Paper 67

A New Method for Data Integration and Integrated Data Interpretation: Self-Organising Maps

Fraser, S. J. [1], Dickson, B. L. [2]

1. CSIRO Exploration and Mining: QCAT
2. Dickson Research Pty Ltd.

ABSTRACT

Today's explorationists have unprecedented capacity to acquire new and historic geological, geochemical and geophysical data; however, the integrated analysis and interpretation of such data remains a significant challenge. A computational approach based on self-organising maps (SOM) can assist in understanding and synthesizing such data. A SOM analysis can highlight subtle relationships and assist in the process of knowledge creation from complex and disparate data. In a SOM analysis each sample is treated as a vector in a data space determined by its variables; and, measures of vector similarity are used to order and segment the input data into meaningful natural patterns. Because SOM is an exploratory data analysis tool that is unsupervised and data-driven, the resulting patterns, boundaries and relationships are internally-derived.

Typically, one of the primary SOM outputs is a rectilinear map. This map is an orderly two-dimensional representation of the multi-dimensional input data set, which displays relationships between the input samples. More importantly, the map can also be used as a framework to display the variables' contributions related to those samples. By judicious application of some form of colour look-up table, it is possible to assess the spatial coherence or context of samples belonging to a particular node, or group of nodes (in the assigned colour of that node). Furthermore, scatterplots of variable values belonging to nodes (also coloured by the node look-up table) are an effective means of identifying subtle trends and relationships such as identifying and separating overlapping geological processes, related to mineralization and subsequent weathering or metamorphic events.

As our technological capacity to acquire data increases, evidence-based knowledge extraction techniques will become increasingly important. Because of its vector-quantization approach and its ability to analyze, integrate and allow an integrated interpretation of complex and disparate data, SOM is an ideal tool to assist in this process.

INTRODUCTION

One of the major challenges facing explorationists today is how to integrate and meaningfully analyse the vast amounts of data they collect during greenfield and brownfield exploration programs. The potential of these data to provide new information and knowledge is exceptional. Not only are the accuracy, precision and signal-to-noise characteristics of measurements improving, but new technologies are allowing us to measure an increasing diversity of chemical, mineralogical and physical properties, rapidly, reliably and with ever-improving spatial coverage. Along with this improved quality and quantity of data there is an increasing need, especially in brownfield environments, not only to locate mineralization but also to derive and provide geotechnical and even geo-metallurgical information to assist with mine and mill planning.

In this paper we introduce self-organising maps (SOM; Kohonen, 2001) as an analysis technique for understanding subtle relationships within and between disparate data sets, and to provide a means of analysing and interpreting disparate data in a meaningful fashion.

Self-organising map analysis procedures are widely used in fields such as finance, industrial control, speech analysis (Kaski et al., 1998) and astronomy (e.g., Garcia-Berro et al., 2003). The approach is also gaining increasing acceptance in the petroleum industry, where it is used to assist in the calibration and interpretation of well-logs and seismic data (e.g., Strecker and Uden, 2002; Briqueu et al., 2002). Apart from petroleum industry applications, the work of Penn (2005), who looked at the relationships between airborne hyperspectral data and surface geochemistry, and the activities of the current authors, the number of papers using SOM as a data analysis tool in the wider geoscience area is limited.

This paper, which is a generalized overview of the SOM approach, is intended to increase awareness and encourage readers to consider SOM as a data analysis methodology for spatially-located exploration data.

In "Proceedings of Exploration 07: Fifth Decennial International Conference on Mineral Exploration" edited by B. Milkereit, 2007, p. 907-910

THE SELF-ORGANISING MAP APPROACH

Overview

The Self-Organising Map (SOM) is a data analysis, visualization and interpretation tool that is based on the principles of vector quantization and measures of vector similarity. While most SOM procedures can be considered exploratory, the method can be used to perform broad categories of operations such as, function fitting, prediction or estimation, clustering, pattern recognition or noise reduction, and classification.

In a SOM analysis, each sample is treated as an n-dimensional (nD) vector in a data space defined by its variables. This sample vector quantization approach means that both continuous (e.g., geochemical assays, structural orientation data) and categorical variables (e.g., observed characteristics, rock types) can be input, making the SOM technique ideal for the analysis of complex and disparate geoscientific data. Because a SOM is unsupervised, no prior knowledge is required as to the nature, or number, of "groupings" within the data set. These features are why the SOM technique has advantages over other more 'conventional' analysis methods such as clustering (both hard and fuzzy), factor analysis, principal components and traditional neural networks.

