Geostatistics and Clustering for Geochemical Data Analysis

by

Carlos Prades

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Mining Engineering

Department of Civil and Environmental Engineering
University of Alberta

Abstract

This thesis addresses challenges in geostatistical analyses of multivariate geochemical data that commonly contain complexities that have a significant influence on geostatistical modeling and cluster analysis.

For geostatistical modeling, the effect of the most common despiking methods is investigated and their problems documented. It is shown that both local average despiking and random despiking lead to bias in the observed variogram and predicted uncertainty. A new despiking method is proposed and implemented to improve variography when the variable has a significant spike. The developed approach combines a random despiking component and a local average despiking component.

Cluster analysis can be applied for mineral exploration purposes. It can be used to find large structures in the data and also to detect multivariate anomalous samples. Data transformations are shown to have a significant impact on clustering results. Guidance and recommendations on appropriate data transformations for improving cluster analysis performance are provided.

Three different methods are developed for identifying multivariate anomalies with cluster and spatial analysis. The first method uses different combinations of clustering and data transformations for finding small anomalous clusters. The second uses different clustering outputs for identifying samples that do not clearly belong to any cluster. The third recognizes samples that are spatially anomalous. Each of these multivariate methods detects anomalies from a different point of view. A combination of these detection methods is recommended. The goal is to obtain more stable and reliable results. Its application in stream silt samples from the Northwest Territories shows that the proposed multivariate anomaly detection methods are capable of identifying several showings (known mineral deposits). Some of these showings are not detected from the histograms of different elements; this supports and motivates the use of multivariate anomaly detection methods for mineral deposit exploration.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

GMM	Gaussian Mixture Models
ICP-MS	Inductively Coupled Plasma Mass Spectrometry
NSS	Normal Scores Preserving Spike Transformation
LUCC	Lack of uniform clustering classification

Chapter 1. Introduction

1.1. Problems and Motivations

This thesis addresses challenges in geostatistical analyses of multivariate geochemical data (Reimann, Filzmoser, & Garrett, 2002; Templ, Filzmoser, & Reimann, 2008). Such data commonly exhibit highly skewed distributions, presence of multimodal distributions, univariate and multivariate outliers, and many samples with values at below detection limit (large spikes). These complexities have a significant influence on geostatistical modeling and cluster analysis.

For geostatistical modeling, the use of normal scores transformation alleviates most of complexities, including skewness and outliers. However, the method selected to transform the spikes to normal scores has an influence in the spatial variability, and thus, it has an impact in all geostatistical models based on variograms modeled after transforming data to normal scores (Pyrcz & Deutsch, 2014; G Verly, David, Journel, & Marechal, 1984). Two methods are commonly used in geostatistics for despiking: random despiking and local average despiking (Rossi & Deutsch, 2013). Random despiking introduces artificial short scale variability (noise) by breaking the ties randomly. Verly (1984) proposed local average despiking where the constant values are ranked according to their local average before being transformed to normal scores. The local average despiking method has been recommended as a better alternative than random despiking when the spikes are significant, to avoid a too-high nugget effect and unrealistic short scale spatial variability (Pyrcz & Deutsch, 2014; Rossi & Deutsch, 2013). Nevertheless, local average despiking could introduce too much short scale continuity. One goal of this thesis is to understand and document the impact of these despiking methods in variography and propose a modified despiking method.

Regarding cluster analysis, this multivariate technique plays an important role in this thesis because it has different applications for exploration purposes. Cluster analysis can be applied to improve metallogenic models, by finding out the metallogenetic meaning of the different clusters; it can be used to find locations where a particular metallogenic unit related to the target ore deposits is present, with the possibility of identifying areas where that geological unit of interest were mistakenly assigned to another unit in the geological map; and it can be applied for looking for ore deposits focusing on patterns rather than univariate outliers, among other possible applications. Cluster analysis has been proposed for geochemical data as an exploratory analysis method (Templ et al., 2008) and for multivariate anomaly detection (Cohen, Kelley, Anand, & Coker, 2010). Another feature that makes cluster analysis suitable for exploration purposes is that it is an unsupervised learning method, meaning that it does not require a response variable to train the algorithm —in exploration there is limited knowledge of where deposits are and are not— but it focuses on finding associations and patterns in data (Hastie, Tibshirani, & Friedman, 2009).

However, multivariate methods like cluster analysis have not been widely used for exploration purposes (Cohen et al., 2010; Filzmoser, Garrett, & Reimann, 2005). One of the concerns about cluster analysis is that it is quite sensitive to different data preparation and clustering methods (Templ et al., 2008). This concern is understandable considering that data complexities and data transformations have a significant impact on clustering results (Massart et al., 2001; Milligan & Cooper, 1988). Consequently, another aim of this thesis is to provide understanding about the impact of different transformations on cluster analysis and guidance on clustering for different purposes for mineral exploration.

Finally, a novel method for anomaly detection with multivariate data is proposed based on the knowledge gained about clustering. This method takes advantage on clustering sensitivity to data transformations, parameters and methods.

The methods and concepts developed in this thesis are not restricted to exploration data. They can be applied to many different multivariate geostatistical applications.

1.2. Regional Geochemical Data for Case Studies

Publicly available data provided by the Government of the Northwest Territories has been used for the case studies developed in the different chapters of this thesis. The geochemical data was collected by the Northwest Territories Geological Survey in partnership with the Geological Survey of Canada across the Mackenzie Mountains in the Northwest Territories, Canada (Figure 1).

Figure 1. Regional geochemical data used in the case studies developed in this thesis, located in the Mackenzie Mountains (Northwest Territories, Canada).

This regional geochemical survey was conducted for the evaluation of mineral potential in the area, based on sample collection and analysis protocols developed by the Geological Survey of Canada for the National Geochemical Reconnaissance program (Falck et al., 2012). Three kinds of samples are available: stream silt samples, stream water samples and bulk stream sediment samples (heavy mineral concentrates). The stream silt samples are the more widely and densely distributed in the area. The case studies are based on the stream silt samples database, which

after removing duplicates is composed by 8959 samples. The database contains the analysis of several elements by using different measurement techniques, including Inductively Coupled Plasma Mass Spectrometry (ICP-MS), Instrumental Neutron Activation Analysis (INAA) and Atomic Absorption Spectroscopy (AAS). The case studies are developed based on a subset of 35 ICP-MS variables measured on the stream silt sediment samples.

1.3. Thesis Outline

In Chapter 2 the impact of spikes and the current despiking methods are evaluated and a new despiking method is proposed to improve geostatistical analysis of data that contain significant spikes. In Chapter 3 the importance of different data transformations on clustering methods are explained and illustrated, providing guidance about how to improve cluster analysis when based on data that contains different complexities. In Chapter 4 different clustering methods are proposed and combined for multivariate anomaly detection. In each of these chapters a case study is developed using the stream silt samples from the Mackenzie Mountains, Northwest Territories, Canada. Conclusions and recommended future work are discussed in Chapter 5. Finally, the software developed for despiking and Python scripts used for anomaly detection are shown in the Appendix.

Chapter 2. Presentation of Data

The purpose of this chapter is to introduce the regional geochemical data from the Northwest Territories that is going to be used for the different case studies developed in this thesis, providing a broader context about some characteristics commonly found in regional geochemical exploration data. To this end, an exploratory univariate and multivariate analysis is performed.

2.1. Exploratory Analysis

A subset of 35 elements analyzed by ICP-MS of the stream silt sediment samples is selected for applying the different methods developed in this thesis. This subset contains 8959 samples. A visual inspection of variables allows identifying an influence of geological units on the background level of different elements, resulting in a general orientation NW-SE for the major continuity axis (Figure 2).

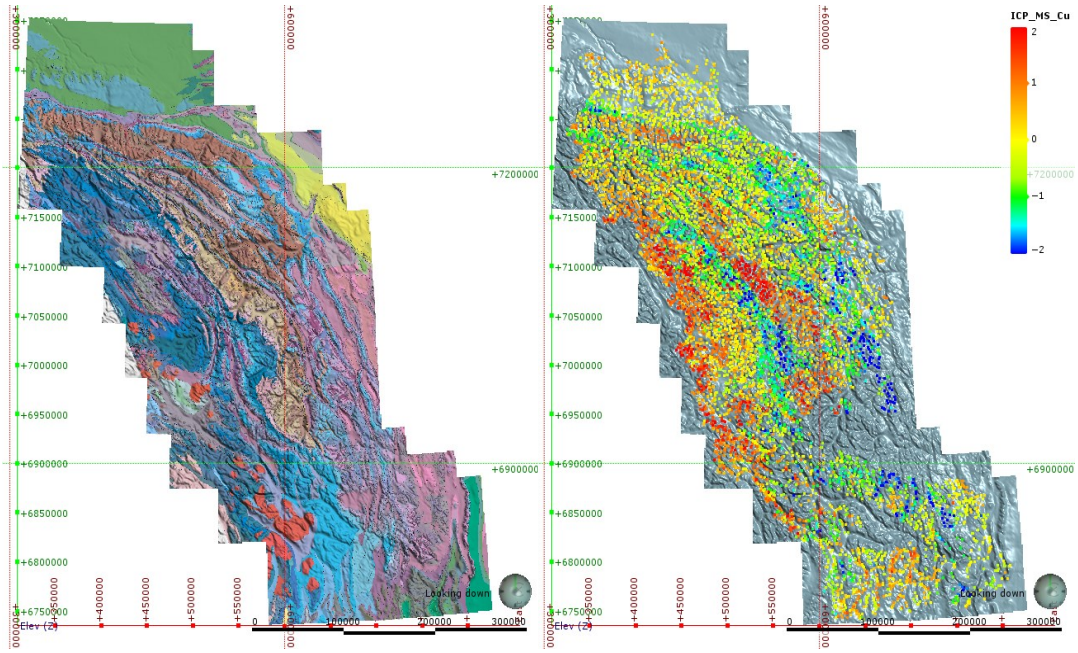

Figure 2. Left: Geological map of the Mackenzie Mountains provided by the Northwest Territories Geological Survey. Right: stream silt samples Cu spatial distribution (normal scores units).

2.1.1. Univariate Analysis

The stream silt sediment data from the Mackenzie Mountains contain different complexities, which makes this dataset suitable for illustration purposes in the different chapters of this thesis. A univariate analysis was performed to help perceive the characteristics and complexities of the distributions for the different variables. Histograms and cumulative probability plots were generated for all variables. Looking at the histograms it was observed that all variables are highly positive skewed and that some of them —Ca and Mg— present bimodal distributions (Figure 3).

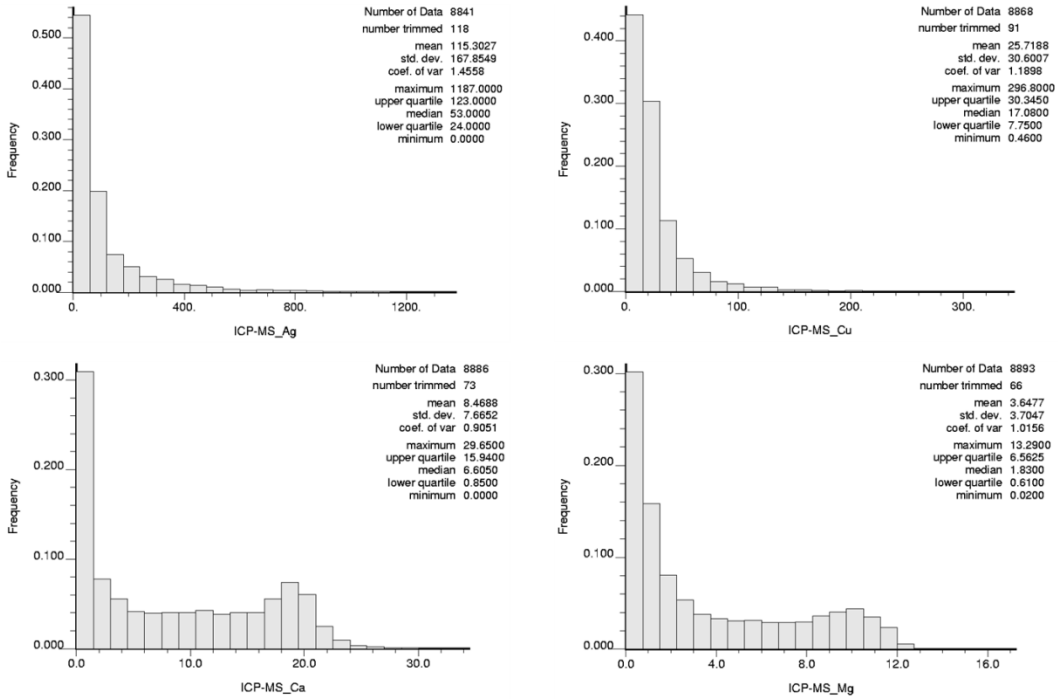

Figure 3. Histograms illustrating skewness (Ag and Cu) and bimodal distributions (Ca and Mg) in silt sediments data.

As illustrated in Figure 4, using cumulative probability plots it is possible to see that most of the variables also show extreme high values (outliers), and further, it is common that variables contain a large percentage of values at below detection limit (large spikes). The spikes corresponding to values at below detection limits for the data analyzed are described in Table 1, where it is shown that 17 elements

contain values at below detection limit and that some of them have large spikes (S, Te and W).

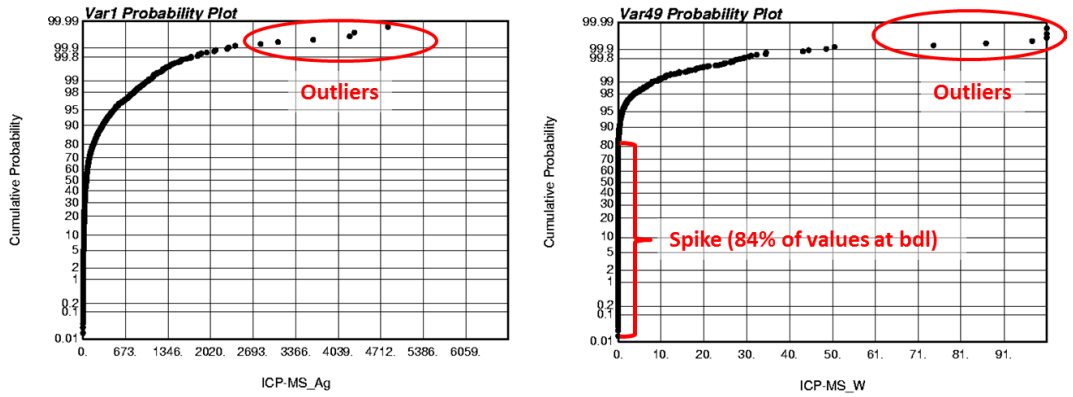

Figure 4. Cumulative probability plots showing two common characteristics observed in the distributions: outliers and spikes.

Element	Constant Values	% Spike
Ag	3	0.03%
As	4	0.05%
Bi	490	5.53%
Ca	4	0.05%
Cd	3	0.03%
Ga	8	0.09%
Hg	717	8.09%
K	24	0.27%
Na	104	1.17%
S	2234	25.22%
Se	600	6.77%
Te	3362	37.95%
Th	4	0.05%
Ti	730	8.24%
Tl	65	0.73%
V	25	0.28%
W	7436	83.94%

Table 1. Elements of the stream silt sediments dataset that contain values at below detection limit. The amount of constant values in the spike and the percentage of the total number of data that is part of the spike are shown.

All these data complexities —skewness, multimodality, outliers and spikes— represent a challenge for applying different (geo)statistical analyses, and must be considered and treated in order to get reasonable results.

2.1.2. Multivariate Analysis

Multivariate analysis was performed to understand the relationship and structure between the data variables. Original units as well as normal scores units are considered. The normal scores transformation considers local average despiking. Normal scores transformation alleviates some of the problems produced by skewness, outliers and spikes. More reliable multivariate analysis may be found. Normal scores transformation is widely used in geostatistics since it is a necessary preliminary step for different workflows.

As seen in Figure 5, the correlation matrix for variables in original units is quite different to the correlation matrix in normal scores. In order to explain this difference, the original distributions were analyzed for some of the variables whose correlations vary the most —like Bi, Hg and Sg— and for variables which correlations are almost the same in either original units or normal scores —like Ca, Sr and Ti. The greatest impact on the difference between them is caused by the presence of outliers. Even one extreme value can have a large influence and distort the correlation between two variables. This is illustrated in Figure 6 where the cumulative probability plot for a variable with similar correlation coefficients in both correlation matrices (Ti) and for a variable that has quite different correlation coefficients (Bi). Accordingly, the correlation structure obtained for data in normal scores is considered more reliable.

Figure 5. Correlation matrix for data in original units (top) and in normal scores (bottom). Note that in the upper correlation matrix the variables are ordered according to their similarity, which is represented by a dendogram. In order to facilitate comparison, the same order was kept for the lower matrix.

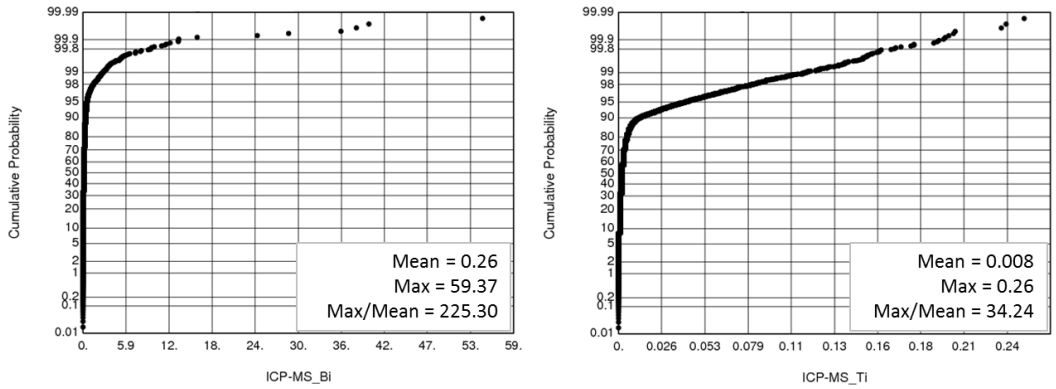

Figure 6. Cumulative probability plot for Bi (left) and Ti (right). Bi correlation coefficient with other variables varies considerably measured in original units vs normal scores, since it contains very high extreme values (outliers) that distort the correlation measure. On the other hand, Ti maximum value is not so high compared to the other values of its distribution, and consequently the correlation coefficient is similar in original units vs normal scores.

Finally, the following groups of correlated variables can be observed in the correlation matrix (Figure 7) and MDS plot (Figure 8) for data in normal scores:

1. Ca, Mg, Na and Sr.
2. Co, Cu and Sc.
3. Al, Bi, Cr, Fe, Ga, K, La, Mn, and Th.
4. Ag, As, P, V and Ni.
5. Hg, Sb, Se and Zn.
6. Cd, Hg, Mo, Sb, Se, Tl and U.

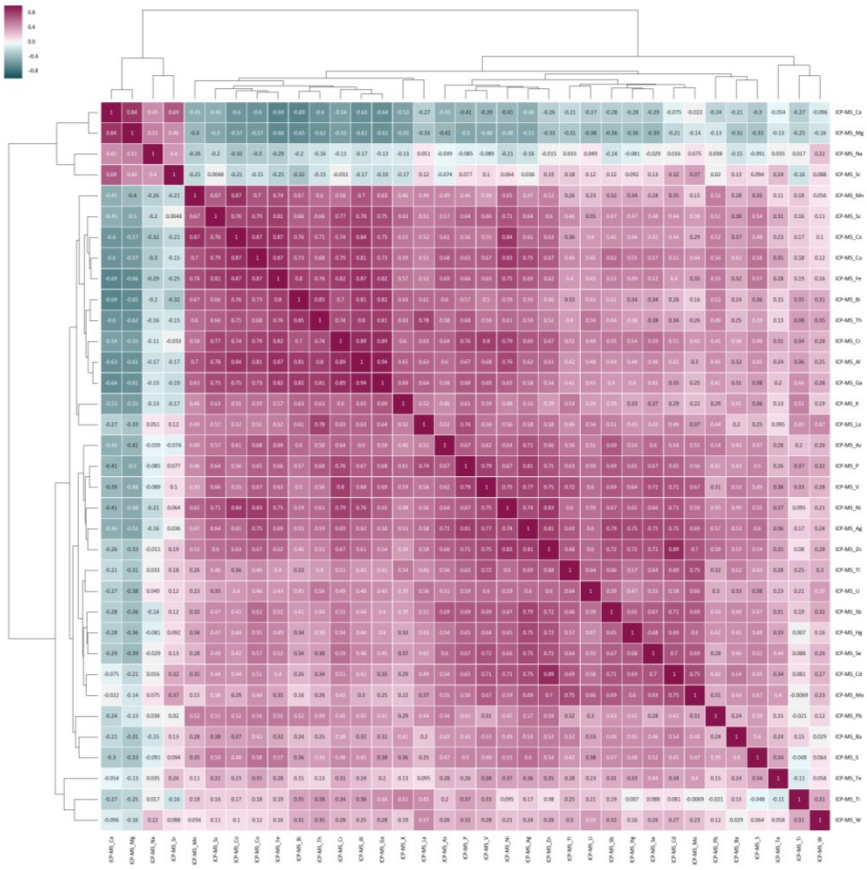

Figure 7. Correlation matrix for data in normal scores. The variables are ordered according to their similarity, which is represented by a dendogram.

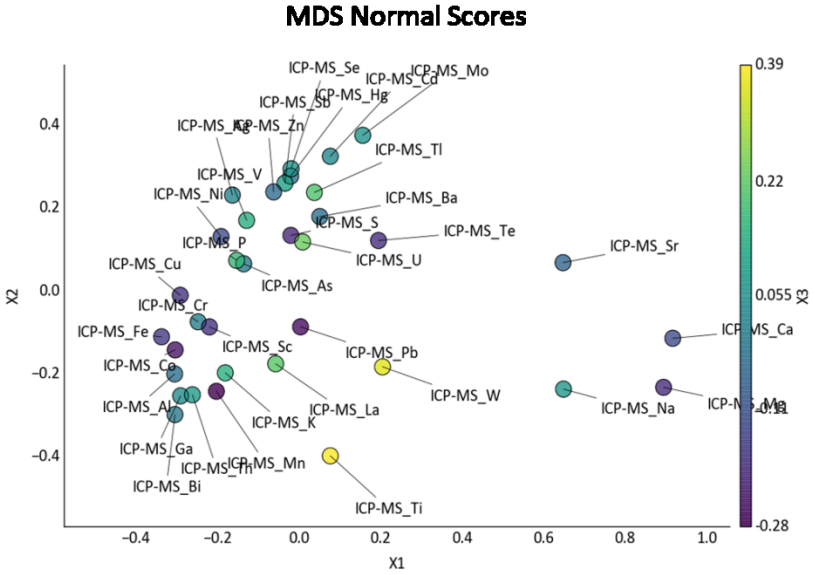

Figure 8. Multidimensional scaling (MDS) plot. This embedding method represents in three dimensions the similarity between variables (the third dimension corresponds to the color). Variables that are closer in the plot are more similar (higher correlation).

2.2. Conclusion

An exploratory analysis has been performed for the stream silt sediments data used in the cases studies developed in this thesis. This data contains different data complexities that make it suitable for testing the different methods proposed herein. It has been shown in this chapter that these complexities have an impact even in simple multivariate exploratory analysis and that data transformation have an important place for geostatistical analysis.

Chapter 3. Despiking Methods for Geostatistical Modeling

3.1. Introduction

Spikes are a common complexity found in regional multivariate geochemical data. There are many samples with measurements at below detection limit. Yet, the principles and software developed here can be applied to any geostatistical model with data containing spikes, which are common in many ore deposits like Au epithermal mineral deposits, where spikes can even correspond to 50 or 60% of data (Rossi & Deutsch, 2013).

