Methodology for Calculating Uncertainty versus Data Spacing

Brandon J. Wilde and Clayton V. Deutsch

Modeling spatial variables involves uncertainty. Uncertainty is affected by the degree to which a spatial variable has been sampled; decreased spacing between samples leads to decreased uncertainty. The reduction in uncertainty due to increased sampling is dependent on the spatial distribution of the variable being modeled. A densely sampled erratic variable may have a level of uncertainty similar to a sparsely sampled continuous variable. A methodology for determining the relationship between data spacing and uncertainty for any spatial variable that is characterized by a stationary variogram is presented. Uncertainty can be expressed in many ways. A number of uncertainty measures have proven useful within geostatistics. The properties and applicability of various uncertainty measures are discussed. An example implementation of the proposed methodology is presented.

1. Introduction

Spatial sampling design has been addressed by a number of authors (McBratney *et al.* 1981; Aspie and Barnes, 1990; Webster and Oliver, 2007). Many methodologies and objective functions have been applied to determine the optimum quantity and locations of samples. Some work has focused on determining the optimum spacing between samples (Deutsch and Beardow, 1999; Boucher *et al*., 2004). This has necessitated choosing an acceptable level of uncertainty. This work presents a methodology for evaluating the relationship between data spacing and uncertainty. This allows the practitioner to consider many data spacings and observe the effect of changing data spacing on uncertainty. No effort is made to define an acceptable level of uncertainty. The methodology could be applied at a greenfield stage to aid a decision regarding data spacing. It could also be applied at a mature stage to assist in the determination of an acceptable level of uncertainty.

The methodology is based on simulation. Sequential Gaussian simulation (SGS) is a popular method for characterizing uncertainty in earth sciences modeling. It allows for the generation of multiple equi-probable realizations that each honor the input data, histogram, and variogram when implemented correctly. Where and by how much the realizations differ provides a measure of uncertainty about the phenomenon being modeled (Journel and Kyriakidis, 2004). The simulation methodology can be extended to evaluate how uncertainty is related to data spacing. A methodology is proposed to evaluate uncertainty for different data spacings which can be applied to a variety of earth science variables.

2. Proposed Methodology

SGS is used to generate realizations of the spatial distribution of $z(\mathbf{u})$, $\mathbf{u} \in A$. Sample data are drawn from these realizations and SGS is then used to generate realizations conditioned to these resampled data. The simulated values are assumed to have the same support as the sample data. Uncertainty at this support is of little practical relevance; typically the assessment of uncertainty at some block scale is required (Journel and Huijbregts, 1978). The conditional realizations are block averaged to some scale of interest and measures of uncertainty are calculated from these block-averaged realizations. Specifically, the methodology for determining uncertainty at a given data spacing consists of the following steps:

1. Simulate realizations of the true distribution
2. Sample the simulated true distributions at a regular spacing and add sampling error
3. Generate realizations conditioned to the simulated data and block average
4. Calculate measures of uncertainty from the block-averaged realizations
5. Summarize uncertainty measures for the given data density

Each of these steps is discussed in greater detail below.

2.1 Simulate the Truth

The first step in evaluating the relationship between uncertainty and data density is to generate realizations of the truth (step 1 in **Figure 1**), denoted $\left\{z^{(l)}(\mathbf{u}), \mathbf{u} \in A, l = 1,...L\right\}$, where $\mathbf{u}$ represents a location and L is the number of

truth realizations. These realizations are characterized by a histogram and variogram which are reproduced within statistical fluctuations from one realization *l* to another *l'* (Journel and Kyriakidis, 2004). These realizations can be conditioned to pre-existing spatial sample data (step 0 in **Figure 1**) when such data are available. They can also be generated unconditionally when no sample data are available. In this case, the practitioner would enter the input histogram and variogram based on expert judgement or an analogue site.

2.2 Simulate Data

The next step is to simulate data, denoted as $\left\{z^{(l)}(\mathbf{u}_i), i=1,...,n_D; l=1,...,L\right\}$ with n_D being the number of data simulated for each realization. These data are drawn from the truth realizations at a specified spacing (step 2 in **Figure 1**).

Random error is added to the samples by Monte-Carlo simulation. This is done to replicate imperfect sampling. Journel and Kyriakidis (2004) discuss the relationship between the error and the truth, noting that the oft-used assumption of homoscedasticity is "extremely congenial and highly unrealistic." In reality, it is likely that both the error variance and error distribution are related to the true value. This work considers a Gaussian error distribution whose spread is proportional to the truth. Let $\{z_{\text{data}}(\mathbf{u}_i), i=1,...,n_D\}$ represent the simulated data with error. The error is randomly drawn from a Gaussian distribution with zero mean and spread specified by c_s such that the simulated data are defined by Equation 1 where $Y_{(0,1)}$ is a random value drawn from the standard normal distribution.

$$z^{(l)}_{\text{data}}(\mathbf{u}_i) = z^{(l)}(\mathbf{u}_i) + Y_{(0,1)} \cdot c_s \cdot z^{(l)}(\mathbf{u}_i),\ i=1,...,n_D;\ l=1,...,L \qquad 1$$

$c_s = 0$ indicates perfect sampling. The mean of the error distribution is zero to prevent the introduction of a bias. The magnitude of c_s depends on the sampling method being imitated. For instance, among the three most popular exploration drilling methods (diamond core, rotary, and percussion drilling)(Peters, 1978) there is significant variation in the precision of the samples obtained by each. There can also be various sampling standards which will vary depending on the nature and stage of the project. For example, a 15% sampling error is an accepted standard for exploration while 5% is typically required for compliance (Neufeld, 2003).

2.3 Conditionally Simulate and Block Average

The next step is to build realizations of the variable of interest conditional to the simulated data (step 3 in **Figure 1**). The pre-existing sample data could also be used to condition these realizations. This would necessitate the calculation of local data spacing for all locations.

