SELF ORGANISING MAPS - A CASE STUDY OF BROKEN HILL

Luke Smith
Macquarie University
18 Wyreema Road Warnervale, NSW AUS
luke.smith58@outlook.com

Tasman Gillfeather-Clark*
Macquarie University
11 Telopea St Booker Bay, NSW
tasman.gc@gmail.com

SUMMARY

The principle aim of the research was to overcome the challenges faced by modern geophysical data analysts, particularly those working with large multivariate datasets using Self Organising Maps (SOM). SOM is an unsupervised learning technique for multivariate data, which works by taking multiple geophysical datasets for an area of interest, and integrating them to illustrate trends. Once developed, our method drastically lowered the time required for an analyst to examine and identify trends and relations across a broad range of geophysical, geochemical and other data layers. It also revealed hidden relations and distinct populations within correlated layers.

Our study shows that SOM continues to be a powerful tool in accelerating the interpretation process. This includes the separation of features into distinct geological units, even without any preliminary map inputs to the SOM process. It also highlights SOM's ability to highlight variation in cover, which has been identified as a key aspect moving forward in Australia's mining future, when considering the vast expanses of Australia covered in sub cropping rock. In the future as data continue to grow and overlap, SOM will play an important role in highlighting these relations in soil cover and outcrop geology.

Key words: Self Organising Maps, Broken Hill, Geophysics, Satellite Data, Remote Sensing

INTRODUCTION

As we enter an age of information rich geoscience datasets – especially large scale mining and environmental programs which accumulate a diverse and expansive range of data, there is an increasing need for automated processing. The overwhelming amount of information enables a comprehensive and detailed analysis in the hands of a large team of skilled analysts, however it is difficult for geoscientists to comprehend relations across so many varied layers. Self Organising Maps automate the discovery and characterisation of such relations.

It should come as no surprise that computers are much better at trend analysis than a typical geoscientist. As an example, people tend to focus their attention on anomalous values, (highs and lows), and may miss correlations of middle range values. Analysts are also only able to consider a finite number of variables at once, whereas a computer's ability to process any number of variables is limited only by the capacity of its components. The end product of this analysis is a spatial map of an area, indicating correlations between all input data. The user of this SOM workflow need only supply the initial data, which may be enhanced through the creation of secondary products such as drainage models or feature identification. The analyst is still required for the identification of the significance of these trends, SOM simply makes them apparent

SOM is by no means a recent innovation, having been used for many different goals; from cluster analysis in data mining (Fraser *et al.*, 2006), being used to overcome gaps in datasets via the production of 'fuzzy' observations (Wang, 2003), to the analysis of ecological communities for exploring the ordination of a species and providing a visualisation of that species' abundance (Giraudel and Lek, 2001). It has even been applied recently to the exploration of rare earth elements (Sarparandeh and Hezarkhani, 2016). However, the later of these works used a dataset of <120 data points that were irregularly sampled, to cluster geochemical datasets and thus determine zones of deposition over an area ~3km^2. This shows one example of the potential of SOM, in comparison, our dataset uses more than 640,800 sampled points over an area greater than 350 km^2 The size of our sampled region allows for a detailed examination of SOM's ability in detecting and locating correlations between varied inputs.

This paper is based on the SOM technique described in (Vesanto, Himberg and Alhoniemi, 1999) which has been used extensively since then (Dickson, 1995; Gulson *et al.*, 2007). The work outlined and provided users with a set of premade tools that allow for the SOM process to operate on any multivariate dataset that could be loaded into it. Our project aimed to establish a simple workflow, which would integrate well with existing exploration industry software. The goal was to transition data from industry standard sources, through the SOM process in MATLAB, and then back out into a usable format for industry end users.

The result of our project is a success, in that we have established a pathway from industry standard software, into the SOM process and back into a spatial map highlighting correlations between the original inputs. The resulting map provides interesting insights and perhaps most impressively they manage to delineate human mapped geological boundaries, whilst also highlighting errors in existing maps and identifying variation in cover.

