Multivariate Exploratory Data Analysis of Spatial Data to Support Geostatistical Modeling

by

Haoze Zhang

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

Geostatistical modeling takes geological data as inputs and builds statistical models for resource prediction. Geostatistics consists of several components, including preprocessing, modeling, and postprocessing. Exploratory data analysis (EDA) is an early step in preprocessing. It provides the characteristics of data and helps identify erroneous or inconsistent data. In the context of geostatistics, missing data and below detection limit (BDL) data are an important anomaly to be understood in EDA. Missing data are problematic in EDA techniques such as principal component analysis (PCA). BDL data also cause problems when conducting cluster analysis and other analysis. Geostatistical models need to be conducted in stationary domains, so multivariate and spatial cluster analysis is another important aspect in EDA. It separates data into smaller groups in which data share similar features.

This thesis covers multiple aspects of geostatistical EDA. A data map examines missing data, and it shows the number of missing data in each variable and location. A combined permutation and Kolmogorov–Smirnov (KS) test identify if the missingness in variables is systematic. BDL data are investigated in univariate and bivariate methods. A BDL statistics table complements histograms. Three methods measure the spikiness of data. Bivariate analysis compares observed distributions with expected distributions which indicate full independence of BDL occurrence. Kullback–Leibler (KL) test quantifies the difference between the distributions, obtaining combinations of variables in which the BDL occurrence can be dependent. This helps the understanding of the reasons for BDL data.

The handling of BDL data in cluster analysis is addressed, including a workflow that finds the optimal number of clusters. Tests on synthetic data examine the compatibility of the workflow with different data transformations and clustering methods. K-means is a suitable clustering method for dealing with BDL spikes. Four transformations compatible with the workflow are combined with k-means to examine clusters in real data. The trade-off between spatial continuity and multivariate continuity in cluster analysis is addressed. A novel classification method is proposed to find the optimal clustering and domain labels. Ensemble clustering labels are used as inputs for the classification. The classification algorithm takes multiple sets of clustering labels as inputs. The domains are assigned based on clustering labels and two hyperparamters - spatial weight and number of domains. The matrix of classification results shows higher spatial weight results in more continuous domains. Flow simulation results show that the domain label assignment has an impact on

the performance of the final geostatistical models, because flow responses are highly sensitive to spatial and multivariate continuity.

Dedication

To PW.

Acknowledgments

I would like to thank my supervisor Dr. Clayton Deutsch. This thesis is not possible without your continued guidance and support. Your brilliance and diligence keep motivating me throughout my studies. I would also like to thank Centre for Computational Geostatistics (CCG) for the financial support. Thank my friends at CCG for the help with my questions and more importantly the great memories.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
2-D	Two-dimensional
3-D	Three-dimensional
BDL	below detection limit
BU	Bayesian Updating
CDF	cumulative distribution function
EDA	exploratory data analysis
GMM	Gaussian mixture model
GSLIB	Geostatistical software library
KL	Kullback–Leibler
KS	Kolmogorov–Smirnov
MAR	missing at random
MCAR	missing completely at random
MNAR	missing not at random
NC	number of clusters
PCA	principal component analysis
PPMT	projection pursuit multivariate transform
WCSS	within cluster sum of squares

Chapter 1

Introduction

1.1 Background

Geostatistical modeling uses geological data for resource estimation and the workflow has several components. Exploratory data analysis (EDA) finds the characteristics of data. If necessary, the analyzed data are cleaned and imputed (Abrevaya & Donald, 2017; Silva & Deutsch, 2018). The next step is to conduct the modeling, including transforming data to Gaussian space, variogram inference, kriging or simulation and back-transformation (M. J. Pyrcz & Deutsch, 2014). Postprocessing verifies the models using statistical tools such as cross-validation (Browne, 2000). To generate accurate models, the quality of the input data is of great importance, and this is more likely when EDA is conducted appropriately.

1.1.1 Exploratory Data Analysis

EDA is an approach to summarize the characteristics of data. The summary can possibly generate suggestions for collecting new data, using suitable data for further analysis or modeling, and reasons for the observed data features (Behrens, 1997; Tukey et al., 1977). The univariate data distribution can be summarized by statistical tables. They provide information such as the mean, variance and quantiles. Quantiles describe the univariate features more robustly when the data are highly skewed (Takeuchi, Le, Sears, Smola, et al., 2006). Data visualization is also important in EDA. It provides direct and concise observations of the univariate and multivariate data distributions. Cumulative distribution function (CDF) and histograms examine the univariate properties of data, which include the range of data, the frequency of data and the data skewness. Box plots examine data characteristics across multiple categories. In each category, the data is summarized with the minimum (q_0), maximum (q_{100}), median (q_{50}), first quartile (q_{25}) and third quartile (q_{75}). Scatter plots show bivariate relations, which can illustrate the correlations or non-linearity between variables.

Real data are rarely homogeneous. Data can be missing because of data collection errors. These missing data can influence the performance of classification or modeling (Ding, Han, Zhao, & Chen, 2015). There are several methods to address the problem, including dropping the missing data, filling in the missing data with mean or median, and predicting the missing values with regression (Little & Rubin, 2019). Adopting which method depends on the nature of missingness. If the missingness is random, dropping the data can be feasible. If the missingness is systematic, regression may be applied (Efron, 1994; Van Buuren, 2018). Outliers are extremely low or high data that appear

far from the majority of the data. Several outliers can drastically influence the results of regression. For example, in logistic regression, outliers can shift the decision boundaries greatly (Menard, 2002). Outliers are often omitted or capped, and there are several methods to identify them. For univariate and bivariate data, outliers can be observed visually using box plots or scatter plots. For high dimensional data, Z-score can be applied. It measures how many standard deviations data are from the mean value. Outliers are often identified as data beyond 3 standard deviations (Rousseeuw & Hubert, 2011).

Advanced multivariate analysis tools explore high-dimensional relations in data, including principal component analysis (PCA) (Hotelling, 1933) and cluster analysis (Fred & Jain, 2005; Romesburg, 2004). PCA is a dimension reduction technique, and also creates new coordinates that decorrelate the original data. For high dimensional data, some dimensions do not show significant variability. By reducing the high dimensional data to fewer dimensions that exhibit the most variations, the modeling can be faster. The new coordinates (principal components) are orthogonal to each other. Starting from the first principal component, the lines are the ones that minimize the average squared distance of points to the lines. The following lines need to be orthogonal to the previous ones and minimize the average squared distance (Abdi & Williams, 2010). Data first need to be standardized to zero mean and a standard deviation of one. The zero mean is necessary as data need to be rotated to the new coordinates. The standard deviation of one makes the interpretation of PCA results easier. Then, the covariance matrix is calculated and decomposed into the eigenvector matrix and the eigenvalue matrix. The resulting PCA transformed data is calculated by multiplying data with the eigenvector matrix. PCA calculates the eigenvalues and eigenvectors of multivariate data and projects data to the eigenvectors. The variability of each eigenvector is reflected by the corresponding eigenvalue. The resulting coordinates may not be the same as the original ones. Projecting data to the first several principal components can represent the majority of the data variations.

Cluster analysis groups similar data together, separating them into subsets that have more distinguishable features. Further analysis conducted on the well separated clusters can provide insight into the data. Data less than 4 dimension may be clustered visually. For higher dimensional data, statistical tools are necessary. The similarity of data is defined differently and this leads to different types of clustering methods. The most common ones include connectivity-based clustering (hierarchical clustering) (Johnson, 1967), centroid-based clustering (k-means) (Krishna & Murty, 1999), distribution-based clustering (Gaussian mixture model (GMM)) (McLachlan & Basford, 1988; Reynolds, 2009) and density-based clustering (Density-based spatial clustering of applications with noise (DBSCAN)) (Shen et al., 2016). Hierarchical clustering groups data based on their connectivity. Data closer to each other are grouped in early stages, forming intermediate groups. Different definitions of distance between groups result in different linkage criteria. The commonly used ones are maximum, minimum and average criteria. Using different linkage criteria can lead to different clustering of the intermediate groups. The clustering results and the number of clusters can be ob-

served using a dendrogram (Sander, Qin, Lu, Niu, & Kovarsky, 2003). K-means clustering groups data based on the distance between clustering centroids and data. Data are assigned to the nearest centroids and the new centroids are calculated for the next iteration of cluster assignment. The process continues until the algorithm converges. The results are determined by the number of clusters and the location of the initially generated centroids, so different initial centroids are generated and the best clustering results are returned as the final results. GMM shares similar process as k-means. The difference is that multiple Gaussian kernels are generated rather than centroids, and data are assigned based on their probabilities in different Gaussian kernels. Gaussian kernels can handle clusters with elongate shapes better than k-means (Lücke & Forster, 2019). The key parameter of DBSCAN is the radius r. For each data point i, DBSCAN search the number of data within r. If the number of data is above a threshold q, the data are defined as core points. Points within r of other core points are called directly reachable. Points are called reachable if there is a path for them to be connected to core points. Points are called noise points when they are not reachable by other points. For a core point, the cluster is defined as all data reachable from it. With many different clustering methods, each clustering type has their own application situation. No clustering can outperform another in all situations.

1.1.2 Problems with Geostatistical Exploratory Data Analysis

Geostatistical EDA focuses on understanding missing data, below detection limit (BDL) data and outliers (Prades, 2017). Geostatistical missing data can come from the high cost of acquiring drill hole data, or data collection errors. Sometimes data are missing in a variable because the data in other variables are below a threshold (Little & Rubin, 2019). Missing data can occur for most variables at some locations or in several variables at many locations. It can cause problems for multivariate analysis such as PCA. If some variables are missing, the coordinates in the multivariate space are unknown, and these data cannot be used in PCA. To handle the missing data, the nature of missingness needs to be understood. BDL data originate primarily measurement equipment limitations (Palarea-Albaladejo & Martin-Fernandez, 2013). The concentrations of some elements are so low that measurement equipment cannot detect them. They are often recorded as 0.0 and form spikes in histograms. These spikes are problematic in quantile transformation (Prades, 2017). The spikes can be distributed from low to high, which is also known as despiking (Verly, 1984). How to assign quantiles for the data in spikes results in multiple despiking methods. Different ways of transforming spikes can lead to different EDA results. To find an appropriate way of despiking the BDL data, the characteristics of the BDL spikes and the dependence of BDL occurrence are examined. Outliers are data with extremely high values, sometimes orders of magnitude larger than the mean of data. They may come from very high concentrations or data collection errors. Outliers can also appear as extreme low values, and orders of magnitude less than the majority of data. The

corresponding distribution is negatively skewed such as Fe or SiO_2. Outliers can cause problems when clustered with centroid-based methods (Chawla & Gionis, 2013; Prades, 2017). They shift the centroids drastically compared with ordinary data, resulting in inaccurate clustering results or an incorrect number of clusters.

Applying advanced EDA methods to geostatistical data helps identify stationary domains. Cluster analysis finds such domains in multivariate space, but there are several factors that affect the performance of cluster analysis. The multivariate clusters cannot be identified simply through univariate or bivariate plots. Different clusters may not be obvious until analyzed in high-dimensional space, so statistical tools are of great importance for cluster analysis. The number of clusters is the most important hyperparameter for popular clustering methods (Milligan & Cooper, 1985; Tibshirani, Walther, & Hastie, 2001). Setting different numbers of clusters can lead to different clustering results. Therefore, robust methods to find the correct number of clusters are needed. The anomaly data mentioned above can also affect the performance of clustering analysis. GMM can falsely assign a Gaussian kernel only for the BDL spikes, so different transformations are compared and the appropriate ones are used to amend the problems caused by anomaly data. Validation methods are also important to ensure the clustering results are trustworthy. Cross-validation applied to clustering results finds if data are clustered or partitioned.

Geostatistical data have two components, the spatial arrangement of the values and the multivariate data values. Cluster analysis groups data in multivariate space. Data are labeled to ensure multivariate continuity, but the corresponding spatial distribution of the labels (domains) may be scattered. The scattered spatial continuity can lead to unstable variograms, and further influence the performance of modeling. To obtain continuous domains, cluster analysis could be considered with spatial data only. In this case, spatial continuity is ensured, but the clustering labels may be scattered in multivariate space. It is not recommended to cluster spatial data directly because of the complexity in the shape and geometry of geological domains. There is a clear trade-off between the multivariate and spatial continuity (Martin, 2019). It would be beneficial to modeling performance if optimal clustering labels ensure reasonable continuity both in multivariate and spatial space.

1.2 Thesis outline

This thesis addresses selected problems in geostatistical EDA, including missing data, BDL data, different transformations of BDL in data cluster analysis, and the trade-off effect of clustering labels between spatial and multivariate continuity. Chapter 2 examines missing data in a geochemical dataset. A data map shows the information about the missingness in variables and locations. Multiple statistical tools determine if the missingness are random or systematic. Chapter 3 explores the BDL data in the same dataset. The data are analyzed in both univariate and bivariate ways. A BDL statistical table complements histograms. Three different methods evaluate data spikiness.

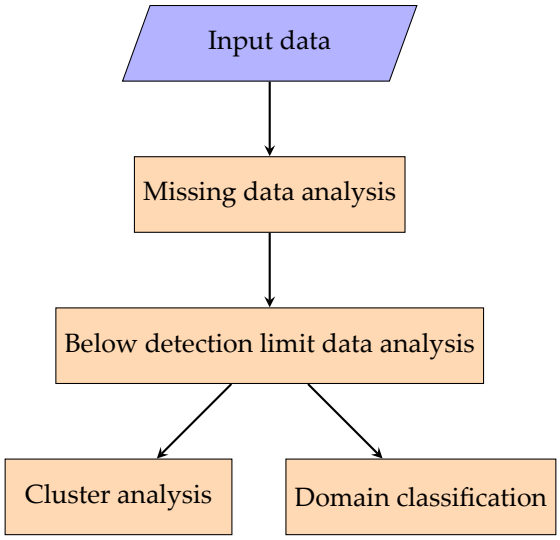


Figure 1.1: Flow chart of Geostatistical EDA workflow

Bivariate analysis of the BDL data examines if the occurrence of BDL between two variables are independent. Chapter 4 compares the effects of different transformations of data on cluster analysis. Different transformations are considered to handle the potential problems caused by BDL spikes and a workflow is proposed to identify the optimal number of clusters. This chapter also investigates the compatibility of the workflow with different transformations and clustering methods. Chapter 5 aims at finding an optimal set of domain labels which ensures both multivariate and spatial continuity. Ensemble clustering is used to cluster multivariate data. Then, a novel classification method classifies domains given the clustering labels and spatial configuration of the data.

The tools covered in this thesis can formulate a flowchart shown in Fig.**??**. Missing data analysis should be conducted first for the imputation of data. BDL analysis should be conducted next before the spikes are despiked or preserved in cluster analysis. Cluster analysis and domain classification can be conducted simultaneously.

Chapter 2

Missing Data Analysis

The increasing availability of multivariate data provides information and challenges for geostatistical modeling. The presence of missing data values may cause problems in exploratory data analysis (EDA) including compositional data calculation and principal component analysis (PCA). The nature of missingness influences the management of missing data. A data map is developed to understand the extent and nature of the missing values. Statistical tools are developed to establish whether the missingness is random or systematic. The comparison is conducted through a quantitative measurement of conditional distributions combining a Kolmogorov–Smirnov (KS) test and a permutation test. Examples demonstrate the robustness of the techniques. The data map distinguishes between missing at random and missing not at random. The statistic tools differentiate missing completely at random from missing at random.

2.1 Introduction

2.1.1 Background

As sampling equipment improves and technical decisions become more challenging, an increasing amount of multivariate data are acquired for geostatistical analysis. Although multivariate data provide extra information, not all of the data are homotopic (equally sampled). Some variables are missing due to cost, data vintage, and other considerations, and this may cause problems during analysis. First, it could lead to undefined values in compositional data calculation. In Aitchison (1982); Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015), the classical definition of compositional data excludes the possibility of missing data. Arbitrarily assigning the missing values mean values or zero values does not satisfy the sum to unity. Another problem with missing data is encountered when conducting PCA (Abdi & Williams, 2010; Hotelling, 1933). PCA is an effective method to reduce the dimensionality of multivariate data. It finds the dimensions that capture the most variability and projects the full-dimensional data onto these dimensions. The projected data are also decorrelated, which simplifies geostatistical modeling. In PCA, the heterotopic (unequally sampled) data cannot be used, because their locations in the full-dimensional space are unknown.

With the potential problems caused by missing data, the missing data need to be imputed. Understanding the nature of missingness is the first step before data imputation. It helps decide which imputation method should be applied. For example, if the data are missing at random, the traditional imputation methods do not introduce bias. There are three types of missingness: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR)

(Rubin, 1976). MCAR means the missing data occurrence is independent of observed and missing samples. It can happen for some systematic reasons but the missing values are completely random with respect to the data values. In this case, simply omitting the missing data does not introduce bias, but MCAR rarely occurs. MAR means the missingness occurrence is dependent on observed data but does not depend on missing data. For example, some variables are not sampled when other variables are below a threshold. The missingness is classified as MNAR when it does not belong to the previous two classifications. In this case, the missingness is dependent both on observed and missing data. This is the most challenging scenario as the imputation based only on observed data can introduce bias.

2.1.2 Geostatistical imputation

To avoid the aforementioned problems, the missing data can be dropped, but this may lead to bias, especially when the nature of missingness is not random. Another solution is to impute the missing data. Different from the well-established theories of imputation in other fields (Enders, 2010), imputed geostatistical data should retain multivariate relationships and spatial structure. One approach is to use Bayesian Updating (BU) (Barnett & Deutsch, 2015; Doyen, Den Boer, Pillet, et al., 1996). The missing data are informed on by different data sources. One source is the data of the same variable from other locations (primary data). The other is the collocated data at the location of the missing value (secondary data). The two results are merged and the final result is sampled to create multiple possibilities. Gaussian mixture model (GMM) imputation (Silva & Deutsch, 2018) considers a non-parametric fitting of the multivariate to adapt to more complex features. The imputed data carry the uncertainty through subsequent analysis.

The real data used in this chapter come from the Government of the Northwest Territories. It is part of the National Geochemical Reconnaissance stream sediment and water survey and the field collected data serve the purpose of building a geochemical database for mineral potential. There are three types of samples: stream silt samples, stream water samples, and bulk steam sediment samples (Falck et al., 2012). The dataset consists of 51 variables (elements) and about 8500 data samples. The dataset has missing data, below detection limit data, and outliers. Here, the missing data are examined.

In this chapter, the difference between MCAR and MAR is examined through numerical analysis. The term systematic missingness refers to MAR. First a data map is generated showing the general information of the missingness, including the missing data location, variables containing the most missing data, and an optimal dataset that contains no missingness. Then, a statistical tool is developed to explore the nature of missingness. It compares the two subsets of complete variables, where the target missing variable is present and absent. The developments are demonstrated with the Northwest Territories dataset. The tool is further validated on a synthetic dataset generated

from the real data where the mechanism of missingness is understood.

2.2 Import and explore the data

In this section, missing data notations are introduced first. Then, basic information about the missingness in the real dataset is presented through a data map. A procedure to find the optimal subset of data containing no missing values is proposed and the optimal data subset are highlighted in the data map.

2.2.1 Missing data notation

Suppose there is a data table (Fig.2.1). Rows represent data and columns represent variables. The data are denoted $\{Z_{\alpha,k}; \alpha = 1, \cdots, N; k = 1, \cdots, K\}$ where N represents the number of data and K represents the number of variables. $Z_{\alpha,k}$ is a number when the data are recorded above or below detection limit (recorded as 0.0), and a null value or not a number (NAN) when the data are missing. Some terms are defined as follows: "data" refers to multivariate observations α $\{\alpha = 1, \cdots, N\}$, "variable" refers to all observations for a specific column k $\{k = 1, \cdots, K\}$, and "value" refers to specific entries $Z_{\alpha,k}$ in the data table. When calculating the number of missing data P in row α,

$$P_\alpha = \sum_{k=1}^{K} 1\{Z_{\alpha,k} = NAN\}, \alpha = 1, \cdots, N,$$

where $1\{True\} = 1$ and $1\{False\} = 0$. When calculating the number of missing data M in column k,

$$M_k = \sum_{\alpha=1}^{N} 1\{Z_{\alpha,k} = NAN\}, k = 1, \cdots, K.$$

The following equation holds for the total number of missing data N_t:

$$N_t = \sum_{\alpha=1}^{N} P_\alpha = \sum_{k=1}^{K} M_k.$$

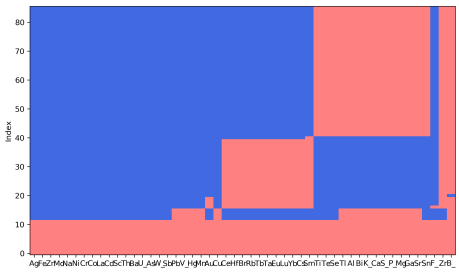


Figure 2.1: An example of data table. Red color represents missing data. Blue color represents observed data. Data index are on the veritcal axis and the variable names are on the horizontal axis.