For a given simulated data set, *l*, taken from a truth realization at a specified spacing (with or without pre-existing samples), a number, *K,* of conditional realizations $\left\{z^{(l)}_k(\mathbf{u}), \mathbf{u}\in A, k=1,...,K\right\}$ are generated by SGS. There are *K* conditional realizations generated for all *L* realizations of simulated data for a total of $K \cdot L$ realizations of the variable of interest for one data spacing. They all reproduce the input histogram and variogram within statistical fluctuations.

The *K* simulated realizations are assumed to have the same support as the sample data when such data exist. If no sample data are used, the realizations are assumed to be point scale. Typically, it is the uncertainty in block grades that is of interest. Journel (1978, 2004) suggests simulating point grades on a dense grid and averaging the point grades to the required block scale to generate simulated realizations at the block scale. Local uncertainty may then be assessed at the relevant scale. Consider the simulation of point support values done on a grid sufficiently dense to discretize a coarser grid of blocks of size *v* by n_v points. The simulated block value can be approximated by the arithmetic average of the n_v simulated point values within $v(\mathbf{u})$ given that $z(\mathbf{u})$ scales arithmetically (Journel and Kyriakidis, 2004) and $v(\mathbf{u})$ is sufficiently large:

$$\overline{z}^{(l)}_k(\mathbf{u}) = \frac{1}{|v|}\int_{v(\mathbf{u})} z^{(l)}_k(\mathbf{u})d\mathbf{u} \simeq \frac{1}{n_v}\sum_{j=1}^{n_v} z^{(l)}_k(\mathbf{u}_j) \qquad 2$$

The resulting block support realizations, $\left\{\overline{z}^{(l)}_k(\mathbf{u}), \mathbf{u}\in A, k=1,...,K; l=1,...L\right\}$, are used to calculate measures of uncertainty.

Block kriging could be used as an alternative to simulation to assess uncertainty as it "*is computationally quicker and provides a reasonable first approximation to the uncertainty*" (Deutsch and Beardow, 1999). SGS followed by block averaging is more flexible and "*provides a joint measure of uncertainty at all locations simultaneously*" (Deutsch and Beardow, 1999).

2.4 Calculate Measures of Uncertainty

The uncertainty at location **u** for one set of simulated data, *l*, is characterized by a probability distribution discretely represented by the *K* simulated block values. This distribution depends on both the volume being simulated and the set of sample information used for simulation. The probability distribution provides a full specification of the uncertainty about the unknown quantity at location **u** (Deutsch and Beardow, 1999). These local distributions are illustrated in step 4 of **Figure 1**. The probability distribution at an unsampled location has non-zero variance, increasing as the distance to nearby samples increases.

The set of probability distributions for all locations in the area of interest provides an assessment of uncertainty. Various uncertainty measures are used to provide a summary of the realizations. These measures are defined in Chapter 2. The standard deviation, coefficient of variation, P_{90}-P_{10}, (P_{90}-P_{10})/P_{50}, precision, and probability of misclassification measures are calculated at each location **u** from the *K* conditional realizations for all *L* data realizations. A single measure of uncertainty for a realization $\overline{U}^{(l)}$ can be calculated as the average of the local uncertainty measures $U^{(l)}(\mathbf{u})$ over all locations as in Equation 3 where $n_{\mathbf{u}}$ is the number of locations. The measure can be averaged over all *L* data realizations to give a single summary measure, $\overline{U}_j$, for a given data density as in Equation 4. This is illustrated by step 5 in **Figure 1**.

$$\overline{U}^{(l)} = \frac{1}{n_{\mathbf{u}}} \sum_{i=1}^{n_{\mathbf{u}}} U^{(l)}(\mathbf{u}_i) \tag{3}$$

$$\overline{U}_j = \frac{1}{L} \sum_{l=1}^{L} \overline{U}^{(l)} \tag{4}$$

The measures of uncertainty depend on the volume being simulated. There is greater uncertainty associated with prediction of small volumes. Uncertainty decreases as more data become available.

2.4.1 Standard Deviation and Coefficient of Variation

Standard deviation is a measure of the spread of a distribution. It is calculated at every block location. The standard deviation at a block location, $\hat{\sigma}^{(l)}(\mathbf{u})$, with *K* point support data is determined by Equation 5:

$$\hat{\sigma}^{(l)}(\mathbf{u}) = \left(\frac{1}{K} \sum_{k=1}^{K} \left(\overline{z}_k^{(l)}(\mathbf{u}) - \hat{\mu}_{\overline{z}}^{(l)}(\mathbf{u}) \right)^2 \right)^{1/2} \tag{5}$$

where $\hat{\mu}_{\overline{z}}^{(l)}(\mathrm{u})$ is the mean of the local distribution of block values:

$$\hat{\mu}_{\overline{z}}^{(l)}(\mathbf{u}) = \frac{1}{K} \sum_{k=1}^{K} \overline{z}_k^{(l)}(\mathbf{u}) \tag{6}$$

The spread of the block distribution is small near conditioning data and increases with increased distance to conditioning data. There are more blocks far from data than there are blocks close to data resulting in more locations with large spread than locations with small spread. The distribution of standard deviations is therefore negatively skewed.

The coefficient of variation of the distribution, $CV^{(l)}(\mathbf{u})$, is the standard deviation divided by the mean:

$$CV^{(l)}(\mathbf{u}) = \frac{\hat{\sigma}^{(l)}(\mathbf{u})}{\hat{\mu}_{\overline{z}}^{(l)}(\mathbf{u})} \tag{7}$$

The average of standard deviation/coefficient of variation over all locations is low for low data spacings (high data densities) and increases for increased data spacings (decreased data densities). The expected standard deviation, $\overline{\sigma}_j$, over all locations and data realizations for a given data density, d_j, is determined by combining Equations 3 and 4 and substituting σ for *U* (Equation 8). The expected coefficient of variation is determined similarly (Equation 9).

$$\bar{\sigma}_j = \frac{1}{n_u \cdot L} \sum_{l=1}^{L} \sum_{i=1}^{n_u} \hat{\sigma}^{(l)}(\mathbf{u}_i) \quad 8$$

$$\bar{CV}_j = \frac{1}{n_u \cdot L} \sum_{l=1}^{L} \sum_{i=1}^{n_u} CV^{(l)}(\mathbf{u}_i) \quad 9$$

2.4.2 Difference between Percentiles

The difference between percentiles is a measure of the spread of a probability distribution. It is determined at a location by first ordering the *K* values from lowest to highest such that $\bar{z}_k^{(l)}(\mathbf{u}) \le \bar{z}_{k+1}^{(l)}(\mathbf{u}), k = 1,...,K-1$. Percentiles of interest can then be located. The difference between the 10th and 90th percentiles provides a reasonable measure of spread. This difference can be standardized by dividing by the 50th percentile. The 10th, 50th, and 90th percentiles at a location are denoted as $\bar{z}_{P10}^{(l)}(\mathbf{u})$, $\bar{z}_{P50}^{(l)}(\mathbf{u})$, and $\bar{z}_{P90}^{(l)}(\mathbf{u})$ respectively.