METHOD

Our dataset was provided as part of the Frank Arnott Award by the Department of Primary Industries NSW. The relative wealth of data in this area is in part due to the Broken Hill Exploration Initiative (BHEI). The BHEI was a collaborative project conducted by Geoscience Australia, the New South Wales Department of Primary Industries and the Department for Manufacturing, Innovation, Trade, Resources and Energy. As such the dataset represents a great subject for our SOM process due to the wealth of overlapping datasets.

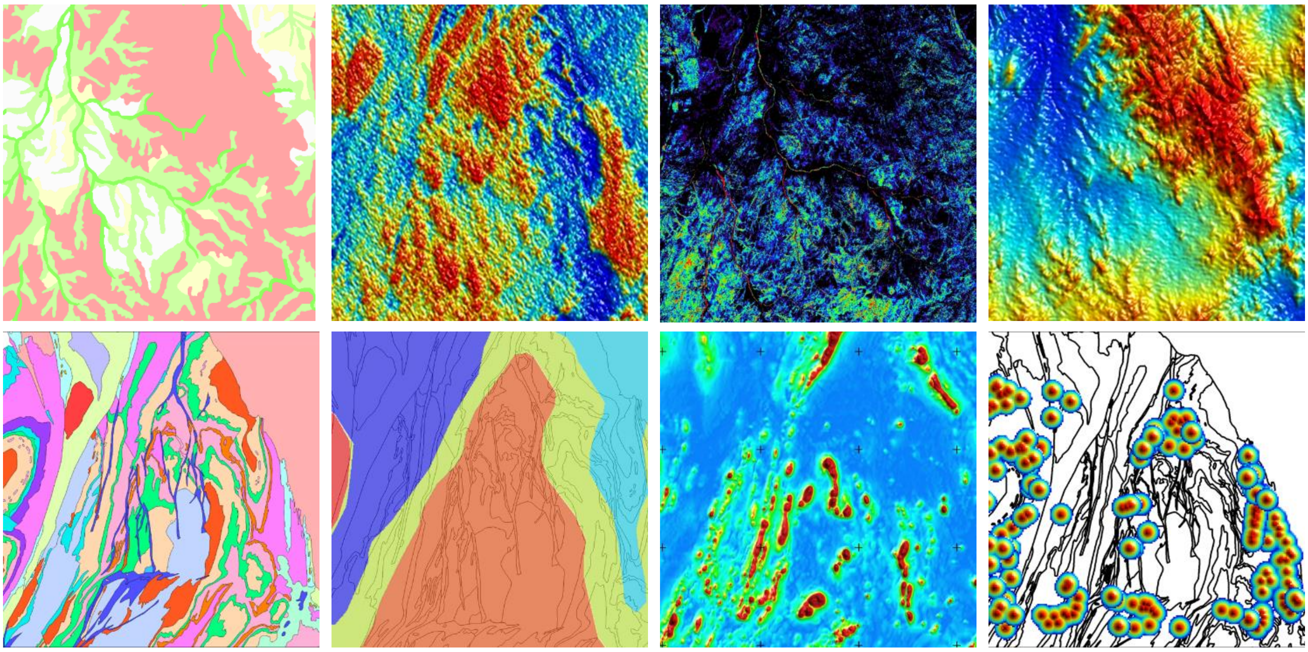


Figure 1 – Examples of the data grids which were developed to be used in the SOM process. From top left to bottom right: Regolith map, Radiometrics (K, U and Th were all used) Hymapper data, DTM, Geology – 100k, Density map, Magnetic Analytical Signal, Mineral occurrence heat maps. These represent only a subset of the data used, for example Radiometrics and Hymapper both provided multiple channels of data.

Due to constraints on computational resources, continual data coverage and time limitation , we selected a subset of the Broken Hill dataset. The limiting factor in computer resources was the data export, in that a satisfactory resolution over the entirety of Broken Hill would be difficult to achieve. On initial examination we were somewhat overwhelmed by the wealth of data. Some of the most intriguing was the Hyperspectral mapping, as it was very high resolution and links well with topography and surface geology. We also learnt several layers would be enhanced by initial processing, for example the gravity map of our subset was developed into a density map, and our TMI map was filtered to become the Analytical Signal a useful proxy for magnetic susceptibility. As an example of discontinuous mapping that can be integrated into the SOM, we processed a list of major mineral occurrences across the region into a spatial distribution map, as shown in Figure 1. Each target has a gaussian probability surrounding the noted location of the measurement, falling to 0 at a set distance.

