University of Alberta

Data Spacing and Uncertainty

by

Brandon Jesse Wilde

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

in

Mining Engineering

Department of Civil and Environmental Engineering

Fall 2010

Edmonton, Alberta

Examining Committee

Clayton V. Deutsch, Civil and Environmental Engineering

Jeff B. Boisvert, Civil and Environmental Engineering

Octavian Catuneanu, Earth and Atmospheric Sciences

Abstract

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 properties of the variable being modeled. A densely sampled erratic variable may have a level of uncertainty similar to a sparsely sampled continuous variable. A simulation based approach is developed to quantify the relationship between uncertainty and data spacing. Reference realizations are simulated and sampled at different spacings. The samples are used to condition additional realizations from which uncertainty is quantified. A number of factors complicate the relationship between uncertainty and data spacing including the proportional effect, nonstationary variogram, classification threshold, number of realizations, data quality and modeling scale. A case study of the relationship between uncertainty and data density for bitumen thickness data from northern Alberta is presented.

Acknowledgements

I would like to thank my supervisor, Dr. Clayton V. Deutsch, whose support and guidance have been invaluable and without whom this thesis would not have been possible. I would also like to thank all other members of the Centre for Computational Geostatistics, particularly Jeff Boisvert for his guidance with respect to effective visual presentation.

I thank the Natural Sciences and Engineering Research Council of Canada as well as the member companies of the Centre for Computational Geostatistics for their financial support.

I would like to thank my parents who have always encouraged me to work hard and follow my own path.

My greatest thanks go to my wife, Bonnie, who has spent too many evenings alone with our children, patiently supporting me in this endeavor. She has been a wonderful source of joy and encouragement and I cannot thank her enough.

Table of Contents

List of Tables

List of Figures

List of Symbols

$\mathbf{u}$	a location in space
$X(\mathbf{u})$	a random variable
V	a volume
A	an area or domain
n	a quantity
s	data spacing
d	data density
$z^{(l)}\left(\mathbf{u}\right)$	simulated truth value from the l^{th} realization
$z_k^{(l)}\left(\mathbf{u}\right)$	simulated value from the k^{th} realization conditioned to data drawn from the l^{th} truth realization
$\overline{z}_k^{(l)}\left(\mathbf{u}\right)$	block average of simulated values
μ	mean
σ	standard deviation
CV	coefficient of variation
Δ	a difference
p	precision
t	a threshold
α	risk alpha, probability of Type I error
β	risk beta, probability of Type II error

Chapter 1

Introduction

Numerical models of geological deposits are often used to make highly consequential decisions. These decisions could be whether to build a mine, whether to drill production wells, whether a site requires environmental remediation and so on. These decisions involve a great deal of money. The numerical models must accurately represent the geologic site to ensure that the best decision is made.

The numerical models are improved by collecting a large quantity of data by sampling. Sampling aims to collect reliable information about a deposit. The spacing between samples decreases as a project progresses through various stages of exploration. Early stages of sampling involve drilling reconnaissance holes spaced at large distances. If favorable results are found, additional holes are drilled at smaller spacings in successive sampling stages. The pattern is adjusted until satisfactory coverage has been attained (Peters, 1978).

There are various ways of determining sampling locations. One method involves statistical pattern sampling on a regular or random stratified grid at predetermined spacings where a geologic model of the expected geology is the dominant consideration (Peters, 1978). Another involves random sampling of a prerequisite quantity of samples (Lowrie, 2002). A third method involves target-area sampling where more samples are collected in areas of particular geologic interest (e.g. areas with a high grade)

and fewer samples are collected in areas of little geologic interest (Peters, 1978).

Sampling the full true distribution of the attribute of interest is not feasible and uncertainty is an inherent aspect of the modeling of geological deposits. The ability to accurately represent the truth is improved as more data are collected. Consider the schematic shown in Figure 1-1. This figure illustrates the ability of a model to represent the truth for different quantities of data. The plot at the top of the figure is the truth (unknown in practice). Three different sample quantities are shown below this plot on the left. Sample quantity ranges from many at the top to few at the bottom. The corresponding models built using the samples are shown on the right. The model built with many data gives a much closer representation of the truth than the model built with few data.

Consider the qualitative cost versus data quantity relationship shown in Figure 1-2. The ability of a model to represent the truth could be summarized in terms of a cost due to modeling error where a good model has low cost and a bad model has high cost. This relationship would be difficult to define explicitly. The relationship between data quantity and sampling cost is linear unless there is a discount when many data are collected. The total cost is the sum of the sampling cost and the cost due to modeling error. The data quantity that minimizes cost could be considered as optimal.

A number of authors have discussed the assessment of sampling schemes (McBratney *et al.* 1981; McBratney and Webster, 1981; Aspie and Barnes, 1990; Bueso *et al.* 1999; Webster and Oliver, 2007), particularly with respect to groundwater monitoring (Carrera *et al.* 1984; Meyer and Brill, 1988; Rouhani and Hall, 1988; Loaiciga, 1989; Andricevic, 1990; Meyer *et al.* 1994; Criminisi *et al.* 1997; Storck *et al.* 1997; Zhang, 2005). Four works are of particular relevance to this one as they have employed similar methodologies for similar problems: Englund and Heravi (1992), Deutsch

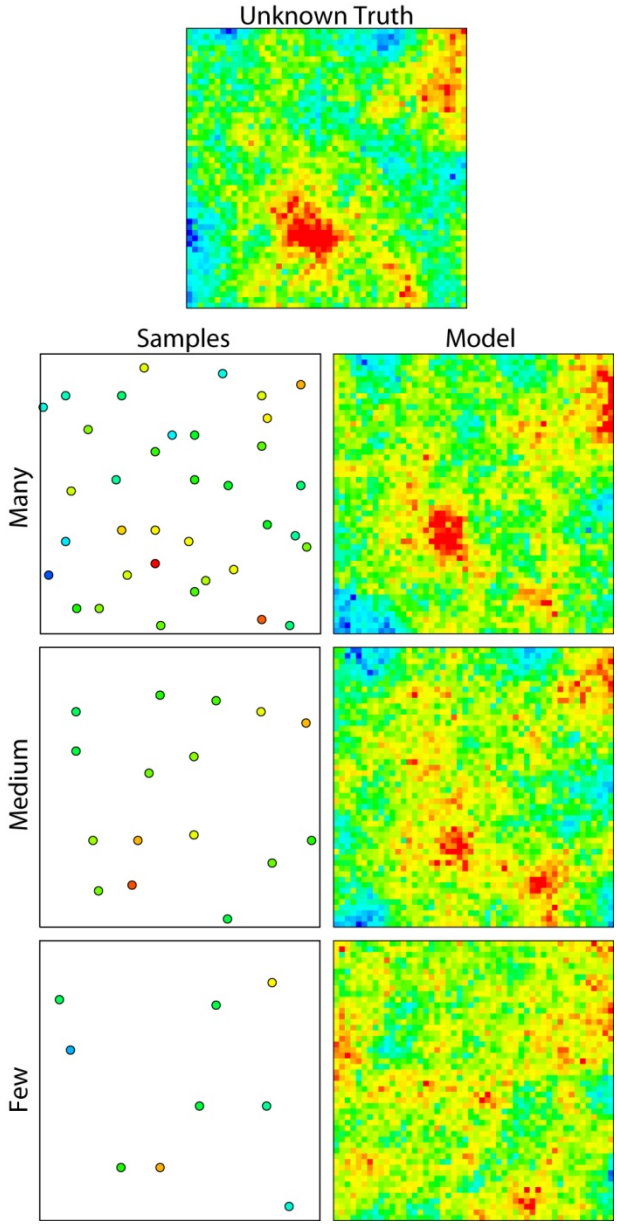

Figure 1-1: Relationship between data quantity and the ability of a model to represent the truth. As the quantity of data increases (top), the model more closely represents the truth.

and Beardow (1999), Boucher *et al.* (2004), and Journel and Kyriakidis (2004).

Englund and Heravi (1992) discuss a methodology for determining the quantity of samples that would result in the lowest total project cost for a

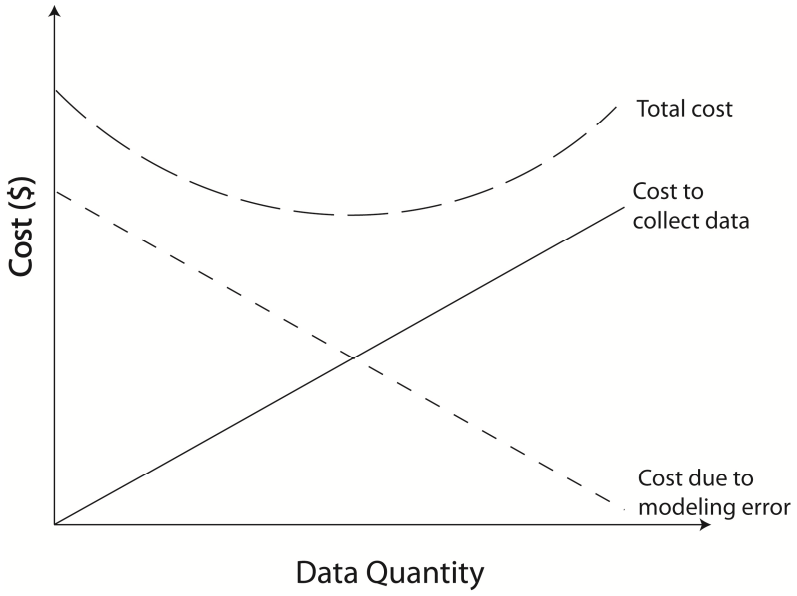

Figure 1-2: Cost versus data quantity for data collection, modeling error and total cost.

remediation project. This methodology uses SGS to generate a simulated site model. This site model is assumed to represent the truth and various quantities of data are sampled from it. Block kriging is then performed using the samples and the total project cost determined. The project cost considers sampling cost and remediation cost as well as the cost of residual contamination. This process is repeated multiple times for different quantities of data resulting in a cost versus data quantity curve. The data quantity that minimizes cost is taken as the optimal.

Deutsch and Beardow (1999) propose a methodology for determining the optimum drillhole spacing in an oil sands deposit. They propose using either block kriging or stochastic simulation to assess uncertainty at all locations within an area of interest for a given data spacing. The expected value of the uncertainty over all locations is retained for the given data spacing. This process is repeated for many data spacings allowing the relationship between uncertainty and data spacing to be discovered.

Accounting for the cost of drilling and the cost of uncertainty allows an optimal drillhole spacing to be determined.

Boucher *et al.* (2004) propose a technique for assessing infill sampling. Their method uses available data to generate one stochastic realization of the geologic attribute. This realization, called the actual deposit, is sampled with the different infill drilling schemes of interest. For each sampling scheme, several realizations of the geologic attributes are generated conditional to the simulated samples. These realizations, including the actual deposit, are block averaged to the desired scale. Economic indicators are used to classify the blocks. The simulated classifications are compared to the actual for many different infill sampling schemes to assess sensitivity of the results. This methodology quantifies the the tradeoff between the cost of misclassification and sampling.

Journel and Kyriakidis (2004) describe a methodology for evaluating mineral reserves. This methodology involves simulating point support realizations of the variable of interest. These point support realizations play two roles in the methodology: 1) they are used as a basis for simulating future selection data, and 2) they are block averaged and called the actual grade of the variable of interest. Future selection estimates are generated from each realization of simulated future data using block kriging. The block kriged future selection estimates are combined with the simulated block grades to determine the profit of the realization. This process is repeated for many realizations to determine the uncertainty in the profit. They suggest that sensitivity analyses be performed by varying the quality and density of future selection data to assess the impact on profit.

The methodology proposed herein is a combination of these approaches. The aim is to understand the relationship between data spacing (or density) and uncertainty. Knowledge about a geological attribute is related to the quantity and quality of observations of the attribute.

Uncertainty exists because of a lack of knowledge and is not an inherent feature of the geological attribute. Thus, uncertainty is reduced as knowledge increases when more and better quality data are collected.

The manner in which uncertainty decreases with decreasing data density is controlled by the spatial variability of the attribute of interest (Deutsch and Beardow, 1999). The variogram is a geostatistical tool that quantifies geological variability. It is a measure of the average dissimilarity between values that are separated by vector distances **h** (Goovaerts, 1997). If the geological attribute is highly continuous, the variability between samples is quite small over large distances and the attribute has a large range of correlation. If the geological attribute is highly variable, the range of correlation is small. An attribute with a long range of correlation will have less uncertainty for a given data density than an attribute with a shorter range of correlation.

The proposed methodology could be used to determine an appropriate data spacing. Doing so would require specifying an acceptable level of uncertainty. The largest data spacing that meets the specified level of uncertainty could then be applied to the site.

One way of quantifying uncertainty is by multiple high resolution geostatistical realizations. These must be constructed with care, checked with cross validation and reconciled with any production data. The uncertainty at the high resolution scale of the geostatistical model is rarely relevant for disclosure or expressing acceptable uncertainty. The high resolution models are scaled to a larger scale relevant for technical and economic decision making. Often, this larger scale represents a nominal time period for production such as a month, quarter or year. Each nominal volume in the area of interest has a distribution of uncertainty in a variable of interest. The variable of interest could be the mass fraction of an important component, a combined economic variable, or the material above a fixed economic threshold. The outcomes of multiple realizations

describe the uncertainty for each nominal volume; however, it is necessary to summarize the uncertainty for comparison and reporting.

An acceptable level of uncertainty is defined for a particular purpose. This purpose is some type of decision making or classification. The decision whether some level of uncertainty is acceptable is made by a qualified person who deems the uncertainty to be reasonably small and suitable for the problem at hand.

The domain for which an acceptable level of uncertainty is applicable must be defined. Within some global site, G, there could be a number of areas, $A_j, j \in G$ as shown in Figure 1-3. Each area would be characterized by regular data spacing as shown for areas 3 and 4. An acceptable level of uncertainty is defined for a given area. Within an area there are a number of volumes, V_i, $i \in A_j$ (area 1 in Figure 1-3). Each volume has a distribution of uncertainty. Establishing whether an area meets an acceptable level of uncertainty involves determining the proportion of volumes that meet a required level. The required proportion of volumes is part of the specification of acceptable uncertainty. The area can then be classified based on whether the acceptable level is met. There are a number of formats for expressing acceptable uncertainty. The choice of format depends on the audience, local customs, the particular problem, transparency, and preferences of the practitioner. It is common for probabilistic uncertainty specification to include:

1. Identification of the population or sample being considered (the volume, V),
2. A defined precision,
3. The probability to be within the defined precision, and
4. The proportion of volumes, V_i, within the area, A_j, required that meet the preceding criteria

An example is *the true grade of monthly production volumes will be within*

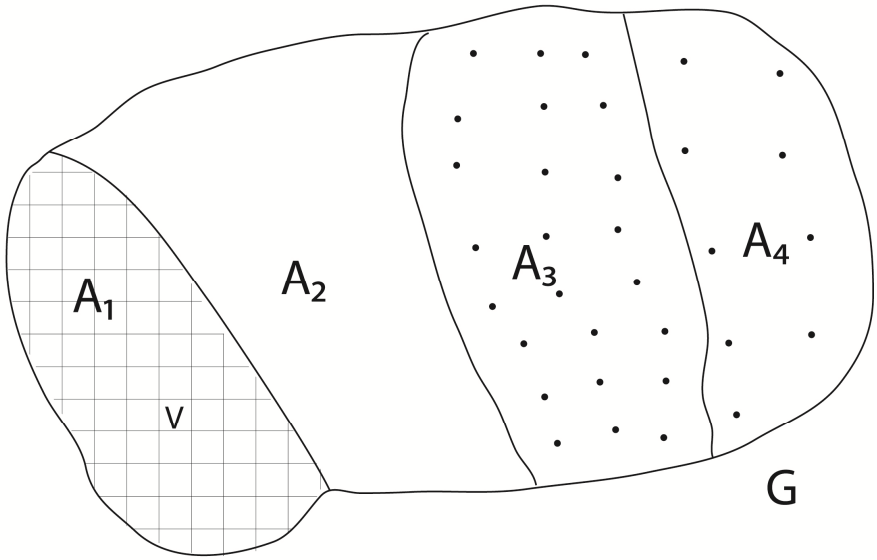

Figure 1-3: Illustration of the domains for which an acceptable level of uncertainty applies.

15% of the predicted grade 19 times out of 20 for at least 90% of the volumes V_i, $i \in A$. (Deutsch *et al.*, 2006). This statement of acceptable uncertainty includes a volume (monthly production volumes), a defined precision (within 15%), a probability to be within the defined precision (19 times out of 20, or 95%), and the proportion of volumes required to meet these criteria (90%). The second and third parameters are illustrated for one volume, V_i, in Figure 1-4. This is just one way of specifying uncertainty. There is nothing special about monthly/15%/ 95%/90%, but values similar to these are commonly mentioned.

Chapter 2 defines measures of spatial arrangement (data density, data spacing) and uncertainty (standard deviation, precision, ...). Chapter 3 describes a new method for determining the relationship between data density/spacing and uncertainty. This method is a combination of the methods previously reviewed. It relies heavily on the sequential Gaussian simulation (SGS) algorithm described in numerous publications including

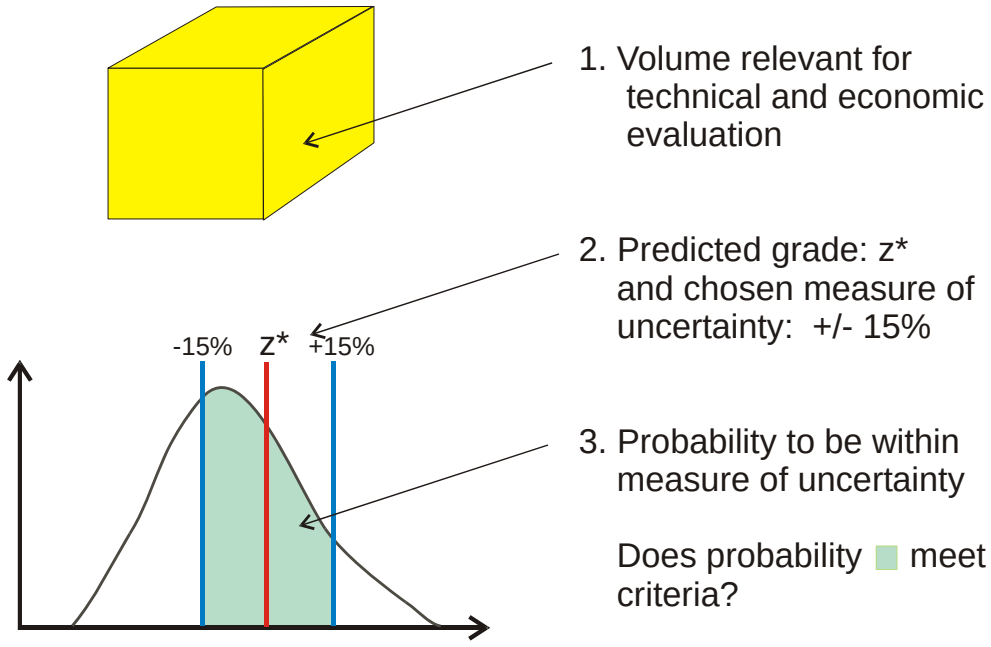

Figure 1-4: Illustration of three of the parameters often used to specify uncertainty (from Deutsch *et al.*, 2006).

Isaaks (1990), Goovaerts (1997), and Deutsch and Journel (1998). An implementation example is presented. Chapter 4 considers the factors that confound the relationship between data density and uncertainty. Uncertainty is not a simple function of data density; it depends on a number of other factors including the histogram of the attribute, the proportional effect, nonstationarity and data quality. This chapter explores these confounding factors. Chapter 5 demonstrates an application of the proposed methodology to an oil sands deposit in northern Alberta. The results obtained using the proposed methodology are compared to results obtained by calculating data spacing directly from the data and determining uncertainty by simulation.

Chapter 2

Measures of Spatial Arrangement and Uncertainty

This thesis discusses the relationship between various measures of spatial arrangement and uncertainty. Data spacing and data density are defined in Section 2.1. Measures of uncertainty include standard deviation, difference between percentiles, coefficient of variation, and probability of misclassification. These are defined in Section 2.2.

2.1 Measures of Spatial Arrangement

2.1.1 Data Spacing

Data spacing is the distance between adjacent data for a representative area. A densely sampled area will have a small spacing relative to an area that is sparsely sampled. Data spacing at a location, $s(\mathbf{u})$, is determined by considering the number of nearby samples, $n_V(\mathbf{u})$, within some volume, $V(\mathbf{u})$. If $V(\mathbf{u})$ is two-dimensional, the square root of $V(\mathbf{u})$ divided by $n_V(\mathbf{u})$ gives data spacing as shown in Equation 2.1. For the three dimensional case, the calculation of data spacing can be performed in two dimensions when the drillholes are all vertical.

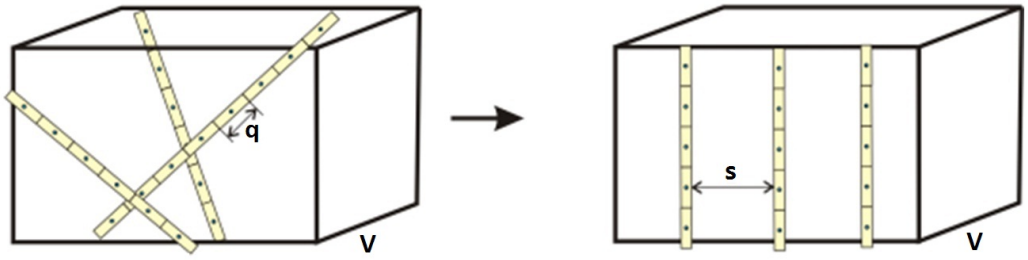

Figure 2-1: Ilustration of the determination of data spacing when data are irregularly spaced in three dimensions.

When the drillholes are not all vertical, as shown in Figure 2-1, the data spacing calculation must consider a three-dimensional volume. The along-hole spacing, q, is included in the calculation as shown in Equation 2.2 thus defining the equivalent regular drillhole spacing.

$$s\left(\mathbf{u}\right) = \left(\frac{V\left(\mathbf{u}\right)}{n_{V}\left(\mathbf{u}\right)}\right)^{1/2} \tag{2.1}$$

$$s\left(\mathbf{u}\right) = \left(\frac{V\left(\mathbf{u}\right)}{q \cdot n_{V}\left(\mathbf{u}\right)}\right)^{1/2} \tag{2.2}$$

To calculate data spacing at a location, either $V(\mathbf{u})$ or $n_V(\mathbf{u})$ are normally fixed. If $n_V(\mathbf{u})$ is fixed i.e. $n_{V}(\mathbf{u}) = n_{V}, \forall\, \mathbf{u} \in A$, the volume $V(\mathbf{u})$ required to encompass the n_V data is calculated and spacing is determined as defined previously. If $V(\mathbf{u})$ is fixed i.e. $V(\mathbf{u}) = V, \forall\, \mathbf{u} \in A$, the number of observations $n_V(\mathbf{u})$ that fall within V is determined and spacing is determined as defined previously. The choice of n_V or V affects the results: too small of a volume or too few samples leads to noisy results; too large of a volume or too many samples leads to over smoothing.

2.1.2 Data Density

Data density is the number of data observations per unit volume, commonly reported as number of data per section or hectare. Data

density at a location, $d(\mathbf{u})$, is determined by considering the number of nearby samples, $n_V(\mathbf{u})$, within some volume, $V(\mathbf{u})$. Dividing the number of samples by their volume gives data density. If the data are arranged such that many observations fall within a small volume, data density is high. If a large volume contains few observations, data density is low.

Data density at a location, $d(\mathbf{u})$, is determined in the same manner as data spacing by fixing either V or n_V and calculating the non-fixed parameter. This is illustrated for a fixed volume in Figure 2-2. The units of density depend on the units of V. For example, if V has units of m^2, then density has units of samples/m^2. It may be desirable to convert the units of density to a more common measure such as samples/hectare or samples/section. There are ten thousand square meters in a hectare and 2,589,988.11 square meters in a section. To convert density from data per square meter to more useful units, simply multiply density by the appropriate constant. Consider an area with side length of 1600m where samples are regularly spaced every 400m. There are 16 samples within the area of 1600^2=2.56x10^6 m^2. The data density is 16 samples/2.56x10^6 m^2=6.25x10^{-6} samples/m^2; a very small number of samples per square meter. This is equivalent to $\frac{6.25\times10^{-6}\ \text{samples}}{\text{m}^2}\times\frac{10000\,\text{m}^2}{\text{ha}}=\frac{0.0625\ \text{samples}}{\text{ha}}$ and $\frac{6.25\times10^{-6}\,\text{samples}}{\text{m}^2}\times\frac{2589988.11\,\text{m}^2}{\text{section}}=\frac{16.2\ \text{samples}}{\text{section}}$.

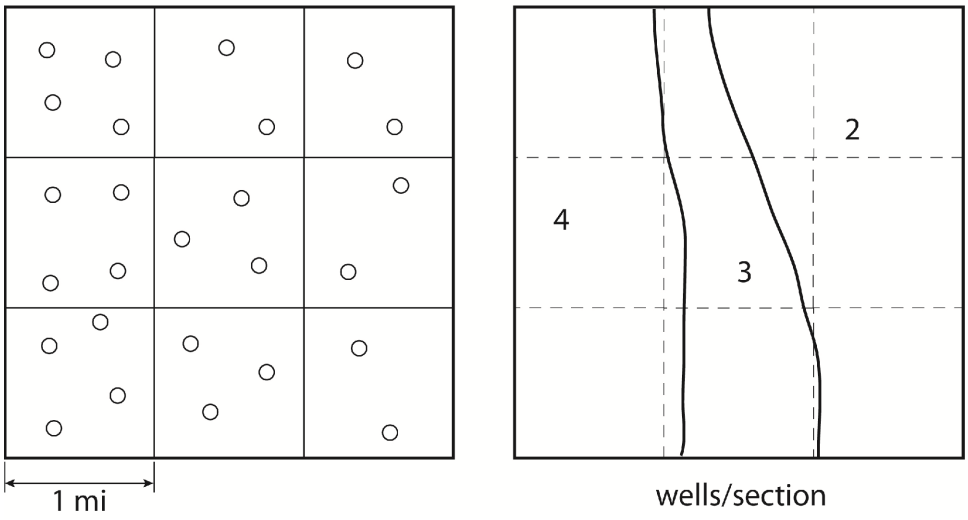

Figure 2-2: Illustration of the calculation of data density for a fixed volume, *V*.

2.2 Measures of Uncertainty

Modeling earth sciences variables involves uncertainty due to our lack of knowledge. It is infeasible to exhaustively sample an area of interest; therefore the uncertainty must be quantified and various measures have been developed to do this. These uncertainty measures have proven useful for geostatistics and include: standard deviation, difference between percentiles, precision, and probability of misclassification (Goovaerts, 1997; Myers, 1997; Barabas et al., 2001; Duggan and Dimitrakopoulos, 2004).

2.2.1 Standard Deviation

A common measure of the spread of a probability distribution is the standard deviation (Figure 2-3). The standard deviation is the square root of the variance and has the same units as the variable. The variance of a distribution is the expected square deviation of the variable from its mean. Consider the random variable, X, with expected value, μ. The standard deviation, σ, of X is:

$$\sigma = \left(E\left\{ (X-\mu)^2 \right\} \right)^{1/2} \qquad 2.3$$

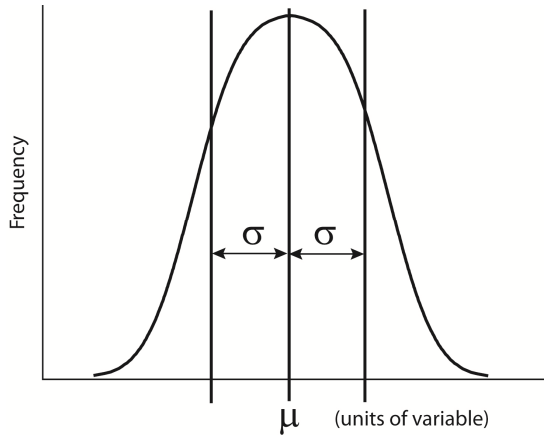

Figure 2-3: Illustration of the standard deviation and mean of a distribution.

