Predictive Mapping with Self-Organising Maps: A Synthetic Study

Angela Carter-McAuslan, Peter G. Lelièvre, Colin G. Farquharson

Department of Earth Sciences, Memorial University of Newfoundland

Introduction

Geological mapping in remote areas or in areas with limited outcrop exposure can be challenging and expensive. The development of predictive pseudo-geological maps from remote sensing data can enable an educated guess to be made about the lithological distribution and may present an avenue by which these challenges can be mitigated. Generation of pseudo geology maps can be accomplished through the application of computer algorithms designed to cluster or classify multivariate datasets. In this project self-organizing map algorithms (SOMs) are being investigated as a possible means of multivariate dataset analysis of remote sensing data for lithological discrimination. Here we present a synthetic study illustrating the capabilities of SOMs.

Self-Organizing Maps Algorithms

SOMs are a class of unsupervised neural network algorithms which use a statistical approach to cluster multivariate datasets. SOMs were developed by Kohonen (1982), and are designed to reduce the dimensionality of multivariate data by clustering the data in a 2D computational map space (i.e. a neural map). This type of algorithm has been applied to a number of different earth science problems, including analysis of geochemical (Iwashita et al., 2011), and remote sensing data (Ji, 2000). The implementation of SOMs used in this project, SiroSOM, was developed specifically for use with geoscience data by Australia's CSIRO.

Synthetic Model and data

The work flow investigated in this poster is summarised in Fig. 1. The first step was to develop a synthetic model (Fig. 2). This model consists of six geological units. Each unit has the form of a simple geological structure, is assigned a general depth from surface, and is assigned five properties (p_1, p_2, p_3, p_4, and p_5) associated respectively with five data types (d_1, d_2, d_3, d_4, and d_5). The form, depth, and properties of the units are summarised in Table 1. The physical properties of geological bodies are not constant and have internal variability. To replicate this variability the properties set for each unit were treated as the mean of a normal distribution and the value for that property at any point was allowed to vary in a Gaussian fashion for a given standard deviation. A standard deviation of 0.1 was used for the examples discussed here. The relationship between a property and its associated data, the depth sensitivity and spatial resolution for each of the five datasets created in this project were based on the behaviour of a specific type of geophysical data and its associated physical property; as such, each dataset can be said to be analogous to a type of geophysical data. The relationship between data and property, the depth sensitivity, and spatial resolution, and analogous geophysical method are summarised for the five data types in Table 2.

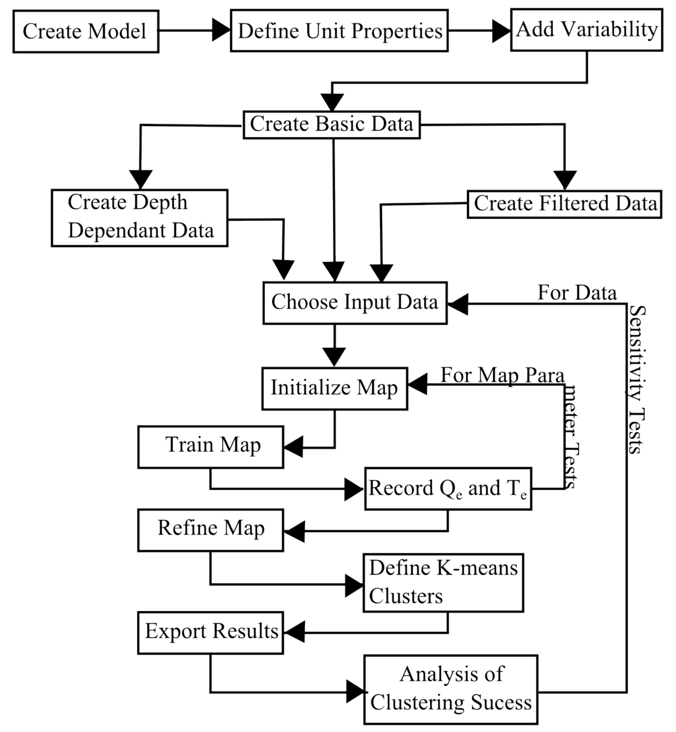

Figure 1: Work flow for this synthetic study showing the progression from model building through analysis of test results.