Figure 2 - Example showing point cloud in a subset of our data produced after using the point sampling tool in Discover. At the figure resolution, individual points are shown as clusters.

These data were loaded as overlapping layers in Discover, and using the point sampling tool, a grid of points was created. Each point in Figure 2 has a long series of data associated with it. Effectively each point comes with two grid co-ordinates (X, Y), as well as 25 different other values (Mag AS, Density, K-Count…etc). This data set is saved as a delimited list and then loaded into the SOM toolbox for Matlab 5. These values aren’t restricted to be being numerical, they can be any standard data type, or Null values. The presence of a null or missing data doesn’t have any weight on the SOM process and is thus ignored. Null values typically occur at map edges or other import conditions.

Each variable or column of the CSV is then normalised, allowing for the varied datasets to be compared with one another. Without weighting each dataset, a profile that contains larger numbers would simply overwhelm the others during SOM computation. The operation of the SOM process is described in (Vesanto, Himberg and Alhoniemi, 1999), and a more practical and simplistic explanation can also be found in Dickson, 1995. The computing demand increases with size, number of variables and resolution, yet finished in seconds on standard laptop.

Understanding SOM
When it comes to interpreting the resulting SOM outputs, it is important to understand the meaning of the colour scale. The main methods for visualizing a SOM involve component planes, distortion patterns and clustering (Agarwal and Skupin, 2008). As the SOM operates, it trains data into best matching units, or BMUs. An over simplified example is given in figure 3. On the left we see 6 variables with their normalised values represented as shades of yellow (high) and blue (low). The more important feature of this component map is the distribution of the colours. Each component map should be visualised as overlying the others, as each hexagon relates to the other components corresponding hexagons. As shown in Figure 3, the top left hexagon of each component map is attributed the Cyan colour in the colour map on the right.

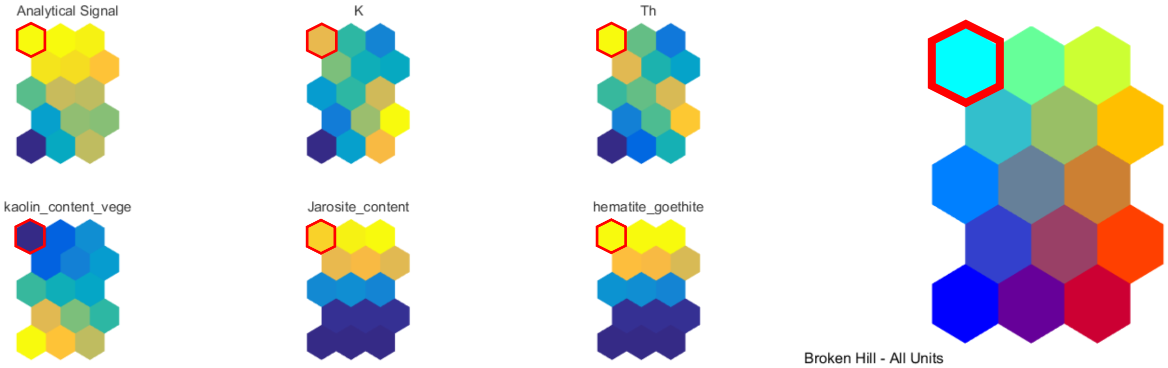
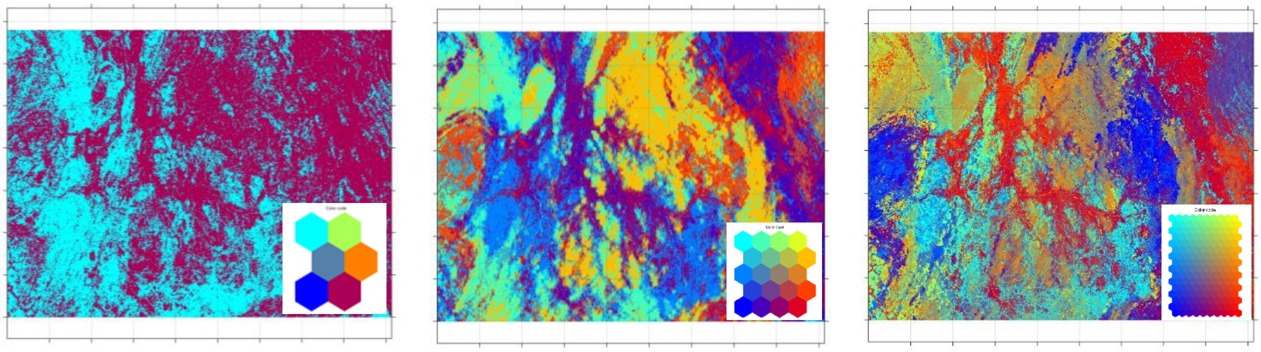