The output of a SOM analysis is typically a 2D rectilinear "self-organised map" that is composed of cells (nodes), each of which represents a "node-vector" in the data space defined by the variables.

Training the "Node-Vectors"

Node-vectors are "trained" to represent the original distribution of samples in the data space by the following process. The n-dimensional (nD) data space defined by the input samples is seeded (typically randomly) by a defined number of seed-vectors. The number of seed-vectors is determined by the size of the required output map: for example, a 12x8 sized map means 96 seed-vectors. In an iterative, two-step process that is applied to each input sample many times, these seed-vectors are subsequently trained to represent the structure and patterns of the input samples. In the first step, which is referred to as the competitive step, a given input sample is compared to all seed-vectors within a particular radius of the input sample and ultimately a winning seed-vector is determined as being the most similar. This process is based on a measure of vector similarity (e.g., dot-product, cosine, Euclidean distance, etc); and once the winning seed-vector is found, its properties are modified by a percentage so that its characteristics more closely resemble those of that nearest input sample. In the second step, which is known as the cooperative step, all the seed vectors within a given radius of the winning seed vector are also modified so that their properties are also changed by a percentage to more closely resemble the input sample in question. This procedure is then repeated for the next input sample. By reducing the radius of influence, and changing the percentage modification applied to the seed-vectors during each iteration, the seed-vectors become trained to represent the structure of the original input data (node-vectors). This may involve the running hundreds or thousands of iterations of the above procedures on each input sample.

The Map

Once the seed-vectors have been trained to represent the structure of the input data (now called node-vectors), all the original input samples closest to that node-vector are represented by that node on the 2D map. A regression is used to map from nD space to the 2D rectilinear representation (the map); and a key feature of this mapping is that it preserves the relative relationships (topology) between the node-vectors. That is, node-vectors that are close in nD space have nodes that are close on the 2D map.

The original input samples are now represented by particular nodes on the self organised map, and these may well form a group or cluster. However, if a node is close to other nodes on the "map", those nodes may be a sub-set of a larger group of similar samples formed by all the samples belonging to nearby nodes as well.

The self-organised "map" is an orderly, 2D representation of a complex multi-parameter data set. It is an ideal framework for subsequent visualization and interpretation purposes. The "unified distance matrix" and component plots, are two examples of such visualizations.

The Unified Distance Matrix (U-matrix)

The "unified distance matrix" (U-matrix; Ultsch and Vetter, 1994) representation of the "map" indicates the closeness, between adjacent nodes on the map, typically in terms of Euclidean distance. A colour-temperature scale is used so that cooler colours (blues) separate adjacent nodes that are closer (similarity), and hotter colours indicate larger Euclidean separations (difference). To assist in this display, alternate "dummy" nodes are added to the U-matrix and these are coloured according to the distance between adjacent nodes; whereas the nodes that represent actual vectors are coloured according to the average of the distances to its neighbours. This representation gives rise to a topographic analogy in that there are valleys of (blue) nodes that are similar, separated by walls of higher-temperature coloured nodes that represent class-boundaries or samples belonging to different groupings. Figure 1 (A) shows a typical U-matrix with white hexagons on those nodes that actually represent input samples; the size of the hexagon is proportional to the number of input samples each node represents.

Quantization Error

Another useful SOM parameter that is recorded for each input sample is the quantization error (QER). This error essentially is a measure of the distance a sample is from its node-vector. Samples with high quantization error represent the outliers in a data set. Such samples tend to be anomalous compared to the

majority of the samples. These high QER samples may also represent the edges or boundaries within a data set, which in geology has implications for lithological contacts etc.

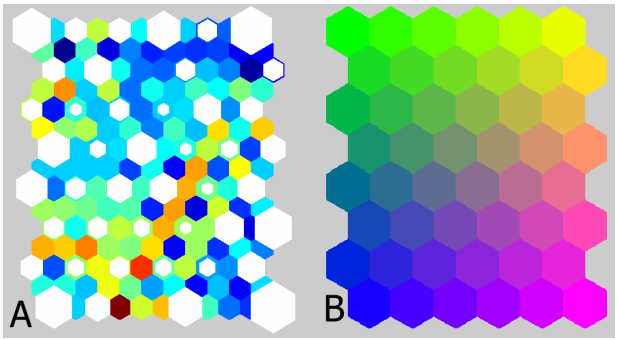

Figure 1: SOM maps for a data set comprising 220 geochemical samples each with 32 variables for samples of igneous rocks from NE Queensland. (A) shows the U-matrix representation with white hexagons sized proportional to the number of samples falling on each node.(B) shows a colour-coding of the nodes which is used in Figure 3.

