Introduction

One of the most frequently asked questions to geoscientists is "how much data do you need?" Knowing how many 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 modeled, the geology under consideration, the scale of investigation, and the objectives of the model.

Methodology

Two different approaches were used to explore the relationship between data spacing and prediction errors. The first is the uncorrelated simulation method of Williams (2003) and Wilks and Deuel (2017). This was used to quantify the interdependence between data spacing and uncertainty based on real data in the Duvernay and Cardium formations. This is a geostatistical simulation-based approach that requires reference realizations at a variety of data spacings to quantify uncertainty in further realizations conditioned to the previous realizations. This approach assumes the universal distribution and spatial data of the spatial distribution and is useful for allocations where the true underlying variable is not well sampled. In addition, synthetic grids were created in the same sampling of 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 variables. These grids are modeled after geostatistics features of varying complexity.

Synthetic Grids

Four synthetic grids of varying complexity were created to explore the effects of data density and sampling arrangements on modeling errors. Figure 7 shows the four grids. Each grid (1-4) consists of 10201 grid nodes that were sampled using a random sampling scheme varying the number of data points (143, 456, 1456, and 4560) using different sampling arrangements (regular, and random). 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 data extracted from grid 3.

The sampling schemes resulted in 45 datasets that were brought into Petrel 2015. Each dataset was reinterpreted to the full spatial extent of the original synthetic grids. The grids (1-201) of all datasets 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.

Randomly sampled error (RMSE) was chosen as a representative statistic for the modeling results. The mean error was used to assess the bias resulting from having limited data in different arrangements.

Duvernay Formation Shale

Three variables were used from the Duvernay Formation: shale thickness, total organic carbon content, and porosity. These data come from field, well, and log analyses and are acquired in West Wilkes (MacCormack et al., 2017). Figure 8 shows the locations of the data points. Figure 9 shows a histogram of the data distribution. Seven data spacings were considered: 0.00, 0.05, 0.09, 0.45, 0.80, 0.90, and 0.95 in Duvernay data. This is a large domain compared to the data spacing, but the data spacing is an important consideration in areas of higher or lower sampling density.

The prediction error of the models created using the artificial data can be compared to true values. Table 1 shows a summary of the number of data points (n=49, 81, 196, and 400) using 3 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 variables. These grids are modeled after geostatistics features of varying complexity.

Cardium Formation Tight Sandstone

The porosity-thickness variable (Phi-H) was chosen as a proxy to compare the univariate distribution had a significant randomness that cannot be predicted, only accounted for. The Cardium data domain is much smaller in area than the synthetic grids (1/10), and there are more gradual changes and a clearer spatial structure. More complex geological structures make more data to achieve the same level of confidence as more simple geological.

Diminishing Returns

The statistics for each variable (Figures 3 & 10) show the first few data points per bin provide the most value. There are diminishing returns in terms of additional data. This is consistent with the synthetic data, in that there is more complicated spatial structure and significant randomness that cannot be predicted, only accounted for. The Cardium data domain is much smaller in area than the synthetic grids (1/10), and there are more gradual changes and a clearer spatial structure. More complex geological structures make more data to achieve the same level of confidence as more simple geological.

Data Clustering

The clustered data sets from the synthetic example (Figure 8) show how RMSE and how variable the data spacing provides the best results (lower RMSE and mean error), although this grid is dominated by data points for the all error in the clustered sampling. Grid 4 is very complex and therefore the clustered sampling produced mixed results.

References


The discussion demonstrated how the relationship between data spacing and prediction accuracy of surfaces of variable complexity. The results show the dependence between data spacing (Figure 4) and errors, with the random sampling results being only slightly worse. The sampling schemes resulted in 45 datasets that were brought into Petrel 2015. Each dataset was reinterpreted to the full spatial extent of the original synthetic grids. The grids (1-201) of all datasets 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.