Figure 3 – This is a simplified explanation of how best matching units (BMUs) work with each hexagon representing a note. The hexagons outlined in red represent a node, with all the associated values noted in the variable fields to the left..

The highlighted node has:

- High Magnetic Analytical Signal response
- Moderately high Potassium count
- High Thorium count
- Low Kaolin content
- Moderate Jarosite content
- High Hematite/Goethite content

All that information is now associated with the unit colour cyan in the far right BMU key, in the top left position. Thus, if you were looking a SOM map and saw cyan you would know all that information about a given unit, as well as the spatial distribution of that group. This process repeats for every other BMU which is mapped. The number of nodes is a user choice with lower numbers resulting in a simpler map. The higher the number of nodes, the less distinct the groups as colours blur together. Thus the appropriate number of BMU's will vary based on the data quality, the geometry of the area being mapped as well as the purpose of the map. The map we finally produced which will be discussed in the results section uses a total of 250 BMU's.

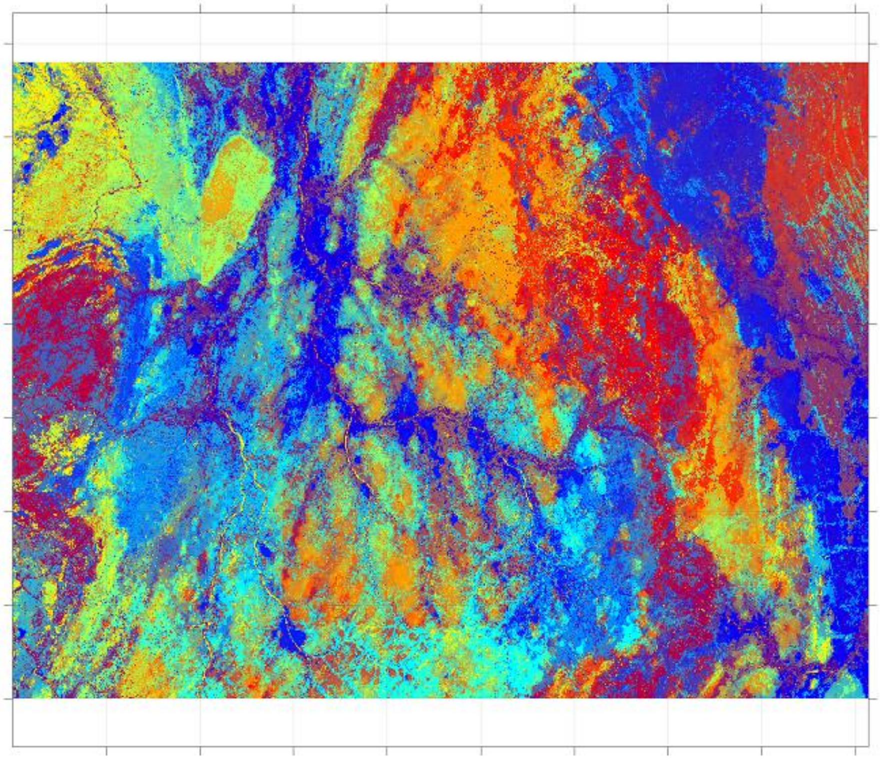


5 BMU 20 BMU 155 BMU

Figure 4 - Comparison of the same area which has been through the SOM process with different numbers of BMU. Note the changing complexity of the maps with different numbers of BMUs. This may be useful if you are looking for more simplistic changes such as cover compared with more complex changes such as geology etc. They key is provided in the bottom right of each map to give an approximate visualisation of the various nodes. This would be used in conjunction to with the component plane key to characterise the different BMUs.

