**QUEEN'S
UNIVERSITY
BELFAST**

An examination of transformation techniques to investigate and interpret multivariate geochemical data analysis – the Tellus Case Study.

McKinley, J., & Leuangthong, O. (2010). An examination of transformation techniques to investigate and interpret multivariate geochemical data analysis – the Tellus Case Study. In P. M. Atkinson, & C. D. Lloyd (Eds.), *GeoENV-VII Geostatistics for Environmental Applications* (pp. 231-242). Springer.

Published in:
GeoENV-VII Geostatistics for Environmental Applications

Queen's University Belfast - Research Portal:
[Link to publication record in Queen's University Belfast Research Portal](#)

General rights

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

An examination of transformation techniques to investigate and interpret multivariate geochemical data analysis: Tellus Case Study.

Jennifer McKinley¹ and Oy Leuangthong²

1 School of Geography, Archaeology and Palaeoecology, Queen's University, Belfast, N.Ireland

²Centre for Computational Geostatistics (CCG), Department of Civil and Environmental Engineering, University of Alberta, Canada

*Corresponding authors: <mailto:j.mckinley@qub.ac.uk>,
oy.leuangthong@ualberta.ca*

Running title: Multivariate geochemical transformation techniques, Tellus study

Abstract

This research aims to use the multivariate geochemical dataset, generated by the Tellus project, to investigate the appropriate use of transformation methods to maintain the integrity of geochemical data and inherent constrained behaviour in multivariate relationships. The widely used normal score transform is compared with the use of a stepwise conditional transform technique. The Tellus Project, managed by GSNI and funded by the Department of Enterprise Trade and Development and the EU's Building Sustainable Prosperity Fund, involves the most comprehensive geological mapping project ever undertaken in Northern Ireland. Previous study has demonstrated spatial variability in the Tellus data but geostatistical analysis and interpretation of the datasets requires use of an appropriate methodology that reproduces the inherently complex multivariate relations. Previous investigation of the Tellus geochemical data has included use of Gaussian-based techniques. However, earth science variables are rarely Gaussian, hence transformation of data is integral to the approach. The multivariate geochemical dataset generated by the Tellus project provides an opportunity to investigate the appropriate use of transformation methods, as required for Gaussian-based geostatistical analysis. In particular, the stepwise conditional transform is investigated and developed for the geochemical datasets obtained as part of the Tellus project. The transform is applied to four variables in a bivariate nested fashion due to the limited availability of data. Simulation of these transformed variables is then carried out, along with a corresponding back transformation to original units. Results show that the stepwise transform is successful in reproducing both univariate statistics and the complex bivariate relations exhibited by the data. Greater fidelity to multivariate relationships will improve uncertainty models, which are required for consequent geological, environmental and economic inferences.

1. Introduction

The Tellus Project, managed by the Geological Survey of Northern Ireland (GSNI) and funded by the Department of Enterprise Trade and Development and

the EU's Building Sustainable Prosperity Fund, involves the most comprehensive geological mapping project ever undertaken in Northern Ireland. The project comprised the collection of both multi-source airborne geophysics and a ground based geochemical survey of soil and streams. The Tellus geochemical survey involved the collection of soil, stream-sediment and stream water samples in rural and urban areas. Rural soil samples were collected at approximately 2km^2 intervals. Each soil sample comprises a composite of five augers collected from a depth interval of 5 to 20cm. The sampling strategy involved the collection of auger samples from each corner of a square with 20m sides as well as at the centre. Site location is recorded at the central auger hole (Figure 1).