The difference between the 10th and 90th percentiles at a location, denoted $\Delta^{(l)}(\mathrm{u})$ and defined in Equation 10, is a measure of the spread of a distribution. The difference between percentiles is small for blocks near data and increases for blocks further from data.

$$\Delta^{(l)}(\mathbf{u}) = \bar{z}_{P90}^{(l)}(\mathbf{u}) - \bar{z}_{P10}^{(l)}(\mathbf{u}) \quad 10$$

The standardized difference between the 10th and 90th percentiles, denote $\Delta_s^{(l)}(\mathbf{u})$ and defined in Equation 11, is a unitless measure of the spread of a distribution. Being unitless makes the standardized difference amenable to comparing multiple distributions with differing units and/or means.

$$\Delta_s^{(l)}(\mathbf{u}) = \frac{\Delta^{(l)}(\mathbf{u})}{\bar{z}_{P50}^{(l)}(\mathbf{u})} \quad 11$$

The average of these measures over all locations is low for small data spacings (high data densities) and increases as data spacing increases (data density decreases). The expected value of the difference between percentiles, $\bar{\Delta}^j$, over all locations and data realizations for a given data density, d_j, is determined by combining Equations 3 and 4 and substituting Δ for *U* as in Equation 12. The determination of the expected value of the standardized difference between percentiles, $\bar{\Delta}_s^j$, over all locations and realizations is shown in Equation 13.

$$\bar{\Delta}^j = \frac{1}{n_u \cdot L} \sum_{l=1}^{L} \sum_{i=1}^{n_u} \Delta^{(l)}(\mathbf{u}_i) \quad 12$$

$$\bar{\Delta}_s^j = \frac{1}{n_u \cdot L} \sum_{l=1}^{L} \sum_{i=1}^{n_u} \Delta_s^{(l)}(\mathbf{u}_i) \quad 13$$

2.4.3 Precision

The precision at a location, $p^{(l)}(\mathbf{u})$, is the proportion of simulated block values, $\bar{z}_k^{(l)}(\mathbf{u})$, that fall within a specified distance, $h^{(l)}(\mathbf{u})$, from the mean, $\mu_{\bar{z}}^{(l)}(\mathbf{u})$, at location **u**. If the spread of simulated block values is narrow, a large proportion of these values fall within the specified distance from the mean and the precision is high. If the spread of simulated block values is large, a small proportion of these values fall within the specified distance from the mean and the precision is low. The spread of simulated block values at a location increases farther from data, meaning fewer values fall within the specified distance from the mean leading to reduced precision.

The tolerance, $h^{(l)}(\mathbf{u})$, is specified by a multiplicative constant, *r*:

$$h^{(l)}(\mathbf{u}) = r \cdot \mu_{\bar{z}}^{(l)}(\mathbf{u}_i') \quad 14$$

Let $\tau_k^{(l)}(\mathbf{u};h)$ be a binary indicator such that Equation 15 is satisfied. The precision at a location, $p^{(l)}(\mathrm{u})$, is defined in Equation 16.

$$\tau_k^{(l)}(\mathbf{u};h) = \begin{cases} 1, \text{ if } \mu_{\bar{z}}^{(l)}(\mathbf{u}) - h^{(l)}(\mathbf{u}) \le \bar{z}_k^{(l)}(\mathbf{u}) \le \mu_{\bar{z}}^{(l)}(\mathbf{u}) + h^{(l)}(\mathbf{u}) \\ 0, \text{ otherwise} \end{cases}, k = 1,...,K \quad 15$$

$$p^{(l)}(\mathbf{u}) = \frac{1}{K}\sum_{k=1}^{K} \tau_k^{(l)}(\mathbf{u};h) \quad 16$$

The expected precision over all locations and data realizations, $\overline{p}_j$, is obtained by combining Equations 3 and 4, substituting *p* for *U* as in Equation 17.

$$\overline{p}_j = \frac{1}{n_{\mathbf{u}} \cdot L}\sum_{l=1}^{L}\sum_{i=1}^{n_u} p^{(l)}(\mathbf{u}_i) \quad 17$$

Precision and data spacing are inversely related. As data spacing increases, precision decreases as the spread of the local distributions become larger. The rate at which precision decreases with increasing data spacing depends on the variogram and histogram of the variable of interest.

2.4.4 Probability of Misclassification

Two categories of misclassification are considered requiring a single threshold, *t*, to define them. To know whether an observation has been misclassified the truth must also be known. The block average of the realization simulated in the first step, $\overline{z}^{(l)}(\mathbf{u})$, $\mathbf{u} \in A$, $l = 1,...,L$, is taken as the truth.

The type of misclassification depends on the value of the truth relative to the threshold. When the truth is greater than or equal to the threshold there is potential for Type I error, or a false positive. For instance, if a block is truly ore, there is potential for it to be falsely classified as waste (lost ore). When the truth is less than the threshold there is potential for Type II error, or a false negative. If a block is truly waste, there is potential for it to be falsely classified as ore (dilution).

