Modeling Local Uncertainty accounting for Uncertainty in the Data

Olena Babak and Clayton V. Deutsch

Consider the problem of estimation at an unsampled location using surrounding samples. The standard approach to this problem is kriging. Kriging uses the spatial correlations provided by the variogram to calculate the weights of the sample values surrounding an unsampled location. The weights obtained from the kriging equations minimize the estimation variance and account for the spatial correlation between the surrounding samples and the estimation location (that is, closeness to the estimation location) and between sample themselves (that is, data redundancy). Kriging results in optimal estimation (in the case of a known variogram model) and provides a model for local conditional distributions; in the Gaussian framework, the kriging estimate and kriging estimation variance are exactly the mean and variance of the local conditional Gaussian distributions. Oftentimes, however, the exact sample data are not known due to measurement errors. In this case simple kriging can not be directly applied to infer the local conditional distributions. A theoretical framework for incorporating data uncertainty into calculation of the local uncertainty distributions is required.

Simple Kriging

The simple kriging estimator predicts the value of the variable of interest $z(\mathbf{u})$ at the estimation location $\mathbf{u}$ as a linear combination of nearby observations $z(\mathbf{u}_i), i = 1, \ldots, n(\mathbf{u})$, (Journel and Huijbregts, 1978):

$$z^*_{SK}(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) z(\mathbf{u}_i) + \left[1 - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})\right] m, \quad (1)$$

where m denotes the stationary mean, $\lambda = (\lambda_1(\mathbf{u}), \ldots, \lambda_{n(\mathbf{u})}(\mathbf{u}))^T$ denotes the vector of the simple kriging weights calculated from the normal system of equations

$$\sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Cov(z(\mathbf{u}_i), z(\mathbf{u}_j)) = Cov(z(\mathbf{u}), z(\mathbf{u}_j)), \quad j = 1, \cdots, n(\mathbf{u}), \quad (2)$$

where $Cov(z(\mathbf{u}_i), z(\mathbf{u}_j)), \quad i, j = 1, \cdots, n(\mathbf{u}),$ denotes the data-to-data covariance values and $Cov(z(\mathbf{u}), z(\mathbf{u}_j)), \quad j = 1, \cdots, n(\mathbf{u}),$ is the data-to-estimation point covariance values. The covariance function is calculated under stationarity through the semivariogram model $\gamma(\mathbf{h})$.

Simple kriging is the best linear unbiased estimator, that is, it provides estimates with minimum error variance $\sigma^2_{SK}(\mathbf{u})$ in the least square sense given by

$$\sigma^2_{SK}(\mathbf{u}) = C(0) - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Cov(z(\mathbf{u}), z(\mathbf{u}_j)), \quad (3)$$

where $C(0)$ is the stationary variance.

In the Gaussian framework the local conditional distributions are derived by simple kriging as follows

Uncertainty at the estimation location u is $Z(\mathbf{u}) \sim N(z^*_{SK}(\mathbf{u}), \sigma^2_{SK}(\mathbf{u}))$. (4)

Accounting for the Uncertainty in Data

Let us assume that each of the observations $z(\mathbf{u}_i), i = 1, \ldots, n(\mathbf{u})$, available for analysis was measured with some measurement error. Further assume that the measurement errors are distributed according to

Gaussian (normal) distribution; thus, uncertainty in each observation (random variable) $Z(\mathbf{u}_i), i = 1, \ldots, n(\mathbf{u})$ can be expressed as follows:

$$Z(\mathbf{u}_i) \sim N(\mu_i, \sigma_i^2), \qquad i = 1, \ldots, n(\mathbf{u}), \tag{5}$$

where μ_i and σ_i^2 denote the mean and variance of the uncertainty distribution in *i*-th data. For now let us assume that the observations $Z(\mathbf{u}_i), i = 1, \ldots, n(\mathbf{u})$, represent independent random variables, e.g., each data location was measured using different measurement tool.

