GEOSTATISTICAL CLUSTER ANALYSIS OF MIXED DATA SETS WITH SPATIAL INFORMATION

Steffen Brammer
sb@brazca.com

Abstract. Geostatistical cluster analysis is routinely applied to decompose mixed data sets, which contain samples with discrete spatial information that puts the data into a relevant geographical context. Various methods exist for this purpose; however, where the individual clusters are intertwined with irregular, discontinuous or complex geometries, conventional methods struggle or fail. Therefore, a new approach has been developed in JMP® and implemented exclusively with JMP® scripts. After an initial estimate of the statistical moments of the underlying components, a series of search trees are built through the sample grid and samples are allocated to one of the conceptual target populations, depending on their probability density functions. Thus the mixed data set is split into its components while maintaining the spatial relationship within and across individual clusters. This method has been developed for the mining industry to domain the phases of multistage mineralizing events of complex ore bodies; but possible fields of application include virtually all disciplines of natural sciences (*e.g.* environmental research, hydrology, biology, agriculture, etc.) and every other discipline where the spatial position of the data matters (such as pattern recognition, image processing, logistics and marketing).

INTRODUCTION

In natural sciences in general and ore geology in particular, bi- or multimodal systems are common and well documented (*e.g.* Brammer, 2012; Dominy *et al.*, 2003; McFarlane *et al.*, 2011). Systematic sampling of such systems leads to correspondingly bi- or multi-modal data sets. In the presence of a relatively small, but very influential outlier population within the principal area of interest, the histogram of the resulting single data set comprising samples from two, or more, underlying populations is highly skewed and long-tailed. Prior to any meaningful analysis, it is required to decompose the data set into its two (or more) components, herein called 'domains'.

Conventionally, domains are delineated physically by creation of regions, commonly through manually digitized polygons or 3-dimensional solids. In practice, the definition of the domains is often based on nominal threshold values. In the field of ore geology, Stegman (2001) and Emery & Ortiz (2005) pointed out the concerns associated with domaining by rigid thresholds, but these concerns are universal and applicable to every other discipline that deals with discretely arranged spatial data. And they are particularly true for bi-modal systems where the histograms of the individual populations are likely to overlap: Samples from the lower tail of the outlier population would be assigned to the principal domain and vice versa, resulting inevitably in inappropriate domain boundaries and geometry.

Alternatively, domains could be demarcated using geostatistical cluster analysis. Cluster analysis measures the degree of similarity, or dissimilarity, between individual observations and then attempts to group these observations into subsets (called clusters, or again – in this context – domains). Several algorithms have been suggested for data sets within a spatial, that is, a geostatistical framework (*e.g.* Allard & Guillot, 2000; Romary *et al.*, 2012). Ambroise *et al.* (1995) proposed model-based clustering through expectation-maximization (EM) that relies on the assumption that the data are drawn from a particular distribution.

Irrespective of the choice of currently available methods, domaining is expected to be difficult,

unreliable or simply impractical when the features of one of the domains (usually the outlier domain) are particularly narrow, discontinuous or complex in their geometry and/or orientation such as, for instance, cross cutting late stage veins or irregular stockworks, as frequently described from so called *high-nugget* gold deposits. For these types of bi-modal systems, this paper proposes an alternative strategy: Domaining is carried out in repetitive trial-and-error searches along paths throughout the sample grid. Samples are progressively allocated to one of the domains according to the statistical properties of the expected underlying components. The method is described below.

CONCEPT & METHODOLOGY

Let us assume a bi-modal system that has been regularly sampled. The distribution of the single data set comprising the two components is equally bi-modal. If both components are approximately normally (or log-normally) distributed, then the mean and spread of the individual components, as well as the probability ratios (that a sample belongs to one or the other domain), can be estimated. Several methods are available to assess the components of a mixed distribution, ranging from simple visual estimates to complex statistical solutions (*e.g.* McLaughan and Peel, 2000).

Once these critical parameters are established, Once these critical parameters are established, probability density functions are generated and used to set up the target histograms of the expected domains. This allows the determination of the number of samples belonging to each domain for any given grade range, as a central condition for the discrimination process during the search – once the maximum of samples for a given grade range has been reached for one of the two target domains, the sample in question must belong to the other domain.

The search itself is conducted on a trial-and-error basis. It starts from a randomly chosen seed location within the sample grid whereby the sample potentially belongs to the smaller outlier domain. Samples within a specified neighbourhood are then investigated and those samples that would fit into the target histogram of the outlier domain are tagged and added to the respective grade bin. Subsequently, the neighbourhoods of the tagged samples are investigated, and so forth, thus building a progressive search tree through the sample grid. The individual branches of the search tree stop when no sample in the neighbourhood satisfies the target histogram, which is the case when the search has entered the principal domain and the lower tail of the target histogram has been filled up.