The standard deviation can be standardized by the mean to give a measure of uncertainty called the coefficient of variation, CV:

$$CV = \sigma / \mu \qquad 2.4$$

This is a relative measure of variability. The standard deviation is often understood in the context of the mean. The coefficient of variation is dimensionless making it useful when comparing distributions with different units or different means (Anderson *et al.* 1994). The coefficient of variation is sensitive to small changes in the mean when the mean is near zero.

2.2.2 Difference between Percentiles

Another measure of spread is the difference between percentiles. Percentiles are a specific form of quantiles. Quantiles are values with probabilistic meaning taken at regular intervals from the cumulative distribution function (CDF) of a random variable. Consider dividing an ordered distribution into q equally sized subsets. The quantiles are the *q-1* values marking the boundaries between consecutive subsets. When the distribution is divided into 100 subsets, the 99 quantiles are called percentiles. Other quantiles have been given specific names.

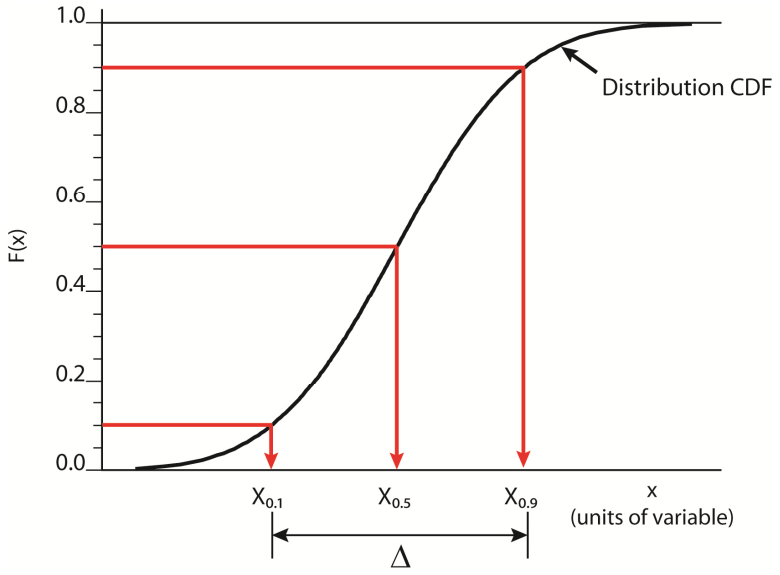

Figure 2-4: Illustration of the three percentiles commonly used to measure uncertainty.

For instance, when $q=2$ the quantile is called the median and when $q=4$ the three quantiles are called quartiles (Johnson and Bhattacharyya, 1996). Quantiles are denoted X_p such that $F\left(x_p\right) = p$ where F represents the cumulative distribution function of X.

The difference between two symmetric percentiles provides a measure of spread, that is, the difference between percentile i $(0<i<q)$ and percentile q-i. One well known percentile difference is the inter-quartile range which is the difference between the first and third quartiles. The difference between the 10^{th} and 90^{th} percentiles of a distribution could also be considered as a measure of uncertainty. These are symmetric percentiles whose difference, Δ, is a measure of the spread of a distribution as defined in Equation 2.5. These percentiles are illustrated in Figure 2-4. The difference between percentiles has the units of the variable under consideration.

$$\Delta = X_{0.9} - X_{0.1} \qquad 2.5$$

$$\Delta_s = \frac{X_{0.9} - X_{0.1}}{X_{0.5}} \qquad 2.6$$

The difference between percentiles can be standardized as in Equation 2.6 to be a unitless measure, Δ_s, similar to the coefficient of variation. This is done by dividing by the median, or 50^{th} percentile (P50), of the distribution. This standardization makes comparison between distributions with different units or means possible.

2.2.3 Precision

Precision is a measure of the narrowness of a distribution; that is, as a distribution narrows, precision increases. Precision is determined by finding the proportion of a distribution that falls within a given distance from the mean. Consider a distribution of the random variable, X, with mean, μ. Also consider a distance from the mean, h, defined by a multiplicative constant, r, as in Equation 2.7. Precision, p, is the probability to be within the specified tolerance as defined in Equation 2.8.

$$h = r \cdot \mu \qquad 2.7$$

If the constant, r, chosen is too large, the precision values will all be near 1.0; if the constant chosen is too small, the precision values will all be near 0.0. Using a reasonable constant provides the best measure of precision.

$$p\left(r\right) = \text{Prob}\left\{\mu - h \le X < \mu + h\right\} \qquad 2.8$$

A distribution with a large spread will have fewer values between μ-h and μ+h leading to a low value for p whereas a distribution with small spread will have more values between μ-h and μ+h leading to a high value for p as illustrated in Figure 2-5 where p is proportional to the shaded area.

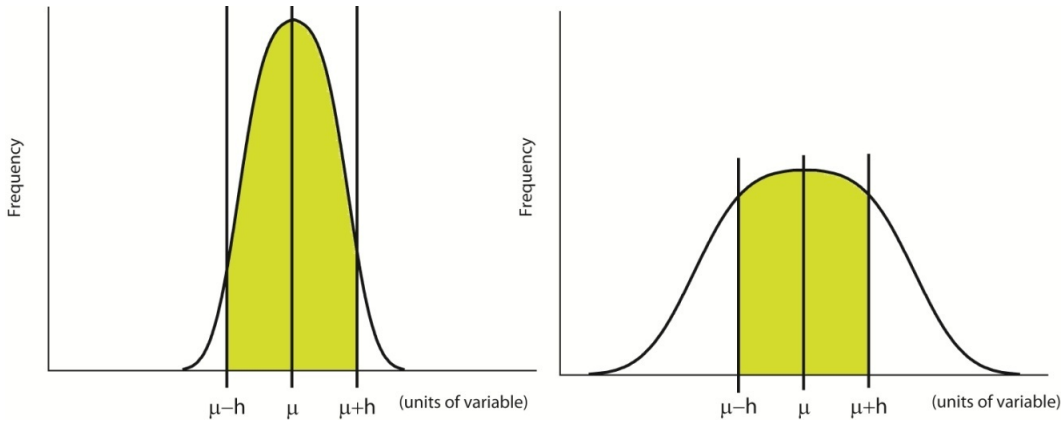

Figure 2-5: Narrow distribution with high precision (left) versus a wide distribution with low precision (right).

2.2.4 Probability of Misclassification

Consider having m groups or classes. The probability of misclassification refers to the probability of classifying an observation from class i as being from class j, $i \neq j$. Classification in the earth sciences is often binary (m=2): ore or waste, net or non-net, requires environmental remediation or does not. In the binary case, misclassification errors have been termed Type I and Type II errors where Type I error is a false positive, or overestimation, and Type II error is a false negative, or underestimation (see Figure 2-6). An example of a Type I error would be classifying waste as ore (dilution) while an example of a Type II error would be classifying ore as waste (lost ore). These errors can have different consequences. The symbols α and β have been adopted to represent the probability of making Type I and Type II errors respectively.

Classification requires m-1 thresholds, denoted t_i, i=1,…,m-1. Values greater than or equal to t_{i-1} and less than t_i are classified as being from class i. When m is two there is one threshold. Values less than the threshold are classified as class 1 while values greater than or equal to the threshold are classified as class 2.

		Actual Condition	
		True	False
Decision	True	True Positive	False Positive (Type I Error)
	False	False Negative (Type II Error)	True Negative

Figure 2-6: Binary classification table.

Myers (1997) describes an example where a site has been contaminated with PCB. The true soil PCB concentration at a location is 22 ppm. A threshold of 25 ppm is determined such that any locations with a PCB concentration greater than 25 ppm must be excavated and treated. If the location is estimated to have a concentration greater than 25 ppm, two errors have been made: estimation error and misclassification error. The misclassification error results in an inappropriate increase in remediation expense and is an example of Type I error. If the location had been estimated to have a concentration less than 22 ppm there would still be estimation error, but no misclassification error.

The true soil PCB concentration at another location is 28 ppm. If the location is estimated to have a concentration less then 25 ppm it goes untreated. This incorrect decision may lead to health risk liabilities the cost of which is difficult to quantify. This example of Type II error has very different consequence from the Type I error.

Figure 2-7 is a visual representation of the two types of misclassification known as a misclassification ellipse. The estimated value is plotted on the x-axis and the true value is plotted on the y-axis. The dashed line at 45° represents values with perfect estimation. Perfect estimation rarely occurs leading to a scatter of points with a roughly elliptical shape (Myers, 1997). The threshold, t, is plotted perpendicular to both axes creating four distinct quadrants. The quadrants labeled I and II represent situations of

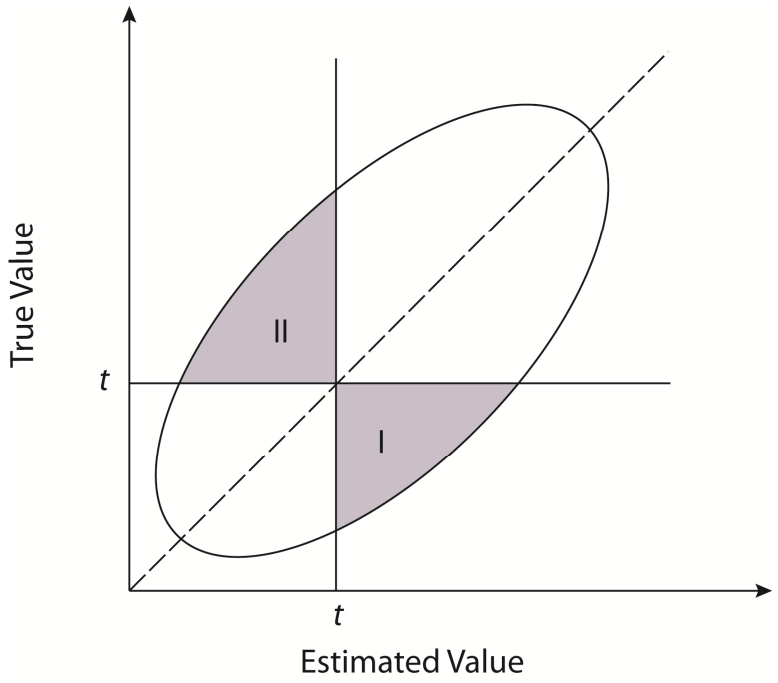

Figure 2-7: Misclassification ellipse.

misclassification corresponding to Type I and Type II errors. The lower right quadrant (Type I) is when the estimated value is greater than t and the true value is less than t. In the upper left quadrant (Type II), the estimated value is less than t but the true value is greater than t.

2.3 Summary

A choice must be made regarding the measures to use for analysis. The choice between data spacing and data density is trivial: if one is known the other can be determined. The choice of uncertainty measure has greater consequence. The measure termed precision herein is common in many applications (Deutsch *et al.*, 2006). The choice of uncertainty measure will depend on the chosen format for expressing acceptable uncertainty. The examples and illustrations provided herein make use of a variety of the measures.

Chapter 3

Simulation Approach to the Determination of Uncertainty versus Data Spacing

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 the determination of an acceptable level of uncertainty.

The methodology is based on sequential Gaussian simulation (SGS). Simulation is a popular method for characterizing uncertainty in earth sciences modeling. It allows the generation of multiple equi-probable

realizations that each honor the input data, histogram, and variogram. 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 that can be applied to a variety of earth science variables.

3.1 Proposed Methodology

The methodology for determining uncertainty at different data spacings is discussed. 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 samples. 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 the goal (Journel and Huijbregts, 1978). The conditional realizations are therefore 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. The procedure outlined assesses uncertainty for one data density d_j.

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.

3.1.1 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 3-1), denoted $\left\{z^{(l)}\left(\mathbf{u}\right), \mathbf{u} \in A, l = 1,...L\right\}$, where $\mathbf{u}$ represents a location and L is the number of truth realizations. These realizations are characterized by a histogram and variogram that 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 3-1) 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 parameters (histogram, variogram) based on expert judgment or an analogue site.

3.1.2 Simulate Data

The next step is to simulate data, denoted as $\left\{z^{(l)}\left(\mathbf{u}_i\right), i = 1,...,n_D; l = 1,...,L\right\}$ 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 3-1).

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

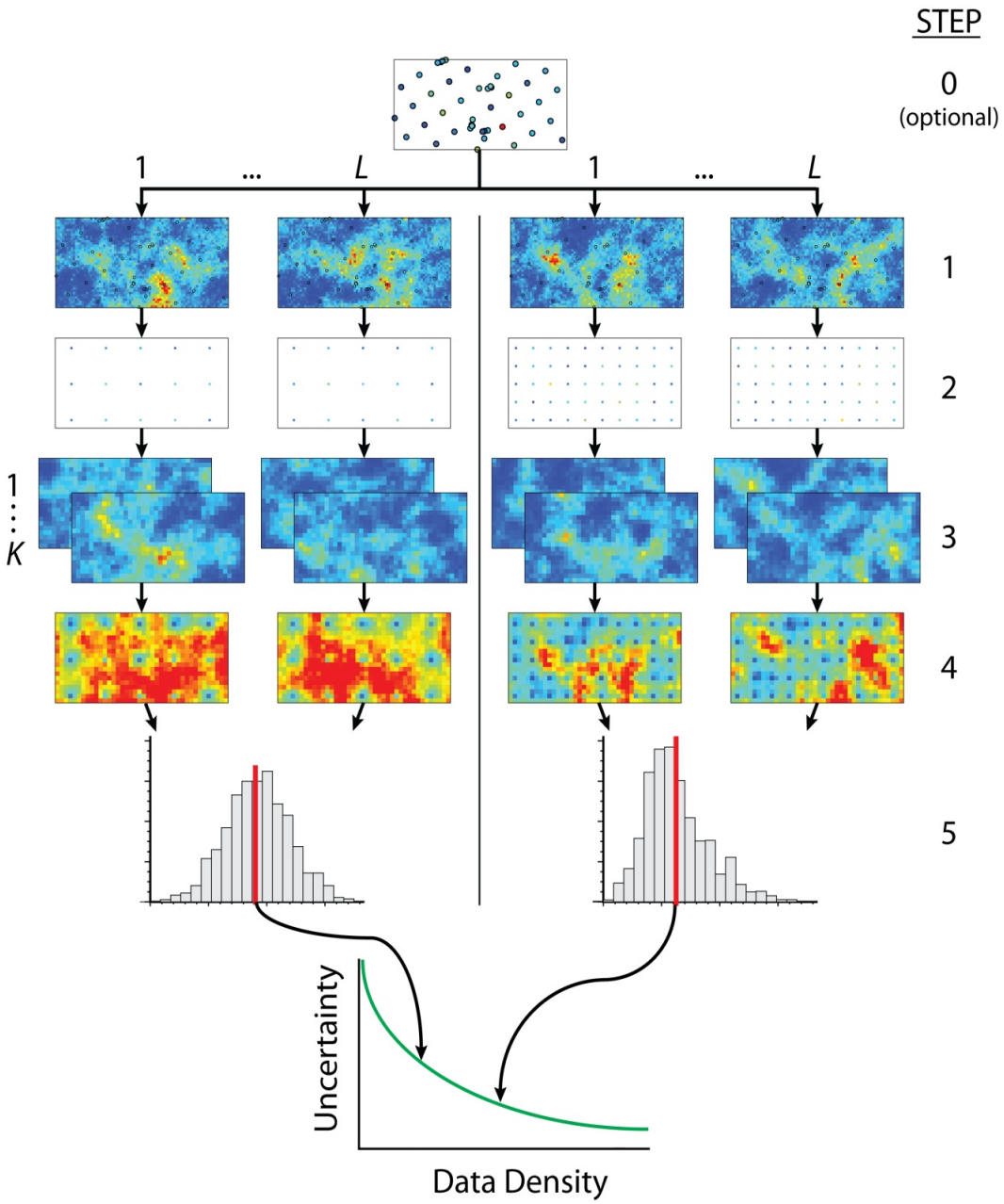

Figure 3-1: Illustration of the proposed methodology: 1) realizations of the truth are generated by sequential Gaussian simulation; 2) the truth is sampled at the desired spacing; 3) realizations are generated conditional to the samples; 4) local measures of uncertainty are calculated from the realizations;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

spread specified by c_s such that the simulated data are defined by Equation 3.1 where $Y_{(0,1)}$ is a random value drawn from the standard normal distribution.

$$z^{(l)}_{\text{data}}\left(\mathbf{u}_i\right) = z^{(l)}\left(\mathbf{u}_i\right) + Y_{(0,1)} \cdot c_s \cdot z^{(l)}\left(\mathbf{u}_i\right),\ i = 1,...,n_D;\ l = 1,...,L \quad 3.1$$

$c_s = 0$ 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 that 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).

3.1.3 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 3-1). 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 $\left\{z^{(l)}_k\left(\mathbf{u}\right),\ \mathbf{u} \in A,\ k = 1,...,K\right\}$ 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 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 then averaging the point grades to the required block dimension to arrive at simulated realizations at the block scale. Local uncertainty may then be assessed at a 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):

$$\overline{z}_k^{(l)}\left(\mathbf{u}\right) = \frac{1}{\left|v\right|}\int_{v\left(\mathbf{u}\right)} z_k^{(l)}\left(\mathbf{u}\right) d\mathbf{u} \simeq \frac{1}{n_v}\sum_{j=1}^{n_v} z_k^{(l)}\left(\mathbf{u}_j\right) \qquad 3.2$$

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

Block kriging could be used as an alternative to simulation to assess uncertainty. It "is computationally quicker and provides a reasonable first approximation to the uncertainty". Simulation, however, is more flexible and "provides a joint measure of uncertainty at all locations simultaneously" (Deutsch and Beardow, 1999).

3.1.4 Calculate Measures of Uncertainty

The uncertainty at location $\mathbf{u}$ 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 3-1. The probability distribution at an unsampled location has non-zero variance, increasing as the location gets farther away from samples.

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. These measures are defined in Chapter 2. The standard deviation, coefficient of variation, P90-P10, (P90-P10)/P50, 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 given truth realization, $\bar{U}^{(l)}$, can be calculated as the average of the local uncertainty measures $U^{(l)}\left(\mathbf{u}\right)$ over all locations as in Equation 3.3 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, $\bar{U}_j$, for a given data density as in Equation 3.4. This is illustrated by step 5 in Figure 3-1.

$$\bar{U}^{(l)} = \frac{1}{n_{\mathbf{u}}} \sum_{i=1}^{n_{\mathbf{u}}} U^{(l)}(\mathbf{u}_i) \tag{3.3}$$

$$\bar{U}_j = \frac{1}{L} \sum_{l=1}^{L} \bar{U}^{(l)} \tag{3.4}$$

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.

3.1.4.1 Standard Deviation and Coefficient of Variation

Standard deviation was defined in Section 2.2.1 as a measure of the spread of a distribution. It is calculated at every block location. The standard deviation at a block location, $\hat{\sigma}^{(l)}\left(\mathbf{u}\right)$, is determined by taking the square root of the variance of the K values comprising the distribution:

$$\hat{\sigma}^{(l)}\left(\mathbf{u}\right)=\left(\frac{1}{K}\sum_{k=1}^{K}\left(\overline{z}_{k}^{(l)}\left(\mathbf{u}\right)-\hat{\mu}_{\overline{z}}^{(l)}\left(\mathbf{u}\right)\right)^{2}\right)^{1/2} \tag{3.5}$$

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

$$\hat{\mu}_{\overline{z}}^{(l)}\left(\mathbf{u}\right)=\frac{1}{K}\sum_{k=1}^{K}\overline{z}_{k}^{(l)}\left(\mathbf{u}\right) \tag{3.6}$$

The spread of the block distribution is small near data and increases as the block gets further from 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, $CV^{(l)}\left(\mathbf{u}\right)$, is the standard deviation divided by the mean:

$$CV^{(l)}\left(\mathbf{u}\right)=\frac{\hat{\sigma}^{(l)}\left(\mathbf{u}\right)}{\hat{\mu}_{\overline{z}}^{(l)}\left(\mathbf{u}\right)} \tag{3.7}$$

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 3.3 and 3.4 and substituting $\boldsymbol{\sigma}$ for U to get Equation 3.8. The expected coefficient of variation is determined in like manner by applying Equation 3.9.

$$\overline{\sigma}_{j}=\frac{1}{n_{\mathrm{u}}\cdot L}\sum_{l=1}^{L}\sum_{i=1}^{n_{\mathrm{u}}}\hat{\sigma}^{(l)}\left(\mathbf{u}_{i}\right) \tag{3.8}$$

$$\overline{CV}_{j}=\frac{1}{n_{\mathrm{u}}\cdot L}\sum_{l=1}^{L}\sum_{i=1}^{n_{\mathrm{u}}}CV^{(l)}\left(\mathbf{u}_{i}\right) \tag{3.9}$$

3.1.4.2 Difference between Percentiles

Percentiles were defined in Section 2.2.2. 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 $\overline{z}_k^{(l)}\left(\mathbf{u}\right) \leq \overline{z}_{k+1}^{(l)}\left(\mathbf{u}\right), k = 1,...,K-1$. Percentiles of interest can then be located. The difference between the 10^{th} and 90^{th} percentiles provides a reasonable measure of spread. This difference can be standardized by dividing by the 50^{th} percentile. The 10^{th}, 50^{th}, and 90^{th} percentiles at a location are denoted as $\overline{z}_{P10}^{(l)}\left(\mathbf{u}\right)$, $\overline{z}_{P50}^{(l)}\left(\mathbf{u}\right)$, and $\overline{z}_{P90}^{(l)}\left(\mathbf{u}\right)$ respectively.

The difference between the 10^{th} and 90^{th} percentiles at a location, denoted $\Delta^{(l)}\left(\mathbf{u}\right)$ and defined in Equation 3.10, 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. There are typically more locations far from data than there are locations close to data causing the distribution of differences between percentiles to be negatively skewed.

$$\Delta^{(l)}\left(\mathbf{u}\right) = \overline{z}_{P90}^{(l)}\left(\mathbf{u}\right) - \overline{z}_{P10}^{(l)}\left(\mathbf{u}\right) \tag{3.10}$$

The standardized difference between the 10^{th} and 90^{th} percentiles, denoted $\Delta_s^{(l)}\left(\mathbf{u}\right)$ and defined in Equation 3.11, is a unitless measure of the spread of a distribution. It's unitless nature makes it amenable for comparing distributions with different units or different means.

$$\Delta_s^{(l)}\left(\mathbf{u}\right) = \frac{\Delta^{(l)}\left(\mathbf{u}\right)}{\overline{z}_{P50}^{(l)}\left(\mathbf{u}\right)} \tag{3.11}$$

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, $\overline{\Delta}^j$, over all locations and data realizations for a given data

density, d_j, is determined by combining Equations 3.3 and 3.4 and substituting Δ for U as in Equation 3.12. The determination of the expected value of the standardized difference between percentiles, $\overline{\Delta}_s^j$, over all locations and realizations is shown in Equation 3.13.

$$\overline{\Delta}^j = \frac{1}{n_{\mathbf{u}} \cdot L} \sum_{l=1}^{L} \sum_{i=1}^{n_{\mathbf{u}}} \Delta^{(l)}\left(\mathbf{u}_i\right) \tag{3.12}$$

$$\overline{\Delta}_s^j = \frac{1}{n_{\mathbf{u}} \cdot L} \sum_{l=1}^{L} \sum_{i=1}^{n_{\mathbf{u}}} \Delta_s^{(l)}\left(\mathbf{u}_i\right) \tag{3.13}$$

3.1.4.3 Precision

The precision of a distribution was defined in Section 2.2.3. The precision at a location, $p^{(l)}\left(\mathbf{u}\right)$, is the proportion of simulated block values, $\overline{z}_k^{(l)}\left(\mathbf{u}\right)$, that fall within a specified distance, $h^{(l)}\left(\mathbf{u}\right)$, from the mean, $\hat{\mu}_{\overline{z}}^{(l)}\left(\mathbf{u}\right)$, at location $\mathbf{u}$. 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, $h^{(l)}\left(\mathbf{u}\right)$, is specified by a multiplicative constant, r :

$$h^{(l)}\left(\mathbf{u}\right) = r \cdot \hat{\mu}_{\overline{z}}^{(l)}\left(\mathbf{u}'_i\right) \tag{3.14}$$

Let $\tau_k^{(l)}\left(\mathbf{u};h\right)$ be a binary indicator such that Equation 3.15 is satisfied. The precision at a location, $p^{(l)}\left(\mathrm{u}\right)$, is defined in Equation 3.16.

$$\tau_k^{(l)}\left(\mathbf{u};h\right) = \begin{cases} 1, \text{ if } \hat{\mu}_{\overline{z}}^{(l)}\left(\mathbf{u}\right) - h^{(l)}\left(\mathbf{u}\right) \leq \overline{z}_k^{(l)}\left(\mathbf{u}\right) \leq \hat{\mu}_{\overline{z}}^{(l)}\left(\mathbf{u}\right) + h^{(l)}\left(\mathbf{u}\right) \\ 0, \text{ otherwise} \end{cases} \tag{3.15}$$

$$p^{(l)}\left(\mathbf{u}\right) = \frac{1}{K}\sum_{k=1}^{K}\tau_k^{(l)}\left(\mathbf{u};h\right) \qquad 3.16$$

The expected precision over all locations and data realizations, $\overline{p}_j$, is obtained by combining Equations 3.3 and 3.4, substituting p for U as in Equation 3.17.

$$\overline{p}_j = \frac{1}{n_{\mathbf{u}} \cdot L}\sum_{l=1}^{L}\sum_{i=1}^{n_{\mathbf{u}}} p^{(l)}\left(\mathbf{u}_i\right) \qquad 3.17$$

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

3.1.4.4 Probability of Misclassification

The probability of misclassification is defined in Section 2.2.4. Two categories are considered requiring one threshold, t, to define them. The only way to know whether an observation has been misclassified is to know the truth. The block average of the realization simulated in the first step, $\overline{z}^{(l)}\left(\mathbf{u}\right)$, $\mathbf{u} \in A$, $l = 1,...,L$, 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 $\alpha^{(l)}\left(\mathbf{u}\right)$ represent the probability of making a Type I error at location $\mathbf{u}$ and let $\beta^{(l)}\left(\mathbf{u}\right)$ represent the probability of making a Type II error at

location $\mathbf{u}$. Consider first the case where the truth at a location, $\overline{z}^{(l)}\left(\mathbf{u}\right)$, is greater than or equal to the threshold, t. The probability of a false positive, $\alpha^{(l)}\left(\mathbf{u}\right)$, is zero while the probability of a false negative, $\beta^{(l)}\left(\mathbf{u}\right)$, may be non-zero. The probability of a false negative at this location is the number of simulated block values that are less than the threshold divided by K. Let $\phi_k^{(l)}\left(\mathbf{u}\right)$ be a binary indicator defined by Equation 3.18. The probability of a false negative at this location is defined by Equation 3.19.

$$\phi_k^{(l)}\left(\mathbf{u}\right) = \begin{cases} 1, \text{ if } \overline{z}_k^{(l)}\left(\mathbf{u}\right) < t \\ 0, \text{ otherwise} \end{cases}, k = 1,...,K \tag{3.18}$$

$$\beta^{(l)}\left(\mathbf{u}\right) = \frac{1}{K}\sum_{k=1}^{K}\phi_k^{(l)}\left(\mathbf{u}\right) \tag{3.19}$$

Next, consider the case where the truth at $\mathbf{u}$, $\overline{z}^{(l)}\left(\mathbf{u}\right)$, 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 $\varphi_k^{(l)}\left(\mathbf{u}\right)$ be a binary indicator defined by Equation 3.20. The probability of a false positive at $\mathbf{u}$ is defined by Equation 3.21.

$$\varphi_k^{(l)}\left(\mathbf{u}\right) = \begin{cases} 1, \text{ if } \overline{z}_k^{(l)}\left(\mathbf{u}\right) \geq t \\ 0, \text{ otherwise} \end{cases}, k = 1,...K \tag{3.20}$$

$$\alpha^{(l)}\left(\mathbf{u}\right) = \frac{1}{K}\sum_{k=1}^{K}\varphi_k^{(l)}\left(\mathbf{u}\right) \tag{3.21}$$

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. As the spacing between data increases, so will this measure of uncertainty.