When the observations are no longer assumed to be known, the mean of the local conditional distributions is a random variable. The variance of the local conditional distributions given in (4) is not a random variable because the simple kriging variance is homoscedastic (see 3). Because the mean of the local conditional distribution is a random variable, the uncertainty at the unsampled location **u** is described the following hierarchical model

$$\begin{aligned} Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u}) &\sim N(Z^*_{SK}(\mathbf{u}), \sigma^2_{SK}(\mathbf{u})), \\ Z^*_{SK}(\mathbf{u}) &\sim N(E[Z^*_{SK}(\mathbf{u})], Var[Z^*_{SK}(\mathbf{u})]), \end{aligned} \tag{6}$$

where

$$Z^*_{SK}(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Z(\mathbf{u}_i) + \left[1 - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})\right] m. \tag{7}$$

Note that distribution of $Z^*_{SK}(\mathbf{u})$ is Gaussian because it is a linear combination of Gaussian random variables. Furthermore, due to (7), the mean and variance of the distribution for $Z^*_{SK}(\mathbf{u})$ can be calculated as follows:

$$\begin{aligned} E[Z^*_{SK}(\mathbf{u})] &= E\left[\sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Z(\mathbf{u}_i) + \left[1 - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})\right] m\right] \\ &= \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) E[Z(\mathbf{u}_i)] + \left[1 - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})\right] m = \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) \mu_i + \left[1 - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})\right] m = \mu_{Z^*_{SK}(\mathbf{u})}; \end{aligned} \tag{8}$$

$$\begin{aligned} Var[Z^*_{SK}(\mathbf{u})] &= Var\left[\sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Z(\mathbf{u}_i) + \left[1 - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})\right] m\right] \\ &= Var\left[\sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Z(\mathbf{u}_i)\right] = \sum_{i=1}^{n(\mathbf{u})} \lambda_i^2(\mathbf{u}) Var[Z(\mathbf{u}_i)] = \sum_{i=1}^{n(\mathbf{u})} \lambda_i^2(\mathbf{u}) \sigma_i^2 = \sigma^2_{Z^*_{SK}(\mathbf{u})}. \end{aligned} \tag{9}$$

Thus, it follows that the local conditional distributions in the case of data uncertainty can be expressed using the following hierarchical model:

$$\begin{aligned} Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u}) &\sim N(Z^*_{SK}(\mathbf{u}), \sigma^2_{SK}(\mathbf{u})), \\ Z^*_{SK}(\mathbf{u}) &\sim N(\mu_{Z^*_{SK}(\mathbf{u})}, \sigma^2_{Z^*_{SK}(\mathbf{u})}), \end{aligned} \tag{10}$$

where $\mu_{Z^*_{SK}(\mathbf{u})}$ and $\sigma^2_{Z^*_{SK}(\mathbf{u})}$) are given in (8)-(9). Moreover, note that the mean and the variance of local conditional distributions are given by:

$$E[Z(\mathbf{u})] = E[E[Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u})]] = E[Z^*_{SK}(\mathbf{u})] = \mu_{Z^*_{SK}(\mathbf{u})}; \tag{11}$$

$$\begin{aligned} Var[Z(\mathbf{u})] &= E[Var[Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u})]] + Var[E[Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u})]] \\ &= E[\sigma^2_{SK}(\mathbf{u})] + Var[Z^*_{SK}(\mathbf{u})] = \sigma^2_{SK}(\mathbf{u}) + \sigma^2_{Z^*_{SK}(\mathbf{u})}. \end{aligned} \tag{12}$$

The shape of the local uncertainty in $Z(\mathbf{u})$ is Gaussian.