Once the search is interrupted, a new, untagged seed sample is chosen and the build-up of a new search tree recommences. This procedure continues until all samples that potentially belong to the outlier domain have been investigated, with paths eventually through, and along, all possible outlier clusters.

Since the result of a single run depends strongly on the sequence of the seed samples, it is recommended to be repeated the entire process sufficiently often, with identical, but ideally also with different search parameters such as varying search directions and anisotropy ratios, in order to capture complex and irregular features.

Finally, the results of the multiple search runs are averaged to produce a map that shows for each sample the probability of belonging to the outlier domain.

EXAMPLE

Given is a data set sampled from a synthetically simulated 2-dimensional grid – see Figure 1a. The principal domain (Domain 1) strikes NNE, with somewhat stronger values along the edges and lower values in the central portion. The upper outlier domain (Domain 2) occurs in narrow, well defined zones that cross-cut the principal domain – see Figure 1b. The bi-modality is clearly exposed, with the two underlying components approximately normally distributed – see Figure 2 - and the critical parameters of the two populations can be estimated. In this case, Domain 2 consists of 75 samples with a mean of 2.5 and a standard deviation of 1.05.

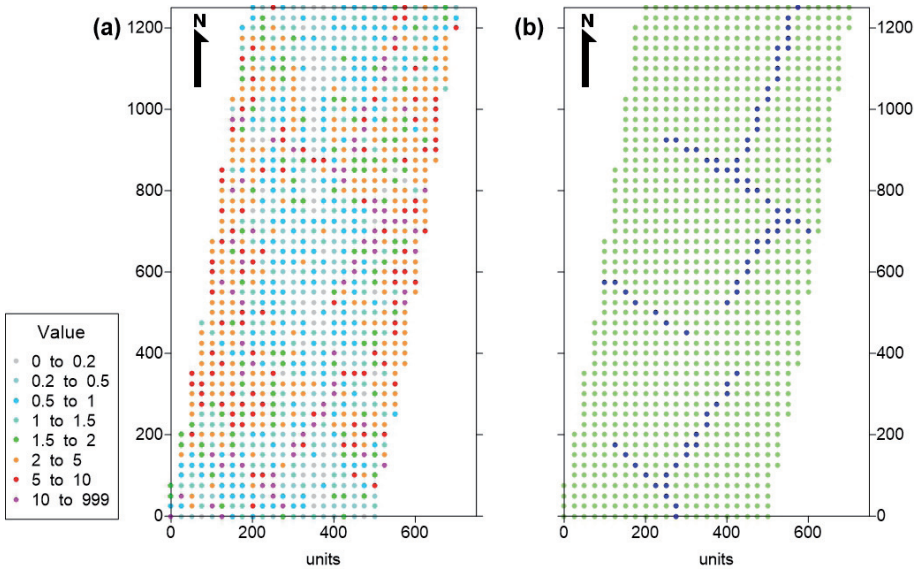

Figure 1
Maps showing the sample data: (a) coloured by value, (b) coloured by domain green: Domain 1, blue: Domain 2

For the search runs, the target histogram of Domain 2 has been set up to consist of 6 grade bins. Each grade bin comprises the range of 1 standard deviation and the number of samples is determined by the 3-sigma rule for normal distributions: 68% (of a total of 75 samples) lie within 1 standard deviation from the mean, 95% (or approximately 70 samples) within 2 standard deviations.