RESULTS

Area Overview



The map in figure 5 is the final output of the SOM processing remapped into the spatial domain. It is important to acknowledge at this point that no human generated maps were used in the production of the grouping of the BMU's displayed here.. In short, these colours map units by identifying correlations..

In the following section, we complete a detailed analysis of two key features the SOM area highlights using currently existing geological map boundaries overlaid on our SOM map. These geological maps were produced using traditional field mapping techniques.

Figure 5 – This is the resultant map which arises from the SOM process The SOM toolbox creates a SOM domain representation of how component clusters correlate. It does not give any spatial information. Thus the grouping of the colours is linked only to 'true correlation'.

Features in the Adelaidean

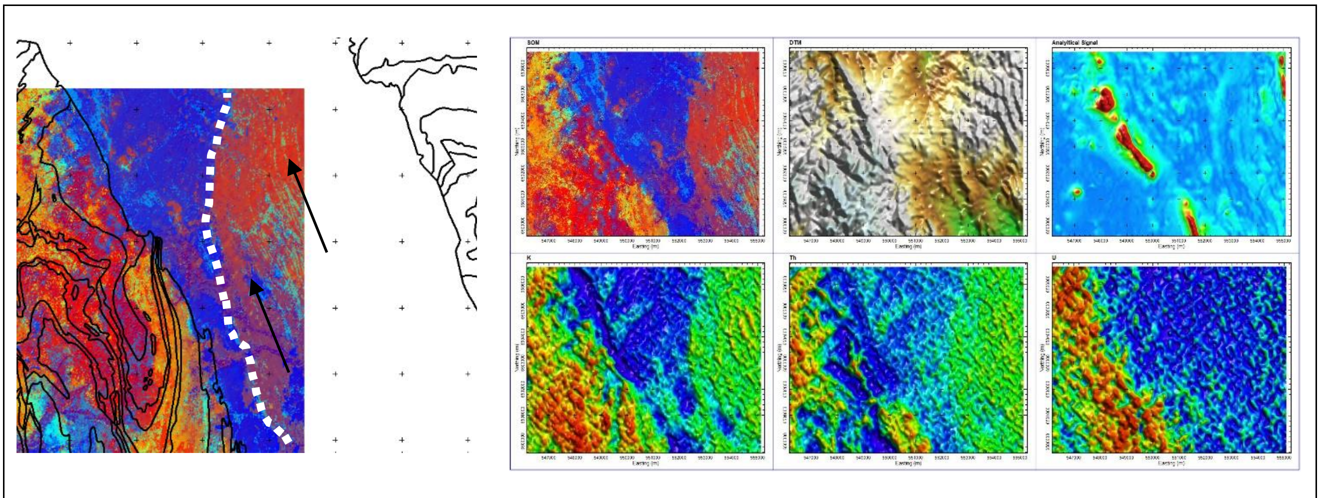


Figure 6 - SOM map overlayed with existing Geological boundaries in the rocks of Adelaidean age which form cover, with data from the same area.

One of the notable features our SOM detects is a missing boundary in one of the map scales which we were provided with. This boundary (marked in dashed white) exists and has been mapped in the 250k scale but just hadn't made the transition to the 100k scale we were using. The Thackaringa group to the North East and the Himalaya formation to the South West form the bulk of the rocks in the area. The North-Eastern unit is un-mapped but contains significant heterogeneities, which may warrant investigation.