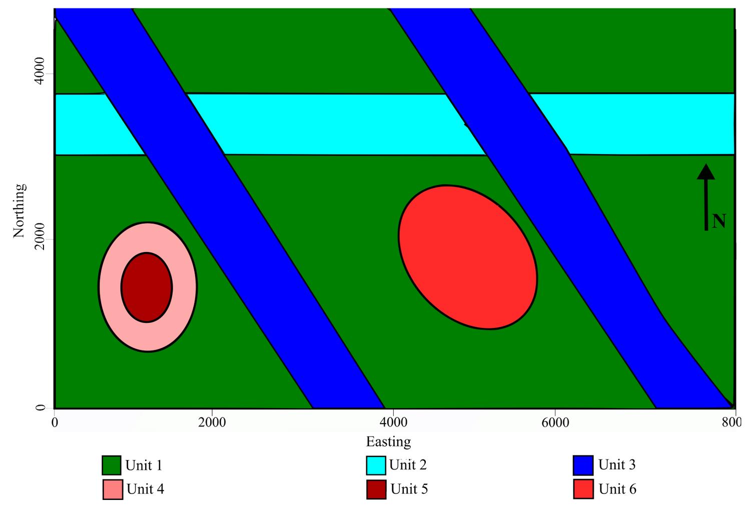

Figure 2: Synthetic model constructed for this project.

Table 1: A summary of the characteristics of each of the units in the model produced for this project.

Unit	Form	Properties	Depth
1	Country Rock	p_1=1 p_2=6 p_3=1 p_4=2 p_5=11	Surface
2	Sill	p_1=2 p_2=5 p_3=1 p_4=5 p_5=11	Shallow
3	Dykes	p_1=3 p_2=4 p_3=2 p_4=7 p_5=10	Surface
4	Alteration Halo	p_1=4 p_2=3 p_3=4 p_4=6 p_5=8	Shallow
5	Intrusion	p_1=5 p_2=2 p_3=7 p_4=4 p_5=5	Deep
6	Intrusion	p_1=6 p_2=1 p_3=11 p_4=1 p_5=1	Surface

Data-type 1, d_1, (Fig. 3b) is analogous to gravity data where p_1 (Fig. 3a) is density. It is sensitive to units of all depths but has poor spatial resolution.
Data-type 2, d_2, (Fig. 3d) is analogous to magnetic data where p_2 (Fig. 3c) is magnetic susceptibility. It is sensitive to shallowly buried units and has moderate spatial resolution.
Data-type 3, d_3, (Fig. 3f) is analogous to the total count from gamma-ray spectrometry where p_3)(Fig. 3e) is the total concentration of K,Th, and U. It is sensitive only to units at the surface and has good spatial resolution.
Data-type 4, d_4, (Fig. 3h) is analogous to seismic traveltime where p_4 (Fig. 3g) is the porosity. It is sensitive to shallow units and has good spatial resolution.
Data-type 5, d_5, (Fig. 3j) is analogous to remote sensing data where p_5 (Fig. 3j) is reflectivity. It is sensitive only to surface units and has good spatial resolution.

Once the basic datasets have been produced the datasets which adhere to the spatial resolution quality of the various datatypes are produced by applying averaging filters. A 3x3 averaging filter was used for those datasets with good spatial resolution, a 5x5 filter was used for those with moderate spatial resolution, and a 9x9 filter was used for those datasets with poor spatial resolution. To create dataset which comply with the depth sensitivity criteria the physical properties over some of the bodies needed to be changed. For data types that could only see surface units the properties for Units 2, 4, and 5 were replaced by those for Unit 1. For data types which could only see shallow units the properteis of Unit 5 was replaced by those of Unit 4.