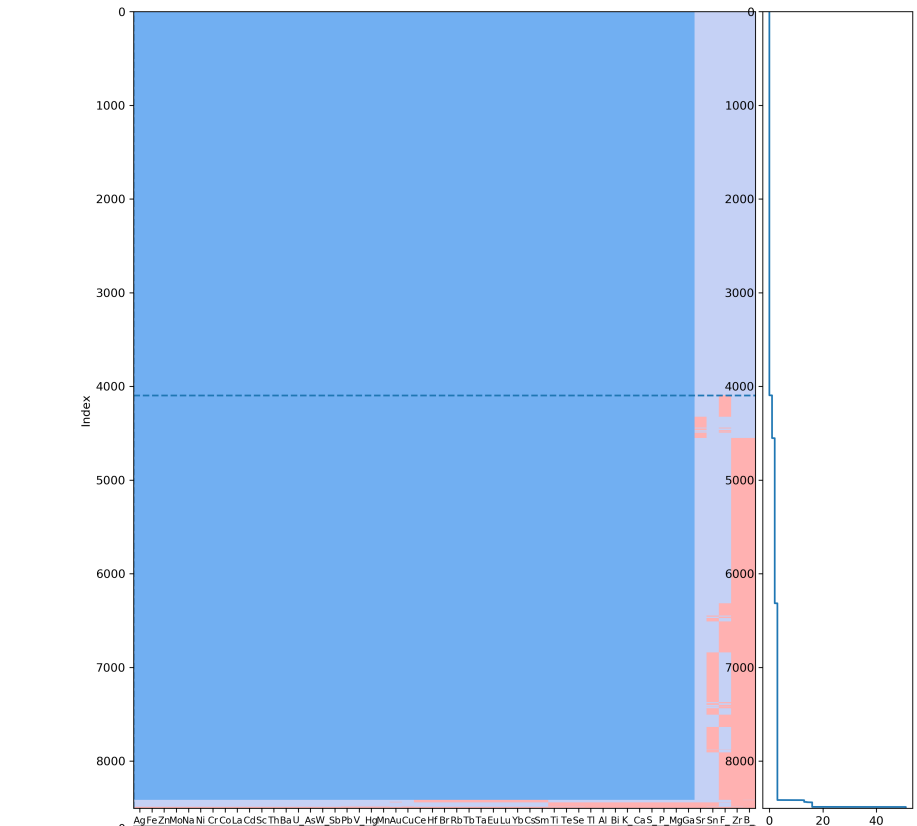
2.2.2 Data map

Consider the data map in Fig.2.2. The data are ordered based on the number of missing data in each row (data observation) and column (variable). Blue represents available data and red represents missing data. The plots on the edges show the marginal distributions of the number of missing data. Data observations closer to the bottom have more missing variables. There are around 50 unsampled locations. Variables closer to the right edge have more missing observations. 5 variables (Sr, Sn, F, Zr, B) contain a large number of missing data, and Zr and B have almost 50% of missing data. The area is divided into four regions: one region with complete rows and columns, two regions with either incomplete rows or columns, and one region with both incomplete rows and columns. Since there are no complete columns, the vertical line representing the complete/incomplete column boundary is overlapped with the left margin. The dash line representing complete/incomplete rows is shown in the figure. The first table in the figure shows the basic information about the missingness in data. For example, there are about half complete data locations and half incomplete data locations.

If a dataset with no missing data is required and imputation is not an option, the missing data can be dropped. Because only a complete row or column can be omitted, there is an optimal dataset that contains the most remaining data. Since it costs more data to drop a column than a row (dropping a column eliminates more than 8500 data, while dropping a row only costs 50 data), and there are less missing data in columns close to the left margin, variables are looped starting from the very left column.

1. In each column, find the rows containing missing data.
2. If dropping these rows costs less data than dropping the column, drop the rows. Otherwise, drop the column.
3. Move to the next column. Repeat the process on the clipped dataset.

In the early stage of the procedure, rows are dropped. For many variables in the middle of the data map, there are no missing data to be found as the missing rows already clipped. When the loop reaches the last five variables, dropping the columns costs less data than dropping the missing rows. The columns and rows to be dropped are highlighted light blue in Fig.2.2. The information about the optimization is tabulated in the second table. There are 51 variables, and 5 variables are dropped. 90% of the variables are preserved. There are 433602 data, and 46466 data are dropped. The optimal dataset has 89% remaining data. This algorithm for choosing a large number of homotopic data aims at preserving the most data after dropping the missing data, and it may be overridden by understanding that some observations or some variables are important so they should not be dropped.



	Num. Complete	% Complete	Num. Incomplete	% Incomplete
Variable (Column)	0	0.0	51	100.0
Locations (Row)	4096	48.18	4406	51.82

	Num. total	Num. dropped	Num. Optimzed	% Available
Variable (Column)	51	5	46	90.2
Locations (Row)	8502	86	8416	98.99
Total data	433602	46466	387136	89.28

Figure 2.2: The plot of the data and highlighted missing part. The figure is divided by dashed lines showing the complete and incomplete datasets.

2.3 The nature of missingness

In this section, a method to explore the nature of missingness is introduced and the results are visualized by three different plots. For comparison, further analysis considering the relevance between variables and the size of missing data is conducted on the observed results. As observed in Fig.2.3, the last five columns (missing variables) are to be dropped. Doing so excludes more than 40,000 potentially useful data.The missing data should be imputed. If the missingness is MCAR, the missing data can be imputed by traditional methods. Otherwise, more advanced techniques (BU, GMM) should be applied.

To understand the nature of missingness, missing variables are compared with non-missing variables (the variables kept after optimization). Consider Fig.2.4. Sn is the missing variable and Ag is the non-missing variable. The two subsets of Ag where Sn is present and absent are compared, and they are denoted $\mathbf{X}_{Ag|Sn} = \{Z_{\alpha,Ag}|Z_{\alpha,Sn} = \mathbb{R}\}$ and $\mathbf{X}_{Ag|NoSn} = \{Z_{\beta,Ag}|Z_{\beta,Sn} = NAN\}$

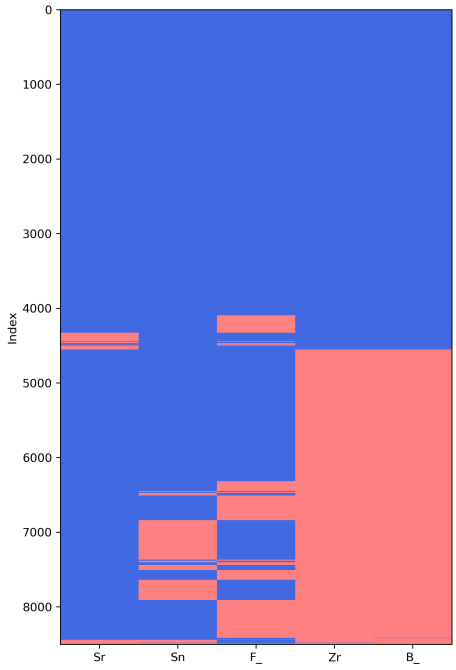
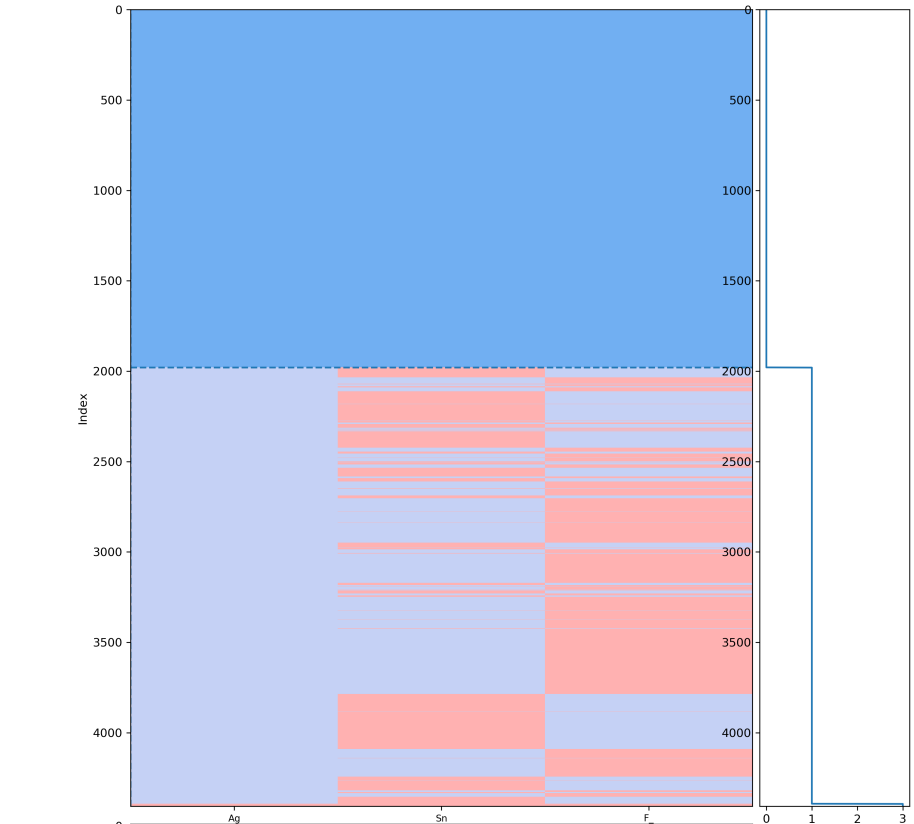


Figure 2.3: The zoomed-in plot of columns (variables) to be dropped.

respectively, where α and β refer to data locations (rows). If the two subsets show very different patterns, we may conclude the missingness is not random. The reason for this comparison is that the collocated data variables are related as they are collected at the same location. The reason for not comparing between the missing variables is clear: when Sn is present F could be absent. This decreases the available data for comparison.

2.3.1 Quantitative measurement

The common quantitative approaches to measure the difference between the two distributions $\mathbf{X}_{Ag|Sn}$ and $\mathbf{X}_{Ag|NoSn}$, such as comparing the mean or the median, ignore the shape of distributions. KS test (Young, 1977) solves this problem. KS test measures the maximum distance d between the cumulative distribution function (CDF) of different distributions. By definition, the KS result $d \in [0, 1]$. The bigger the d value, the more different the two distributions are. As shown in Fig.2.5, the maximum distance between the CDFs is marked by the blue line. The two distributions are Gaussian distributions with different means and standard deviations. The red distribution has a mean of 1.04 and a standard deviation of 0.96, and the gray distribution has a mean of 1.82 and a standard deviation of 2.04. The two distributions are different, so the maximum distance d is equal to 0.31. Since the CDFs are compared, the center and the shape of the distributions are both considered. The problem of KS-test is that different sample sizes could lead to artificial errors such as high value due to few data. Another issue is that different variables retain different baseline d, so it is difficult to set a threshold d to distinguish MCAR and MAR for all variables.



	Num. Complete	% Complete	Num. Incomplete	% Incomplete
Variable (Column)	0	0.0	3	100.0
Locations (Row)	1980	44.94	2426	55.06

	Num. total	Num. dropped	Num. Optimzed	% Available
Variable (Column)	3	0	3	100.0
Locations (Row)	4406	2426	1980	44.94
Total data	13218	7278	5940	44.94

Figure 2.4: The partial observations of three variables from the complete dataset.

The permutation test (Odén, Wedel, et al., 1975) is proposed, combined with KS test to solve the issues above. Suppose there are two subsets $\mathbf{X_1}$ and $\mathbf{X_2}$ with n_1 and n_2 data respectively. The whole sample $\mathbf{X} = \mathbf{X_1} + \mathbf{X_2}$ with $n = n_1 + n_2$ data. First, $\mathbf{X_1}$ and $\mathbf{X_2}$ are compared. Then n_1 and n_2 samples are randomly drawn with no replacement from $\mathbf{X}$ for m times as if they were from the same population. For each iteration i, the n_1 subsample is denoted $\mathbf{X_1^i}$ and the n_2 subsample is denoted $\mathbf{X_2^i}$. The m pairs of $\mathbf{X_1^i}$ and $\mathbf{X_2^i}$ are compared. If the comparison of the observed subsamples $\mathbf{X_1}$ and $\mathbf{X_2}$ shares similar features with the comparison of m pairs , $\mathbf{X_1}$ and $\mathbf{X_2}$ may come from the same population. Otherwise, they belong to different populations. KS test is conducted on $\mathbf{X}_{Ag|Sn}$ and $\mathbf{X}_{Ag|NoSn}$, obtaining d_{obs}, and the sample sizes are n_1 and n_2 respectively. Then n_1 and n_2 data are sampled from $\mathbf{X}_{Ag}$, obtaining $\mathbf{X}_{Ag1}$ and $\mathbf{X}_{Ag2}$. The KS test is conducted on the random sampled subsets to obtain d_i. The resampling is iterated 1000 times (adequate in this case), and a pool of $\mathbf{d} = [d_1, d_2, \cdots, d_{1000}]$ is obtained. The observed d_{obs} is compared with $\mathbf{d}$ to decide how different the two observed samples really are. Fig.2.6 shows the results of the permutation test. The vertical

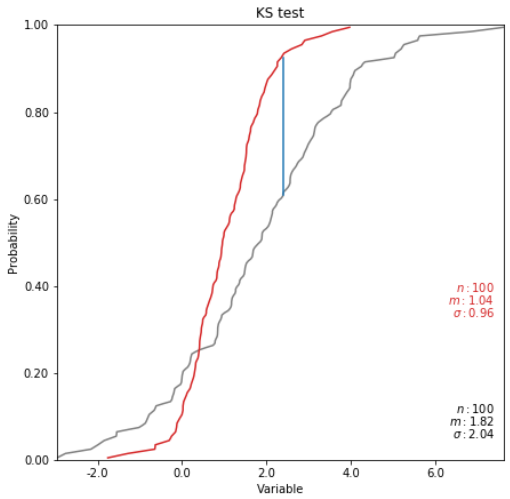


Figure 2.5: KS test for two distributions. The blue vertical line is the result d.

orange line shows the value of d_{obs}, and the sampled $\mathbf{d}$ form the blue histogram. The histogram of $\mathbf{d}$ shows if two distributions are sampled from the same population, their difference in KS test should be around 0.03. The observed value is far from the cluster of random samples, so the subsets $\mathbf{X}_{Ag|Sn}$ and $\mathbf{X}_{Ag|NoSn}$ are not randomly drawn from the same population, and the missingness in Sn may be systematic.

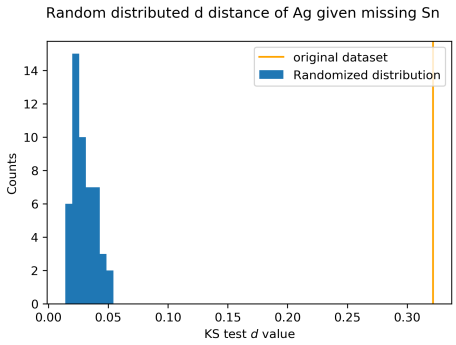


Figure 2.6: The distibution of $\mathbf{d}$. The orange line represents d_{obs}

Sn can be compared with other 46 non-missing variables and obtain a cumulative measurement of the systematic missingness. The same procedure is conducted on the other 4 missing variables. The cumulative results of different missing variables show which missing variable has the most systematic missingness. Since different combinations of missing and non-missing variables can generate multiple sets of $\mathbf{d}$, a universal quantitative measurement of the missingness that applies to all combinations is needed. This measurement is denoted as p and p is calculated as:

$$p = \frac{d_{obs} - d_{mean}}{\sigma_d}$$

where d_{mean} and σ_d are the mean and standard deviation of $\mathbf{d}$. p measures how many standard deviations the observed d is from the mean of $\mathbf{d}$. $\mathbf{d}$ represents the set of the KS test results when two subsets are drawn from the same population. In this way, the d_{obs} is standardized and a threshold value p can be set as the distinction between MCAR and MAR. Algorithm 1 summarized procedure to calculate the cumulative measurement of the missingness in pseudocodes.

Algorithm 1 Measaurement of missingness

```
L_m is a list of missing variables, and L_n is a list of non-missing variables
for every element B in the list L_m do
    Create an empty list p_B
    for every element A in the list L_n do
        Find the subset X_{A|B} and X_{A|NoB}, and their size n_1 and n_2
        Calculate d_obs using KS test
        Create an empty list d
        for i from 1 to 1000 do
            Sample two subsets X_{A1} and X_{A2} with size n_1 and n_2
            Calculate d_i using KS test on X_{A1} and X_{A2}
            Add di to the list d
        end for
        Calculate the mean and variance of d
        Calculate p_{B|A}
    end for
    Add p_{B|A} to the list p_B
end for
Configure a Dataframe D using p_B as columns
return D
```

2.4 Results visualization

2.4.1 p measurement

	Sr	Sn	F	Zr	B
Ag	22.361275	33.367028	31.659201	17.397683	17.397683
Fe	23.506660	32.704326	19.658435	25.118166	25.118166
Zn	17.654142	30.638435	24.279118	18.707961	18.707961
Mo	15.892016	23.819519	20.156665	7.322113	7.322113
Na	18.245105	13.962434	2.526203	11.726063	11.726063

Table 2.1: The first 5 rows of the table of p value.

Table.2.1 shows the dataframe obtained from Algorithm 1. Columns are the missing variables and rows are the non-missing variables. p shows how different the two subsets of the non-missing variables are. $p_{Sn|Ag} > p_{Sn|Na}$ means Ag indicates the systematic missingness more strongly than Na. The dataframe is also plotted in three different formats (Fig.2.7). Each of the three plots focuses on different aspects. The first plot shows the top 5 non-missing variables giving the highest p for each missing variable, and the cumulative p is also shown. The second plot arranges the non-

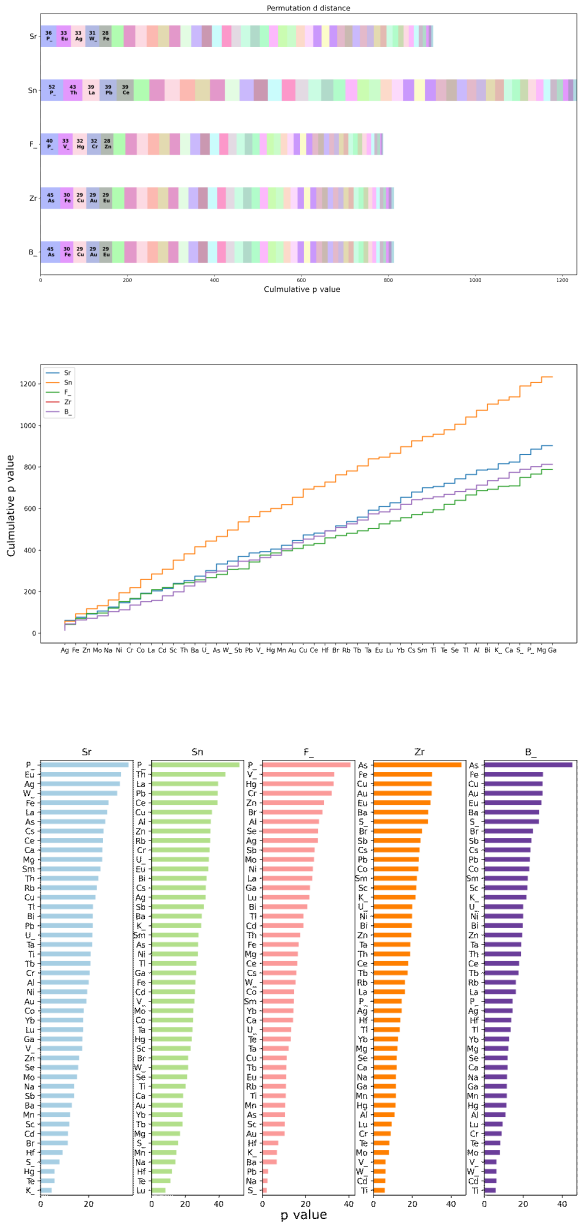
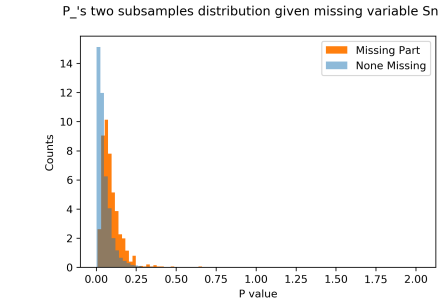


Figure 2.7: The p value for all the combinations of missing and non-missing variables.

missing variables in the same order. The magnitude of p value for a specific combination is easy to find. The third tornado chart ranks the non-missing variables. It is convenient to find the ranking of the non-missing variables and their corresponding p value, but it is hard to find a specific combination and the cumulative p value on the tornado chart. As observed from the figure, Sn exhibits the most systematic missingness. P is the variable that indicates the missingess in Sr, Sn and F most. The difference of the subsets of P is examined in Fig.2.8. The blue histogram shows the distribution of $\mathbf{X}_{P|Sn}$, and the orange histogram shows the distribution of $\mathbf{X}_{P|NoSn}$. Note the numbers of

data are different, and the histogram is normalized. The statistic table shows the centers of the two distributions are significantly different, which validates the high p value of this combination



	Mean	Standrard deviation	Num used
Missing Part	0.09	0.0037	930
None Missing	0.05	0.0041	7486

Figure 2.8: The histograms of the subsets of P given the missing variable Sn.

