Introduction

The importance of handling uncertainty of data and being able to present the ambiguity of geo-models of any kind have got more attention the last couple of years.
The same way it is important for a geo-modeler to understand the uncertainty and limitations of data to make an adequate geo-model, it is important for a decision maker to understand the uncertainty and limitations of a geo-model to perform adequate decisions.
Whether the geo-model is made to manage infrastructure projects or nature resources like oil, gas, or groundwater, being able to understand the uncertainty and limitations of a geo-model potentially have great economic value.

One way to combine information and data is to fine one optimal model, typically the simplest or smoothest model, that will fit data. However, due to the underdetermined nature of the inverse problem to be solved (Making a geo-model is an inverse problem), several solutions will fit the available data and information about the problem. One way of presenting the uncertainty and ambiguity of a geo-model is to present a suite of geologically realistic geo-models, that represent the uncertainty of all available information.

Multiple-Point Statistics (MPS) is an overall methodology representing a series of simulation algorithms all utilizing the available information (geophysics, boreholes, background knowledge, etc.) to produce a series of geo-model realizations all fitting the available information (see e.g. Guardiano and Srivastava 1993, Strebelle 2002, Mariethoz et al. 2010, and Hansen et al. 2016). Generating many of these realizations allow for any kind of statistical questions to be answered – What is the probability of having sand at a specific location and depth? What is the probability that location A and B are sitting in fully connected sand layer? etc.

Depending on the available data and background knowledge you have, these geo-models can be more or less informative. Some questions might be answered based on the available models, others might not be answered with a satisfying confidence. To achieve a higher level of certainty in your geo-models more information needs to be added.

In this study we demonstrate a methodology where we utilize a series of geo-model realizations to tell us where we should obtain more data – e.g. drill another borehole, to update our probabilistic model such that it adds most information, and hence reduces the uncertainty the most.

Method and Theory.

The different 'questions' you can ask your MPS realizations are numerous and very problem-specific. One question is, however, relevant to ask regardless of the task you are trying to take decisions based on MPS modelling – Where do we lack information?

We present here a methodology not only to answer this question but also to the question where information should be obtained to update the probabilistic model the most.

The methodology consists of the following 5 steps:

1) Generate a series of MPS realizations
2) Compute the Shannon Entropy
3) Convolve the Entropy Grid with Gaussian Kernel
4) Compute the weighted accumulated Entropy at each location.
5) Find the location with the maximum weighted accumulated entropy.

After generating a series of MPS realizations the Shannon entropy (eq. 1) (Shannon 1948) of those realizations is computed. Mathematically the Entropy S takes the form:

$$S = -\sum_i^N (\mathrm{P}_i \log_N \mathrm{P}_i) \quad (1)$$

where N is the number of geological categories in the realizations, P_i is the probability of category i, and $\log_N$ is the logarithm of base N. The logarithmic function does not need to have the base N, but this will constrain the entropy value to be between 0 and 1, which is convenient when comparing entropy values between different cases with a varying number of geological categories. The Shannon entropy comes from information theory and is a measure of information. The output is a number between 0 and 1 and represents lack of information. An entropy of 1 represent complete ignorance. If all realizations contain the same category at a specific node in the simulation grid, the entropy of that node will be 0 – meaning no uncertainty or fully informed. On the other hand, if half of the realizations contain one category and the other half the other category, in a binary case, the entropy will be 1 – meaning no information or full uncertainty.

When the entropy has been computed, the next step is to convolve the entropy grid with a Gaussian Kernel. This is done to include the spatial varying entropy around each node. This way we do not just get a measure of nodes with low information, but a measure of low informative nodes surrounded by other low informative nodes.

The final step of the methodology is to locate the point in the convolved entropy grid with the highest value. This point will be the place where information should be added in order to update the probabilistic model the most.

Case Study

The methodology described above is carried out on a dataset from an area just south of Aarhus in Denmark. The data used in this case study are real, however, the training image and the simulation grid is set up to showcase the methodology and not performed properly in a geologic context.

The MPS Simulation was set up as a binary hydrogeologic problem, assuming only two different hydrogeologic materials – one resistive containing sand and gravel, and one more conductive including clays. For simplicity, these two categories will be called sand and clay. The simulation grid is a 3D grid with 61x61x21 voxels with voxel sizes of 50x50x10 meters. The simulations are made between the terrain and an arbitrary subsurface boundary layer with varying topography.
The MPS realizations was obtained based on three sources of information - a borehole data set acting as hard data, a ground-based time domain electromagnetic, TEM40, dataset, acting as soft data, and a training image representing geological prior information.