It worth noting that when the observations $Z(\mathbf{u}_i), i = 1,\ldots,n(\mathbf{u})$, do not represent independent random variables, but are correlated with a prescribed correlation structure, the mean and variance of the local conditional distributions can be calculated following the same steps as before expect variance of $Z^*_{SK}(\mathbf{u})$ needs to be calculated as

$$\begin{aligned} Var[Z^*_{SK}(\mathbf{u})] &= Var\left[\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})Z(\mathbf{u}_i) + \left[1-\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})\right]m\right] \\ &= Var\left[\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})Z(\mathbf{u}_i)\right] = \sum_{i=1}^{n(\mathbf{u})}\sum_{j=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})\lambda_j(\mathbf{u})Cov[Z(\mathbf{u}_i),Z(\mathbf{u}_j)] = \sigma^2_{Z^*_{SK}(\mathbf{u})}. \end{aligned} \quad (13)$$

Moreover note that the above derivations heavily rely on the assumption that the variogram model for the study domain is known; uncertainty in the data does not impact the assumption of the stationary variogram model.

Small Examples

Example 1: Consider the data configuration shown in Figure 1. In total, there are 4 conditioning data available for inference of the local conditional distribution at the unsampled location. All conditioning data are known subject to measurement errors; the distributions of the conditioning data are Gaussian with different means μ_i and variances σ^2_i, $i = 1,\ldots,4$, see Table 1 below. Study domain of size 10 by 10 units is assumed to be stationary; stationary mean and variance are 0 and 1, respectively. The variogram of the data is a single structured spherical with nugget effect of 0 and range of correlation of 10 units.

Table 1: Data locations and values

	Data 1	**Data 2**	**Data 3**	**Data 4**	**Unsampled Location**
X position	1	5	9	3	5
Y position	3	7	8	2	5
Value	$N(\mu_1,\sigma^2_1)$	$N(\mu_2,\sigma^2_2)$	$N(\mu_3,\sigma^2_3)$	$N(\mu_4,\sigma^2_4)$	?

We will vary the means and variances of the conditional data distributions to assess the impact of data uncertainty on the resulting local uncertainty distribution inferred from simple kriging. First, let us fix μ_i's as follows

$$\mu_1 = 0.8; \quad \mu_2 = 0.2; \quad \mu_3 = -0.4; \quad \mu_4 = -0.1;$$

and examine the effect of σ^2_i's. Table 2 show results for five different scenarios for σ^2_i's. Note that Table 2 shows only results for the variance of the local distribution of uncertainty accounting for data uncertainty, that is, $Var[Z(\mathbf{u})]$; this is because the mean of the local conditional distribution is independent of σ^2_i's and equal to 0.0884.

Table 2: Effect of σ^2_i's on the local uncertainty distribution.

	Case 1	**Case 2**	**Case 3**	**Case 4**	**Case 5**
σ^2_1	0	0.1	0.3	0.5	0.8
σ^2_2	0	0.2	0.4	0.6	0.9
σ^2_3	0	0.1	0.2	0.2	0.6
σ^2_4	0	0.3	0.4	0.4	0.7
$Var[Z(\mathbf{u})]$	0.4094 (σ^2_{SK})	0.5073	0.5871	0.6574	0.7915

It can be clearly noted from Table 2 that with an increase in the data uncertainty (that is, increase in the variance of the conditional data distributions), the variance of the local conditional distribution at the unsampled location increases. Moreover, when there is no uncertainty in the conditioning data; the variance of the local conditional distribution at the unsampled location is equal to simple kriging variance.

On the other hand, if we fix σ_i^2's as:

$$\sigma_1^2 = 0.8; \quad \sigma_2^2 = 0.2; \quad \sigma_3^2 = 0.3; \quad \sigma_4^2 = 0.4;$$

we can observe that with increase in the mean of the conditional data distributions, the mean of the local conditional distribution at the unsampled location increases, see Table 3.

Table 3: Effect of μ_i's on the local uncertainty distribution.

	Case 1	**Case 2**	**Case 3**	**Case 4**
μ_1	-0.8	-0.2	0.2	1
μ_2	-0.2	0.2	0.2	1
μ_3	-0.4	0.4	0.4	1
μ_4	-0.1	0.1	0.1	1
$E[Z(\mathbf{u})]$	-0.1933	0.1654	0.1765	0.9780

Note that Table 3 shows only results for the mean of the local distribution of uncertainty accounting for data uncertainty, that is, $E[Z(\mathbf{u})]$; this is because the mean of the local conditional distribution is independent of μ_i's and equal to0.5176.