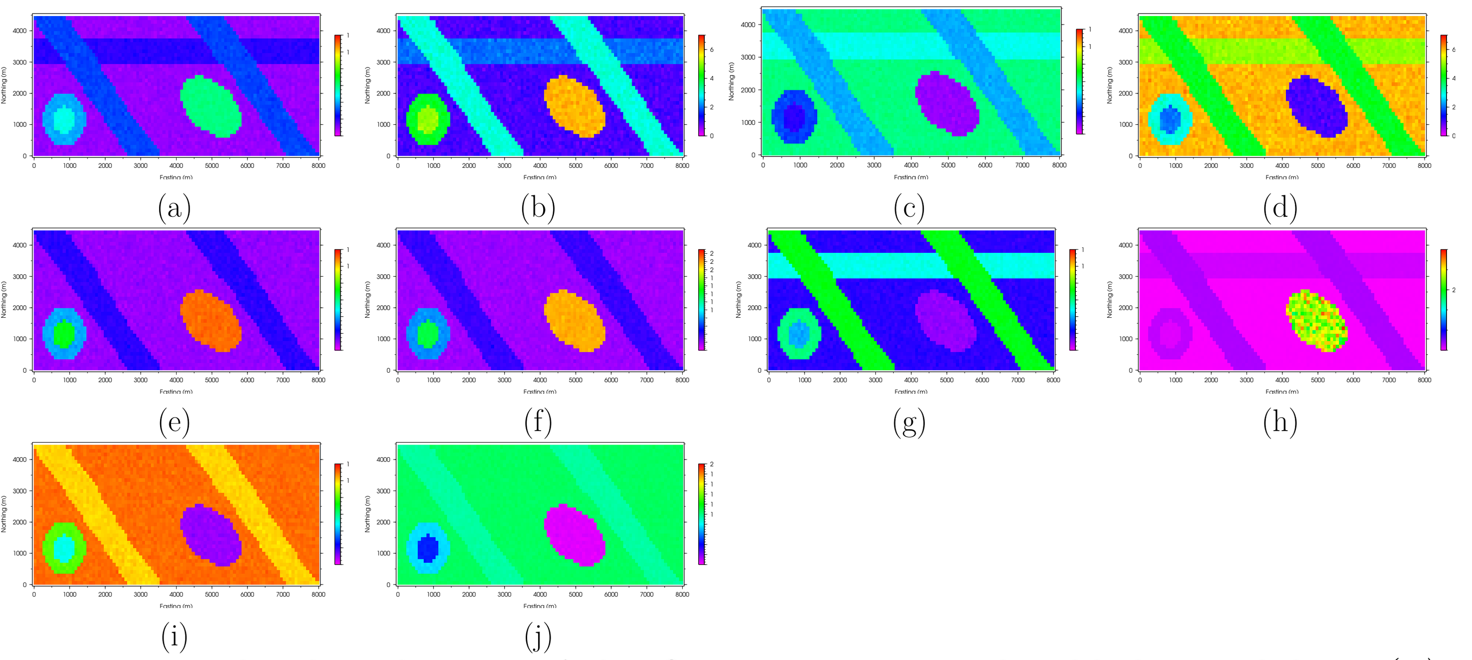

Figure 3: The distribution of the five properties with added variability, (a) p_1, (c) p_2, (e) p_3, (g) p_4, and (i) p_5, and their associated basic datasets (b) d_1, (d) d_2,(f) d_3, (h) d_4, and (j) d_5 which don't include any smoothing or depth sensitivity information.

Mesh Size and Shape Tests

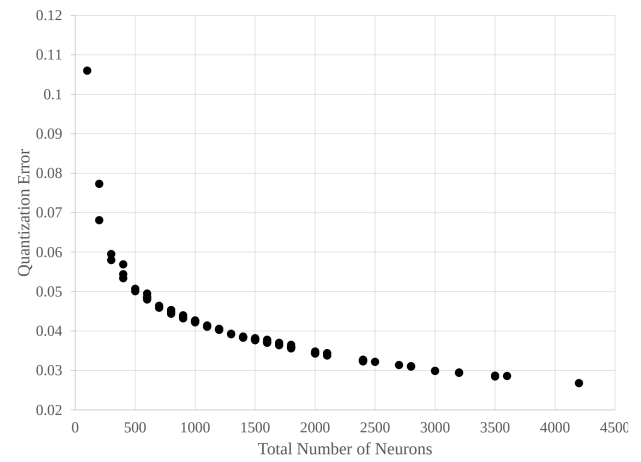

Figure 4: The Q_e values for each of the 81 test training runs plotted with respect to the size of the neural map being trained.

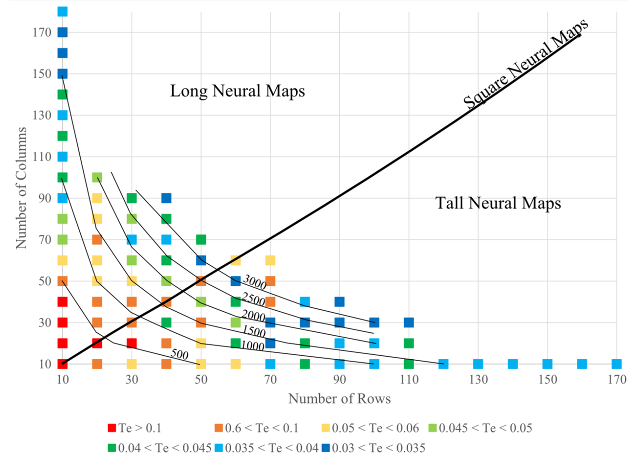

Figure 5: Topographic errors for each of the 81 test training runs plotted with respect to the numbers of rows and columns in their respective neural maps. Thin black lines are contours of maps with 500, 1000, 1500, 2000, 2500, and 3000 cells.