Let $\alpha^{(l)}(\mathbf{u})$ represent the probability of making a Type I error at location **u** and let $\beta^{(l)}(\mathbf{u})$ represent the probability of making a Type II error at location **u**. Consider first the case where the truth at a location, $\overline{z}^{(l)}(\mathbf{u})$, is greater than or equal to the threshold, *t*. The probability of a false positive, $\alpha^{(l)}(\mathbf{u})$, is zero while the probability of a false negative, $\beta^{(l)}(\mathbf{u})$, may be non-zero. The probability of a false negative at this location is the number of simulated block values which are less than the threshold divided by *K*. Let $\phi_k^{(l)}(\mathbf{u})$ be a binary indicator defined by Equation 18. The probability of a false negative at this location is defined by Equation 19.

$$\phi_k^{(l)}(\mathbf{u}) = \begin{cases} 1, \text{ if } \overline{z}_k^{(l)}(\mathbf{u}) < t \\ 0, \text{ otherwise} \end{cases}, k = 1,...,K \quad 18$$

$$\beta^{(l)}(\mathbf{u}) = \frac{1}{K}\sum_{k=1}^{K} \phi_k^{(l)}(\mathbf{u}) \quad 19$$

Next, consider the case where the truth at **u**, $\overline{z}^{(l)}(\mathbf{u})$, is less than the threshold. The probability of a false negative is zero while the probability of a false positive may be non-zero. The probability of a false positive at this location is the number of simulated block values greater than or equal to the threshold divided by *K*. Let $\varphi_k^{(l)}(\mathbf{u})$ be a binary indicator defined by Equation 20. The probability of a false positive at **u** is defined by Equation 21.

$$\varphi_k^{(l)}(\mathbf{u}) = \begin{cases} 1, \text{ if } \overline{z}_k^{(l)}(\mathbf{u}) \geq t \\ 0, \text{ otherwise} \end{cases}, k = 1,...K \quad 20$$

$$\alpha^{(l)}(\mathbf{u}) = \frac{1}{K}\sum_{k=1}^{K} \varphi_k^{(l)}(\mathbf{u}) \quad 21$$

Probability of misclassification depends on the variability of the random variable. A variable with high variability has a higher probability of misclassification than a variable with low variability. It also depends on how near the local value is to the threshold. Values near the threshold have a higher probability of misclassification.

3. Implementation Example

Consider the determination of uncertainty vs. data density for a normally distributed variable with mean of 3.0 and variance of 1.0 and a one structure spherical variogram with no nugget effect and a range of 100m. To evaluate the relationship between uncertainty and data density for this variable the following parameters are used. Point-scale values are simulated at a spacing of 10m within a 600m x 600m area. This is a fairly coarse scale, but is useful

for illustrative purposes. These values are averaged into blocks with size 20m x 20m. There are *L*=10 unconditional data realizations each associated with *K*=100 conditional realizations for data spacings of 50, 70, 90, 110, and 130m. For each data spacing, 10 unconditional realizations are generated in Gaussian space. The samples drawn from these realizations are used to condition a set of 100 realizations. These realizations are block averaged to the desired scale and uncertainty measures are calculated at every location.

This process is illustrated in **Figure 2**. The top left plot shows one of the 50 (10 truth realizations for each of the five spacings) truth realizations generated. This truth realization is sampled at 90m spacing for a total of 36 samples (top middle). 100 realizations are generated that are conditioned to the 36 samples. Two of these realizations are shown in the top right plot. These point-scale realizations are arithmetically averaged to 20m square blocks (bottom right). The block averaged realizations are used to calculate local uncertainty measures (bottom middle). The distribution of local uncertainties is shown in the bottom left plot. The uncertainty measure shown is the coefficient of variation.

This same process is repeated for each of the five spacings. The result is five sets of distributions of uncertainty. The five distributions of the coefficient of variation are shown in **Figure 3**. The uncertainty distribution for 50m spacing has the lowest expected uncertainty and is positively skewed; most of the locations have low uncertainty. Two things happen to the distribution as data spacing increases. The first is an increase in the mean of the uncertainty. There is greater uncertainty associated with widely spaced data. The second is a change in the shape of the distribution. The shape changes from being positively skewed for spacings less than the variogram range to being negatively skewed for spacings greater than the variogram range. For spacings less than the variogram range, the majority of locations are close enough to data to be well informed whereas for spacings greater than the variogram range, the majority of locations fall outside the range of correlation.

Another method for visualizing this relationship is shown in **Figure 4**. This is a plot of uncertainty versus data spacing and shows the same five distributions as **Figure 3**. The relationship between uncertainty and data spacing is more easily discerned viewing the distributions in this manner. In addition to a vertical line representation of the histogram, an erased box plot (Tufte, 2001) shows the values of the 10^{th} percentile, first quartile, mean, third quartile, and 90^{th} percentile. This plot is useful in the context of an acceptable level of uncertainty. Assume that the acceptable level of uncertainty is specified as *the coefficient of variation will be less than 0.3 for 90% of the volumes within A*. The plot shows that this level of uncertainty is met at a data spacing of 70m.

The other measures of uncertainty which measure spread (standard deviation, P90-P10, and (P90-P10)/P50) exhibit a relationship with data spacing similar to the relationship between the coefficient of variation and data spacing (**Figure 6**). In all four cases, the measure of spread increases more rapidly for spacings less than the variogram range than for spacings greater than the variogram range. This mimics the variogram shape.

One aspect to note is that the non-standardized measures (standard deviation and P90-P10) start to show a bimodal distribution for data spacings approximately twice the block size whereas the standardized measures do not. Consider the plots in **Figure 5**. The left plot is the non-standardized P90-P10 uncertainty measure and the right plot is the standardized (P90-P10)/P50 uncertainty measure; both for 50m data spacing. The bimodal nature of the non-standardized measure can be seen. The uncertainty is low near data and high far from data with few values in between. The standardized measure varies smoothly due to its dependence on the local P50, eliminating this bimodal feature. Recall that this example is for a data spacing of 50m and a block size of 20m. The bimodal nature of the uncertainty distribution is most pronounced for a data spacing twice the block size.