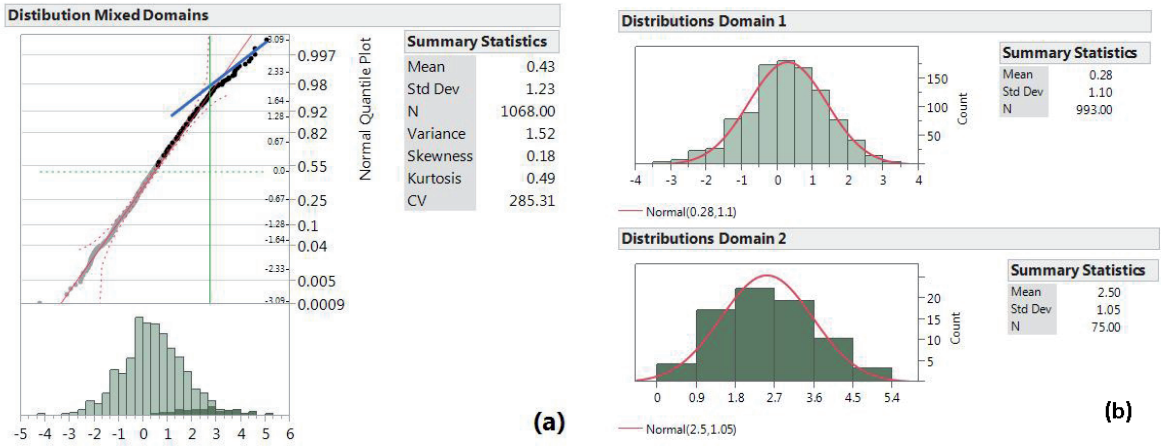

Figure 2
Distributions of the data: (a) histogram and normal quantile plot of the mixed population. Green line: estimated mean of Domain 2, (b) histogram and probability density function (red curves) of the component populations. Light green: Domain 1, dark green: Domain 2

Three omni-directional search runs have been carried out, using JMP® scripts and the JMP® platform, with 10, 25 and 50 search repetitions respectively. Results are shown in Figure 3, with those samples selected and highlighted that have the highest probability of belonging to Domain 2 when the probability cut-off was set to yield a sample size that corresponds approximately with the estimation (*i.e.* 75 samples). The maps illustrate that the quality of the result increases with the number of search repetitions, as expected. After 50 search runs, Domain 2 has been mapped out well, as the probabilities attached to samples from the low grade Domain 1 have been gradually reduced to a level where they fall below the sample size cut-off.

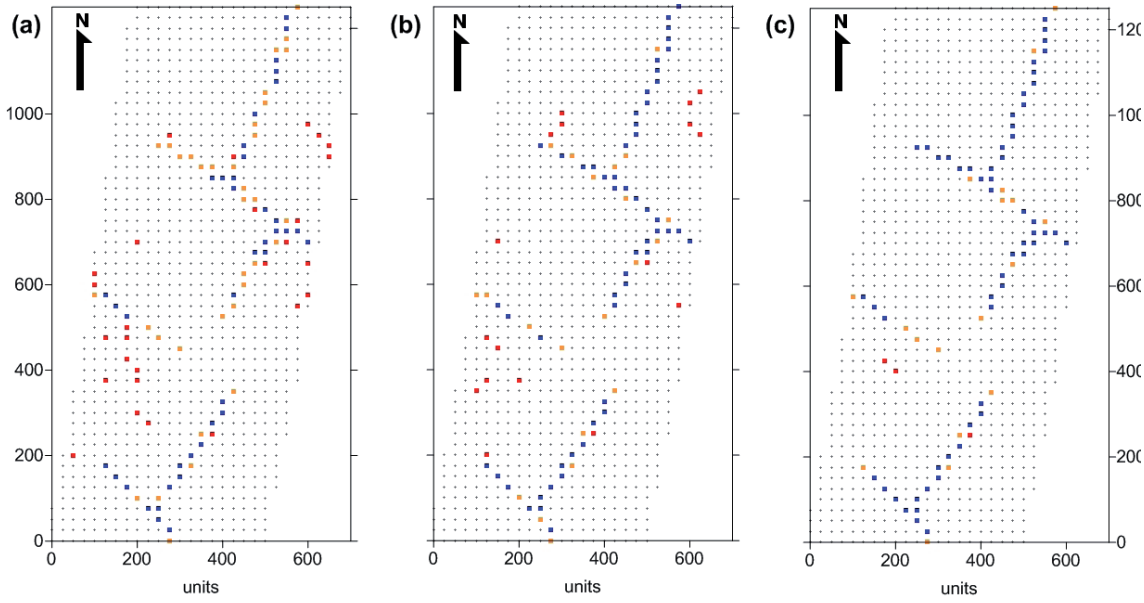

Figure 3
Maps showing samples with highest probability after (a) 10, (b) 25 and (c) 50 search runs. Blue dots: samples correctly allocated to Domain 2, red dots: samples from Domain 1 misallocated to Domain 2, orange dots: samples from Domain 2 misallocated to Domain 1

DISCUSSION

In this paper, a new strategy has been proposed to decompose highly skewed and long-tailed, bi-modal data sets. It is based on the premise that bi-modality, where present, is correctly identified and that the distributions of the two underlying component populations can be adequately estimated prior to the actual domaining process. Domaining is then conducted during repetitive, progressive searches through the sample grid by discriminating samples according to the expected target histograms of the domains. The proposed method is computational inexpensive and incorporates three properties that are considered favourable for its practical implementation:

i. besides the initial estimate of the component populations, the process does not require any expert input. This makes it suitable for universal and routine everyday use;

ii. results are not presented as categorical sets of clusters, but in form of probability maps that acknowledge the possibilities of misallocation. This enables the user to keep editorial control over the final product; and

iii. the method is easily applicable with commercial off-the-shelf statistical software (such as JMP®)

The proposed method has been developed for the mining industry to domain the various phases of multistage mineralizing events of complex ore bodies; but possible fields of application include virtually all disciplines of natural sciences (*e.g.* environmental research, hydrology, biology, agriculture, etc.) and every other discipline where the spatial position of the data matters, such as pattern recognition, image processing, logistics and marketing.