Normal scores transformation is a preliminary requirement for many important workflows in geostatistics. When transforming data containing spikes to normal scores, it is necessary to previously break the ties in order to perform the quantile-to-quantile transformation. Two methods for breaking the ties are commonly used in geostatistics: random despiking and local average despiking (Pyrcz & Deutsch, 2014; Rossi & Deutsch, 2013). For performing random despiking, a small random number is added to each data in the spike, and then the modified number is used for sorting them before performing quantile transformation (Pyrcz & Deutsch, 2014). For applying local average despiking, the idea introduced by Verly (1984) is to calculate the average of data values close to the location of each sample in the spike, and then sort them according to that average before preforming the transformation.

The local average transformation has been recommended as a better approach for despiking, because random despiking can introduce a too-high nugget effect and unrealistic short scale variability, while local average avoids introducing artificial spatial variability (Pyrcz & Deutsch, 2014; Rossi & Deutsch, 2013; G Verly et al., 1984). The effect of these two methods has not been documented. Furthermore, the impact of local average despiking in the variography and subsequent

geostatistical calculations could be important. In principle there could be exaggerated short scale continuity producing an error in variogram modeling.

The aim of this chapter is to understand and document the effect of the two despiking methods mentioned. A new approach will be proposed to improve despiking for improved variogram modeling.

3.2. Effect of Despiking Methods in Variography

Two case studies are developed: a synthetic case study with unconditional simulation and a case study based on the stream silt sediments data from the Northwest Territories.

3.2.1. Case Study 1: Synthetic Case

The goal is to understand the influence of spikes and despiking methods on measures of spatial continuity. A synthetic case is created using a known variogram model and unconditional simulation. For creating reference results (Figure 9, left plot) a two dimensional grid of 256x256 points are simulated in a 1 by 1 resolution. The variogram model has a nugget effect of 20% and a single spherical structure with a range of 16. Finally, the simulated realization was transformed to normal scores. The experimental variogram considered the "true variogram" is shown in Figure 10.

Then, six different cases with varied amount of spikes are created to understand the effect of despiking methods for different spike sizes. The percentages of spike are: 5%, 10%, 20%, 30%, 40% and 50%. For illustration, the reference realization and the 30% spike case is shown in the right plot in Figure 9. The histogram of the 30% spike case is displayed in Figure 11.

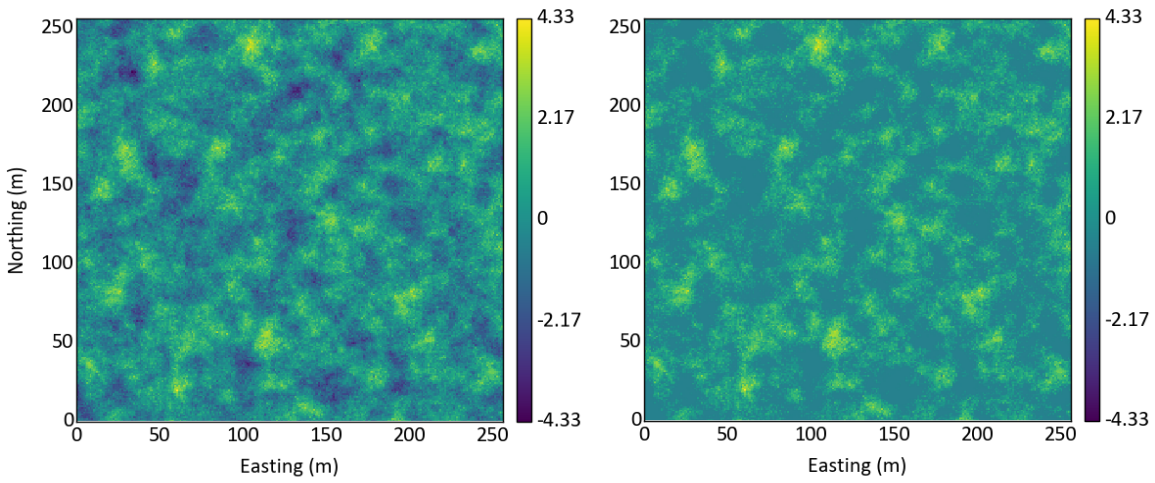

Figure 9. Left: "true" realization created through unconditional simulation. Right: an example of case generated with 30% spike, where the 30% lowest values are given a constant value, to understand the effect of despiking methods in cases with many values at below detection limits.

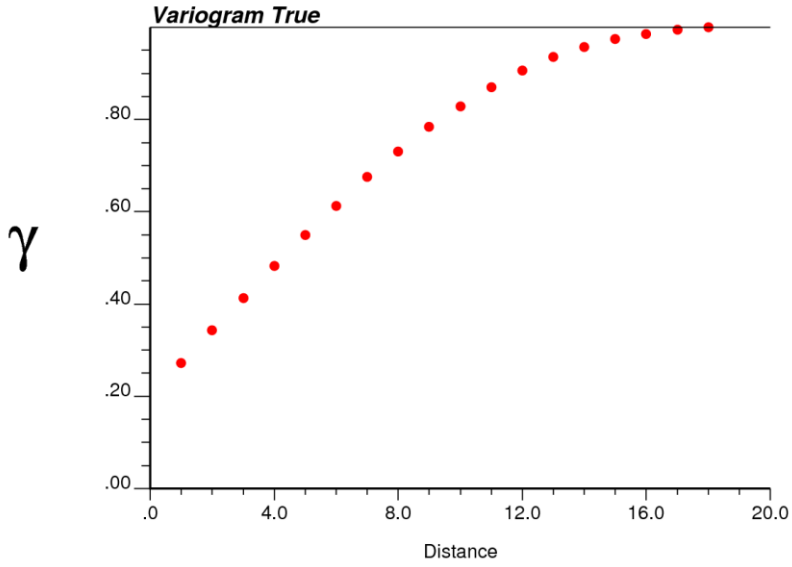

Figure 10. Experimental variogram of the realization considered the "true", which is the base case for generating different cases with different percentages of constant values.

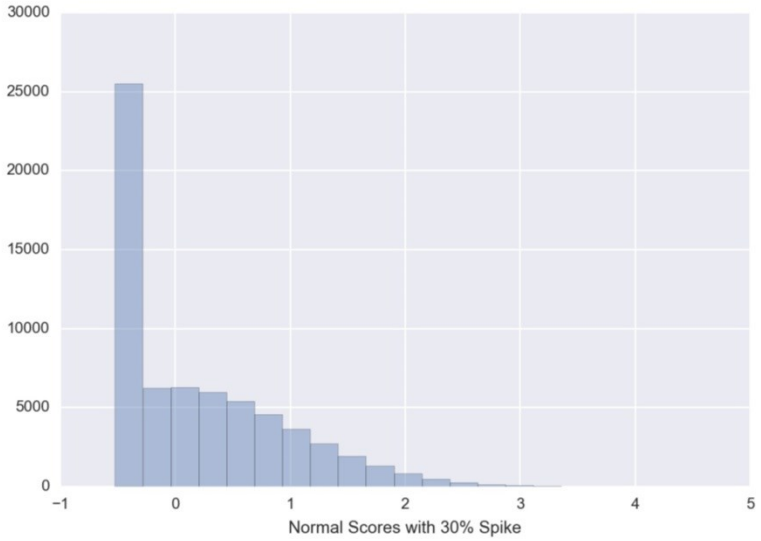

Figure 11. Histogram of realization where the 30% of lowest values are considered constant value (corresponding to the right plot in Figure 9)

The two despiking methods are applied to each case. Then, the variogram is calculated to understand the effect on the spatial continuity.

The effect of different percentages of constant values on the variogram for random despiking and local average despiking is shown in Figure 12, where the experimental variogram is shown for all cases using both despiking methods. The effect of the spike is noticeable on the variogram when the spike is greater than 10%. In the left plot the random despiking increases the short scale variability. In the right plot the local average despiking increases the short scale continuity for short distances relative to the true variogram. This result is also summarized in Figure 13 for the 30% spike case, where it is possible to see that both methods lead to a similar degree of error but in opposite directions, one overestimating the spatial variability and the other underestimating it.

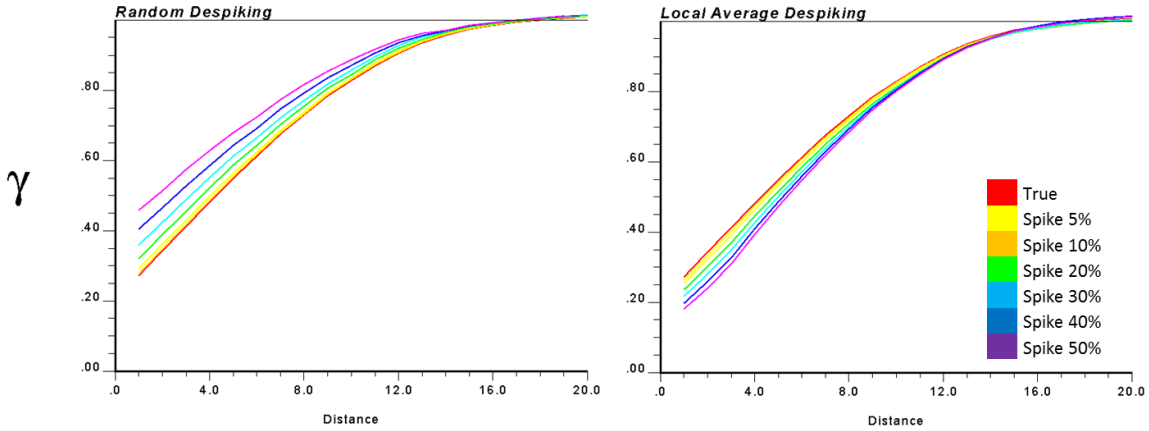

Figure 12. Effect of random despiking (left) and local average despiking (right) on spatial continuity for different percentages of constant values.

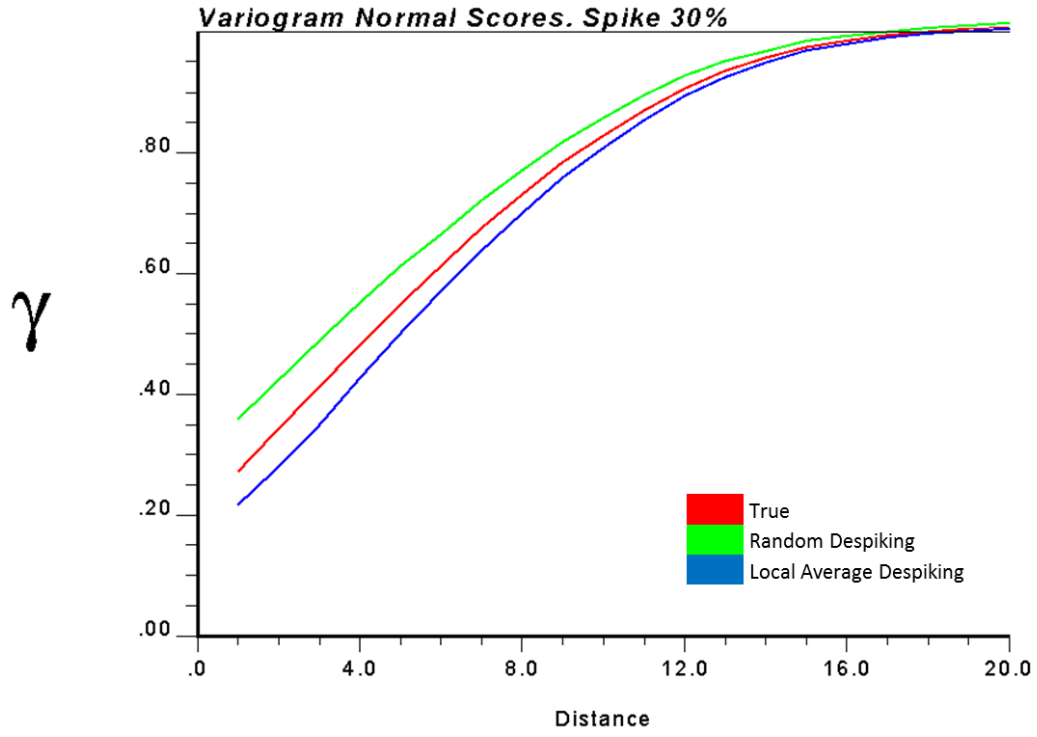

Figure 13. Comparison between true experimental variogram (red) and the ones obtained by applying random despiking (green) and local average despiking (blue) for the 30% spike case.

3.2.2. Case Study 2: Northwest Territories Data

In order to observe the despiking method impact on a case study with a real grade spatial distribution, the stream silt samples from the Northwest Territories are used. The Cu element is selected, since it does not have values at below detection limits in the database, so the true variogram can be calculated as a reference. Then, its distribution is transformed to normal scores, and the 30% lowest values are assigned a constant value equal to the 30th percentile of the original normal scores distribution (-0.524530), as shown in Figure 14.

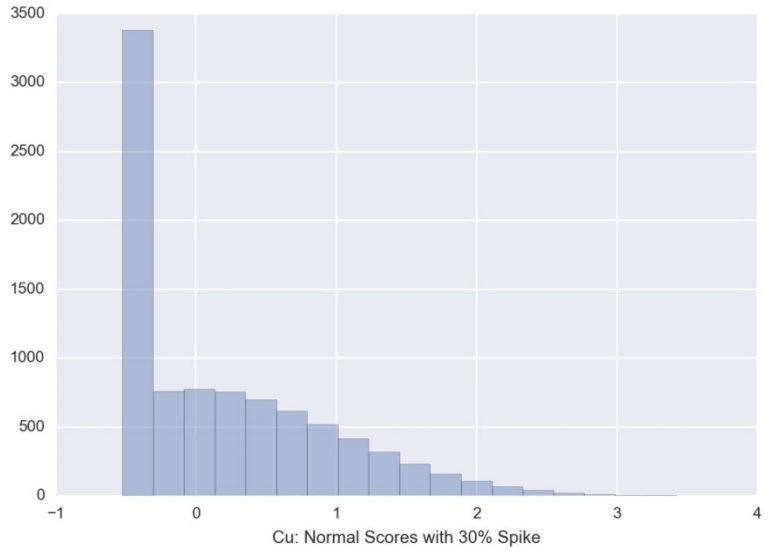

Figure 14. Histogram of the Cu normal scores values for the stream silt samples from the Northwest Territories. To generate a spike the 30% lowest values were assigned a constant value equal to the 30th percentile.

Then, both despiking methods are applied on the samples with the created spike and the result is transformed to normal scores. The experimental variogram obtained from both methods are compared to the true variogram. The data is plotted in Figure 15 and the experimental variograms are shown in Figure 16.

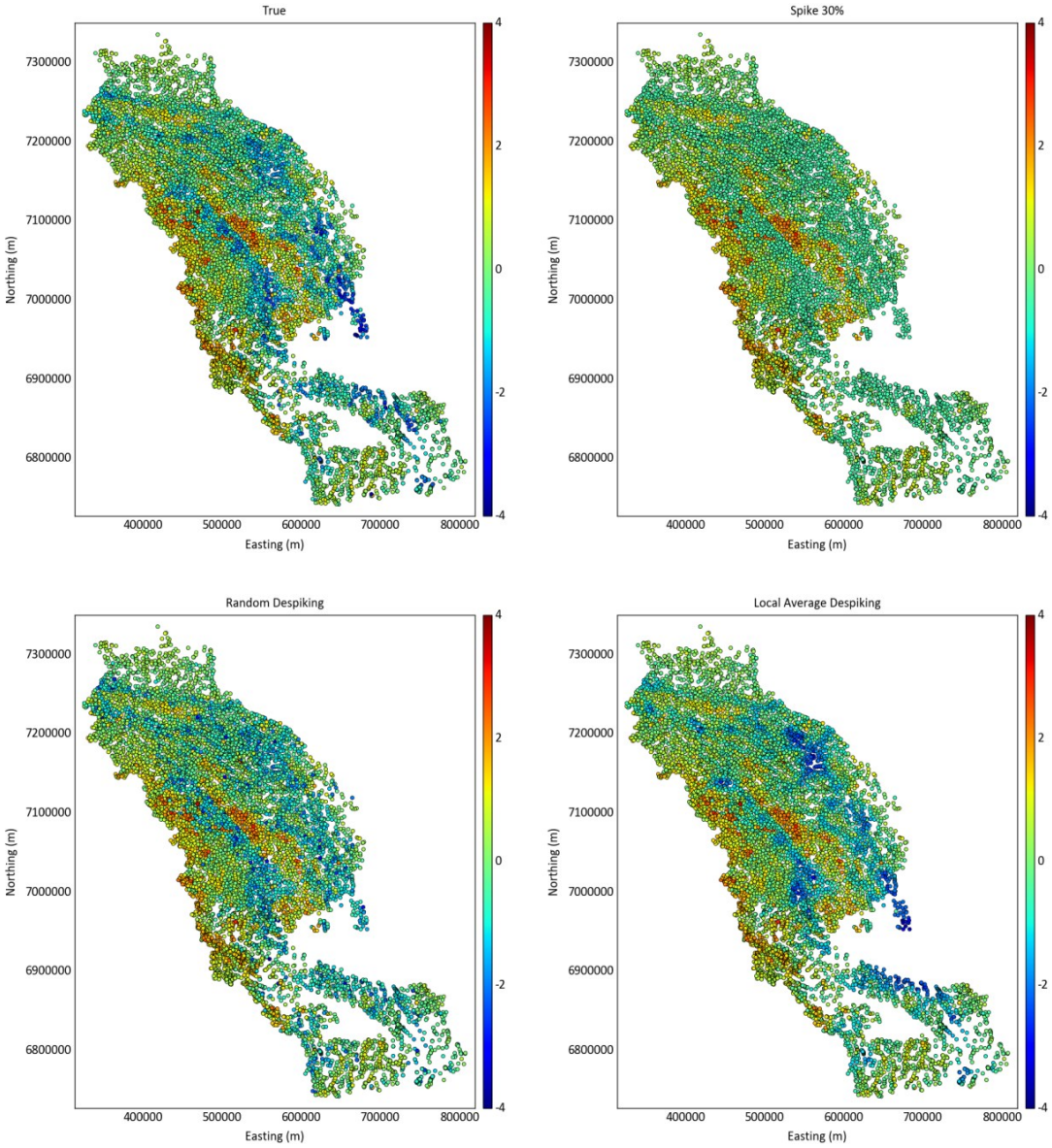

Figure 15. Stream silt samples from the Mackenzie Mountains, Northwest Territories. Top left: True Cu normal scores values. Top right: data with 30% spike created. Bottom left: data obtained using random despiking. Bottom right: data using local average despiking,

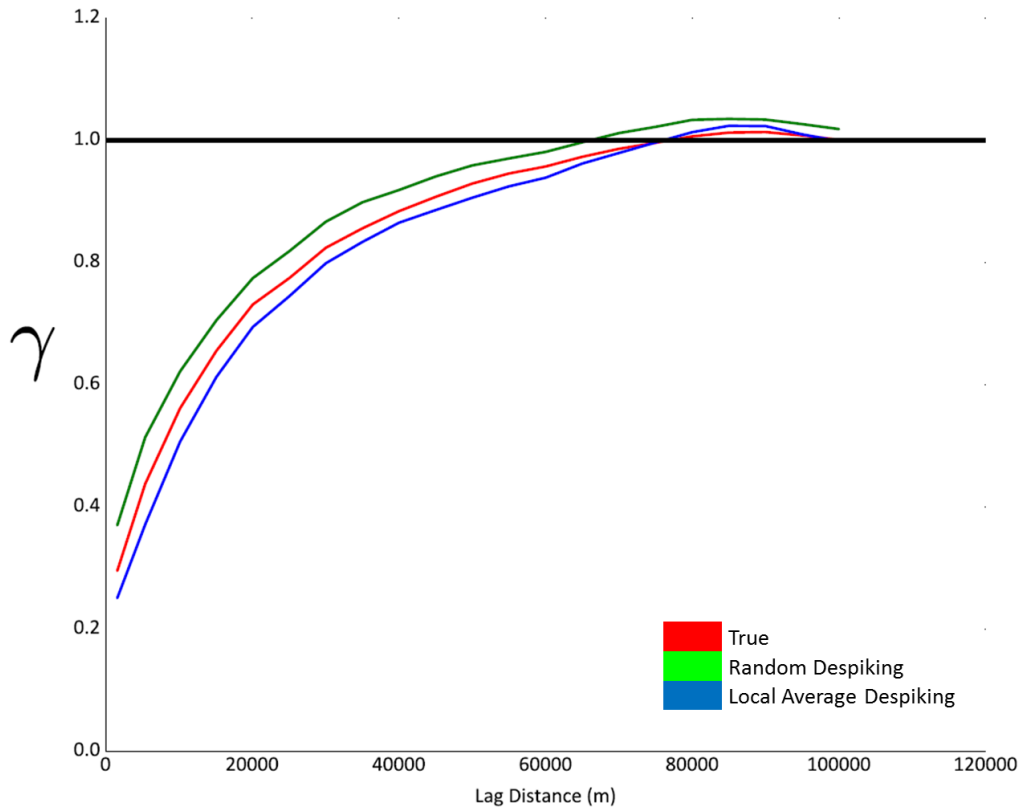

Figure 16. Experimental variograms for true Cu data (red) for the 30% spike data using random despiking (green) and for the 30% spike data using local average despiking (blue).

Both methods lead to an error in the estimation of the true spatial variability, one overestimating the spatial variability and the other overestimating the spatial continuity. These errors in the variogram model would lead to underestimation or overestimation in local uncertainty. To illustrate this effect on uncertainty, simulation is performed using both despiking methods to compare the results. For better visualization, the difference between the variances calculated based on both despiking methods (in normal scores units) for every cell is shown in Figure 17. The difference tends to be positive: the uncertainty estimated using random despiking is greater. This result is expected, since the short scale variability is overestimated when using this despiking method, leading to more uncertainty.