3.2 Implementation Example

Consider the determination of uncertainty versus 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 for an area of this size, 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 3-2. 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.

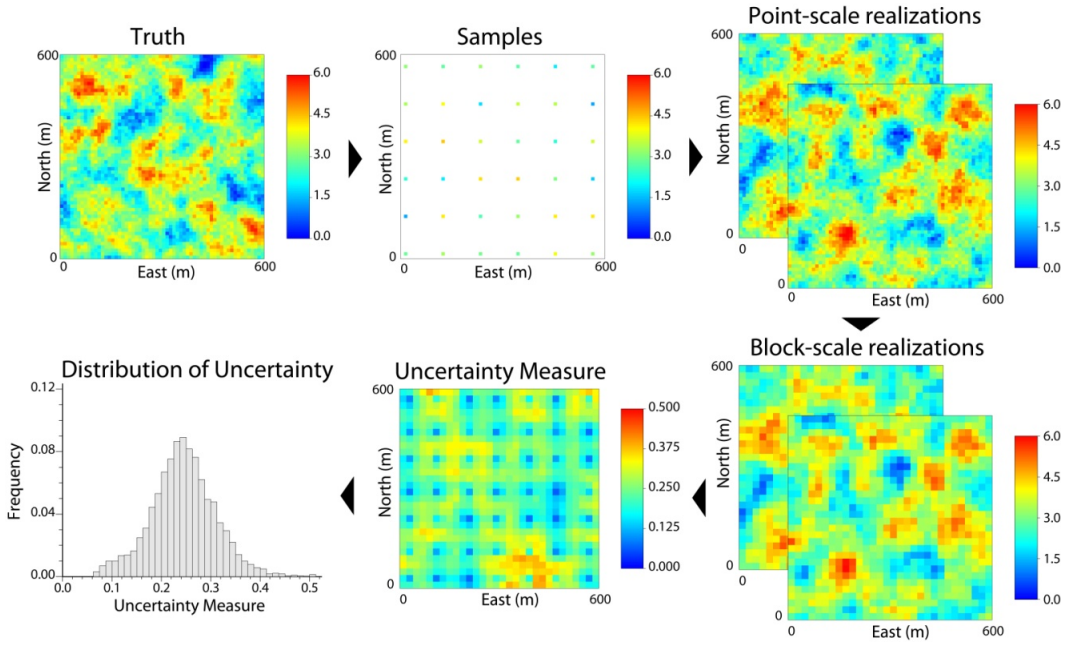

Figure 3-2: Illustration of the steps for the implementation example.

This same process is repeated for each of the five spacings. The result is five distributions of uncertainty. The five distributions of the coefficient of variation are shown in Figure 3-3. 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.

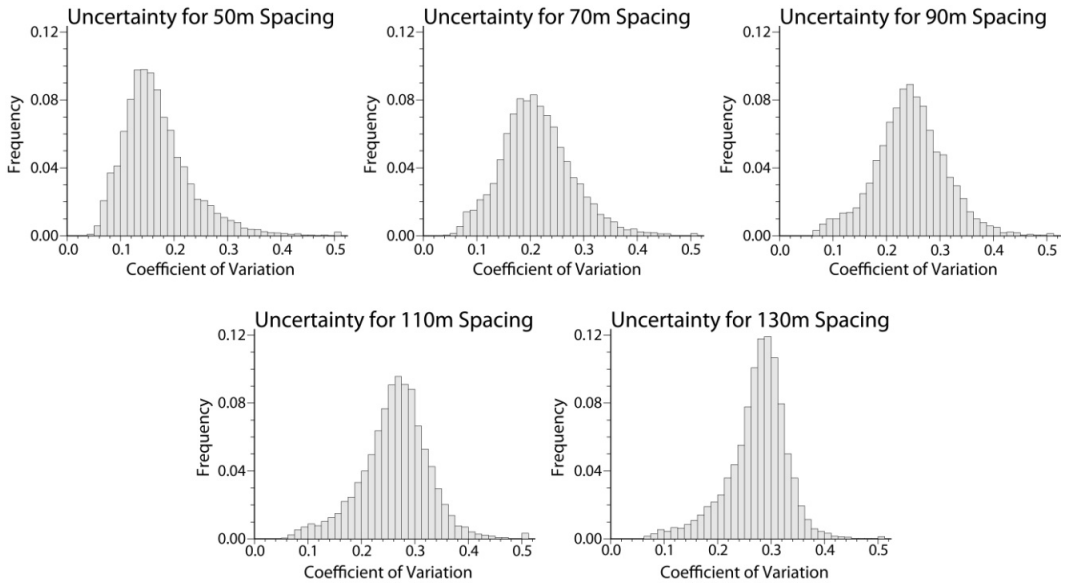

Figure 3-3: Uncertainty distributions for data spacings of 50, 70, 90, 110, and 130m respectively.

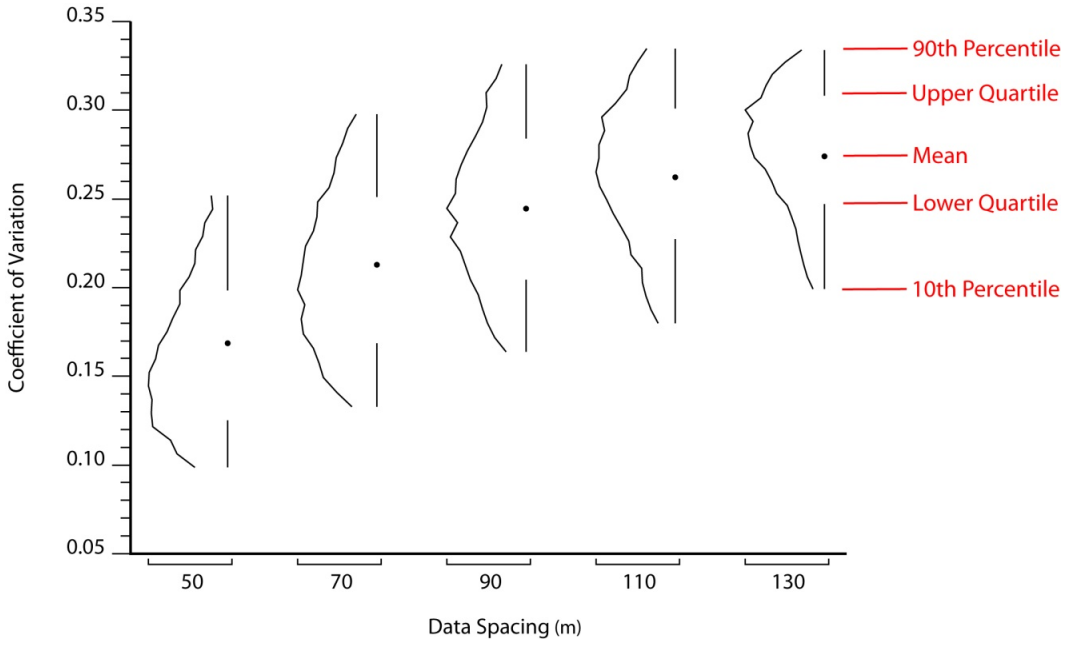

Figure 3-4: Uncertainty distributions for different data spacings with a description of the various markings on the plot (red).

Another method for visualizing this relationship is shown in Figure 3-4. This is a plot of uncertainty versus data spacing and shows the same five distributions as Figure 3-3. The relationship between uncertainty and data spacing is more easily discerned viewing the distributions in this manner. The markings that summarize the distributions are described in the figure. 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 that 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 3-5). 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 3-6. 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

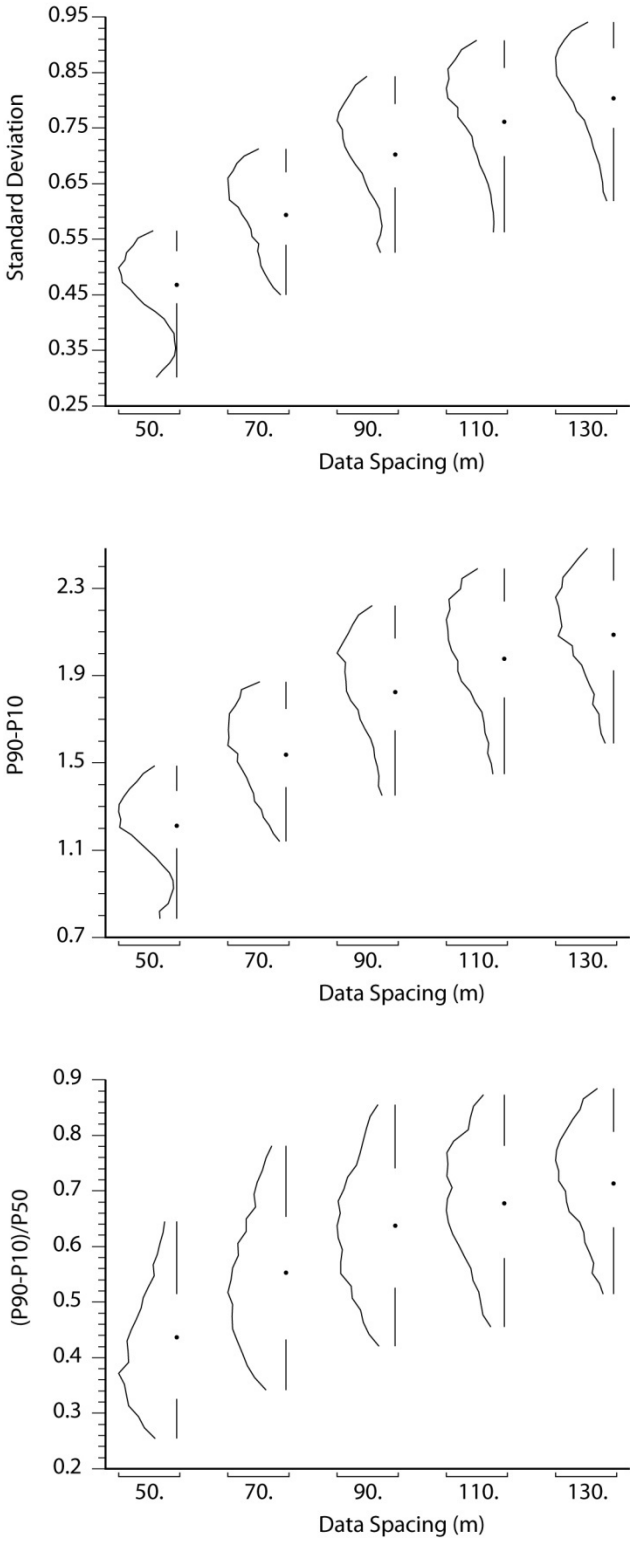

Figure 3-5: Relationships between standard deviation, P90-P10, (P90-P10)/P50, and data spacing.

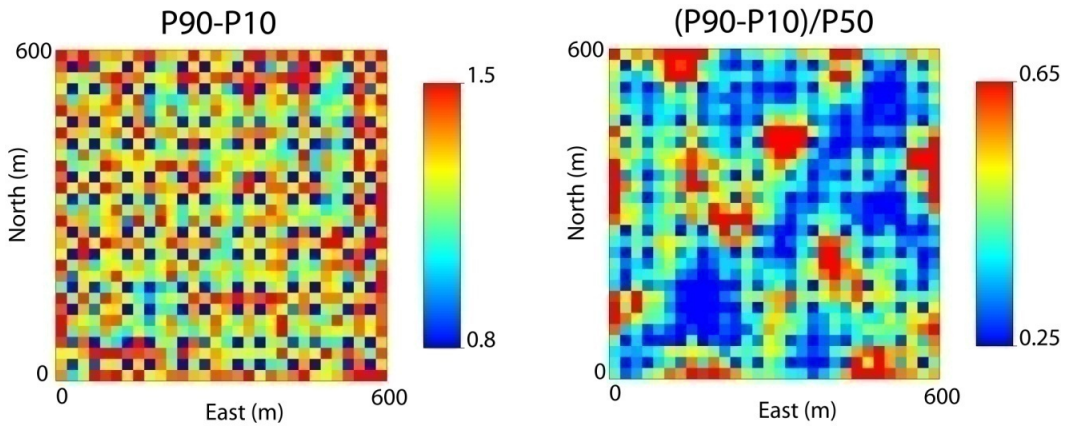

Figure 3-6: Non-standardized and standardized measure of uncertainty for 50m spacing.

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 3-7). 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 3-7; only the erased box plot is shown. The line representing the 10^{th} and 25^{th} percentiles does not appear

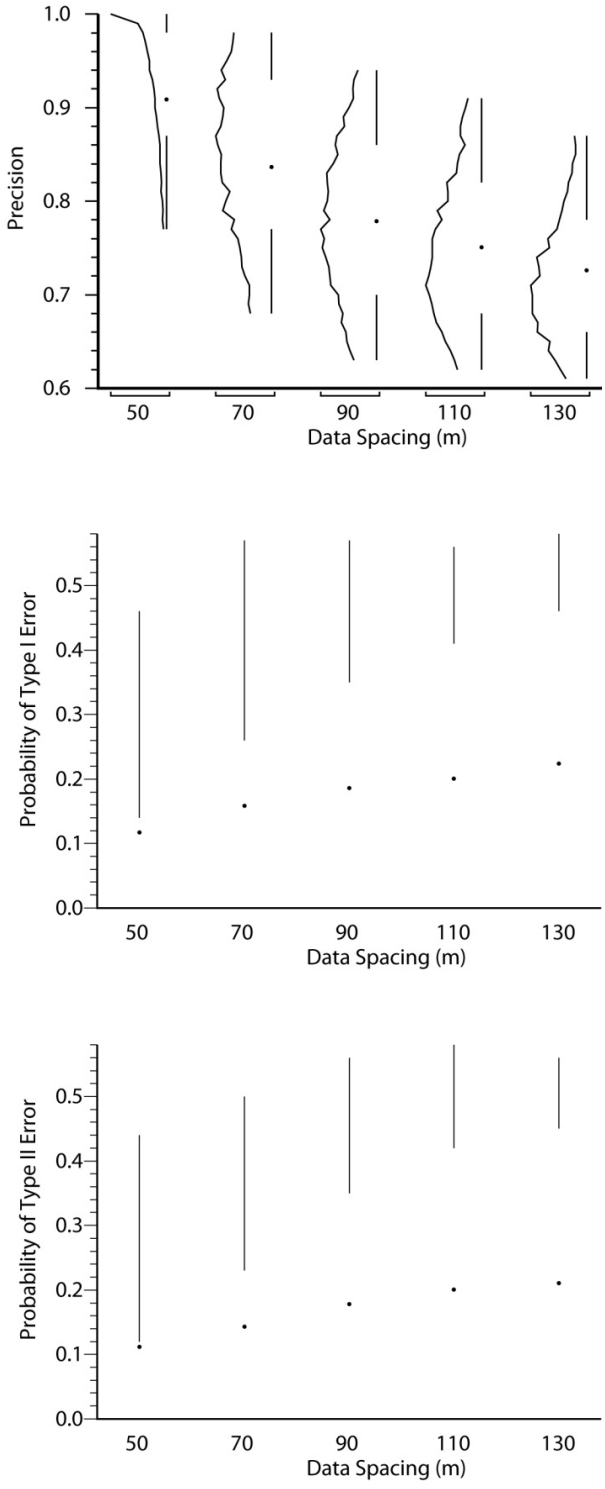

Figure 3-7: Relationships between precision, Type I error, Type II error, and data spacing.

as both values are zero. Those summaries that are visible (mean, upper quartile, 90^{th} percentile) show that the occurrence of each type of misclassification increases as data spacing increases. 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%. Another example is, 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.

3.3 Limitations

The proposed methodology is constrained by a number of limitations. One of these limitations is the need for a sufficient quantity of data to allow the analysis to proceed. Applying the methodology requires an understanding of the spatial distribution of the attribute which in turn requires an initial quantity of data. It must be determined whether sufficient data is available to adequately characterize the attribute prior to performing the proposed methodology. Parameters such as the histogram and variogram are required. It is difficult to know if the attribute is sufficiently described by the existing data.

Another limitation of the methodology is its inability to evaluate uncertainty in the geology. This SGS-based methodology is suitable for evaluating uncertainty due to the histogram and variogram of an attribute. Assessing uncertainty in the geology would require implementing a similar methodology based on multiple-point statistics (MPS) or object-based modeling. This would allow the relationship between data quantity and geological uncertainty to be assessed.

Another limitation of the methodology is its applicability to settings where the calculation of data spacing cannot be easily reduced to two dimensions such as in an underground mining context. In underground mining data are often collected from drillholes which are not parallel. Determining a relationship between data spacing and uncertainty for such sampling schemes is not straightforward.

Chapter 4

Confounding Factors

Uncertainty and data density are closely related. As data density increases, more is known about the variable of interest and uncertainty decreases. When data density is low, less is known about the variable of interest and uncertainty is higher. Data density is not the only factor controlling uncertainty. A number of other confounding factors play a role in determining uncertainty for a given data spacing. These factors include stationarity, parameter uncertainty, the proportional effect, nonstationarity in the variogram, classification thresholds, scale, number of realizations, and data quality. The effect of each of these factors is discussed.

4.1 Stationarity, Parameter Uncertainty and Model Uncertainty

Prior to performing the methodology described herein, decisions must be made regarding the pooling of data and the input parameters. In addition, various geostatistical assumptions must be made such as the adoption of the multivariate Gaussian model. These decisions and assumptions can affect uncertainty.

Stationarity refers to the decision made regarding pooling of the data. Any statistical analysis requires a decision of stationarity. This decision allows inference. It may be appropriate to subdivide data based on

geological facies; however, dividing data into too many categories can lead to unreliable statistics. The decision of how to subdivide the data will affect uncertainty.

Uncertainty in the input parameters such as the histogram and variogram will have an effect on uncertainty in the model. It is common to consider these global statistical parameters as fixed with no uncertainty. This can lead to underestimation of the uncertainty. For example, it is common to consider the declustered histogram as known and fixed. A bootstrap procedure could be implemented that would allow the uncertainty in the input distribution to be assessed. This uncertainty could then be transferred to the geostatistical models. It has been shown that considering parameter uncertainty can increase model uncertainty (Deutsch, *et al.*, 2006).

The spatial continuity parameters such as the nugget effect can also affect uncertainty. It can be difficult to establish the short-scale variability for distances less than the smallest data spacing, yet the choice of the variograms' behavior at the origin can greatly impact the geostatistical models (Dubrule, 1994), at times in a non-intuitive and non-transparent manner. For example, an increase in the nugget effect could intuitively imply an increase in uncertainty. It has been shown that an increase in the nugget effect can, in fact, decrease uncertainty (Deutsch *et al.*, 2006), particularly at large scale.

Various geostatistical assumptions such as the assumption of the multivariate Gaussian distribution in Gaussian simulation can impact uncertainty. This is a common assumption in geostatistical modeling. Most geostatistical estimation and simulation techniques rely on a covariance model as the sole descriptor of the spatial distribution of the attribute being modeled and the multivariate Gaussian distribution for all high order distributions. The methodology described herein is based on sequential Gaussian simulation. One advantage of using this distribution

is that it maximizes entropy beyond the statistics that are considered known. This minimizes unwarranted structural properties. However, this does not lead to maximum entropy in response variables calculated from modeled variables. As noted in Journel and Deutsch (1993), "maximum entropy of the random function model does not entail maximum entropy of the response distributions; in fact, the contrary is observed for most response variables." This could cause the space of response uncertainty to be too small.

4.2 Proportional Effect

The proportional effect is a well documented aspect of earth sciences modeling (Journel and Huijbregts, 1978; Goovaerts, 1997). It refers to the phenomena of the spread of a distribution being related to the magnitude of the distribution center. It occurs when a random variable has a skewed distribution. The proportional effect increases uncertainty for positively skewed distributions when local estimates are high. It increases uncertainty for negatively skewed distributions when local estimates are low.

To examine the effect of the proportional effect on uncertainty, the proposed methodology is implemented using the same parameters as the implementation example in Section 3.2. In addition to considering a symmetric reference distribution with mean=3.0 and standard deviation=1.0, a lognormal reference distribution with the same mean and standard deviation is considered. These distributions are shown in Figure 4-1. The skewness of this lognormal distribution is small, but sufficient for this illustration. Recall that the methodology implementation is within an area 600m x 600m and populated with point-scale simulated values at 10m x 10m spacing averaged into 20x20m blocks. Data spacings of 50, 70, 90, 110, and 130m are considered. The variogram is single structure spherical with no nugget and 100m range.

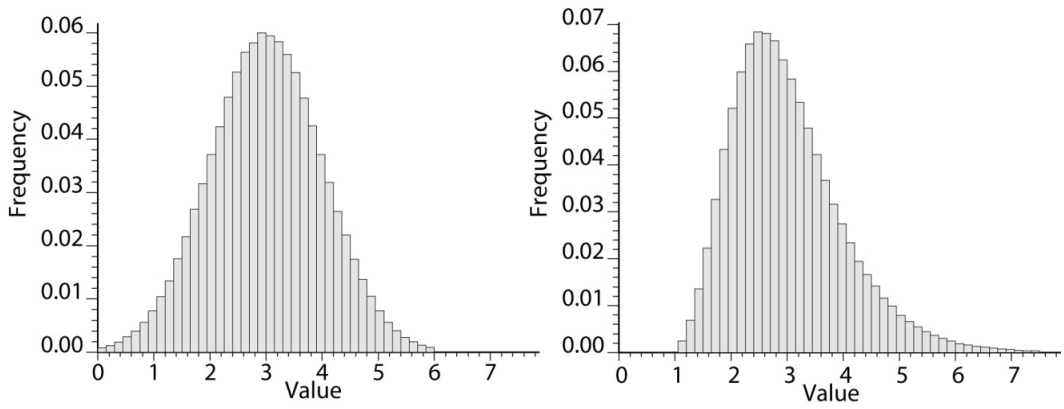

Figure 4-1: Symmetric and skewed reference distributions with mean=3.0 and variance = 1.0.

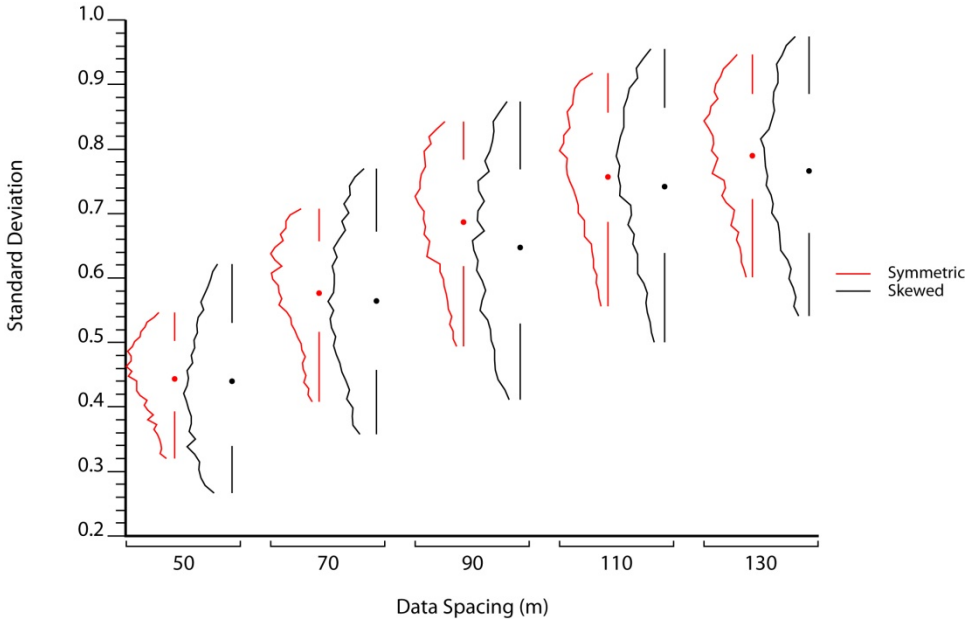

Figure 4-2: Standard deviation versus data spacing for different reference distributions.

For the two cases, the truth realizations are generated unconditionally in Gaussian units and are back transformed according to the reference distribution. These realizations are sampled at the desired spacing and these samples are used to generate conditional realizations. The conditional realizations are then used to assess uncertainty for the given data spacings.

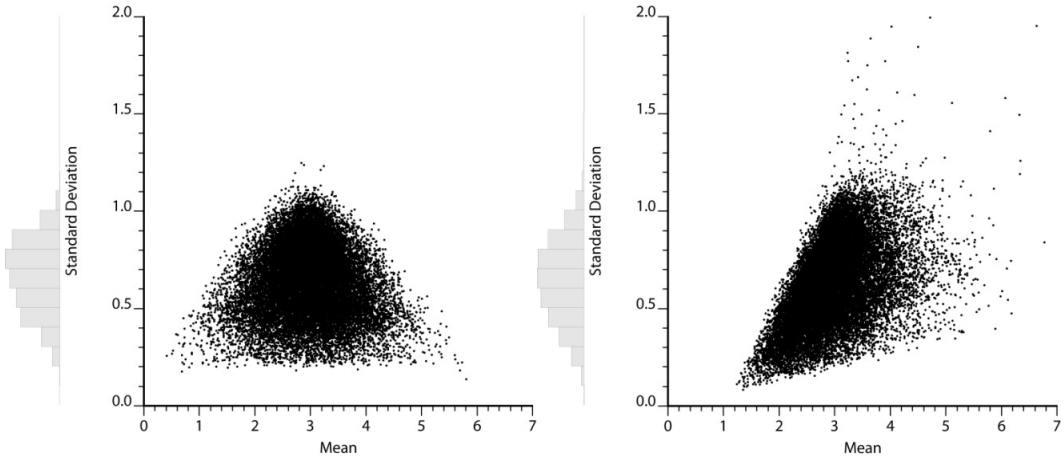

Figure 4-3: Standard deviation versus mean for a symmetric reference distribution (left) and a skewed reference distribution (right).

The pair of reference distributions produces the results shown in Figure 4-2. The spread of uncertainty for the skewed distribution is greater than the spread for the symmetric distribution while the expected uncertainty value is less for the skewed distribution than for the symmetric. A positively skewed distribution with the same mean and variance as a symmetric distribution has a larger proportion of low values than the symmetric distribution. For example, the symmetric distribution considered, exactly 50% of the values are less than the mean while 56.4% of the values in the lognormal distribution are less than the mean. This increased proportion of low values means that most locations have low uncertainty. This large proportion of low uncertainty reduces the expected uncertainty value for a skewed distribution. There are some high-valued areas that are associated with large uncertainty. These values of high uncertainty increase the spread of the uncertainty distribution.

Consider the plots of standard deviation versus mean in Figure 4-3. These plots correspond to the distributions shown in Figure 4-2 for a spacing of 70m. The spread in uncertainty due to the skewed reference distribution is more than the spread from a symmetric distribution. The plots in

Figure 4-3 also illustrate the proportional effect. The standard deviation is dependent on the mean for a skewed reference distribution.

4.3 Nonstationarity in the Variogram

The term stationarity refers to the decision to pool data together for subsequent analysis and the location-dependence of statistical parameters (Deutsch, 2002). In geostatistics, the two statistics commonly assumed constant across a domain are the mean and the variogram. This assumption may not always be valid. Variations in these parameters can affect uncertainty. Variations in the mean will lead to increased uncertainty if the distribution is skewed; see previous section on the proportional effect. Variations in the variogram will also affect uncertainty. Areas where the attribute is more continuous will have less uncertainty than areas where the attribute is less continuous.

To examine nonstationarity in the variogram, the proposed methodology is implemented with three different variograms. All three models are one structure spherical with no nugget. The ranges of the variograms are 50, 100, and 200m (top of Figure 4-4). The realizations are built using the different variograms and used to calculate the measures of uncertainty. The bottom plot in Figure 4-4 shows the uncertainty distributions for the different variograms at different data spacings. As expected, the short range variogram results in the greatest overall uncertainty. Figure 4-4 demonstrates that uncertainty decreases as continuity increases. Note the similarity in the pattern of these distributions with the variogram models.