REFERENCES

Allard, D. and Guillot, G. 2000. Clustering geostatistical data. *Proceedings of the VI geostatistical conference*, Cape Town, South Africa, 10-14 April 2000. The Southern African Institute of Mining and Metallurgy, Johannesburg. pp. 49-63.

Ambroise, C., Dang, M. and Govaert, G. 1995. Clustering of spatial data by the EM algorithm. *In:* Soares, A. *et al*. (*eds*). Quantitative Geology and Geostatistics, v. 9. Dordrecht, Kluwer. pp. 493-504.

Brammer, S. 2012. Resource estimate of a bi-modal gold deposit: merging simulated high-grade ore shoots with kriged background mineralization – a case study. *Proceedings XXXIV International Geological Congress*, Brisbane, Australia, 5-10 August 2012

Brammer, S. 2014. Domaining Bi-modal Data Sets Geostatistically Using a Directional Neighborhood Search. *Proceedings of the XV Annual Conference of the International Association for Mathematical Geosciences*, Madrid, Spain, 2-6 September 2013. Pardo-Igúzquiza, E. *et al.* (*eds*). Heidelberg, Springer. pp. 779-782

Brammer, S. 2015. Domaining of long-tailed bimodal data-sets with statistical methods. *In:* The Danie Krige Geostatistical Conference. SAIMM, Johannesburg. pp. 281-286

Brammer, S. 2015. A self-guiding domaining tool for long-tailed bi-modal data sets. *Proceedings of the 17th annual conference of the International Association for Mathematical Geosciences*. Sept 5-13, 2015, Freiberg, Germany

Clark, I. 2000. Erratic highs - a perennial problem in resource estimation. *Proceedings 2000 SME Annual Meeting*, Salt Lake City, Utah, USA, 28 February – 1 March 2000. Society for Mining, Metallurgy & Exploration. Preprint 00-110. pp. 1-6

Dominy, S.C., Annels, A.E., Platten, I.M. and Raine, M.D. 2003. A Review of Problems and Challenges in the Resource Estimation of High Nugget-Effect Lode-Gold Deposits. *Proceedings of the V International Mining Geology Conference*, Bendigo, Australia, 17-19 November 2003. The Australasian Institute of Mining and Metallurgy, Melbourne. pp. 279-298

Emery, X. and Ortiz, J.M. 2005. Estimation of mineral resources using grade domains: critical analysis and a suggested methodology. J.S.Afr.Inst.Min.Metall, v. 105. pp. 247-256

Krige, D.G. and Magri, E.J. 1982. Studies of the effects of outliers and data transformation on variogram estimates for a base metal and a gold ore body. Math.Geol., v. 14, no. 6. pp. 557–564

Leuangthong, O. and Nowak, M. 2015. Dealing with high-grade data in resource estimation. J.S.Afr. Inst.Min.Metall, v. 115. pp. 27-36

McFarlane, C., Mavrogenes, J., Lentz, D., King, K., Allibone, A. and Holcombe, R. 2011. Geology and Intrusion-Related Affinity of the Morila Gold Mine, Southeast Mali. Econ.Geol., v. 106. pp. 727-750

McLaughlan, G.J. and Peel, D. 2000. Finite Mixture Models. New York, Wiley

Romary, T., Rivoirard, J., Deraisme, J., Quinones, C. and Freulon, X. 2012. Domaining by Clustering Multivariate Geostatistical Data. *Proceedings of IX International Geostatistics Congress*, Oslo, Norway, 11-15 June 2012. Abrahamsen, P. *et al*. (*eds*). Dordrecht, Springer. pp. 455-466

Stegman, C.L. 2001. How domain envelopes impact on the resource estimate – case studies from the Cobar Gold Field, NSW, Australia. Mineral Resource and Ore Reserve Estimation. *In:* Edwards, A.C. (*ed*). The AusIMM Guide to Good Practice. Australasian Institute of Mining and Metallurgy, Melbourne. pp. 221-236.