Example Application of Uncertainty versus Data Spacing

Brandon J. Wilde and Clayton V. Deutsch

A methodology for evaluating the relationship between uncertainty and data spacing is proposed in Papers 108 and 403 in this report. This methodology is applied to oil sands data from northern Alberta. The methodology is validated by evaluating the uncertainty/data spacing relationship directly from the data. The methods are compared. For data spacings less than 2500m, the first method shows greater uncertainty than the second and for spacings greater than 2500m the reverse is true. This is due to preferential sampling (thick areas have smaller spacing) and the proportional effect (thick areas have greater uncertainty).

1 Introduction

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 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. Bitumen thickness is laterally continuous; the horizontal omnidirectional variogram of the normal scores of the thickness data is shown in the top left of Figure **2**. The variogram model is isotropic with three structures summarized in Table 1.

Table 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

The relationship between uncertainty and data spacing/density is evaluated in two ways. First, the methodology proposed in Paper 108 is applied. 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 is done 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 for the establishment of the relationship between data spacing and uncertainty.

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 Paper 403. 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.

2 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 the top right of Figure 2. The distribution shapes are primarily negatively skewed. 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 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 such as the proportional effect.

The center left plot in Figure 2 demonstrates the behavior of the difference between percentiles. 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 as there are more locations far from data than close to data when samples are on a regular grid. The variance of 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 standard deviation while the other 45% is due to other factors.

The behavior of the coefficient of variation is shown in the center right of Figure 2. 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 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 standard deviation while the other 83% is due to other factors.

The behavior of the standardized difference between percentiles, shown in the bottom left of Figure 2, 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 0.44. This means that data spacing is responsible for about 12% of the uncertainty captured by the standard deviation 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 the bottom right of Figure 2. 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 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 the standard deviation while the other 75% is due to other factors.

The probabilities of the two types of misclassification errors are shown in Figure 3. 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 10th and 25th 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 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 3 than in Figure 4. 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.

3 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 with n_V=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 4. The histograms associated with these maps are shown in Figure 5. Data density is overall very low with a few small areas being 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.

SGS is used to generate 100 conditional realizations of normal scored bitumen thickness on a 100m grid (1120 x 1710). The normal score variogram in Figure **2** is used to define the spatial continuity. The normal score values are back-transformed to units of bitumen thickness according to the declustered distribution. The back-transformed values are then block averaged to 400m square blocks. Figure 6 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 7 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 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 8. 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 9. 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. The precision values shown in Figure 9 are determined using a multiplicative constant of 15%. This choice of parameters leads to the relationship between precision and spacing.

4 Comparison of Methods

A comparison of the uncertainty vs. data spacing results is presented. The five measures are considered in Figure 11 through Figure 13. 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 10th, 25th, 75th, and 90th percentiles and the expected uncertainty is represented by the black line.

The non-standardized measures of spread shown in Figure 11 show reasonable agreement between the uncertainty 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 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 10 left. 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 10 right) 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 vs. 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 12. 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 8, 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 vs. data spacing relationship is the same for the two methods.

Precision is also affected by preferential sampling and the proportional effect as shown in Figure 13. As shown in Figure 9, 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 vs. data spacing relationship is the same for the two methods; precision decreases as data spacing increases.

5 Conclusions

Two methods have been used to evaluate the relationship between uncertainty and data spacing. The two methods generally show good agreement for the uncertainty vs. data spacing relationships examined. This serves to validate the proposed methodology.