The bottom plot in Figure 4-4 reveals an additional aspect of the uncertainty versus data spacing relationship: the spread in the distributions of uncertainty is related to the magnitude of the data spacing relative to the variogram range. There is greater variability in uncertainty for data spacings near the variogram range and less variability in

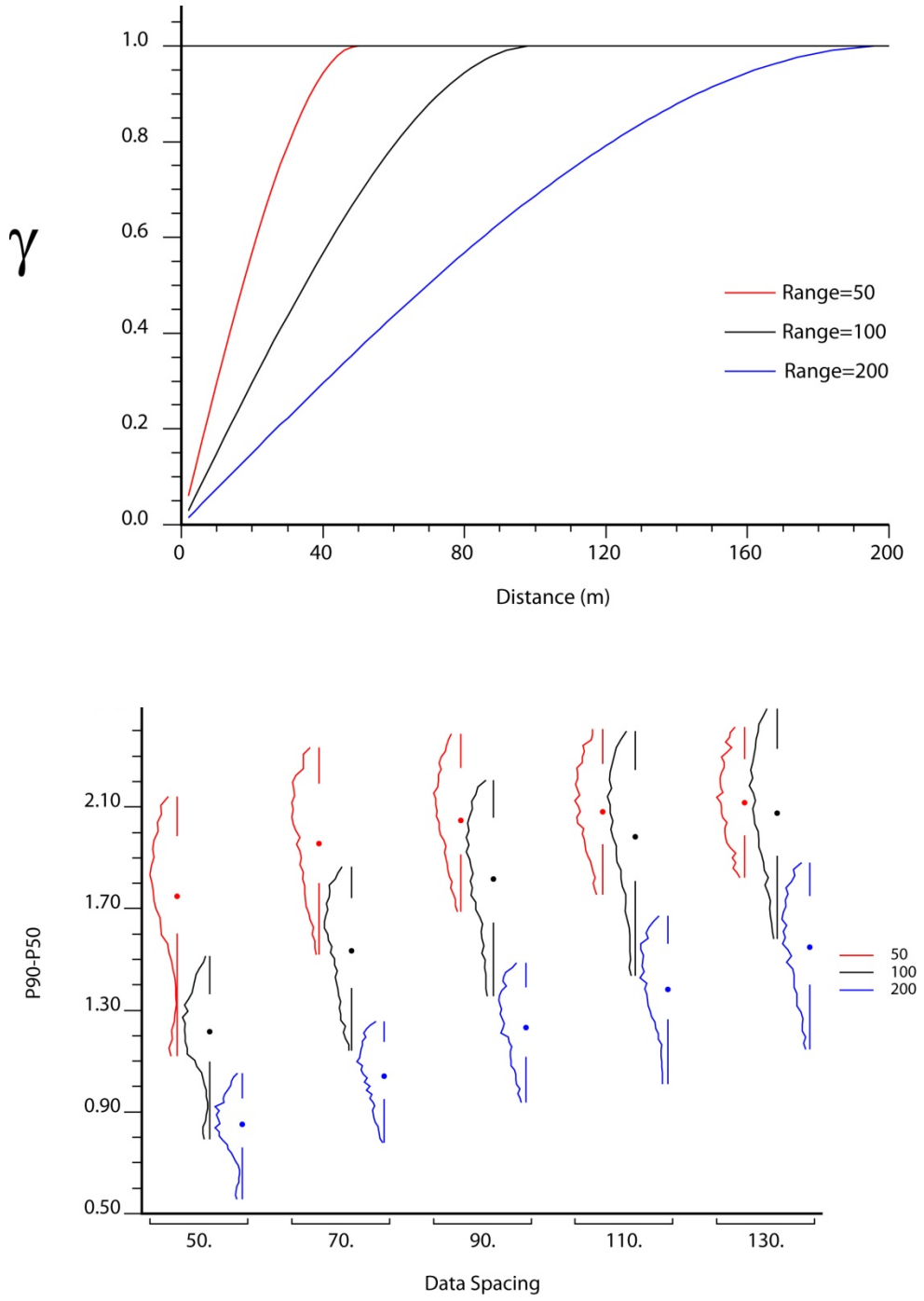

Figure 4-4: Three variogram models (top) used to examine the effect of a nonstationary variogram on uncertainty and the comparison of P90-P50 versus data spacing results for different data spacings (bottom).

uncertainty for spacings less than and greater than the range. Consider first the uncertainty distributions resulting from a variogram with a range of 50m (red). The spread in these distributions is greatest when the spacing is equal to the variogram range and decreases as the spacing becomes greater than the range. Next consider the uncertainty distributions resulting from a variogram with range of 200m (blue). The spread in these distributions is least when the spacing is much less than the variogram range and increases as the spacing approaches the variogram range.

An example of nonstationarity in the variogram comes from bitumen thickness data of the McMurray formation in northern Alberta (Warren, 2003). Two areas are considered, one in the north and one in the south as shown in Figure 4-5. Omnidirectional variograms are calculated from the data in these two areas. The experimental and modeled variograms for these two areas are shown in Figure 4-6. The south area has better correlation at the distances shown than the north area. This increased correlation translates into reduced uncertainty as shown by the P90-P10 versus data spacing plot in Figure 4-7. For the four data spacings considered, uncertainty is lower for the south than for the north.

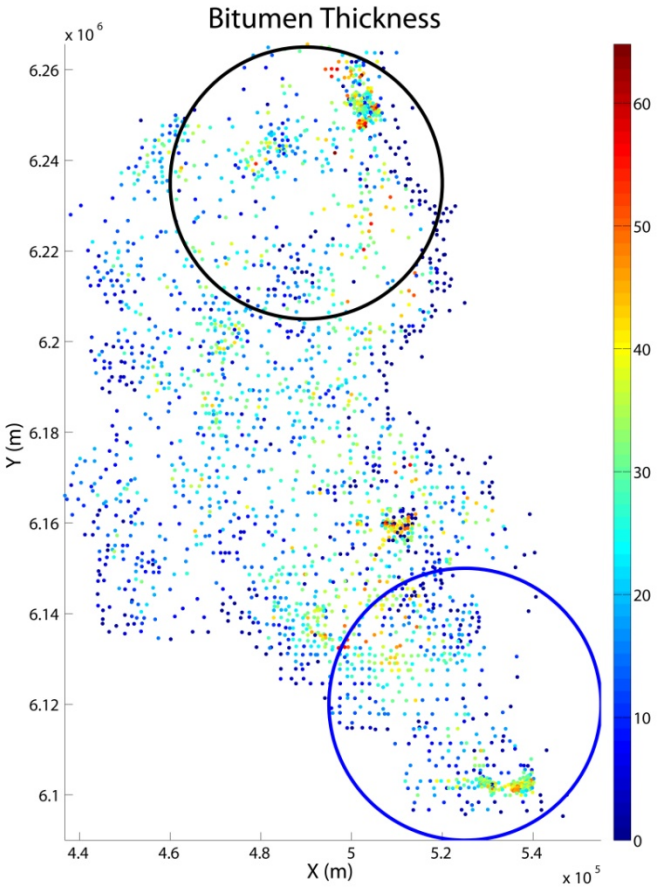

Figure 4-5: Location map of bitumen thickness data showing the two areas considered for nonstationarity in the variogram.

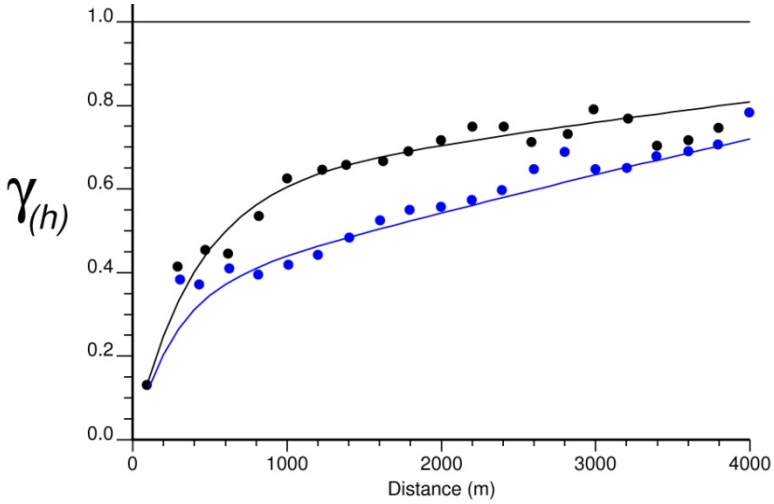

Figure 4-6: Omnidirectional normal-score variograms of the north (black) and south (blue) areas shown in Figure 4-5.

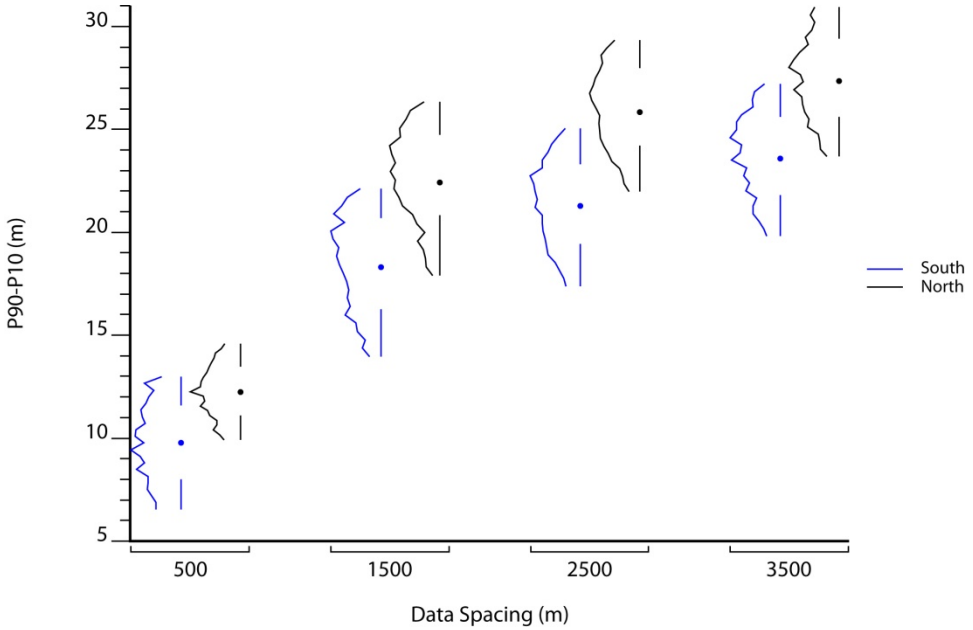

Figure 4-7: P90-P10 versus data spacing for the variogram models shown in Figure 4-6.

4.4 Classification Threshold

Estimates can be classified based on their value with respect to a classification threshold. The value of the classification threshold has an effect on uncertainty. Most of the uncertainty measures considered herein are unaffected by changes in the value of the classification threshold. Only the two probability of misclassification measures are affected. The relationship between the classification threshold and these measures is discussed and illustrated.

To know whether an estimate has been misclassified requires knowledge of the true classification. The truth is, of course, inaccessible without exhaustive sampling. The truth is assumed to be represented by the simulated point-scale realization from which the samples are taken. At each location, a probability of misclassification can be determined. The type of misclassification depends on the value of the truth with respect to the threshold while the probability of misclassification depends on the local distribution. A high probability of making a Type I error signifies

that the truth is greater than the threshold while a large proportion of the simulated values falls below the threshold. A high probability of making a Type II error signifies that the truth is less than the threshold while a large proportion of the simulated values are greater than the threshold.

To illustrate, consider one slice from a 2-D simulation of normally distributed values with a mean of 3.0 units. Values were simulated every 10m within a 600m x 600m area. The red line shown in the top left of Figure 4-8 represents the truth for the given slice. The truth is sampled every 90m as represented by the black dots. These samples are used to generate 100 conditional realizations shown in the top right of Figure 4-8. The probability of misclassification at each location is determined by considering the truth and the conditionally simulated values and applying a threshold (black) as shown in the middle plot of Figure 4-8. When simulated values fall on the opposite side of the threshold than the truth, there is non-zero probability of misclassification. The blue line in the bottom plot of Figure 4-8 represents the probability of Type I error for each location while the green line in the same plot represents the probability of Type II error for each location. These probabilities are determined by counting the number of simulated values that fall on the opposite side of the threshold than the truth and dividing by the total number of simulated values.

The threshold for this example is the mean value of 3.0. The relative occurrence of each type of misclassification is controlled by the value of the threshold relative to the distribution of true values. In this case, the values are normally distributed and the threshold falls in the middle of the distribution. This leads to an approximately equal number of Type I and Type II errors.

As discussed, a different threshold leads to different possibilities of each type of misclassification error. Consider the effect of reducing the threshold to a value of 2.0 as shown in Figure 4-9. The truth at a

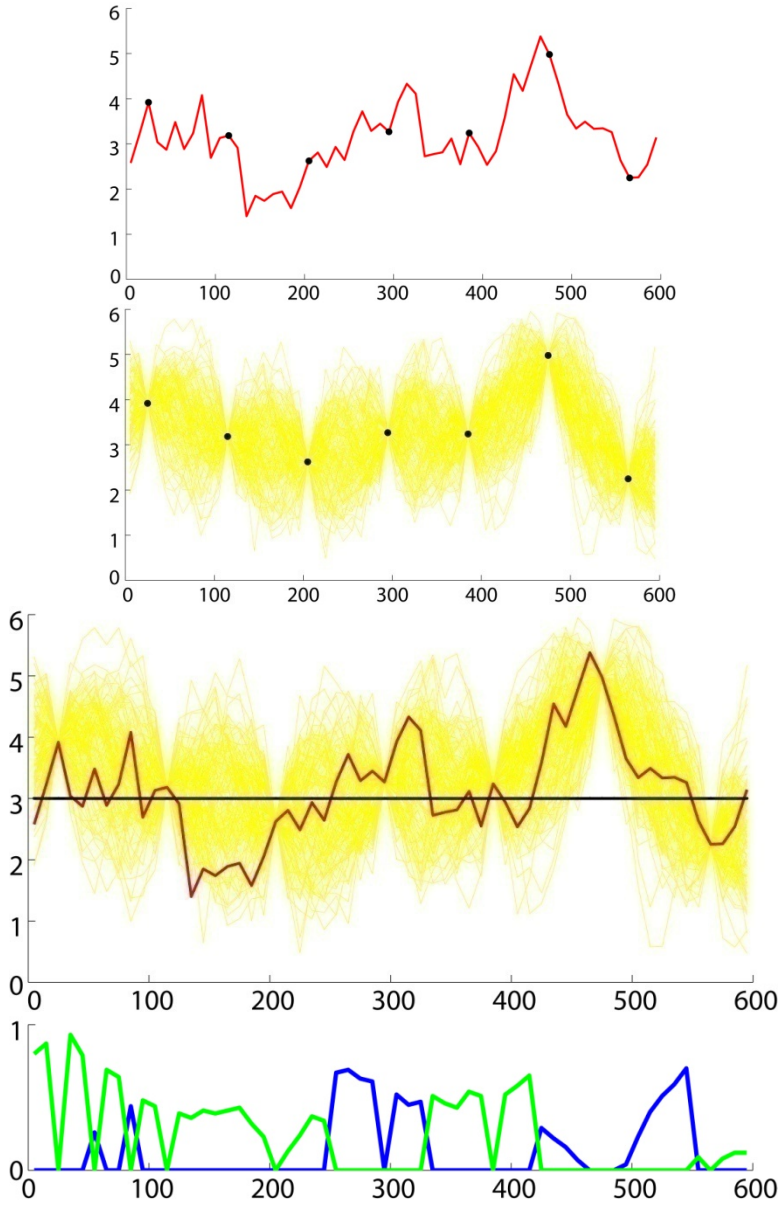

Figure 4-8: One slice from a 2-D simulation example to demonstrate the factors that influence the probability of misclassification.

majority of locations is greater than the threshold. At each of these locations the possibility of Type I misclassification error exists. However, the probability of Type I error is small as the number of simulated values falling below the threshold is small. Conversely, there are few locations where the truth is less than the threshold and therefore a small possibility

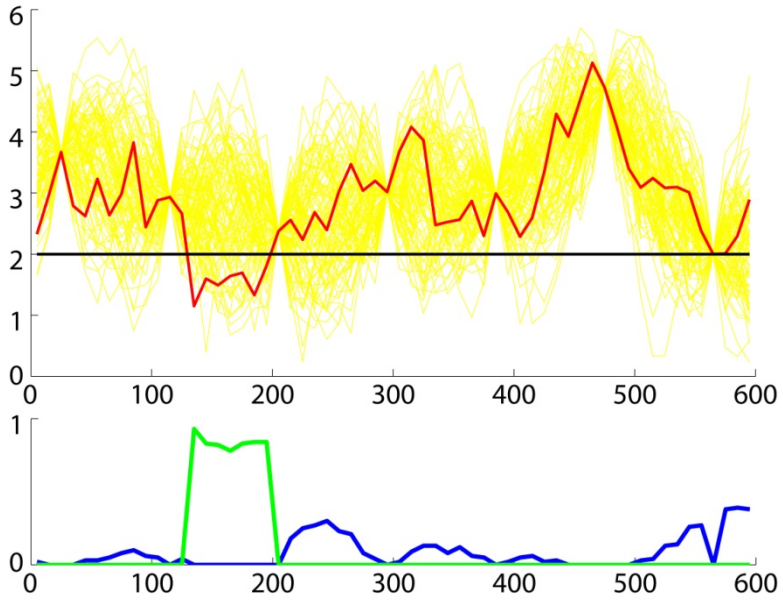

Figure 4-9: Truth (red), simulated values (yellow), and a threshold (black) of 2.0 leading to the probabilities of Type I (blue) and Type II (green) error.

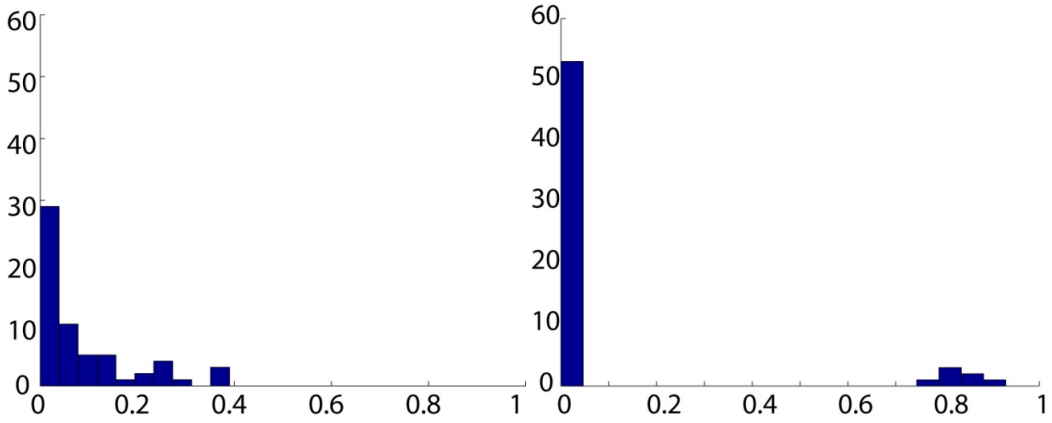

Figure 4-10: Distributions of Type I error (left) and Type II error (right) for a threshold of 2.0 relating to the blue and green lines in Figure 4-9 respectively.

of Type II misclassification error. However, when the truth does fall below the threshold, there are a large number of simulated values greater than the threshold and the probability of Type II error is high. The distributions of Type I and Type II error for a threshold of 2.0 are shown

in Figure 4-10. There are many low-valued Type I probabilities for this threshold while Type II has mostly zeros with a few high values.

Increasing the threshold has the opposite effect. Consider the result of increasing the threshold to a value of 4.0 as shown in Figure 4-11. The truth at the majority of locations is less than the threshold leading to an increased possibility of Type II error and a decreased possibility of Type I error. When the threshold is greater than the truth, there are typically few simulated values greater than the threshold leading to low probabilities of Type II error. When the truth is greater than the threshold there are typically many simulated values less than the threshold leading to high probabilities of Type I error. The distributions of Type I and Type II error for a threshold of 4.0 are shown in Figure 4-12. This figure is the reverse of Figure 4-10. For Type I error the probabilities are mostly zero with a few high probabilities while for Type II error there are many low probabilities.

The proposed methodology further validates these results. The methodology is applied using the same parameters as the implementation example in Section 3.2 with classification thresholds of 2.0, 3.0, and 4.0. The relationship between Type I error and data spacing for the three thresholds is demonstrated in Figure 4-13. The expected error probabilities are highest for a cutoff of 3.0 reflecting the fact that the probability of misclassification is highest for a threshold at the center of the global distribution. The expected probabilities of Type I error are approximately equal for thresholds of 2.0 and 4.0 while the tails of these distributions are very different. For a threshold of 2.0, there are many locations that have probability of Type I error, but these probabilities are low. A threshold of 4.0, on the other hand, has few locations with probability of Type I error, but the probability is high.

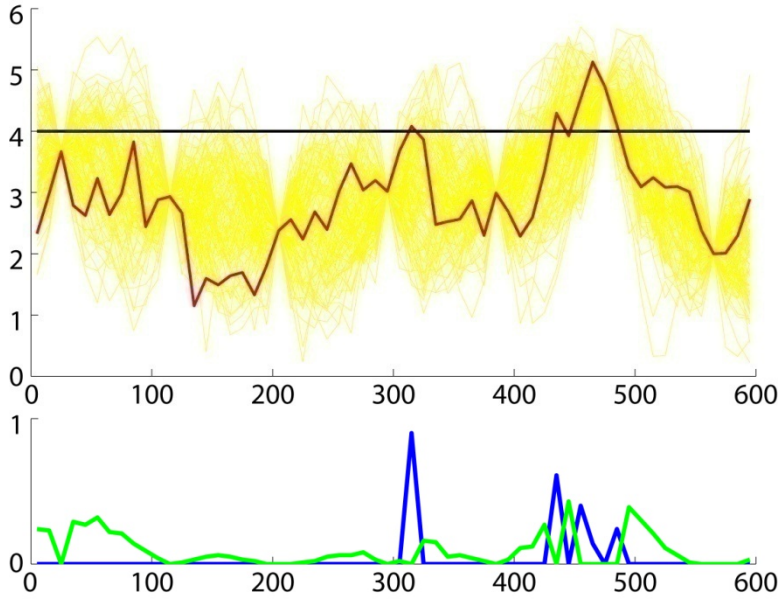

Figure 4-11: Truth (red), simulated values (yellow), and a threshold (black) of 4.0 leading to the probabilities of Type I (blue) and Type II (green) error.

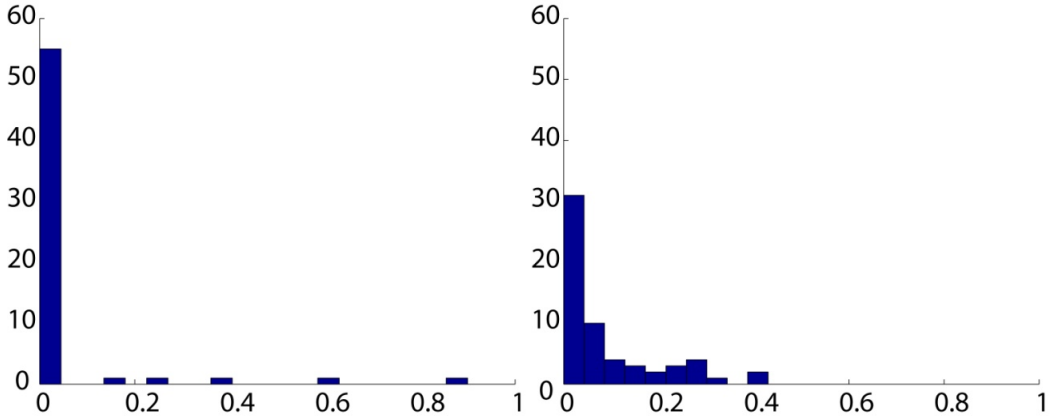

Figure 4-12: Distributions of Type I error (left) and Type II error (right) for a threshold of 4.0 relating to the blue and green lines in Figure 4-11 respectively.

Figure 4-14 demonstrates the relationship between Type II error and data spacing for the three thresholds. As with Type I error, the expected error probabilities are highest for a cutoff of 3.0 and the expected probabilities are approximately equal for thresholds of 2.0 and 4.0. However, the behavior of the tails has been reversed. There are few locations with probability of Type II error for a cutoff of 2.0, but these probabilities are high while the opposite holds true for a cutoff of 4.0.

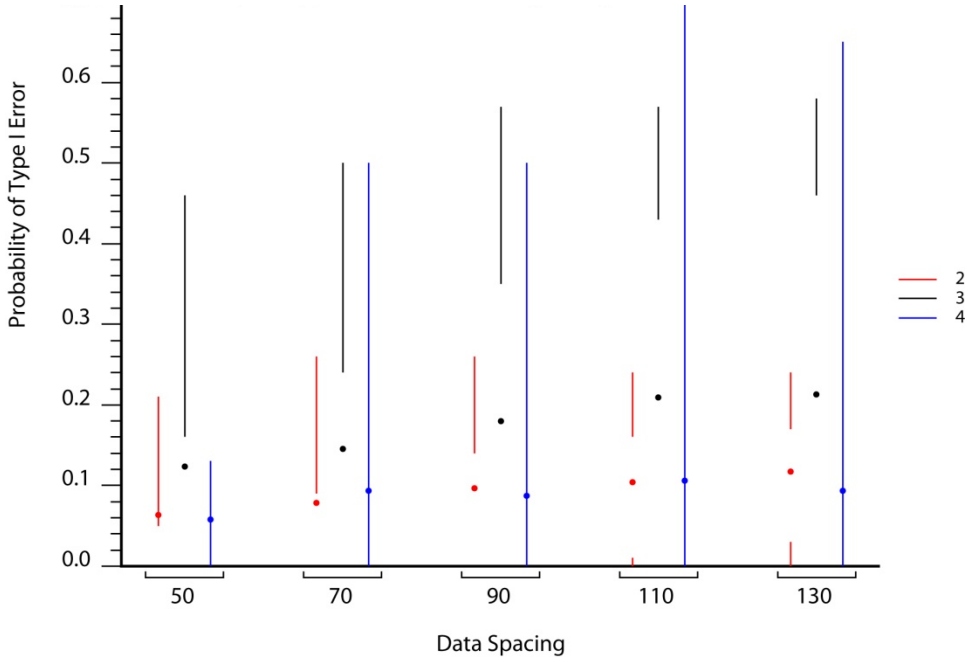

Figure 4-13: Probability of Type I error versus data spacing for thresholds of 2.0, 3.0, and 4.0.

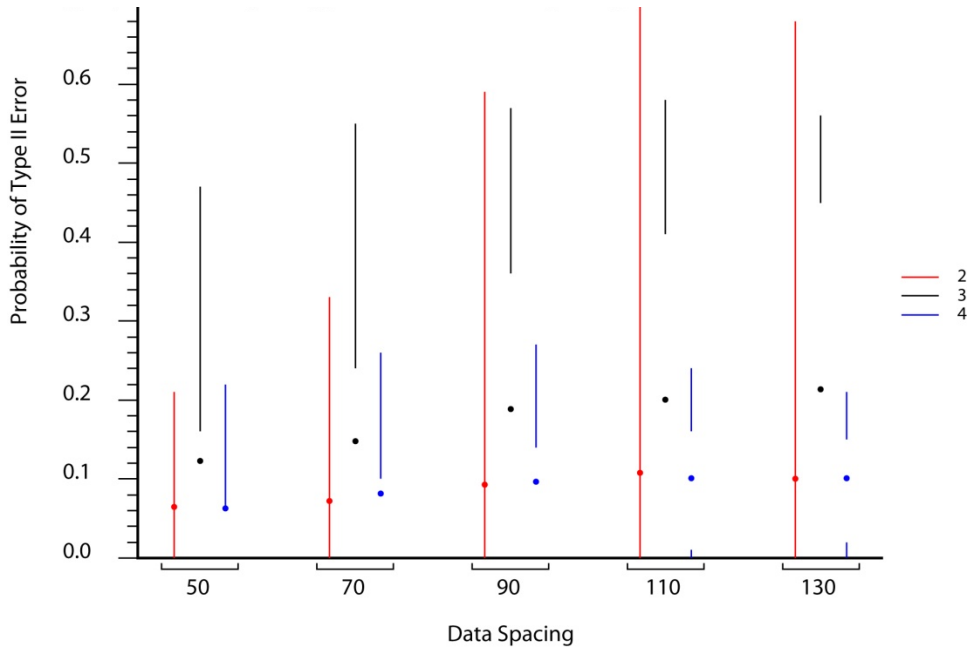

Figure 4-14: Probability of Type II error versus data spacing for thresholds of 2.0, 3.0, and 4.0.