It is worth noting that the results shown in Tables 2-3 were theoretically calculated from Equations (11)-(12). There is, however, another much more computationally intensive approach based on Monte Carlo simulation to obtain the same result. Specifically, in order to calculate the mean and variance of the local uncertainty distribution accounting for parameter uncertainty via Monte Carlo simulation approach the following steps need to be undertaken:

1. At each of the conditioning data locations draw a value from the conditioning data distribution using Monte Carlo simulation approach;
2. Apply simple kriging to calculate the mean and variance of the local conditional distribution using the conditional data generated in 1;
3. Draw a value from the local conditional distribution obtained in 2. Add to the database;
4. Repeat steps 1-3 many times, say 20000.

To show the equivalence of the theoretically derived local conditional distributions of uncertainty and the ones obtained using Monte Carlo simulation, let us repeat analysis of Table 2. Results are shown in Tables 4.

Table 4: Theoretically-derived approach vs. Monte-Carlo simulation: Variance of the local uncertainty distribution

$Var[Z(\mathbf{u})]$	**Case 1**	**Case 2**	**Case 3**	**Case 4**	**Case 5**
Theory	0.4094	0.5073	0.5871	0.6574	0.7915
Simulation	0.4071	0.5078	0.5884	0.6545	0.7974

The results of theoretically-derived approach vs. Monte-Carlo simulation approach match perfectly; the difference between results of both approaches could have been even smaller if instead of 20000 drawings in Monte-Carlo approach 100000 or more were used.

Example 2: To further understand the influence of the data uncertainty on the local conditional distributions at the unsampled locations, let us asses the change in the variance of the local conditional distributions

(accounting for data uncertainty) over the study domain. Let us consider the same data configuration as before; set the means of the conditioning data distributions at:

$$\mu_1 = 0.8; \quad \mu_2 = 0.2; \quad \mu_3 = -0.4; \quad \mu_4 = -0.1;$$

and consider three different cases, that is, case 3, case 4 and case 5, for σ_i^2's, see Table 2. In present study let us also consider two different variogram models, both single structured spherical with nugget effect of 0, but one with range of correlation equal to 10 units and the other one with a range of 5 units and let us compare results.

Figure 2 shows results obtained in each case. It can be clearly noted from Figure 2 that with increase in the range of continuity, the variance of the local conditional distributions decreases. The variance of the local conditional distributions is usually lies in the interval from 0 to 1. However, it can be also higher than 1, see Table 5.

Table 5: Maximum variance f the local conditional distributions over the study domain.

Maximum $Var[Z(\mathbf{u})]$	**Case 3**	**Case 4**	**Case 5**
Range 5	1	1	1.0021
Range 10	0.9967	0.9993	1.0446

Conclusions

In this paper a new interesting framework for incorporation of the data uncertainty into geostatistical estimation is presented. The theory behind the methodology was developed in detail; theoretical results were compared with practical results obtained via direct Monte Carlo simulation. Two small examples illustrating the change in the local uncertainty when incorporating data uncertainty were presented.

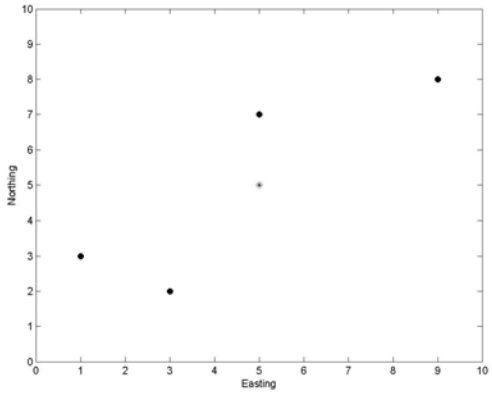

Figure 1: Data configuration for Examples 1.