Figure 1: Location map, counties of Northern Ireland and sample locations for Tellus geochemical survey sampling scheme

1.2 Geological background

Northern Ireland, despite occupying a limited proportion of the land area represents an almost unparalleled diversity of geology (Mitchell 2004). The range

of rocks presented includes examples of all geological systems up to and including the Palaeogene (comprising basalt lavas and lacustrine sedimentary rocks formed between c. 55 and 62 million years ago). The last 100,000 years of the Northern Ireland's history involves the advance of ice sheets and their meltwaters resulting in a cover of alluvium and peat deposits over at least 80% of bedrock. The economic significance of the Tellus project and the opportunity the multivariate geochemical data offers to decipher and investigate the geological underlay relates to the history of hydrocarbon exploration and mineral prospecting in Northern Ireland. However if geological, environmental and economic inferences are to be made then the integrity of the geochemical data is paramount and manipulation of the soil geochemistry data must honour any inherent geochemical constraints. This study uses the geochemical data to examine the use of transformation methods to maintain the integrity of the data and inherent constrained behaviour in the multivariate relationships. The aim of the research is to enable greater accuracy in the interpretation of the nature of the geochemical variability and consequently any geological, environmental and economic inferences.

2 Previous research

Previous work (Rawlins et al. 2007, McKinley et al. 2006, 2008 in prep) involved the use of Tellus geochemical data to investigate methods of integrating the geochemical data with the airborne geophysical data to maximise information collected from the ground geochemical survey. The aim of the research was to enable greater interpretation of geological, environmental and economic aspects of Northern Ireland. A Gaussian-based Bayesian updating approach was used by McKinley et al. (2006, 2008 in prep) as a means to improve the resolution of the widely sampled soil geochemistry data by integrating the more closely sampled airborne geophysical data. The advantage of the approach is that multiple variables of different types and different sources (in this case radiometric and soil geochemistry) can be simultaneously integrated and applied to mapping the geochemical variables of economic interest.

2.1 Rationale for the present study

The Bayesian updating approach (Deutsch and Zanon 2004, Ren et al. 2005) used by McKinley et al. 2008 (in prep) to improve the resolution of the soil geochemistry data, is a Gaussian-based technique. Hence transformation of data is an integral stage of the approach (normal score transformation was used in this case). Geological data rarely conform to Gaussian behaviour (Leuangthong and Deutsch 2003), likewise multivariate distributions rarely exhibit Gaussian characteristics such as homoscedasticity and linearity. Common non-Gaussian behaviour for geochemical data is heteroscedasticity, non-linearity and mineralogical constraint. However, Gaussian techniques are often used to represent models of continuous variables. Common practice in geostatistical analysis of multiple-related variables is to transform each variable to a univariate Gaussian distribution one at a time. This ensures each variable is univariate but the multivariate distributions (involving two or more variables at a time) are not explicitly transformed to be multivariate Gaussian and hence does not address the case when the multivariate Gaussian

assumption is violated. An alternative transformation technique must be considered.

2.2 The Tellus geochemical data

The multivariate geochemical dataset collected by the Tellus project comprise multiple variables that are dependent on each other. This provides an opportunity to investigate the appropriate use of transformation methods such as normal score transform and the stepwise conditional transform (Leuangthong, 2003; Leuangthong and Deutsch, 2003). The transforms need to be implemented with the central aim to maintain the integrity of the geochemical data and honour the inherently constrained behaviour between multiple variables. These relationships often show complex features such as nonlinear relations and/or stoichiometric constraints. This is especially relevant for geochemical data collected as part of the Tellus project and any subsequent geological, environmental and economic inferences. With this mind, the Clogher Valley area comprising Co. Fermanagh and the southern part of Co. Tyrone, was taken due to the inferred relationship between basement faulting and base metals and renewed interest in mineral prospecting in the area. The Clogher Valley dataset comprised 589 points for seven variables of interest; Cu ppm, Ni ppm, Zn ppm, K₂O%, Pb ppm, Co ppm and Cr ppm.

Figure 2 Crossplot matrix of original variables: Cu, Ni, Zn, K₂O, Pb, Co and Cr.