4.5 Modeling Scale

Modeling scale is an important case-specific issue in the modeling of geological attributes. It is not practical to base decision on attributes modeled at the scale of the data. Some intermediate modeling scale must be chosen by the geostatistical practitioner. The choice of scale must consider the goals of the modeling as well as strike a balance between computational efficiency and sufficient detail (Deutsch, 2002).

The choice of modeling scale influences uncertainty. Classic dispersion variance theory (covered in Journel and Huijbregts, 1978; Isaaks and Srivastava, 1989; Deutsch, 2002; and Wackernagel, 2003) defines the relationship between scale and variability. Large blocks will show less variability than small blocks because the high and low values will be averaged out within the block.

To investigate the effect of modeling scale on uncertainty, the same parameters used in the implementation example in Section 3.2 are used, changing the size of the blocks that the point-scale estimates are averaged into. Recall that the point-scale spacing is 10m and that all previous work has considered averaging the simulated values to blocks of size 20x20m. The effect of scale is demonstrated by considering the point-scale values at 10m spacing directly as well as averaging the values to blocks of size 20x20m and 40x40m. The results are shown in Figure 4-15. Scale clearly has an impact and the well known results are verified here. The smallest scale shows the largest uncertainty and uncertainty decreases with scale. Increasing the scale from 10x10m to 20x20m decreases the expected standard deviation by approximately 15% and increasing the scale from 20x20m to 40x40m reduces the expected standard deviation again by approximately 15%. High and low values are being averaged out resulting in reduced uncertainty for large blocks.

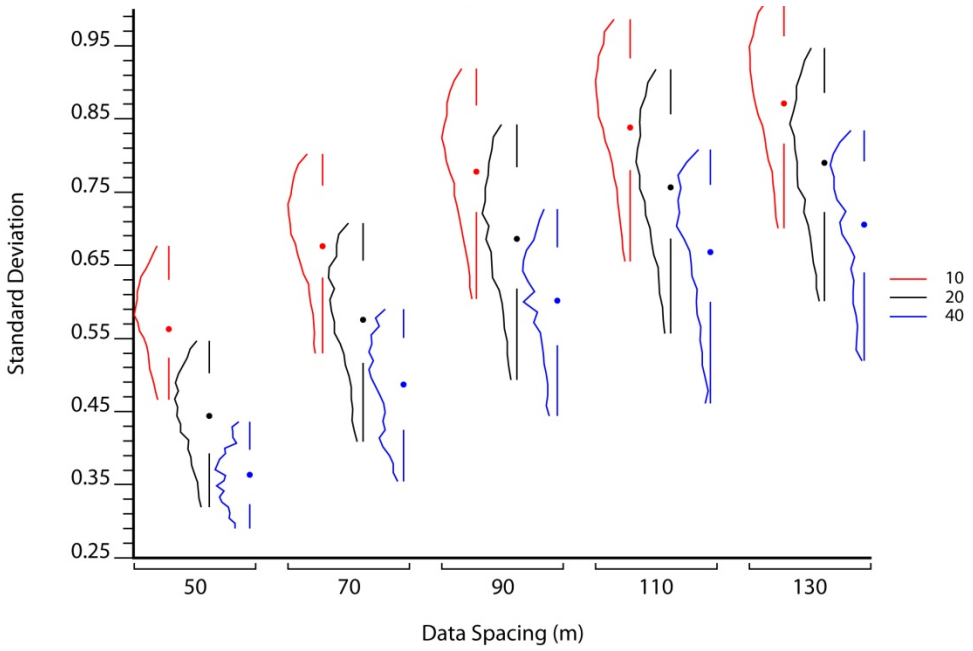

Figure 4-15: Uncertainty versus data spacing results for block sizes of 10x10m, 20x20m, and 40x40m.

Changes in the spread of the uncertainty distributions are also of note. For data spacings greater than or equal to the variogram range the variability in uncertainty is approximately equal between scales. The variability in uncertainty for 40m blocks is about the same as that for 20m and 10m blocks. This is not the case for data spacings less than the variogram range. The variability in uncertainty is reduced for large blocks. The variability in uncertainty for 40m blocks is less than the variability in uncertainty for 20m and 10m blocks.

4.6 Number of Realizations

The proposed methodology requires the generation of L truth realizations. Each truth realization is sampled at the desired spacing and the samples are used to generate K conditional realizations for a total of $L \cdot K$ realizations. The choice for K and L affects uncertainty. The analysis

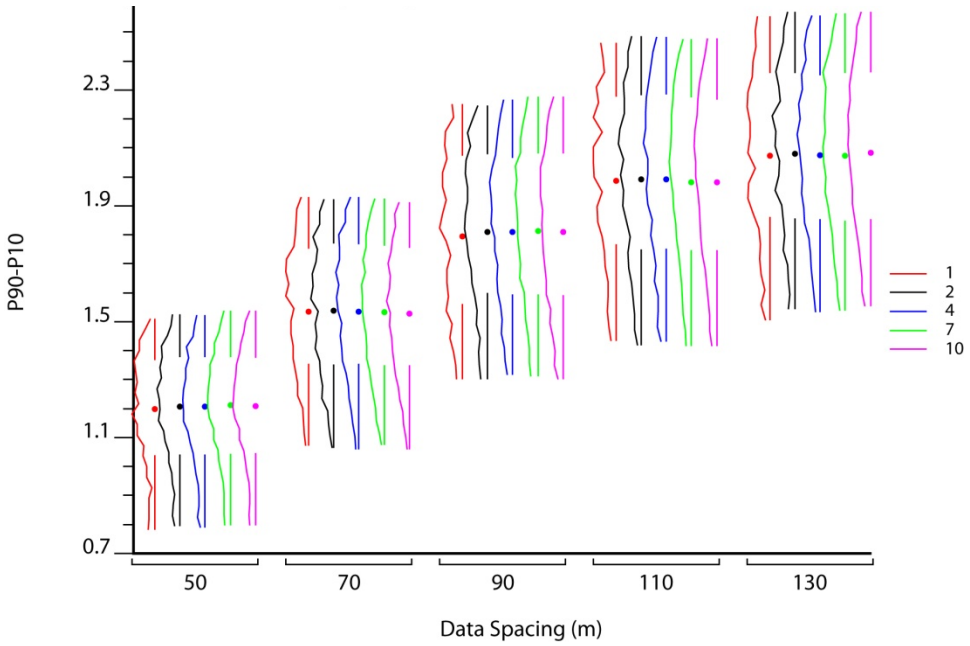

Figure 4-16: P90-P10 versus data spacing for L values of 1, 2, 4, 7, and 10.

presented here deals with the effect on uncertainty of using different values for K and L.

The same parameters used in the implementation example in Section 3.2 are used here, varying K and L from their base case values of 100 and 10 respectively. The effect of varying L is investigated first. The methodology is implemented using values for L of 1, 2, 4, 7, and 10 with K held constant at 50. The results are shown in Figure 4-16. For each data spacing the expected uncertainty as well as the spread of the uncertainty is constant over all values of L.

The effect of varying K is examined by considering values of 10, 20, 50, 70, and 100 holding L constant at 4. The results are shown in Figure 4-17. For each data spacing, the expected uncertainty is relatively constant while the spread in the uncertainty decreases with increasing K until K equals approximately 70 where the spread in uncertainty stabilizes. This is consistent with the discussion in Deutsch (2002) where he shows that

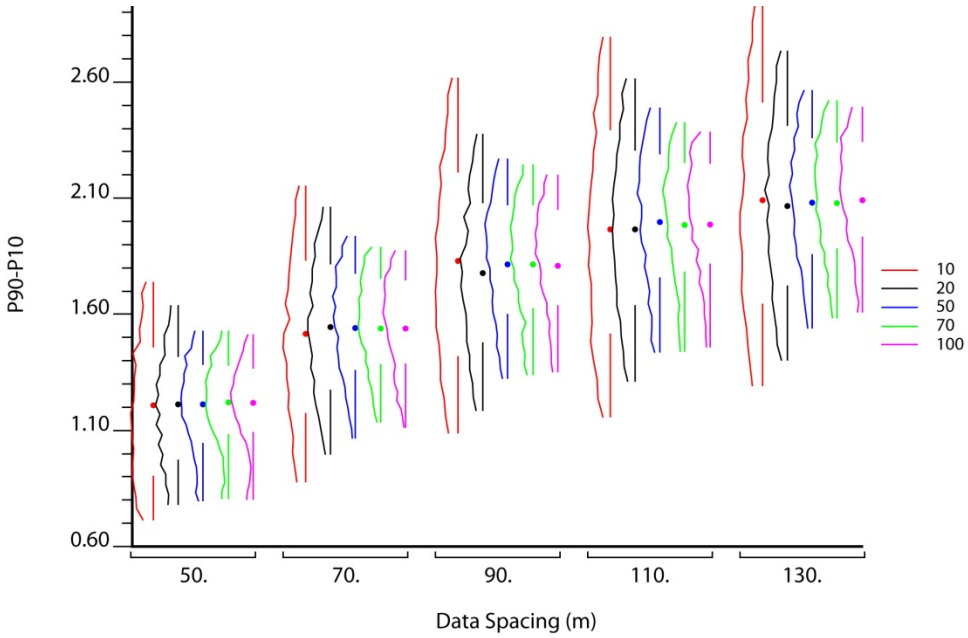

Figure 4-17: P90-P10 versus data spacing for K values of 10, 20, 50, 70, and 100.

precision is proportional to the number of realizations. The proposed methodology allows the sensitivity to the number of realizations to be assessed. The number of realizations should be chosen sufficiently high such that the spread in uncertainty is stable.

4.7 Data Quality

Data is acquired by sampling. Sampling is the process of measuring some geologic attribute using a representative portion of a larger mass. A number of errors can be introduced during the various stages required for sampling. These errors have been classified as fundamental, delimitation, extraction, and accidental (Pitard, 1993). Fundamental error cannot be removed by modifying the sampling practice. These errors are random with a mean of zero. Fundamental errors arise due to differences between the compositions of fragments within the lot, or constitution heterogeneity, which is a function of the material being sampled. The constitution

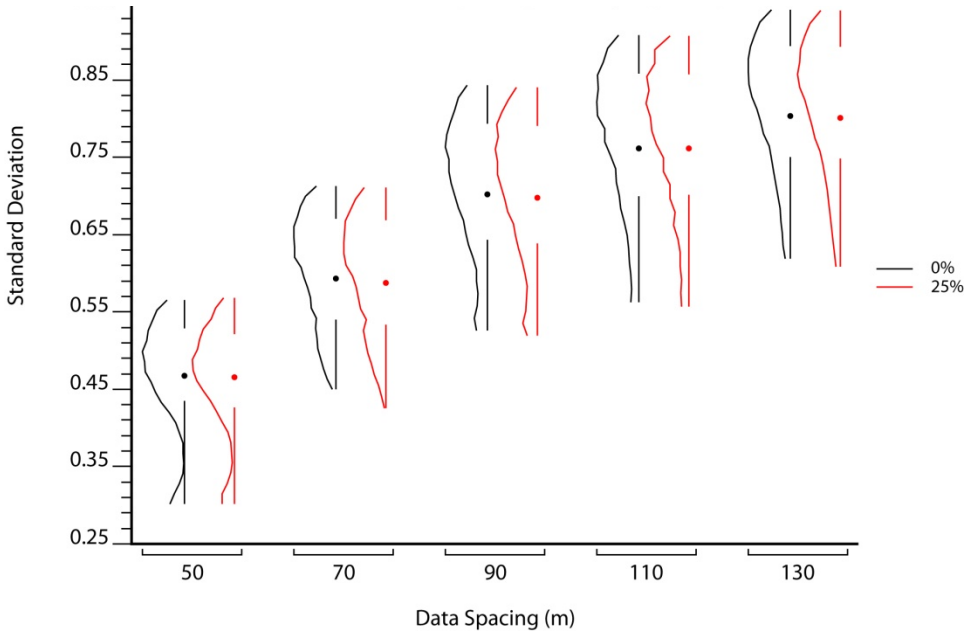

Figure 4-18: Standard deviation versus data density for varying sampling error.

heterogeneity of a material can be quantified allowing the quantification of fundamental error (Pitard, 1993). Delimitation and extraction errors typically have non-zero mean introducing a bias to the sampling program. Accidental errors cannot be analyzed statistically since they are usually non-random events (Neufeld, 2003).

This work considers only fundamental error as it introduces no bias. As mentioned, unbiased sampling error is predominantly controlled by the nature of the material being sampled. A heterogeneous material will have greater sampling error than a homogeneous material.

The realizations of the truth are sampled at the desired spacing and those samples are used to generate conditional realizations from which uncertainty is determined. When the truth is sampled, random error could be added to each sample. The magnitude of these random errors is controlled by specifying the variance of the error distribution.

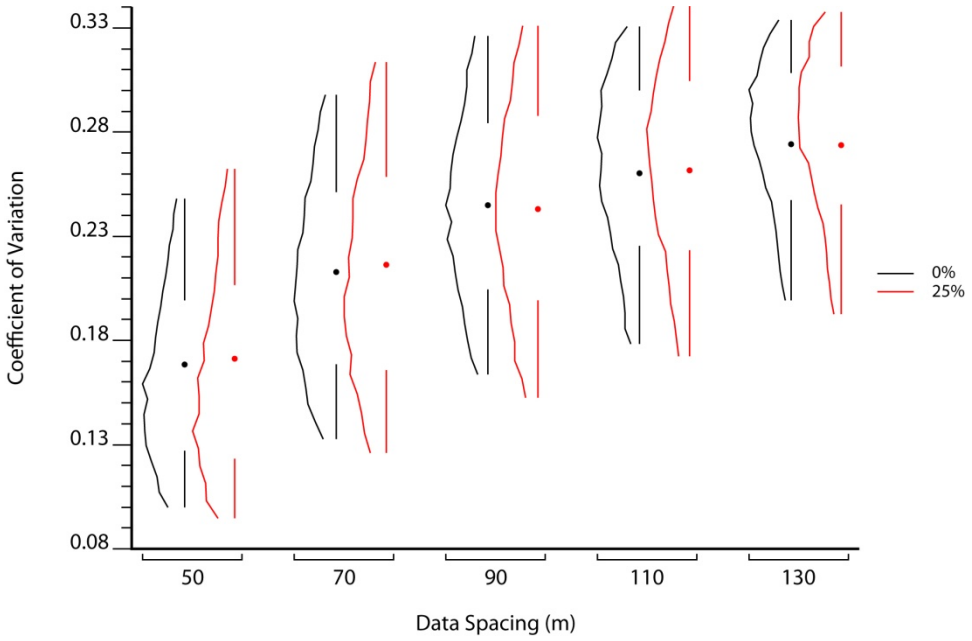

Figure 4-19: Coefficient of variation versus data spacing for varying sampling error.

The effect of data error on uncertainty can be isolated and analyzed. This is done using the same parameters as the implementation example in Section 3.2, but varying the variance of the sampling errors from no sampling error to 25% sampling error.

The degree of sampling error affects different measures of uncertainty in different ways. The non-standardized measures of spread (standard deviation and P90-P10) are largely unaffected by changes in data quality, see Figure 4-18. The errors are unbiased and cancel out quickly with averaging to a larger scale.

The standardized measures (coefficient of variation and (P90-P10)/P50), on the other hand, are affected by changes in data quality as demonstrated by the coefficient of variation versus data spacing plot in Figure 4-19. The expected uncertainty is approximately the same for no sampling error as for 25% sampling error. The difference is in the spread of the distribution of uncertainty values for a given data spacing; the spread is greater when there is sampling error. This is due to fluctuations in the center of the local distribution caused by sampling error. Consider

a location close to, but not at, a data location. This location will have the similar spread in the simulated values whether the data it is near has sampling error or not, but the mean (or median) of the simulated values will be different. If the nearby sample has a value lower than the truth, then the mean of the simulated values is reduced increasing the value of the standardized measures of spread. If the nearby sample has a value higher than the truth, then the mean of the simulated values is increased reducing the standardized measures. Sampling error therefore increases the spread of the standardized uncertainty measures.

Precision can also be affected by sampling error. Recall that precision is determined by considering some distance, h, from the mean and that this distance can be proportional to the mean. When this is the case, precision is affected by sampling error. This effect is similar to that seen for the coefficient of variation and (P90-P10)/P50; sampling error leads to an increase in the spread of the distribution of precision values for a given data spacing. This is caused by variations in the local mean. A reduction in the local mean leads to a reduction in h which, for constant spread, leads to a reduction in precision. An increase in the local mean leads to an increase in h which, for constant spread, leads to an increase in precision. The presence of these higher and lower precision values increases the spread of the precision distribution.

The probability of misclassification is relatively unaffected by sampling error as shown in Figure 4-20 and Figure 4-21 where the expected misclassification probabilities show little change when sampling error is introduced. There is a slight increase in the P90 values of the distributions due to sampling error, but overall these results show that an increase in sampling error has little effect. The spread in the uncertainty distributions shown in these figures is very large. There are a number of

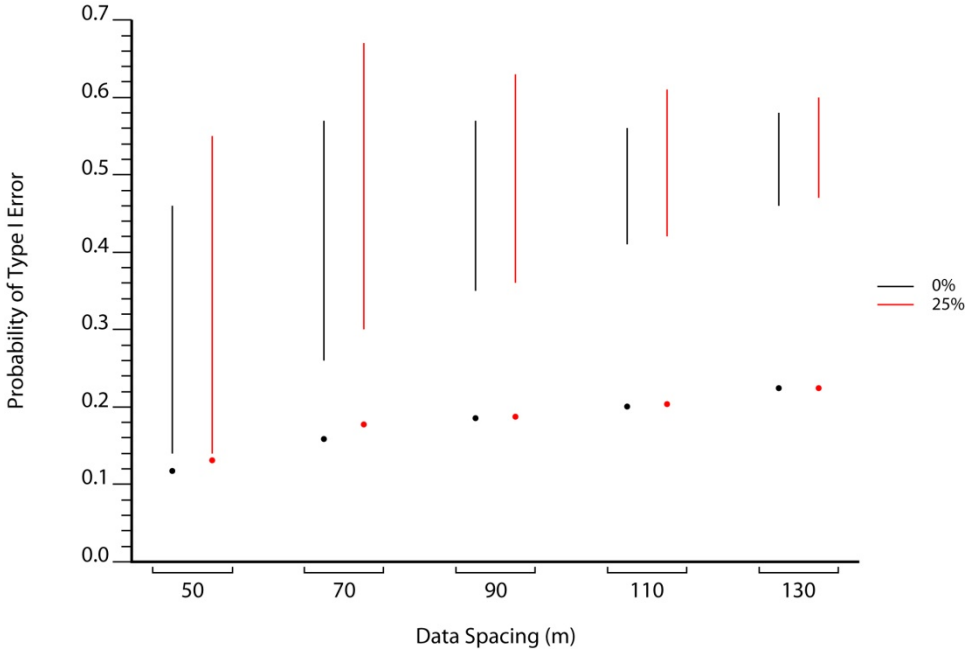

Figure 4-20: Probability of Type I error versus data spacing for varying sampling error.

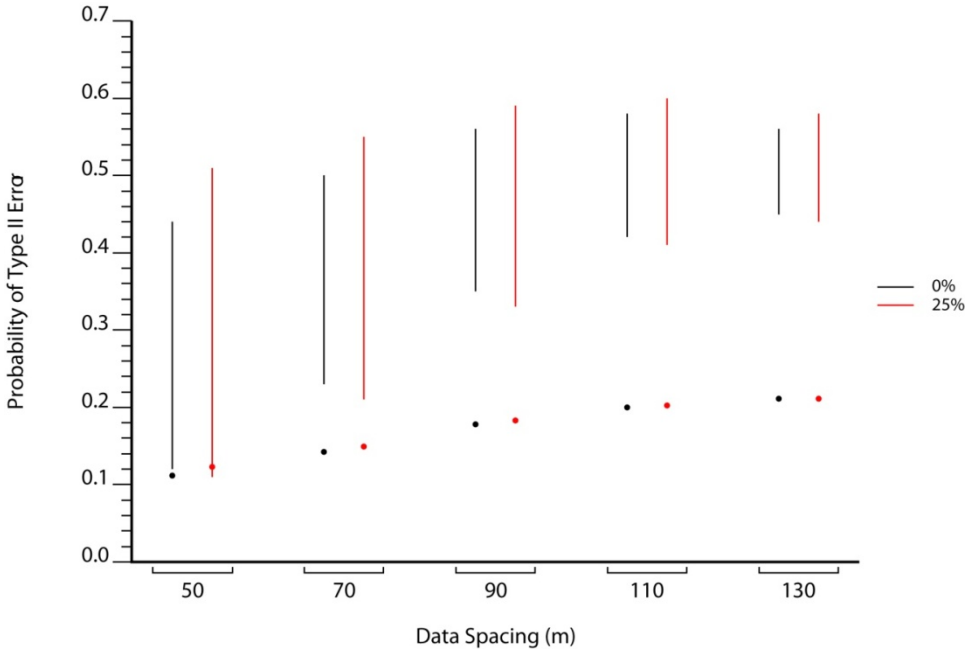

Figure 4-21: Probability of Type II error versus data spacing for varying sampling error.

locations where the majority of the local grade distribution falls on the opposite side of the threshold from the truth. These locations lead to the instances of probability of misclassification greater than 0.5. There are more of these locations when sampling error is present leading to an increase in the P90 uncertainty values.

Locally, sampling error could either increase or decrease the probability of misclassification. The areas where the probability of misclassification is increased will average with those areas where probability is decreased keeping the expected value the same. Instances of these local effects are illustrated in Figure 4-22. The red line in the plots in this figure represents the truth, the black dots represent a sample, and the yellow lines represent the simulated values conditioned to the samples. The blue and green lines represent the probabilities of Type I and Type II error respectively.

In Figure 4-22a, the sample with error falls below the threshold while the truth is greater than the threshold increasing the probability of Type I error locally. Figure 4-22b shows a location where the sample with error falls further from the threshold than the truth reducing the local probability of Type I error. In Figure 4-22c and Figure 4-22d the sample with error falls opposite the threshold than the truth leading to an increase and a decrease in the local probability of Type II error respectively.

4.8 Summary

This chapter has mentioned some of the 'known unknowns' (Maluf *et al.*, 2005), things that are known to be unknown. It is known that the decisions of stationarity and input parameters as well as various geostatistical assumptions can affect uncertainty while the exact nature of their effect is unknown. This chapter has also considered the 'known

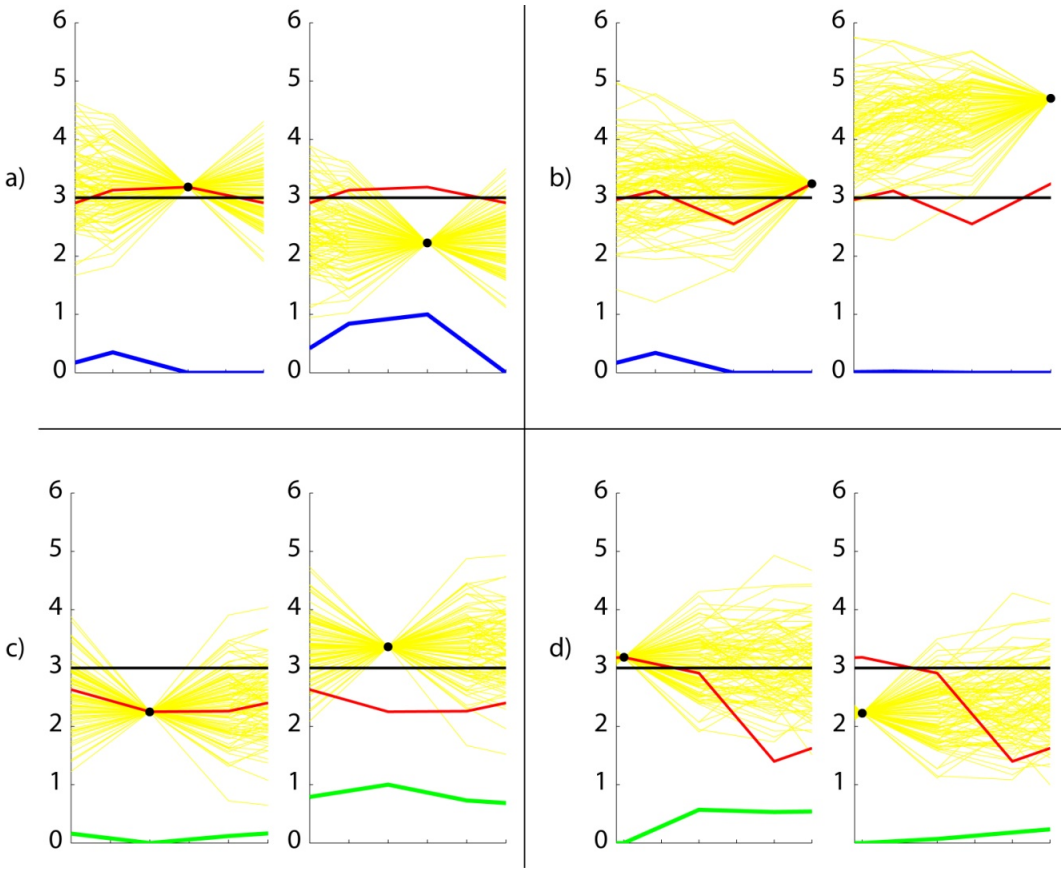

Figure 4-22: Instances where sample error has a) increased local Type I probability, b) decreased local Type I probability, c) increased local Type II probability, and d) decreased local Type II probability.

knowns'. Factors such as the proportional effect, nonstationarity, and classification thresholds are known to affect uncertainty and the nature of these effects can be quantified, or made known. It must also be acknowledged that there are 'unknown unknowns' when it comes to uncertainty, things that are not known to be unknown. After all possible causes for the relationship between data spacing and uncertainty have been examined, it may have to be conceded that there are aspects of uncertainty that elude even the most detailed study.

Chapter 5

Case Study

The proposed methodology is implemented using oil sands data from the McMurray formation in northern Alberta (Warren, 2003). The data is bitumen thickness data within an area 112 x 171 km in size (Figure 5-1). Within this area data density is highly variable ranging from very low ($<$ 1 well per section) to almost 20 wells per section in select areas. There are 2514 data with an equal-weighted average thickness of 20.8m; accounting for data clustering yields an average thickness of 16.1m (Figure 5-2). Bitumen thickness is laterally continuous; the horizontal omnidirectional variogram of the normal scores of the thickness data is shown in Figure 5-3. The variogram model is isotropic with three structures summarized in Table 5-1.

Table 5-1: Variogram model parameters for bitumen thickness normal scores.