Precision and the two types of misclassification exhibit their own unique relationships with data spacing (**Figure 7**). Precision is high when data are closely spaced and decreases as data spacing increases. The decrease in precision slows as the spacing between data exceeds the variogram range. Precision is very high for small data spacings resulting in a negatively skewed distribution with a large number of precision values at or near 1.0. Various statements regarding the level of uncertainty associated with each data spacing can be made. For example, at a spacing of 50m, 90% of the volumes have a greater than 77% probability of falling within 15% of the estimate.

The two types of probability of misclassification exhibit a relationship with data spacing similar to the measures of spread previously discussed; that is, the probability of misclassification increases with increased data spacing. These distributions are characterized by a large number of zero values creating a large spike in the histogram. As such, the histogram is not shown for these measures in **Figure 7**; only the erased box plot is shown. The line representing the 10^{th} and 25^{th} percentiles does not appear as both values are zero. Those summaries that

are visible (mean, upper quartile, 90th percentile) show that the occurrence of each type of misclassification increases as data spacing increase. The increase is more rapid for spacings less than the variogram range than for spacings greater than the variogram range. A variety of statements can be made regarding the level of uncertainty associated with the various data spacings. For example, *at a spacing of 90m the expected probability of Type I error is 18%*, or *for a spacing of 70m less than 10% of the volumes have a probability of Type II error greater than 50%*.

The relative occurrence of each type of misclassification error is controlled by the classification threshold and the reference distribution. The threshold for this example is 3.0 which is the mean of the reference distribution. Since the reference distribution is normally distributed, the occurrence of each type of misclassification is approximately equal. A cutoff below the mean would lead to an increase in the occurrence of Type I errors and a decrease in the occurrence of Type II errors. This threshold dependence is investigated further in Chapter 4.

4. Conclusions

The methodology for evaluating the relationship between uncertainty and data spacing involves simulating realizations of the truth, sampling these realizations at specified spacings, and simulating conditional to the chosen samples. These conditional realizations are then block averaged to some relevant scale and local measures of uncertainty determined for each location. This results in measures of uncertainty for the specified data spacings. The data spacing at which an acceptable level of uncertainty is met can then be determined.

There are limitations with this methodology. There is heavy reliance on the multivariate Gaussian modeling methodology. Often more data permits an improved understanding of the spatial distribution and the area of interest is subset into different domains; a global Gaussian model is not considered appropriate. Uncertainty in the modeling parameters is also important, but not considered here. Multiple sets of input parameters (histograms and variograms) could be considered.

References

Aspie D and Barnes RJ, 1990. Infill-sampling design and the cost of classification errors. Mathematical Geology. 22(8):915-932.

Boucher A, Dimitrakopoulos R and Vargas-Guzman JA, 2004. Joint simulation, optimal drillhole spacing and the role of the stockpile. In Leuangthong O and Deutsch CV (eds.) Geostatistics Banff 2004, Springer. 35-44.

Deutsch CV and Beardow AP, 1999. Optimal drillhole spacing for oil sands delineation, CIM Annual Meeting, Calgary, Alberta.

Journel AG and Huijbregts CJ, 1978. Mining Geostatistics. Blackburn Press, New Jersey. 600p.

Journel AG and Kyriakidis PC, 2004. Evaluation of Mineral Reserves. Oxford University Press, New York. 232p.

McBratney AB, Webster R, and Burgess TM, 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables – I. Computers and Geosciences. 7(4):331-334.

Neufeld CT, 2003. Beginner's Guide to Sampling. Centre for Computational Geostatistics, University of Alberta. 30p.

Webster R and Oliver MA, 2007. Geostatistics for Environmental Scientists 2nd edition. Wiley. 330p.

