Estimating Concentrations Based on Proxy Data

Obtaining a spatially representative sample data set is important for characterizing complex environmental sites. Often, cost, time, and site access limitations can prevent an adequate sampling density needed to confidently characterize the extent, migration pathways, and source of a contaminant. Proxy, or secondary, sampling technologies, such as membrane interface probe (MIP), LiDAR, geophysics, portable x-ray fluorescence spectroscopy, and other similar proxy sampling technologies are becoming widely used to map sparse, more expensive direct sample observations (for example, soil samples or geologic borings). Proxy sampling technologies are attractive because they collect relatively inexpensive, rapid, and spatially dense data that can correlate with sparse direct observations. Consequently, these technologies can quickly and economically fill sampling gaps in the direct sampling design to generate spatial estimates or simulations. Note that proxy sampling technologies are not a substitute for direct sampling, but rather a support measure.

Proxy sampling technologies can be applied using two approaches:

1. In a phased sampling approach, proxy sampling technologies are used as a screening-level assessment to provide a first approximation of spatial heterogeneity of the site. This approach optimizes the subsequent direct sampling regarding sample number, interval, and location, by modeling the spatial heterogeneity in the proxy data set.
2. In a nonphased sampling approach, proxy and direct sampling technologies are implemented simultaneously.

Understanding the Results: ▼*Read more*

There are no simple geospatial methods that can use proxy information. The specific applications of appropriate more complex and advanced geospatial methods are described below.

Variography (EDA): If sampling is implemented in a phased approach, variogram analysis can be applied to the initial phase proximal sensing data to evaluate the sampled spatial heterogeneity of the site. If the data meet the following criteria, then spatial autocorrelation can be determined:

- A sill and range are present in the cross-variogram.
- As a rule of thumb, the nugget to sill ratio is <50%.
- Cross-validation statistics are acceptable.

The variogram model can be used to determine a defensible sample number and spacing for direct sampling implemented in the second phase. As a conservative rule of thumb, the sampling interval for direct sampling should be half the fitted range value. Based on the area and number of sampling depths, a total number of samples can be derived using the sampling interval. If anisotropy is present, then orient a subset of sampling locations to account for this directional variation.

Multiple Regression (More Complex Method): Multiple regression can be used to incorporate the proxy information into the estimate of the quantity of interest. An advantage of regression methods is that the proxy information can include all types of data, including a combination of categorical data (such as soil type) and quantitative data (such as MIP results). If a nonparametric regression approach is used, then a flexible surface based on splines or kernel methods can be used to implicitly model the autocorrelation of the primary data, in addition to the correlation between the secondary and the primary data. Because a fit to a spatial correlation model (such as a cross-variogram) is not necessary, it is more straightforward to use a regression method than an advanced method. However, the prediction error estimates from regression methods may not be as accurate as those from advanced methods such as kriging or co-kriging. An example illustrates the use of secondary information as explanatory variables. Regression outputs can be used to create visualizations (using ArcGIS, EVS, or other software) that can help to identify the extent of potential impact, source areas, environmental controls, and potential preferential migration pathways.

Kriging with External Drift (Advanced Method): An extension to the regression approach is to combine a regression model with an explicit spatial autocorrelation model. The regression model relates the secondary data to the primary data, and the spatial correlation model represents the spatial autocorrelation of the primary variable. This type of kriging is called kriging with external drift (trend).

Co-kriging (Advanced Method): Integrating densely sampled secondary measurements can lead to more consistent descriptions of more sparse primary measurements and help reduce the uncertainty in modeled estimates. If a spatial

correlation exists between the secondary and primary sampled data, then it is possible to fuse the two data sets using co-kriging. Spatial correlation can be tested using the cross-variogram. Using the following criteria, it is possible to determine if spatial structure exists in the data: a sill and range are present in the cross-variogram; as a rule of thumb (see Variograms), the nugget to sill ratio is <50%; and cross-validation statistics are acceptable. The cross-variogram is used to optimize the search neighborhood for generating co-kriged estimates. Determine if point or block co-kriging is appropriate; when the primary and secondary variables have different support, block co-kriging is the recommended approach. To generate spatial estimates at a finer resolution than the sampled resolution, use point co-kriging. To generate average spatial estimates or volumes, use block co-kriging. The co-kriging spatial estimates can be integrated into a digitized CSM (EVS, GIS, or other software platform) to guide interpretation of the extent of potential impact, source areas, and identify potential preferential migration pathways, for example. The co-kriging variance can be used to assess the uncertainty in the co-kriged spatial estimates.

Co-simulation (Advanced Method): Co-simulation generates multiple values for the primary variable at unsampled locations using both the primary and secondary sampled data. Co-simulation uses a probability simulation technique. While co-kriging only provides standard error estimates, co-simulation provides a more complete description of the uncertainty that can be used to generate exceedance probability maps and other descriptions of uncertainty. Integrating densely sampled proxy measurements can lead to more consistent descriptions of sparse direct measurements and can generate more realistic simulated uncertainties. Spatial correlation must be present between the direct- and proxy-sampled data, which can be tested using the cross-variogram. The data should be transformed so that it is normally distributed (Gaussian) prior to fitting the cross-variogram model. Spatial structure exists when a sill and range are present in the cross-variogram, the nugget to sill ratio is <50% (see Variograms), and cross-validation statistics are acceptable.

The cross-variogram is then used to optimize the search neighborhood for generating predictions and uncertainties. Determine a probability threshold value of interest, which may be related to an upper confidence limit, maximum concentration limit, or other remediation or human-health risk-driven threshold. Determine if point or block co-simulation is appropriate. To generate spatial mean estimates, uncertainties, or probabilities at a finer spatial resolution than sampled, use point co-simulation. To generate spatial mean estimates, uncertainties, or probabilities at a coarser spatial resolution than sampled, use block co-simulation.

The probability maps correspond to the risk of occurrence of contamination and can be used to localize areas of potential concern or remediation implementation. Spatial mean estimates can be used to map the extent of potential impact and pinpoint potential source areas. Maps of uncertainty can be used to optimize the placement of future sampling with the objective of reducing spatial uncertainty in simulated estimates and probabilities. In such cases, additional sampling targets areas of elevated uncertainty. Co-simulation outputs can be used to create visualizations (using ArcGIS, EVS, or other software) to help identify the extent of potential impact, source areas, environmental controls, risk of exposure, and potential preferential migration pathways.