Of note are the NNW striking aqua features (marked with arrows), which based on the SOM visualisation are rich in Hematite and Goethite and Ferric Oxide and correlate with the Analytical Signal.

Silver King Formation

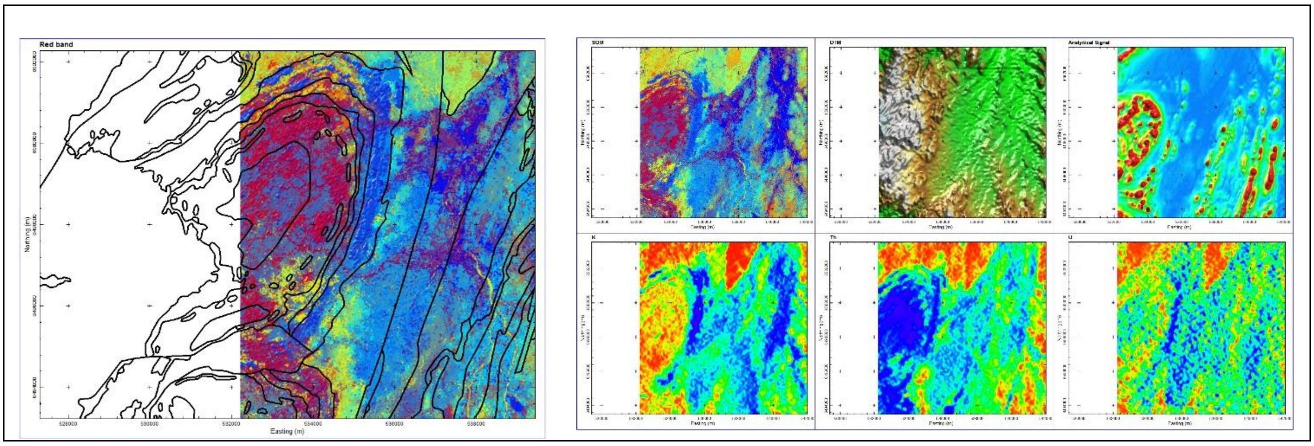


Figure 7 - SOM map overlayed with existing Geological boundaries and the area surrounding the Silver King formation, with data correlation to the same area.

The Silver King formation is part of the Broken Hill Group, appearing in dark blue as an arch in the west section of the image. There is a high mineral occurrence density in the hinge zones, which is examined in more detail below. No particular component maps this unit in detail, however there is good resolution in the SOM spatial map. This section was of note, as it clearly details sections of maps the Silver King formation.

The SOM process produces a Q error map and the grayscale image in Figure 9 shows the quantization error, a measure of how well the unit fits into the BMU it has been matched with. The brighter spots being high Q error values. The Silver King formation is anomalously high, indicating the formation does not entirely fit with the best matching unit it was mapped with. This means that there is an incongruous component to the formation that is unique among its best matching unit.

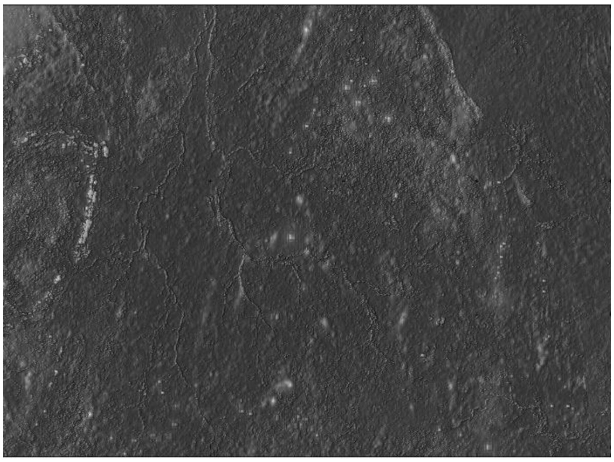


Figure 9 - Q error map of the SOM area.

