Data Mining Geoscientific Data Sets Using Self Organizing Maps

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

(1) CSIRO Exploration & Mining , QCAT PO Box 883 Kenmore 4069, Australia, Stephen.Fraser@csiro.au
(2) Dickson Research Pty Ltd, 47 Amiens St, Gladesville, 2111, Bruce.Dickson@optusnet.com.au

Abstract

Geoscientists are increasingly challenged by the joint interpretation of ever-expanding amounts of new and historic, spatially-located exploration data (e.g., geochemistry, geophysics, geology, mineralogy, elevation data, etc.). And because, we can gather data faster than it can be interpreted, the availability of geographic information systems (GIS) has, to some extent, compounded, rather than reduced this problem. Research into the analysis and interpretation methods for data held in a GIS is in its infancy. A limited number of "advanced" interpretation methods have been developed; however, these often rely on *a priori* knowledge, training, or assumptions about mineralisation models. Objective, unsupervised methods for the spatial analysis of disparate data sets are needed.

We have investigated and developed a new computational "tool" to assist in the interpretation of spatially located mineral exploration data sets. Our procedures are based on the data-ordering and visualization capabilities of the Self Organizing Map (SOM), combined with interactive software to investigate and display the spatial context of the derived SOM "clusters". These computational procedures have the capacity to improve the efficiency and effectiveness of geoscientists as they attempt to discover and understand the often subtle signals associated with specific geological processes (e.g., mineralization), and separate them from the effects of overprinting noise caused by other processes such as metamorphism or weathering.

Based on the principles of "ordered vector quantization", the SOM approach has the advantage that all input data samples are represented as vectors in a data-space defined by the number of observations (variables) for each sample. The SOM procedure is an exploratory data analysis technique whereby patterns and relationships within a database are internally derived (unsupervised) based on measures of vector similarity (e.g., Euclidean distance and the dot product). The outputs of a SOM analysis are highly visual, which assists the analyst in understanding the data's internal relationships.

Keywords: Self Organizing Maps, Data Mining, Geosciences

Introduction

Geoscientists in general and explorationists in particular, commonly suffer data overload. Volumes of open-file reports, digital geological maps, geophysical data sets, and remotely sensed data are typically available. When these data are combined with the results of current exploration activities, serious data-overload problems can occur.

Geographic Information Systems (GIS) and their capacity to store spatially located (digital) data have not necessarily assisted in the interpretation of data. More often than not, the incorporation of data into a GIS is seen as the 'goal', whereas in reality, it is only the first step in the data analysis and interpretation procedure. GIS are important in that they allow spatially located exploration data to be stored in a database (ideally with checks as to the data's validity and integrity). However, the mechanisms for data interpretation in such systems, have not kept pace with the enthusiasm with which data can be collected and stored.

Traditional multivariate statistical approaches are often confused by data sets with variable relationships that are non-linear, by data distributions that are non-normal (typically with multiple populations), and by the data sets themselves that may be disparate, sparsely-filled, (contain "nulls"), with both continuous and discontinuous numeric data and text. The SOM, ordered vector-quantization approach can overcome many of these problematic issues.

A number of "advanced" interpretation methods have been developed for the GIS environment, such as "Weights of Evidence" (see, Bonham–Carter and Agterberg, 1999), "Neural Networks" (see Brown *et. al*, 2000) and other "Expert Systems". These "advanced" methods often rely on *a priori* knowledge, training, or a subjective approach (assumptions about mineralization models, and the probabilities as to the significance of particular occurrences or features), which may or may not exist, be relevant, or available. There are very few techniques that enable a user to explore and analyze the relationships between the various data-layers stored in a GIS in an objective quantitative fashion.

The authors have an ongoing interest in the development of tools and techniques to assist in the integrated analysis, interpretation and visualization of various exploration and mining related data sets, especially those with spatial or geographic attributes. For some time, they have been promoting the use of Self Organizing Map (SOM - Kohonen, 2001) as a knowledge discovery or exploratory, data analysis tool.

The Self Organizing Map procedures are described in detail elsewhere (Kohonen, 2001). Briefly, however, if one represents all sample points as vectors in a data-space defined by the number of observations, the SOM procedure provides a non-parametric mapping (regression) that transforms an n-dimensional representation of these high dimensional, nonlinearly-related data items to a typically two-dimensional representation, in a fashion that provides both an un-supervised clustering and a highly visual representation of the data's relationships. SOM procedures are used in a range of applications, but they are having a major impact in the fields of data exploration (Kaski, 1997) and data mining (Vesanto, 2000).

The SOM procedures being developed by the authors are aimed at providing geoscientists with access to new methods for determining the intricate relationships, within and between multiple, spatially-located and complex data sets. The analysis and visualization provided via SOM has the potential to be significant to both spatial and non-spatial investigations involving both resource discovery and utilization.

Except for some specific applications, SOM procedures have not been widely accepted in the geosciences nor used by the exploration and mining industries. Because of the relatively recent development of the technique, there is much to learn about the potential and application of SOM for analyzing resource related data sets. SOM has been widely used for data analysis in the fields of finance, speech analysis, astronomy (see Kaski *et. al*, 1998; Garcia-Berro *et al*., 2003) and more recently in petroleum well log and seismic interpretation (Strecker & Uden, 2002; Briqueu *et al*., 2002) and geochemisrtry and hyperspectral data (Penn, 2005).