Figures

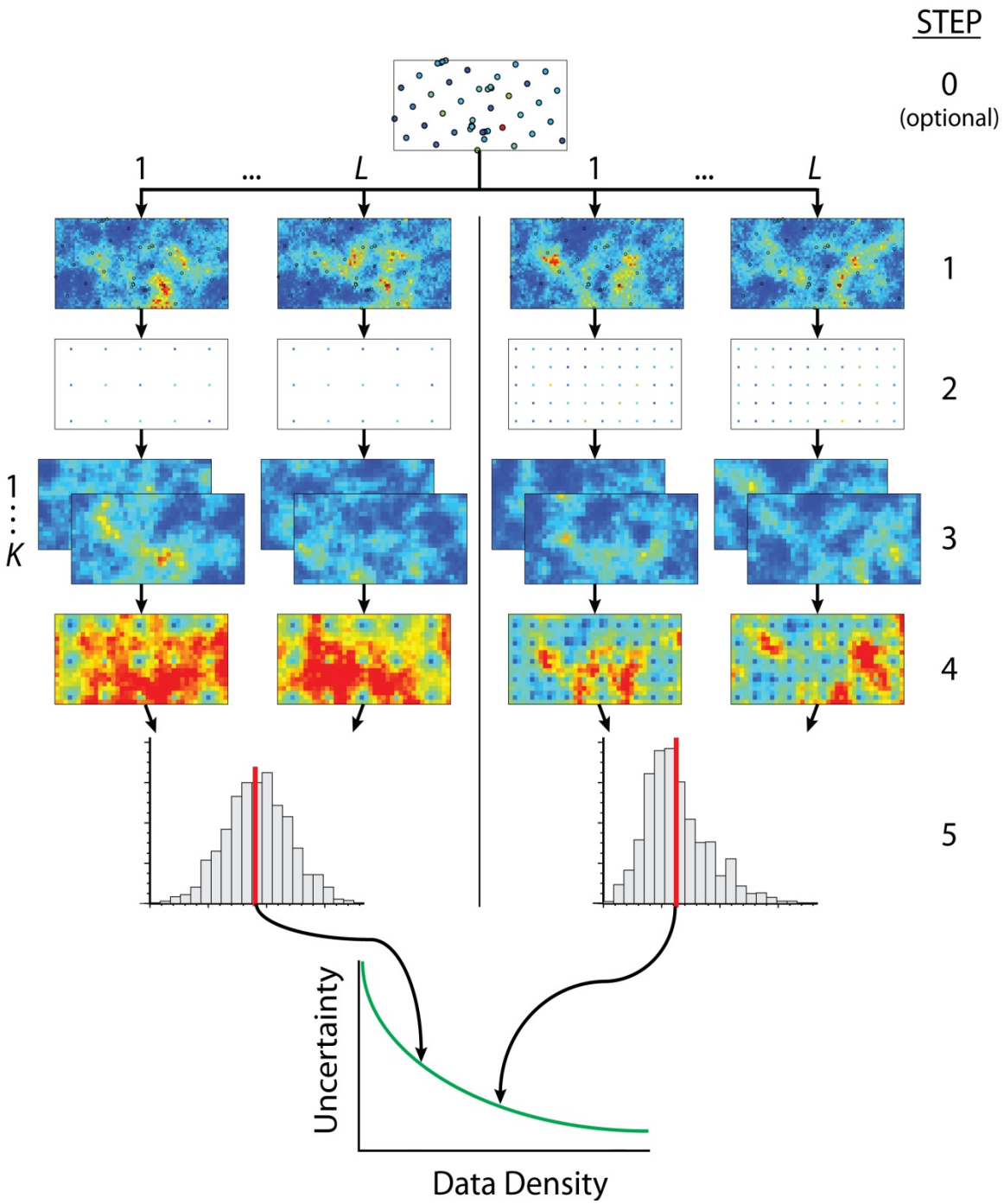

Figure 1: Illustration of the proposed methodology: 1) realizations of the truth are generated by sequential Gaussian simulation; 2) the truth is sampled at the desired spacing; 3) realizations are generated conditional to the samples; 4) local measures of uncertainty are calculated from the realizations; 5) the local measures of uncertainty are summarized for each data density.

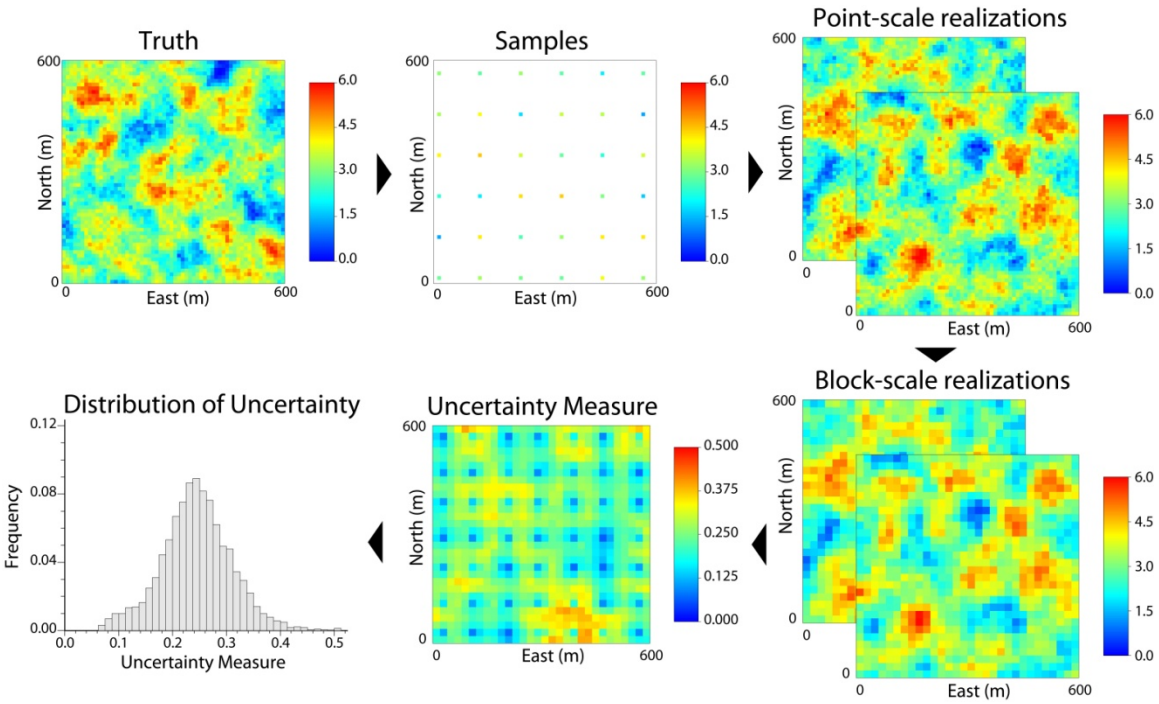

Figure 2: Illustration of the steps for the implementation example.

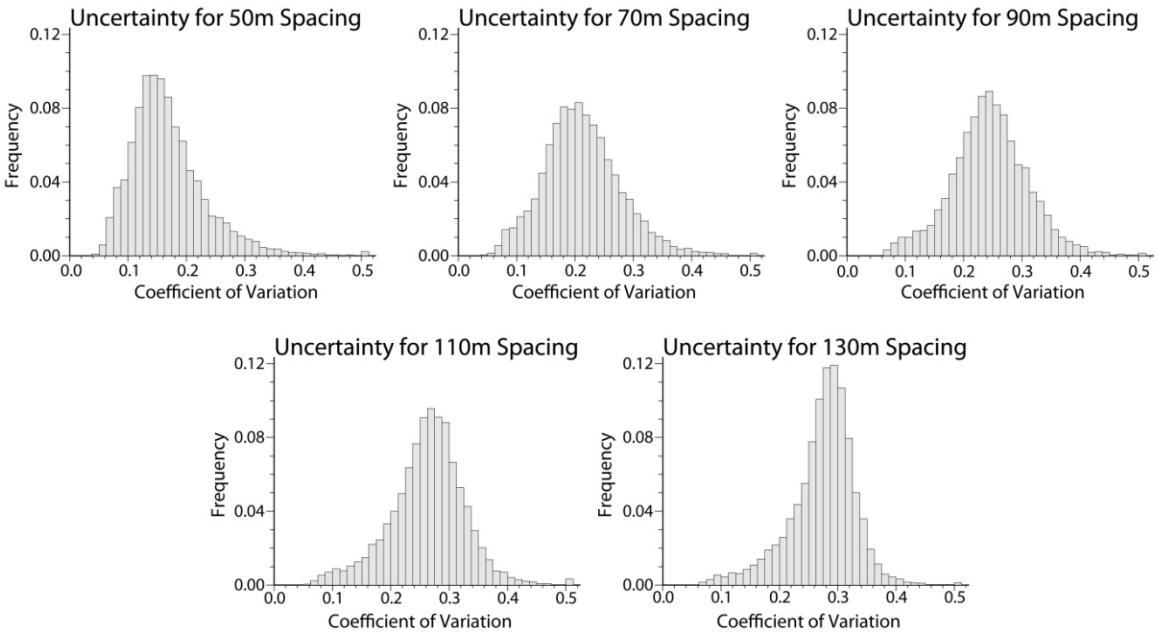

Figure 3: Uncertainty distributions for data spacings of 50, 70, 90, 110, and 130m respectively.