The success of the SOM training process used in the SiroSOM algorithm is quantitatively determined using the Quantization Error (Q_e) and Topographic Error (T_e). Q_e is a measure of how well the data vectors map into the neural map. T_e is a measure of how well structured the neural map is after the training process. Ideally, both Q_e and T_e should be minimized through the training process.

Eighty-six test training runs with 3686 input data were conducted using a range of different mesh sizes and shapes. For each training run all other training parameters were kept the same. The Q_e and T_e values for each trial were recorded.

The results of these tests are summarised in Figures 4 and 5. Figure 4 shows that an increased total number of nodes in the neural mesh is effective in reducing Q_e. Based on this result it is suggested that the minimum number of nodes in a neural map be 25% of the number of input data.

Figure 5 shows that T_e value is minimized both by having an increased number of nodes and by having a rectangular mesh which is taller than it is long (in this case between 70 and 100 rows and between 10 and 50 columns). Perfectly square meshes are to be avoided.

Spatial Resolution and Depth Sensitivity Tests

Sixteen tests were carried out to determine the effect of the different depth sensitivities and spatial resolutions inherent to the data types on the accuracy with which the data points can be clustered by the SOM process. The first test was a baseline using only the basic synthetic data for all five datasets. The next five tests investigated the effect of depth sensitivity but progressively switching from the basic datasets to the depth sensitive datasets (i.e. in test 1 only d_1 was switched, in test 2 d_1 and d_2 were switched ect.). In a second set of five tests the effects of spatial resolution were investigated by switching progressively from the basic datasets to those that included the spatial resolution filtering. The last set of five tests started with the depth sensitive datasets and progressively switched to datasets that included both depth sensitivity and spatial resolution. The SOM process results in each of the data points being assigned to a cluster. Each cluster was paired to a known unit based on the spatial distribution of the data in the cluster; this is now the 'correct' cluster for that unit. The fraction of the data points which clustered correctly was determined based on these assignments (Figure 6).

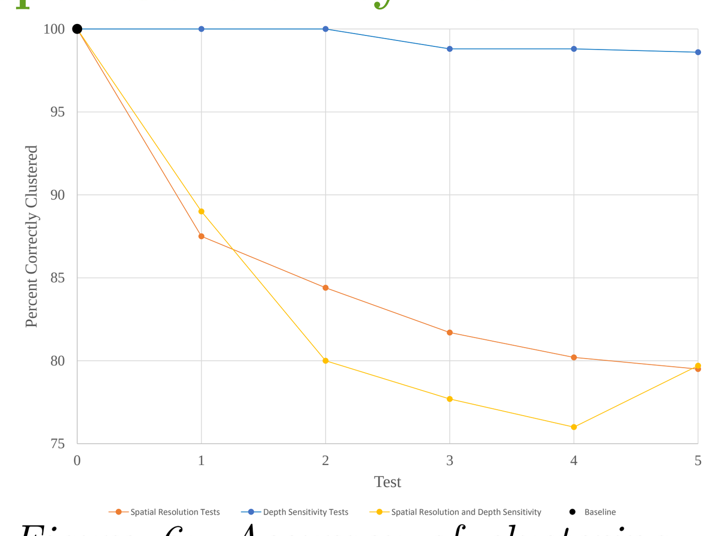

Figure 6: Accuracy of clustering for all data points for each of the 16 tests.

The results of these tests can also be visualised by plotted the data coloured based on their cluster. In Fig. 7 the input data (top five panels) and results of the final test (where all datasets met the spatial resolution and depth sensitivity criteria) are presented showing that much of the mis-clustering is occurring around the edges of the units.