2.4.2 p considering variable relevance

The results above measure the difference between the two observed subsets of the collocated variables, but it does not necessarily indicate the difference between observed and missing data. For example, $\mathbf{X}_{P|Sn}$ and $\mathbf{X}_{P|NoSn}$ are significantly different but the difference between $\mathbf{X}_{Sn}$ and $\mathbf{X}_{NoSn}$ may not be significant if P and Sn are unrelated. Since the collocated data are related, this relevance can link the measurement of non-missing variables with missing variables. At the locations where P and Sn are both present, equal size KS test is applied to find the relevance between the two variables. Since different variables share different units, to compare their CDFs, each variable is standardized. In this case, KS test mainly identifies the shape difference. For variable pairs sharing low d value, the two distributions share similar shape, and the relevance between variables is calculated as $r = 1 - d$. The p value considering the relevance is denoted as pr and is calculated as $pr = r \cdot p$ (e.g. $pr_{Ag|Sr} = r_{Ag|Sr} \cdot p_{Ag|Sr}$). The pr value is plotted in Fig.2.9. The relative cumulative pr value between missing variables is adjusted. Zr has the lowest cumulative pr value, which means Zr has less relevance with the other variables. Sn still shows the most systematic missingness, but the sequence of the important non-missing variables have adjusted. This can be revealed from the tornado chart. The most significant non-missing variable for Sr changes to Ag, which means the difference between $\mathbf{X}_{Ag|Sr}$ and $\mathbf{X}_{Ag|NoSr}$ are more representative for $\mathbf{X}_{Sr}$ and $\mathbf{X}_{NoSr}$.

2.4.3 pr considering missing size

Furthermore, the relative size of missing and non-missing data can be considered. If the missing data take up only 5% of the total data, the missingness may not be a major concern, so the measurement of missingess should be low and this measurement is denoted as pn. Note, the low value does not mean the missingness is random. In Fig.2.9, Sr has a higher cumulative pr value than Zr

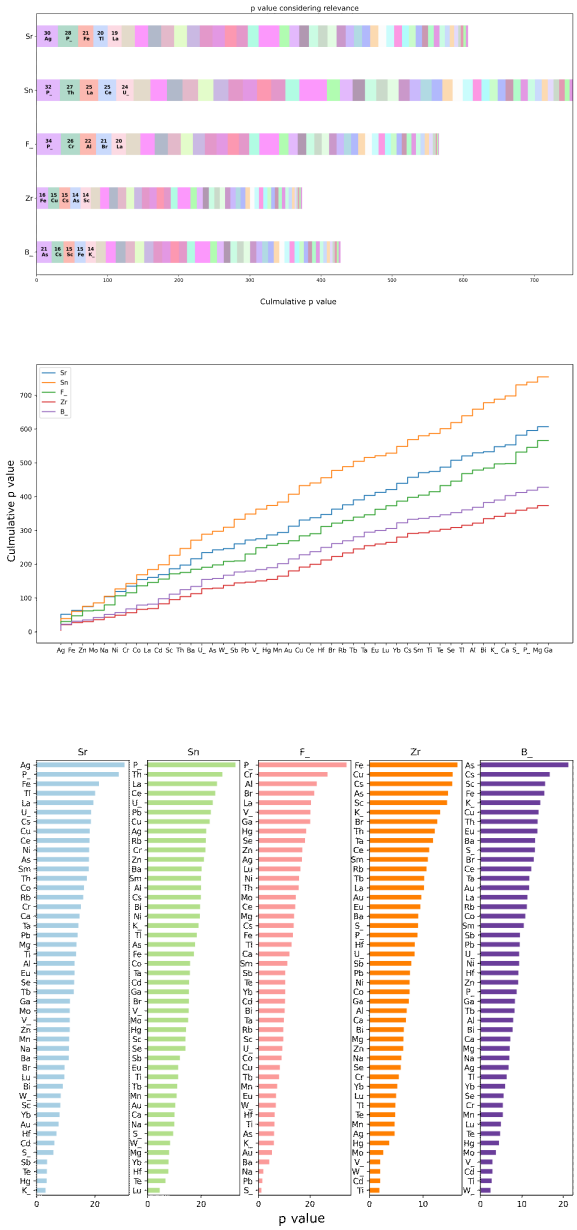


Figure 2.9: The pr dataframe after combining the p value with the relevance.

while the missing data size of Sr is 20 times less than that of Zr. Taking this factor into consideration, the pr value of each non-missing variable is divided by the ratio of non-missing and missing size. For example, Sr has 8000 non-missing data and 200 missing data, so the ratio is 40. Each pr value in Sr is divided by 40. Fig.2.10 shows the plots after this adjustment. The relative importance of non-missing variables does not change for missing variables. The ranks of cumulative pn change significantly compared with that of pr. Sr and Sn have the least pn, indicating their systematic missingness is not very concerning considering their missing data size. Observing the middle plot, the

magnitudes of missingness can be divided into three parts. The minimum is Sr. It has the least missing data and the cumulative pn is smaller than 10. The second part includes F and Sn. Their cumulative pn range from 50 to 100. Zr and B have the most pn which is above 250, indicating their systematic missingness is the most concerning among missing variables.

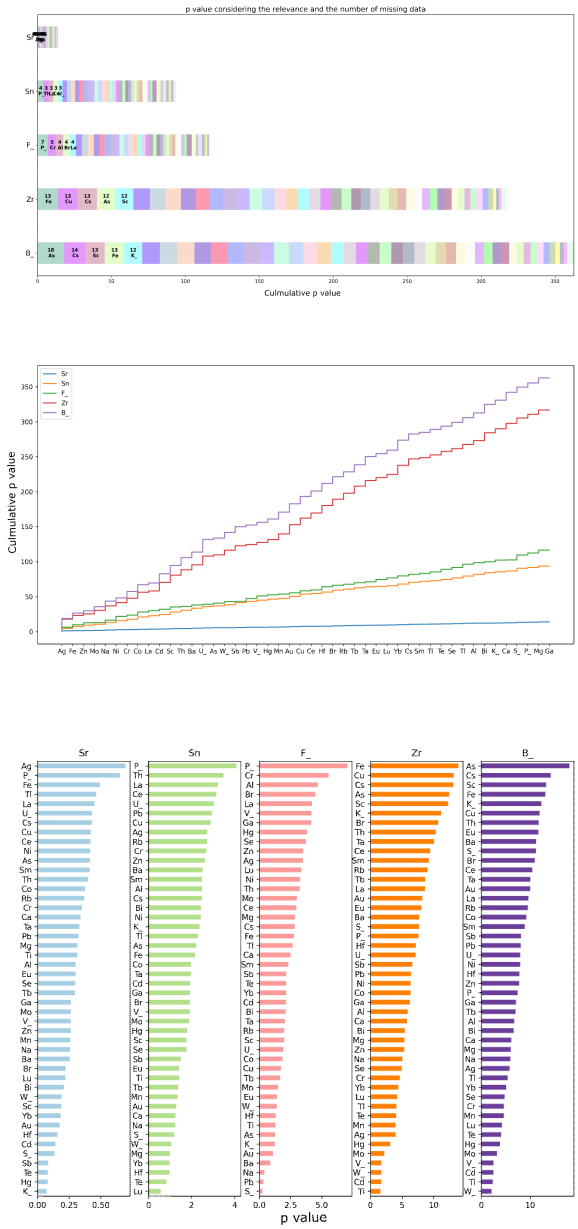


Figure 2.10: The plots of dataframe showing the level of missingness considering the missing size and relevance.

2.5 Method Validation

In this section, a synthetic dataset containing systematic and random missing data is used for validating the previous methods. A full dataset which contains no missing data is generated. Then, data are dropped randomly and systematically. If the method is robust, it should identify the basic information and the mechanism of the missingness correctly. The results show the method is robust and the mechanism of systematic missingness is also understood.

The full dataset to start with is the optimal dataset obtained in Fig.2.2, because the variables in the synthetic data should be intrinsically related. 20 variables are randomly drawn from the non-missing variables, and the data observations are shuffled. Variable Sc and W are dropped randomly with the missing size of 500 and 3000 respectively. Variable Th, Cu, Ca and La are dropped systematically as follows:

$$\mathbf{z}_{Th} \in (-\infty, m_{Th}/3] \cup [1.5m_{Th}, +\infty)$$

$$\mathbf{z}_{Cu} \in (-\infty, m_{Cu}/4] \cup [1.7m_{Cu}, +\infty)$$

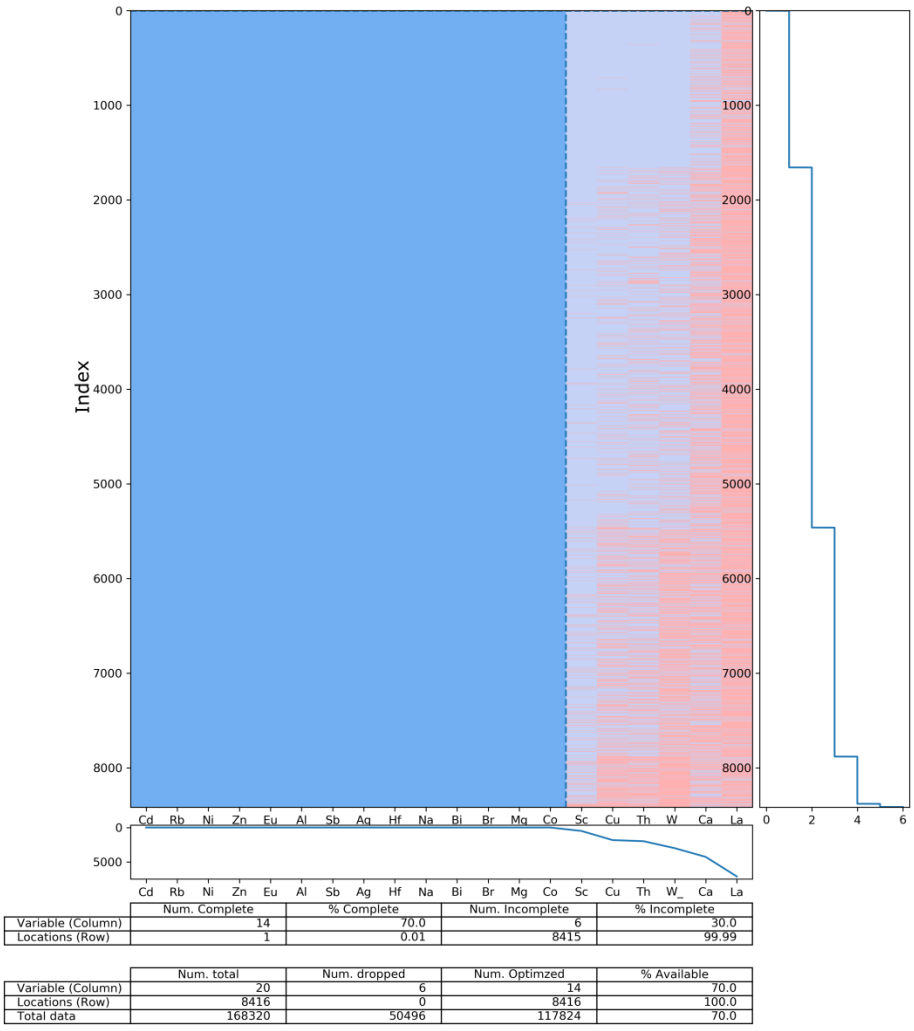
$$\mathbf{z}_{Ca} \in [\frac{3}{4}m_{Ca}, 1.5m_{Ca}]$$

$$\mathbf{z}_{La} \in [\frac{1}{4}m_{La}, 1.2m_{La}]$$

where $\mathbf{z}_i$ represents the data list dropped in variable i (i= Th, Cu, Ca, La) and m_i is the full dataset mean of variable i. For example, in variable Th, the data below one third of the mean and above 1.5 times of the mean are dropped.

2.5.1 Missing data map

The synthetic dataset is ordered based on the size of missingness, and Fig.2.11 shows the data map of the general missingness information. The right side of the figure has the columns with the most missing data, and the bottom of the figure have locations containing the most missing data. 14 variables are complete and 6 variables have missing data. The smallest size of missingness is in variable Sc as only 500 data samples are dropped. Ca and La have the most missing data as the majority of the data around mean value are dropped. The general missingness information is consistent with the way the synthetic data are generated. To obtain the optimal dataset with no missingness, all missing columns are dropped, and no rows need to be dropped. The synthetic data map does not have patterns in the observed one where Zr and B have missing data at the same locations, because data are dropped independently in each variable.



	Num. Complete	% Complete	Num. Incomplete	% Incomplete
Variable (Column)	14	70.0	6	30.0
Locations (Row)	1	0.01	8415	99.99

	Num. total	Num. dropped	Num. Optimzed	% Available
Variable (Column)	20	6	14	70.0
Locations (Row)	8416	0	8416	100.0
Total data	168320	50496	117824	70.0

Figure 2.11: The synthetic data map. The highlighted columns are to be dropped.

2.5.2 Permutation test

Table 2.2 shows the KS test results of d_{obs} for all combinations of missing and non-missing variables. The smaller the d is, the closer the two CDFs are. Sc and W have relatively low d value as the missing data are dropped randomly, and other missing variables have d value at least 10 times larger, which implies the missingness is systematic. The results are consistent with the design of the synthetic data. The permutation test results (Fig.2.13) show a relatively low p value in variables Sc and W. p implies how much non-missing variables reflect the difference between the missing and observed data of a missing variable. The results also reveal that the variables missing the middle data ranges (Ca, La) have higher p than the variables missing the outer data ranges (Cu, Th). The magnitude of Cu and Th are similar to that of the missing variables in the real dataset. This implies the missing data in the five variables (Sr, Sn, F, Zr, B) lie in the outer data ranges. The difference of d_{obs} and the set of random sampled subsets are shown in Fig.2.12. d_{obs} of Sc fall within the majority of the

	Sc	Cu	Th	W	Ca	La
Cd	0.022855	0.174078	0.115534	0.026534	0.135357	0.428173
Rb	0.027880	0.237570	0.532445	0.018968	0.644970	0.944102
Ni	0.035398	0.348219	0.147874	0.013417	0.397736	0.760646
Zn	0.037868	0.246505	0.080256	0.036658	0.339642	0.560598
Eu	0.015316	0.287281	0.379795	0.007999	0.547218	0.442209
Al	0.047457	0.312095	0.362104	0.017139	0.522110	0.876179
Sb	0.035239	0.216004	0.112636	0.033488	0.214670	0.481931
Ag	0.063809	0.233299	0.105138	0.025822	0.400627	0.664889
Hf	0.036257	0.212964	0.257828	0.014200	0.615349	0.839151
Na	0.023399	0.137106	0.189646	0.015000	0.449543	0.613481
Bi	0.028937	0.233261	0.542790	0.025964	0.612153	0.874565
Br	0.042346	0.127733	0.123594	0.029850	0.220423	0.436009
Mg	0.047230	0.170895	0.264166	0.015042	0.779581	0.867649
Co	0.040629	0.364067	0.299488	0.014464	0.505220	0.884321

Table 2.2: Observed KS test results d_{obs} for the synthetic dataset.

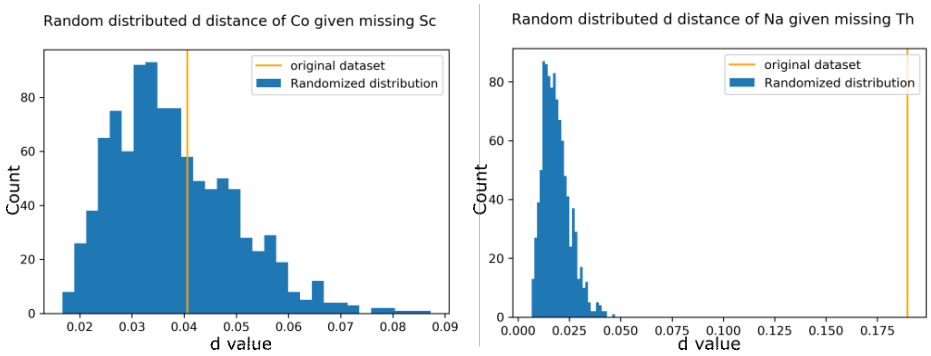


Figure 2.12: The histogram plots of **d** distributions for different variables

permutation subsets (**d**) whereas the d_{obs} of Th is far from the permutation subsets **d** because the missing data in Th are dropped systematically. It also explains the higher cumulative p in systematic missing variables compared with random missing variables.

Fig.2.13 shows the plots of the cumulative p of the synthetic data, and they illustrate the missingness in variables correctly. Take the relevance between variables and the size of missing data into consideration. The pn results are plotted in Fig.2.14. For example, La has the most missing data and the missingness in La is the most systematic, so La has the largest cumulative pn. The two randomly dropped variables have fairly small cumulative pn. Each one of the results in the section captures different aspects of the missingness in the synthetic dataset. The number of missing variables, random and systematic missingness, and the missing sizes are illustrated. Thus, the proposed method is robust at assessing the property of missingness in a dataset.

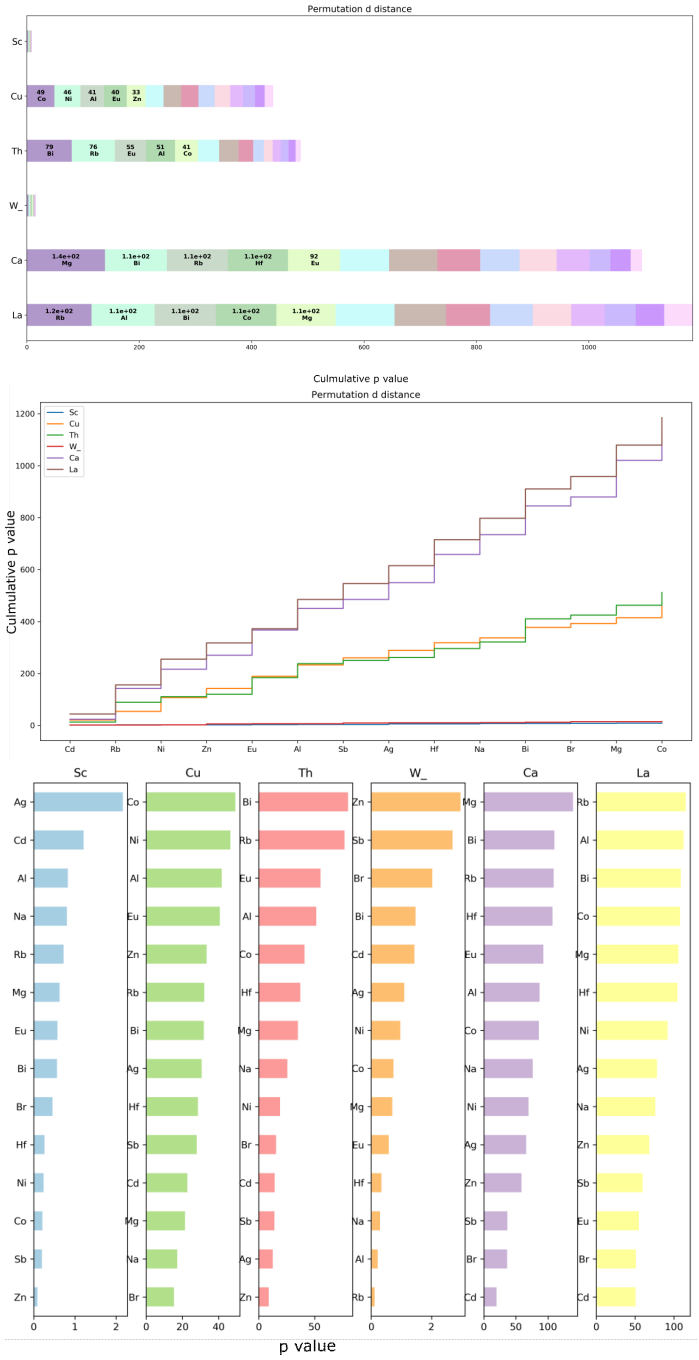


Figure 2.13: The results of the synthetic data p from permutation test.