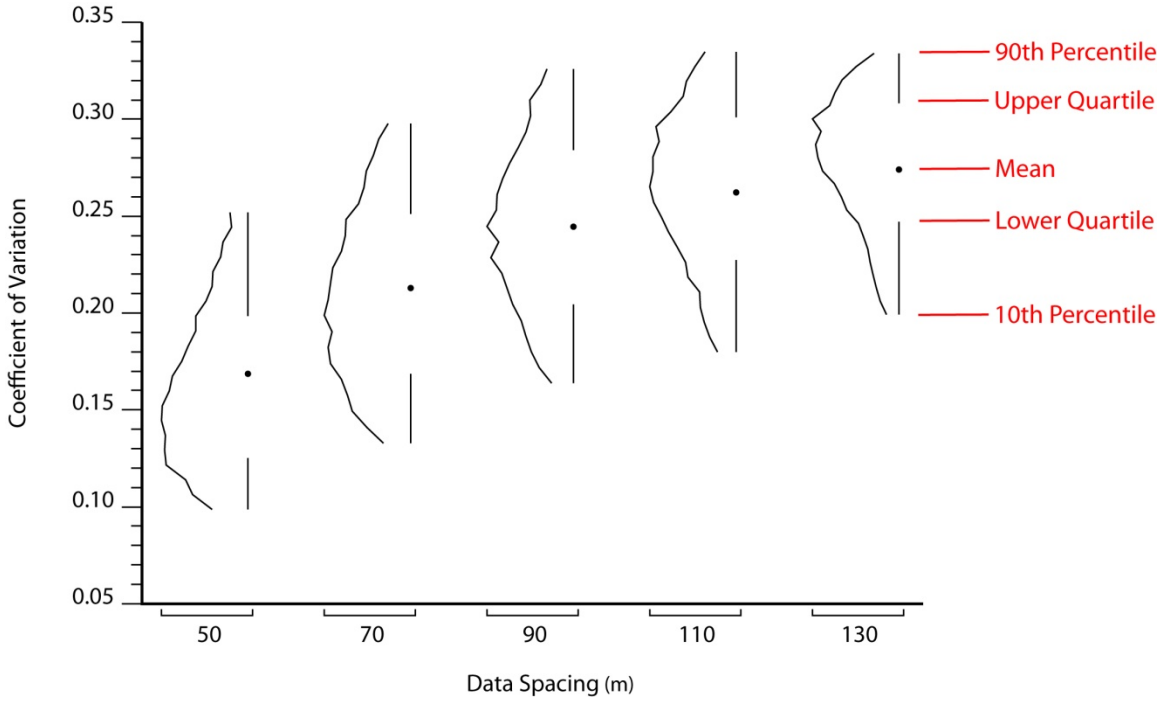

Figure 4: Uncertainty distributions for different data spacings with a description of the various markings on the plot (red).

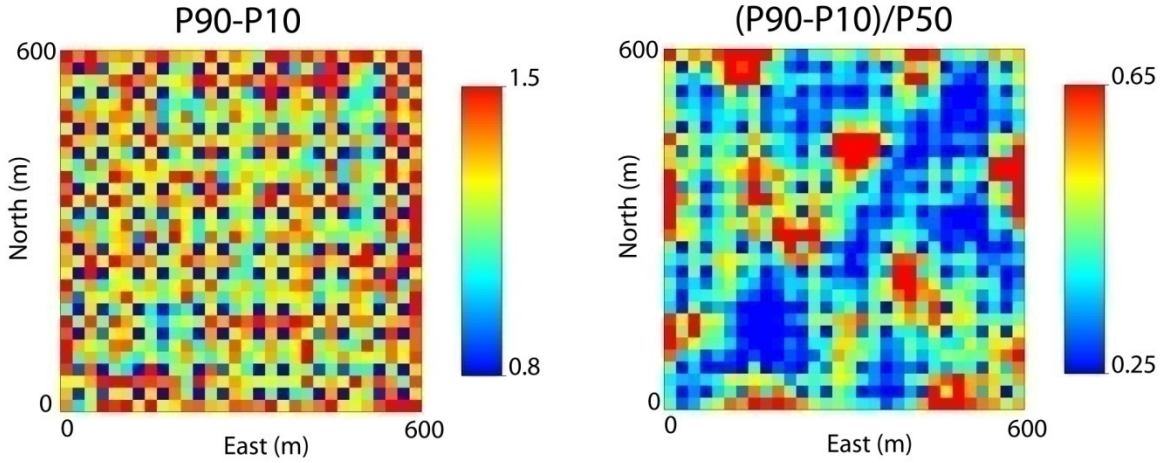

Figure 5: Nonstandardized and standardized measure of uncertainty for 50m spacing.