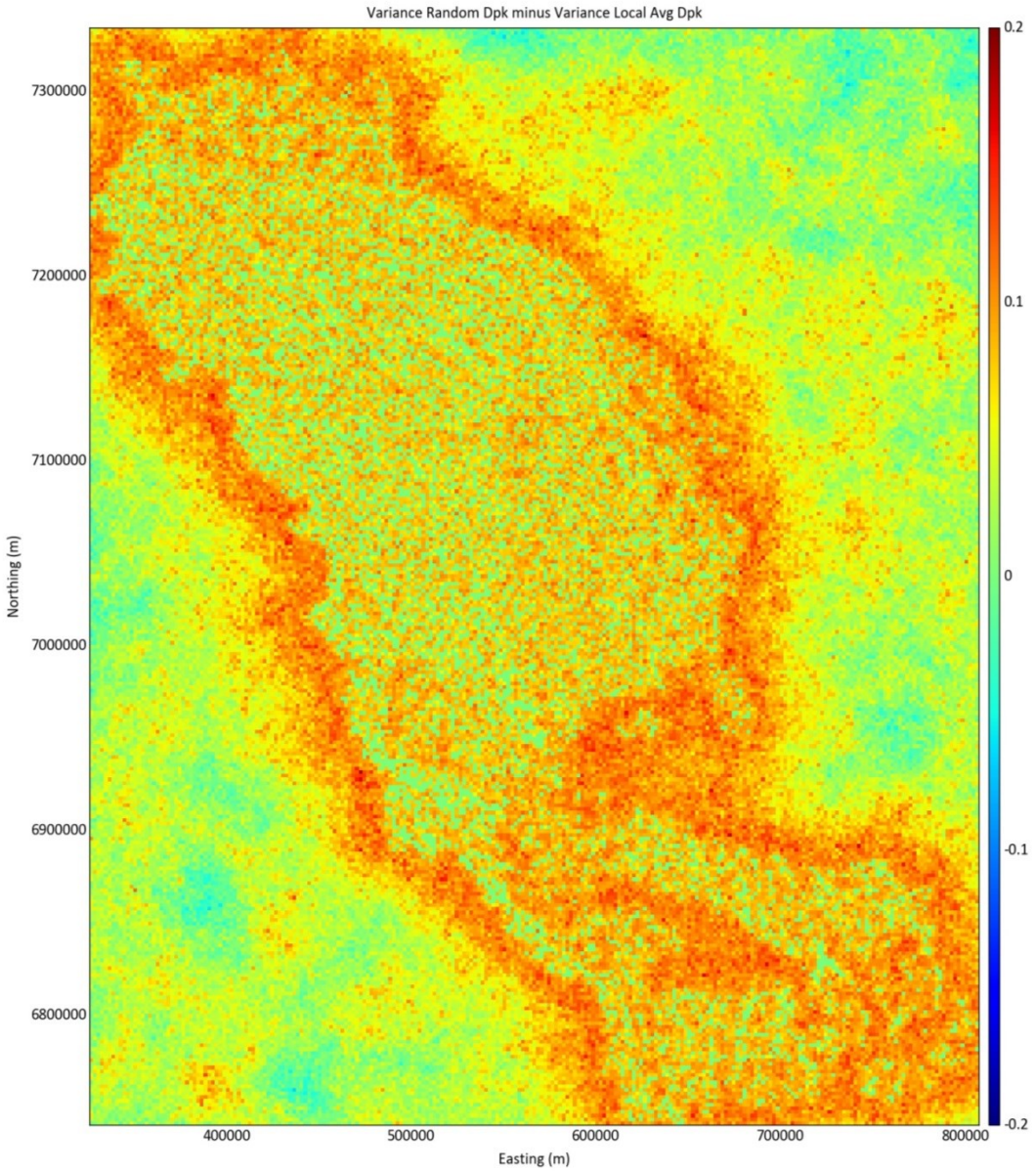

Figure 17. Difference between local variance calculated in normal scores units using random despiking minus the one calculated using local average despiking.

The impact of despiking methods is also seen when back-transforming the realizations. As shown in Figure 18, the different variogram models lead to differences in the average and variance, particularly near the samples with higher Cu values.

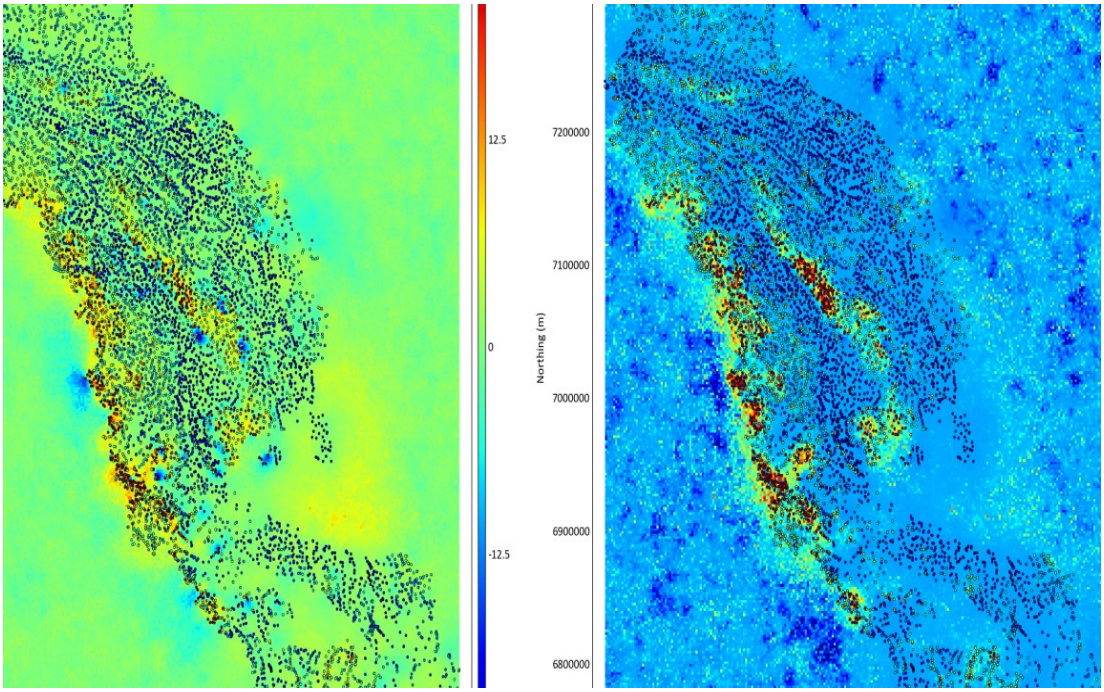

Figure 18. Left: Difference between averages estimated using random despiking and local average despiking. Right: Difference between variances estimated using random despiking and local average despiking. The samples are colored by Cu value but the color scale is not shown (reddish colors are higher values and bluish colors are lower values).

The two common despiking methods produce overestimation or underestimation of the spatial variability, which has an impact on geostatistical analysis. An algorithm to improve the despiking process is proposed.

3.3. Proposed Despiking Method

The true variogram is in between the variogram obtained by using random and local average despiking. Since the former overestimates the spatial variability and the later overestimates the spatial continuity, a despiking method that combines both of them is proposed. The idea is to break the spikes by considering the local average while adding a random component.

The proposed algorithm is summarized by:

1. Find the constant values (spikes). The value of the samples in the spike is X_1 and the next higher value in the distribution is X_2. There are N samples in the spike.

2. Sort the constant values according to their local average.
3. Calculate a constant increment $dinc$ used in step 4 for the local average component:

$$dinc = (1 - W_1) * ((X_2 - X_1)/(N + 1))$$

where W_1 is the weight for the random component, and therefore $(1 - W_1)$ is the weight for the local average component.

4. For every of the *N* data in the spike, calculate the despiked value vr considering a local average component ($dval$) and a random component. For this calculation use the following loop:

do i = 1, N:

$$dval = dval + dinc$$

$$vr(i) = dval + rand * (W_1 * (X_2 - X_1))$$

end do

where *rand* is a random number between 0 and 1.

The performance of this method is tested using the same case studies shown above. In both cases the variogram obtained is very close to the real variogram. A new version of the *despike* CCG software was generated for the proposed algorithm. The parameter file is explained in the Appendix A.

It is important to keep in mind that this method leads to non-unique results due to the random component. A full multiple imputation workflow would have to be considered to pass multiple non-unique data through geostatistical modeling. Another interesting approach could be using the Gibbs sampler algorithm (Geman & Geman, 1984; Silva & Deutsch, 2016) to simulate the values in the spike reproducing the correct spatial correlation. However, this solution requires knowing the correct variogram, which is circular. One possibility is to average the variograms obtained by using random and local average despiking, which was tested and also leads to a good approximation of the true variogram in the case studies. This variogram could be used for the Gibbs sampler to generate multiple realizations, which considers uncertainty in the despiking process. However, the proposed method is a simple and practical solution.

3.3.1. Case Study 1: Synthetic Case

The 30% spike synthetic case is used to demonstrate the performance of the proposed method. The spike is broken using the new *despike* program. The result is shown in Figure 19, where the true experimental variogram is shown in red, the experimental variogram calculated using random despiking (RD) is plotted in green, the experimental variogram calculated using local average despiking (LAD) is shown in blue, and the experimental variogram calculated using the proposed method (RD+LAD) is in black. In this synthetic case, where the spike influence was isolated, it is possible to see a very close match to the true variogram.

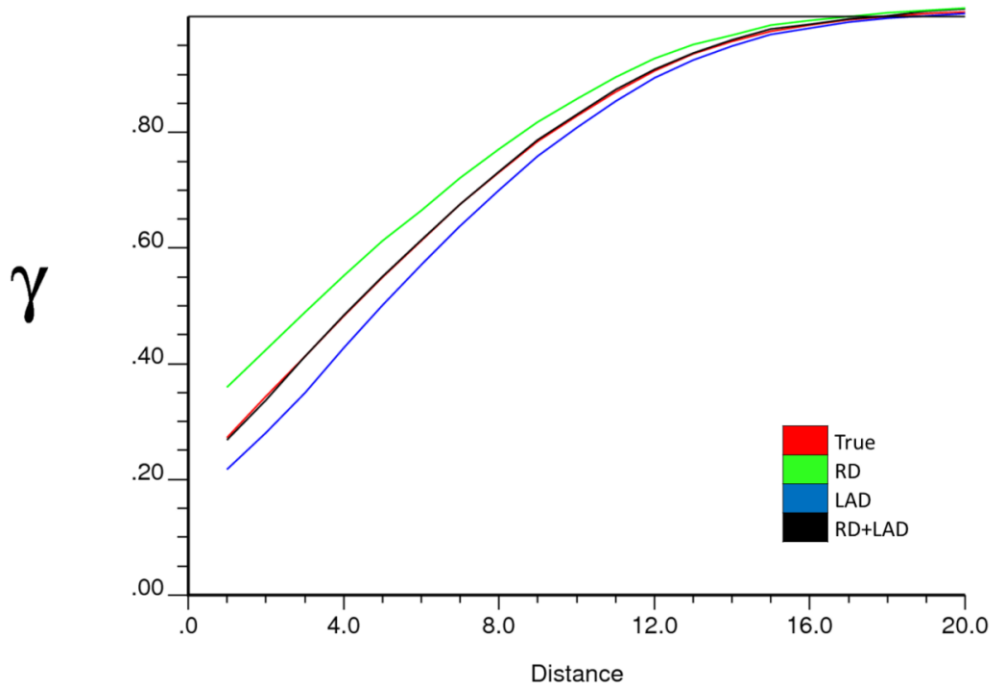

Figure 19. Experimental variograms for the 30% spike synthetic case study. The true experimental variogram is shown in red, the experimental variogram obtained using random despiking (RD) is shown in green, the experimental variogram obtained local average despiking (LAD) is shown in blue, and the experimental variogram using the proposed despiking method (RD+LAV) is shown in black.

Note that the *despike* program allows varying the weight assigned to the random component and the local average component, by varying the weight W_1 in the parameter file. A default value of 0.5 is recommended unless additional information is available that would indicate more or less randomness. An example is shown in Figure 20, where the *despike* program was used to break the spikes

giving more weight to the local average in the left plot ($W_1 = 0.2$) and more to the random component in the right plot ($W_1 = 0.8$).

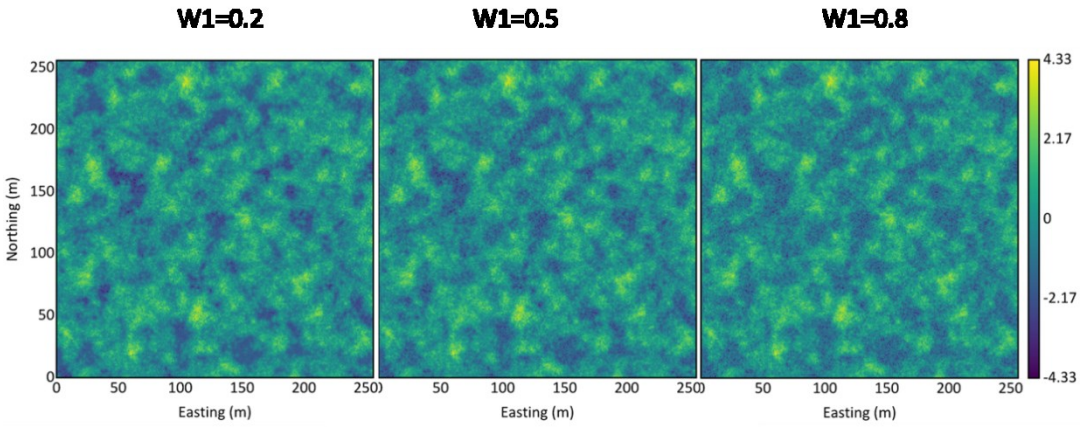

Figure 20. Synthetic case grid colored by normal score value for data despiked using different weights W_1 for the random component. It is possible to see an increment of randomness from the left plot to the right plot in the low grades zone.

As expected, the experimental variogram calculated for the left plot in Figure 20 ($W_1 = 0.2$) is closer to the one calculated using random despiking, while the experimental variogram calculated for the right plot in Figure 20 ($W_1 = 0.8$) is more similar to the one calculated using local average despiking (Figure 21).

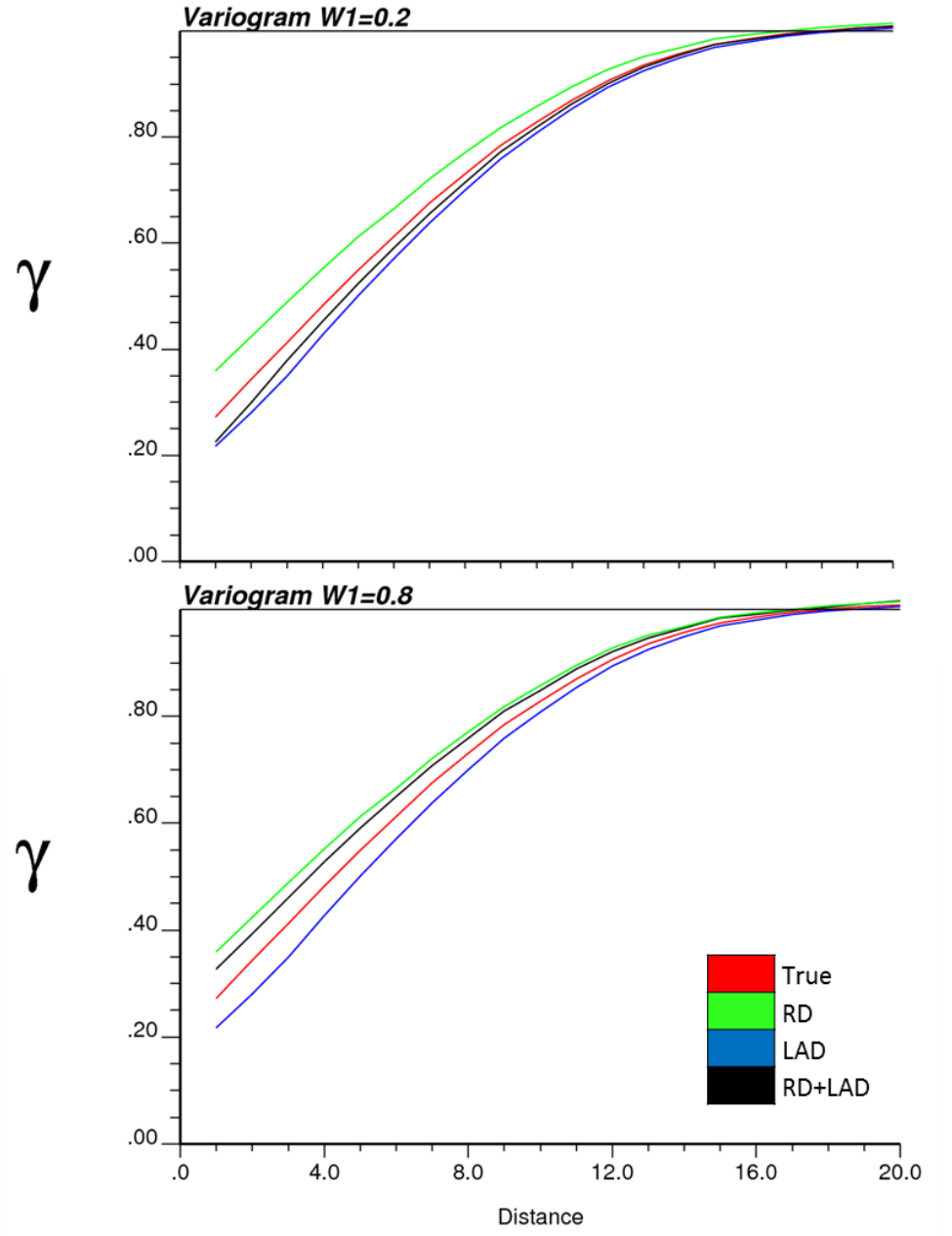

Figure 21. Experimental variograms for the 30% spike synthetic case study. Top: variogram in black calculated using a random component W1 equal to 0.2, so it is more similar to the one calculated using local average despiking (LAD). Bottom: variogram in black calculated using a random component W1 equal to 0.8, being more similar to the one calculated using random despiking (RD).

3.3.2. Case Study 2: Northwest Territories Data

The experimental variogram obtained by despiking using the proposed method (using W_1 equal to 0.5) is shown in black in Figure 22. The proposed method leads to a better approximation of the true variogram.

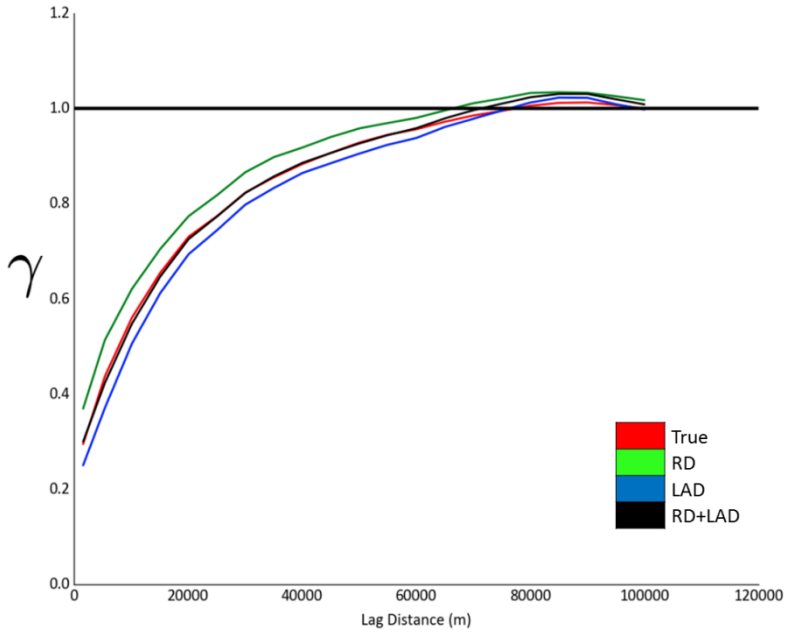

Figure 22. Experimental variograms for stream silt samples Cu values. The true experimental variogram is shown in red, the experimental variogram obtained using random despiking (RD) is shown in green, the experimental variogram obtained local average despiking (LAD) is shown in blue, and the experimental variogram using the despiking method proposed (RD+LAV) is shown in black.

3.4. Conclusion

The effect of the most common despiking methods —random despiking and local average despiking— is investigated. It has been shown that local average despiking and random despiking lead to bias in the observed variogram and predicted uncertainty. A new despiking method is proposed to improve variography when the variable has a significant spike, which combines a random despiking component and a local average despiking component. This method was tested on two different case studies, a synthetic case and one based on the Northwest Territories stream silt samples, showing an improvement in the estimation of the variogram. A new version of the *despike* program is documented in the Appendix A.

Chapter 4. Data Transformation for Cluster Analysis

A critical aspect of clustering is to find groups of samples that are close to each other and far from other clusters in the multivariate space. This is done by distance calculations or by fitting multivariate kernel distributions, where data transformation has a significant impact. The appropriate transformation of the original data units for these calculations is investigated in detail in this chapter. A normal scores transformation preserving spike (NSS) is proposed for improving distance-based cluster analysis. A case study is developed based on stream silt samples from the NWT to illustrate the process of selecting an appropriate transformation for clustering purposes.

4.1. Introduction

4.1.1. Motivation

Cluster analysis mainly depends on three basic choices: the clustering technique, the similarity measure and the magnitude or scale of the different variables (Massart et al., 2001). The third point involves the decision of using original units or transforming data for the analysis. It has been demonstrated that this decision has a significant influence on the clustering performance (Massart et al., 2001; Milligan & Cooper, 1988; Templ et al., 2008). Transformation of variables plays an especially important role for complex geochemical data, since geochemical data is usually highly skewed, multimodal, with presence of spikes —values at below detection limit— and outliers (Templ et al., 2008).

In this chapter different transformation methods are reviewed. The pros and cons of different alternatives are evaluated according to the data characteristics to provide practical recommendations for data transformation depending on the complexities of the data.

4.1.2. Similarity Measures for Clustering Methods

Most clustering techniques are based on "distances" in order to measure how similar or dissimilar are different samples in the data (Templ et al., 2008). A classic way to measure distance is the Euclidean distance. The Euclidean distance between sample i and sample j is defined as:

$$Dij = \sqrt{\sum_{k=1}^{K} \left(X_{ik} - X_{jk}\right)^2}$$

Where X_{ik} and X_{jk} are the values of sample i and j respectively for variable k, and K is the total number of variables. Two popular clustering methods based on distances are Hierarchical Clustering and K-Means (Hastie et al., 2009).

As an alternative, there are clustering methods based on models that fit the data with multivariate distributions (Reynolds, 2009; Templ et al., 2008). A Gaussian Mixture Model (GMM) is a popular model-based or distribution-based clustering method, which is a weighted sum of component Gaussian distributions. A GMM is defined by three parameters for all component Gaussian distributions: the mixture weights, the mean vectors, and the covariance matrices. The idea is to find the parameters that best match the distribution of the data, which can be achieved either by maximum likelihood parameter estimation or maximum a posteriori parameter estimation (Reynolds, 2009). The most common algorithm to fit the GMM is maximum likelihood parameter estimation, which iterates between two steps in order to find the optimized parameters, given the data and the number of clusters —number of components. This iterative procedure is known the expectation-maximization (EM) algorithm, which iterates between computing the maximum-likelihood parameter estimates given the conditional expectation of the Gaussian component (label) assigned to each data instance (M-Step) and then

recalculating this conditional expectation given the parameter estimates computed the M-step (E-Step) (Fraley & Raftery, 1998).

Barnett & Deutsch (2015) provide a description of Hierarchical clustering, K-means and GMM algorithms. The units of the data and possible transformations of the data are important because they change distances and densities; therefore, the results of clustering.

4.2. Data Transformations

The basic idea when transforming data for cluster analysis is to improve clustering performance by affecting the similarity measures, since they are sensitive to the magnitude and variability of the input original variables (Milligan & Cooper, 1988). The alternative transformations commonly considered are explained below.

4.2.1. Standardization Methods

Z-scores

Standardization to z-scores (standard scores) is a common transformation method that measures for each a data instance how above or below it is related to the mean considering the standard deviation of its distribution. Standardized variables have zero mean and unit variance. The equation for this transformation is:

$$Z_1 = \frac{(X - m)}{s}$$

where X is the original data value, m is the variable mean and s is the variable standard deviation. This standardization method preserves the shape of the distribution, translating it to be centered at zero by subtracting the mean, and changing the scale or magnitude of the variable by dividing by the standard deviation. In other words, standardization equilibrates the magnitude of the different variables, while keeping the shape.

Rescaling by Range

Rescaling variable X to a range is performed as follows:

$$Z_2 = (X - Min(X))/(Max(X) - Min(X))$$

where Min(X) and Max(X) are the minimum and maximum values of variable X. This method scales all variables to the range [0, 1], although it is possible to rescale to any range. The mean and standard deviation after rescaling are not the same for the different variables, and they are respectively calculated as follows:

$$\overline{Z_2} = (\bar{X} - Min(X))/(Max(X) - Min(X))$$

$$\sigma_{Z_2}^2 = \sigma_X^2/(Max(X) - Min(X))^2$$

These standardization methods are linear transformations that preserve the shape of the distribution including outliers and spikes. Other non-linear transformations could also be considered.