Figures

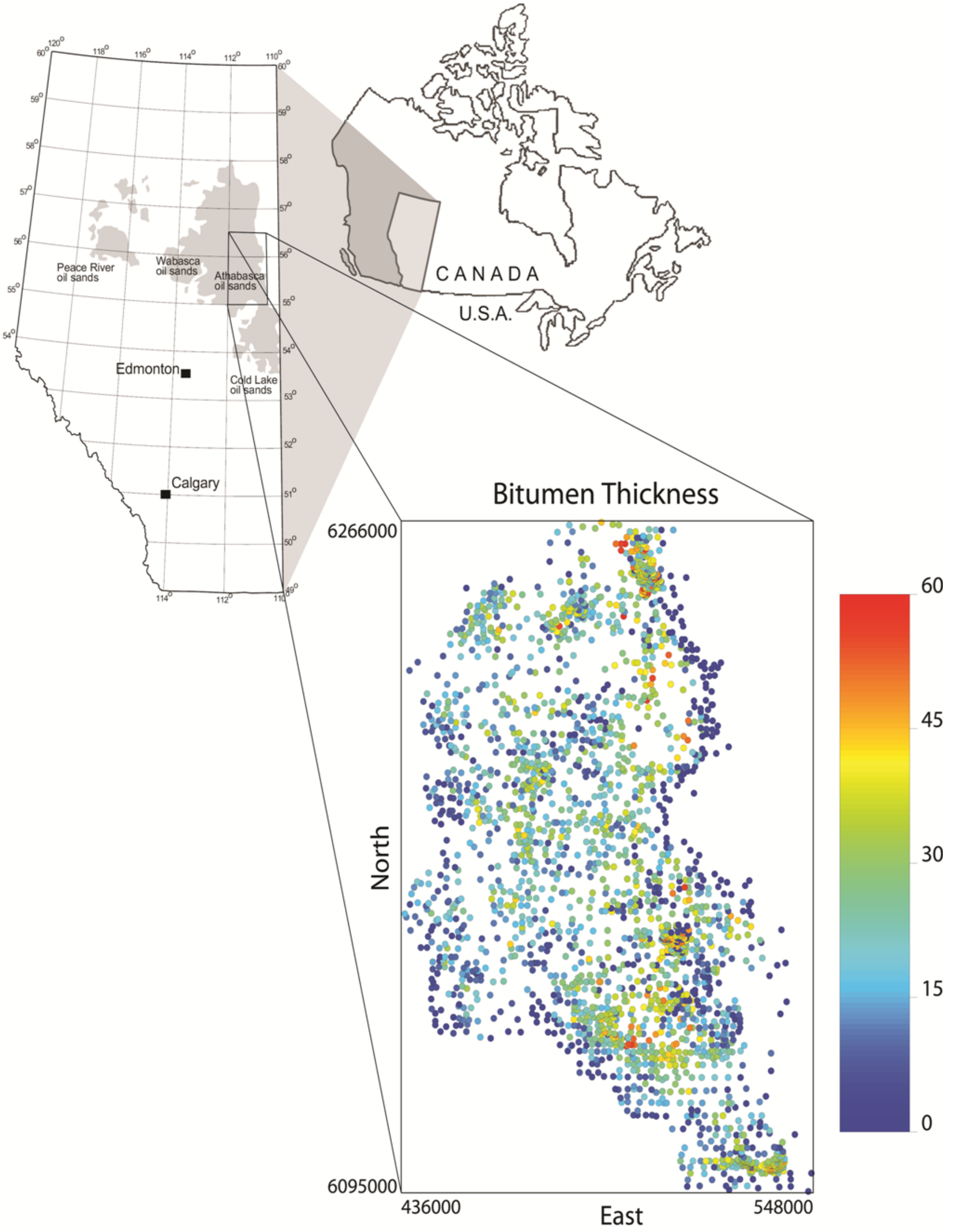

Figure 1: Location of bitumen thickness data. (adapted from Alberta, 2000)

