Steve Lyster
Kelsey MacCormack
Alberta Geological Survey
402, Twin Atria Building
4999-98 Avenue
Edmonton, AB
www.ags.gov.ab.ca

Data Spacing for Geological Modelling

Introduction

One of the most frequently asked questions to geomodellers is, "How much data do you need?" Knowing how many picks are necessary, how much log analysis should be done, how many core samples must be collected is an open question. The answer is not straightforward due to changes in the variables being modelled, the geology under consideration, the scale of investigation, and the objective of the model.

Methodology

Two different approaches were used to explore the relationship between data spacing/density and uncertainty. The unconditional simulation method of Wilde (2010) and Wilde and Deutsch (2013) was used to quantify the relationship between data spacing and uncertainty based on real data in the Duvernay and Cardium formations. This is a geostatistical simulation-based approach that resamples reference realizations at a variety of data spacings to quantify uncertainty in further realizations conditioned to the previously-simulated grids. This approach accounts for the univariate distribution and spatial structure of the data and is useful for situations where the true underlying variable is not exhaustively sampled.

In addition, synthetic grids were created to allow sampling at different spacings and for different sampling schemes. This approach has the advantage of using spatial variables that are defined at all locations, and so the models created using the artificial data can be compared to true values. The grids are modelled after real geologic features of varying complexity.

Duvernay Formation Shale

Three variables were used from the Duvernay Formation: net shale thickness, total organic carbon content, and porosity. These data come from well picks and log analyses used to assess the Duvernay in Rokosh et al. (2012); Figure 1 shows the locations of the data points. Figure 2 shows histograms of the data distributions. Net shale thickness is the most skewed and has the highest coefficient of variation, and TOC is the most symmetrical and has the lowest coefficient of variation.

Six data spacings were considered using the method of Wilde and Deutsch (2013): 800, 1600, 3200, 4800, 9600, and 19200 m. These correspond to data densities of 144, 36, 9, 4, 1, and 0.25 data per township. The spatial extent of the simulated grids was 134 km by 150 km, with about half of the cells within the grids being active (i.e., within the Duvernay extents). This is a large domain compared to the data spacing, meaning that local uncertainty is an important consideration in areas of higher or lower sampling density.

Figure 3 shows graphs of data spacing and data density vs. the average local coefficient of variation. The coefficient of variation was chosen to represent the local uncertainty because it implicitly corrects for the proportional effect, that is, higher variance in areas of higher data values.

Figure 1. Location map of Duvernay data.

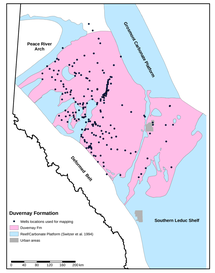


Figure 3. Graphs of data spacing and density vs. average local uncertainty.

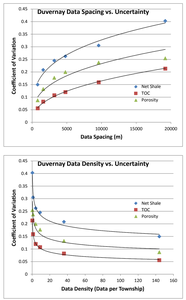
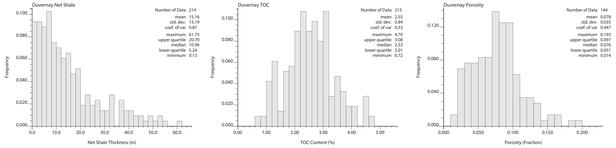


Figure 2. Histograms of Duvernay data.



Cardium Formation Tight Sandstone

The porosity-thickness variable (Phi-H) was chosen as a proxy to represent the uncertainty in the Cardium Formation (D. Bechtel, unpublished data, 2013). Figure 4 shows the locations of the data points. Figure 5 shows a histogram of the data distribution. Seven data spacings were considered: 800, 1600, 3200, 4800, 6400, 8000, and 9600 m; corresponding to data densities of 144, 36, 9, 4, 2.25, 1.44, and 1 data per township. Figure 6 shows graphs of data spacing and data density vs. average local uncertainty. The spatial extent of the simulated grids was slightly larger than four townships in area, 21.2 km by 23.2 km. The domain is relatively small compared to the data spacing intervals considered, meaning the sampled univariate distribution had a significant impact on the results relative to the spatial distribution of data.

Figure 4. Location map of Cardium data.

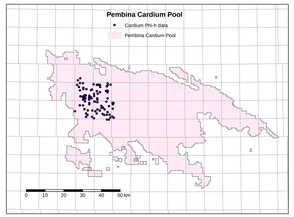


Figure 5. Histogram of Cardium data.

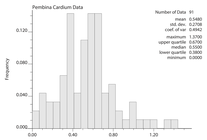
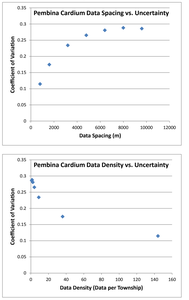


Figure 6. Graphs of data spacing and density vs. average local uncertainty.



Discussion

The examples presented here explore the relationship between data spacing (or density) and uncertainty. While it is very difficult to determine broadly applicable rules, some general guidelines and pitfalls are presented below.

Geological Complexity

The Duvernay example is similar to the more complex grids (3 and 4) from the synthetic data, in that there is more complicated spatial structure and significant randomness that cannot be predicted, only accounted for. The Cardium example is similar to the simplest synthetic grid (1) in that there are more gradual changes and a clearer spatial structure. More complex geology requires more data to achieve the same level of confidence as simpler geology.

Diminishing Returns

From the data density vs. uncertainty graphs (Figures 3, 6, and 10), it can be seen that the first several data per township provide the most value; that is, there are diminishing returns as more data are added. The more complex geological scenarios have a greater decrease in uncertainty within the first few wells per township. The simpler scenarios have a shallower curve, suggesting that the returns diminish quicker when the geology is well-behaved and more predictable as long as the underlying univariate distribution has been sufficiently sampled.