3 Stepwise Conditional Transform (SCT) of Tellus Data

Before any multivariate conditional simulation with dependent variables, we need to understand the univariate distribution of each variable, and any second and higher order relations between the variables. Figure 2 shows the matrix of crossplots illustrating the bivariate relations between the seven variables. Figure 3 shows the same relations as Figure 2, with the exception that the variables are now normal score transformed. Following a univariate normal score transform, the complex features (e.g. heteroscedasticity, constraints, non-linearity) that are apparent in the original variable crossplots are visibly transferred into Gaussian units; the presence of these relations after Gaussian transform indicates that they may be challenging to reproduce in a conventional Gaussian simulation framework. Normal score transform can usually be effective in mitigating heteroscedastic features but in several of the cross plots it is observed that this clearly not the case. The bivariate distributions are clearly not bivariate Gaussian. This is most evident for crossplots involving K2O, Cr, Cu and Ni. For this reason, an alternative transform is considered for these four variables.

Figure 3 Crossplot matrix of Normal Score transformed values for Cu, Ni, Zn, K2O, Pb, Co and Cr.

For complex multivariate relations, a number of transformation approaches can be considered. Principal components or factor analysis could be used to generate uncorrelated variables; however, a lack of correlation does not ensure independence.

A log ratio transform (Aitchison, 1981,1999) is another alternative, but is primarily aimed at accounting for compositional data that exhibit constrained behavior. Minimum/maximum autocorrelation factors (Switzer and Green 1984, Vargas-Guzmán and Dimitrakopoulos 2003) is yet another technique that is available and is an extension of PCA to a lag $h \neq 0$ scatterplot. There are many other multivariate transformation approaches available for different purposes; however, in most cases and for those identified here, a transformation to Gaussianity is still required and there is no assurance that even bivariate Gaussianity can be achieved.

The stepwise conditional transform (SCT) was introduced by Rosenblatt (1952) and is described in detail by Leuangthong and Deutsch (2003). The technique applies a quantile transformation technique of observed univariate conditional distributions to standard Gaussian distributions. For the univariate case the SCT technique is identical to the normal score transform. In a bivariate situation, the first variable (Z_1) is transformed using normal scores to yield Y_1 . The normal score transformation of the second variable (Z_2) is conditional to the probability class of the first or primary variable (Y_1). In essence, Z_2 is partitioned into classes conditional to Y_1 . A normal score transform is then undertaken for each class of Z_2 . For the k -variate case the k^{th} variable is conditionally transformed based on the $(k-1)$ first variables. All multivariate distributions are Gaussian in shape at distance lag $h = 0$. The covariance at $h > 0$ may not be zero.

3.1 Implementation of SCT in a nested fashion for Tellus data

With less than 600 samples available in the entire dataset for the Clogher Valley area (589 data points), stepwise transformation of four variables will yield poor results for the third and fourth transformed variable due to paucity of information to infer the conditional distributions. Since the transform requires successive conditioning as we increase the number of variables, this effectively means that we are sub setting the data into finer and finer classes, leaving fewer data within each class. For example, in the case of two variables, if we had 100 Cu data points and established that we would subdivide into 10 classes, this would mean we had 10 Cu data within each class. To define a conditional distribution based on 10 data is at the very limit of what would be considered reliable. As a result, a nested transform order (Leuangthong et.al. 2006) up to two variables is considered (see Table 1). Figure 4 shows the crossplots corresponding to each transform order. Modelling is focussed on the three transformed variables. Note that for the first variable K₂O%, SCT K₂O is the same as the normal score (NS) K₂O.

SCT Order	Primary	Secondary	ρ_{os}	ρ_{ns}	ρ_{sct}
1	K ₂ O	Cr	0.652	0.543	0.003
2	K ₂ O	Cu	0.223	0.357	0.000
3	Cr	Ni	0.858	0.879	0.050

Table 1 Summary of SCT transform orders and corresponding bivariate statistics.

The choice of the primary variable and the ordering of transformation are based on correlation coefficients (Table 1).

Figure 4 SCT order 1 (left), 2 (middle) and 3 (right) using 10 classes.

For the forward transformation, SCT is run three times with the order shown in Table 1. Following transformation, the variograms corresponding to the four transformed variables are calculated and modeled (see Figure 5).