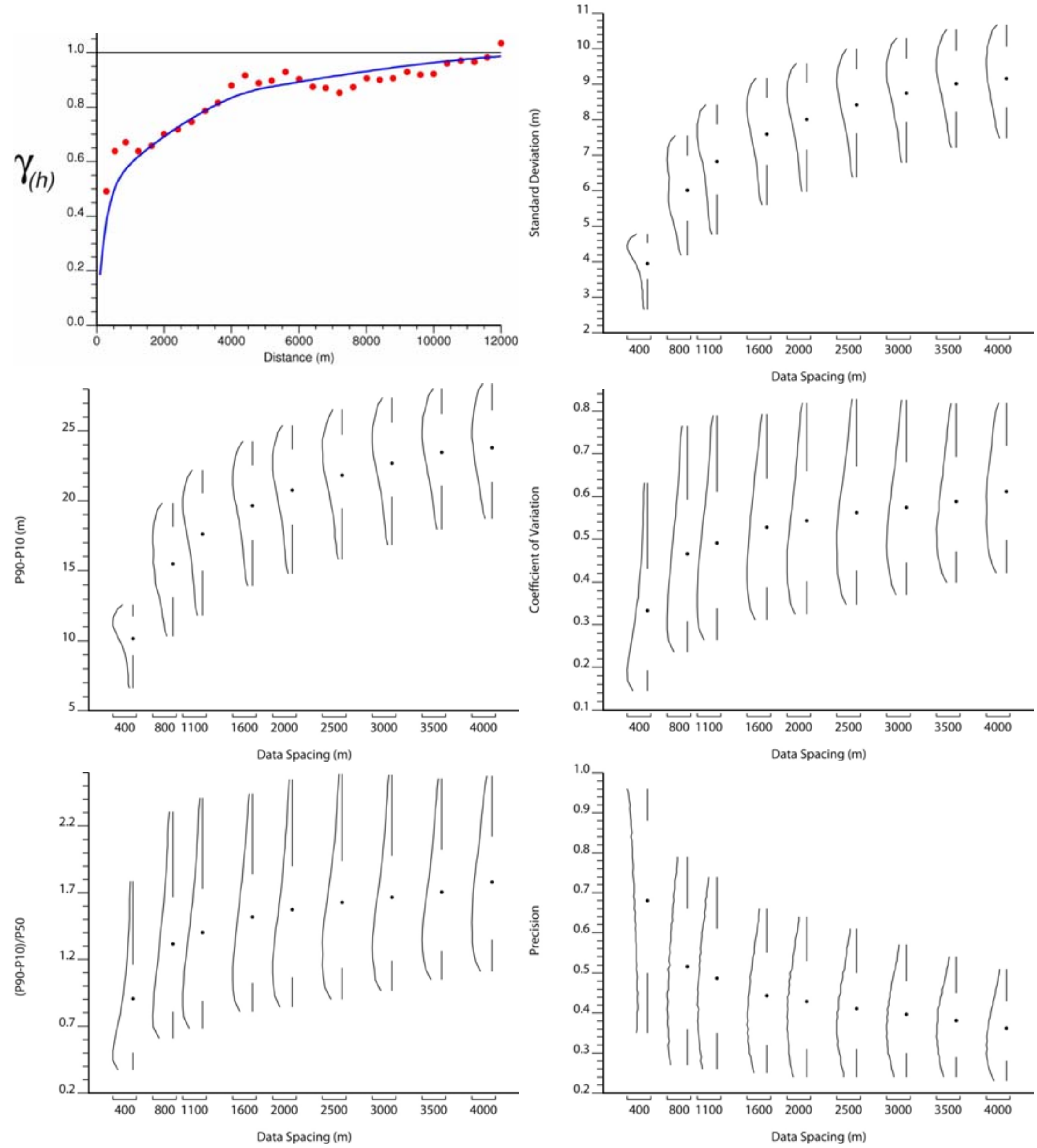

Figure 2**:** Variogram of the normal scores of the bitumen thickness (top left), standard deviation versus data spacing (top right), difference between percentiles versus data spacing (center left), coefficient of variation versus data spacing (center right), standardized difference between percentiles versus data spacing (bottom left), and precision versus data spacing (bottom right).