2.6 Discussion

Although MNAR is difficult to analyze statistically, it can be observed from the data map. Comparing the two data maps (Fig.2.2 and Fig.2.11), the major differences are in the ordering of the missing data. In the real data map, missing data can be well clustered into the right bottom corner while the synthetic data do not possess this feature. That is because the synthetic missing data are

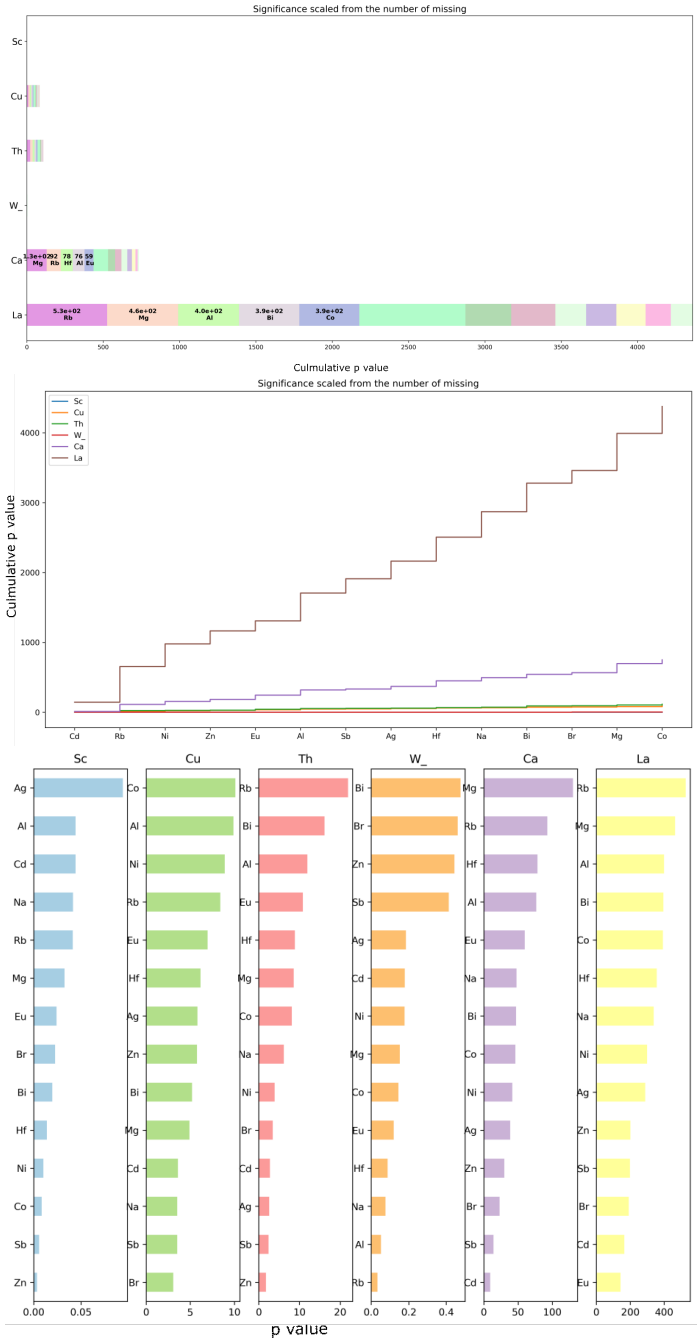


Figure 2.14: The results of the synthetic data from permutation test considering missing size and relevance.

dropped independently for each variable, while the missingness in real data occur at the same locations. Especially for Zr and B that are missing together, their availability depend both on observed and missing data. When the missing data can be clustered as in Fig.2.2, the missingness is likely to be MNAR.

The reasons for using permutation test rather than bootstrap are that permutation is used to test null hypothesis whereas bootstrap is used to obtain confidence intervals. The result of KS

test (d) is the property between two subsets, while the confidence interval is calculated within one subset. The null hypothesis here is the two distributions $\mathbf{X}_{A|B}$ and $\mathbf{X}_{A|NoB}$ are drawn from the same population. If the hypothesis holds, the missingness in B is random. Otherwise, the missingness may be systematic. Moreover, the reason for using the permutation test combined with the KS test is that different sample sizes and variables generate multiple sets of observed $\mathbf{d}$. When $d_{obs} = 0.1$ represents significant difference in one variable, it may not be significant in another. $d \in [0, 1]$ but a threshold is needed to distinguish the random and systematic missingness for all variables. When combined with the permutation test, the value of d_{obs} is calibrated and the p value is a universal measurement. So a threshold can be set for p. From the observations of this dataset, it is recommend that the missingness can be viewed as systematic when $p > 10$.

The robustness of the method is validated by the synthetic data. Different patterns of systematic missingness lead to various magnitudes of p value, and this helps understand the mechanism of missingness in the real data. For example, the variables missing outer data ranges (Cu, Th) share similar p value with missing variables in the real dataset (Sr, Sn, F, Zr, B). This implies the missing variables in the original dataset may also miss the outer data ranges. So the mechanism (missing which data ranges) of the missingness may be explored by dropping different data ranges of the synthetic data and comparing the p values with the original ones. Note, this only works when the synthetic missing data are generated from the same real data as the relations between variables remain the same.

2.7 Conclusions

Missing data come from multiple sources. They can cause problems in data analysis such as PCA. Different data imputation methods can be applied depending on the types of missingness, so it is important to understand if the missingness in data is random or systematic. The technique of missing data exploratory analysis performs well on the Northwest Territory data and it has two major tools. First, it illustrates the general information of the missing data, that is, the number of missing variables, boundary of missing and non-missing data, and the optimal dataset. The second tool is the KS-permutation test that identifies the nature of missingness. The three types of results, p, pr, and pn, convey different information about the missingness. The major information conveyed in Fig.2.7 is the cumulative p value that indicates if the missingness is random. The pr value (Fig.2.9) identifies the most relevant non-missing variable that could be used for data imputation. The scaled pn plots (Fig.2.10) show the most concerning missing variable considering the missing size. The robustness of the method on this dataset is validated through a synthetic dataset generated from the real data with no missing data. Variables dropping missing data randomly and systematically are well identified.

Chapter 3

Below Detection Limit Data

Below detection limit (BDL) data are data lower than the minimum detection limit of the laboratory measurement equipment and recorded as 0.0 or identified as "BDL" in the database. These values may form a spike in histograms and can cause problems when conducting further analysis. Random despiking and local average despiking are applied in practice. Before despiking, the nature of BDL data and data spikiness should be understood. In this chapter, univariate and bivariate analyses are conducted on the data from the Northwest Territories (Falck et al., 2012). A tabulated summary of the univariate data analysis complements conventional histograms. Important information includes the proportion of BDL occurrence in each variable, which is used in bivariate analysis. The spikiness of data are also examined. Three quantitative measurements are developed to show different aspects of spikiness. The purpose of the bivariate analysis is to determine if the BDL occurrence between variables are independent. A bivariate plane is divided into four regions by BDL boundaries. The observed bivariate proportions are compared with the expected probabilities. The expected probabilities represent independent BDL occurrence and are obtained by sampling a standard Gaussian space considering correlations. Kullback–Leibler (KL) divergence is used to measure the difference between the distributions. The resulting measurement is standardized by the practical maximum, and multiple combinations of variables show strong dependence in BDL occurrence.

3.1 Introduction

In geochemical data, there are often data below detection limit. The concentration of some elements is so low that they are beyond the detection capability of the measurement equipment, such as inductively coupled plasma (ICP) (Thompson, 2012), thermal ionization mass spectrometry (TIMS) (Richter & Goldberg, 2003) and Energy-Dispersive X-ray spectroscopy (EDS) (d'Alfonso, Freitag, Klenov, & Allen, 2010). The BDL data are recorded as either 0.0 or the minimum detectable value. In either case, there are many duplicated data and they may form spikes in the distribution, which could be problematic for exploratory data analysis and modeling (M. J. Pyrcz & Deutsch, 2014). In cluster analysis, data are often normal score transformed to eliminate the effects of outliers and scale variables to the $N(0,1)$ range (Prades, 2017). A problem arises in how to handle the BDL data during transformation so that cluster analysis gives reasonable results. Since normal score transformation uses quantile to quantile transform, there are two evident ways to handle the BDL spikes. One is to spread the spike over a range of the normal distribution, in which case, each data

has a unique Gaussian data value. The other way is to retain the spikes, and the BDL data share the same rank. For example, if 30% of the data are BDL, they are all at the 0.3 quantile or some arbitrarily low value. With spikes preserved, the transformed units behave similarly to the original units, so they are suitable for centroid based clustering (k-means), but not suitable for distribution based clustering (Gaussian Mixture Model) (Prades, 2017). When spikes are spread, the transformed data are not suitable for centroid based method, because the cluster centroids are shifted, and the spread spikes distort the relative distance between data. Suppose a spike consist of 50% data is transformed to Gaussian space. The original data are all 0, but most of the spread BDL data are distributed from -3 to 0.

There are two major methods to spread the spikes: random despiking and local average despiking (M. J. Pyrcz & Deutsch, 2014; Verly, 1984). These two methods have their limitations of a too high or too low nugget effect respectively. With random despiking, the quantiles of BDL data are assigned randomly. Thus, the variogram have a high nugget effect. With local average despiking, the BDL data are ranked based on averages of surrounding data. High local averages give the BDL data high ranks, while low local averages give data low ranks. This may cause the transformed data to be too smooth spatially, and a low nugget effect. Prades (2017) proposes to combine the local and random despiking methods. BDL data are ranked based on local average first, and data value are assigned from BDL value X_1 to the nearest value X_2 incrementally. A random value ranges between X_1 and X_2 is added. Its weight is controlled by a hyper-parameter W_1. The results show setting W_1 to 0.5 can achieve a suitable trade-off between random despiking and local despiking.

Before applying the methods to handle spikes in data, the first step is to understand the nature of spikes and this helps choose the despiking method. In this chapter, BDL spikes and duplicate data spikes are analysed with univariate and bivariate methods. First, the univariate distribution is summarized by an information table to overcome the binning effect of histogram plots. Three measurements of spikiness are developed to reveal different types of spike distributions, including few spikes containing many data in each spike, many spikes containing few data, and spikes different from the expected distribution. The bivariate analysis uses KL divergence (Kullback & Leibler, 1951) to measure the discrepancy between the observed bivariate BDL distribution and an independent bivariate distribution. When the occurrence of BDL data are dependent, further investigation can be done to understand the relation between variables. The same Northwest Territory data from Section.2.1.2 (Falck et al., 2012) are used for demonstration.

3.2 Univariate Analysis

For univariate analysis, it is easy to plot histograms and examine the distributions visually. Histograms show the data distribution over a range of values. BDL data are binned with surrounding low value data, which makes BDL spikes less obvious. A statistics table is created to show informa-

Variables	Min Value	BDL Num.	Aval. Data	Sec. Min. Value	Sec. Num.	Average	Ave. Exclude Min.
Au	0.0	6981	8486	0.30	1	0.96	5.41
B	0.0	2416	4554	1.00	342	1.79	3.81
Ba	0.0	462	8490	50.00	10	1357.60	1435.73
Bi	0.0	469	8441	0.02	328	0.26	0.28
Br	0.0	183	8466	0.50	18	4.11	4.20
Ce	0.0	191	8466	5.00	43	58.76	60.11
Cs	0.0	600	8466	0.50	102	4.23	4.55
Eu	0.0	5276	8466	1.00	1493	0.68	1.83
Hf	0.0	944	8466	1.00	438	6.47	7.28
Hg	0.0	703	8486	5.00	126	41.06	44.77
Lu	0.0	3831	8466	0.20	534	0.25	0.46
Na	0.0	122	8490	0.001	144	0.01	0.01
Rb	0.0	122	8466	5.00	37	71.75	72.80
S	0.0	2159	8441	0.01	158	0.09	0.12
Se	0.0	594	8445	0.10	585	0.95	1.02
Sn	0.0	2983	7515	0.10	6	1.34	2.22
Ta	0.0	2745	8466	0.50	256	0.75	1.11
Tb	0.0	3165	8466	0.50	344	0.62	1.00
Te	0.0	3269	8445	0.02	1114	0.029	0.04
Ti	0.0	694	8445	0.001	2008	0.007	0.008
W_	0.0	5384	8490	1.00	1517	1.51	4.14
Yb	0.0	4618	8466	2.00	1190	1.47	3.23
Zr	0.0	3085	4557	200.00	31	133.04	411.88

Table 3.1: Univariate distribution informaion for each variable. The shortened column names are explained in the context.

tion regarding the BDL data to complement the histograms.

There are other smaller spikes due to the round-off errors. These spikes cause similar problems as BDL spikes. Three quantitative measurements are developed to reveal the spikiness of data, including the number of spikes and the sizes of spikes. The three measurements focus on different aspects of the spikiness.

3.2.1 BDL table

The Northwest Territory data consist of 51 variables and about 8500 data samples. Most of the variables contain BDL data. Table.3.1 shows information about the variables containing more than 100 BDL data. The relative size of the BDL spike influences the performance of quantile transform. The first column shows the names of variables. The second column represents the recorded value of BDL data. Here, all BDL values are recorded as 0.0. The third and fourth columns show the number of BDL data and the available data. The number of available data varies as some variables contain missing data. Fig.3.1 shows the percentage of the BDL data in each variable, along with the number of BDL data and the number of available data. The variables are ranked based on their BDL percentage. Au has the most BDL data, so it is ranked first. Zr has less BDL data but also ranked the second because the available data in Zr is only around 5000. The proportions of the BDL data are used in bivariate analysis.

The 'Sec. Min. Value' column means the minimum detectable value. Note different variables

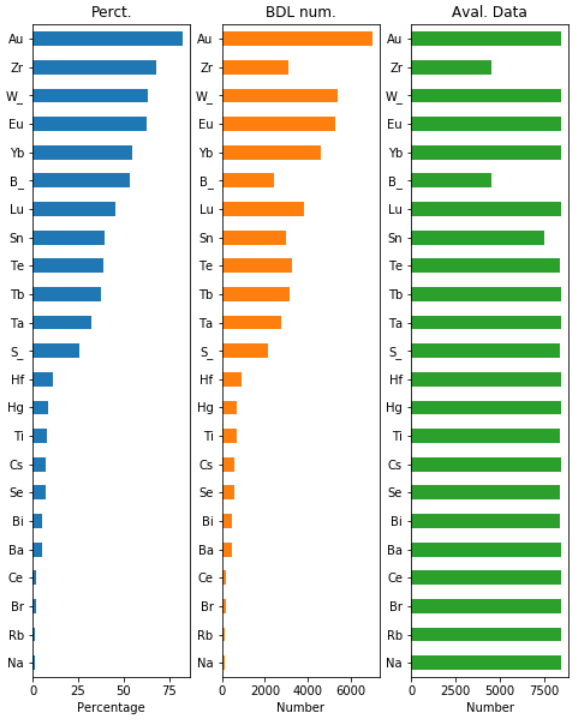


Figure 3.1: The tornado chart of the number of BDL data.

may use different units. Some use ppm and some use ppb. The 'Sec. Num.' column is the number of the second minimum data. These two columns are a measurement of the precision of equipment and the general spike size in the variables. When the number of duplicated data in measured values is small, the BDL spike can be problematic.

The second last column shows the average of data, considering the BDL data. The last column shows the average value excluding the BDL data. In variables which contain many BDL, the average excluding the BDL are significantly higher than the overall average. It measures how far the detectable mean are from the BDL spike. It can also help with despiking. The despiked BDL data should not be far from the detectable mean to retain a realistic data distribution. If the BDL data is too far from the detectable mean, the spike may need to be preserved close to the second minimum value. If the BDL data is close to the detectable mean, the spike may be spread between the minimum and the second minimum value.

3.2.2 Measurement of univariate spikiness

Similar to BDL data, there are other spikes originating from the round up or round down effect. If the precision of equipment is 0.01 ppm, the mineral content of 1.122 ppm and 1.123 ppm are both

recorded as 1.12 ppm. Unlike the large spikes created by BDL data, these smaller spikes are not easy to identify from histograms. So the number of spikes and the size of spikes need to be measured quantitatively. These features are the spikiness of data.

There are two major types of spikes, which are demonstrated in Fig.3.2. One is the "few spikes but many data" type. In the distribution of Au, there are only several spikes but they take more than 70% of the available data. The other type is the "many spikes but few data" type. In the distribution of Fe, there are multiple spikes and each spike takes only about 5% of the data. The histogram illustration may look different because of data are binned together. Different measurements are developed for the two types of spikes.

Quadratic and log method

Suppose there are K variables, and N_i is the number of data at data value i. The number of unique data value is L and each data value is denoted as l. For example, if there are 50 data with a value of 0.3, $N_{0.3} = 50$. Only data with N_i larger than 1% of the total number of data are considered as spikes. To identify the variables with few spikes but many data in each spike, the quadratic equation is applied:

$$M_g(k) = \sqrt{\sum_{i=l}^{L} N_i^2(k)}, \quad k = 1, \cdots, K$$

and to identify the variables with many spikes but few data in each spike, the log equation is applied:

$$M_s(k) = \sum_{i=l}^{L} log(N_i(k)), \quad k = 1, \cdots, K$$

where $M_g(k)$ and $M_s(k)$ are the scores of spikiness for variable k. g stands for giant spikes and s stands for small spikes. The value of $M_g(k)$ is dominated by the size of large spikes, while the value of $M_s(k)$ is dominated by the number of spikes. The difference between the two measurements is the relative importance of the size of spikes and the number of spikes. The quadratic term gives large spikes more weight in $M_g(k)$ measurement, and the logarithm term give the number of spikes more weight in $M_s(k)$ measurement.

When using these two methods, the number of duplicate data is more important than the proportion of duplicate data. Fig.3.3 shows the scores of spikiness for all variables, using different measurements. Variables are ranked based on their scores in each method. The top figure shows the scores using the quadratic method, and the bottom figure shows the scores using the log method. Au ranks the first in the quadratic method, which means it has the largest spike. The rank of variables using the quadratic method generally follows the proportion of BDL data in each variable. The higher the BDL proportion, the higher the $M_g(k)$ spikiness score, which is consistent with the purpose of the quadratic method. Cs ranks the first in the log method, which implies it has the most spikes. The variable ranks are generally different than the quadratic method. It is anticipated as

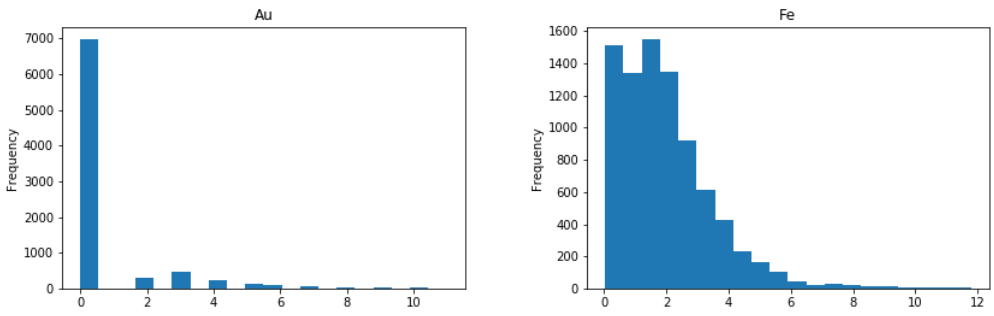


Figure 3.2: Illustration of spikes in Au and Fe.

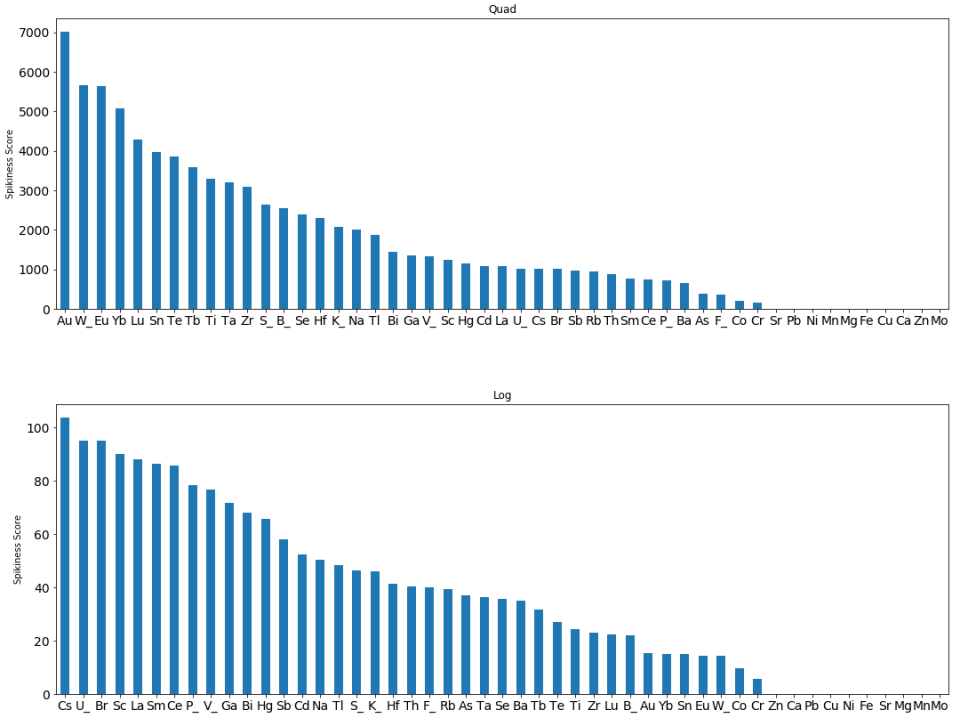


Figure 3.3: Measurement of spikiness using the quadratic and logrithm methods.

there are not many spikes with great size when the number of data is limited. The zero values in both figures indicates there are no spikes in the corresponding variables. Table.3.2 also shows the methods work as expected. The left table shows the quadratic method can find large spikes. The five highest ranked variables all have giant spikes with $N_i > 3000$. The right table shows the log method finds evenly distributed spikes. When the size of spikes are small and evenly distributed, the number of spikes can be large.

	Mes. Spikiness	Five biggest count		Mes. Spikiness	Five biggest count
Au	7,008.18	[6981, 468, 309, 235, 139]	Cs	103.62	[600, 224, 161, 158, 143]
W	5,672.25	[5384, 1517, 917, 216, 120]	U	95.12	[227, 225, 214, 201, 196]
Eu	5,631.11	[5276, 1493, 1252, 294, 84]	Br	95.04	[203, 199, 192, 192, 192]
Yb	5,070.30	[4618, 1543, 1190, 751, 194]	Sc	90.20	[275, 254, 252, 251, 247]
Lu	4,294.60	[3831, 1280, 1042, 663, 534]	La	88.05	[263, 258, 246, 237, 232]

Table 3.2: Results from the two methods. Left using the quadratic equation and right using the log equation.