Figure 5 Variograms corresponding to four SCT Variables. A slight trend is apparent in Ni but this was not explicitly modelled in the results shown.

The variograms of K2O and Cr are calculated from the first order transforms, the variogram for Cu from the 2nd order transform and for Ni based on the 3rd order transform. The cross variograms are also calculated to verify that correlation at $h > 0$ remains relatively small to allow independent modeling to be carried out. Using the SCT values and variograms (based on the orders described in the previous

step and Table 1), sequential Gaussian simulation (SGS) is then performed for each of the four SCT variables. Similar to the forward transform, the back transform must also be performed in a stepwise conditional manner.

Figure 6 Comparison between E-type estimate (left) and simulated SGS realization (right).

Using the order outlined in Table 1, the following steps are followed: (a) K₂O and Cr are back transformed first using transform table from order 1; (b) Cu is back transformed based on back transform values of K₂O values given from (a) and the

transform table from order 2; and finally Ni is back transformed based on the back transformed values of Cr (from (a)) and transform table from order 3. This way, we avoid multiple values of K₂O (although they would have been the same given the order established) and multiple values of Cr, which would be a more critical issue. In total, 100 realizations were generated. For the purposes of comparison, the E-type estimate of the simulations and an arbitrarily chosen realization are plotted in Figure 6. As expected, the E-type estimate yields a smooth map that is similar to a kriged result while the simulated realization is clearly more variable. Regions of high and low concentrations are easily identifiable in both cases. Moreover the relationship between zones of higher elemental concentration and fault orientation is evident. Following back transformation each model was checked for data reproduction, histogram reproduction, variogram reproduction and multivariate distribution reproduction. Figure 7 shows the reproduction of the bivariate relations (as seen in Figure 2) following simulation using SCT.

Figure 7 Reproduction of crossplot features following simulation using SCT.

3.2 Data related issues with SCT

There are three important issues that need to be addressed with the use of the SCT technique: 1) cross variance for $\mathbf{h} > 0$, 2) the effect of ordering on covariance models, and 3) inference of multivariate distributions with sparse data. 1) There is no guarantee that there is independence beyond $\mathbf{h} = 0$ (i.e. at $\mathbf{h} > 0$). The cross variogram of the transformed variables is checked to ensure that if there is correlation beyond $\mathbf{h} = 0$, that this correlation is relatively negligible (i.e. $\rho(\mathbf{h}>0) \leq 0.2$). If this is not the case then some form of cosimulation may have to be considered.

However, experience with many data sets (e.g. Leuangthong, 2003; Leuangthong and Deutsch (2003, 2004); Leuangthong et.al. 2006) has shown that this has generally not been required. 2) Leuangthong and Deutsch (2003) found that the effect of transformation ordering was observable in the departure of the variogram of the transformed variable from the original variable. The mismatch can be minimized by choice of the most continuous variable as the primary for the SCT. In the Tellus Clogher Valley data, K_2O forms the most continuous variable and is used as the first variable for the SCT technique. 3) Sparse data leads to erratic and non-representative conditional distributions. A general rule is that 10^N to 20^N number of data is acceptable where N is the number of variables (Leuangthong and Deutsch 2003). A limited data set can be supplemented by the use of smoothing algorithms such as kernel densities to ‘fill-in’ the gaps; this was not required in this study.

SCT was implemented in a nested fashion for the Tellus Clogher Valley data. In this case a data set totalling less than 600 samples would have yielded poor results for the third and fourth transformed variable due to limited information to infer the conditional distributions. There is an implicit assumption in the implementation of SCT that all data variables are available at all data locations. Therefore the greatest limitation to SCT is non-isotropic sampling. One solution is to transform and simulate the first variable at all locations (Leuangthong and Deutsch 2003). However, there remains no unique transformed value for the secondary data at all locations of non-isotropic sampling. This was not an issue in the current research.