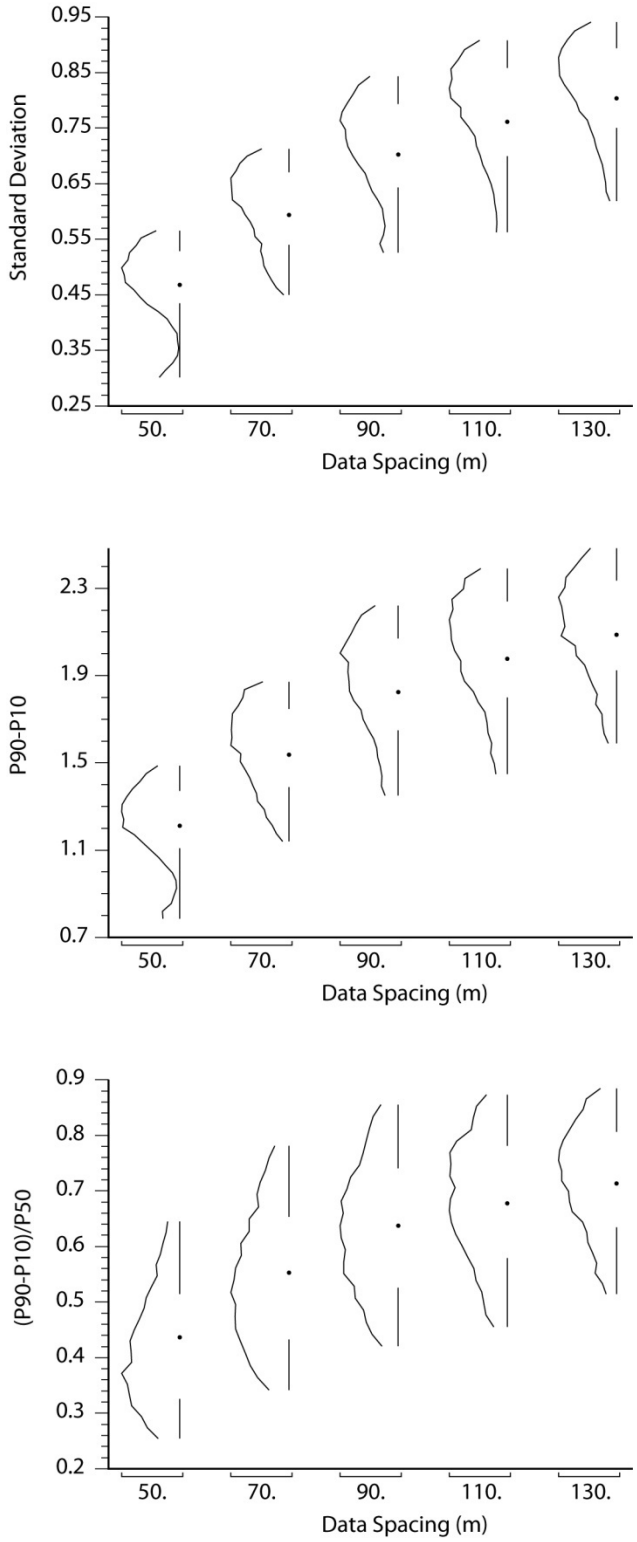

Figure 6: Relationships between standard deviation, P90-P10, (P90-P10)/P50, and data spacing.

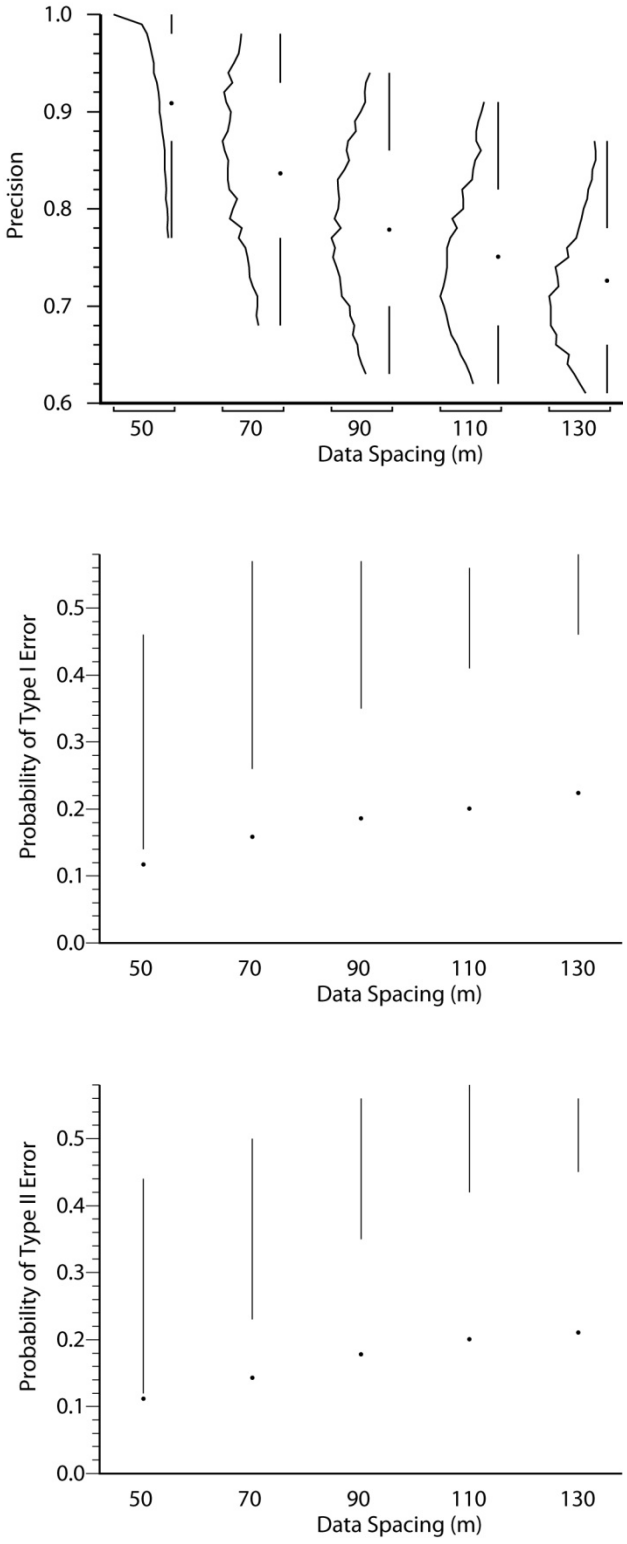

Figure 7: Relationships between precision, Type I error, Type II error, and data spacing.