Structure	Type	Contribution	Range
1	Exponential	.5	700
2	Spherical	.25	5000
3	Spherical	.25	15000

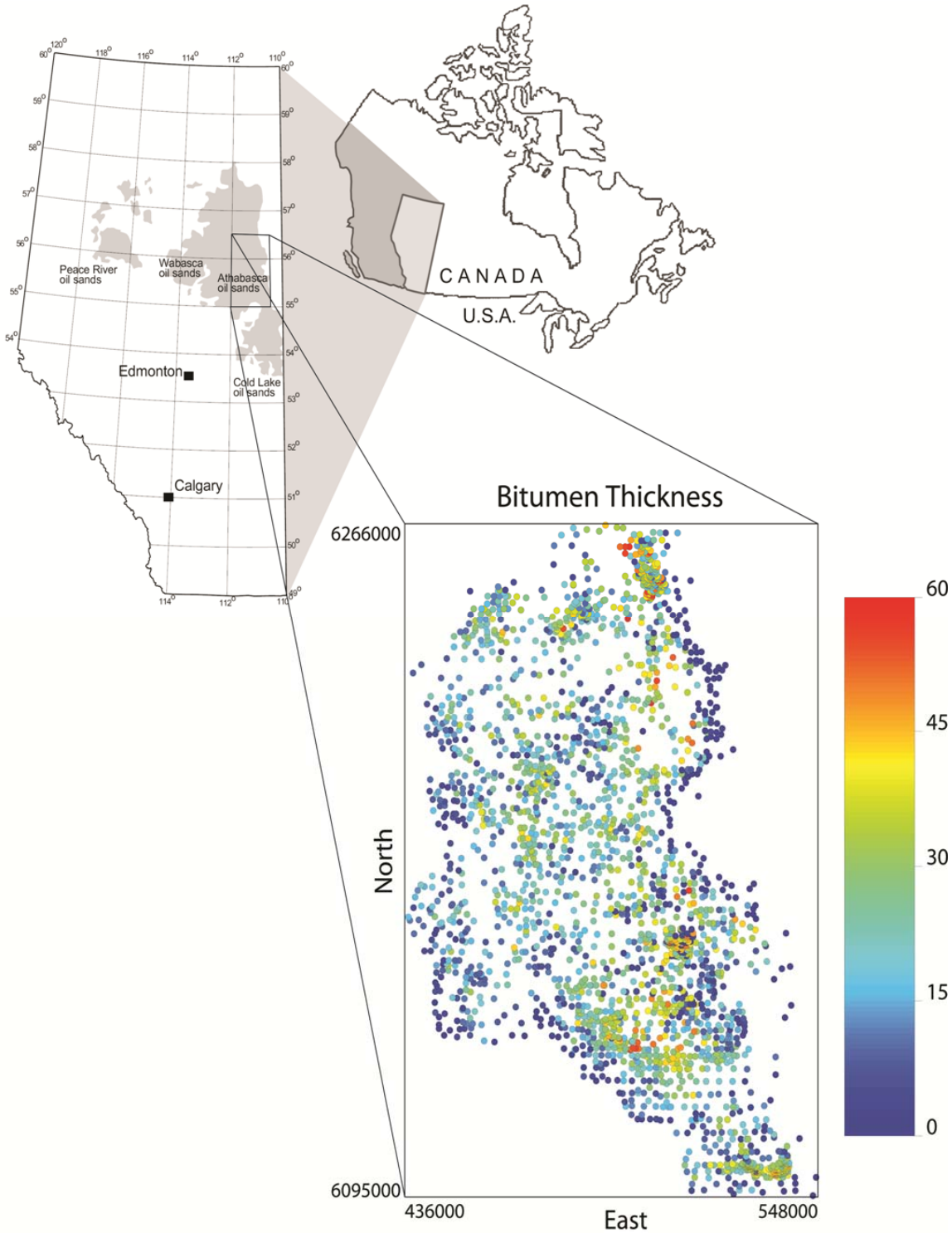

Figure 5-1: Location of bitumen thickness data. (adapted from Alberta, 2000)

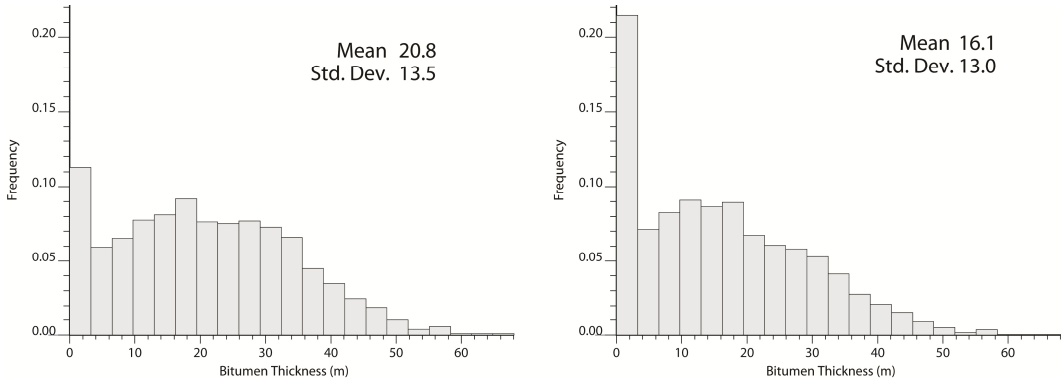

Figure 5-2: Bitumen thickness distributions: left - equal weighted; right - declustered.

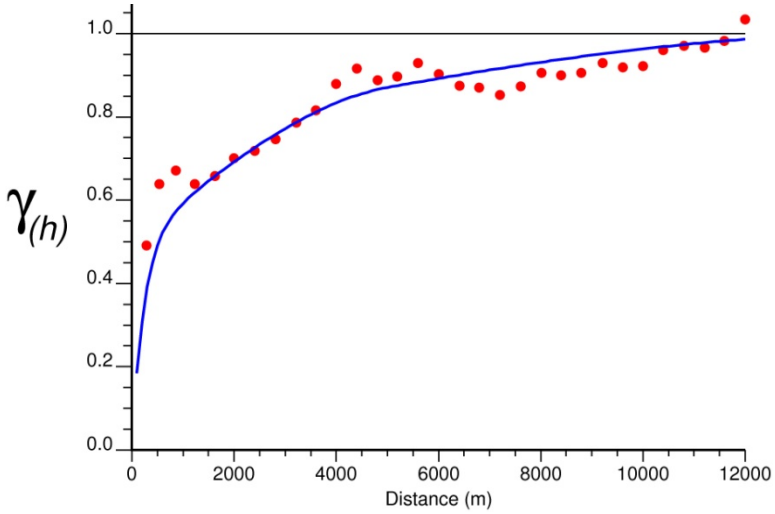

Figure 5-3: Variogram of the normal scores of the bitumen thickness.

The relationship between uncertainty and data spacing/density is evaluated in two different ways. The first method applies the methodology proposed herein. Truth realizations are generated conditional to the bitumen thickness data that are then sampled at spacings from 400m to 4000m. This range is much larger than would normally be considered in practice, but suffices for illustrative purposes. This range also does not consider spacings less than 400m that would also be considered in practice. The samples are used to generate additional realizations from which measures of uncertainty are determined. This

allows the establishment of the relationship between data spacing and uncertainty for this histogram and variogram.

For the second method, measures of uncertainty are determined from simulated realizations generated conditional to the bitumen thickness data. Data spacing is determined on a regular grid using the constant n method described in Chapter 2. The measures are then compared to their corresponding data spacing to arrive at the relationship between data spacing and uncertainty. For both cases, data-scale values are simulated at a spacing of 100m which are then block averaged to 400m square blocks.

5.1 Method One

Reference realizations are generated conditional to the pre-existing thickness data. Values are simulated every 100m. This realization is then sampled at the desired spacing. A 1% random sampling error is added to each sample and these samples are used to condition 100 realizations of thickness. The point-scale values in these realizations are averaged into blocks 400m square. There are about 120,000 400m blocks within the area of interest. Uncertainty measures are calculated from the 100 realizations at each block location. Data spacings from 400m to 4000m are evaluated.

Results for standard deviation are shown in Figure 5-4. Uncertainty is highest for the largest spacing. The distribution shapes are primarily negatively skewed. There is a hint of a bimodal distribution at a spacing of 800m confirming the earlier observation that the distributions of non-standardized uncertainty measures tend to be bimodal for spacings approximately twice the block size. Uncertainty is reduced by reducing the distance between data. The magnitude of this reduction is controlled by the variogram. For the thickness variable, halving the data spacing from 1600m to 800m decreases the expected standard deviation

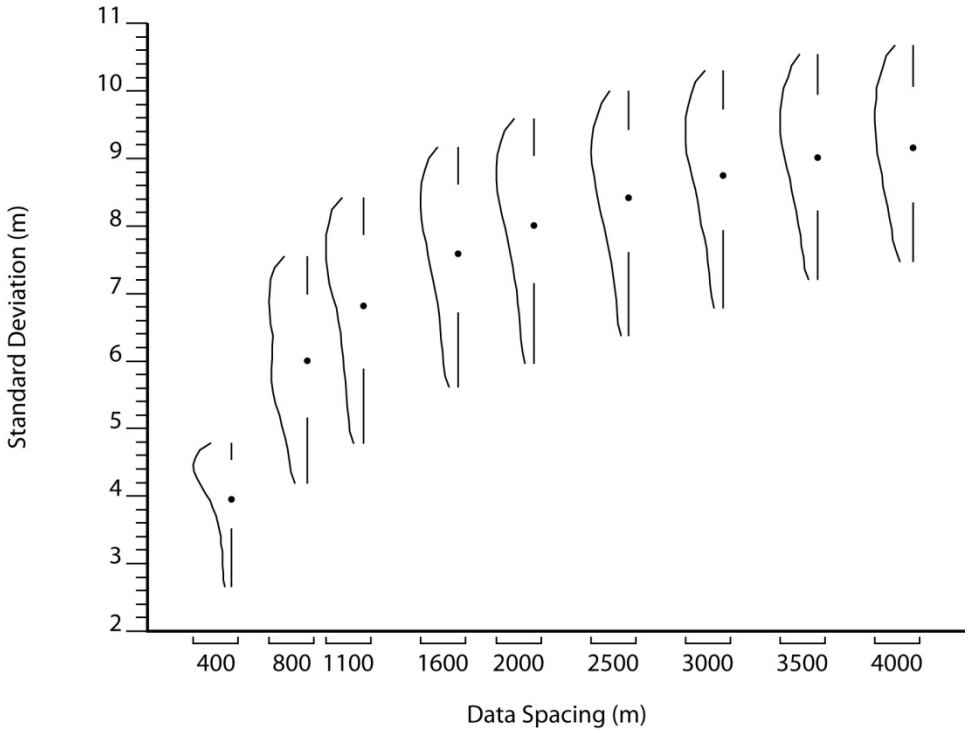

Figure 5-4: Standard deviation versus data spacing for the oil sand example for data spacings from 400m to 8000m.

from approximately 7.6m to 6.0m. Halving the spacing again from 800m to 400m decreases the expected standard deviation from 6.0m to slightly less than 4.0m. The variance of the nine expected standard deviation values is approximately 2.6 while the expected variance of the nine distributions of standard deviation is approximately 1.7. This means that data spacing is responsible for about 60% of the uncertainty captured by the standard deviation while the other 40% is due to other factors.

Figure 5-5 demonstrates the behavior of the difference between percentiles measure. Its behavior is similar to the standard deviation. The expected difference is lowest for a spacing of 400m at approximately 10m and increases to 24m at a spacing of 4000m. The distributions are predominantly negatively skewed there are more locations far from data than close to data when samples are on a regular grid. The variance of

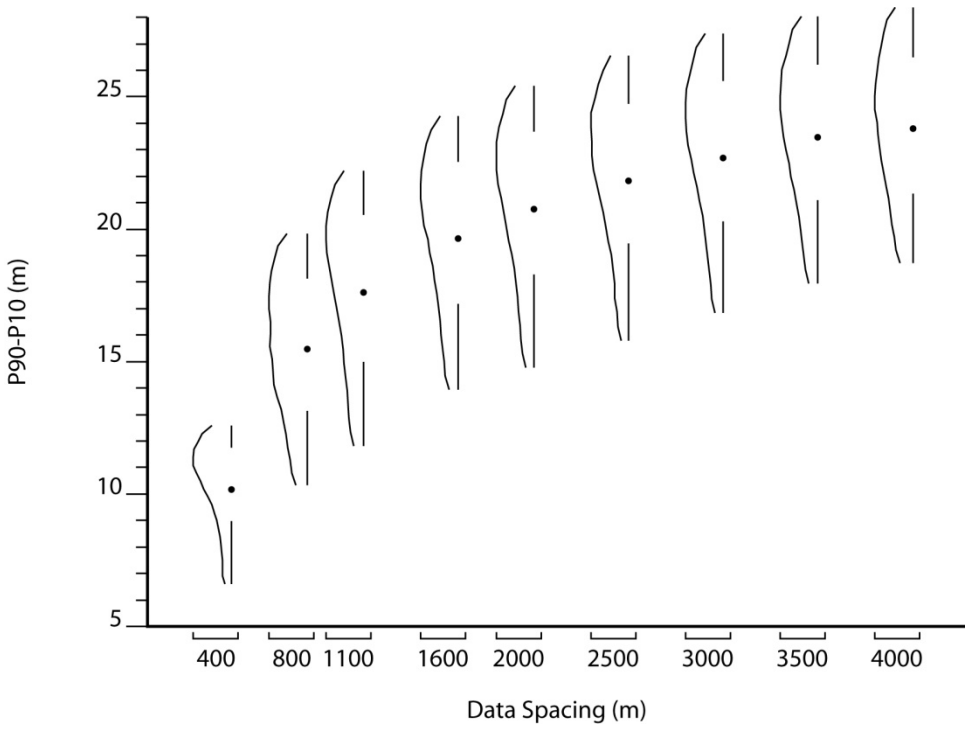

Figure 5-5: Difference between percentiles versus data spacing for spacings from 400m to 4000m.

the nine expected difference between percentiles values is approximately 17 while the expected variance of the nine distributions of difference between quantiles is approximately 14. This means that data spacing is responsible for about 55% of the uncertainty captured by the difference between percentiles while the other 45% is due to other factors.

The behavior of the coefficient of variation is shown in Figure 5-6. Its expected value increases with increasing data spacing similar to the measures previously examined. This increase is steep for small data spacings and flattens off at spacings greater than 700m. This reflects the variogram model used which has a range of 700m for the first structure. The distributions are positively skewed as there are few instances where the standard deviation is high for a low-valued mean. The variability between distributions is much lower than the variability within the distributions relative to the measures observed earlier. The variance of

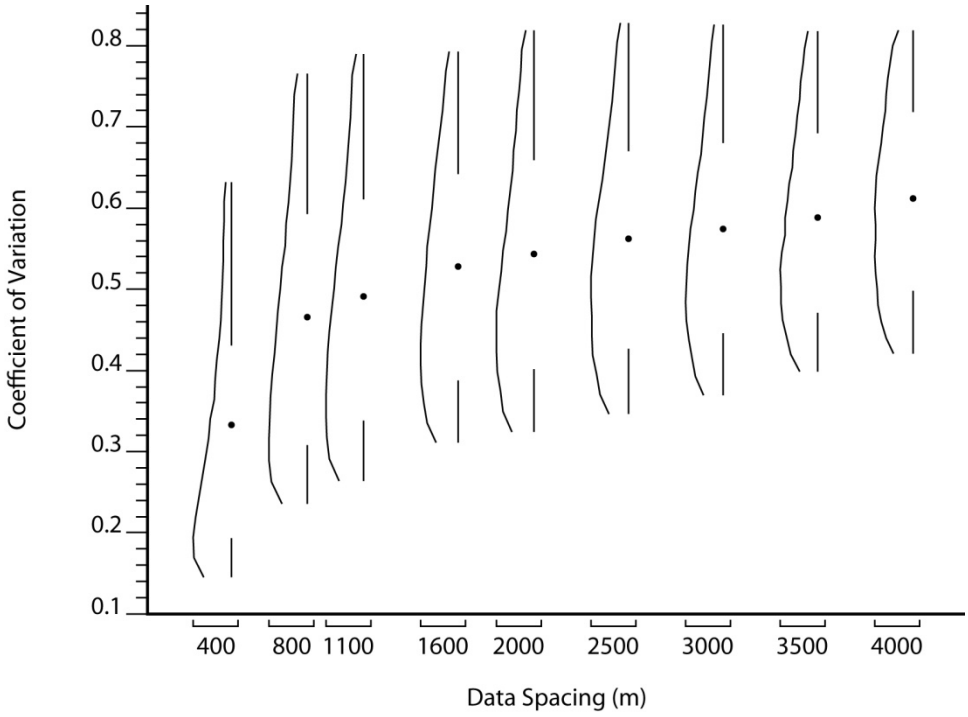

Figure 5-6: Coefficient of variation versus data spacing for spacings from 400m to 4000m.

the nine expected coefficient of variation values is approximately 0.006 while the expected variance of the nine distributions of coefficient of variation is approximately 0.03. This means that data spacing is responsible for only about 17% of the uncertainty captured by the coefficient of variation while the other 83% is due to other factors.

The behavior of the standardized difference between percentiles, shown in Figure 5-7, is similar to the coefficient of variation. It increases with increasing data spacing in approximately the same manner, increasing more at small spacings and flattening off beyond a spacing of 800m. This reflects the influence of the variogram model used which has a range of 700m for the first structure. This measure is also positively skewed due to their being few instances of large spread for low median values. Again, the variability between distributions is much lower than the variability within distributions. The variance of the nine expected values is approximately 0.06 while the expected variance of the nine distributions is approximately

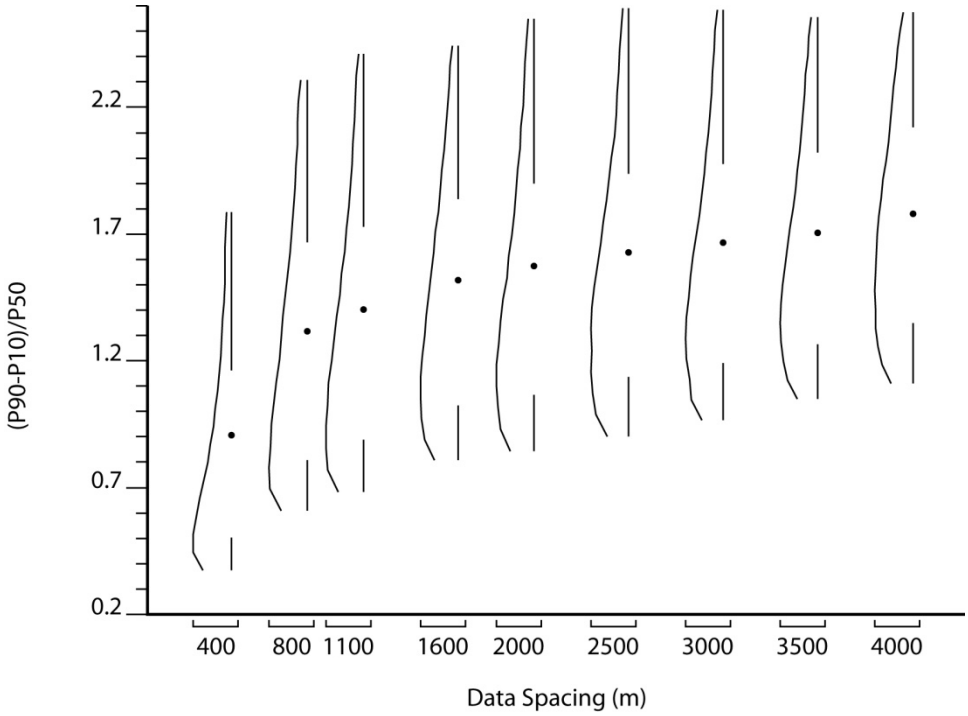

Figure 5-7: Standardized difference between percentiles versus data spacing for spacings from 400m to 4000m.

0.44. This means that data spacing is responsible for about 12% of the uncertainty captured by the standardized difference between percentiles while the other 88% is due to other factors.

Precision is a measure of the narrowness of a distribution and therefore decreases with increasing data spacing as shown in Figure 5-8. Precision for this study is defined as the proportion of a distribution that falls within 15% of the mean of that distribution. The expected precision for a spacing of 400m is approximately 0.68 and decreases to approximately 0.36 for a spacing of 4000m. For small spacings the distribution of precision values is negatively skewed with a large number of precision values near 1.0. As spacing increases, the distribution changes to being positively skewed reflecting the increase in the number of locations far from data. The variance of the nine expected precision values is approximately .008

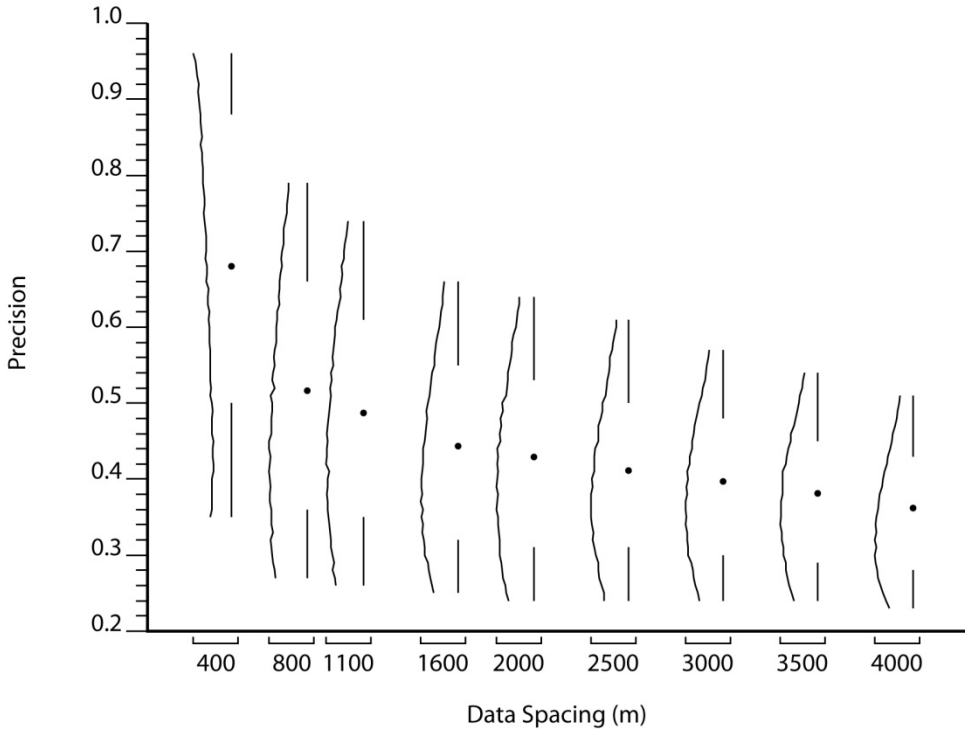

Figure 5-8: Precision versus data spacing for spacings from 400m to 4000m.

while the expected variance of the nine distributions of precision is approximately .03. This means that data spacing is responsible for about 25% of the uncertainty captured by precision while the other 75% is due to other factors.

The probabilities of the two types of misclassification error are shown in Figure 5-9 and Figure 5-10. The classification threshold for this study is 20m. When the truth is greater than or equal to 20m there is potential for Type I misclassification to occur and when the truth is less than 20m there is potential for Type II misclassification error to occur. For the nine data spacings considered many locations have 0% probability of being misclassified as is shown by both the 10^{th} and 25^{th} percentiles being zero. The expected probability of misclassification is the most useful summary here. As is shown in both plots, the expected probability of misclassification increases with increasing data spacing. The relative

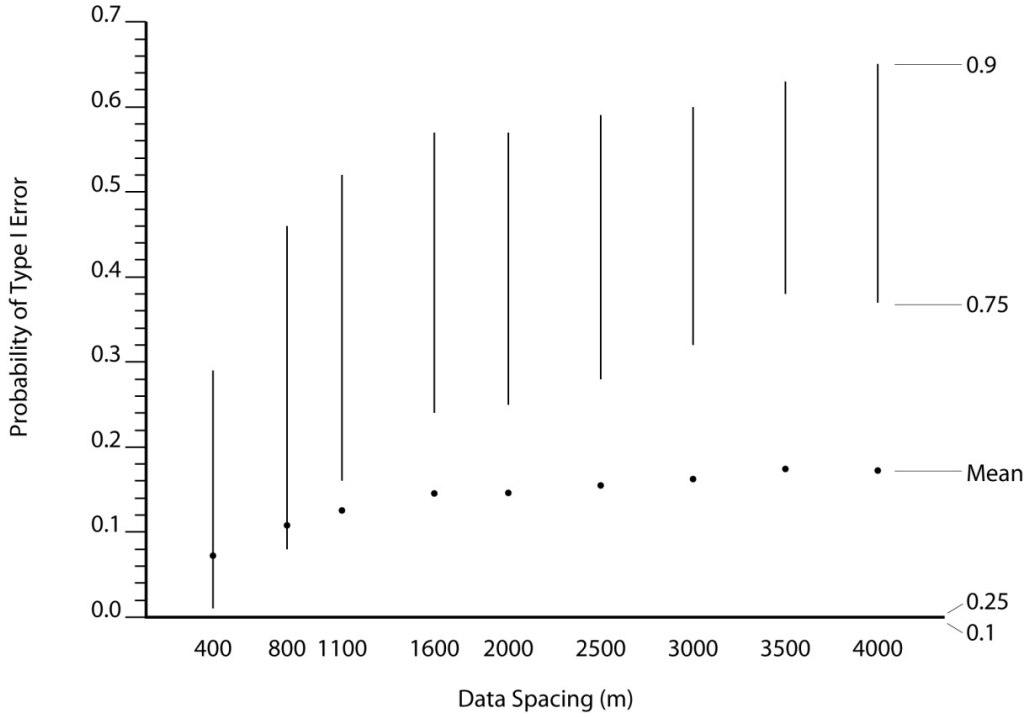

Figure 5-9: Probability of Type I error versus data spacing for spacings from 400m to 4000m.

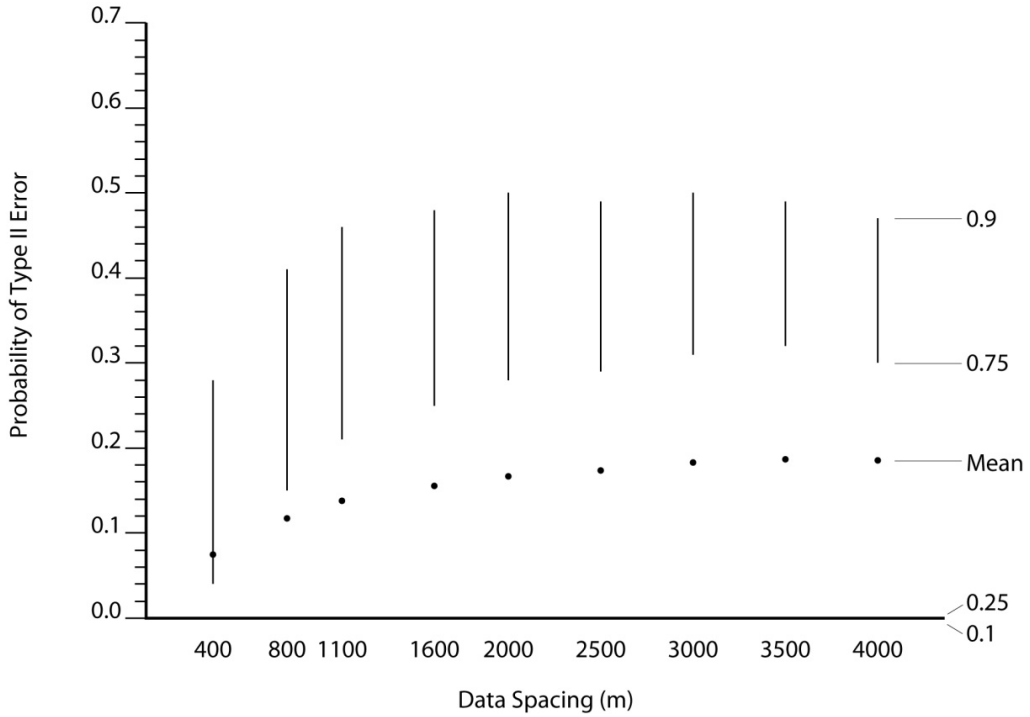

Figure 5-10: Probability of Type II error versus data spacing for spacings from 400m to 4000m.

probability of each type of misclassification error is dependent on the value of the threshold with respect to the reference distribution. The threshold of 20m is greater than both the mean and median of the reference distribution. This means there are more locations where the truth is less than 20m increasing the possibility of Type II misclassification errors. The possibility of Type I errors is reduced for this threshold, but when a Type I error is possible, it is more probable. This increased probability is communicated by the higher P90 values in Figure 5-9 than in Figure 5-10. The higher possibility of Type II error means that the expected probability of Type II error is greater than the expected probability of Type I error. For a data spacing of 4000m the expected probability of Type I error is approximately 0.17 while the expected probability of Type II error is approximately 0.19. The difference in probabilities is small due to the threshold being close to the center of the reference distribution. The variability among the nine expected values for these two measures is very low relative to the variability within the distributions. Data spacing accounts for only 2% of the variability while the remaining 98% is due to other factors.