4.2.2. Normal Scores Transformation

The normal scores transformation is widely used in geostatistics, being a preliminary step for different workflows. The normal scores transformation is a quantile transformation that matches the p-quantile of the variable univariate distribution to the p-quantile of a standard normal distribution (Pyrcz & Deutsch, 2014), so the original data distribution is non-linearly transformed to a Gaussian distribution with zero mean and unit variance (Figure 23).

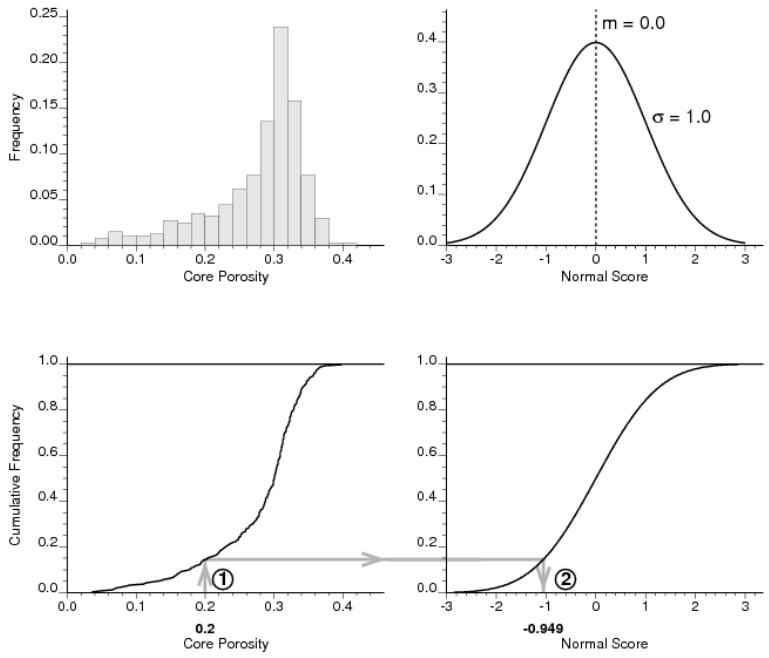

Figure 23. Procedure for transforming original units to normal scores. The original and transformed histograms are show at the top of the figure for core porosity. The cumulative distributions are used for transformation. The procedure for transforming any core porosity value (say 0.2) is the following: 1) read the cumulative frequency corresponding to the porosity, and 2) read the normal score value (-0.949) corresponding to that cumulative frequency on the normal distribution (Pyrcz & Deutsch, 2014).

For distance-based clustering algorithms transforming data to normal scores is considered unrealistic for variables with a large spike, because it distributes the spike along a range of the Gaussian distribution even though they are almost the same value in original units. For illustration, consider a variable with 60% data at or below detection limit. It means that there is 60% data very close to zero in original units, and thus their difference is very low. If we transform to normal scores, these values are going to be spread on the left 60% of a standard normal distribution. Therefore, while one sample at below detection limit can have a value around -3 in normal scores, another value at below detection limit can have a value of +0.25 in normal scores. This implies that the distance between two values at below detection limit is so distorted after transforming to normal scores that in some cases is even larger than the distance between some values at below detection limit and the outliers of the distribution. In order to solve this problem, a normal scores transformation preserving the spike is proposed.

Normal Scores Preserving Spike (NSS)

NSS transformation is proposed with the aim of solving the problem of unrealistic distances between values at below detection limits when transforming to normal scores data with large spikes, to improve distance-based clustering performance. The idea is to transform data to normal scores, and then to assign to every value at below detection limit the highest normal scores value assigned to one of them (Figure 24). A lower value could be assigned to the spike. However, nowadays detection limits are very low (Thompson, 2012). Thus, the difference between the values at below detection limit and the lowest value over detection limit is very small, reason why it is preferred to keep this difference small when assigning a value to the spike in NSS, even though it leads to a mean not equal to zero and a variance not equal to one, which is not a problem for clustering purposes.

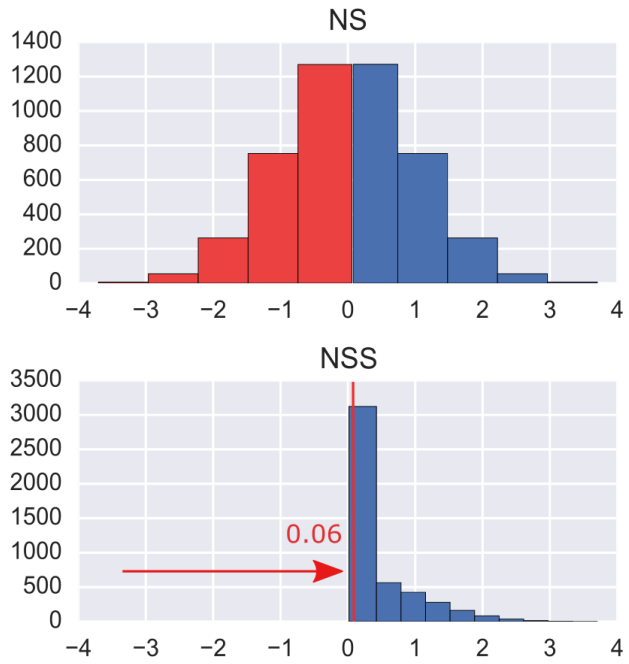

Figure 24. Normal Scores Preserving Spike transformation (NSS). First, the data is transformed to normal scores (top). In this example there is around 50% of data at below detection limit, which is spread in the red part of the normal scores distribution shown in the top of the figure. Then, all data at below detection limit is assigned to the highest normal scores value assigned to a sample at below detection limit —in this case the highest normal score value assigned to a sample on the spike was 0.06 (bottom).

The NSS transformation allows solving the problem caused by outliers and can reduce the problem caused by skewness, while keeping a more realistic distance between samples at below detection limits.

4.3. Data Complexities for Clustering Analysis

The purpose of this section is to explain the possible data complexities that impair clustering algorithms performances. In order to demonstrate the impact of the complexities, some explanatory examples are illustrated. Three elements are used for the examples: calcium (Ca) magnesium (Mg) and boron (B). Both Ca and Mg present bimodal distributions, while B presents a large spike —around 50% of the data at below detection limit — and also outliers (Figure 25).

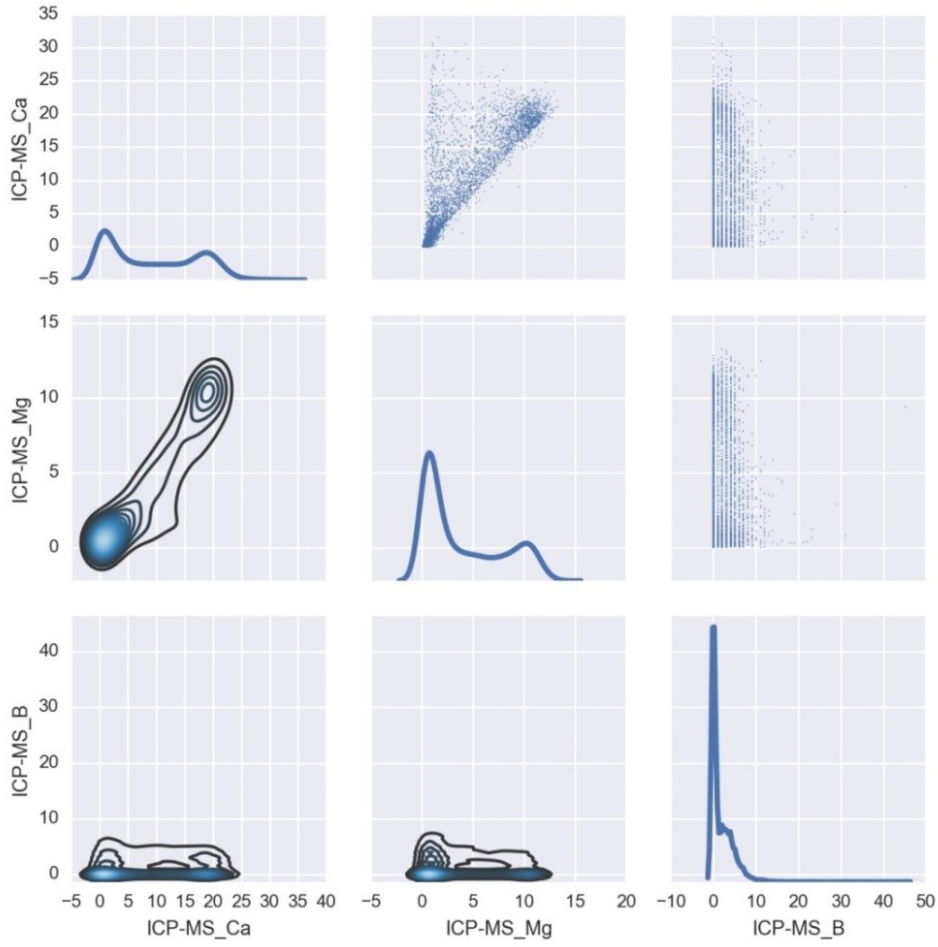

Figure 25. Univariate and bivariate distributions Ca, Mg and B. The plots in the diagonal are the univariate distributions. The plots below the diagonal are bivariate kernel density plots and the plots above the diagonal are bivariate scatterplots (the stripes in scatter plots are due to limited precision in the measurement of B).

4.3.1. Outliers

Outliers are anomalous high (or low) values that have a significant effect on the performance of clustering methods. The distance of outliers to the distribution is very large compared to the distance between the lower values of the distribution, which makes them seem excessively dissimilar for clustering. Accordingly, clustering algorithms do not work well in presence of outliers (Barnett & Deutsch, 2015; Milligan & Cooper, 1988).

Outliers also have an important influence on the effect of some transformations for cluster analysis. Z-scores standardization and rescaling by the range can make low values appear very similar, reducing the overall influence of the difference between those values —and thus, the influence of the variable— when performing distance-based clustering methods, while making outliers seem too dissimilar. This effect is illustrated in Figure 26, which shows the Hierarchical clustering dendogram based on Ca, Mg and B. In this example, Agglomerative Hierarchical clustering method was run in order to determine two clusters (A and B). It is possible to see that Boron outliers (variable Z_3 in the figure) are comparatively so large, that they are assigned to cluster A, while all other samples are assigned to cluster B.

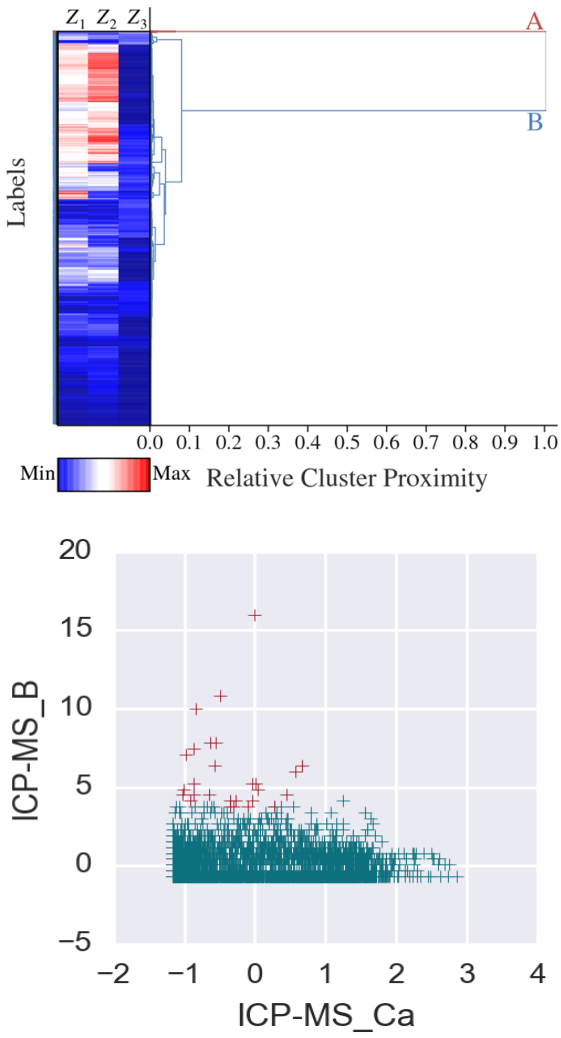

Figure 26. The upper image corresponds to the Hierarchical clustering dendogram for variables Ca, Mg and B (Z_1, Z_2 and Z_3 respectively). The colors on the left of the dendogram represent the values of the three elements. The lines on the right represent the distances between all possible clusters. The clustering method was used to select two clusters. Cluster A is composed by very few samples: as shown in the bottom image they correspond to the variable B outliers.

Even though both, z-scores standardization and rescaling by range suffer with the presence of outliers, the impact is lower when rescaling by range as shown by G. Milligan & Cooper (1988), because the impact of high values is reduced in comparison to the influence of other variables. This is illustrated in Figure 27, where it is possible to see that even though Hierarchical clustering based on both standardizations methods separate outliers in cluster A, the cluster B is better separated when rescaling by the range, giving more influence to variables Z_1 and Z_2 (Ca and Mg), while standard scores still gives too much importance to the high values of variable Z_3 (B).

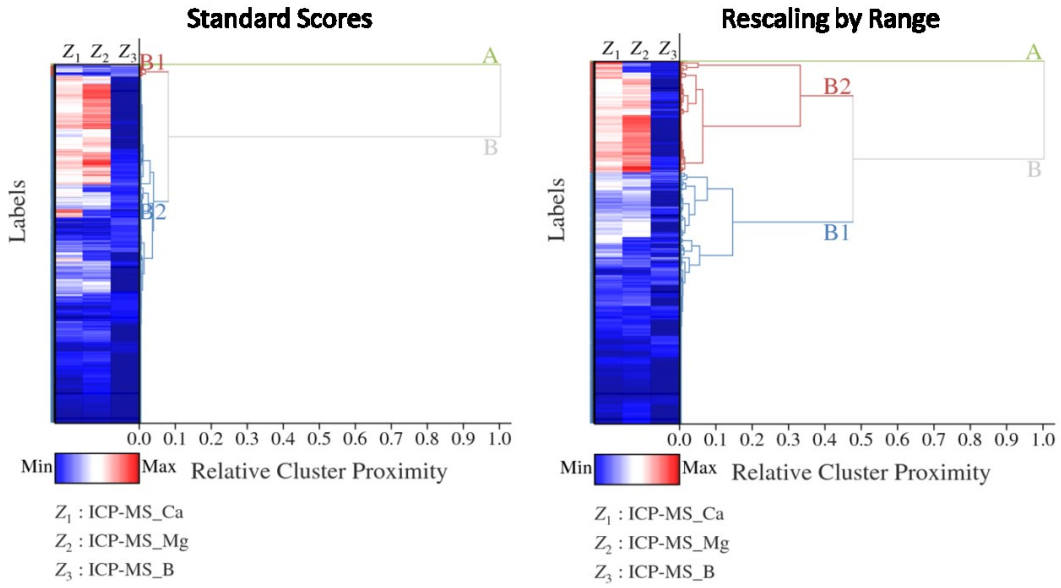

Figure 27. Hierarchical clustering dendograms based on two different transformations: standard scores (left) and rescaling by range (right). Cluster B is better separated when recalling by the range, since the separation is based on variables Z_1 and Z_2 (Ca and Mg), while when based on standard scores too much importance is given to high values of variable Z_3 (B).

The non-linear normal scores transformations alleviate the problem caused by outliers, by reducing the distance of outliers to the distribution, which is useful when the purpose is to find large structures (large clusters) on data.

4.3.2. Skewness

The main problem of highly skewed distributions is that the distances between high values compared to the distance between low values may be disproportionately large, while the distance between low values may seem exaggeratedly small with the Euclidean distances used in distance-based clustering methods. On the other hand, distribution-based clustering methods are also affected by highly skewed distributions because the density around low values may seem excessively high compared to the density around high values. The normal score transformation solves this problem by eliminating skewness when transforming the distribution to standard normal. However, there is the possibility that preserving the data distribution could lead to more realistic results, since they are based on the real similarity/dissimilarity between data values.

4.3.3. Spikes

Spikes can have a significant impact on distribution-based clustering methods like GMM. Large spikes have high density, and this high density can make the variable more important than the other variables used in clustering. This impact is shown in Figure 28, where the difference between GMM clustering results based on three elements —Ca, Mg and B— transformed to normal scores (on the left) is compared to the clusters assigned when B is transformed to normal scores but its spike is preserved (on the right). When B is also transformed to normal scores (left) the separation is dominated by the two Ca-Mg visual clusters observed in original units (Figure 25), but when B spike is preserved it dominates the clustering assignment: there are no red cluster samples outside the spike. In this example, the spike in B makes it appear more important than Ca and Mg for the cluster assignment, even though Ca and Mg may be more important from a geologic point of view. On the other hand, the influence and treatment of spikes is different for distance-based clustering methods, where it may be more realistic preserving the spike to avoid distorting the Euclidean distance between values at below detection limit.

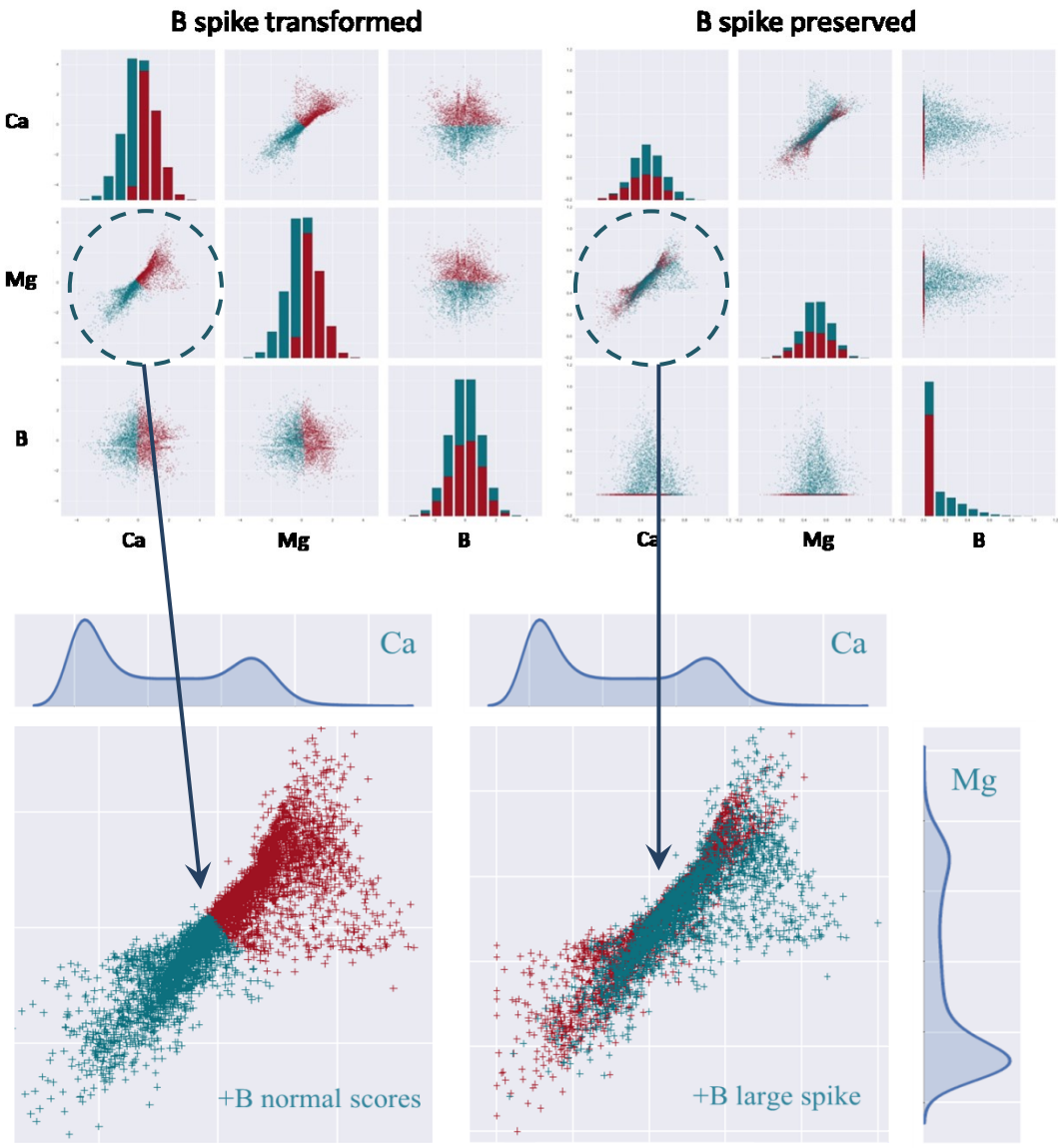

Figure 28. Univariate and bivariate distributions colored by the two clusters assigned by GMM based on two scenarios: 1) Ca, Mg and B are transformed to normal scores removing variable B spike (on the left), and 2) Ca, Mg and B are transformed to normal scores but B spike is preserved (on the right). Note that when the spike is preserved its influence is so important for the density-based clustering method, that the spike dominates the separation and the two Ca-Mg clusters seen in original units are not separated anymore as seen in the dashed circle.

4.3.4. Multimodal Distributions

Geochemical measurements may show multimodal distributions. For instance, Ca and Mg are two important silt sediment variables for differentiating lithologies, and they both show bimodal distributions. The normal scores transformation transforms a multimodal distribution into a unimodal one (Figure 29). This may hinder the ability of clustering algorithms to separate samples corresponding to different populations represented by the different modes. However, it is important to consider that even when transforming variables to normal scores the modes may be identified in higher dimensions by clustering methods.

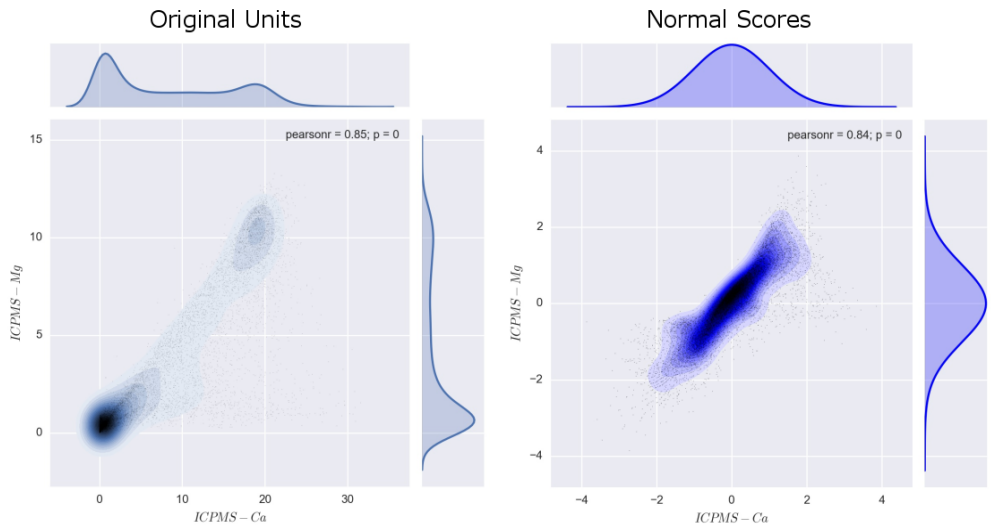

Figure 29. Ca-Mg bivariate —bimodal— distribution in original units (left) versus bivariate distribution in normal scores (right).

4.4. Recommendations for Transformation Decisions

As illustrated above, there is no a single transformation that works well under all possible data complexities and all possible clustering methods. Each of the transformations analyzed has pros and cons. They may alleviate some problems and cause concerns in other conditions. The following summarizes some points to consider for choosing the correct transformation under different circumstances. The focus here is to find large structures in data. Therefore, some of these concepts should be applied differently when the aim is to detect anomalies, as seen in the next chapter.