Data Clustering

The clustered data sets from the synthetic example (Figure 8) show worse RMSE and bias over the entire modelled area than the other sampling schemes (Figures 9 and 10). In particular, additional sampling data from the more complex grids (3 and 4) adds little value when that data is clustered. If the important features (channels in this case) are not found already, clustering only reinforces existing data biases. Clustering adds risk to the modelling unless some other information, such as seismic, is available to guide the sampling. For simpler scenarios without the all-or-nothing channel features there is less risk in the clustering.

Domain Size

The Duvernay data domain is quite large compared to the data spacing (Figure 1). This means that there are subareas within the domain that effectively have different data densities. The spatial uncertainty in this case is a very important factor, and sparser areas can be infilled to provide more value. There are also significant changes in the variables of interest over the domain that make some minimum level of sampling in all areas necessary to detect a trend. The Cardium domain is relatively small compared to the data spacing (Figure 4). In this case the univariate sampling distribution has a larger impact and the spatial uncertainty is less of a concern.

References

Rokosh, C.D., Lyster, S., Anderson, S.D.A., Beaton, A.P., Berhane, H., Brazzoni, T., Chen, D., Cheng, Y., Mack, T., Pana, C. and Pawlowicz, J.G. (2012): Summary of Alberta's shale- and siltstone-hosted hydrocarbon resource potential; Energy Resources Conservation Board, ERCB/AGS Open File Report 2012-06, 327 p.

Wilde, B.J. (2010): Data Spacing and Uncertainty; M.Sc. thesis, University of Alberta, 103 p.

Wilde, B.J. and Deutsch, C.V. (2013): Methodology for quantifying uncertainty versus data spacing applied to the oil sands; CIM Journal, vol. 4, no. 4, p. 211-219.

Synthetic Grids

Four synthetic grids of varying complexity were created to explore the effects of data density and sampling arrangements on modelling errors. Figure 7 shows the four grids. Each grid (1-4) consists of 10201 grid nodes that were sampled using multiple sampling schemes varying the number of data points (n=49, 81, 196, and 400) using 3 different sampling arrangements (regular, random, and clustered). These sampling schemes are meant to mimic the number of data available and sampling arrangements for a real project. Table 1 shows a summary of the sampling. Figure 8 shows the sample data extracted from grid 3.

Table 1. Data sampling schemes for synthetic grids.

Data Points (Total # of data points)	Nominal Data Spacing (km)	Nominal Data Density (Wells per township)
49	12.3	0.6
81	9.6	1.0
196	6.2	2.4
400	4.3	4.9

The sampling schemes resulted in 48 datasets that were brought into Petrel 2013. Each dataset was interpolated to the full spatial extent of the original synthetic grids. The grids nodes (n=10201) of all 48 surfaces were extracted and compared back to the original synthetic grids. This allowed us to assess the impact of data spacing and sampling distribution on the prediction accuracy of surfaces of variable complexity.

Root-mean-square-error (RMSE) was chosen as a representative statistic for the modelling results. The mean error was also used to assess the bias resulting from having limited data in different arrangements.

Figure 7. Synthetic grids used in the example.

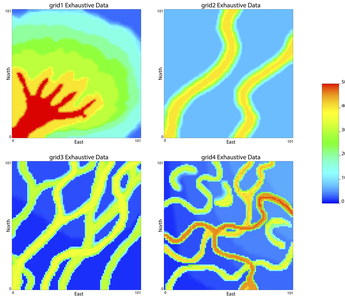


Figure 9 shows the cumulative histograms of errors for the modelled surfaces of grid 3. Similar histograms were made for all of the models, but are omitted for space. The regular sampling patterns have the narrowest and most symmetrical distributions of errors, with the random sampling results being only slightly worse. The clusted data sets produced the largest errors and most biased results, although there is a noticeable vertical portion of the distribution near zero error that represents the modelled cells near the data clusters.

Figure 8. The sample data extracted from grid 3, at four different data densities and three different arrangements.

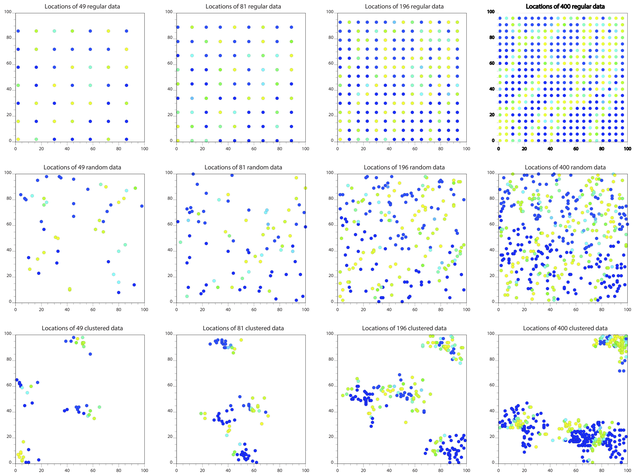


Figure 9. Cumulative histograms of the modelling errors for grid 3.

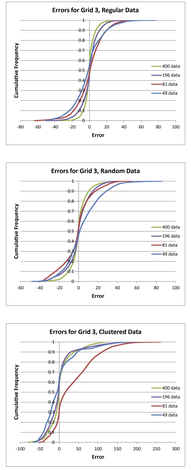


Figure 10 shows the relationship between number of data, RMSE, and bias. In general more data produced better results (lower RMSE and mean error), although for grid 3 the clustered 81 data point set the sampling happened to miss several of the channels. Grid 4 is very complex and therefore the clustered sampling produced mixed results.

Figure 10. Number of data vs. RMSE and mean error (bias) .
Blue is regular sampling; Red is random sampling; Green is clustered sampling.