Component Plots

"Component Plots" (Figure 2) are another visualization of the nodes on the self-organised map. Because each node is a vector in the data space defined by the input variables, it is possible to visualize each node's contribution for a particular variable, and display the values again using a colour-temperature scale so that low values (of the variable in question) are displayed blues and high are in red. It is also possible then to use standard image processing procedures, such as principal components analysis, to determine relationships and trends amongst these images.

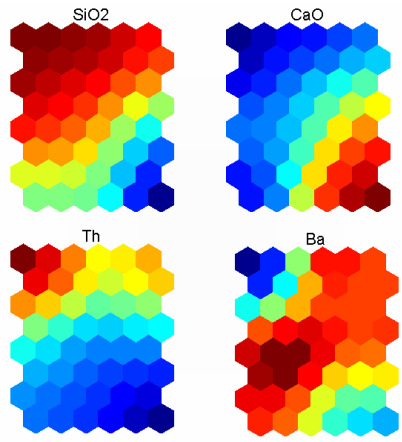

Figure 2: Component plots for SiO2, CaO, Ba and Th. Th and SiO2 behave similarly, Ca is antipathetic to SiO2 and Ba is different again.

Spatial Plots of Samples belonging to Map Nodes

A colour-mapping can be applied to the map, so that each node, or group of nodes, of interest can be coloured uniquely (Figure 1 (B)). Another way of achieving this colour-mapping is to cluster the node vectors using an approach such as K-means.
There are advantages in displaying the spatial distribution of samples belonging to a particular node based on the above colour-mapping. If spatially coherent patterns emerge, there is a high probability that the analysis is producing meaningful results. If there is no spatial coherence to the samples, one would have to attempt to explain the distribution in terms of the properties for the node(s) in question.

Node Scatterplots

Another use of the colour-mapping of the map nodes is to colour nodes in variable scatterplots (Figure 3). Because the SOM is a segmentation technique, scatterplots can show both linear and non-linear relationships and trends which are difficult to observe in the input data.

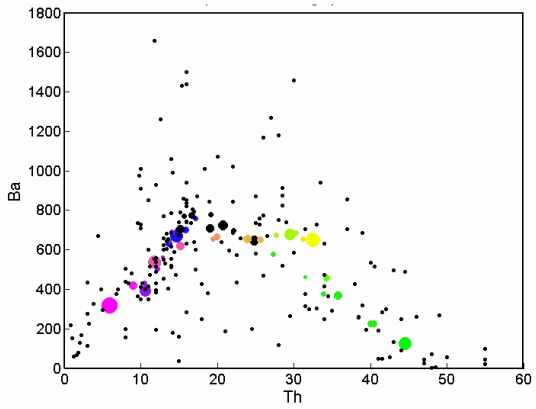

Figure 3: Scatterplot of Th versus Ba . Black dots represent original samples; coloured circles are plots of the SOM nodes. The different trends in Ba and Th concentrations due to fractional crystallization are apparent from the SOM node data.

Strategies for Including Spatial and Categorical Data into a SOM Analysis

Spatial information, and or labels from a categorical variable, may be included as input into a SOM analysis. However, one must be aware of the consequences of such inclusion. For example, if one is hoping to see spatial patterns related to geochemical zoning, it may be better not to include the spatial co-ordinates in the actual SOM analysis, because if you include the coordinates, you will impose an ordering on the samples based on location. Instead it is often better not to include the spatial information in the analysis, but then to examine spatial plots of the SOM-coloured data to see if coherent spatial patterns exist. If coherent spatial patterns are found, or if particular outputs correlate with particular labels, there is added

weight to the proposal that relationships based on location, or label information exist.