Distance-based Clustering Methods:

- Outliers: transforming data containing outliers to normal scores is an effective alternative and it avoids manual outlier management, which is time consuming when working with data with many variables. If cluster analysis is performed in original units, outlier management is recommended as a preliminary step. Outlier management is also important if data are going to be standardized to z-scores or rescaled by the range before clustering, since these transformations do not appear to work well in presence of anomalous high values.
- Skewness: normal scores transformation solves problems caused by skewed data by removing it. However, if preserving the distribution shape is considered important, rescaling by the range is recommended rather than standardization, since the former reduces the influence of high values on the multivariate distances while keeping the shape. This is in accordance to the conclusions obtained by Milligan & Cooper (1988) when comparing different standardization methods for cluster analysis.
- Spikes: preserving spikes is a realistic choice for distance-based cluster analysis. Accordingly, if data is transformed to normal scores it is a good option to preserve the spike by considering the NSS transformation approach.
- Multimodal distributions: transforming data to normal scores does not seem reasonable in presence of a multimodal distribution. However, if normal scores transformation is required to alleviate other complexities, it is suggested to proceed with normal scores and to check that the clustering method is still recognizing the different modes.

Distribution-based Clustering Methods:

- Outliers: data outliers appear to cause the GMM algorithm difficulty in finding the optimal Gaussian components that fit the multivariate distribution. Accordingly, if data is transformed to normal scores, the higher variability introduced by outliers does not have to be explained by

the Gaussian components, and thus, the clustering method can focus on understanding the multivariate relations of data without giving excessive importance to the anomalous values.

- Skewness: in presence of skewness the Gaussian components fitted by the GMM algorithm are affected; ideally, the algorithm should use a limited number of Gaussian components to explain the multivariate variability. Consequently, in presence of highly skewed data the normal scores transformation should be considered, which will help equalize the weight of each variable in the choice of optimal components, means and covariance matrices.
- Spikes: spikes influence the choice of Gaussian components for GMM, since the algorithm can give too much weight to the variable with a spike in the fitting process. Therefore, transforming to normal scores can be a good idea despite of the fact that it is introducing artificial distances between data.
- Multimodal distributions: in order to make easier for the distribution-based algorithm to recognize the zones with more density in the multivariate space, it is a good idea to keep the original shapes of multimodal distributions. However, consider that even though normal scores transformation removes multiple modes in the univariate space, it is possible that they are still recognizable in the multivariate space, which could be partially checked using scatter plots.

A multivariate dataset will contain a combination of the complexities discussed above, and therefore the decision about the optimal transformation for every data requires balancing its pros and cons according to the different complexities observed. Finally, a sensitivity analysis performing a combination of transformations could be a good alternative, for instance transforming data to normal scores in case of highly skewed data, but just rescaling by the range [-4,4] the multimodal variables.

4.5. Case Study

In this case study the aim is to use clustering for separating data into geochemical populations related to geological units identified in the Mackenzie Mountains stream sediment samples. Different lithologies and alterations contain different distributions of elements. The expected Cu value in a plutonic felsic rock is not the same as in a mafic volcanic rock, and accordingly, what is considered an outlier in the former may not be an outlier in the latter. For this and other reasons —like defining stationary domains for geostatistics— it is often required to separate different geological populations into geochemically similar groups.

GMM and Hierarchical clustering were applied in order to separate the silt sediment samples into similar geochemical populations. Given the complexities observed in this data, which contain outliers, highly skewed distributions, spikes and multimodal distributions, a normal scores transformation was applied as a preliminary step since it solves most of them. For applying GMM the data was despiked and transformed to normal scores. For applying Hierarchical clustering it was transformed to NSS to get a more correct distance calculation between samples at below detection limits. It was also tested for GMM rescaling by the range [-3.5, 3.5] the multimodal variables (to keep the modes in those univariate distributions) and transforming the other variables to normal scores, but it did not improve the result, since in this case the modes were still recognized by GMM in the multivariate space when transforming to normal scores.

The next step was to select the number of populations or clusters, since it is a required input for most clustering methods. This is a challenging task when performing clustering, because there is no single measure that indicates unambiguously the number of clusters for all types of data, requiring experimentation and critical review of the results. Besides, a different number of clusters can be used for different purposes: if the purpose is to find anomalies, then a large number of clusters may need to be specified. The purpose here is to separate data into relatively few large scale populations, similar to the 8 large scale

geological units mapped in the area (Ootes et al., 2013): Proterozoic Epicratonic Basin (Mackenzie Mountains Supergroup), Neoproterozoic Extension and Rift-related Successions (Windermere Supergroup), Lower Paleozoic Mackenzie Platform, Lower Paleozoic Selwyn Basin, Upper Paleozoic Siliciclastic Basin, Upper Paleozoic Siliciclastic/Carbonate Shelf, Cretaceous Intrusions and Mesozoic Foreland Basin.

There are some methods and measurements that can give an idea of the number of clusters contained in the data. Some measurements evaluate clustering performance without requiring that the ground truth classes are known, like the Silhouette Coefficient (Rousseeuw, 1987) that can be used for orientating the clustering method to find better defined clusters. The embedding method t-SNE can also be useful in such circumstances, which has been applied effectively for a variety of datasets as a tool to visualize the data structures and to observe the natural amount of clusters on data (L. Van Der Maaten, 2009; L. J. P. Van Der Maaten & Hinton, 2008). When the ground true class is known for each sample, some evaluations are used to measure the similarity between two assignments, like the Rand Index or Adjusted Rand Index (Hubert & Arabie, 1985). The Hierarchical clustering dendogram is another useful tool for understanding the natural clusters in data, which is based on Euclidean distances. In this graph, each row is a sample and each column is a variable. The color that varies from blue to red represents the values of every sample for every variable. On the right it has a dendogram (tree) that shows how similar are groups of data to each other. The longer the distance between a split point to the next one, the larger the Euclidean distance between the clusters. Figure 30 shows the dendogram for the stream silt data. It is possible to see that from the red dashed line to the right —8 clusters or less— the data is separated into fairly different clusters. But if we move the dashed line to the left to more than 8 clusters, the clusters become similar in the multivariate space.

Some of these methods were considered together with the general understanding of the geological units of the area in order to select a range of possible number of clusters, and the final decision was made according to the result that provided the best match with the geologic map. This leads to a result that is a good compromise between mathematics-computer science and professional judgement to find patterns and group natural processes. It was considered reasonable to try between 4 to 10 large scale clusters.

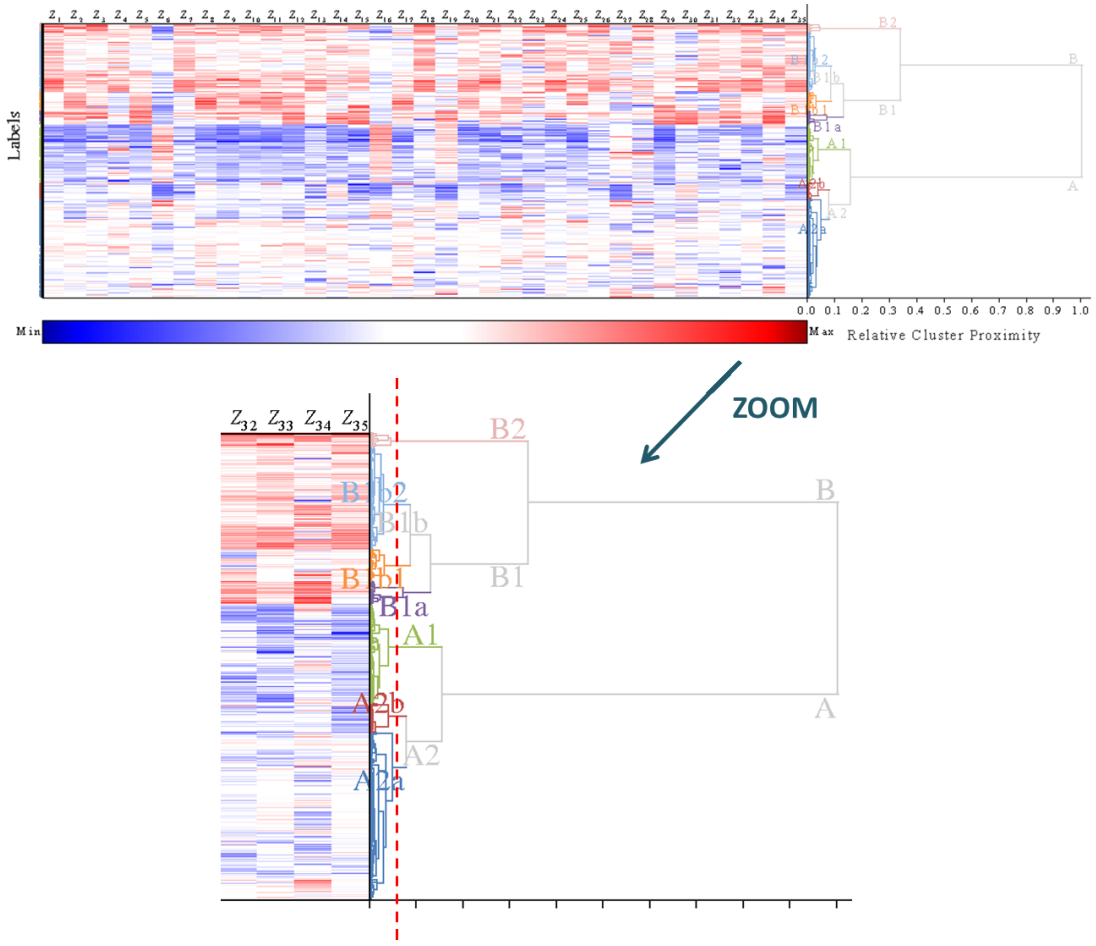

Figure 30. Hierarchical clustering dendogram based on NSS transformation of data. The red dashed line seems to separate data into fairly different clusters. If we move the line to the right the clusters become more different from each other. On the contrary, if we move the dashed line to the left (more clusters) the clusters become too similar to each other in the multivariate space.

Consequently, GMM and Hierarchical clustering are applied on the stream silt sediments data considering 4 to 10 clusters. Other clustering methods like K-means could also be considered. Given that in this case study the goal is separating

data into geochemical populations related to geological units identified in the area, GMM applied on data transformed to normal scores and assigning 8 clusters is visually considered to give the best match between clustering and the geologic map. The result is shown in Figure 31, where a geologic map of the Mackenzie Mountains is shown with the samples plotted and colored according to the different clusters assigned. It is possible to see a good match between the geological units and the clusters. One possible application of this result is to identify areas where there are differences between the map and the clustering output in order to check the geological mapping and the data values, for example see some highlighted in Figure 32.

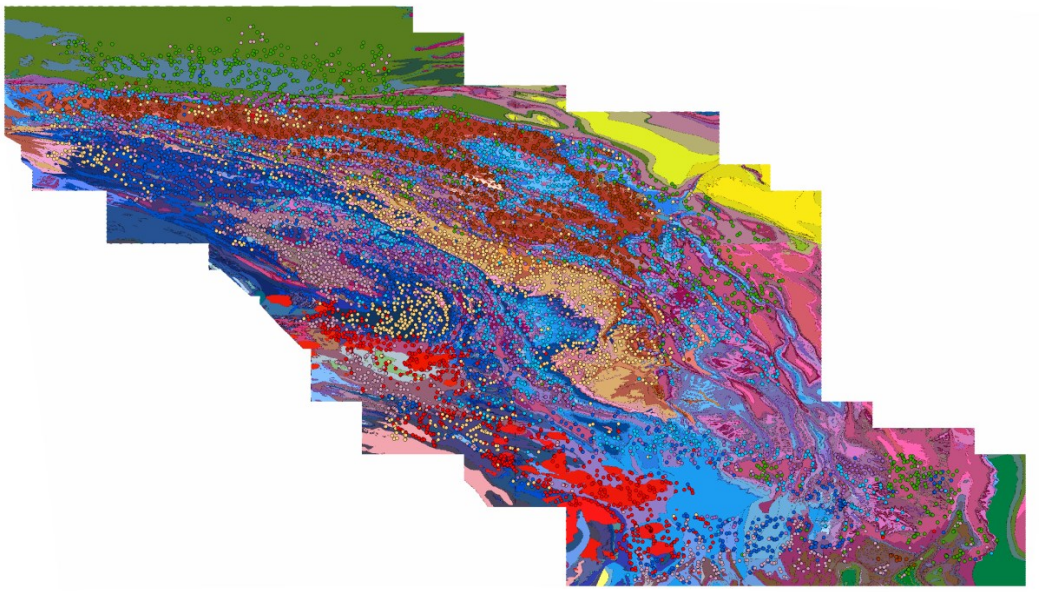

Figure 31. Geologic map of the Mackenzie Mountains (provided by the Northwest Territories Geological Survey) with the silt sediments samples plotted and colored by clusters assigned by GMM method on data transformed to normal scores.

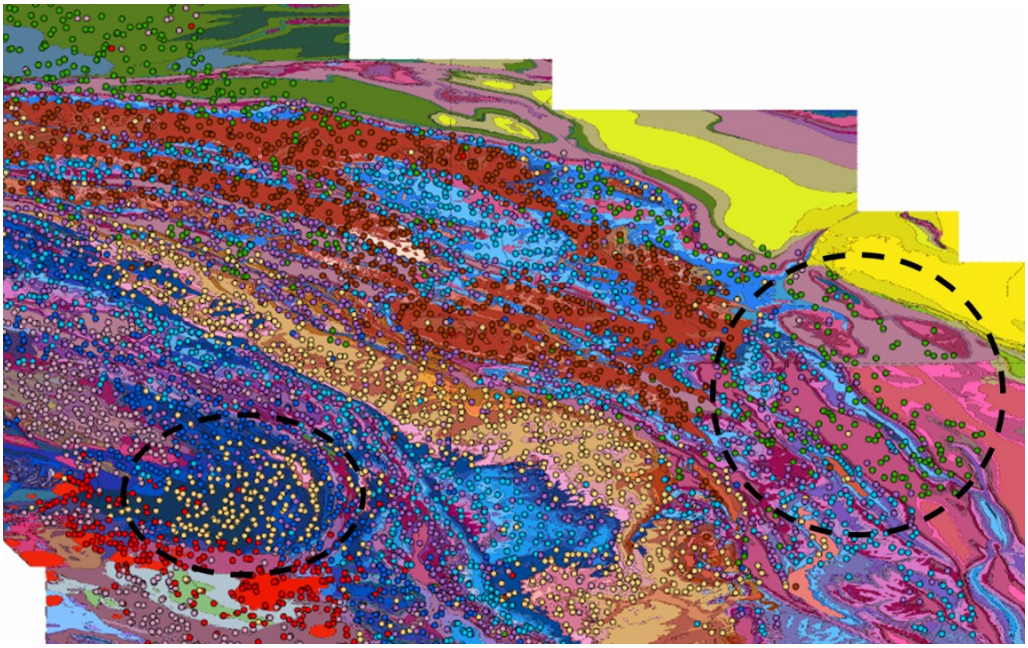

Figure 32. Geologic map of the Mackenzie Mountains with the silt sediments samples colored by cluster, highlighting areas where there are major differences between the map and the clustering result.

4.6. Conclusion

Data transformations have a significant impact on clustering results. Accordingly, an understanding of the impact that different transformations have on clustering is essential for its correct application. Some common data transformations have been reviewed, explaining their effects on clustering and pointing out the data complexities they alleviate for improving clustering performance. This chapter provides guidance for finding large structure in data and for detecting anomalous multivariate data instances, which is developed in the next chapter.

Chapter 5. Spatial Anomaly Detection with Multivariate Data

Three methods for multivariate anomaly detection are proposed in order to identify multivariate samples that are anomalous in the multivariate space for mineral deposit exploration. The first method uses different combinations of clustering and data transformations for finding small anomalous clusters; the second uses different clustering outputs for identifying samples that do not belong clearly to any cluster; the third recognizes samples that are spatially anomalous, that are surrounded by samples assigned to a different cluster. The method is applied to stream sediment samples from the Northwest Territories for illustration. The multivariate methods are capable of recognizing deposits that are not identified in the univariate space.

5.1. Introduction

Anomaly detection has played an important role in many different areas like fraud detection, cyber-security, health care, military surveillance, finance and law enforcement (Akoglu, Tong, & Koutra, 2015; Chandola, Banerjee, & Kumar, 2009). Anomaly detection can be defined as the identification of data instances or samples with a patterns that do not conform to expected behavior (Chandola et al., 2009). A natural application in geoscience is for exploration purposes: to find the locations where there are anomalous concentrations of ore minerals on the earth.

The goal is to develop new methods to detect multivariate anomalies, going beyond the anomalies recognizable from individual measurements. For this purpose, different algorithms are combined to identify multivariate anomalous samples that provide guidance to define interesting locations for exploration, detecting anomalies in multivariate space.

The method for multivariate anomaly detection proposed here is based on data transformations and clustering techniques, discussed in the previous chapter. Three methods for multivariate anomaly detection are combined:

- Small anomalous cluster detection: using the appropriate transformations and clustering methods for finding small clusters that correspond to an anomalous group of samples with similar patterns.
- Lack of Uniform Clustering Classification (LUCC): using different clustering techniques and parameters for detecting samples that are far from clusters centroids in the multivariate space. The idea is that since anomalous samples are far from the cluster centroids, they are not clearly part of a cluster, and therefore, there is disagreement between different clustering methods applied on different data transformations.
- Spatial anomalies: identifying samples that are different from the surrounding samples in the geographic space.

These methods are explained below and a case study is shown based on stream sediment samples from the Canadian North West Territories.

5.2. Multivariate Techniques for Anomaly Detection

5.2.1. Small Anomalous Clusters

The idea of using cluster analysis for finding small anomalous clusters for exploration purposes is not new (Cohen et al., 2010; Garrett & Grunsky, 2001). The contribution here is to provide some guidance and propose a way to use different data transformations and clustering methods to identify multivariate anomalies from different point of views.

Three decisions have a major impact when performing clustering for anomaly detection: data transformation, the clustering technique and the number of clusters to be identified. Others parameters, such as the linkage calculation, also have some impact and can be varied for sensitivity analysis. For detecting anomalies it is

necessary to find the combination of parameters that lead to the identification of small clusters with unique patterns. The basic idea is that samples related to ore deposits have some similar patterns that make them different —far in multivariate space— from the common patterns, so they can be detected as a different cluster. This is illustrated in Figure 33, where the green cluster represents a small anomalous cluster.

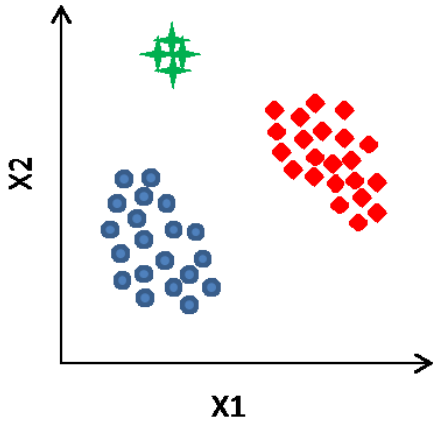

Figure 33. Synthetic bivariate case for illustrating the idea of identifying small anomalous clusters like the one symbolized with green stars. X1 and X2 represents two pathfinder elements.

As shown in the previous chapter, data transformations that keep the shape of the distribution —like standardizing or rescaling by the range— can make low values seem closer, while outliers seem very far in the multivariate space, influencing the clustering methods to separate small anomalous clusters containing samples with the highest distances to the clusters centroids. On the other hand, transforming the shape of the data distribution to normal scores is also recommended, because it adds another point of view for detecting anomalies, based more on multivariate position. Normal scores transformation can potentially detect different anomalies because it reduces the influence of those samples that are very far from cluster centroids to separate them into different groups, allowing clustering methods focusing more in the patterns of samples for dividing data into clusters.

Different clustering methods allow observing data and looking for anomalies from different perspectives, since they are based on different algorithms and similarity measurements. Additionally, different target numbers of clusters should be tried,

because a small number of clusters may only lead to the identification of large structures in data, which is inevitably subjective and requires experimentation.

Finally, once different anomalous samples have been detected by trying different combinations of data transformations, clustering techniques and number of clusters, they can be filtered by computing the mean of pathfinder elements for each small anomalous cluster, discarding the samples corresponding to clusters that do not match candidate geochemical signatures.

5.2.2. Lack of Uniform Clustering Classification (LUCC) Anomalies

Multivariate anomalous samples are likely far from cluster centroids in multivariate space. Consequently, applying different clustering configurations on different data transformations will probably lead to disagreement when assigning these samples to clusters. There will be more agreement between different clustering methods/data transformation combinations for samples that are very close together in the multivariate space. Based on this concept, a novel algorithm is proposed for finding anomalous. The idea is to compute a *lack of uniform clustering classification* (LUCC) measure for each sample, and then identify the outliers of the LUCC distribution as anomalous samples. The algorithm proceeds as follows:

1. Use different clustering techniques, linkage calculations, target numbers of clusters and other reasonable variations of clustering parameters for obtaining multiple clustering outputs. The target numbers of clusters should be kept relatively small to identify the samples that are not clearly part of large clusters.

   The following notation is considered:
   - nd: number of data
   - $nc(k)$: target number of clusters for clustering output k.
   - K: total number of clustering outputs (number of clustering results).

- $L(i,k)$: cluster label assigned to data instance i by clustering output k.

2. For each combination of clustering outputs, calculate a clustering persistence matrix that measures the agreement between clusters for each possible pair of clustering outputs. In those matrices the number of rows and columns correspond to the number of clusters for each clustering output respectively. For instance, if the cluster output 1 considers 5 clusters and the second 8 clusters, the matrix dimension is going to be 5x8, and for each cell [i, j] of the matrix, count the number of samples that fall in cluster i for clustering output 1 and in cluster j for clustering output 2. The algorithm is the following (comments start with hash character #):

   #for each matrix

   for a=1, K:

   for b=a+1, K:

   #for each cell of the matrix

   for x=1, nc(a):

   for y=1, nc(b):

$$M_{ab}[x,y] = \sum_{i=1}^{nd} I_{ab}[x,y](i)$$

   where:

$$I_{ab}[x,y](i) = \begin{cases} 1 & ;if\ (L(i,a)=x)\ and\ (L(i,b)=y) \\ 0 & ;otherwise \end{cases}$$

3. Standardize the matrices by dividing every cell by the total number of data:

   #for each matrix

   for a=1, K:

   for b=a+1, K:

   # for each cell of the matrix

   for x=1, nc(a):

   for y=1, nc(b):

$$S_{ab}[x,y] = \frac{M_{ab}[x,y]}{nd}$$

4. Store for each data instance its corresponding cell value in each matrix:

   #for each matrix
   for a=1, K:
     for b=a+1, K:
      #for each cell of the matrix
       for x=1, nc(a):
         for y=1, nc(b):
          #for each data instance i
           for i=1, nd:
               if (($L(i,a) = x$*) & (*$L(i,b) = y$*)):*
                   $P_{ab}(i) = S_{ab}[x,y]$

5. Compute the LUCC value for each data instance, which is a measure of how unlikely the sample falls into consistent clusters considering the different clustering outputs:

$$LUCC(i) = -\sum_{a=1}^{K}\sum_{b=a+1}^{K} \log(P_{ab}(i)) \qquad \forall i = 1, \dots, nd$$

   Note that natural logarithm is applied in order to highlight low values, so the samples with higher LUCC values are considered anomalies.

6. Plot the LUCC value in a cumulative probability plot and select the anomalous high values of the distribution (outliers).