5.2 Method Two

For the second method, the relationship between uncertainty and data spacing is determined by generating 100 realizations of thickness conditional to the thickness data. These realizations are block averaged and uncertainty measures are calculated from the block averaged values. This method requires a measure of data spacing at all locations. Data spacing is determined on a 400m grid by applying Equation 2.1 where n_V is 20. Once V has been determined it is a simple matter to calculate density and spacing. Maps of data density and data spacing are shown in Figure 5-11. The histograms associated with these maps are shown in Figure 5-12. Data density is overall very low with a few small areas being

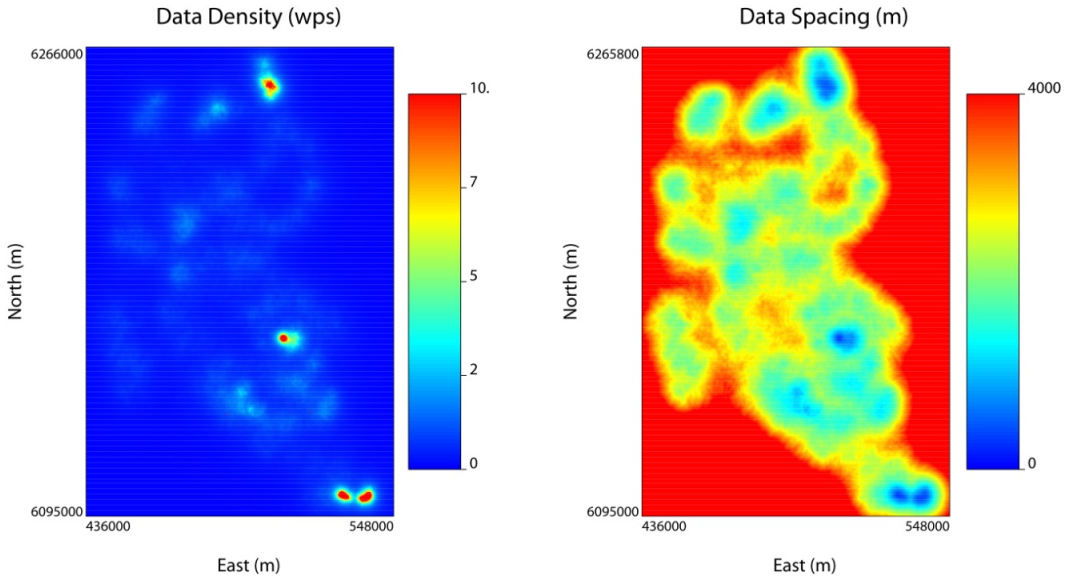

Figure 5-11: Data density and data spacing on 400m grid.

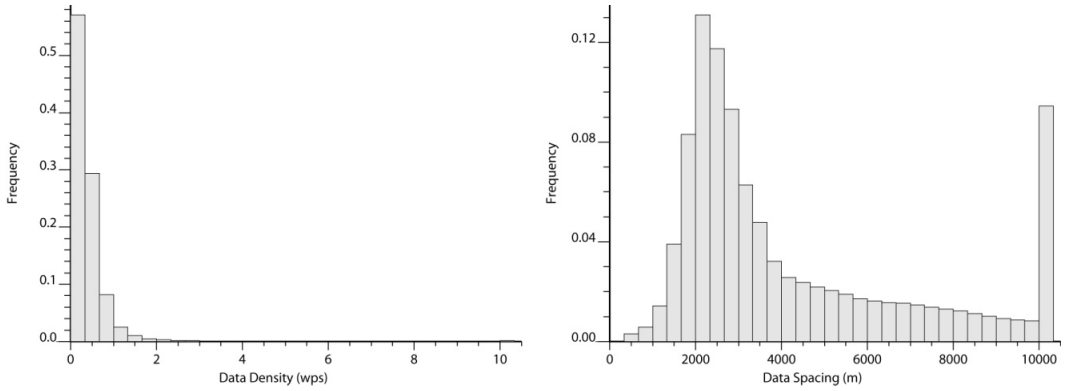

Figure 5-12: Histograms of data density and data spacing.

densely sampled. The majority (>60%) of the data spacing values are less than 4000m. The smallest spacing observed is just under 400m. Examination of the relationship between data spacing and uncertainty is restricted to data spacings in this range.

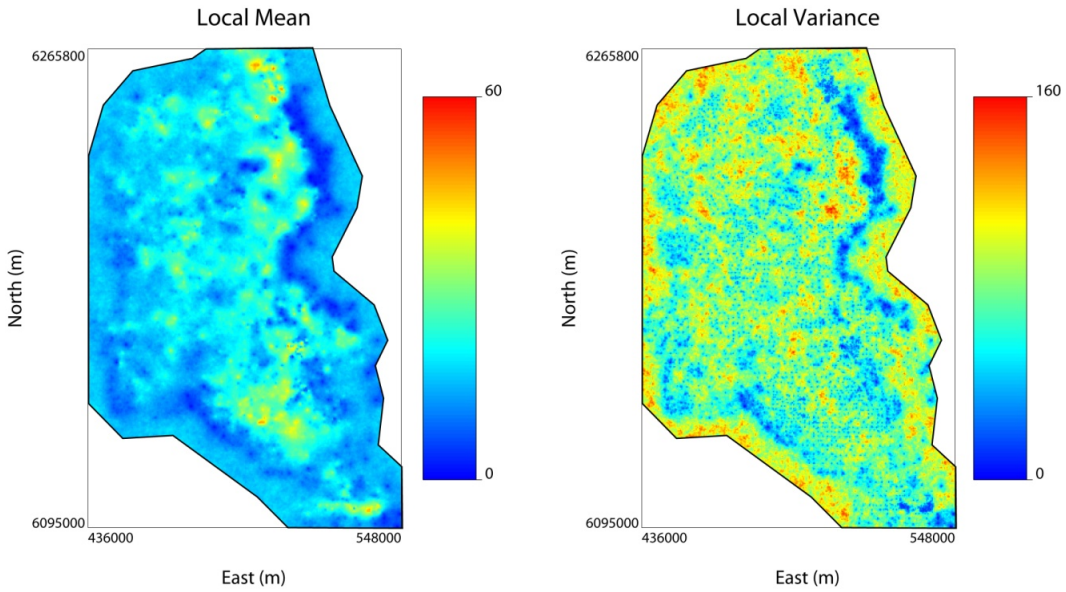

Figure 5-13: Local mean and variance of the 100 simulated realizations at 400m scale.

SGS is used to generate 100 conditional realizations of normal scored bitumen thickness. Each realization has 1,915,200 simulated values (1120 x 1710) spaced every 100m. The normal score variogram in Figure 5-3 is used to define the spatial continuity. The normal score values are back-transformed to units of bitumen thickness according to the declustered distribution shown in Figure 5-2 right. The back-transformed values are then block averaged to 400m square blocks. Figure 5-13 shows the local mean and variance of these 400m blocks.

Five of the seven previously utilized uncertainty measures are calculated for all 280 x 427 block locations. The two probability of misclassification measures cannot be calculated in the absence of a realization of the truth. Figure 5-14 shows the relationship between the non-standardized measures of spread and data spacing. The points are colored according to bitumen thickness and the line represents the mean uncertainty measure. The direct relationship between these measures and data spacing is evident. The measures increase rapidly with increasing data spacing at first before leveling off at a spacing of approximately 700m, mimicking the variogram

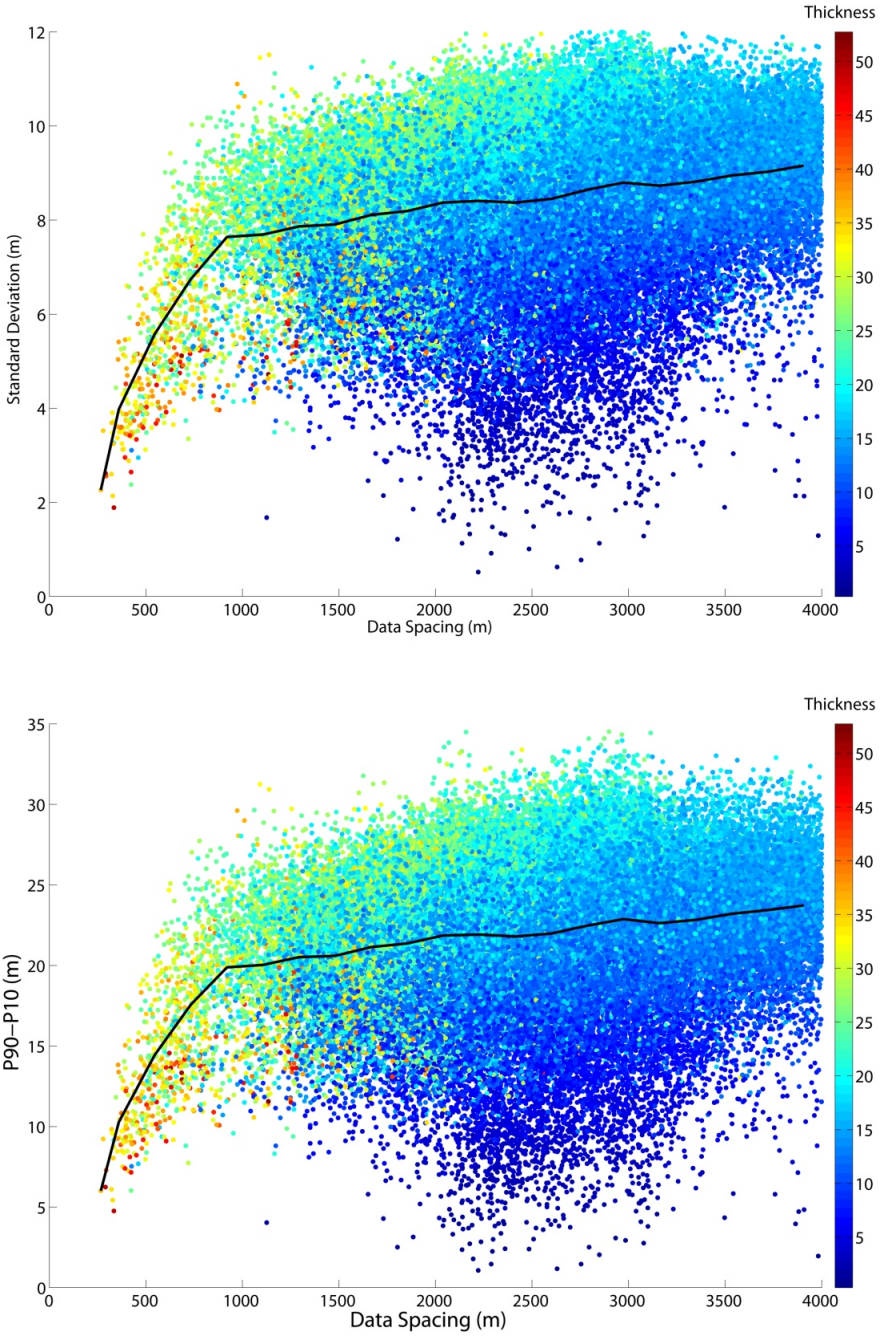

Figure 5-14: The relationship between the non-standardized measures of spread (standard deviation and P90-P10) and data spacing for spacings from 0 to 4000m.

model. The deposit has been densely sampled in areas where bitumen thickness is greatest as is evidenced by the low data spacings being dominated by high thickness values. The proportional effect has an

influence on the results as the reference distribution is positively skewed. This is evidenced by a large spread in uncertainty for most data spacings where the uncertainty is clearly proportional to the bitumen thickness.

The proportional effect has a markedly different impact on the standardized measures of spread shown in Figure 5-15. The relationship between uncertainty and thickness is reversed with the thickest values having the smallest uncertainty and the thinnest values having the largest uncertainty. This is due to the standardization step: dividing by a large thickness results in a small measure; dividing by a small thickness results in a large measure. Standardizing also results in a more uniformly increasing relationship between these measures and data spacing.

Precision has an indirect relationship with data spacing as shown in Figure 5-16. Precision drops dramatically for small data spacings before leveling off at a spacing of approximately 700m, mimicking the covariance. For a given data spacing, precision is high where thickness is large and low where thickness is small. The calculation of precision requires a distance from the mean, h, defined by a multiplicative constant as in Equation 2.7. The precision values shown in Figure 5-16 are determined using a multiplicative constant of 15%. Using a relative measure like this leads to the relationship between precision and thickness exhibited in this figure.

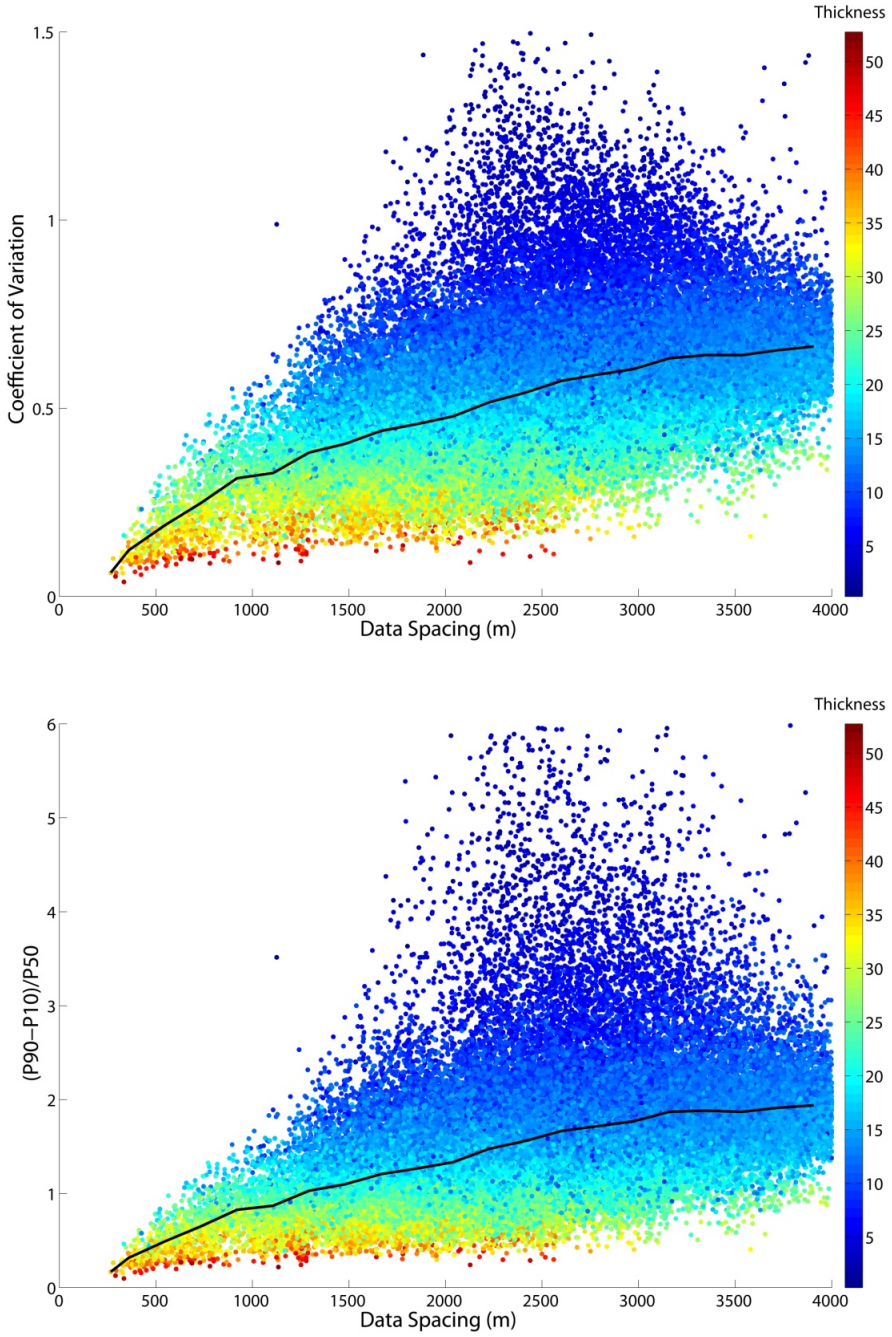

Figure 5-15: The relationship between the standardized measures of spread (coefficient of variation and (P90-P10)/P50) and data spacing for spacings from 0 to 4000m.

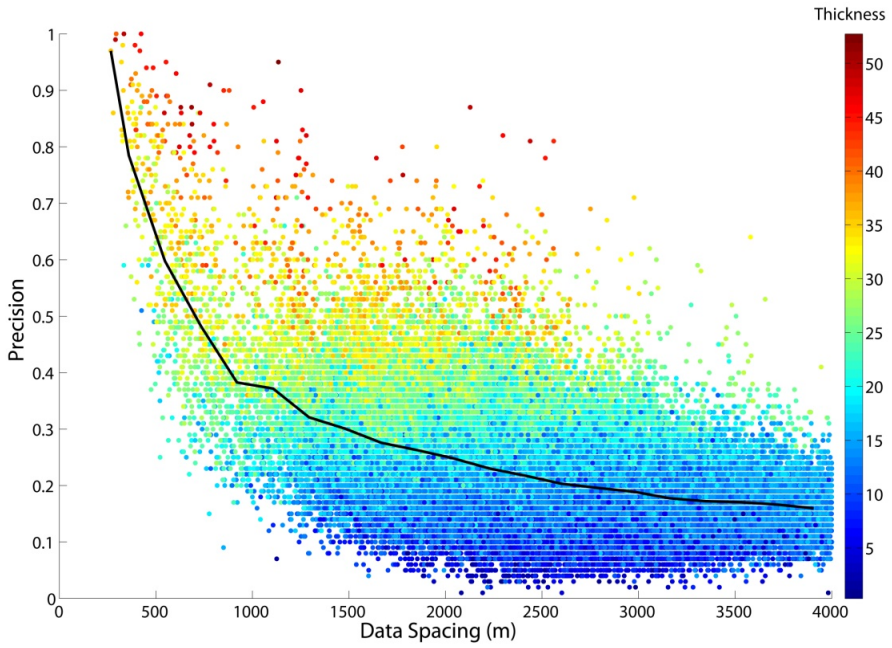

Figure 5-16: The relationship between precision and data spacing for spacings from 0 to 4000m.

5.3 Comparison of Methods

A comparison of the uncertainty versus data spacing results is presented. The five measures are considered in Figure 5-17 through Figure 5-21. The uncertainty determined by method one is represented by the black line histogram and erased box plot with the expected uncertainty being represented by the black dot. The uncertainty determined by method two is represented by the five horizontal colored lines. The dark blue, light blue, yellow, and red lines correspond to the 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentiles and the expected uncertainty is represented by the black line.

The non-standardized measures of spread shown in Figure 5-17 show reasonable agreement between the uncertainties determined using the two methods for most spacings. For the spacings between 400m and 2500m, the uncertainty determined by method two is greater than the uncertainty

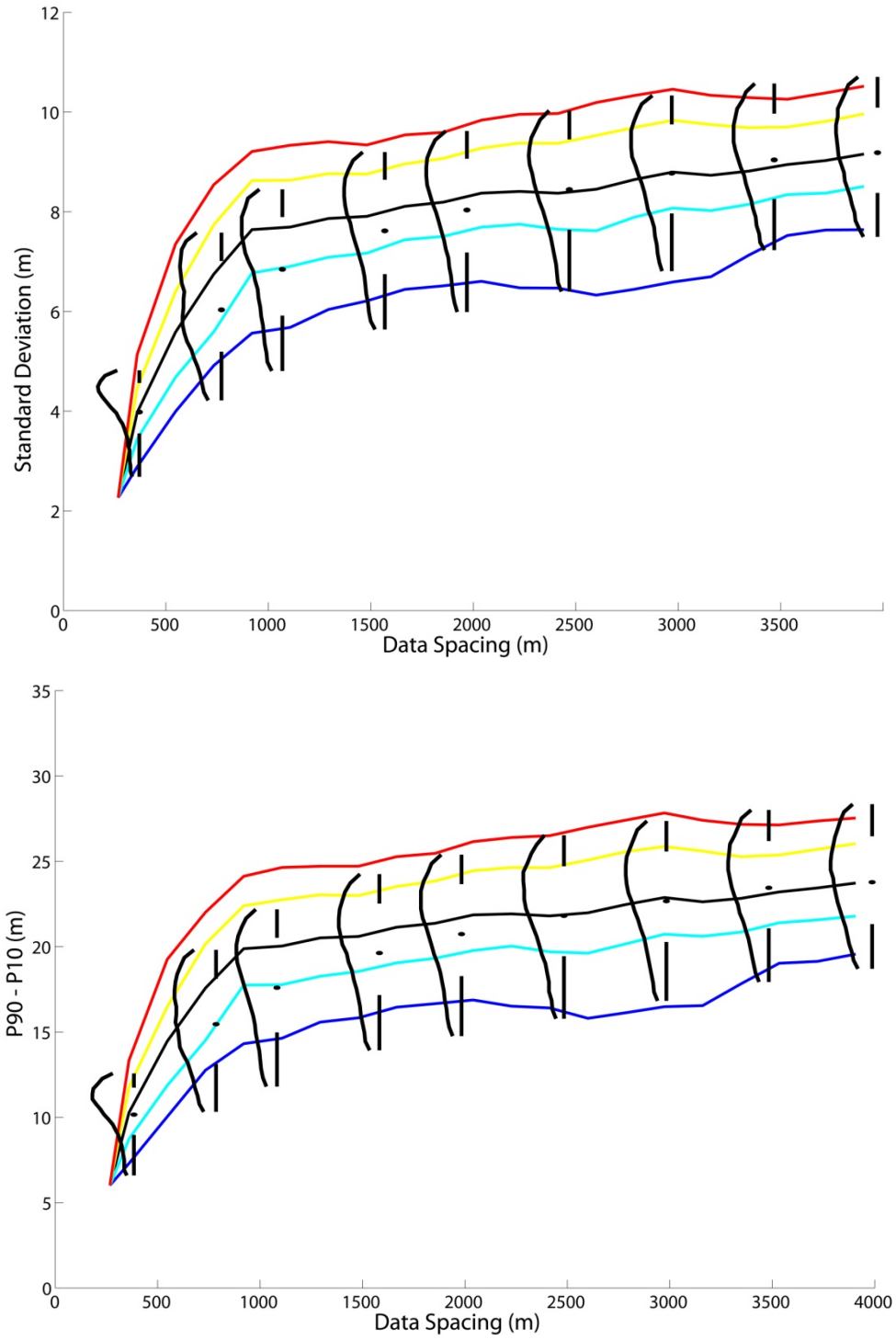

Figure 5-17: A comparison of the relationship between standard deviation and P90-P10 versus data spacing for the two methods considered.

determined using method one. The bitumen thickness data was preferentially sampled, that is, more samples were taken in areas where the bitumen layer is thick. There are no thin values with small data spacings nor are there any thick values with large data spacings as is shown in Figure 5-18. The absence of values in these ranges causes the uncertainty determined by the two methods to be different. The proportional effect (increased uncertainty in areas of large thickness as shown in Figure 5-19) causes the uncertainty determined by method two to be higher for small data spacings where only thick values occur. Method one is not subject to effects caused by preferential sampling. There are thin values with small spacing and thick values with large spacing. The uncertainty is low for the thin values, due to the proportional effect, reducing the expected uncertainty for small data spacings. The overall trend of the uncertainty versus data spacing relationship is the same for the two methods.

Preferential sampling and the proportional effect have a different effect on the standardized measures of spread shown in Figure 5-20. For spacings less than 2500m the uncertainty determined by method one is substantially greater than that determined by method two while for spacings greater than 2500m the opposite is true. As shown in Figure 5-15, these measures are highest for small thickness values and lowest for large thicknesses. The absence of any thin values at small spacings ($<$2500m) and the resulting absence of large uncertainty values for these spacings causes the uncertainty determined by method two to be less than that determined by method one. Similarly, the absence of any thick values at large spacings ($>$2500m) and the resulting absence of small uncertainty for these spacings causes the uncertainty determined by method two to be greater than that determined by method one. The overall trend of the uncertainty versus data spacing relationship is the same for the two methods.

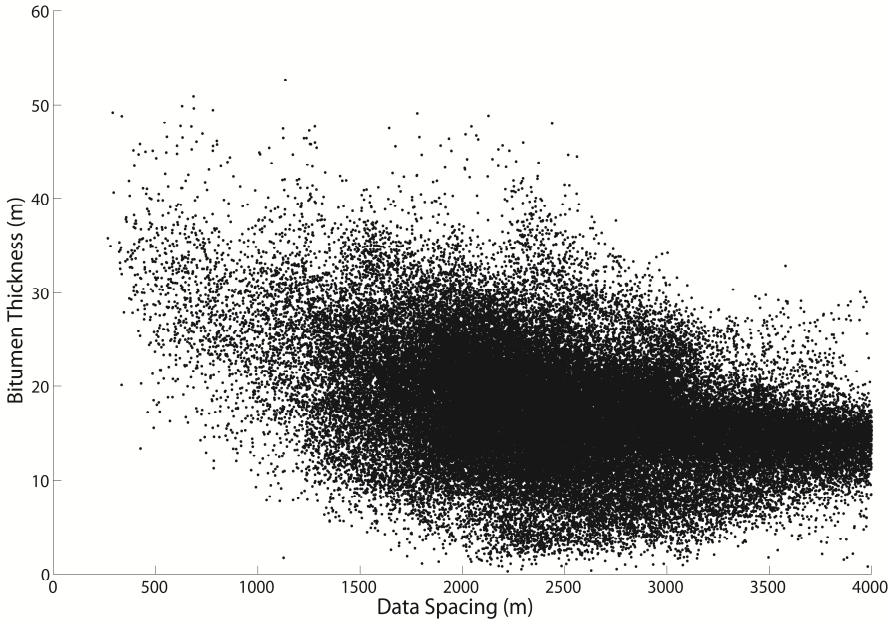

Figure 5-18: Bitumen thickness versus data spacing.

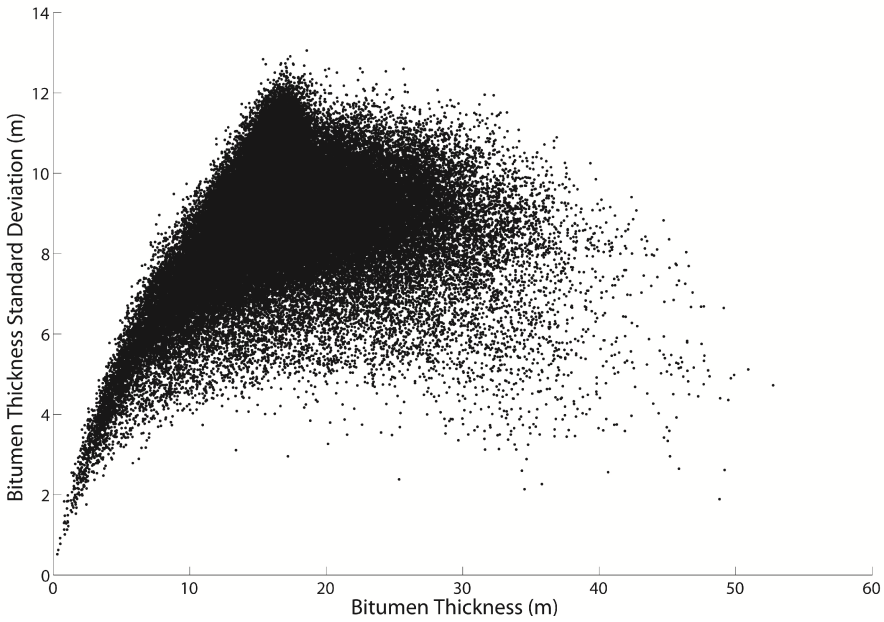

Figure 5-19: Bitumen thickness standard deviation versus bitumen thickness.