Kullback–Leibler divergence

Before introducing another type of measurement of data spikiness, Kullback–Leibler divergence (Kullback & Leibler, 1951) needs to be explained. It measures the relative entropy between different distributions. Entropy is also referred to as Shannon's entropy which measures the average level of information conveyed by random variables (Shannon, 2001). The equation is as follows:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{3.1}$$

where X is a discrete random variable and $x_1, \cdots, x_n$ are the possible outcomes. $P(x_i)$ is the probability of x_i. The more information each observation can provide, the higher $H(X)$ is. The equation for the relative entropy is as follows:

$$D(P||Q) = \sum P(x) log(\frac{P(x)}{Q(x)}) \tag{3.2}$$

where $P(x)$ is the observed distribution and $Q(x)$ is the expected distribution. It compares the difference between two discrete distributions. The higher the $D(P||Q)$, the more different they are. When two distributions are the same, the log term is equal to zero, and $D(P||Q)$ is equal to zero, which is also the minimum value. The divergence indicates how much information is lost if Q(x) is used to represent P(x). For example, suppose there is a coin and it is assumed to be fair, so the probability of head and tail are both 0.5 ($Q(head) = Q(tail) = 0.5$). To verify the assumption, the coin is tossed 1000 times with 700 heads and 300 tails ($P(head) = 0.7, P(tail) = 0.3$). The observation diverts from the assumption, and the divergence is calculated using Eq.(3.2).

$$D(P||Q) = 0.7 \log(0.7/0.5) + 0.3 \log(0.3/0.5) = 0.082$$

In the following sections, KL divergence is used to compare the difference between discrete distributions. The distributions can be univariate or multivariate.

Scaled method

The scores of the quadratic and the log methods become higher when the available data increase. To amend this issue, the scaled method can be applied. It measures how different the observed spike distribution is from the expected spike distribution. Here, spikes are defined differently. Instead of using the number of data at value i (N_i), the probability of data at value i (P_i) is used. Suppose

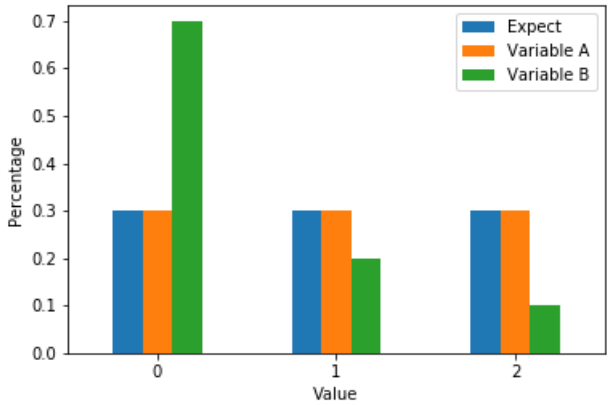


Figure 3.4: An illustration of spikes in the scaled method. Variable A and B represent two random variables.

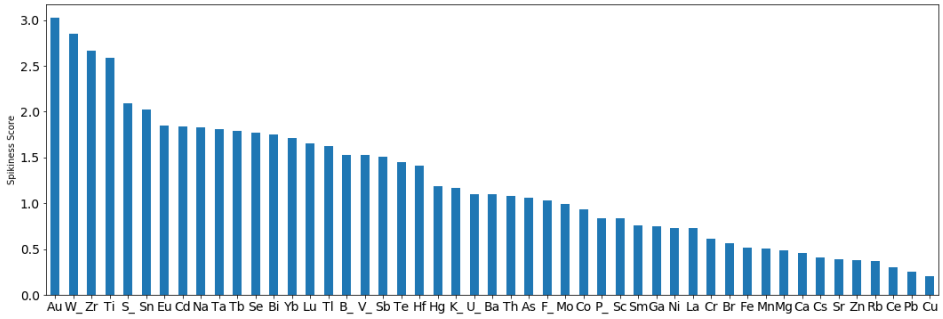


Figure 3.5: The results of the measurement of spikiness, using the scaled method.

there are 1,000 data in total and 200 data have the value of 4.3 ppm. Then $P_{4.3} = 0.2$. If there are 50 unique values in the 1,000 data, the expected probability for each value i is $P_i = 0.02$. This is defined as the expected distribution, which is compared with the observed distribution. For example, in Fig.3.4, there are three possible values. The expected probability for each value (the blue bars) is 0.33. Variable A has the same distribution as the expected one, so there is no spike in A. For variable B, P_i distribution is different from the expected distribution and the difference in measured using KL divergence. The same measurement can be applied to the variables in the real data and each variable has an expected distribution and an observed distribution. The spikiness score for a variable is calculated using Eq.(3.2).

Fig.3.5 shows the results of comparing the observed distribution with the expected distribution using KL divergence. The ranks are similar to those in the quadratic method, but they do not follow the same principle. The first ranked variable Cs in the log method is ranked low in this case. It means the observed spike distribution in Cs is similar to the expected spike distribution.

3.3 Bivariate analysis

The purpose of bivariate analysis is to identify if the occurrence of BDL data between variables are dependent. If so, more investigation can be conducted on the pairs of variables to understand the reasons for BDL occurrence. The bivariate distribution is divided by BDL boundaries into four regions - one region where both dimensions are at BDL, two regions where either one of the dimensions is at BDL, and one region where both dimensions are not at BDL (Fig. 3.6). Based on the univariate analysis in the previous section, the proportions of BDL in variables are known. Assuming two variables are independent, the bivariate distribution of BDL occurrence can be calculated. If the observed probabilities are far from the independent distribution, we may conclude the BDL data occurrence is dependent.

3.3.1 Expected distribution

From probability theory, if two events A_1 and A_2 are independent, the joint probability is simply the multiplication of the probabilities of each event.

$$f(A_1, A_2) = f(A_1)f(A_2)$$

where f represents the probability of an event. Since the probability of BDL for each variable is known, it is easy to calculate the probabilities $P_{ind}(B_1, B_2)$, $P_{ind}(B_1, N_2)$, $P_{ind}(N_1, B_2)$ and $P_{ind}(N_1, N_2)$, where P_{ind} is the probability when two events are independent, and B_i and N_i represent variable i is at BDL and not at BDL respectively. Fig.3.6 shows the four regions in a 2D plane. The univariate BDL proportions are marked by red dashed lines. However, Fig.3.7 shows some variables are correlated. When calculating the expected probability, the correlations have to be considered. The highest negative correlation is -0.16. In this case, P_{ind} should be replaced by P_{exp}, which means the expected probability of BDL considering correlations.

To obtain P_{exp}, a simple way is to sample in multi-Gaussian space and count the number that falls in each region. Dividing the number in each region by the sample size gives the expected probability. Now, the BDL boundaries in Gaussian units given the univariate BDL proportions need to be calculated. This is achieved through quantile transform. The boundary of BDL in Gaussian space is calculated by taking the inverse of Gaussian cumulative distribution function (CDF) given the BDL proportions in the original space

$$b_k = G^{-1}(P(B_k))$$

where b_k is the BDL boundary of variable k in Gaussian space, $P(B_k)$ is the probability of BDL in variable k. Table.3.3 shows the BDL boundaries in Gaussian space for Sn and S. The column "Probability" refers to the probability of BDL. "Gauss Unit" shows the converted BDL boundaries in Gaussian space. Note the Gaussian space is standard. To calculate the expected probability of each region, for example the region where both Sn and S are BDL, samples are below -0.26 for Sn and

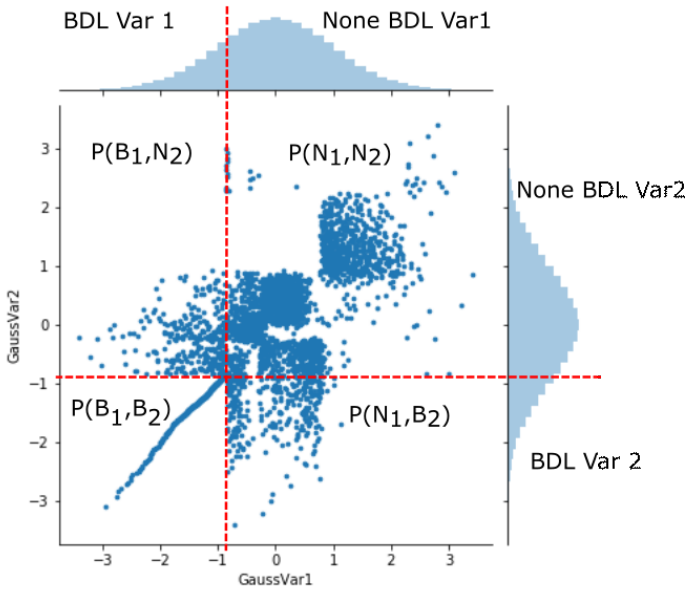


Figure 3.6: Illustration of the four BDL regions in a bivariate setting. The original data are quantile transformed and the spikes are spread. The random despiked BDL data are shown as the diagonal line. The marginal distributions are Gaussian.

	BDL Num.	Aval. Data	Probability	Gauss Unit
Sn	2982	7511	0.39	-0.26
S	2031	7511	0.27	-0.61

Table 3.3: The BDL boundaries for Sn and S in Gaussian space.

	GBoth_bdl	GVar1_bdl	GVar2_bdl	GNon_bdl
Sn S	0.10	0.28	0.15	0.44

Table 3.4: Expected probability for Sn and S.

-0.61 for S are counted, and divided by the total number of samples. The expected probabilities using 10000 samples are shown in Table.3.4. "GBoth_bdl" means the expected probability of data are BDL for both variables. "GVar1_bdl" means the expected probability of data are BDL for variable 1 (Sn). "GVar2_bdl" means the expected probability of data are BDL for variable 2 (S). "GNon_bdl" means the expected probability of no data being BDL. Since the correlation between Sn and S is 0.009, the expected probabilities of the four regions are similar to the independent probabilities.

3.3.2 KL divergence results

To obtain the observed probabilities $P_{obs}(B_1, B_2)$, $P_{obs}(B_1, N_2)$, $P_{obs}(N_1, B_2)$ and $P_{obs}(N_1, N_2)$, the number of data in the four regions are counted and divided by the number of available data. Table.3.5 shows the results of the observed probabilities for Sn and S. "Perct_both" means the percentage of data are BDL for both variables. "Perct_Col1" means the percentage of data are BDL for

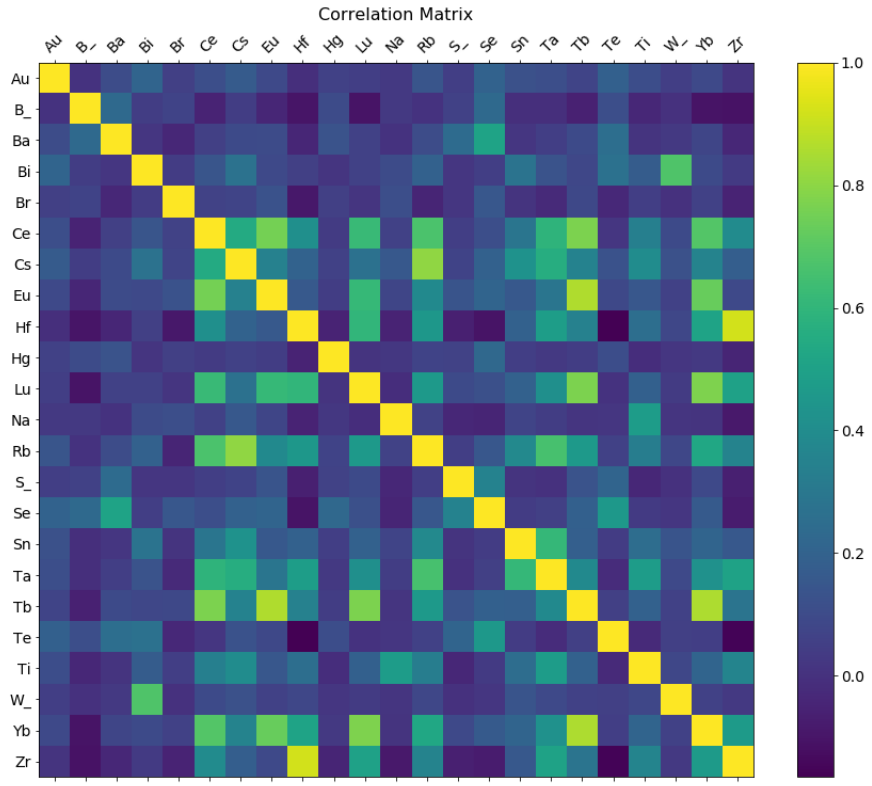


Figure 3.7: The correaltion matrix between variables with over 100 BDL data. The minimum correlation is -0.16.

variable 1 (Sn). "Perct_Col2" means the percentage of data are BDL for variable 2 (S). "Perct_None" means the percentage of data are not BDL for both variables. The observed bivariate distribution is different from the expected distribution. It has higher probabilities for $P_{obs}(B_1, B_2)$ and $P_{obs}(N_1, N_2)$. KL divergence measures the difference between the two discrete distribution (P_{exp} and P_{obs}) quantitatively. Given the data from Table.3.4 and Table.3.5, the difference is calculated using Eq.(3.2):

$$D = 0.17log(0.17/0.108) + 0.21log(0.21/0.28) + 0.09log(0.09/0.15) + 0.51log(0.51/0.44) = 0.046.$$

The same procedure of sampling expected bivariate distribution, calculating observed distribution, and calculating the difference using KL divergence is applied on combinations of two variables having more than 1000 BDL data in Table.3.1. The variables include Au, B, Eu, Lu, S, Sn, Ta, Tb, Te, W, Yb and Zr. If the resulting D_{obs} is large, we may conclude that the BDL occurrence between the variables are not independent and further inspection can be conducted. Fig.3.8 shows D_{obs} for different combinations. Only combinations with D_{obs} larger than 0.1 are shown in the figure. The right figures show the probability distributions of four regions for observed (blue) and expected

	Perct_both	Perct_Var1	Perct_Var2	Perct_None
Sn S	0.17	0.21	0.09	0.51

Table 3.5: Observed probability for Sn and S.

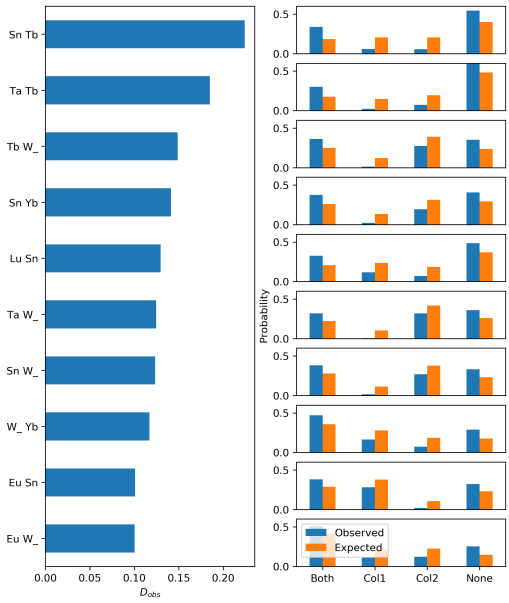


Figure 3.8: D values for each combination of the two variables. Only the combinations with D value larger than 0.1 are shown.

(orange) distribution. Col1 refers to the first variable in combination and Col2 refers to the second variable. The largest D_{obs} value is 0.22 for Sn and Tb and only 10 combinations are larger than 0.1 in all 66 ($12 \times 11/2$) combinations. Considering the theoretical maximum of D_{obs} is $+\infty$, the difference may not seem to be large, but the distributions are fairly different in the right column of Fig.3.8. To reveal more combinations of variables, simply lowering the threshold of D_{obs} can work, for example using 0.01 rather than 0.1, but different datasets have different range of D. It is difficult to set a threshold suitable for all cases. In fact, the range of D_{obs} is constrained by the possible observed bivariate distribution P_{obsp}. The possible bivariate distributions are constrained by the univariate BDL proportions. The maximum D_{max} can be found to standardize the observed D_{obs}, and the scaled D should provide more combinations that show difference between the observed and expected distributions. A threshold can also be set as the scaled D is between 0 and 1.

Find D_{max}

The range of D_{obs} can be calculated when treating P_{exp} as constant and P_{obsp} as variable. P_{exp} represents independent BDL occurrence. When P_{obsp} is close to P_{exp}, D_{obs} is small and the observed BDL occurrence is also independent. When P_{obsp} is very different from P_{exp}, D_{obs} is large and the BDL occurrence is dependent. To find the range of D_{obs}, we need to find the range of P_{obsp}. First,

how the univariate BDL probabilities constrain the bivariate distributions is shown. Suppose the probability of BDL in variable 1 is x and the probability of BDL in variable 2 is y. The following equations hold:

$$P_{obsp}(B_1, B_2) + P_{obsp}(B_1, N_2) = x$$

$$P_{obsp}(B_1, B_2) + P_{obsp}(N_1, B_2) = y$$

$$P_{obsp}(N_1, N_2) + P_{obsp}(N_1, B_2) = 1 - x$$

$$P_{obsp}(N_1, N_2) + P_{obsp}(B_1, N_2) = 1 - y$$

This is visualized on Fig.3.6. $x,\ y,\ 1-x,\ 1-y$ are the marginal probabilities. Although there are 4 equations, only 3 are independent (any three equations can derive the fourth). Since there are 3 independent equations with 4 unknown parameters, once one of the four probabilities($P_{obsp}(B_1, B_2)$, $P_{obsp}(B_1, N_2)$, $P_{obsp}(N_1, B_2)$ and $P_{obsp}(N_1, N_2)$) is set the other three are determined. The possible range for each region is also bounded as $P \in [0, 1]$. To show the range of possible bivariate distributions, it is convenient to treat the region with $P \in [0, min(x, y, 1-x, 1-y)]$ as the independent variable and the other three as the dependent variables. Now it is shown when D_{max} can be reached. To make the demonstration convenient, assume $x < y < 0.5$.(If they are set larger than 0.5, $1-x$ and $1-y$ are less than 0.5. The demonstration remains the same and only symbols change.) In this case, x is the minimum, and the independent probability region is $P_{obsp}(B_1, B_2) \in [0, x]$. Denote $P_{obsp}(B_1, B_2) = v$, and the probabilities of the other regions are

$$P_{obsp}(B_1, N_2) = x - v,$$

$$P_{obsp}(N_1, B_2) = y - v,$$

$$P_{obsp}(N_1, N_2) = 1 - x - y + v.$$

The independent variable of expected distribution is denoted as $P_{exp}(B_1, B_2) = p$ (Note p is constant) and the other regions are denoted as

$$P_{exp}(B_1, N_2) = x - p,$$

$$P_{exp}(N_1, B_2) = y - p,$$

$$P_{exp}(N_1, N_2) = 1 - x - y + p.$$

Putting them into Eq.(3.2):

$$D(v) = v \ln \frac{v}{p} + (x - v) \ln \frac{x - v}{x - p} + (y - v) \ln \frac{y - v}{y - p} + (1 - x - y + v) \ln \frac{1 - x - y + v}{1 - x - y + p} \tag{3.3}$$

In the definition of Eq.(3.2) (D(P||Q)), P is the observed distribution and Q is the assumed distribution. Here, P_{obsp} is treated as P and P_{exp} as Q. This convention is consistent with the definition of

KL divergence. Take the second derivative of Eq.(3.3):

$$\frac{\partial^2 D}{\partial v^2} = \frac{p}{v} + \frac{x-p}{x-v} + \frac{y-p}{y-v} + \frac{1-x-y+p}{1-x-y+v} \tag{3.4}$$

Since $v, p \in [0, x]$ and $x < y < 0.5$, Eq.(3.4) is larger than 0, so Eq.(3.3) is a convex function. By definition, the minimum of $D(v)$ is reach at $v = p$, and the maximum is reached at either $v = 0$ or $v = x$.

Take the data in Table.3.3 for example. The independent variable is $P_{obsp}(B_1, B_2) \in [0, 0.27]$. The minimum D value is 0 and can only be obtained when $P_{obsp} = P_{exp}$, and the largest D value is obtained at either $P_{obsp}(B_1, B_2) = 0$ or $P_{obsp}(B_1, B_2) = 0.27$. D is calculated in both cases to find out the maximum value. In this case, the theoretical maximum D_{max} is 0.32 and the maximum implies strong dependence in the BDL occurrence of two variables. Dividing the observed D_{obs} by D_{max} scales the measurement and it indicates how dependent the BDL occurrences are. Note since Eq.3.3 is not linear, the percentage of the scaled D is not a linear indication of the dependence. The same procedure is applied to all the combinations of variables with more than 1000 BDL data in Table.3.1 and the results are shown in Fig.3.9. Here, the dependent threshold is set to 10%, so only the combinations with more than 10% of dependence are shown. There are 20 combinations showing the dependency of BDL occurrence. There are more combinations than the previous case where D_{obs} is used. On the right side of the figure, the probability distribution of each region is also demonstrated. The blue bar represents the observed probabilities and orange bar represents the expected probabilities assuming BDL occurrences are independent. From the histograms, the observed and the expected distributions are fairly different, but they are not revealed using the unscaled D_{obs}. The maximum scaled D is 50% dependence between Ta and Tb. This implies the BDL occurrences in the these variables are strongly dependent. When data are BDL in one variable, they are likely to be BDL in the other variable. Five combinations have a percentage exceeding 20%. These combinations can also be worth further investigation.