4 Conclusions

Geostatistical analysis and interpretation of the Tellus geochemical datasets requires use of an appropriate methodology that reproduces the complex multivariate relations that are inherent to the data. SCT is investigated and developed for the geochemical datasets. The transform is shown to reproduce the heteroscedastic, non-linearity and constrained behaviours evident in the data. A nested transform order is considered given the relatively few samples that are available for a multivariate study. These findings are of interest in particular because of the previously recorded relationship between base metals and basin faulting in the Clogher Valley area (Mitchell 2004, McKinley et al. 2006). The value of the research is reduced uncertainty in modeling of soil geochemistry data, and honouring of inherent geochemical constraints. This will enable more meaningful interpretation of the nature of the geochemical variability in data and consequent geological, environmental and economic inferences. Future work will involve comparison with other transformation methods such the use of the log ratio transform to deal with a greater number of multiple variable and consideration of the constant sum constraint given the geochemical nature of the data.

5 References

- Aitchison J (1981) A new approach to null correlations of proportions. *Mathematical Geology*, 13, 2, 175-189.

- Aitchison J (1999) Logratios and natural laws in compositional data analysis. *Mathematical Geology*, 31, 5, 563-580.
- Deutsch CV and Zanon S (2004) Direct Prediction of Reservoir Performance with Bayesian Updating under a Multivariate Gaussian Model. Paper presented at the Petroleum Society's 5th Canadian International Petroleum Conference (55th Annual Technical Meeting), Calgary, Alberta, Canada, June 8 – 10, 2004.
- Leuangthong O (2003) Stepwise Conditional Transform for Multivariate Geostatistical Simulation, PhD Thesis, University of Alberta, Edmonton.
- Leuangthong O and Deutsch CV (2003) Stepwise Conditional transformation for simulation of multiple variables. *Mathematical Geology*, 35, 2, 155-172.
- Leuangthong O and Deutsch CV (2004) Transformation of residuals to avoid artefacts in geostatistical modelling with a trend. *Mathematical Geology*, 36, 3, 287-305.
- Leuangthong O, Hodson T, Rolley P, and Deutsch CV (2006) "Multivariate Simulation of Red Dog Mine, Alaska, USA", *CIM Bulletin*.
- Mitchell WI (2004) The Geology of Northern Ireland- Our Natural Foundation, Geological Survey of Northern Ireland, Belfast.
- McKinley JM, Ruffell A, Lloyd C, van Dam C and Smyth D (2006). Use of geostatistics in the integration of multi-source geophysical and geochemical data in mapping the geological underlay of Northern Ireland. *Proceedings of International Congress of the IAMG 2006 Liège, Belgium September 3–8, 2006*. (eds) Picard E, Dassargues A and Havenith HB) IAMG.
- McKinley JM, Ruffell A, Deutsch CV, Neufield C and Young ME (2008 in prep) Use of geostatistics in the integration of multi-source geophysical and geochemical data generated by the Tellus Project, Northern Ireland.
- Rawlins BG, Lark RM and Webster R (2007) Understanding airborne radiometric survey signals across part of eastern England. *Earth Surface Processes and Landforms*, 32, 1503–1515.
- Ren W, Leuangthong O and Deutsch CV (2005) Global Resource Uncertainty Using a Spatial/Multivariate Decomposition Approach presented at the Petroleum Society's 6th Canadian International Petroleum Conference (56th Annual Technical Meeting), Calgary, Alberta, Canada, June 7 – 9.
- Rosenblatt M. (1952) remarks on a multivariate transformation: *Ann. Math. Stat* v.23 no. 3 p.470-472.
- Switzer P and Green AA (1984) Min/Max autocorrelation factors for multivariate spatial imaging. Technical Report No. 6, Department of Statistics, Stanford University, Stanford, CA, 14pp.
- Vargas-Guzmán JA and Dimitrakopoulos R (2003) Computational properties of min/max autocorrelation factors. *Computers and Geosciences*, 29, 715–723.