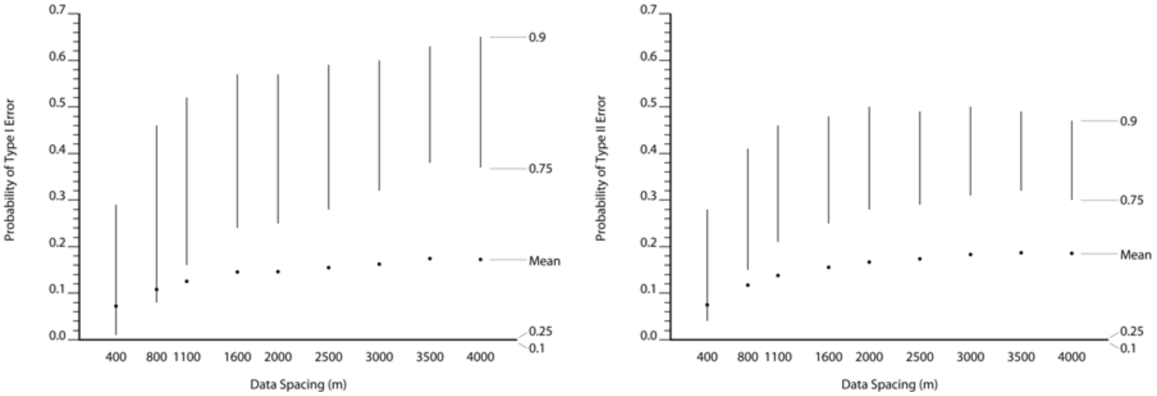

Figure 3: Probability of Type I error versus data spacing (left) and probability of Type II error versus data spacing (right) for spacings from 400m to 4000m.

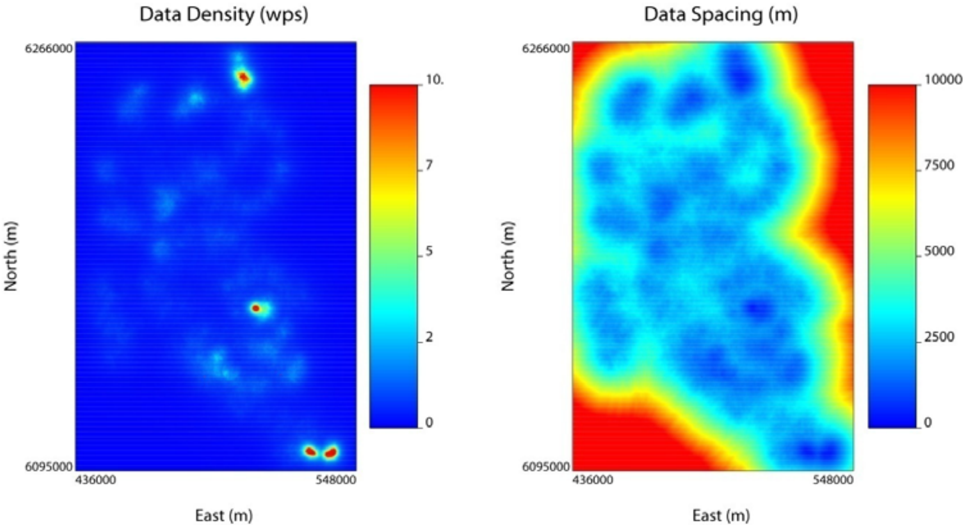

Figure 4: Data density and data spacing on 400m grid.

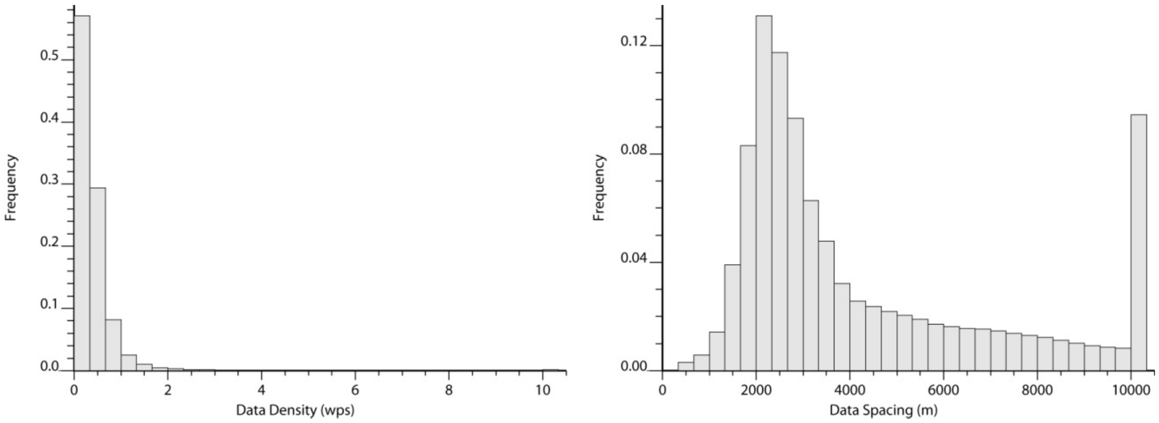

Figure 5: Histograms of data density and data spacing.