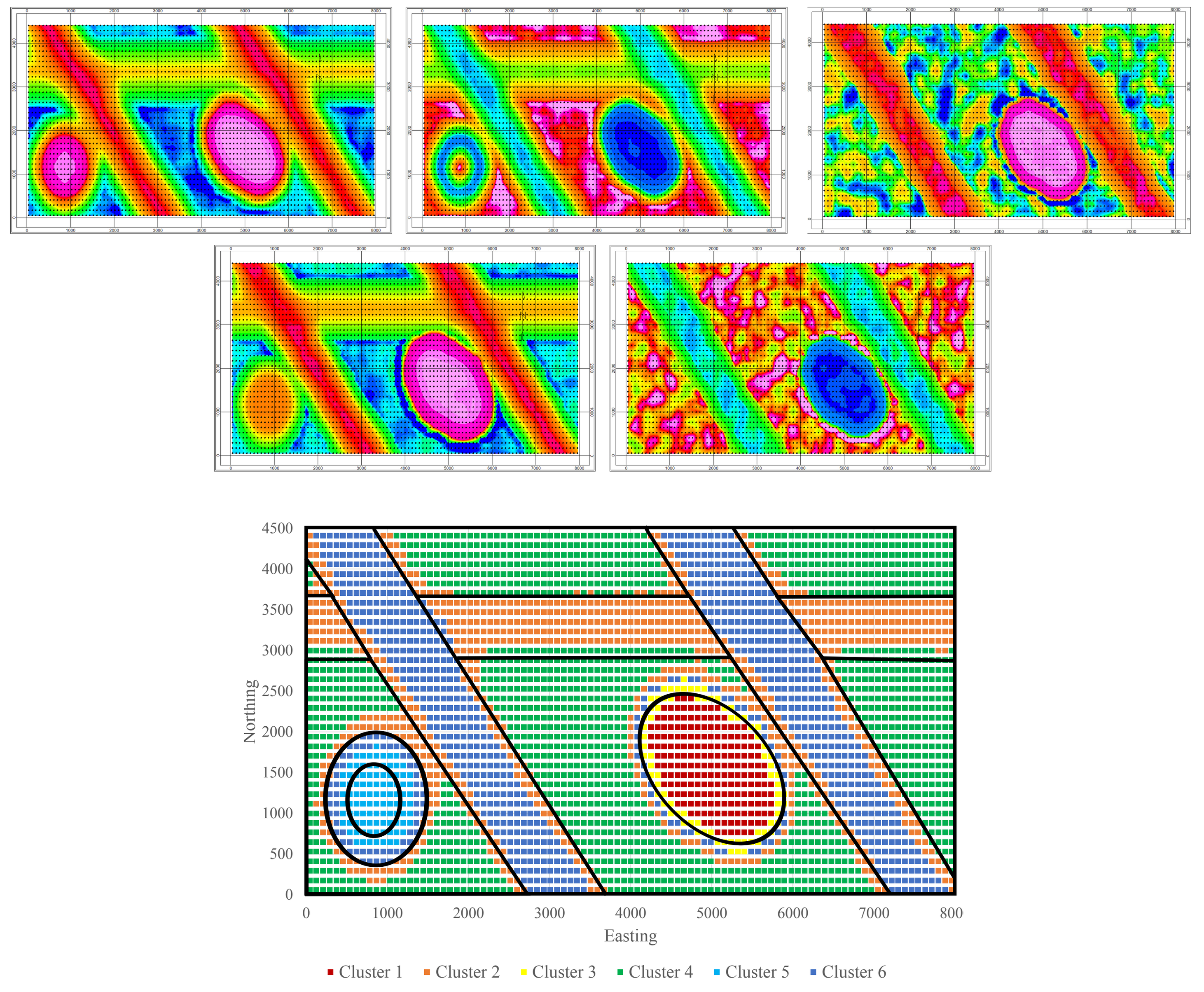

Figure 7: The input data (top five panels), with observation location indicated by black does, and final clustered results (bottom panel) for the final test where all datasets are consistent with their spatial resolution and depth sensitivity characteristics

Conclusion

These tests show that choices made in the initialization prior to training are crucial to obtaining good results for training the SOM. However, the numbers of rows and columns in the neural map need to be chosen carefully. Also, better results will be attained from a rectangular mesh with more rows than columns. The integration of data complexity through addition of depth sensitivity and spatial resolution criteria did influence the success of clustering the data. However, even the most complex scenario attempted led to more than 75% of the data points clustering correctly and was able to replicate the original model rather well. Further work needs to be done to investigate the effects of increased intra-unit property variability before this method can be extended to testing with real data.

References

F. Iwashita, M.J. Friedel, C.R.deS. Filho, and S.J. Fraser, 2011. Using self-organizing maps to analyse high-dimensional geochemistry data across Paranna, Brazil. Proceedings of the 15^{th} Annual Brazilian Symposium on Remote Sensing, April 3-5 2011, Curitiba, Brazil

C.Y. Ji, 2000. Land-use classification of remote sensed data using Kohonen self-organizing feature map neural networks. Photogrammetric Engineering and Remote Sensing 66 (12)

T. Kohonen, 1982. Self-organised formation of topologically correct feature maps. Biological Cybernetics Vol. 43 Pg. 59-69

Acknowledgements

We would like to thank the Nalcor Petroleum Exploration Enhancement Project for funding this research and CSIRO for providing the SiroSOM software at reduced cost. A particularly thank you to Stephen Fraser at CSIRO for his technical help with using the SiroSOM software.