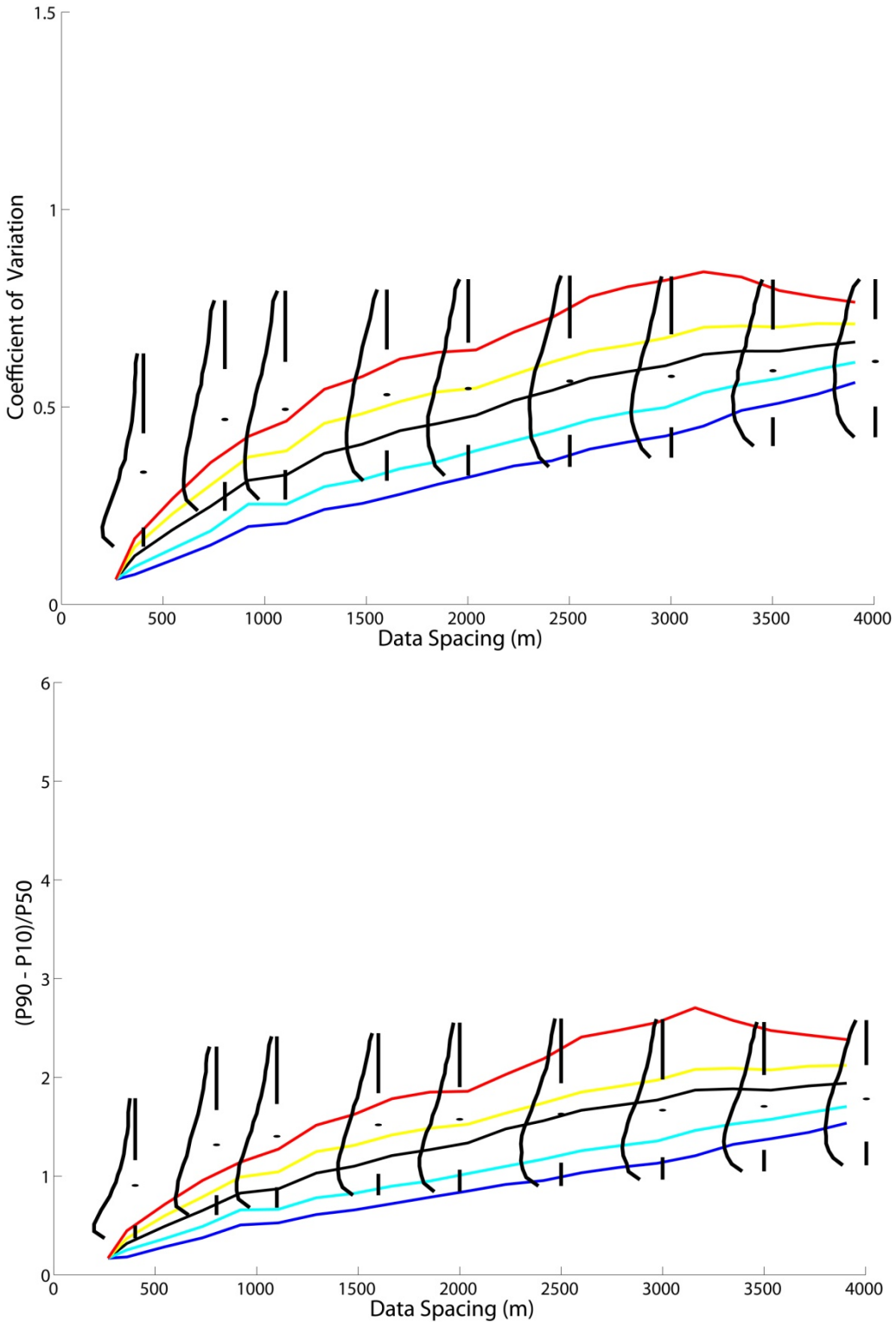

Figure 5-20: A comparison of the relationship between the coefficient of variation and (P90-P10)/P50 versus data spacing for the two methods considered.

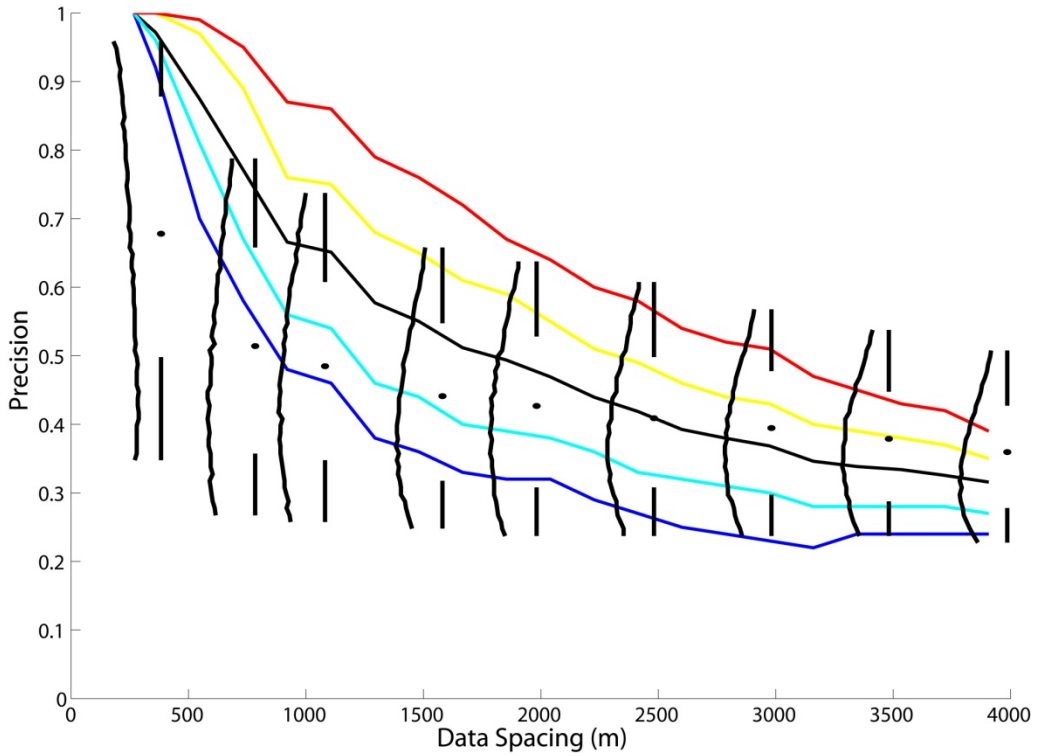

Figure 5-21: A comparison of the relationship between precision versus data spacing for the two methods considered.

Precision is also affected by preferential sampling and the proportional effect as shown in Figure 5-21. As shown in Figure 5-16, precision is highest for large thickness values and smallest for small thickness values. The lack of thin bitumen samples at small spacings (<2500m) and the resulting lack of low precision at these spacings results in the precision determined using method two being substantially higher than the precision determined using method one. The lack of thick bitumen samples at large spacings (>2500m) and the resulting lack of high precision at these spacings leads to the precision determined by method two being lower than the precision determined by method one. The overall trend of the precision versus data spacing relationship is the same for the two methods; precision decreases as data spacing increases.

The two methods generally show good agreement for the uncertainty versus data spacing relationships examined. This serves to validate the proposed methodology.

Chapter 6

Final Comments

These closing comments provide a summary of the main points, highlight the contributions and discuss possible areas for future work.

6.1 Summary

One of the primary purposes of geological modeling is to aid and support decisions. These decisions are best supported by models that meet some acceptable level of uncertainty. Uncertainty is decreased as the spacing between data decreases. This work has presented a methodology for evaluating the uncertainty as a function of data spacing.

In order to quantify the relationship between data spacing and uncertainty, geometric measures for describing the spatial arrangement of data were defined. These measures include data spacing and data density. Determining these measures at a location involves counting the number of samples that fall within some volume. These measures are a function of the number of samples and the volume that encompasses them.

It was also necessary to define measures of uncertainty. Uncertainty can be communicated in a variety of ways. The measures described herein have been found to be useful in a geostatistical context and include standard deviation, coefficient of variation, difference between specific percentiles, precision, and probability of misclassification.

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.

Uncertainty depends on more than just data spacing. A number of other factors include the decision of stationarity, uncertainty in the input parameters, the proportional effect, nonstationarity in the variogram, classification thresholds, number of realizations, and data quality. The effect of each of these factors was discussed.

6.2 Contributions

This work provides three main contributions. The first is an integrated approach for determining the relationship between data spacing and uncertainty. A program called ADUDS, which stands for the Automatic Determination of Uncertainty versus Data Spacing, implements this methodology. It takes the necessary input parameters, performs the evaluation, and outputs the distributions of uncertainty measures for the specified data spacings. A process that previously would have necessitated running multiple programs with extensive file manipulation has been consolidated into one program.

The second contribution is clear documentation and definition of various geometric and uncertainty measures. The calculations of geometric measures presented herein are robust and applicable to any spatial data. The definitions of various measures of uncertainty are useful as they are

not well defined in the literature with respect to their calculation from multiple simulated realizations.

The third contribution is an understanding of the confounding factors. It is understood that uncertainty is not entirely controlled by data spacing. The proposed methodology allowed the influence of a number of these factors to be isolated and analyzed.

6.3 Future Work

The benefits of an integrated approach to determining the relationship between data spacing and uncertainty have been demonstrated. The next step is to apply the methodology to a real problem as it relates to regulatory requirements. Codes for public disclosure suggest that the error associated with estimation for classification be quantified. The petroleum resource/reserve classification scheme presented in NI 51-101 (CSA, 2007) and the accompanying CIM guidelines mention specific probabilities. This is less common in the mining industry. Codes such as the JORC code (JORC, 2004), SAMREC code (SAMREC, 2000), SEC Industry Guide 7 (USSEC, 2006) and NI 43-101 (CSA, 2005) make no mention of specific probabilities; however, there is an expectation that the uncertainty would be quantified and used to support the final classification decision. The methodology has been applied to real data as demonstrated in Chapter 5, but no effort has been made to consider regulatory requirements.

Another area of future work is the determination of an acceptable level of uncertainty, or a procedure for making this determination. The format of the uncertainty statements could appear as universal and independent of the deposit type; however, the level must be customized for each deposit. There are no clear guidelines for choosing the parameters or thresholds of acceptable uncertainty.

The issues of uncertainty in the model and the input parameters were mentioned. These decisions affect uncertainty; however their impact was not assessed. Examining the influence of these decisions on uncertainty would be useful.

Bibliography

Alberta, 2000. Alberta Geological Survey. An Atlas of Lithofacies of the McMurray Formation Athabasca Oil Sands Deposit, Northeastern Alberta: Surface and Subsurface. Edmonton: Alberta Energy and Utilities Board. Earth Sciences Report 2000-07.

Anderson DR, Sweeney DJ, and Williams TA, 1994. Introduction to Statistics: Concepts and Applications 3rd edition. West Publishing Company. 901p.

Andricevic R, 1990. Cost-effective network design for groundwater flow monitoring. Stochastic Hydrology and Hydraulics. 4(1):27-41.

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

Barabas N, Goovaerts P, and Adriaens P, 2001. Geostatistical assessment and validation of uncertainty for three-dimensional dioxin data from sediments in an estuarine river. Environmental Science & Technology. 35(16):3294-3301.

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.

Bueso MC, Angula JM, Cruz-Sanjulian J, and Garcia-Arostegui JL, 1999. Optimal spatial sampling design in a multivariate framework. Mathematical Geology 31(5): 507-525.

Canadian Securities Administrators (CSA), 2005. National Instrument 43-101: Standards of Disclosure for Mineral Projects.

Canadian Securities Administrators (CSA), 2007. National Instrument 51-101: Standards of Disclosure for Oil and Gas Activities.

Carerra J, Usunoff E, and Szidarovszky F, 1984. A method for optimal observation network design for groundwater management. Journal of Hydrology. 73(1-2):147-163.

Criminisi A, Tucciarelli T and Karatzas G.P, 1997. A methodology to determine optimal transmissivity measurement locations in groundwater quality management models with scarce field information. Water Resources Research. 33(6):1265-1274.

Deutsch CV and Journel AG, 1998. GSLIB: Geostatistical Software Library and User's Guide 2nd edition. Oxford University Press, New York. 369p.

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

Deutsch CV, 2002. Geostatistical Reservoir Modeling. Oxford University Press, New York. 376p.

Deutsch CV, Leuangthong O and Ortiz J, 2006. A case for geometric criteria in resources and reserves classification. Centre for Computational Geostatistics Annual Report Eight, University of Alberta.

Dubrule O, 1994. Estimating or choosing a geostatistical model. In Dimitrakopoulos R (ed.) Geostatistics for the Next Century, Kluwer. 3-14.

Duggan S and Dimitrakopoulos R, 2004. Application of conditional simulation to quantify uncertainty and to classify a diamond deflation deposit. In Leuangthong O and Deutsch CV (eds.) Geostatistics Banff 2004, Springer. 419-428.

Englund EJ and Heravi N, 1992. Conditional simulation: practical application for sampling design optimization. In Soares A (ed.) Geostatistics Troia '92, Kluwer. 613-624

Goovaerts P, 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York. 483p.

Isaaks EH and Srivastava RM, 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York. 561p.

Isaaks EH, 1990. The application of Monte Carlo methods to the analysis of spatially correlated data. Ph.D. Thesis, Stanford University. 226p.

Johnson RA and Bhattacharyya GK, 1996. Statistics: Principles and Methods 5th edition. Wiley, New York. 688p.

Joint Ore Reserves Committee (JORC), 2004. Australasian Code for Reporting of Exploration Results, Mineral Resources and Ore Reserves (JORC Code), http://www.jorc.org/pdf/jorc2004web.pdf, accessed on June 21, 2010.

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

Journel AG and Deutsch CV, 1993. Entropy and spatial disorder. Mathematical Geology. 25(3):329-355.

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

Loaiciga HA, 1989. An optimization approach for groundwater quality monitoring network design. Water Resources Research. 25(8):1771-1782.

Lowrie RL, 2002. Mining Reference Handbook. Society for Mining, Metallurgy, and Exploration Inc, Littleton. 448p.

Maluf DA, Gawdiak YO, Bell DG, 2005. On space exploration and human error: a paper on reliability and safety. Hawaii International Converence on Systems Science. Accessed from http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20060016368_2006006769.pdf on July 27, 2010.

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.

McBratney AB and Webster R, 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables – II. Computers and Geosciences. 7(4):335-365.

Meyer PD and Brill ED, 1988. A method for locating wells in a groundwater monitoring network under conditions of uncertainty. Water Resources Research. 24(8):1277-1282.

Meyer PD, Valocchi AJ and Eheart JW, 1994. Monitoring network design to provide initial detection of groundwater contamination. Water Resources Research. 30(9):2647-2659.

Myers JC, 1997. Geostatistical Error Management: Quantifying Uncertainty for Environmental Sampling and Mapping. Van Norstrand Reinhold, New York. 571p.

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

Peters WC, 1978. Exploration and Mining Geology 2nd ed. Wiley, New York. 685p.

Pitard FF, 1993. Pierre Gy's Sampling Theory and Sampling Practice 2nd ed. CRC Press LLC, Boca Raton, Florida.

Rouhani S amd Hall TJ, 1988. Geostatistical schemes for groundwater sampling. Journal of Hydrology. 103(1-2):85-102.

South African Mineral Resource Committee (SAMREC) within the South African Institute of Mining and Metallurgy, 2000. South African Code for Reporting of Mineral Resources and Mineral Reserves, http://www.geolsoc.org.uk/webdav/site/GSL/shared/pdfs/Fellowship/South%20Africa%20Code.pdf, accessed on June 21, 2010.

Storck P, Eheart JW and Valocchi AJ, 1997. A method for the optimal location of monitoring wells for detection of groundwater contamination in three-dimensional heterogeneous aquifers. Water Resources Research. 33(9):2081-2088.

Tufte ER, 2001. The Visual Display of Quantitative Information 2nd edition. Graphics Press, Connecticut. 197p.

United States Securities and Exchange Commission (SEC), Industry Guide 7, http://www.sec.gov/about/forms/industryguides.pdf, accessed on June 21, 2010.

Wackernagel H, 2003. Multivariate Geostatistics 3rd edition. Springer-Verlag Berlin. 387p.

Warren A, 2003. Report 2003-A: EUB Athabasca Wabiskaw-McMurray Regional Geological Study. Alberta Energy and Utilities Board. 187p.

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

Zhang Y, Pinder GF and Herrera GS, 2005. Least cost design of groundwater quality monitoring networks. Water Resources Research. 41. 12p.

Appendix – Practical Guide

The methodology proposed herein results in a quantification of the relationship between uncertainty and data spacing as exemplified in Figure A-1 where uncertainty is measured by the standard deviation. Some direction regarding the interpretation of this plot is useful, particularly with respect to an acceptable level of uncertainty.

Defining an acceptable level of uncertainty is not trivial. It depends on the attribute being considered, the applicable stage of the project, and the motive for its definition. Determining an acceptable level requires a great deal of analysis. It involves choosing an appropriate measure and determining what level of this measure is acceptable. This is done in the context of the decision to be made, considering the consequences of the possible choices for various levels of uncertainty. The acceptable level of uncertainty is the level when enough is known be comfortable with the consequences of the decision.

Assume that for the relationship shown in Figure A-1 an acceptable level of uncertainty has been determined at a standard deviation of 0.75. This value could be in units of thickness or grade or any other variable, but specific units are omitted from this discussion without loss of applicability. The goal is to choose a data spacing that will meet the required level of

uncertainty. Simply specifying 0.75 as the required level is insufficient. Must the *expected* uncertainty be 0.75? Must 90% of the locations meet this level? 75%? 25%? Each case yields dramatically different results.

To illustrate, consider that the expected standard deviation must be 0.75. Interpolating between the expected standard deviation for a data spacing of 90m and the expected standard deviation for a spacing of 110m (green line in Figure A-2) shows that the data spacing that meets this criteria will be slightly less than 110m (blue arrow in Figure A-2).

Next, consider that the acceptable level of uncertainty requires 90% of the locations to have a standard deviation less than 0.75. Interpolating between the 90^{th} percentile for a spacing of 70m and the 90^{th} percentile for a spacing of 90m (green line in Figure A-3) reveals that a data spacing of ~75m (blue arrow in Figure A-3) meets the acceptable level.

Finally, consider an acceptable level of uncertainty where only 25% of the locations must have a standard deviation less than 0.75. This requirement is met at a spacing of 130m (see green and blue lines in Figure A-4).

These three examples illustrate the necessity of stipulating a probability as part of an acceptable level of uncertainty. Stating only a desired level of uncertainty can lead to multiple interpretations, each with different results as shown by the large spread in data spacing values from ~75m to 130m.

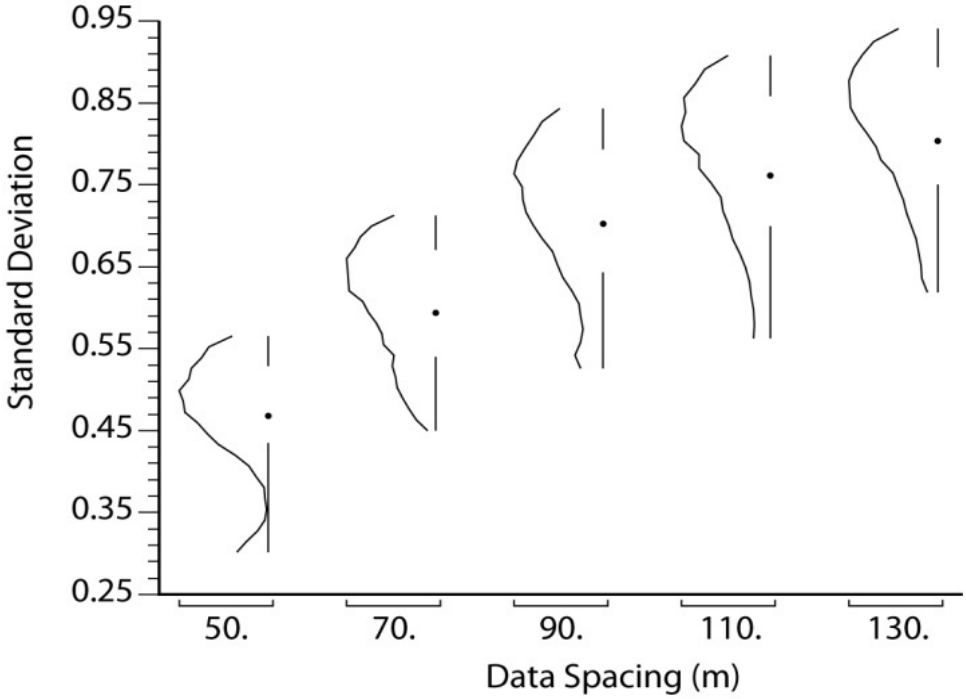

Figure A-1: An example of the relationship between uncertainty and data spacing.

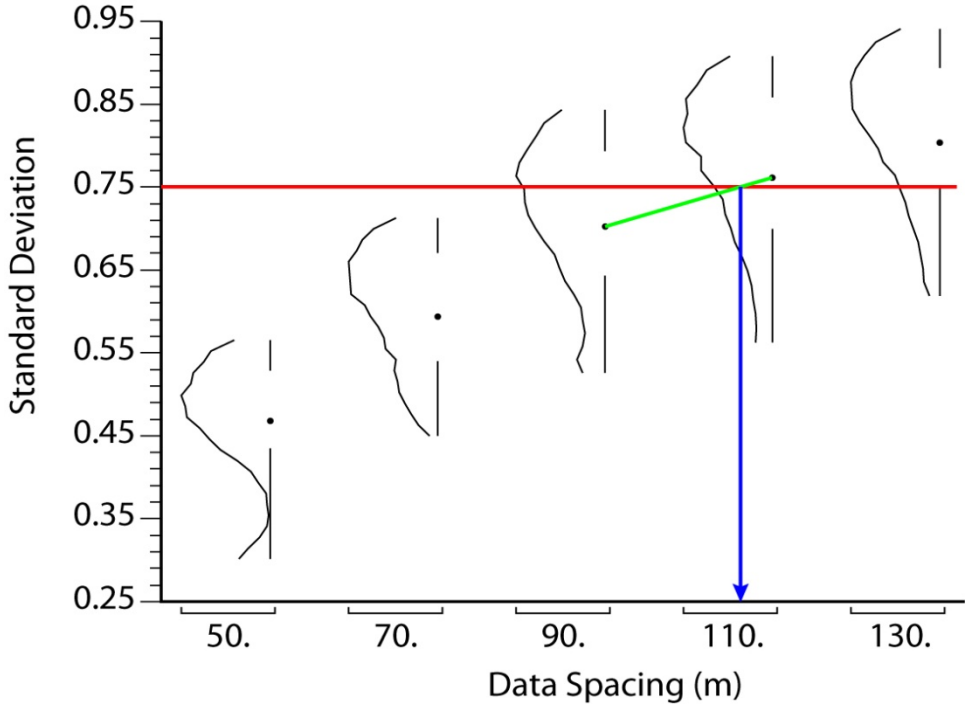

Figure A-2: Determining data spacing when expected value of 0.75 is the acceptable level of uncertainty.

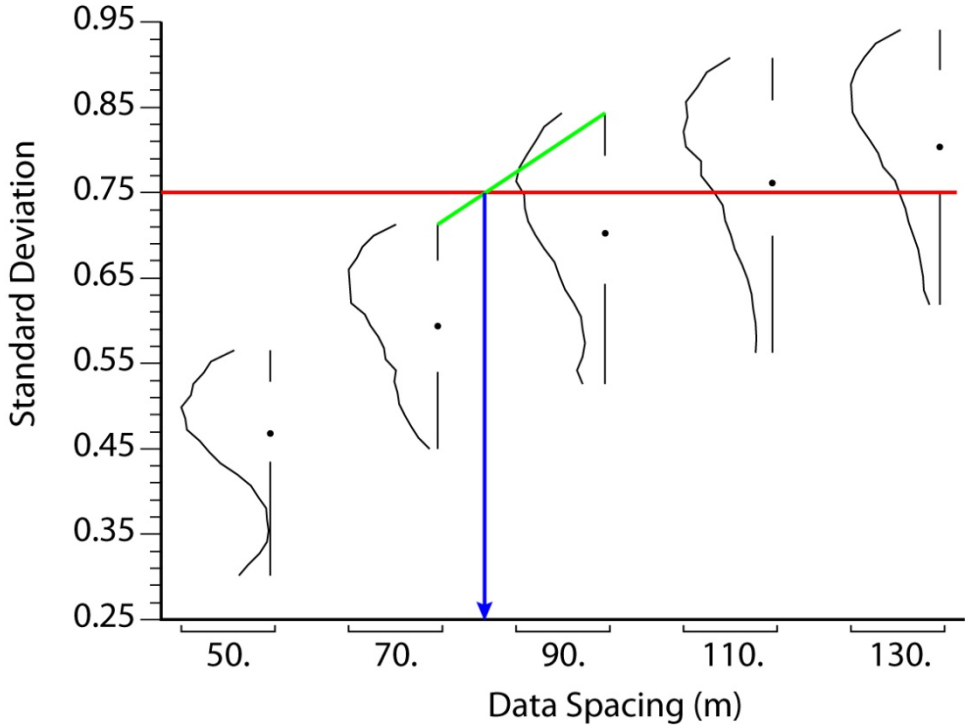

Figure A-3: Determining data spacing when 90% of locations must have standard deviation less then 0.75.

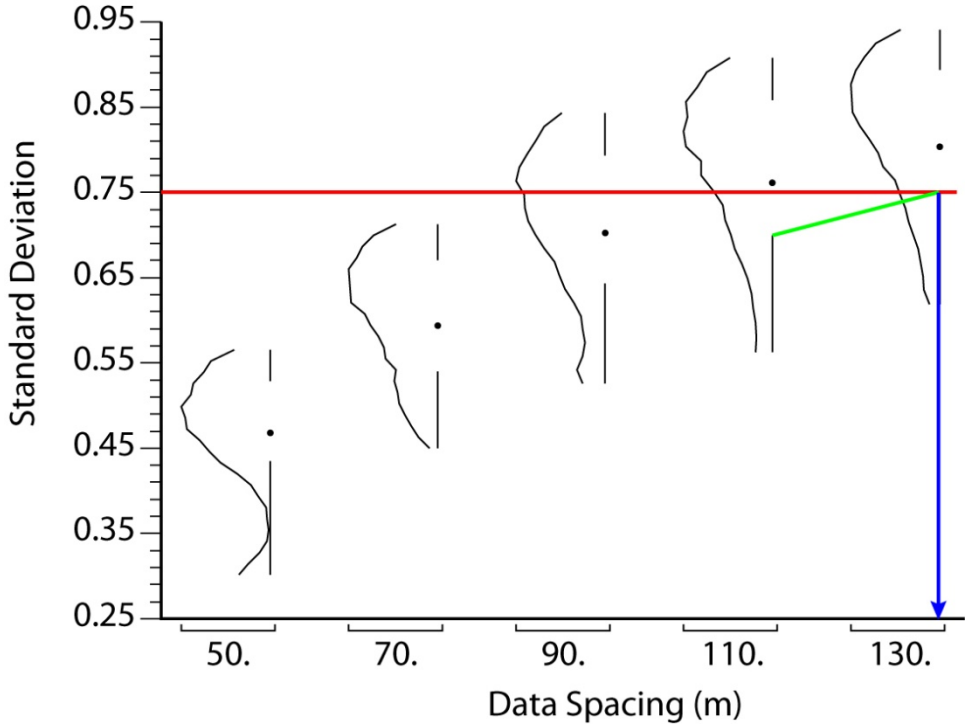

Figure A-4: Determining data spacing when 25% of locations must have standard deviation less than 0.75.