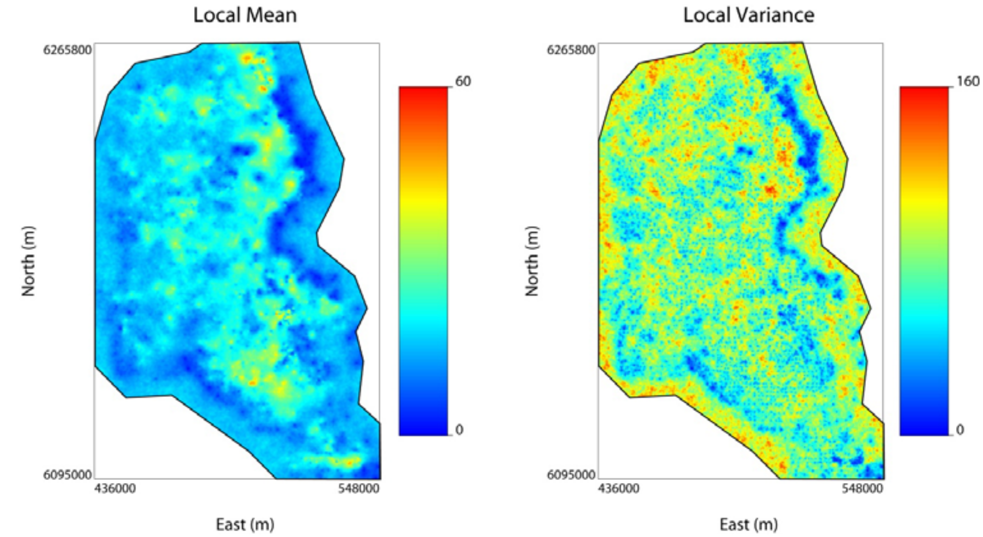

Figure 6: Local mean and variance of the 100 simulated realizations at 400m scale.

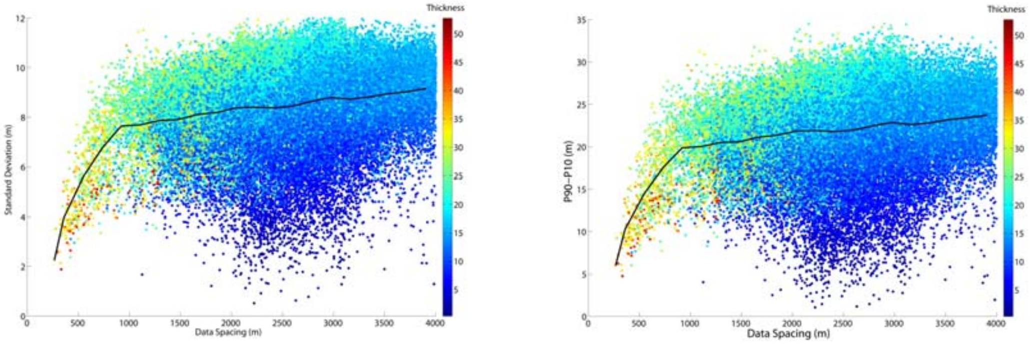

Figure 7: The relationship between the non-standardized measures of spread (standard deviation and P90-P10) and data spacing for spacings from 0 to 4000m.