You can also see streams and other heterogeneous features across the extent of the map arising from poor fits to their respective BMUs. You would expect this for rivers and streams as they act as the sink for all the weathering products from the units they act as catchments for. Thus, you are seeing a mix of different BMU's all in a specific location. Why then would we expect a high quantisation error from a geological unit?

To further examine this unit, we ran the SOM map only over the Silver King formation to see if there was any zoning within the unit. Relative to the rest of this unit (Fig 8), the eastern half has high Ferric Oxide, Thorium, and Hematite/Goethite content, middle to low mineral occurrence and a low magnetic response.

We took this question to those familiar with the area and one input which we found interesting was that the eastern and western arm of the silver king formation have different amounts of cover. This is one possible explanation of why the continuous unit had one section mapped with such a high Q-error.

CONCLUSIONS

Our project took a large amount of complex data and reduced it to a single map with software and tools. We also showed the value and practical application of applying unsupervised learning techniques to mapping trends in the GIS space. Ultimately this technique's application is limited to areas with high quality overlapping datasets. We foresee however that such areas will become more and more common and SOM may represent an effective way to interpret and rapidly characterise areas of interest as datasets cover more and overlap each other. Thus, we find that SOM is a useful tool for any analyst working with in GIS mapping and interpretation of numerous layers of spatial information..

ACKNOWLEDGMENTS

This work was completed as part of the Frank Arnott Award under the apprentice category, in which we ultimately placed 3rd. As such we'd like to thank all those who've contributed to the completion of this work. David Pratt of Tensor Research who mentored us through this project and not only shared his time and council but also his software. We'd also like to thank our co-mentor Bruce Dickson who, through his thorough knowledge of the SOM technique, gave invaluable insight into the process of SOM. We'd also like to thank Pitney Bowes and Geosoft who provided their software free of charge for the duration of this competition, as well as providing us which great technical and operational assistance to fast track our ability to utilise the software. Finally, we'd like to thank the NSW Dept. of Primary Industry including Dr John Greenfield, Astrid Carlton and Rosemary Hegarty, who provided us with access to the dataset which made this work possible but also for their feedback on the output maps.

REFERENCES

Agarwal, P. and Skupin, A. (2008) *Self-Organising Maps: Applications in Geographic Information Science, Self-Organising Maps: Applications in Geographic Information Science*. doi: 10.1002/9780470021699.

Dickson, B. L. (1995) 'Analysis and Visualization of Multiple Data Sets using Self- Organizing Maps', *CSIRO Exploration & Mining*, pp. 1–4.

Fraser, S. J., Mikula, P. A., Lee, M. F., Dickson, B. L., Kinnersly, E. and Snowden (2006) 'Data mining mining data - Ordered vector quantisation and examples of its application to mine geotechnical data sets', *6th International Mining Geology Conference, Rising to the Challenge*, (August), pp. 259–268. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-38349116959&partnerID=40&md5=d27badd7bc2344542046be364f2a1051.

Fraser 2012 ...... see comment above with first Fraser reference.

Giraudel, J. L. and Lek, S. (2001) 'A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination', *Ecological Modelling*, 146(1–3), pp. 329–339. doi: 10.1016/S0304-3800(01)00324-6.

Gulson, B., Korsch, M., Dickson, B., Cohen, D., Mizon, K. and Michael Davis, J. (2007) 'Comparison of lead isotopes with source apportionment models, including SOM, for air particulates', *Science of the Total Environment*, 381(1–3), pp. 169–179. doi: 10.1016/j.scitotenv.2007.03.018.

Sarparandeh, M. and Hezarkhani, A. (2016) 'Application of Self-Organizing Map for Exploration of REEs ' Deposition', 2016(July), pp. 571–582.

Vesanto, J., Himberg, J. and Alhoniemi, E. (1999) 'Self-organizing map in Matlab: the SOM Toolbox', *Proceedings of the Matlab DSP conference*, 99.

Wang, S. (2003) 'Application of self-organising maps for data mining with incomplete data sets', *Neural Computing and Applications*, 12(1), pp. 42–48. doi: 10.1007/s00521-003-0372-1.