In order to visually illustrate how this method works, consider the bivariate case shown in Figure 34, where there are two clusters (blue circles and red squares) and one anomalous sample (star shape). If different clustering configurations are applied to assign 2 clusters, there will likely be agreement that the circles are part

of a cluster and the squares correspond to another cluster. On the other hand, there is likely to be disagreement about the cluster to which the anomalous sample belongs. Thus, that sample is going to be an outlier in the LUCC distribution, and consequently, it will be considered as a LUCC anomaly.

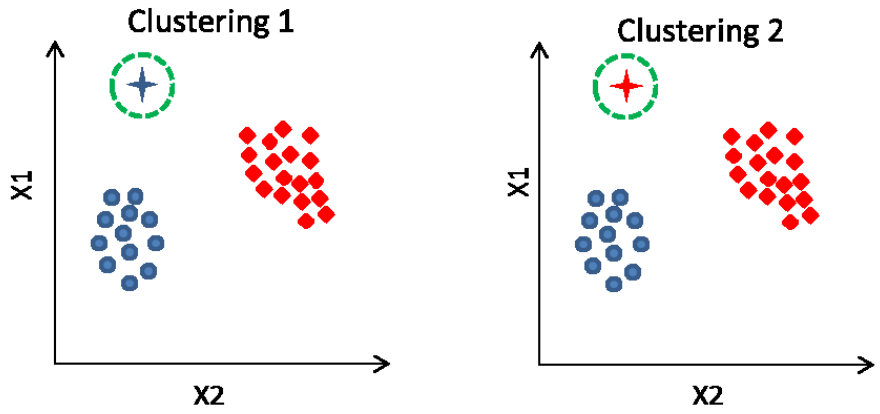

Figure 34. Example bivariate case for illustrating the LUCC method. Consider 2 clustering outputs for finding 2 clusters (blue and red). The clustering 1 assigns the anomalous sample to cluster blue while the clustering 2 assigns it to cluster red.

5.2.3. Spatial Anomalies

The first step is to apply a clustering technique aimed at large structures in the multivariate data. Then, the distance d to the closest sample assigned to the same cluster is calculated (Figure 35) as well as the number of samples assigned to a different cluster before reaching d (called *n_before*). Finally, a cumulative probability plot of the *n_before* value is generated and the outliers of the distribution are considered as multivariate spatial anomalies.

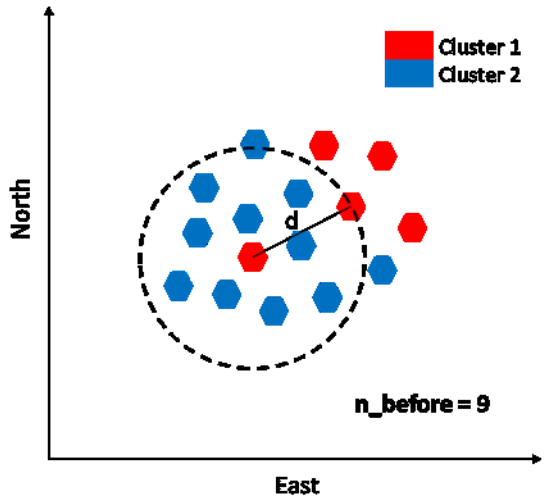

Figure 35. Illustration of n_before value calculation for detecting spatial anomalies. This value is calculated for each sample. For the sample in the middle of the circle there are 9 blue samples closer than d (distance of the closest sample assigned to the same cluster).

5.3. Case Study

The stream silt sediment samples from the Mackenzie Mountains are used to illustrate the method. Some elements deemed pathfinder elements for Sedimentary Hosted Copper/Base Metals deposits (Ootes et al., 2013) are selected for the analysis: Ag, Cd, Cu, Ga, Mo, Pb and Zn. The group of elements to use for applying the method is important. They should be selected in accordance to the type of exploration target. Univariate anomalies are considered as well as the multivariate methods described above. They are easy to use and are useful for comparison purposes.

5.3.1. Univariate Anomalies

Univariate anomalies (outliers) are selected based on the cumulative probability plots as show in Figure 36. Other methods for finding anomalies in the univariate space have been developed (Deutsch & Deutsch, 2010), yet the focus of this study is on multivariate methods for anomaly detection. Thus, graph-based visual univariate outlier detection is considered adequate.

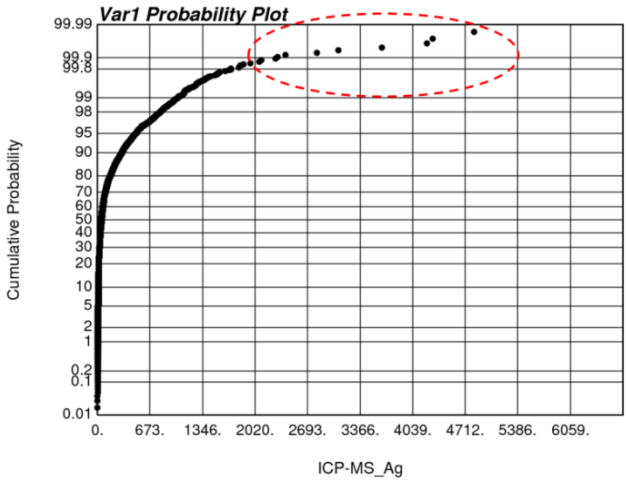

Figure 36. Cumulative probability plot for Ag. The dashed red circle shows the samples considered outliers of the distribution.

The selected thresholds for defining univariate anomalous values are: Ag ≥ 2020, Cd ≥ 44, Cu ≥ 355, Ga ≥ 10, Mo ≥ 63, Pb ≥ 350 and Zn ≥ 3900.

5.3.2. Multivariate Anomalies - Small Anomalous Clusters

The basic idea of this approach is to find small anomalous clusters by using different clustering techniques and data transformations considering different numbers of clusters as output. In this case, three different data transformations (standard scores, standardized by the range and normal scores transformation) and two clustering techniques (Hierarchical clustering and GMM) are used for identifying small anomalous clusters. Hierarchical clustering is performed using the average and Ward's method for linkage calculation. More transformations and/or clustering techniques —like K-means— could be added to the process.

For each combination of data transformation and clustering technique, the number of clusters was varied with the aim of identifying small groups of samples with anomalous concentrations of pathfinder elements. As an example, when standardizing by the range and using hierarchical clustering, if the data was grouped in only two clusters, the smallest contains 840 samples, which is fairly large. When selecting three clusters, one of them contains 54 samples with relatively high mean of Ag, Cd, Cu, Mo and Zn (Figure 37 and Figure 38), which

could be an interesting cluster for mineral exploration. Following this process, different small anomalous clusters are selected for each combination of transformation-clustering technique. Finally, the mean for each element in those small clusters is explored, in order to select only those with relatively high content of some pathfinder elements. This step requires experimentation and basic statistical analysis of the small clusters detected.

Figure 37. Number of samples for each cluster considering: 2 clusters (left) and 3 clusters (right). Using hierarchical clustering with Ward's linkage calculation on data rescaled by the range.

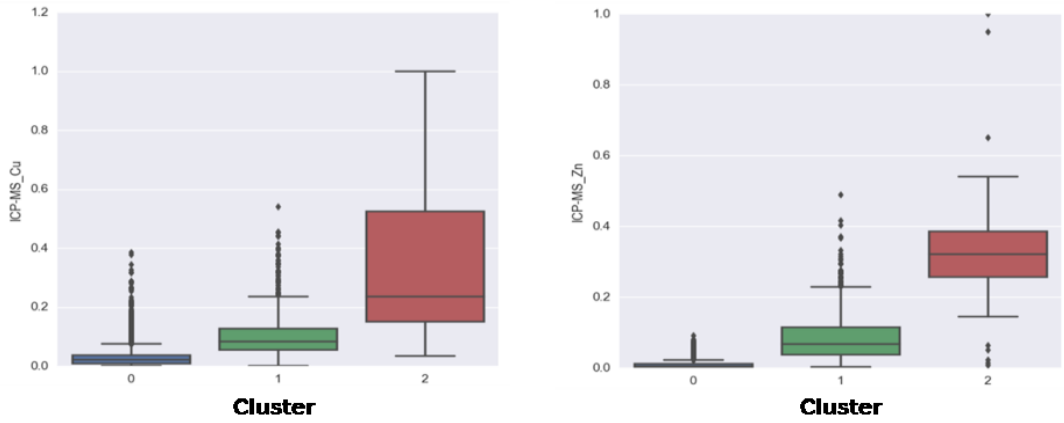

Figure 38. Box plot showing the relatively high mean of Cu and Zn of cluster 2, based on data standardized by the range and using hierarchical clustering with Ward's linkage

To illustrate the impact of the data transformation for finding small anomalous clusters, the results obtained by rescalinging by the range versus normal scores transformation are compared. Using Hierarchical method (Ward's linkage) for separating data into four clusters, when applied on data standardized by the range

leaded to the recognition of one relatively small cluster containing 54 samples (Cluster 2 in the left plot in Figure 39). On the other hand, when performing the same clustering method but on data transformed to normal scores, all four clusters contain more than 1500 samples (right plot in Figure 39).

Finally, the impact of the clustering method used for identifying small clusters is illustrated. When applying GMM on the data transformed to normal scores, even when asking it to assign 20 clusters, all of them were relatively large containing not less than 200 samples. But when using Hierarchical clustering with average linkage, even selecting 4 clusters recognized three small clusters containing 35, 1 and 1 samples. In the cases studies developed in this study is noted that GMM tends to be less affected by high values, which allows it to find the large structures in the data. On the other hand, Hierarchical clustering tends to give more importance to separate those groups of samples that are far from the large clusters centroids than GMM. This confirms the high sensitivity of distance-based clustering methods to skewness and outliers. In line with this, the average linkage calculation is more prone to highlight the more distant samples than Ward's linkage calculation.

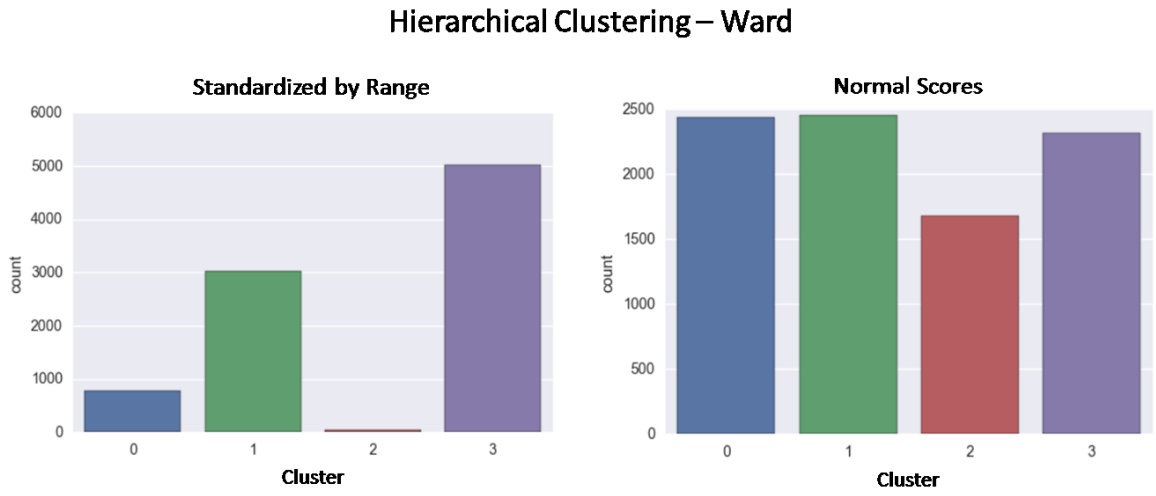

Figure 39. Bar plots illustrating the impact of the standardization method on the clustering output. The same clustering technique and parameters is used in both cases, but on the left the data was standardized by the range while on the left it was transformed to normal scores.

After exploring basic statistics of the different small clusters, the ones with relatively high mean of pathfinder elements are selected. The following small clusters are selected, which are shown in red in Figure 42:

- Normal scores:
    - Hierarchical clustering – average linkage:
        - 4 clusters: clusters selected 0, 2 and 4
        - 16 clusters: clusters selected 1, 6, 7, 8, 9, 12, 13 and 15.
        - 20 clusters: clusters selected 0, 6, 8, 10, 12, 13, 15, 16, 17 and 19.
- Standard Scores:
    - Hierarchical clustering – average linkage:
        - 10 clusters: clusters selected 0, 1, 3, 4, 5, 6 and 7.
    - Hierarchical clustering – Ward's linkage:
        - 6 clusters: cluster selected 0.
- Standardized by range:
    - Hierarchical clustering – average linkage:
        - 12 clusters: clusters selected 1, 2, 3, 4, 7, 8, 9, 10 and 11.
    - Hierarchical clustering – Ward's linkage:
        - 3 clusters: cluster selected 2.

Almost all univariate anomalous data are included as small cluster anomalies, but there are also many small cluster anomalies that are not univariate outliers, as seen in Figure 42.

5.3.3. Multivariate Anomalies – LUCC Anomalies

The LUCC value measures how a sample falls into different clusters. This method should work better if the clustering is set up to find large structures in the data. Therefore, clustering techniques based on data transformations and numbers of clusters that lead to small clusters are not used. Ward's linkage is used instead of average linkage. Other clustering methods like K-means can be used as well.

Accordingly, the following clustering methods, data transformations and numbers of clusters are used for identifying the LUCC anomalies:

- Normal scores:
    - GMM: 2, 3 and 4 clusters
- Standard scores:
    - Hierarchical clustering – Ward's linkage: 2 clusters.
    - GMM: 2 clusters.
- Standardized by range:
    - Hierarchical clustering – Ward's linkage: 2 clusters.
    - GMM: 2 clusters.

The second step is to calculate the clustering persistence matrix for each pair of clustering outputs. Twenty one matrices are generated. One of the matrices is shown in Table 2 as an example, in which the 2 clustering outputs are generated using hierarchical clustering with Ward's linkage assigning two clusters, but they are based on different data transformations. Note that both assign almost 90% of data to one cluster and 10% to the other, but there is disagreement in some samples: there are 56 samples assigned to the smaller cluster for data transformed to standard scores that are considered part of the larger cluster when standardizing dividing by range. That means that they are not clearly part of a particular group of samples in the multivariate space. They could be anomalous. The approach highlights the samples with the greatest inconsistency between different clustering classifications.

		Dividing by Range		
		0	1	
Standard Scores	0	**784**	**56**	840
	1	**158**	**7895**	8053
		942	7951	8893

Table 2. One of the twenty one matrices generated combining the seven clustering outputs performed. In this matrix the rows correspond to hierarchical clustering on data transformed to standard scores and the columns to hierarchical clustering on data standardized dividing by the range.

Then these matrices were standardized by dividing each cell by the total number of samples (8893) obtaining the $S_{ab}[x, y]$ values, as shown in Table 3.

		Dividing by Range		
		0	1	
Standard Scores	0	**0.088**	**0.006**	0.094456
	1	**0.018**	**0.888**	0.905544
		0.106	0.894	1

Table 3. An example of a matrix standardized by the total number of samples.

Finally the LUCC measure was calculated for each sample. Outliers of the LUCC distribution are visually selected using a cumulative probability plot (Figure 40). The samples with LUCC values greater than 50 are considered anomalies.

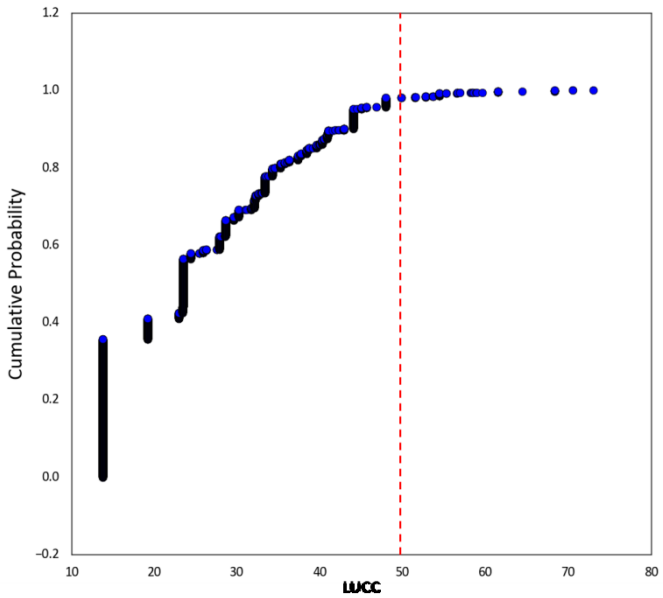

Figure 40. Cumulative probability plot of the LUCC value computed for the stream silt samples from the NWT. The samples with LUCC greater than 50 were considered LUCC anomalies.

Note that not all samples that this method highlights are anomalous because of their high values. Some of them have a high LUCC because the clusters are not clearly separated but they are transitional. The goal here is ore deposit exploration so all samples selected as LUCC anomalies with normal scores values lower than 2.0 for every element are discarded. The LUCC anomalies are shown in light blue in Figure 42.

5.3.4. Multivariate Anomalies – Spatial anomalies

The *n_before* value is calculated considering the clustering output shown in the case study developed in the previous chapter, which is based on 35 elements and performed with the aim of separating the geological background into large scale stationary domains. Then, a cumulative probability plot was generated (Figure 41). Based on this plot, a threshold of *n_before* equal to 250 was chosen to select the spatial anomalies (the *n_before* outliers). Seventeen samples are selected as spatial anomalies which are shown in green in Figure 42.

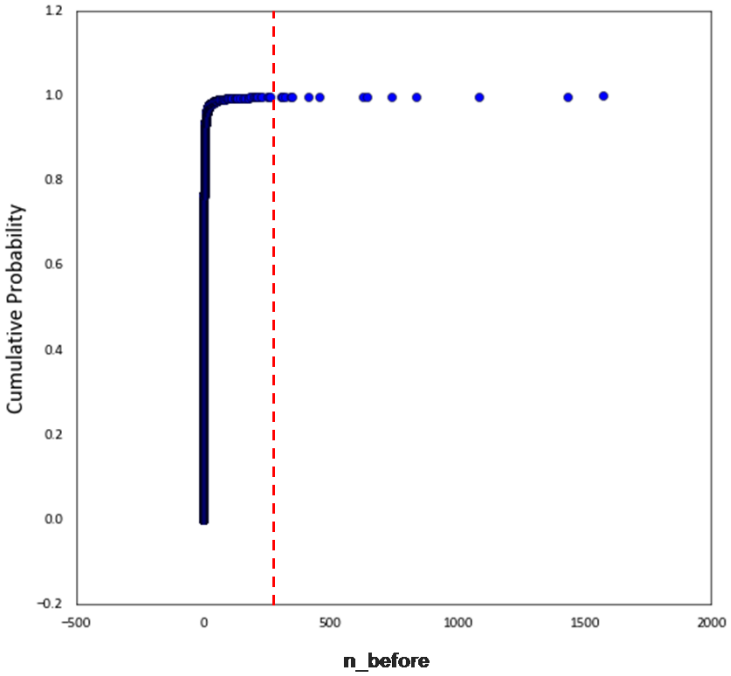

Figure 41. Cumulative probability plot for n_before value, which measures how different is a sample in the geographic area in which it is located. The threshold used for considering a sample spatial anomaly is shown in red dashed line.

5.4. Anomaly Detection Results

5.4.1. Visual Analysis

The anomalies identified by the different methods are summarized in Figure 42, where the pie charts symbolize with colors the corresponding methods that consider a sample to be anomalous. Yellow symbolizes univariate anomalies, red corresponds to small cluster anomalies, light blue represents LUCC anomalies and green symbolizes spatial anomalies. The more methods that consider a sample to be anomalous the greater the size of the pie charts. First of all, it can be noticed that anomalies are not randomly distributed in the sampled area (small grey dots are not anomalous samples), but they tend to be concentrated in some zones of the Mackenzie Mountains, which can be used for defining interesting areas for further investigation.

Another observation from this plot is that almost all univariate anomalies are considered multivariate anomalies, but on the other hand, there are many multivariate anomalies that are not identified as anomalies in the univariate space.

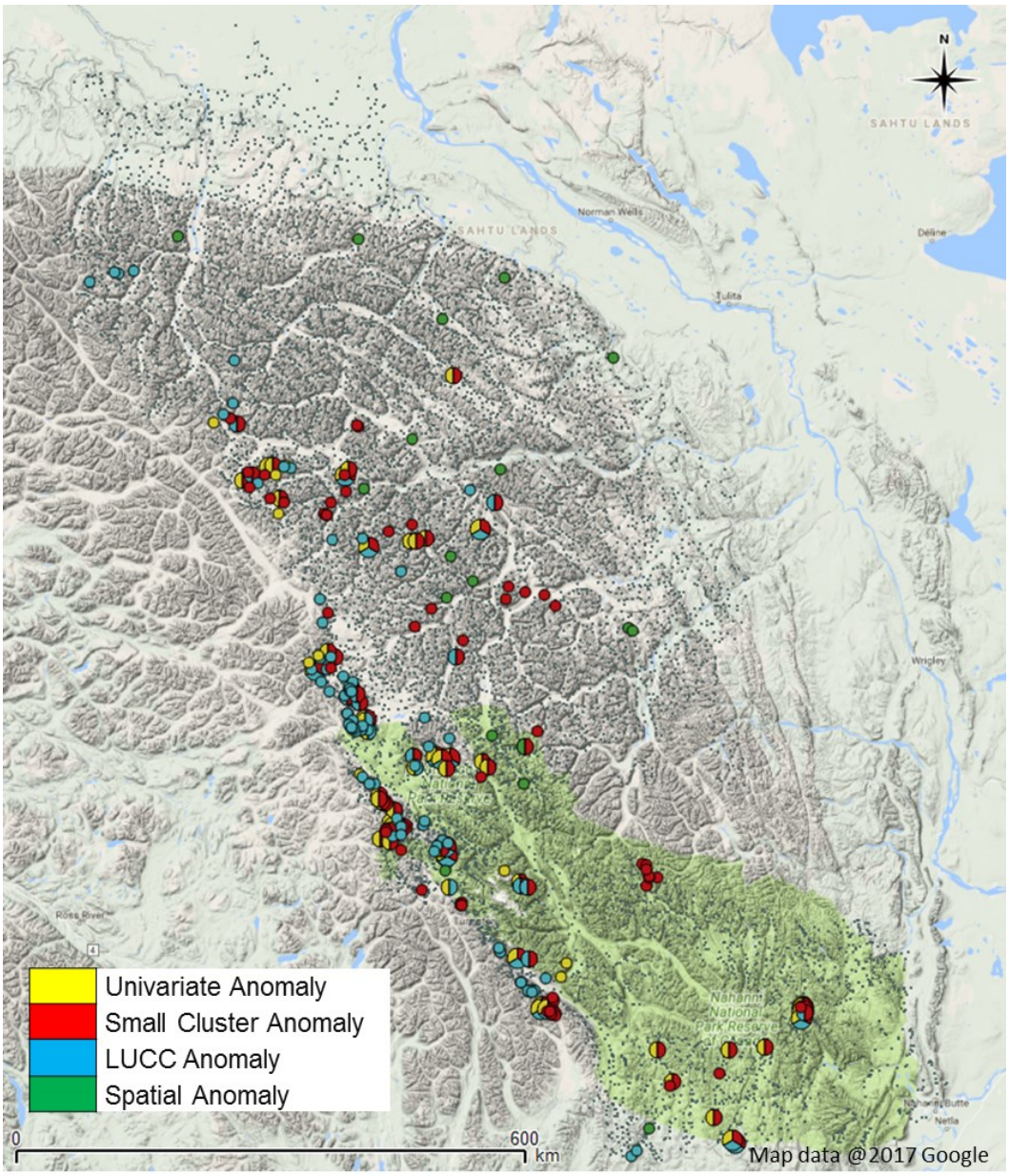

Figure 42. Stream silt samples anomalies plotted as pie charts on Google Physical image of the Mackenzie Mountains. The color in the pie charts symbolizes the type of anomaly. The more methods for finding anomalies agree, the more colors in the pie chart and the larger the area of the circle. Samples not considered anomalies are plotted as small grey dots

5.5. Validation

It is not straightforward to validate the performance of anomaly detection methods for mineral deposits exploration purposes, given that an intrinsic characteristic of exploration is that we do not know for every location whether there is a deposit nearby or not. This motivates unsupervised learning methods like clustering rather than supervised learning methods.