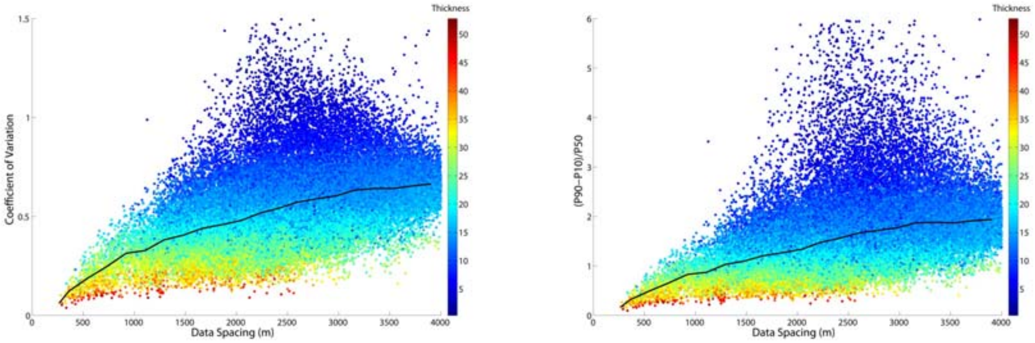

Figure 8: The relationship between the standardized measures of spread (coefficient of variation and (P90-P10)/P50) and data spacing for spacings from 0 to 4000m.

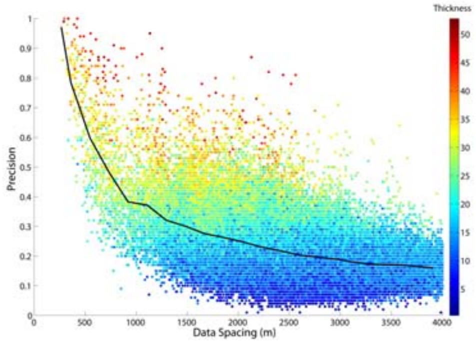

Figure 9: The relationship between precision and data spacing for spacings from 0 to 4000m.

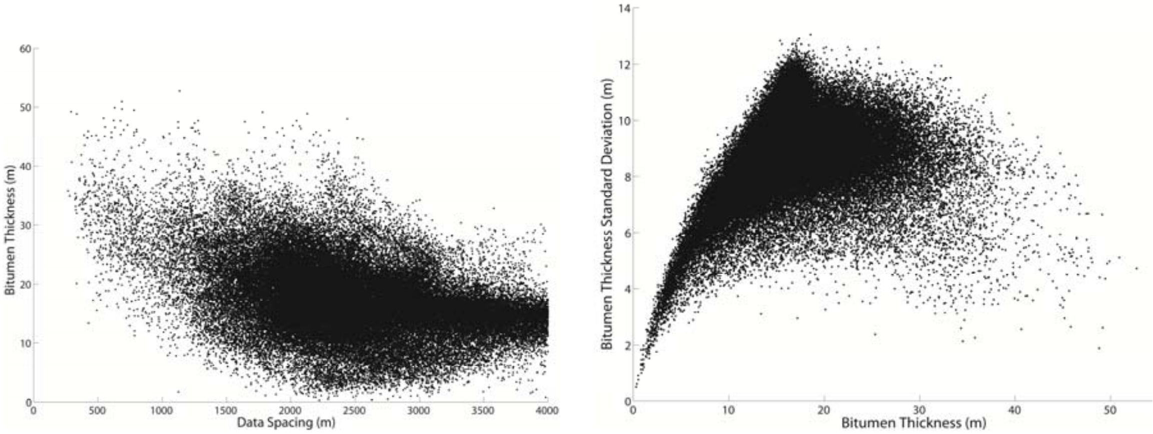

Figure 10: Bitumen thickness versus data spacing (left) and bitumen thickness standard deviation versus bitumen thickness (right).