3.3.3 Discussion

Although the problem of small D_{obs} value is solved by scaling D_{obs} using D_{max}, the actual threshold for determining the dependence is unclear. 100% represents full dependence and 0 means complete independence. The intermediate percentage needs to be examined and a threshold for concern set.

When calculating the expected probability, only the linear relation between variables in the original space is considered. There are two reasons for that. First, when sampling in a multi-Gaussian space, besides mean, only the correlations impacts the shape of the Gaussian distribution. Another reason is, no matter how non-linear the data are, the number of data in each region does not change, and non-linearity can only be observed in the region where data are not at BDL. It could be a problem when the area is divided finer, but here only considering the linear dependence is reasonable.

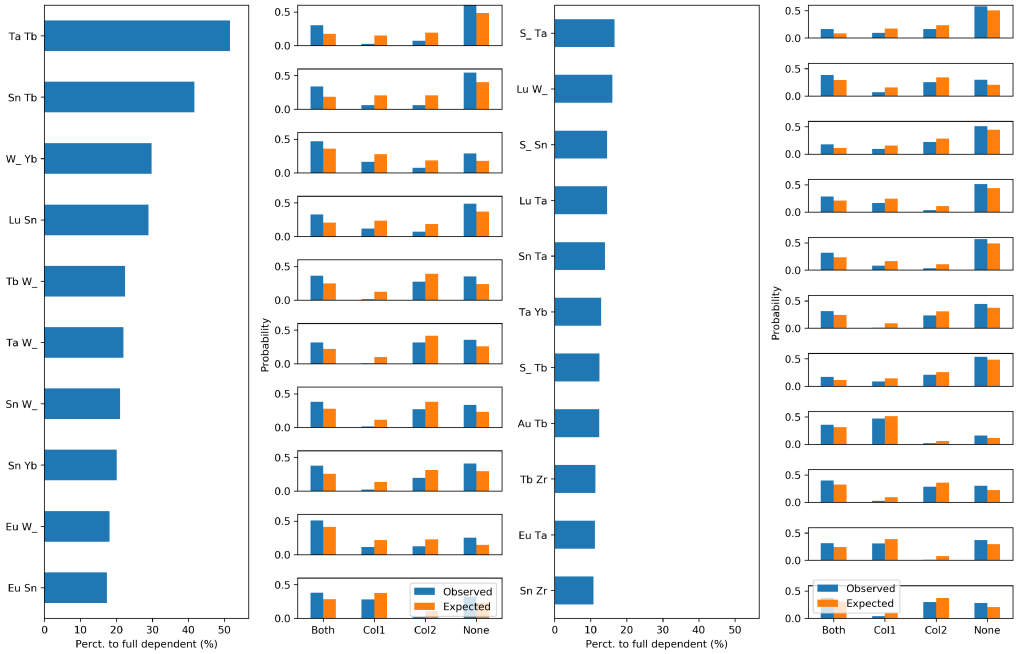


Figure 3.9: Percentage of dependence for each combination of the two variables. Only the combinations have percentage larger than 10% are shown.

3.4 Conclusion

BDL data come from the detection limit of equipments. They are recorded as the same value and form a spike. The spike can be problematic for geostatistical modeling and multiple despiking methods are proposed. Before despiking the BDL data, the characteristics of data spikes need to be understood. The BDL data table, which reveals more details on the BDL side, complements histograms. When measuring the spikiness of data, three different methods are applied. Each method has a unique application. Quadratic method reveals variables with large spikes, log method reveals variables with many spikes and the scaled method focuses on variables with different spike distribution from the expected one.

The proposed bivariate method detects the dependence of BDL occurrence between variables. Observed bivariate distributions are compared with expected distributions which assume independent BDL occurrence. When calculating the expected distribution, correlations between variables are considered. Since joint probabilities are difficult to calculate directly, expected distributions are simulated in a multi-Gaussian space considering correlations. The observed probabilities are compared with the expected one, using KL divergence and obtaining D_{obs}. Considering the theoretical value of KL divergence ranging from 0 to infinity, it is difficult to set a threshold to determine whether BDL occurrence is dependent. The bivariate distributions are bounded by univariate BDL probabilities, so a maximum D_{max} can be obtained, representing the full dependence of BDL occurrence. The observed results are scaled by the maximum value to evaluate the level of BDL

occurrence dependency. The workflow is summarized as follows:

- Choose two variables having over 1000 BDL data
- Find the univariate probability of BDL for each variable P_{BDL}
- Convert the probability boundary to standard Gaussian unit boundary b_k through quantile transform
- Sample data in a standard bivariate Gaussian space considering the correlation between variables
- Obtain the expected distribution
- Calculate the observed distribution by counting the number of data in each region
- Obtain the observed KL divergence value D_{obs}
- Calculate the range of possible observed probability P_{obsp}
- Obtain the maximum KL divergence value D_{max}
- Calculate the scaled D, $D = D_{obs}/D_{max}$

With the scaled D, 20 combinations show dependence of BDL occurrence when the threshold is set to 0.1. In the real data, 5 combinations having 20% of BDL occurrence dependence, which can be worth further investigating. The strong dependence of the BDL occurrence can come from the dilution during the measurement. When the one variable is diluted, other variables contained in the same solution are diluted at the same time. If the concentrations of some variables are already low, the dilution can result in the concentrations of these variables at BDL simultaneously, which is reflected as the dependence of the BDL occurrence.

Chapter 4

Multivariate Cluster Analysis

Data from different domains possess different features. If they are analyzed together, the combined statistics may not provide as much meaningful information, so it is important to conduct cluster analysis to support the definition of domains. The data quality affects the performance of clustering. Missing data, below detection limit data and outliers can decrease the accuracy of clustering analysis. Spikes formed by below detection limit (BDL) data can change the centroids of clusters. The number of clusters is another important parameter that influences clustering performace. When the number of clusters is wrong, data are partitioned rather than clustered. This chapter introduces a workflow for detecting the optimal number of clusters. The clustering methods used are k-means and Gaussian mixture model. Different data transformations, including linear transform, uniform transform and Gaussian transform are used to deal with outliers and spikes. The chapter examines the performance of different combinations of the workflow, clustering methods and data transformations. The statistical analysis indicates two clusters in the real data. The multivariate data are projected to a 2D plane for validation.

4.1 Introduction

The purpose of exploratory data analysis is to understand the nature of data. It includes exploring univariate and multivariate distributions, finding outliers and duplicated data, and evaluating summary statistics of the data. If data have multiple clusters, the statistical inference may not provide accurate information, because they are averaged. Therefore, it is important to apply cluster analysis. Further statistical calculations and geostatistical modeling can be conducted on representative groups.

4.1.1 Types of clustering methods

The idea of cluster analysis is straightforward: keeping similar data together. Different ways of measuring similarity between data result in different types of clustering methods, such as distance based and distribution based clustering. Distance based methods use distance between data as a measurement of similarity. The closer the distance, the more similar the data are. Most distance based clustering methods use the Euclidean distance to quantify the similarity of data:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{||\sum_{i=1}^{k} x_i - y_i||^2}$$

where $d(\mathbf{x}, \mathbf{y})$ is the distance between data vector $\mathbf{x}$ and $\mathbf{y}$, and i represents variable i and k is the number of variables. Distribution based methods fit multiple kernels to data distribution. The probabilities of data belong to a same kernel show their similarity. Gaussian kernels are often used because a small number of parameters can formulate multivariate Gaussian distributions.

K-means clustering is the most commonly used distance based method (Abubaker & Ashour, 2013; MacQueen et al., 1967; Pedregosa et al., 2011). K-means clusters data by assigning them to their closest centroids. The number of centroids is the number of clusters. The procedure is as follows: centroids are assigned randomly at the beginning. With the centroids assigned, data are clustered to their closest centroids. Then the cluster centroids are recalculated. Data are assigned again to the closest new centroids. The iterations continue until the centroids remain unchanged. The process is equivalent to finding the clusters giving the minimum within cluster sum of squares (WCSS):

$$WCSS = \sum_{i=0}^{N} \min_{\mu_j \in C} (||\mathbf{x}_i - \boldsymbol{\mu}_j||^2)$$

where N is the number of data, μ_j is the mean of cluster j, and C is the number of clusters. The minimum term shows the sum of squares is only calculated for data to their closest centroids (within the same cluster). The algorithm can settle in local minima, and the final centroids are determined by the initial assignment of centroids, so multiple realizations of initial centroids are generated, and the final clusters are the ones giving the minimum WCSS.

Gaussian mixture model (GMM) is a common distribution based clustering method (Pedregosa et al., 2011; Reynolds, 2009). It fits Gaussian density kernels to clusters and data are assigned based on their probabilities in each kernel. The clustering procedure is similar to k-means. Gaussian kernels are assigned in the initial state, and data are assigned to the kernels in which they have the maximum probability. Then new Gaussian kernels are generated that give the maximum likelihood of data in the same cluster. Data are assigned again based on the new kernels. This process continues until the Gaussian kernels stay the same.

Although the fitting process is similar for k-means and GMM, each method has its own advantages. k-means is more stable than GMM. When there are too few data to calculate a non-singular covariance matrix, GMM can diverge (Yamazaki & Watanabe, 2003). The cluster shape is more flexible for GMM. The covariance matrix in GMM can handle elongated cluster shapes, while k-means assumes clusters have isotropic shape. Different methods need to be adopted for different situations.

There are new clustering methods which can handle data with special shapes shown in Fig.4.1 (Fred & Jain, 2005). Using robust and efficient clustering methods such as k-means and GMM should be sufficient for identifying clusters in geostatistical data. In this chapter, k-means and GMM are used to examine clusters in data.

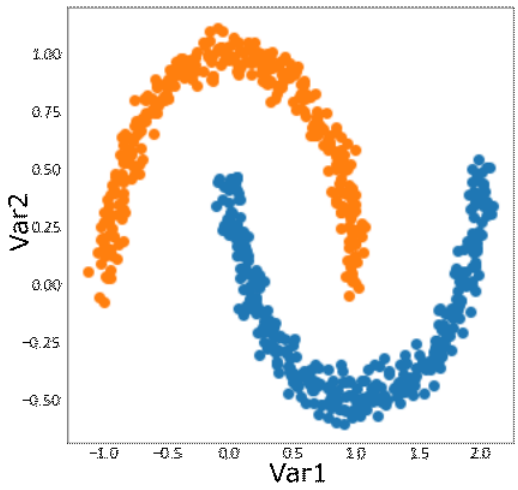


Figure 4.1: New clustering methods can handle complex clusters such as the moon shape clusters (Fred & Jain, 2005).

4.1.2 Data transformation and number of clusters

Most data need to be processed prior to cluster analysis. One common practice is rescaling the data to [0,1] for equal scale in every dimension. This is important especially for distance based clustering methods. If the scale of one dimension is orders of magnitudes larger than others, the clustering of data would be dominated by this dimension. The advantage of rescaling is that it keeps the original shape of data, but it is not reliable when outliers and spikes are present. Outliers are the data with extremely high value, sometimes 100 times higher than the mean. This can be problematic for clustering, because the cluster centroids can be shifted drastically by outliers. Spikes are duplicated data, mainly coming from the precision limitations of the measurement equipment. The biggest spike in real data can be the BDL spike. They are the data below the detection limit of the measurement equipment, and can be recorded as 0.0 or the minimum detectable value. The BDL data form a large spike at the low value region. This can cause distribution based clustering to fit a narrow kernel only for the BDL data, while they could belong to other groups if the true values were known.

Quantile transformation is another data transform used commonly in geostatistical modeling to address outliers and skewed distributions. Suppose z represents data in the original units and y represents the transformed units. $F(z)$ is the cumulative distribution function (CDF) of data in the original units and $G(y)$ is the CDF in the space to be transformed to. When data are normal score transformed, $G(y)$ is the Gaussian density (M. Pyrcz & Deutsch, 2018). The following equation is used to transform the univariate z to y:

$$y = G^{-1}(F(z)) \tag{4.1}$$

Eq.(4.1) can transform the original univariate distribution to a Gaussian distribution. In multivariate cases, the transformation is conducted in each dimension separately. For example, in normal score transform, the 1D marginal distribution has Gaussian shape but the multivariate data do not necessarily have a multi-Gaussian shape. This is desired as cluster analysis depends on the multivariate relations. The transformed outliers are closer to the main data. The transformation of spikes is also important. If spikes are preserved, data in spikes are assigned the same quantile. If spikes are spread, data in spikes are assigned different quantiles.

Prades (2017) proposes to normal score transform data with the spikes spread when using GMM and with the spikes preserved when using k-means. However, transferring data to a uniform distribution could be better, since it does not create an artificial cluster in the center of the space. In this chapter, data rescaling, normal score transform and uniform transform of data are compared, along with different treatments for data spikes.

The number of clusters (NC) is another important aspect that affects clustering performance. If the NC is not optimal, data are partitioned rather than clustered. It is relatively easy to visualize NC in a space less than 4 dimensions. For higher dimensional data, statistical tools are required. The statistical tools used in this chapter are Hopkins statistic (Lachheb, 2021; Lawson & Jurs, 1990), gap statistic (Tibshirani et al., 2001), silhouette coefficient (Aranganayagi & Thangavel, 2007) and prediction strength (Tibshirani & Walther, 2005). Hopkins statistic determines if there is any cluster in the distribution by comparing data with uniform samples. Gap statistic and silhouette coefficient find the optimal NC. Prediction strength uses cross-validation to verify if data are truly clustered or partitioned. A workflow combining these four tools is proposed to determine the optimal NC.

4.1.3 Chapter structure

In this chapter, a series of tools are introduced for handling spikes when conducting cluster analysis. First, a workflow to find the optimal NC is introduced with a detailed explanation of the statistic tools used. Its application is demonstrated using synthetic data. Then the workflow's compatibility with different transformations and clustering methods is examined. The clustered results are compared to the true labels and a measurement of the correctness rate is used to determine the appropriate transforms and clustering methods. The synthetic data show k-means clustering combined with linear transform, uniform transform and Gaussian transform with spikes preserved are feasible. The workflow and the appropriate transforms are used on high-dimensional real data. The real data come from the Northwest Territories with the missing data eliminated (Falck et al., 2012). From the previous chapters, the number of BDL data is significant. The large spikes are handled when conducting cluster analysis.

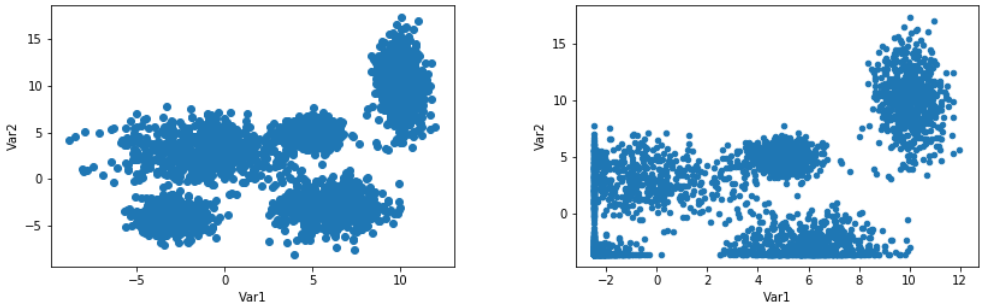


Figure 4.2: The original synthetic data (left) and the data with synthetic spikes (right).

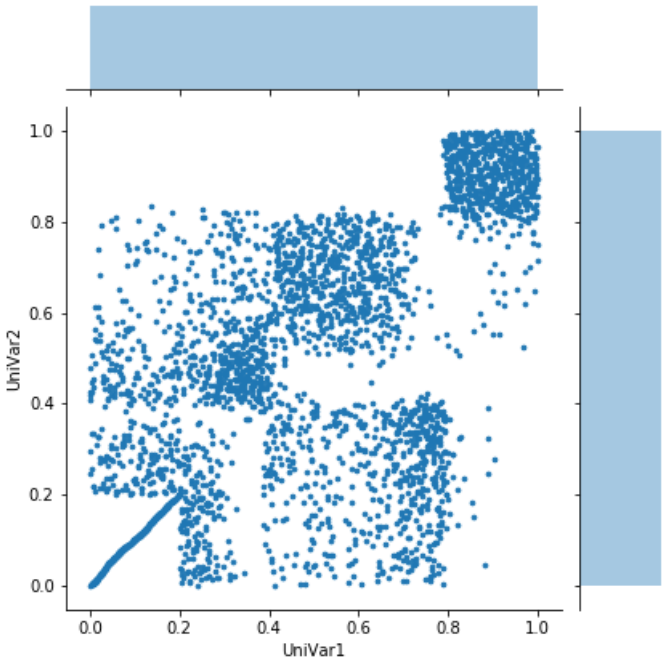


Figure 4.3: Uniform transformed data, spreading out the spikes. The marginal distributions are shown on the edges.

4.2 Workflow to determine the optimal NC

In most clustering methods, NC is an important parameter for correct clustering results. This section introduces a workflow of determining NC. It has 3 aspects: the detection of clusters existence using Hopkins statistics, finding NC using gap statistic and silhouette coefficient, and the cross-validation of the resulting NC.

4.2.1 Data Preparation

Fig.4.2 shows the synthetic data used for illustration purposes. The left figure shows the original data. The data consist of 3000 samples and 5 clusters with Gaussian shape, and the 5 clusters have

different mean and variances. The right figure shows the synthetic data with spikes. 20% low value data are treated as BDL. To handle spikes, the data are uniform transformed with spikes spread. Fig.4.3 shows the uniform transformed data. The range of the uniform space is from 0 to 1. The BDL data quantiles are randomly assigned. The left bottom line represents the data that are BDL in both dimensions. They appear as a single dot at the left bottom in Fig.4.2. Data with only one dimension BDL are on the bottom or left margins in Fig.4.2. The BDL boundary is at 0.2 because the uniform space range is [0,1] and 20% data are set to be BDL. If the range is [0,2], the BDL boundary is at 0.4. If 30% data are BDL, the boundary is at 0.3. From the marginal histogram distribution, the data are uniformly transformed. The five clusters are still distinguishable. The transformed clusters shapes are relatively isotropic, which is important for applying k-means.

4.2.2 Clustering Tendency

The first step of cluster analysis is to determine if there are any clusters in data. It is also called clustering tendency. Lawson and Jurs (1990) proposed the Hopkins statistic for this task. The idea is to compare the dataset with uniform distributed samples (used as references). If the results are very different from the reference, there can be clusters.

Suppose there are N data ($dt_j \quad j = 1, \cdots, N$). N samples are simulated in a uniform space that shares the same range with the data, which are denoted as $s_i, \quad i = 1, \cdots, N$. The Hopkins statistic H is calculated as

$$H = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i + \sum_{j=1}^{N} x_j}$$

where y_i is the distance of random sample s_i to its nearest neighbor s_k (k denotes the nearest neighbor), and x_j is the distance of data dt_j to its nearest neighbor dt_k. From the equation, if there are clusters, the sum of x_j is much smaller than the sum of y_i, so H is close to 1. When there are no clusters, the sum of x_i is similar to the sum of y_i. H is close to 0.5. Since the reference $\mathbf{y}$ is sampled throughout the uniform sapce, Hopkins statistic is not very accurate when outliers are present. The empty space between outliers and main data is uniformly sampled, and the sum of reference data is much larger than the original data, resulting in H close to 1. The Hopkins statistic is more informative when the data are uniform transformed. The uniform transformed synthetic data in Fig.4.2 have H value of 0.8, so it indicates clusters existence.

4.2.3 Optimal number of clusters

Silhouette coefficient and gap statistic are used to determine the optimal NC. Silhouette coefficient measures the sparsity of clusters. For each data i, the silhouette coefficient S_i is calculated as

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

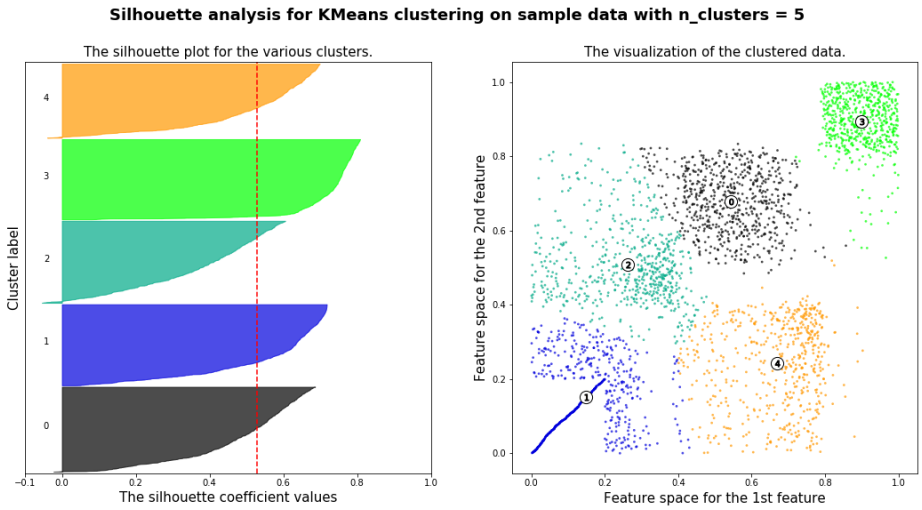


Figure 4.4: The silhouette coefficient for data and the corresponding clusters when using k-means and NC=5.

where a_i is the average distance between i and the data in its own cluster, and b_i is the average distance between i and the data in the nearest cluster. Here, the distance means the Euclidean distance between data. From the equation, $S_i \in [-1, 1]$. When S_i is close to 1, it indicates data i should belong to the current cluster whereas S_i close to -1 indicates the data may belong to other clusters. The silhouette coefficient for the dataset S_N is the average S_i over all data, where $i = 1, 2, \cdots, N$. S_N close to 1 means the clusters are well separately. When S_N is close to 0, it indicates there may not be distinguishable clusters. When using Euclidean distance, silhouette coefficient assumes clusters have isotropic shapes, so when S_N is negative, it does not necessarily mean clusters are wrongly grouped. It is also possible that clusters have irregular shapes. As shown in Fig.4.4, the transformed data are clustered into 5 groups using k-means. In the left figure, each horizontal line represents S_i for each data (3000 lines in total). The vertical red dashed line represents the averaged silhouette coefficient S_N. Data in cluster number 3 have a high silhouette coefficient because they are compact and distant from other clusters. There are some negative values in cluster number 2 and 0. As observed from the right figure, they may be the data to the top-left corner of cluster 0 and bottom-right corner of cluster 2, which are wrongly clustered.