However, it is possible to see if the method is capable of detecting the known deposits in the area of study. A database with the showings (known ore deposits) that have been recognized in the area is provided by the Northwest Territories Geological Survey. The showings not related to the pathfinders selected in the case study were discarded and deposits that have been or are in production or advanced exploration are considered. The following filter was considered:

- Dev_label: 'Drilled' , 'Advanced Exploration' , 'Minor Past Producer' , 'Producer' , 'Past Producer' or 'Minor Producer'.
- Deposit_type: 'Carbonate-hosted Zn-Pb' , 'SEDEX' or 'Intrusion-related' .

The showing and anomalies are shown together in Figure 43 where the black dots symbolize the deposits of the area. Visually it is possible to observe that many areas with showings match areas with anomalies. It can be also noticed that many anomalies do not have showings next to them, which is not a problem but an opportunity.

To validate using numbers, for each showing the distance to the nearest anomaly is computed. This measure does not indicate if the deposits were detected by an anomaly downstream, but it is an indication of the capacity to identify interesting areas for exploration. A boxplot of the distance to the nearest anomaly for every showing is shown in Figure 44, in which it is possible to see that most of the showings are closer than 2.5 km to an anomaly.

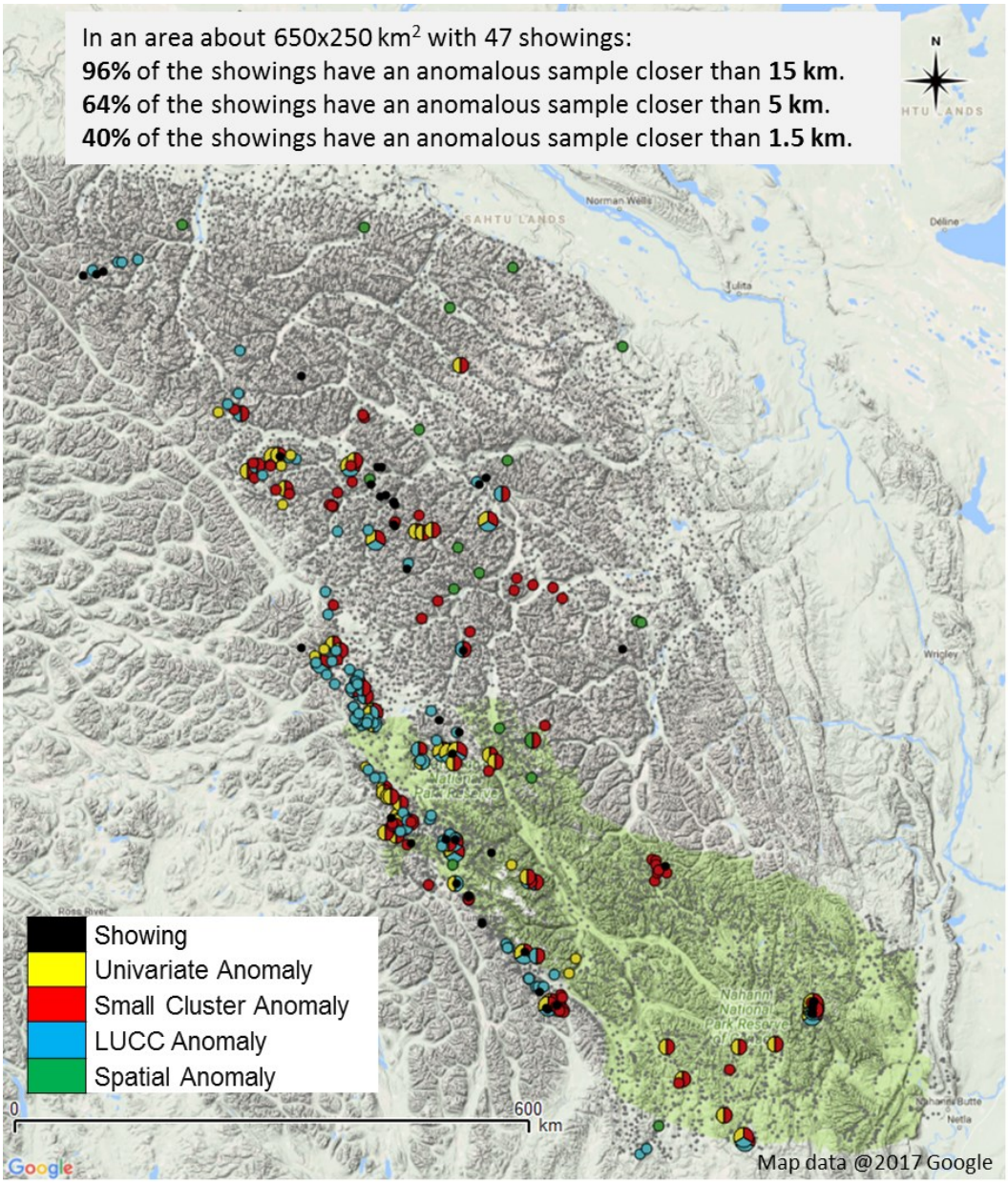

Figure 43. Showings (black dots) and anomalies (pie charts) on Google physical image of the Mackenzie Mountains. The color in the pie charts symbolizes the type of anomaly. The more methods for finding anomalies agree, the larger the area of the circle. Samples not considered anomalies are plotted as small grey dots.

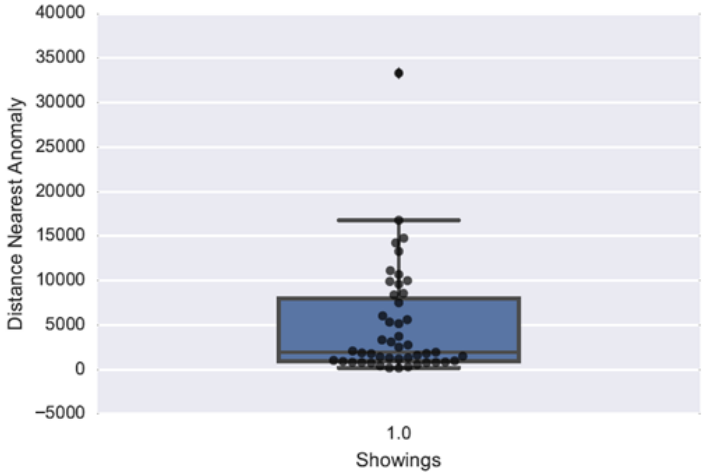

Figure 44. Boxplot of the distance to the nearest anomaly identified. The y axis is the distance of the closest anomaly to the showing. The black dots correspond to the 44 showings, which are spread in the x-axis just for visualization purposes.

To better understand this result it is necessary to consider the data spacing. In order to provide an idea about the data spacing for the stream silt samples, the average distance of the five nearest neighbors for each sample is calculated, which cumulative distribution is shown in Figure 45. The result supports the good performance of the method for detecting areas with presence of showings.

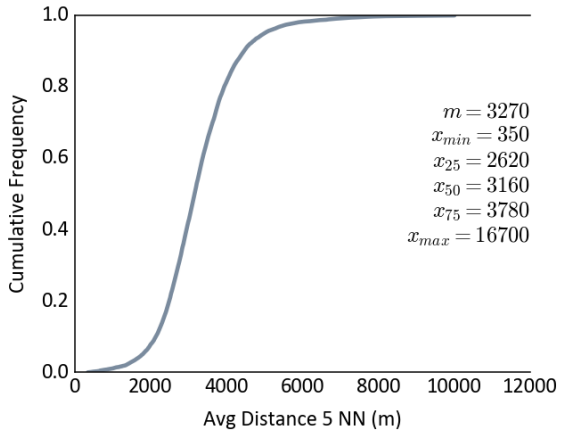

Figure 45. Cumulative probability plot of the average distance of the five nearest neighbors for each stream silt sample, to provide an idea of the data spacing.

Multivariate anomaly detection outperforms the univariate method for identifying anomalies related to showings in the area. Several deposits are detected just by using the multivariate methods proposed. The deposits identified as multivariate anomalies but not as univariate anomalies are highlighted with red circles in Figure 46. Some of these cases are also shown in Figure 47 in more detail, where it

is possible to observe that the multivariate anomaly detection methods not only give a general idea of interesting areas for exploration, but many of the anomalies correspond to the nearest sample downstream (Figure 47).

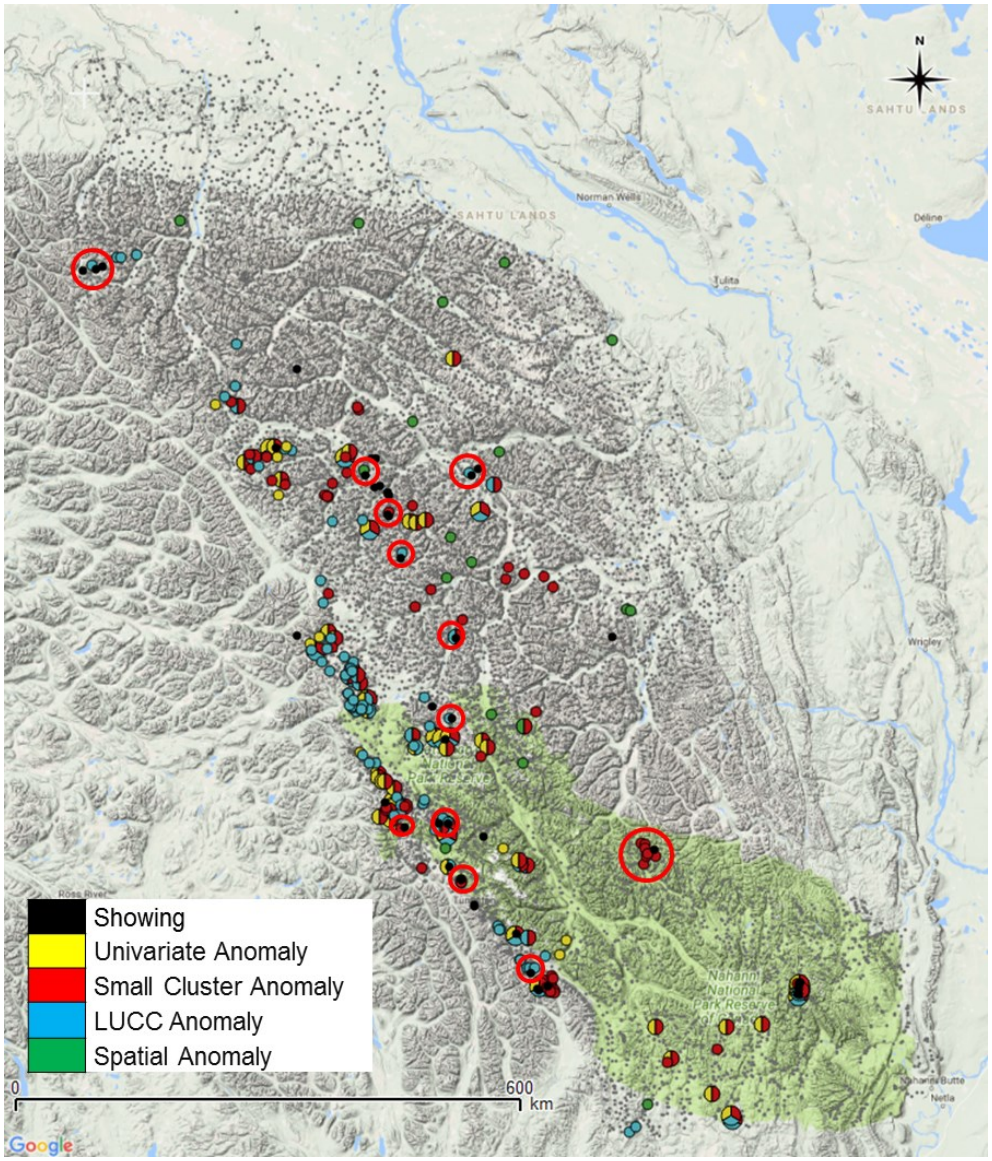

Figure 46. Red circles highlighting some of the cases in which the showings were just identified by the multivariate methods proposed. Showings (black dots) and anomalies (pie charts) on Google Physical image of the Mackenzie Mountains. The color in the pie charts symbolizes the type of anomaly. The more methods for finding anomalies agree, the larger the area of the circle. Samples not considered anomalies are plotted as small grey dots.

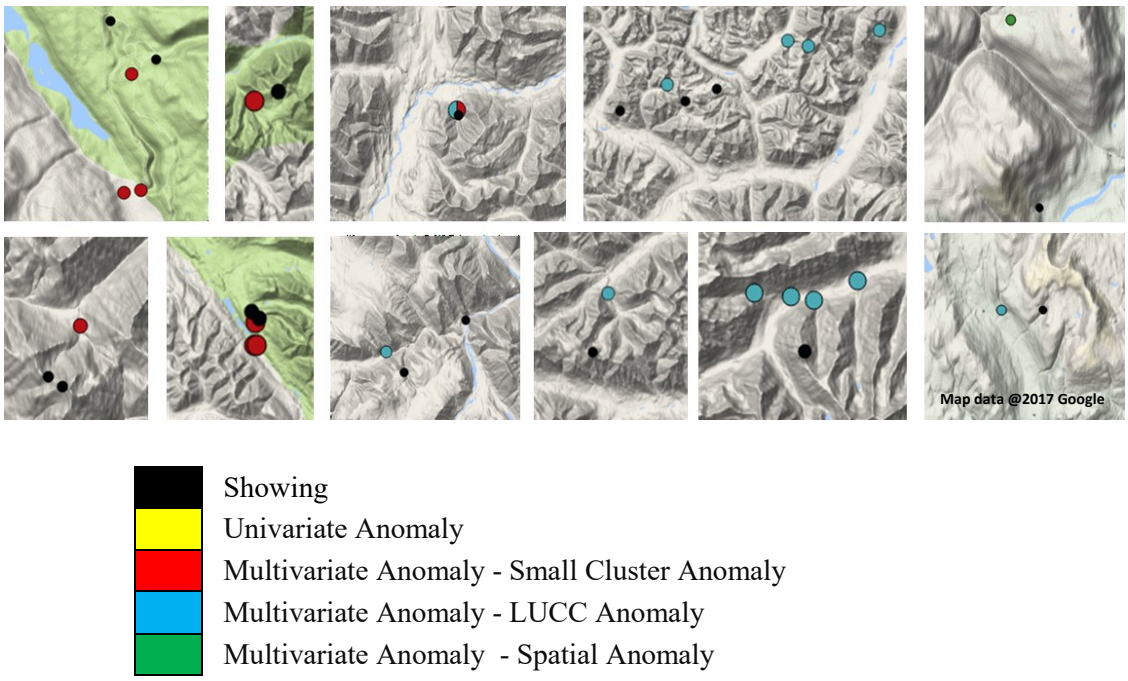

Figure 47. Some examples of showings detected just by using the multivariate methods proposed.

Another interesting observation is that different deposits were detected when different clustering techniques were applied on different data transformations. This supports the recommended approach of using different multivariate methods and varied combinations of clustering methods and transformations to identify anomalies from different perspectives. For example, the showing highlighted with a red circle in Figure 48 was just detected when transforming data to normal scores and applying hierarchical clustering with average method. In Figure 48 is also shown the position of these identified anomalies in the univariate cumulative distribution of the most important pathfinders. It is possible to see that they are not extremely high values of these univariate distributions. In fact, the pathfinder element with highest relative values is Pb and even for this element these multivariate anomalies are not univariate outliers.

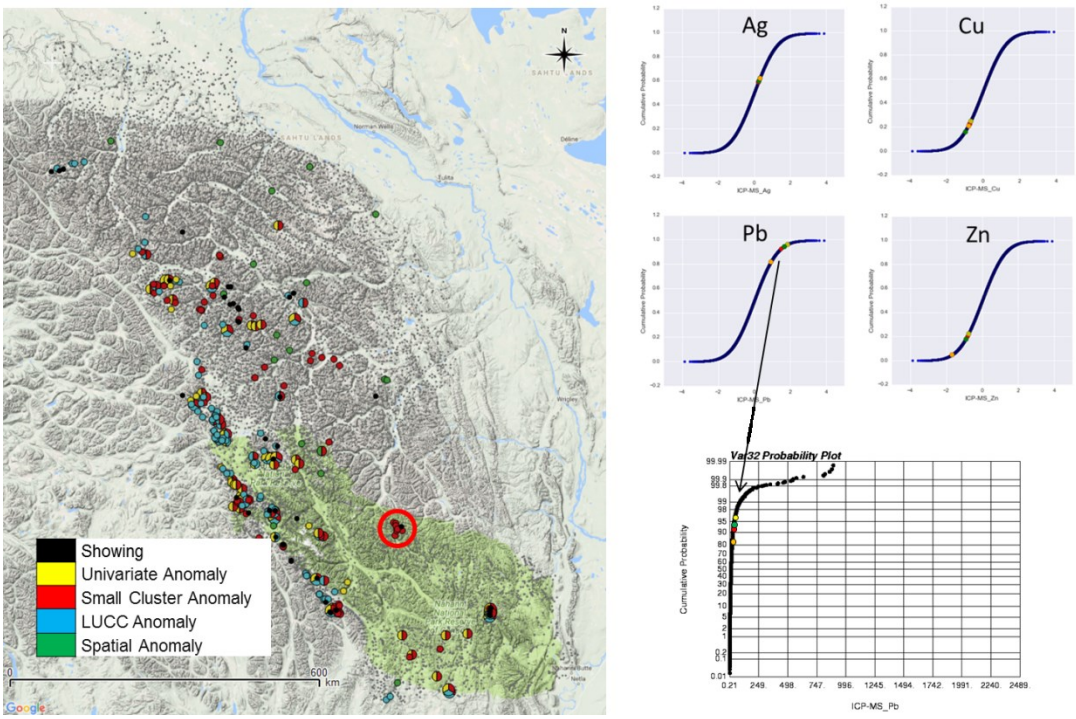

Figure 48. Left: red circle highlighting showing just detected using clustering to find a small anomalous cluster on normal scores data. Right top: cumulative probability plots of pathfinder elements in normal scores. Right bottom: Pb cumulative probability plot in original units. The multivariate anomalies are not univariate anomalies.

The amount of anomalies identified using each method is not a factor that was considered when performing the different methods. There are 73 univariate anomalies detected, 128 small cluster anomalies, 117 LUCC anomalies and 18 spatial anomalies. The fact that univariate anomalies are less than small cluster anomalies and LUCC anomalies is in line with the observation that almost all univariate anomalies are detected by the multivariate methods and several anomalies are just identified in multivariate space. It is possible that the threshold used to define outliers of the *n_before* distribution to define spatial anomalies is too high, but the results were not modified after the database of showings was provided. However, to understand this point a lower threshold were tried and it was possible to see that some more showings in the area are detected by the spatial anomaly method. Regarding the amount of anomalies for each method, it is important to note that frequently a deposit is detected by several anomalies of the same method around it. Therefore, dividing the amount of showings identified by

the number of anomalies detected for each method is not an accurate measure of their performance.

To understand the performance of the different methods developed regarding to their capacity to detect a deposit upstream by the closest samples downstream, a visual inspection around each of the 47 showings is performed. 25 deposits are detected by the nearest sample(s) downstream by the LUCC method, 22 by the small anomalous clusters method, 17 by the univariate method and 1 deposit is detected by the spatial anomaly method. From this point of view, the LUCC method outperformed the other methods.

It is important to consider that due to the low mobility of zinc under neutral to alkaline conditions, zinc deposits are difficult to detect using stream silt samples in carbonate-dominated regions like the Mackenzie Mountains, reason why the Northwest Territories Geological Survey decided to collect bulk stream samples for heavy mineral concentrate analysis (Falck et al., 2012). It is possible that some of the deposits not identified could be detected by applying the proposed methods on another type of data.

5.5.1. Comparing Performance with Other Multivariate Methods

The aim of this study is to develop novel multivariate methods for anomaly detection. One of the ideas is to use diverse techniques to find anomalies from different point of view. The focus is not to directly compare with existing multivariate methods that have already been proposed for exploration, like principal component analysis (PCA), factor analysis and weighted sums, among others (Cheng, Agterberg, & Bonham-Carter, 1996; Cohen et al., 2010; Garrett & Grunsky, 2001; Jimenez-Espinosa, Sousa, & Chica-Olmo, 1993; Zuo, 2011).

Nevertheless, comparison to the common method of PCA is of interest. PCA is used to identify anomalies for the case study, to see if the methods proposed here are capable to identify anomalies not detected by PCA. For this purpose data was

firstly transformed to normal scores, since PCA is sensitive to outliers, and then the principal components 1 and 2 (PC1 and PC2) were used to find anomalous samples. For defining the anomalies, an elliptic envelope was fitted to the central data points and then the samples with largest Mahalanobis distance to the center of the fitted envelope were considered outliers. Finally, just the anomalies with low PC2 values were selected, since they correspond to high values of the pathfinder elements, as seen in Figure 49.

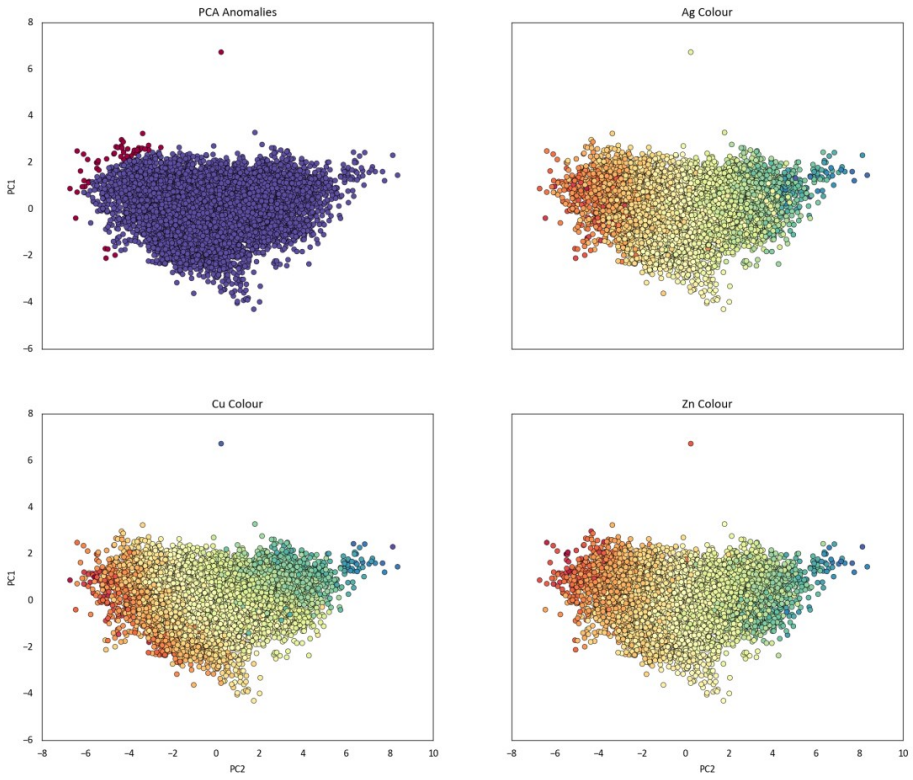

Figure 49. Scatter plots of principal component 1 (PC1) versus principal component 2 (PC2) colored different ways. Top-left: PCA anomalies selected in red color. Top-right: colored by Ag content. Bottom-left: colored by Cu content. Bottom-right: colored by Zn content.