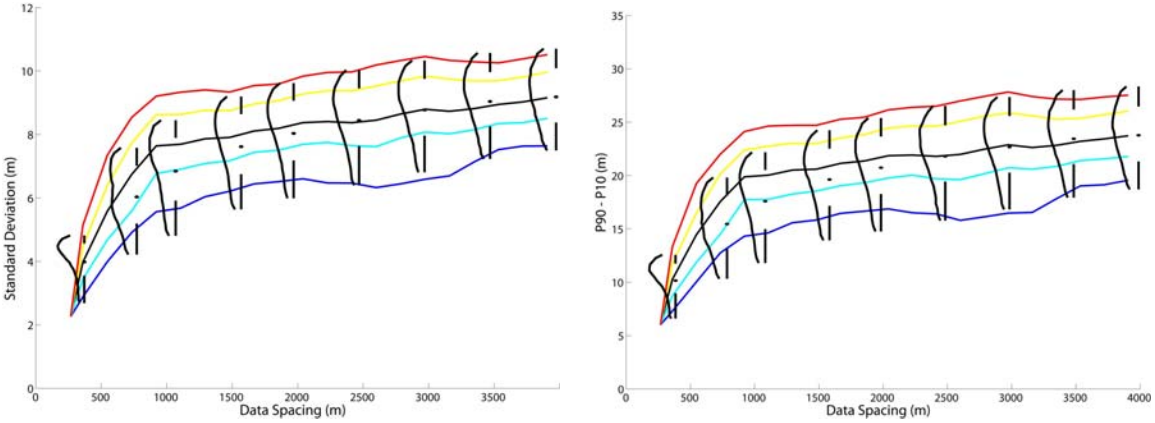

Figure 11: A comparison of the relationship between standard deviation and P90-P10 vs. data spacing for the two methods considered.

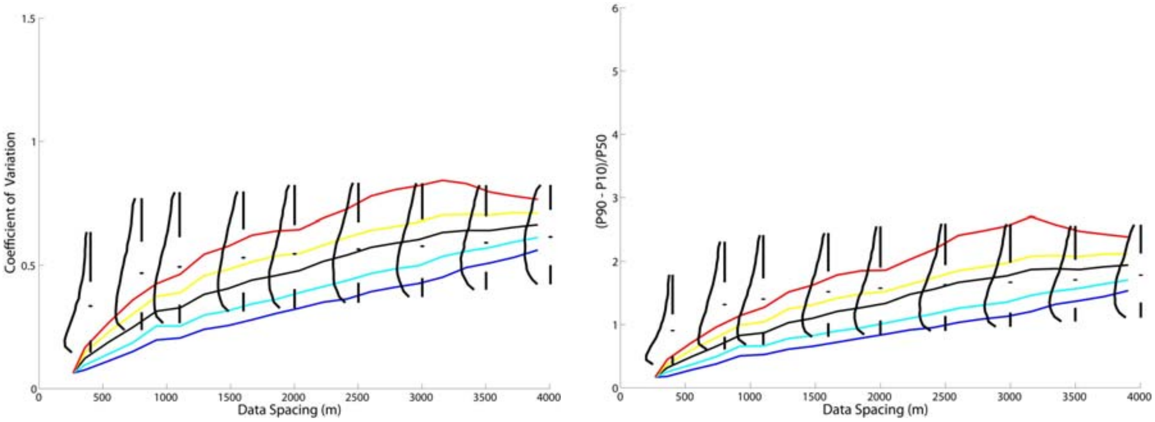

Figure 12: A comparison of the relationship between the coefficient of variation and (P90-P10)/P50 vs. data spacing for the two methods considered.

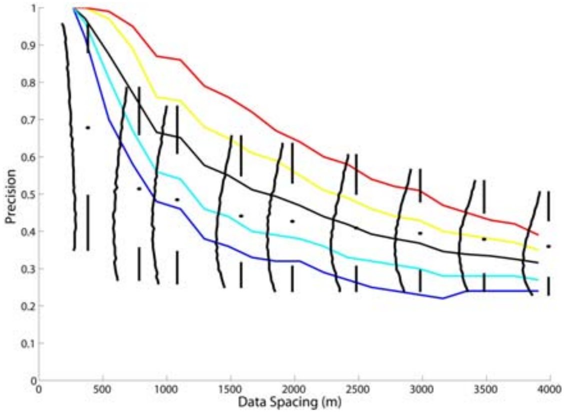

Figure 13: A comparison of the relationship between precision vs. data spacing for the two methods considered.