The silhouette coefficients are plotted against a range of NC, and the optimal one is indicated by the highest silhouette coefficient. Fig.4.5 shows the optimal NC is 5, which is corresponding to the true NC of the synthetic data. Note "Clustering Number" in the figures means the number of clusters. The silhouette coefficient range from 0.43 to 0.53. The difference is not very great considering how differnt the data can be clustered using different NC. Especially when NC is 4, the silhouette coefficient is fairly close to that of NC equal to 5. If the true NC is unknown, it is difficult to decide between 4 or 5 clusters, so another measurement to validate the results is needed.

Similar to Hopkins statistic, the gap statistic uses uniform distributed samples as a reference to

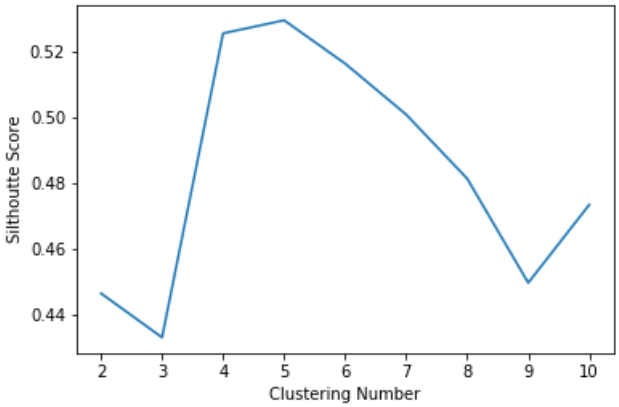


Figure 4.5: The silhouette coefficient for different NC.

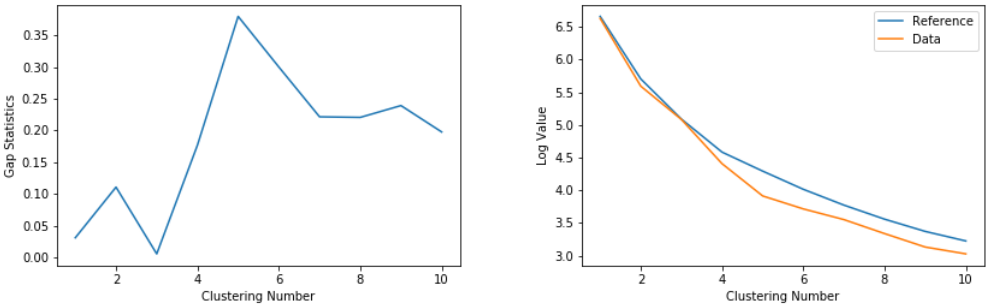


Figure 4.6: The gap statistic and the corresponding $log(W_k)$ for reference and data in a range of NC.

compare with data (Tibshirani et al., 2001). Suppose N data are divided into K clusters. Within a specific cluster r ($r = 1, \cdots, K$), the sum of pairwise squared Euclidean distance D_r is calculated as

$$\begin{aligned} D_{i,j} &= ||\mathbf{x}_i - \mathbf{x}_j||^2 \\ D_r &= \sum_{i,j \in C_r} D_{i,j} \end{aligned} \tag{4.2}$$

where $D_{i,j}$ is the squared Euclidean distance between data i and j, and C_r represents cluster r. For all K clusters,

$$W_K = \sum_{r=1}^{K} \frac{1}{2N_r} D_r,$$

where N_r is the number of data in C_r. Gap statistic is equal to

$$Gap(K) = E^* log(W_K) - log(W'_K),$$

where the first term represents the expected value of $log(W_K)$ providing multiple realizations of N samples from the multivariate uniform distribution. The second term $log(W'_K)$ is calculated from the observed data. The optimal K is the NC which gives the maximum $Gap(K)$.

Fig.4.6 shows the results of the gap statistic on the uniform transformed data. The NC is from 1 to 10. The left figure implies the optimal NC is 5, and $Gap(4)$ is much smaller than $Gap(5)$. It provides another measure to help choose the optimal NC when silhouette coefficents are close. The right figure gives an intuitive explanation for the gap statistic. Increasing NC decreases the overall W_k, and the reference data provide a reference to illustrate the tendency of the decrease. When K is not the optimal NC, the $log(W'_K)$ of observed data decreases similarly to the reference data. When K is the optimal number, observed data are well separated and the W_K decreases faster than other cases, thus giving the maximum gap between the observed $log(W'_K)$ and the reference $log(W_K)$. Note the optimal NC selection criterion is modified to give a more intuitive explanation. The disadvantage of the gap statistic is similar to that of the Hopkins statistic. When outliers are present, the decreasing of W_K may not be in the same magnitude for reference data and observed data. Gap statistic is more reliable when data are uniform transformed, so it is mainly used as a complement to the silhouette coefficent.

4.2.4 Cross Validation

The cross validation for cluster analysis is also called "prediction strength". It is used to validate the NC obtained from previous steps. The main idea is to conduct clustering twice on different proportions of data and compare the results. The first time can be the whole data and the second time randomly sampled 80% of the data. Since the proportion of data are randomly sampled, the general shape of data should remain the same. If the results are similar, data are well clustered. If the results are fairly different, data can be partitioned because the NC is not optimal or the clustering method is not suitable.

First, all data are clustered (training data), obtaining cluster centroids. Then part of data are sampled (testing data). They can be a proportion of data with or without replacement. 80% of data without replacement are used here. Testing data are clustered, obtaining cluster labels. For testing data sharing the same label, they are classified using cluster centroids from the training data. If data in the same testing cluster are classified into the same group, the prediction strength for this cluster is high. If they are classified into multiple different groups, the prediction strength is low. The resulting prediction strength is the lowest among all clusters.

Suppose the training and testing data are clustered into K clusters separately. The prediction strength:

$$ps(K) = min(\frac{1}{N_r(N_r-1)} \sum_{i \neq i' \in C_r} D[C(X_{tr},k), X_{te}]_{i,i'}), \quad r = 1, \cdots, K$$

where N_r is the number of data in testing cluster C_r, i and $i^{'}$ are the data in cluster C_r, $[C(X_{tr},K), X_{te}]$ means cluster the training data X_{tr} to K clusters, and use the K centroids to cluster the testing data X_{te}. $D[C(X_{tr},k), X_{te}]_{i,i'} = 1$ if testing data i and $i^{'}$ are clustered together using the training cen-

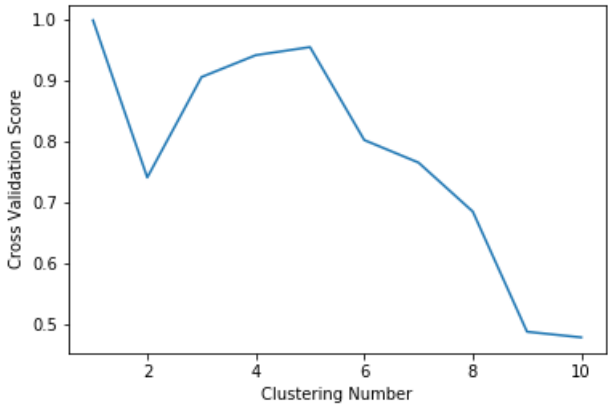


Figure 4.7: The prediction strength over a range of NC.

troids, and $D[C(X_{tr}, k), X_{te}]_{i,i'} = 0$ otherwise. The summation term calculates the pairwised similarity in testing cluster C_r. From the equation, $ps(K) \in [0, 1]$. The reason for using the minimum value is because the cluster giving the minimum prediction strength is where partitioning happens. The more similar training and testing data are clustered, the more robust the resulting NC.

Fig.4.7 shows the prediction strength over a range of K values on the uniform transformed data. $ps(1)$ is always 1.0, because training and testing data have only one cluster, all data in the one cluster of testing data are grouped together by the only centroid. The 5 clusters has a high prediction strength of 0.95. It validates 5 as the optimal NC and the suitability of k-means for the data. The plot should not be used to determine the optimal NC as different testing data size changes the K giving maximum prediction strength. The optimal NC does not always promise the highest prediction strength. It should be used to validate the choice of NC and the compatibility of clustering methods with data. A prediction strength higher than 0.9 indicates the NC is trustworthy.

4.2.5 Clustering results

Fig.4.8 shows the results of k-means clustering using 5 clusters. In general, the 5 clusters are well identified. Some data are miss-clustered such as the left top corner data in the blue cluster. The k-means results are reasonable when data are uniform transformed with spikes spread. For high dimensional data, there is no luxury of visualizing the results, so the workflow is more important. Its compatibility with different clustering methods and different data transformation is examined in the next section.

4.3 Compare Different Transformations

In this section, the compatibility between the proposed workflow and different transformations is compared. The transformations include linear scaling, uniform score transformation with spikes

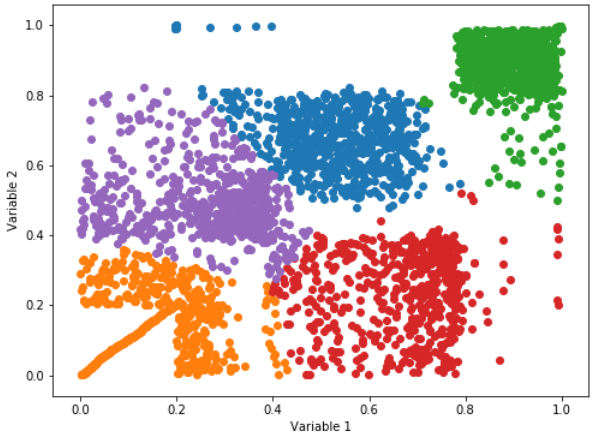


Figure 4.8: K-means results of the transformed data.

spreaded and preserved, and normal score transformation with spikes spreaded and preserved. The clustering methods used are k-means and GMM. Synthetic data containing spikes from the previous section is used here. Deemed appropriate clustering and transformations are used to examine real data.

4.3.1 Linear transformation

The linear transform rescales variables into the same range. Otherwise, dimensions with larger scale may distort distance based clustering. Suppose a dataset with N samples and M dimensions $d_{ij}, i = 1, \cdots, N, j = 1, \cdots, M$. Consider in dimension j:

$$dr_{ij} = \frac{d_{ij} - min(\mathbf{d_j})}{max(\mathbf{d_j}) - min(\mathbf{d_j})}$$

where dr_{ij} is the rescaled data, and $\mathbf{d_j}$ represents all N samples in dimension j. The transform rescales all variables to a range of [0,1]. The advantage of the linear transform is that it preserves the original shape of data, which is preferred in cluster analysis. The method also has several limitations. It does not have alternative ways to handle spikes, and it is not robust to centroid shifting caused by outliers.

Fig.4.9 shows the rescaled data. The shape remains the same as in Fig.4.2 but the data are rescaled from 0 to 1. The Hopkins statistic of the data is 0.94, indicating cluster existence. Gap statistic and silhouette coefficients are calculated over a range of NC. Fig.4.10 shows the results when data are clustered using k-means and GMM. Both clustering tools show the optimal NC is 5, except for the silhouette coefficient plot when using GMM. The results of k-means and GMM clustering using 5 clusters are shown in Fig.4.11. It is obvious that k-means method gives better results. GMM clusters the BDL data as one single cluster, which is shown at the bottom left corner in the

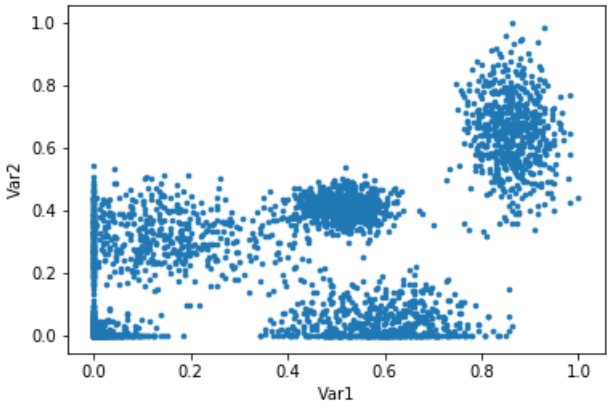


Figure 4.9: Linearly rescaled synthetic data.

right figure. This comes from GMM fitting a kernel specific to the spike formed by BDL data. The spike also results in the wrong optimal NC using silhouette coefficient and GMM in Fig.4.10. The cross-validation results of k-means is 0.98 while the results of GMM is 0.52, which means the results of using GMM change drastically when using different proportion of data. GMM is not compatible with the transformation when spikes present.

Since the true label is available for the dataset, the percentage of data correctly clustered can be calculated. More specifically, consider one cluster at a time and calculate the percentage of data that truly belong to that cluster. The minimum percentage among all clusters is used as a measurement for clustering performance and refer to it as the correctness rate. Suppose there are K clusters after clustering. In a specific cluster $C_r,\ r = 1, \cdots, K$, there are T_r true labels and each label is denoted as t. The correctness rate R_r in this cluster is calculated as:

$$R_r = \max(P(t))\ t = 1, \cdots, T_r \tag{4.3}$$

where $P(t)$ is the proportion of the true label t in cluster C_r. The correctness rate of the clustering results R is calculated as

$$R = \min(R_r)\ r = 1, \cdots, K \tag{4.4}$$

The k-means results have a correctness rate of 0.90 while the correctness rate of GMM results is only 0.72. It means in the cluster labeled orange, only 72% of the data truly belong to that cluster. GMM does not appear appropriate for the linear transformed data when spikes are present.

Now consider the scenario of outliers, which is very common in real data. Fig.4.12 is the same synthetic dataset but with some outliers added. Some outliers are 10 times larger than the main data. Since GMM does not work well when spikes are present, only the performance of k-means is examined here. The Hopkins statistic is 0.97, showing clustering existence in the data. Fig.4.13 shows the plots of gap statistic and silhouette coefficient, indicating the optimal NC should be 8. Fig.4.14 shows the eight clusters using k-means. Although the optimal NC is not the same as the

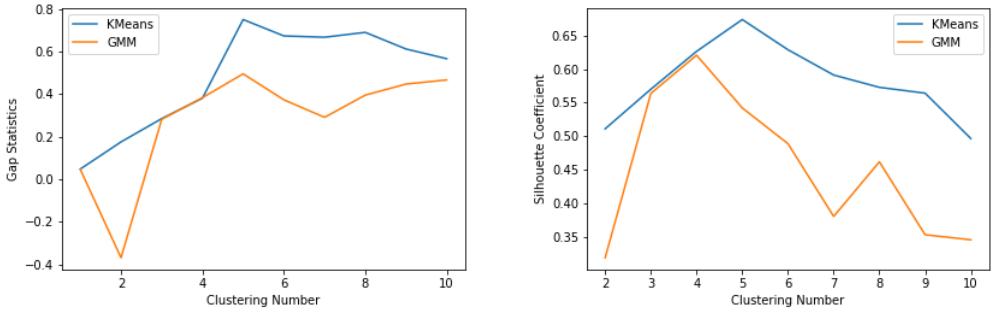


Figure 4.10: Gap statistic (left) and silhouetter coefficient (right) on the linearly transformed data.

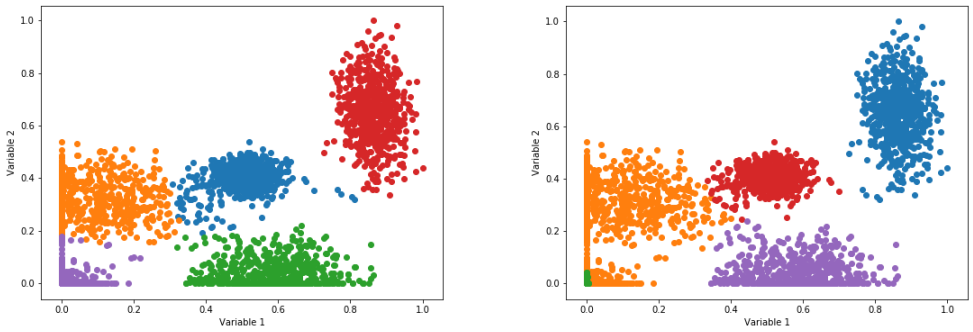


Figure 4.11: K-means (left) and GMM (right) clustering results on the linearly transformed data.

true one, all five major clusters are well clustered and the outliers are separately clustered. By using k-means, data with outliers can still be well clustered. The outliers only affects the analysis of NC. However, the cross-validation results are 0.0 when using 8 clusters. It could come from the testing samples containing outliers. The number of data in outlier clusters is small, so it is possible that the outliers in testing data are clusterd very differently from those in the training data. If the real NC is used, the data are poorly clustered as shown in Fig.4.15. It comes from the outliers change the cluster centroids drastically. Thus, using k-means for linearly transformed data is reasonable, but when outliers present, the number of data in each cluster may need to be examined.

4.3.2 Uniform transformation

There are two reasons for uniform transformation. Firstly, it scales different dimensions to a range of [0,1]. The quantile transformation also handles the outliers. The outliers distort the clustering by switching the cluster centroids drastically or changing the real NC as shown in the previous section. Through quantile transformation, outliers appear as normal data and their effects on the clustering results can be minimized. The quantile transformation can also treat spikes differently, which increases its compatibility with different clustering methods.

First the synthetic data are uniform transformed with spikes spread. The robustness of using

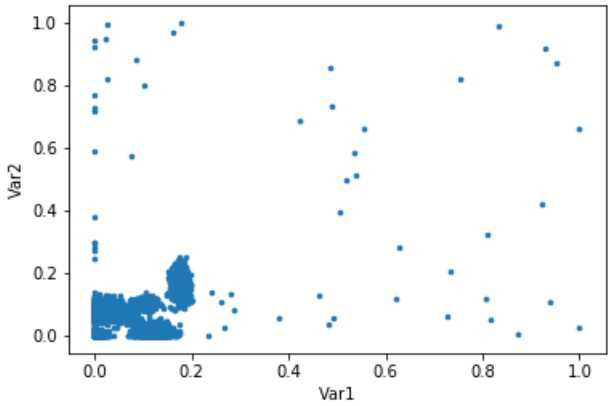


Figure 4.12: Synthetic data with outliers.

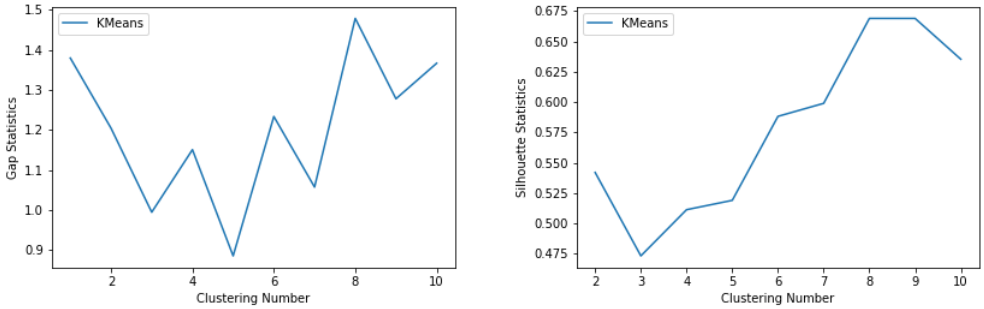


Figure 4.13: Gap statistic (left) and silhouetter coefficient (right) on linearly scaled data containing outliers.

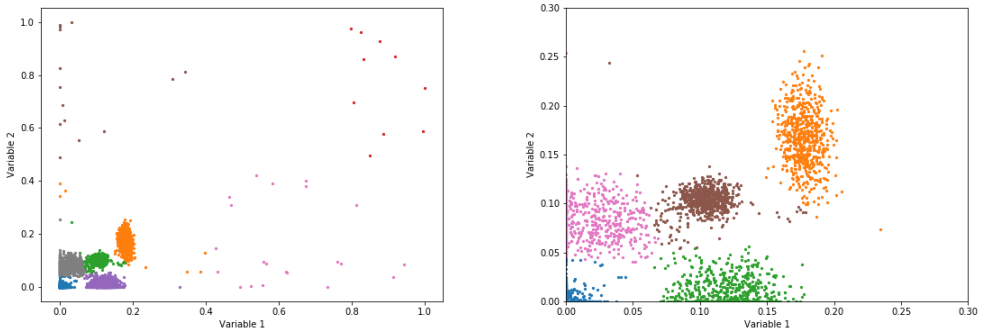


Figure 4.14: Results of k-means clustering using cluster number of 8. Right one is the zoomed in scatter plot of the region of interests.