The PCA anomalies were finally plotted together with the showings and anomalies detected previously (Figure 50), were it is possible to observe that: 1) some of the deposits in the area are detected by PCA method; 2) PCA did not detect new showings that had not been identified by the other anomalies; and 3) there are still several showings just detected by the multivariate methods developed in this study.

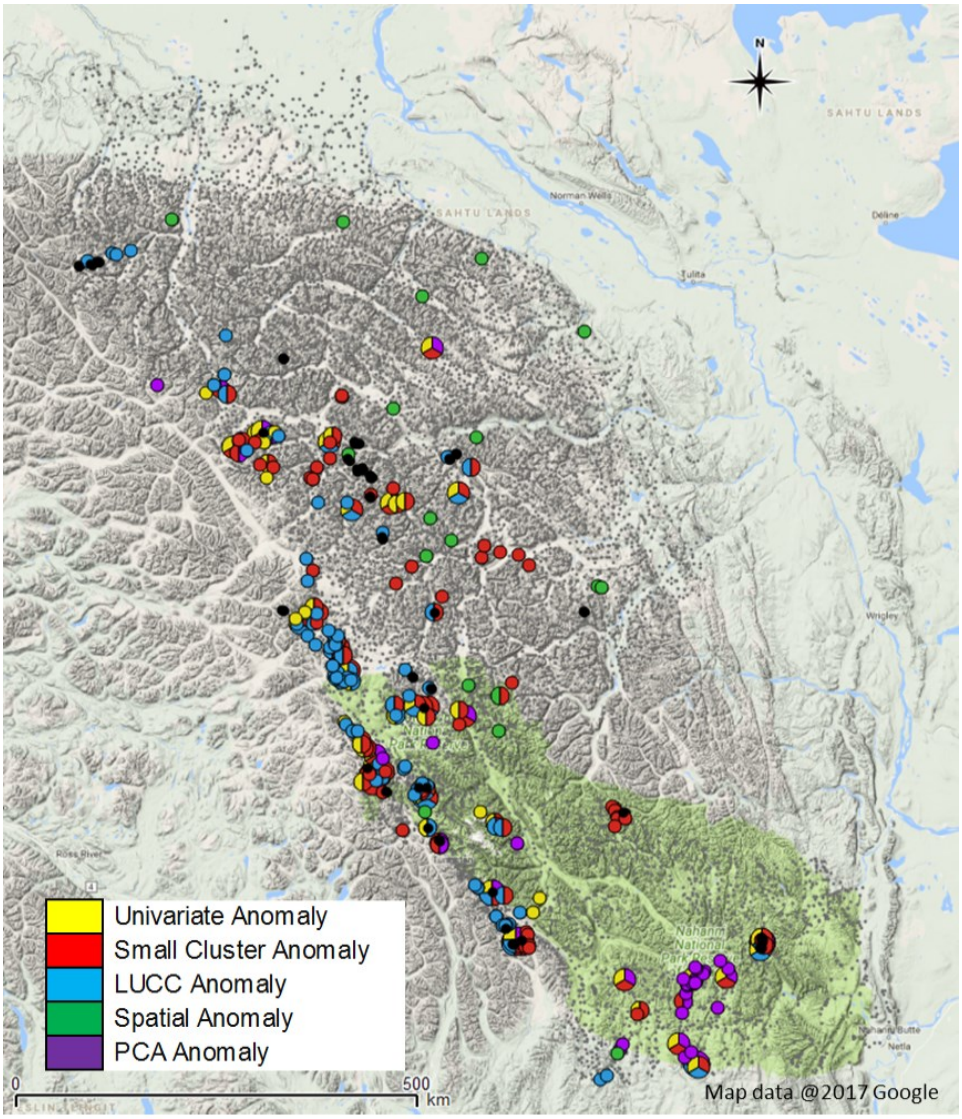

Figure 50. Showings (black dots) and anomalies (pie charts) on Google Physical image of the Mackenzie Mountains. PCA anomalies shown in purple. The color in the pie charts symbolizes the type of anomaly. The more methods for finding anomalies agree, the larger the area of the circle. Samples not considered anomalies are plotted as small grey dots.

5.6. Conclusion

Different ways for finding anomalies that go beyond univariate anomaly identification have been proposed and implemented for exploration purposes. The method allows detecting anomalous values from different points of view — varying the multivariate method, the clustering technique and the data transformation— with the idea of increasing the number of deposits that can be identified. This is supported by the case study. In this example, it is possible to

observe that the multivariate methods are capable of identifying several deposits that are not detected by univariate anomalies.

Many of the showings in the area were recognized, demonstrating the capacity of the method to detect anomalous samples related to ore deposits. The results suggest that the multivariate anomaly detection methods can lead to the identification of new ore deposits.

Chapter 6. Conclusions and Future Work

This thesis addresses challenges commonly faced when working with multivariate geochemical data. The goal is to provide guidance and methods to improve geostatistical analysis and cluster analysis for mineral deposit exploration. The main issue covered for geostatistical analysis is the effect of spikes and the selected despiking method in variography. For cluster analysis, the influence of different complexities and the appropriate data transformation are investigated. Finally, based on the knowledge gained about data transformations and clustering, a novel multivariate anomaly detection method is proposed.

6.1. Contributions

The effect of spikes in variography is addressed in Chapter 3, where problems with the commonly used despiking methods are documented. It is shown that local average despiking produces a similar error in variogram modelling compared to random despiking, but in the opposite direction: it leads to an overestimation of spatial continuity. The impact of the common despiking methods in the variogram model and uncertainty estimation is documented. A modified despiking method is proposed to improve variography. The proposed method combines a random despiking component and a local average despiking component, balancing both methods in order to avoid an excessive overestimation or underestimation of the spatial variability. The proposed despiking method is applied to two case studies, a synthetic case and a case based on the Northwest Territories stream silt samples, showing an improvement in the estimation of the variogram.

In Chapter 4 it is shown that data transformations have a significant impact on clustering results. The effects of outliers, skewness, multimodality and spikes on cluster analysis methods are reviewed, as well as the way that different data transformations alleviate the problems caused by these complexities. Finally, some guidance and recommendations are provided for improving cluster analysis performance. This chapter is oriented to find large clusters that could be used as

geochemical domains for improving the metallogenic model or for checking the geologic map, among other applications. A normal scores transformation preserving the spike (NSS) is proposed for improving distance-based clustering methods.

The principles developed in Chapter 4 are also that could be used as applied to identify multivariate anomalous samples. In Chapter 5, three different methods for identifying multivariate anomalies are developed. The first method uses different combinations of clustering and data transformations for finding small anomalous clusters. The second uses different clustering outputs for identifying samples that do not clearly belong to any cluster. The third recognizes samples that are spatially anomalous. Each of these multivariate methods detects anomalies from a different point of view. A combination of these detection methods is recommended. The goal is to obtain more stable and reliable results. If different anomaly detection methods agree that there are anomalies in a particular geographic area, that zone could be ranked with higher priority for mineral exploration. Finally, in the case study developed in Chapter 5 it is shown that the multivariate anomaly detection methods are capable of identifying several showings, that is, known mineral deposits in advanced exploration or production stage. Some of these showings are not detected from the histograms of different elements; this supports and motivates the use of multivariate anomaly detection methods for mineral deposit exploration.

On the whole, this thesis is a contribution to the mineral deposit exploration in the Northwest Territories, Canada. A report, high quality graphs and images, as well as the location of the detected anomalies have been delivered to the Government of the Northwest Territories.

6.2. Future Work

Another algorithm could be developed for despiking that considers uncertainty in the imputation of the values in the spike. Perhaps the Gibbs sample algorithm (Silva & Deutsch, 2016) could be used to simulate the values for the samples in the spike while considering the spatial correlation between samples. The input variogram model could be calculated using the despike algorithm developed in this thesis.

Data transformations and clustering techniques have been used in this thesis in combination either for the identification of large structures in data or for anomaly detection. It is not clear if some of the multivariate anomaly detection methods proposed would work better if they are applied separately for different domains. Additional research is required to provide more definitive guidance on different transformation methods.

Regarding the multivariate anomaly detection methods proposed, the interesting performance of the LUCC method motivates further investigation. It is probable that if the same clustering configurations used for the small anomalous clusters method are included as input for the LUCC method, the small anomalous cluster anomalies are identified as LUCC anomalies. It would be interesting to investigate if this method can be used in a more automatized way, using a default set of clustering algorithms, data transformations and amount of clusters, to address some concerns about the inevitable subjectivity to select the clustering configurations to find small anomalous clusters. The spatial anomalies method also requires more research. The method could be improved to consider anisotropy. It could also be improved to identify anomalies in cases where there is not just one spatial anomalous sample, but two or three samples are anomalous in a neighborhood. Finally, the decision of the threshold for defining spatial anomalies based on the *n_before* distribution could also be enhanced.

An intrinsic characteristic of exploration data is that there is limited knowledge of where deposits are located. That is a reason why unsupervised learning methods like clustering are useful for exploration purposes; they do not require a label for each sample for training the algorithm. However, it would be interesting to find a way to use some supervised learning algorithms to detect anomalies. One possibility could be using one-class classification (Martinus & Tax, 2001) which only requires samples from one class to train the algorithm.

Another possible future work for anomaly detection is performing sensitivity analysis to find the most influential variables for a target, since the selection of the group of elements to which apply the anomaly detection method proposed has an important impact on the results.

Bibliography

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, *29*(3), 626–688. http://doi.org/10.1007/s10618-014-0365-y

Barnett, R. M., & Deutsch, C. V. (2015). Conventional Clustering Algorithms and a Program for their Application. *CCG Annual Report 17*, (404), 1–18. Retrieved from http://www.ccgalberta.com

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58. http://doi.org/10.1145/1541880.1541882

Cheng, Q., Agterberg, F. P., & Bonham-Carter, G. F. (1996). A spatial analysis method for geochemical anomaly separation. *Journal of Geochemical Exploration*, *56*(3), 183–195. http://doi.org/10.1016/S0375-6742(96)00035-0

Cohen, D. R., Kelley, D. L., Anand, R., & Coker, W. B. (2010). Major advances in exploration geochemistry, 1998–2007. *Geochemistry: Exploration, Environment Analysis*, *10*(1), 3–16. http://doi.org/10.1144/1467-7873/09-215

Deutsch, J. L., & Deutsch, C. V. (2010). Fitting Probability Plots to Identify Multiple Populations and Outliers. *CCG Annual Report 12*, (310), 1–6. Retrieved from http://www.ccgalberta.com

Falck, H., Day, S. J. A., Pierce, K. L., Rentmeister, K., Ozyer, C. A., & Watson, D. M. (2012). *A Compilation of Heavy Mineral Concentrates : Results from Stream Sediment Samples Collected 2007-2010 , Mackenzie Mountains , NWT. NWT Open Report 2012-001*. Yellowknife, NT.

Filzmoser, P., Garrett, R. G., & Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers and Geosciences*, *31*(5), 579–587. http://doi.org/10.1016/j.cageo.2004.11.013

Fraley, C., & Raftery, A. (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, *41*(8), 578–588. http://doi.org/10.1093/comjnl/41.8.578

Garrett, R., & Grunsky, E. (2001). Weighted sums–knowledge based empirical

indices for use in exploration geochemistry. *Geochemistry: Exploration, Environment, Analysis*, *1*(2), 135–141. http://doi.org/10.1144/geochem.1.2.135

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. http://doi.org/10.1109/TPAMI.1984.4767596

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics. http://doi.org/10.1007/b94608

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. http://doi.org/10.1007/BF01908075

Jimenez-Espinosa, R., Sousa, A. J., & Chica-Olmo, M. (1993). Identification of geochemical anomalies using principal component analysis and factorial kriging analysis. *Journal of Geochemical Exploration*, *46*(3), 245–256. http://doi.org/10.1016/0375-6742(93)90024-G

Martinus, D., & Tax, J. (2001). *One-class classification: Concept-learning in the absence of counterexamples*. Delft University of Technology.

Massart, B., Guo, Q., Questier, F., Massart, D. L., Boucon, C., De Jong, S., & Vandeginste, B. G. M. (2001). Data structures and data transformations for clustering chemical data. *TrAC - Trends in Analytical Chemistry*, *20*(1), 35–41. http://doi.org/10.1016/S0165-9936(00)00058-3

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*(2), 181–204. http://doi.org/10.1007/BF01897163

Ootes, L., Gleeson, S. A., Turner, E., Ras-, K., Gordey, S., & Martel, E. (2013). Metallogenic Evolution of the Mackenzie and Eastern Selwyn Mountains of Canada's Northern Cordillera, Northwest Territories: A Compilation and Review. *Geoscience Canada*, *40*, 40–69.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: machine Learning in Python. *Journal of*

Machine Learning Research, *12*(Oct), 2825–2830. http://doi.org/10.1007/s13398-014-0173-7.2

Pyrcz, M. J., & Deutsch, C. (2014). *Geostatistical Reservoir Modeling* (Second Edi). Oxford university press.

Reimann, C., Filzmoser, P., & Garrett, R. (2002). Factor analysis applied to regional geochemical data: problems and possibilities. *Applied Geochemistry*, *17*(3), 185–206. http://doi.org/10.1016/S0883-2927(01)00066-X

Reynolds, D. (2009). Gaussian Mixture Models. *Encyclopedia of Biometrics*, (2), 659–663. http://doi.org/10.1007/978-0-387-73003-5_196

Rossi, M. E., & Deutsch, C. V. (2013). *Mineral resource estimation*. Springer Science & Business Media.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. http://doi.org/10.1016/0377-0427(87)90125-7

Silva, D. S. F., & Deutsch, C. V. (2016). Software for Gaussian Data Assignment in Truncated PluriGaussian Simulation. *CCG Annual Report 17*, (109), 1–12. Retrieved from http://www.ccgalberta.com

Templ, M., Filzmoser, P., & Reimann, C. (2008). Cluster analysis applied to regional geochemical data: Problems and possibilities. *Applied Geochemistry*, *23*(8), 2198–2213. http://doi.org/10.1016/j.apgeochem.2008.03.004

Thompson, M. (2012). *Handbook of inductively coupled plasma spectrometry*. Springer Science & Business Media.

Van Der Maaten, L. (2009). Learning a Parametric Embedding by Preserving Local Structure. In *12th International Conference on Artificial Intelligence and Statistics (AISTATS)* (Vol. 5, pp. 384–391). Clearwater Beach, Florida, USA.

Van Der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605. http://doi.org/10.1007/s10479-011-0841-3

Verly, G. (1984). *Estimation of Spatial Point and Block Distributions: The MultiGaussian Model*. Standford University, Standford, CA.

Verly, G., David, M., Journel, A., & Marechal, A. (1984). *Geostatistics for Natural Resources Characterization*. Dordrecht, Holland: D. Reidel Publishing Company.

Zuo, R. (2011). Identifying geochemical anomalies associated with Cu and Pb-Zn skarn mineralization using principal component analysis and spectrum-area fractal modeling in the Gangdese Belt, Tibet (China). *Journal of Geochemical Exploration*, *111*(1–2), 13–22. http://doi.org/10.1016/j.gexplo.2011.06.012

Appendix

- Appendix A: Despike program

The CCG *despike* program was modified to break the spikes by considering a local average component and a random component. The weight used for combining both components can be tuned in the parameter file, although the default weight W_1 equal to 0.5 is recommended.

Despike_2000 Parameter File

The required parameters for the despike program version 2.0 are:

- Line 4: the start point to read the parameter file. It must be present.
- Line 5: the data file in GeoEAS file format.
- Line 6: columns for coordinates, the variable to despike and the rock type.
- Line 7: number of valid rock types and their integer codes.
- Line 8: trimming limits for variable.
- Line 9: number of nearest neighbors to consider for computing the local average.
- Line 10: weight W_1 assigned to the random component. The weight (1-W_1) is given to the local average component. The default value W_1=0.5 is recommended.
- Line 11: seed number used for the random component.
- Line 12: name output file.

```
1                   Parameters for Despike
2                   **********************
3
4  START OF PARAMETERS:
5  spk_30.dat                    -file with data
6  1  2  0  3  0                 -columns for X, Y, Z, Var, and rock type
7  0  0  0                       -number valid RTs and their integer codes
8  -1.0e21   1.0e21              -trimming limits
9  10                            -number of NN for local average
10 0.5                           -Weight W1 for random component (between [0-1])
11 69069                         -Random Seed
12 dspk_30.out                   -file for output
```

- Appendix B: Anomaly Detection Method Scripts

1. Lack of Uniform Cluster Classification (LUCC) script:

The first step for calculating the LUCC value for each sample is performing different cluster analyses. For that purpose the Python package scikit-learn (Pedregosa et al., 2011) was used, which is recommended for data mining and data analysis. For illustration purposes consider a dataset with 2 coordinates and 7 clustering outputs:

	x	y	hier2_mmax	gmm_full2_mmax	hier2_st	gmm_full2_st	gmm_full2_ns	gmm_full3_ns	gmm_full4_ns
0	62.85193	-128.51200	1	0	1	1	0	2	2
1	62.82269	-128.56859	1	1	1	1	1	1	1
2	62.80669	-128.44852	1	1	1	1	1	1	1
3	62.81320	-128.48480	1	1	1	1	1	1	1
4	62.78361	-128.56381	0	1	0	1	1	1	1

Then, the following Python code for calculating the clustering persistence matrix for each pair of clustering outputs:

```
clusters = ['hier2_mmax', 'gmm_full2_mmax','hier2_st', 'gmm_full2_st','gmm_full2_ns',
            'gmm_full3_ns','gmm_full4_ns']
matrices = []
#For each pair of clustering outputs:
for n in range(len(clusters)):
    for m in range (n+1,len(clusters)):
        c = np.zeros((df[clusters[n]].unique().max()+1,df[clusters[m]].unique().max()+1))
        #Calculating number of data in each cell of the matrix.
        for i in range (df[clusters[n]].unique().max()+1):
            for j in range (df[clusters[m]].unique().max()+1):
                temp = df[(df[clusters[n]]==i)&(df[clusters[m]]==j)]
                c[i,j] = temp.shape[0]
        matrix = pd.DataFrame(data=c).astype(int)
        matrices.append(matrix)
```

The matrices are standardized:

```
#Standardize
matrices_st = []
for item in matrices:
    matrix_st = np.zeros(item.shape)
    for i in range(item.shape[0]):
        for j in range (item.shape[1]):
            matrix_st[i,j] = item.iloc[i,j]/(np.sum(item.values))
            if(matrix_st[i,j] == 0):
                matrix_st[i,j] = 1  # transform 0 (no samples) to 1 so logarithm(1)=0
    matrix_st = matrix_st
    matrices_st.append(matrix_st)
matrices_st
```

Finally the following code is used for assigning to each sample the corresponding value for each matrix (in this case 21 matrices). The LUCC value is calculated for each sample as the negative value of the sum of the logarithm of the different matrices values:

```
#Generating a column for each matrix and assigning p(j,j') value to every sample
col_mat = ['mat1','mat2','mat3','mat4','mat5','mat6','mat7','mat8','mat9','mat10',
           'mat11','mat12','mat13','mat14','mat15','mat16','mat17','mat18','mat19',
           'mat20','mat21']
k=0
for n in range(len(clusters)):
    for m in range (n+1,len(clusters)):
        df[col_mat[k]] = np.nan
        for i in range(len(df[clusters[n]].unique())):
            for j in range (len(df[clusters[m]].unique())):
                df[col_mat[k]][(df[clusters[n]] == i) & (df[clusters[m]] == j)] = matrices_st[k][i,j]
        k=k+1
```

```
df['LUCC'] = -(np.log(df['mat1']) + np.log(df['mat2']) + np.log(df['mat3']) + np.log(df['mat4'])
            + np.log(df['mat6']) + np.log(df['mat7']) + np.log(df['mat8']) + np.log(df['mat9'])
            + np.log(df['mat11']) + np.log(df['mat12']) + np.log(df['mat13']) + np.log(df['mat14
            + np.log(df['mat16']) + np.log(df['mat17']) + np.log(df['mat18']) + np.log(df['mat19
            + np.log(df['mat21']))
```

The outliers of the LUCC distribution are considered LUCC anomalies.

2. Spatial Anomalies Script:

A Python script used for calculating spatial anomalies is described below. There are not many samples in the case study developed, so the code works in a reasonable time even though it uses pandas dataframes. However, if more speed is required it is recommended using numpy arrays. Consider a pandas dataframe called *datafl* with coordinates and the clustering output:

	X	Y	gmm_full8
0	524843.3855	6969186.006	5
1	521984.3272	6965907.732	5
2	528118.1682	6964171.843	7
3	526262.5980	6964881.815	7
4	522257.3922	6961555.390	5

A function was generated to compute for each sample the distance for the closest *n_samples*. It creates a pandas dataframe storing those *n_samples* distances followed by the corresponding *n_samples* cluster labels assigned by the clustering method.

```
def build_spatial(data,n_samples=25):
    start_time = time.time()
    d=[]
    df4=pd.DataFrame()
    #For each sample:
    for i in range (data.shape[0]):
        #Calcluate distance to the other samples
        df1 = data.loc[i,:]
        df2 = data.drop(i)
        dist=((df2['X']-df1['X'])**2 + (df2['Y']-df1['Y'])**2)**0.5
        df = pd.DataFrame({'dist':dist,'gmm_full8':df2['gmm_full8']})
        #Sort df by distance
        df = df.sort_values(by=('dist'), ascending=True).reset_index(drop=True)
        #Keep the first n_samples rows and write the distances and labels in a row
        df = df[:n_samples]
        df3 = df['dist'].append(df['gmm_full8']).reset_index(drop=True)
        df4 = df4.append(df3, ignore_index=True)
    df5 = pd.concat((data,df4), axis=1)
    print("--- %s seconds ---" % (time.time() - start_time))
    return df5
```

Then the function is run indicating as input the dataframe (*datafl*) and the number of closest *n_samples* to consider. A dataframe called *df_dist* with the closest 1800 samples distances and cluster labels is obtained as follows:

```
n_samples=1800
df_dist = build_spatial(data=datafl,n_samples=n_samples)
```

To calculate how many samples assigned to a different cluster are closer than the closest sample with the same cluster label the following code is used:

```
start_time = time.time()
d_array = []
for i in range (datafl.shape[0]):
    distance=-99
    n_before=-99
    for j in range(n_samples):
        if (df_dist.iloc[i,2] == df_dist.iloc[i,j+(n_samples+3)]):#if same cluster
            distance = df_dist.iloc[i,j+3]
            n_before = j
            break
    d_array.append((distance, n_before))
d6 = pd.DataFrame(d_array, columns = ['dist','n_before'])
print("--- %s seconds ---" % (time.time() - start_time))
```

Finally, the original dataframe information can be concatenated with the distance of the closest sample with the same label and the *n_before* value:

```
d7 = pd.concat((df_dist[['X','Y','gmm_full8']],d6), axis=1)
print(d7.shape)
d7.head(5)
```

```
(8693, 5)
```

	X	Y	gmm_full8	dist	n_before
0	524843.3855	6969186.006	5	607.294980	1
1	521984.3272	6965907.732	5	2204.740422	1
2	528118.1682	6964171.843	7	894.583430	0
3	526262.5980	6964881.815	7	1986.756404	0
4	522257.3922	6961555.390	5	627.579266	0

The outliers of the *n_before* distribution are considered spatial anomalies.