more complex and advanced methods make assumptions about the statistical distribution of a sampled population and can provide estimates of prediction error. This error derives from the inability to sample everywhere, and from the inability of the method to match the data. More complex methods include regression type methods, which are based on statistical assumptions about the data but do not include an explicit representation of spatial correlation. See GSMC-1, Section 5.5.1 for information about linear regression. The advanced methods, including kriging and conditional simulation, are based on more stringent statistical assumptions and include an explicit model of spatial correlation. The advanced methods are also commonly known as geostatistical methods.

The output of every geospatial interpolation method is the set of interpolated values for unsampled locations of interest. The inputs to the method are the observed data. The method can be thought of as a mathematical equation that converts the sampled inputs into the interpolated outputs. This equation or model of the data will be complicated in some cases, but the mathematical details are handled by the software and do not need to be fully understood by the user. One way of thinking about how the geospatial interpolation methods differ is to distinguish between the different components (spatial trend, spatial correlation, error) that are represented within the methods. The different components for the categories of methods are illustrated here:

Simple Method	= spatial trend component
More Complex Method	= spatial trend component + (secondary data component) + error component
Advanced Method	= spatial trend component + spatial correlation component + error

Simple Methods ▼*Read more*

For simple methods, the model used to interpolate the sample data consists of only one component—a simple spatial trend component. The spatial trend is calculated using a fixed formula or algorithm and represents the spatial average value of the data. The spatial average estimate can be weighted, meaning that sample observations closer to the location being estimated holds more weight than other, more distant, sample locations. Spatially weighted averages use a moving window interpolator (for example, IDW, natural neighbor) to generate spatial estimates. In other cases, the average can represent an area or volume based on boundaries defined by the user (for example, Voronoi polygons). Simple geospatial methods do not make any assumptions about the distribution of the sampled data and provide no measure of uncertainty of the interpolated values.

More Complex Methods ▼*Read more*

More complex methods are regression methods, which include spatial trend and error components. The spatial trend component represents the local average of the data as a function of the spatial coordinates or location. The error component allows the method to provide a measure of prediction uncertainty. Regression methods can also include an optional component for secondary data (such as distance from a source, when the primary data are concentrations), which would include additional explanatory variables that are correlated with the primary variable of interest.

For example, a polynomial regression model might be defined as follows:

Z = a1*x + a2*x^2 + b1*y + b2*y^2 + e

In this model, the parameters (a1, a2, b1, b2) are estimated from the data, and e is the error term, which is modeled as a spatially uncorrelated random variable. There are many variations on the regression model. These models can be categorized as parametric regression (global regression) or nonparametric regression (local regression). Global techniques calculate predictions using the entire data set and fit the data to a parametric function, such as the polynomial given in the example. Local techniques calculate predictions from the measured locations within subregions or neighborhoods, which are smaller spatial areas within the larger study area. Local nonparametric regression methods are more flexible than global methods and can represent the spatial trend of practically any process. Many local nonparametric regression methods are based on different ways of combining simple functions to achieve a flexible and smooth function. These methods are called nonparametric because they do not depend on the data fitting a particular predefined distribution, such as the normal (Gaussian) distribution. Due to their flexibility, nonparametric regression methods are able to capture practically all of the

variation in the data, producing excellent interpolated maps.

Since spatial regression models include a noise or error term, they can quantify uncertainty in the model predictions similar to a regular nonspatial regression model. In many cases, however, the predictions and estimates of uncertainty from spatial regression models may not be exactly correct, because in regression models it is assumed that the noise or errors are not spatially correlated. If the spatial trend component of the model has not captured all of the spatial correlation present in the data, then the error term is spatially correlated. This residual correlation violates one of the assumptions of the regression model that the noise or error terms are uncorrelated. As a result, the regression model tends to underestimate uncertainty in the predictions.

Advanced Methods ▼*Read more*

Advanced methods are also known as geostatistical methods. In the geostatistics literature, spatial trend is often called drift. The spatial trend component of the data model represents the overall trend over the area of interest. The spatial correlation component represents the fluctuation around this trend. The error component includes both measurement error and fluctuations at a scale that are too small to observe given the sample spacing (called microscale variation). There is no clear-cut division between what is represented by the spatial trend component and spatial correlation component. One approach might represent most of the data variation in the spatial trend component of the model, while another approach might use a simple model for the spatial trend component and represent most of the data variation using the spatial correlation component.

Interpolation using models of spatial dependence is referred to as kriging, after the South African mining engineer Danie Krige. Kriging models can be thought of as spatial regression models that include an extra component to account for the spatial dependence of the spatial correlation. The type of kriging used depends on the model used for the spatial trend:

- Simple kriging: the spatial trend is assumed to be a known constant.
- Ordinary kriging: the spatial trend is assumed to be an unknown constant estimated from the data.
- Universal kriging: the spatial trend is modeled using a simple polynomial of the spatial coordinates (regression).

Usually, the purpose of geospatial modeling is to predict the value of a variable of interest at a set of points that have not been sampled (called point prediction, or if kriging is being used, point kriging). In geostatistics, the term sample support is used to describe the larger mass, length, area or time represented by a smaller sample or group of composite samples. Most samples have point support. In some cases, however, the goal is to not to predict the variable at individual points, but rather to predict a function of the variable (such as the average) over a set of points or an area. Calculating the average over a set of areas using kriging is called block kriging. Block kriging allows a change from point support to block support.

If a nonlinear function of the predicted variable (such as the maximum) is of interest, or the area over which the variable exceeds some prescribed value, then a method such as indicator kriging or conditional simulation may be appropriate. Indicator kriging is a nonlinear, nonparametric form of kriging in which continuous variables are converted to binary (indicator) variables. Because it is a nonparametric method, indicator kriging can handle distributions of any kind and can handle nondetected concentrations. Conditional simulation is used to produce a series of randomly simulated predictions that match both the data points and the geospatial model. Conditional simulation is the only method that attempts to reconstitute the intrinsic heterogeneity of the sampled environment and can therefore provide a more robust measure of overall uncertainty in generating spatial predictions.

Measures of Uncertainty ▼*Read more*

Regression and kriging methods can also provide a measure of uncertainty for each prediction called the prediction variance or standard error. The prediction variance does not indicate the full extent of the uncertainty in the predictions because it assumes that the geospatial model parameters are known perfectly, when in fact they must be estimated from the data. The prediction variance is still useful, however, as a relative measure of uncertainty; it shows, for example, where additional data collection would be most useful.

Almost all geospatial methods (except conditional simulation) have a smoothing effect, meaning that the predicted surface is much smoother than the actual surface. As a result, the predictions tend to under-predict the high values and over-predict the low values. One of the benefits of conditional simulation is that the simulated surfaces are not overly smooth, which is particularly important when the focus of the investigation is on identifying values that are below or above a threshold.

More recent research in geostatistical prediction methods has focused on the development of methods that can account for the additional prediction uncertainty that results from estimating the model parameters. These methods are generally Bayesian methods, which assume that the model parameters are also random variables. These methods are extensions of regression and kriging that are called hierarchical models or model-based geostatistics. Bayesian methods (Cressie and Wikle 2011; Diggle and Ribeiro 2007) are not described in this document because these methods are rarely used for

remediation optimization.

Additional information is provided elsewhere in this guidance on sources of uncertainty and the measures of uncertainty for the different geospatial methods.