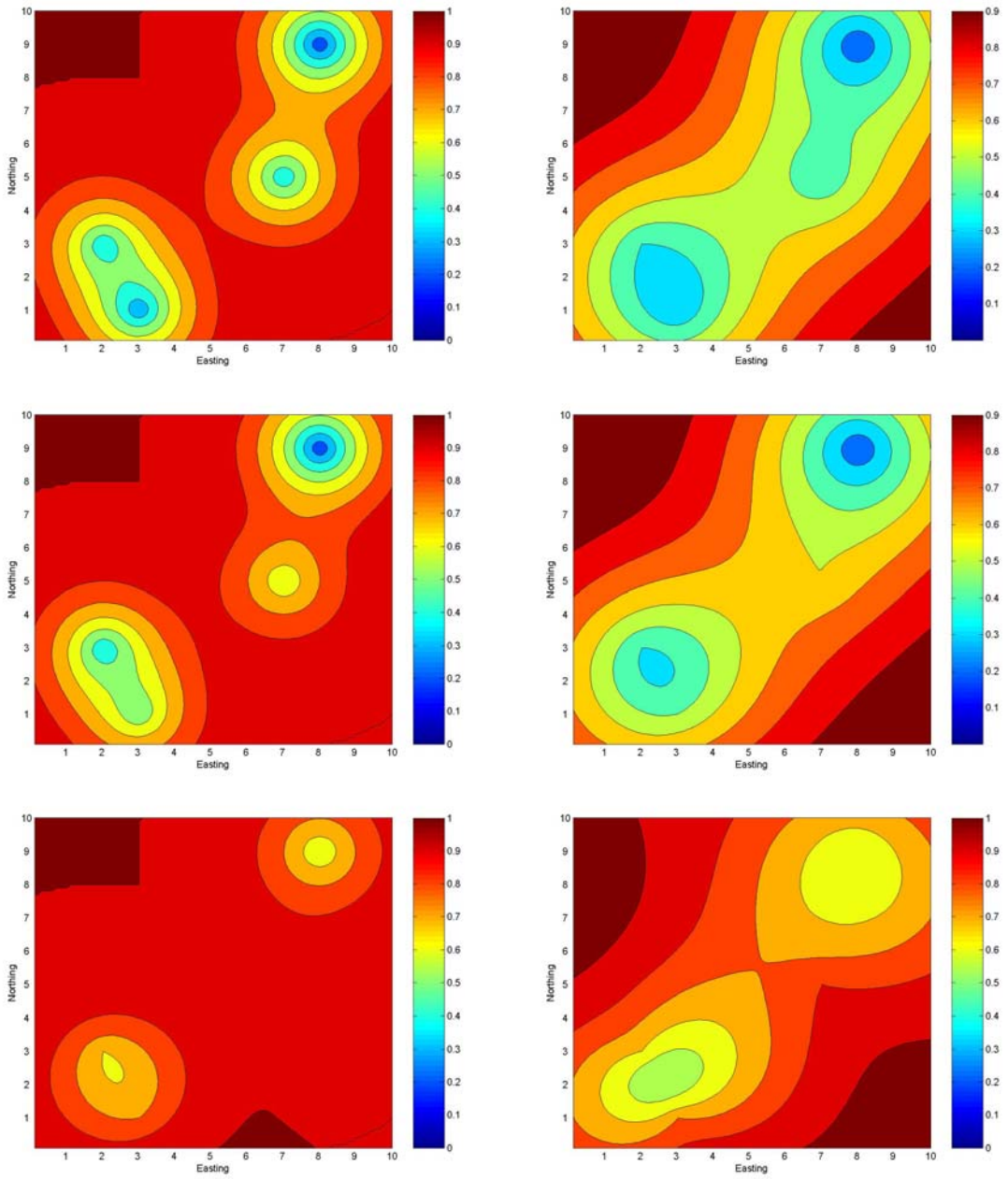

Figure 2: Variance of the local conditional distributions accounting for the uncertainty in the data obtained based on a single structured spherical variogram with nugget effect of 0 and range of continuity 5 (left) and 10 (right) : case 3 (top), case 4 (middle) and case 5 (bottom).