The SOM technique has characteristics and capabilities that make it ideal for geoscience applications, including:

- An ability to identify and define subtle relationships within and between diverse data, such as continuous (e.g., geophysical logs) and categorical (e.g., rock-type) variables;
- No required prior knowledge about the nature or number of clusters within the data (unsupervised);
- No assumptions about statistical distributions of variables or linear correlations between variables;
- Robust handling of missing and noisy data;
- Additional analysis tools, such as component analysis, spatial analysis and the ability to use a pre-computed SOM as a classification framework for a new dataset.

Results

Three examples using our SOM approach for the analysis of geoscience data sets shall be presented.

The first study used SOM to perform an analysis on some 40,000 located geochemical samples from drill-holes around a known copper-gold deposit. Each sample was assayed for up to 13 elements; however 60% of the variable cells were nulls. A consequence of the data being collected over a 10 year period, with different element suites being used as the paragenetic model for the deposit evolved. Three main gold populations were highlighted using the SOM procedure within the data set. The first we propose relates to transported particulate gold within overlying Mesozoic sediments, the second to hydromorphically transported gold that is being moved into the overlying sediments; and the third relates to gold at the interface/ unconformity between the overlying sediments and the basement lithologies. The SOM procedure was also able to highlight three spatial groupings of anomalous gold values. One was considered to be extensions to the known mine mineralization; the second related to a known prospect some four kilometers away from the mine; while the third occurs some 25km away in a scout drill hole that was part of a regional grid, drilled during regional evaluation. The geochemical samples from this third region, were not assayed for the same element suite for the holes around the known mine, but were based on an earlier, superseded model for mineralization in the area. The SOM procedure however, was able to assign those samples to similar groupings of samples around the mine, despite the fact that key elements were not assayed for in those samples.

In the second study over another Au-prospect, geochemical assay measurements were supplemented by a geologist's logged alteration descriptions. In this case the alteration descriptions were used as labels and not actually included in the SOM analysis. Two distinct high Au associations were delimited by the SOM analysis of some ten elements for each sample. One Au-association was related to high Ag values; the other Au association was related to only moderate Ag values. These two Au populations when plotted spatially form coherent spatial patterns. On a scatter plot of Au and Ag, values coloured by their SOM-assigned groupings, a distinct trend could be observed that we believe indicates the "process of mineralization". This information can be used on spatial plots as a "vector-to-ore". When the alteration

labels were overlain onto points on the scatter plot, there is a general trend evident from poorly mineralized propylitic samples through to highly mineralized samples exhibiting silica flooding; though not all samples logged as "silica-flooding" were highly mineralized.

Our third study involves the use of SOM to assist in the analysis of hyperspectral reflectance data acquired by the HyLogger core–logging system on coal cores. The SOM procedure was applied to approximately 40,000 spectra, each with 522 channels of spectral values, to find natural "groupings" within these data that could be related to facies within the sediments and layering within the coal. In this case the SOM was used to simplify a very complex data set by "clumping" the data into meaningful packages that could be related to the geology by the domain analyst.

Discussion

In each study, the SOM procedure provided fundamental new knowledge, or assisted in simplifying the complexity present within these data, to assist in their analysis and interpretation. In both the first and second examples, the SOM visualizations alerted the analyst to evidence of geological processes present in the data, which assisted in their interpretation. In the third example, the SOM was used to simplify a complex data set into patterns that could be related to the coal sequence sedimentary packages. These capabilities are valuable contributions towards the analysis of geoscientific data sets, which are further enhanced by an ability to display the SOM outputs in their spatial context.

The SOM procedure is an exploratory data analysis technique that derives the patterns and relationships within a data set in an unsupervised fashion based on measures of vector similarity (e.g., Euclidean distance and the dot product). The outputs of a SOM analysis are highly visual, which assists the analyst in understanding the data's internal relationships, and relating them to geological processes.

References

G.F. Bonham–Carter and F.P Agterberg: Arc-WofE: a GIS tool for statistical integration of mineral exploration datasets. pp. 497–500, Proceedings International Statistical Institute, Helsinki, August 11–16, 1999.

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

W.M. Brown, T.D. Gedeon, D.I. Groves, and R.G. Barnes; Artificial neural networks: a new method for mineral prospectivity mapping. Australian Journal of Earth Sciences; (2000) 47, 757-770, 2000.

B.L. Dickson, D.A. Clark D.A., and S.J.Fraser: New techniques for interpretation of aerial gamma-ray surveys. Final Report Project P491. CSIRO Exploration and Mining Report 653R, 20 pages, includes a CD ROM. 1999.

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

S. Kaski: Data exploration using self-organizing maps; Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82, Espoo 1997, 57 pp. Published by the Finnish Academy of Technology, 1997.

S. Kaski, J. Kangas, and T. Kohonen: Bibliography of Self-Organizing Map (SOM) Papers: 1981--1997, Neural Computing Surveys, 1: 102-350. Available from http://www.icsi.berkeley.edu/~jagota/NCS/. 1998.

T. Kohonen: Self-Organizing Maps. Third Extended Edition, Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 2001.

T. Kohonen: Self Organized Formation of Topological Correct Feature Maps. Biol. Cyberbetics. Vol 43, 1982, pp.59-96, 1982.

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

U. Strecker, and R. Uden: Data mining of poststack seismic attribute volumes using Kohonen self-organizing maps. The Leading Edge, October 2002, pp1032 -1037. 2002