The lithologic borehole data was categorized into the two categories sands and clays, and since treated as hard data, all realizations were conditioned and fixed with the lithologic information from the boreholes.
Setting up the geophysical data to act as a soft data is done in three steps. First the TEM soundings are interpolated into a 3D grid. Then a transfer-function between resistivity and lithology is computed based on information from boreholes and the co-located resistivity values in the resistivity grid. The resistivity grid is then translated into a probability grid of the two categories sand and clay using the transfer function. The 3D training image is displayed in **Figure 1**.

The MPS realizations are obtained using the Simulation Extension in the geoScene3D software (www.geoscene3d.com), which builds on top of the open source MPS code MPSLib (Hansen et al. 2016). The simulation is based on a random simulation path, five multiple layers and a 3D template of size 5x5x3. The specific details and theory concerning the MPS algorithm is outside the scope of this paper. The algorithm was set to simulate 10 realizations, and **Figure 2** shows a cross-section through 4 of them. Based on these 10 realizations, an entropy grid was computed using equation (1).

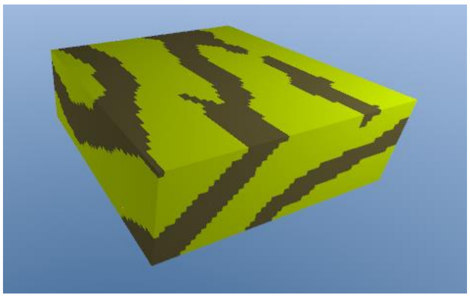

Figure 1: *The 3D training image used to represent the geologic background knowledge of the study area.*

This grid was in turn convolved with a Gaussian Kernel – a smoothed Dirac delta function in 3D - with a grid geometry of 11 by 11 by 5 cells, and a standard deviation in the three dimensions of 3,3, and 1, respectively.

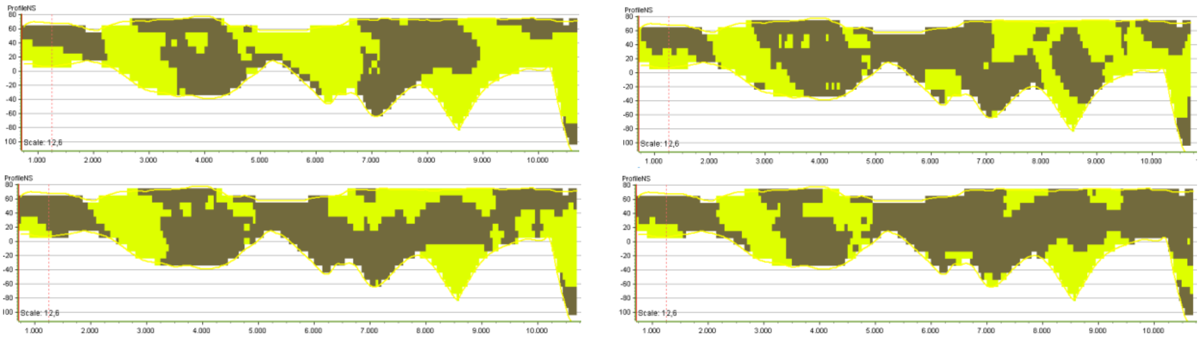

Figure 2 *Cross-sections through 4 of the 10 realizations generated using Multiple-Point Statistics.*

The result is a 3D grid with values ranging from 0 to 1. To represent the results with respect to certainty instead of *un*certainty, 1 minus the convolved entropy grid is computed. The low values, now, represent low certainty and values close to 1 represent high certainty - or information content. Now the weighted accumulated certainty is computed. This measure is the accumulated (1-entropy) values with depth for each xy location in the grid, and in turn weighted with the thickness of the simulation layer. Mathematically, the weighted accumulated certainty in each xy location is computed as

$$WAC = \frac{1}{N}\sum_{n}\left(1 - S(n)\right) \qquad (2)$$

with *N* being the number of grid cells within the simulation layer at each xy cell location and S(*n*) being the entropy computed in each cell using equation (1). The lowest *WAC* in the resulting grid, represents the position with the lowest amount of information. This position is plotted together with the 1-Entopy values in the uppermost 10 meters below terrain (i.e. the thickness of the voxel in the 3D grid) and displayed in left panel of **Figure 3**. The place where you gain the most information to the probabilistic model by adding more information is found.
Now, a lithologic borehole log is added to this location and the MPS simulations and in turn the computations described above is remade on this updated model. See results in **Figure 3 Right.**

Discussions

It the right plot in **Figure 3** it is seen that the least informative position is found close to a set of data points. This might seem curios, but several factors might influence the results in this way. First of all, the data points close to the lowest *WAC* position are all TEM40 soundings. They are treated as soft data, and not necessarily very informative. Another reason might be due to the fact that not only the pointwise most uninformed location is found, but the most uninformed location in a low informed neighborhood. A third reason might be to the limited number of realizations in the MPS simulation. In this case only 10 realizations were generated. This number should ideally be much higher to correctly be able to represent the full posterior probability distribution.