DISCUSSION & CONCLUSION

SOM is a useful tool for the analysis of complex, disparate and spatially-located exploration data. We have used SOM to address a wide range of issues and problems; and the following are some examples of its application. Dickson and Taylor (2003) used SOM as a noise reduction method for aerial gamma-ray data. Sliwa et al. (2003) used SOM to assist in the interpretation of lithologies from a suite of geophysical borehole logs; Zhou et al. (2005) estimated rock strength from geophysical borehole logs; Fraser et al. (2005) and Fraser and Dickson (2005) reported geochemical applications; and, Fraser et al. (2006) used SOM to investigate mine geotechnical data.

Technology has improved the ease and hastened the speed with which data can be collected and stored. SOM is a data mining technique that has the capacity to improve the effectiveness and efficiency of explorationists as they seek to discover subtle clues within data sets that may be associated with mineralization or other geological processes.

ACKNOWLEDGEMENT

We thank Dr P. Blevin NSW DPI for permission to use the igneous rock chemistry used to illustrate this paper.

REFERENCES

Briqueu, L, Gottlib-Zeh, S, Ramadan, M and Brulhet, J, 2002. Traitement des disgraphies a l'aide d'un reseau de neurons du type <carte auto-organizatrice>: application a l'etude lithogique de la couche silteuse de Marcoule (Gard France), C R Geoscience, 334 (2002): 31-337.

Dickson, B.L. and G. M. Taylor, G.M., 2003, Quietening the noise: an evaluation of noise reduction methods applied to aerial gamma-ray survey data. Exploration Geophysics (2003) 34, 97–102.

Fraser, S. J. and Dickson, B L, 2005, Ordered vector quantization for the integrated analysis of geochemical and geoscientific data sets, Paper presented to IGES 2005 – 22nd International Geochemical Exploration Symposium incorporating the 1st International Applied Geochemistry Symposium, Perth, 19 - 23 September.

Fraser, S., Dickson, B., Kowalczyk, P. and Sparks, G., 2005, And now for "SOM" thing completely different: Spatial Data Mining. Program with Abstracts, Geological Society of Nevada Symposium 2005, Reno/Sparks, Nevada, May, 2005, pp. 45.

Fraser, S. J., Mikula, P.A., Lee, M. F., Dickson, B.L. and Kinnersly, E., 2006, Data Mining Mining Data – Ordered Vector Quantisation and Examples of its Application to Mine Geotechnical Data Sets. In, Dominy S. (Editor): Sixth International Mining Geology Conference, "Rising to the Challenge", August 21-23, 2006 Darwin. AusIMM. Publication Series No 6/2006, pp 259-268. (CSIRO Exploration & Mining Report No P2005/207

Garcia-Berro, E, Santiago Torres, S. and Isern, J., 2003, Using self-organizing maps to identify potential halo white dwarfs, Neural Networks, 16(2003):405-410.

Kaski, S., Kangas, J., and Kohonen, T., 1998: Bibliography of self-organizing map (SOM) papers: 1981-1997, Neural Computing Surveys, 1: 102-350 [online]. Available from: <http://www.cse.ucsc.edu/~jagota/NCS/vol1.html>

Kohonen, T., 2001, Self-Organizing Maps, third extended edition, Springer Series in Information Sciences, Vol 30 (Springer: Berlin, Heidelberg, New York).

Penn, B.S., 2005, Using Self-Organizing maps to visualize high-dimensional data. Computers and Geosciences 31, 531-544.

Sliwa, R., Fraser, S. J. and Dickson, B. L., 2003, Application of self-organising maps to the recognition of specific lithologies from borehole geophysics, in Proceedings 35th Sydney Basin Symposium: Advances in the Study of the Sydney Basin (eds: A C Hutton, B G Jones, P. F. Carr, B. Ackermann and A. D. Switzer), pp 105-113 (University of Wollongong).

Strecker, U. and Uden, R., 2002, Data mining of poststack seismic attribute volumes using Kohonen self-organizing maps, The Leading Edge, October:1032-1036.

Ultsch, A. and Vetter C., 1994, Self-organising Feature Maps versus Statistical Clustering A Benchmark, Technical Report No. 9, Dept. of Mathematics and Computer Science, University of Marburg, Germany, (1994)

Zhou, B., Fraser, S., Borsaru, M., Aizawa, T., Sliwa, R. and Hashimoto, T., 2005, New approaches for rock strength estimation from geophysical logs, in Proceedings Bowen Basin Symposium 2005: The Future for Coal — Fuel for Thought (ed: J. W. Beeston), pp 151-164 (Geological Society of Australia Inc, Coal Geology Group and the Bowen Basin Geologists Group)