Conclusions

In this study we have presented a workflow allowing you to answer the question: Where do we lack information? The workflow is based on the computation of the information content in a series of Multiple-point statistics realizations, and in turn the least informative location within a neighbourhood of low information content is computed by convolving the entropy grid with a Gaussian 3D Kernel. The result is the location where you should obtain additional information in order to update the probabilistic model the most.

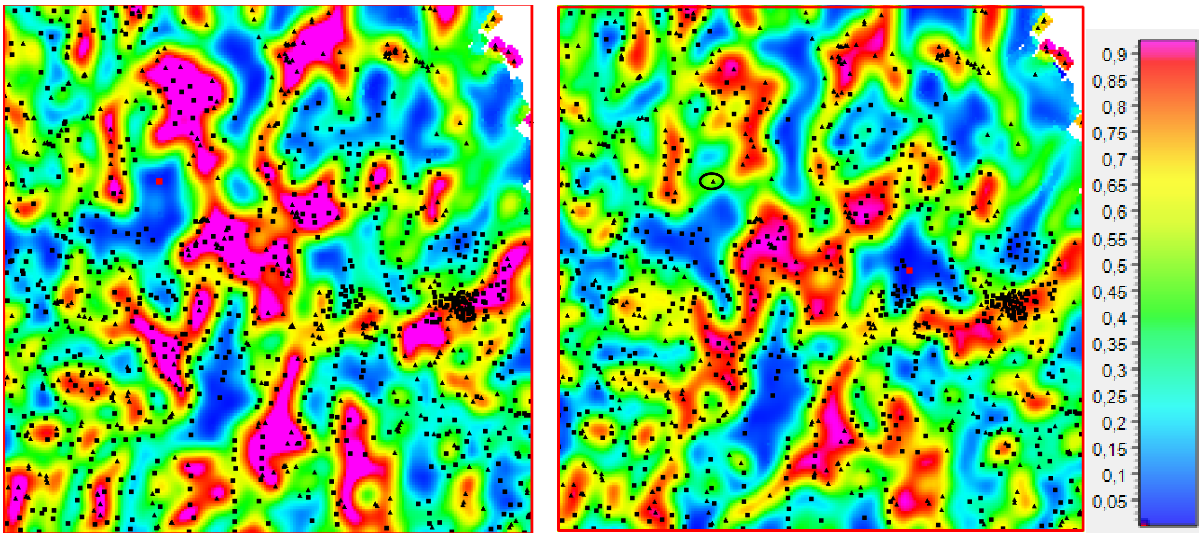

Figure 3: *1-Entropy in the uppermost 10 meters below terrain. Warmer colors represent high certainty - based on the MPS realizations - and cooler colors less certainty. Black squares represent positions with TEM40 soundings, black triangles represent borehole locations, and the red squares represent the positions with the lowest weighted accumulated certainty.* ***Left****: the results from the first MPS simulation.* ***Right****: the results when rerunning the simulation after adding borehole information (position indicated with an ellipsoid) to the computed lowest WAC position.*

Acknowledgement

We would like to thank the Innovation Fund Denmark (Case 7039-00006B) and the Danish Environmental Technology Development and Demonstration Program (MUDP) for funding this study.

References

Guardiano, F. and Srivastava R. [1993] Multivariate geostatistics: beyond bivariate moments. *geostatistics-Troia*, **1**, 133-144

Hansen, T.M., Vu, L.T., and Bach, T. [2016] MPSLIB: A C++ class for sequential simulation of multiple-point statistical models. *SoftwareX*, **5**, 127-133

Mariethoz, G., Renard, P., Straubhaar, J. [2010] The Direct Sampling method to perform multiple-point geostatistical simulations, *Water Resources Research*, **46** (11)

Shannon, C.E. [1948] A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**(3), 379-423

Strebelle, S. [2002] Conditional simulation of complex geological structures using multiple-point statistics, *Mathematical geology*, **34** (1), 1-20