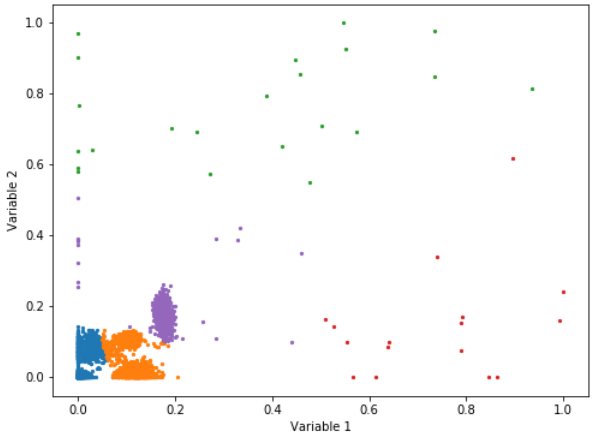


Figure 4.15: Results of k-means clustering using cluster number of 5.

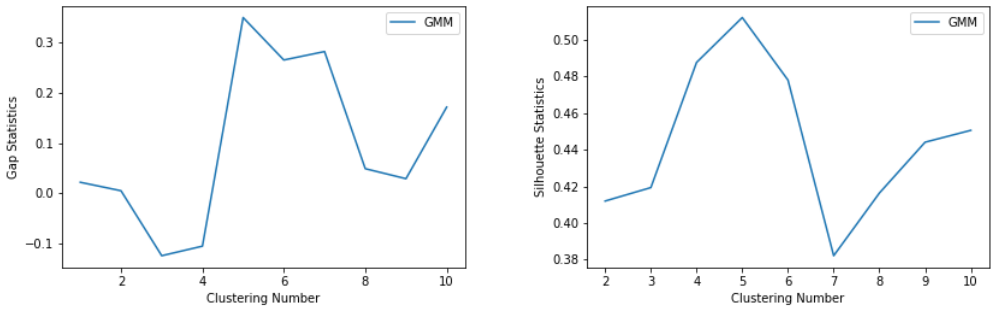


Figure 4.16: Results of gap statistic and silhouette coefficient using GMM when data are uniform transformed with spikes spread.

k-means on this transformation is demonstrated in the previous section, so only GMM is examined in this part. Fig.4.3 shows the transformed data. The Hopkins statistic of the data is 0.8. Fig.4.16 shows the plots of the gap statistic and silhouette coefficients against different NC, and the results indicate the optimal NC is 5. Fig.4.17 shows the results of GMM clustering. The 5 clusters have similar shape as in the k-means results (Fig.4.8), but there are more wrongly clustered data in the right upper corner and right bottom corner of red cluster. The cross-validation score is 0.84 for GMM results, which is lower than that of k-means. The correctness rate of the resulting clusters is 0.78. In the wrost scnario (the red cluster), only 78% of data truly belong to that cluster. On the contrary, the correctness rate of k-means is 0.9. Therefore, when data are uniform transformed with spikes spread out, it is better to use k-means method.

Now data are uniform transformed with spikes preserved. Fig.4.18 shows the transformed data. Outliers are at the margins, and they appear much closer to the main data compared with the original units. The spikes are obvious on the marginal histograms, and 5 clusters are visually distinguishable. The Hopkins statistic is 0.83. It is higher than the spikes spreading case as spike data remain

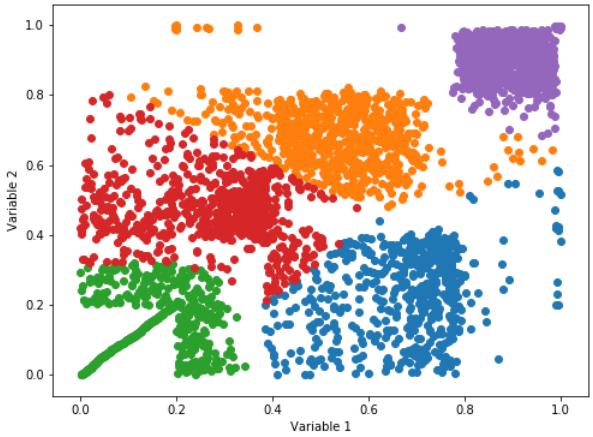


Figure 4.17: Results of GMM clustering when data are uniform transformed with spikes spread.

close together. Fig.4.19 shows the gap statistic and silhouette coefficients plots using k-means and GMM. The k-means results indicate the optimal NC should be 5 but the GMM results have problems identifying an optimal NC. Fig.4.20 shows the results of the two clustering methods. K-means performs better than GMM. The 5 clusters are separated better, although some miss-clustered data in the red cluster. The outliers do not influence the clustering as much as in Fig.4.15. Their contributions in shifting cluster centroids are similar to other data. The clusters of GMM show similar features with the linear transformed data. The spike is independently separated rather than being grouped with other data. The corresponding cluster is shown as a single dot in the figure. This leads to the wrong grouping of green and orange clusters. The cross-validation score for k-means is 0.96 while the score for GMM is only 0.46, which means data are partitioned by GMM. The correctness rate for k-means is 0.91 while the rate is only 0.61 for GMM. When clustering uniform transformed data with spikes preserved, it appears better to use the k-means method.

4.3.3 Gaussian transformation

Gaussian transform is another type of quantile transformation and widely used in geostatistical modeling. It shares similar advantages with uniform transformation. The transformed data are clustered in the center of multi-Gaussian space. This can cause the Hopkins statistic and gap statistic inaccurate as they use uniform distribution as their reference distribution. These two statistics can be adapted to use Gaussian distribution as reference in future work.

Fig.4.21 is the Gaussian transformed data with spikes spread as shown on the marginal histogram. The Hopkins statistic is 0.93. This high value comes from the 5 clusters and the Gaussian transformation clusters data in the center of the space. The five clusters are not as visually distinguishable compared to the previous transforms, because data are centered around (0,0) co-

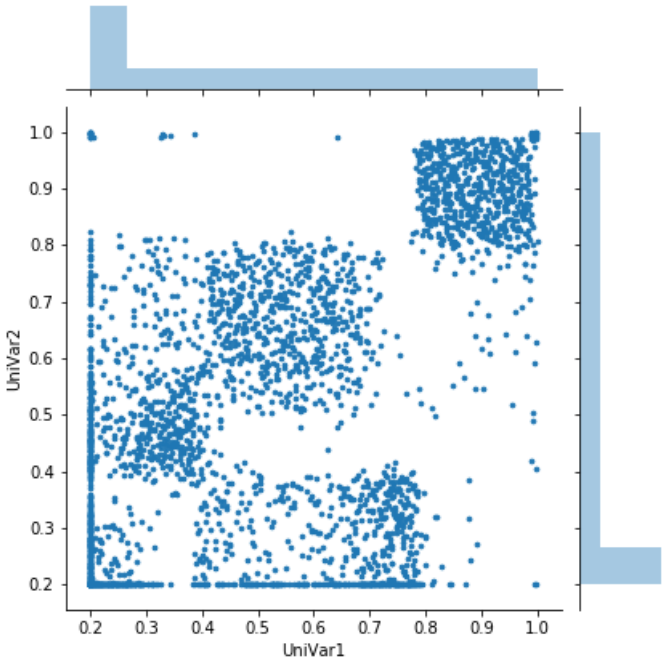


Figure 4.18: Synthetic data after uniform transform and spikes preserved.

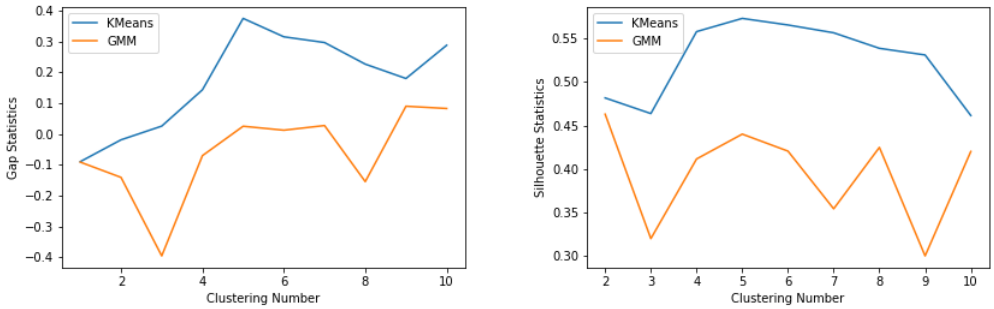


Figure 4.19: Results of gap statistic and silhouette coefficient using k-means and GMM when data are uniform transformed with spikes preserved.

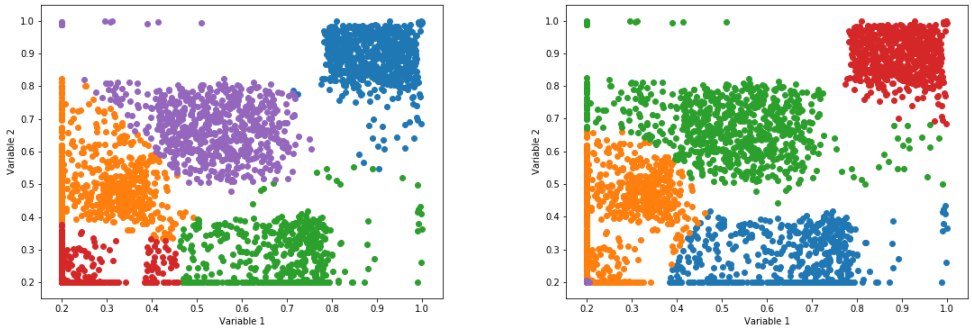


Figure 4.20: Resulting clusters from k-means (left) and GMM (right) when data are uniform transformed with spikes preserved.

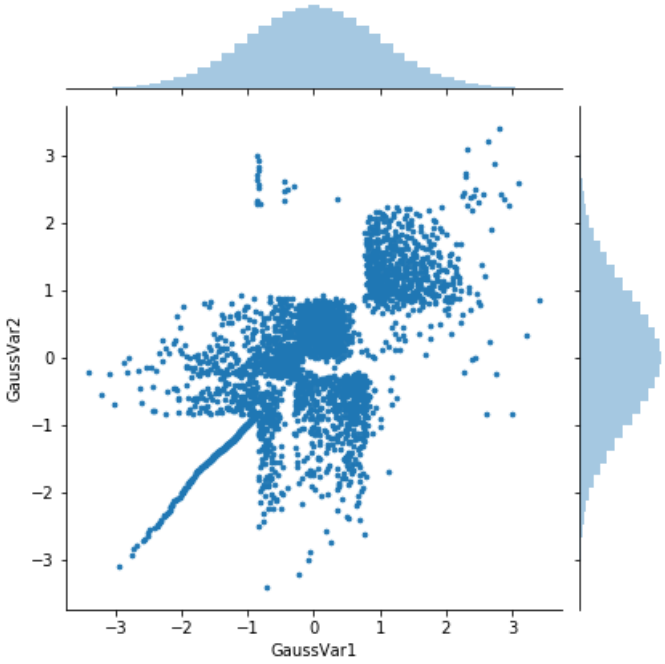


Figure 4.21: Gaussian transformed synthetic data with spikes spread out.

ordinate. Fig.4.22 shows the gap statistic and silhouette coefficients plots. Only the gap statistic on k-means results gives the right NC, which indicates this transformation is not very compatible with the workflow. Fig.4.23 shows the clustering results using 5 clusters. Neither approach gives reasonable clustering results. GMM gives worse results by clustering the transformed spike as a single cluster. Although k-means performs relatively better, many data are miss-clustered into the orange cluster. The clustering does not separate data well especially at the center of space where the data distribution is dense. The cross-validation scores are 0.54 and 0.52 for k-means and GMM respectively, which means the transformed data are not reasonably clustered using these two methods. The correctness rate of k-means is only 0.77, which is much lower than that of the uniform transformed data, and the correctness rate for GMM is 0.43. So transforming data to Gaussian units with spike spread may not be compatible with the workflow and clustering methods.

Fig.4.24 is the Gaussian transformed data with spikes preserved. The five clusters are more visually distinguishable compared with the spike spread case. The Hopkins statistic is 0.92. Fig.4.25 has similar patterns as in Fig.4.22. Only the gap statistic on k-means results gives the correct NC. Other methods cannot give a reasonable estimate of the optimal NC. Fig.4.26 shows the GMM and k-means clustering results using 5 clusters. K-means clusters the data relatively better. The five clusters are well separated, despite some data miss-clustered in the blue and purple clusters. GMM fits a specific kernel for the spike, which is shown as the red cluster on the left margin. Four major clusters are mixed into two clusters, and outliers are clustered into one group. The GMM clustering on the transformed data is not successful. The cross-validation results are 0.95 and 0.48 for k-means

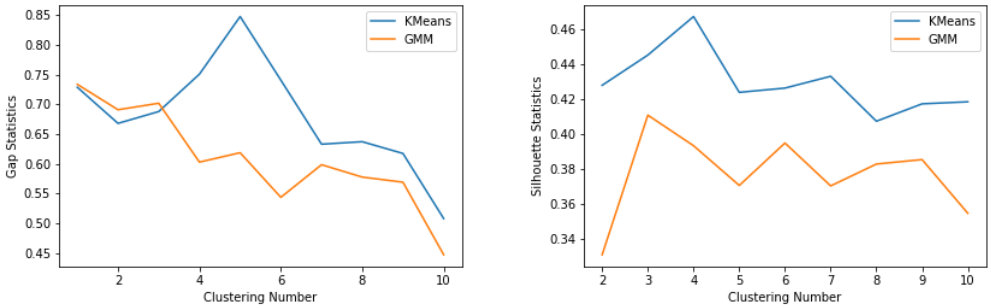


Figure 4.22: Results of gap statistic (left) and silhouette coefficient using k-means and GMM (right) when data are Gaussian transformed with spikes spread out.

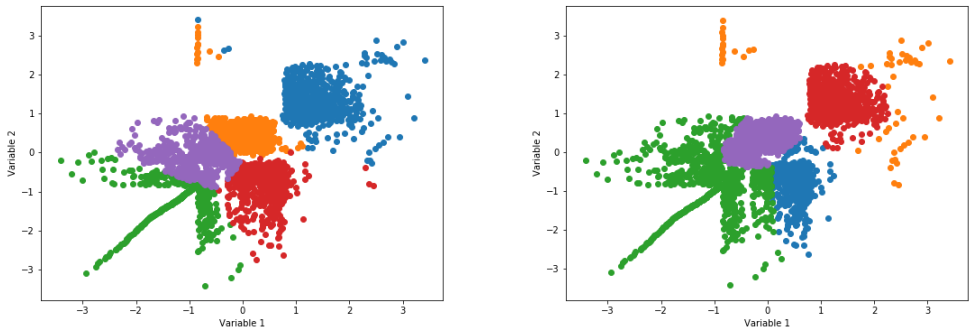


Figure 4.23: Clustering results using k-means (left) and GMM (right) when data are Gaussian transformed with spikes spread.

and GMM respectively. GMM only partitions the data. The correctness rate for k-means is 0.91, while the correctness rate for GMM is 0.60. The workflow on GMM results cannot provide correct optimal NC and clustering results are poorly separated, so GMM may not be compatible with the Gaussian transformed and spikes preserved data.

4.3.4 Summary for Different Transformations

Table.4.1 summarizes the compatibility of data transformations and clustering methods with the workflow. The correctness rates of k-means are generally higher than that of GMM. When BDL data and outliers present in data, k-means is more compatible with the transformations than GMM. Gaussian transformation with spikes spread is not an appropriate transformation as both clustering methods give low correctness rate. Linear transformation of data does not handle outliers well as it either gives incorrect NC or fails the cross-validation test. Although Gaussian transformation with spikes preserved performs well using k-means, only the gap statistic results give the correct NC. When analyzing real data, the transformation may have difficulty determining the optimal NC. For the methods having low correctness rate, they either do not give the optimal NC or partition data rather than cluster them. Since the synthetic data do not have complicated distributions to

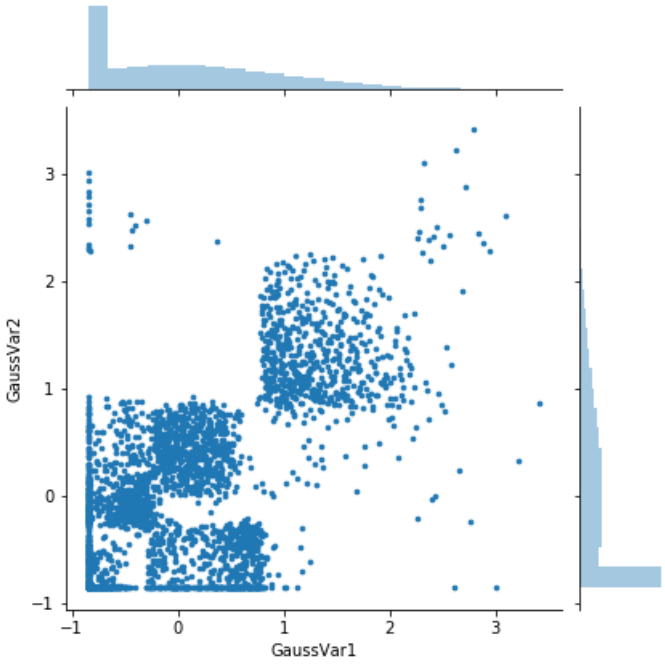


Figure 4.24: Gaussian transformed synthetic data with spikes preserved.

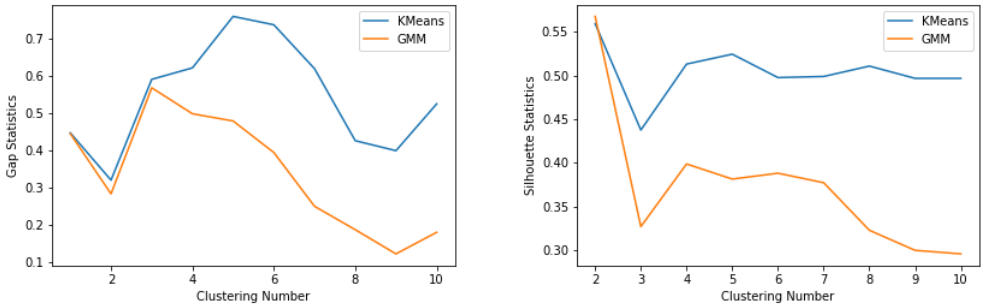


Figure 4.25: Results of gap statistic (left) and silhouette coefficient using k-means and GMM (right) when data are Gaussian transformed with spikes preserved.

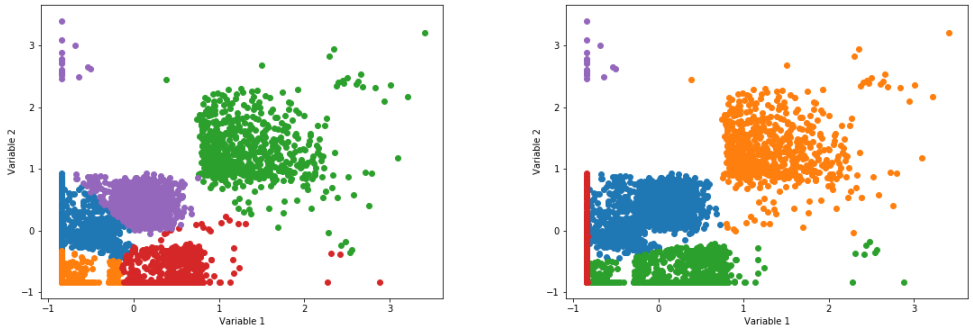


Figure 4.26: Clustering results using k-means (left) and GMM (right) when data are Gaussian transformed with spikes preserved.

	k-means	GMM
Linear Transform	0.90	0.72
Uni. Spread	0.90	0.78
Uni. Preserved	0.91	0.61
Gauss. Spread	0.77	0.43
Gauss. Preserved	0.91	0.60

Table 4.1: The correctness rates for each transform and the clustering methods.

be clustered, the appropriate transformations that can be applied on real data should have correctness rates above 0.9, which include linear transform, uniform transform with spikes spread and preserved, and Gaussian transform with spikes preserved.

There are two features worth noting: the poor clustering performance of GMM when spikes preserved, and the poor compatibility of Gaussian transformation with the workflow. GMM is a distribution-based clustering method. When spikes are preserved, their distribution is very different from the rest data. Therefore, GMM tends to fit a sharp Gaussian model with almost zero variance to spikes. The rest data cannot be grouped into this Gaussian kernel, even though they are close to the spike data. The synthetic clusters are relatively isotropic. GMM may have a better performance when the clusters are highly anisotropic. The poor compatibility between Gaussian transform and the workflow could come from the mechanism of the transformation. The reference state of the workflow is the uniform distribution, but the Gaussian transformation pushes data around the center of the space while the marginal data are separated farther, which changed the reference distribution of data. Thus, the Gaussian transformed data are not very compatible with the workflow.

4.4 Real Data Application

In this section, the appropriate transformations are applied to examine clusters in real data. As shown above, k-means outperforms GMM when BDL data are present, so only k-means clustering is applied here. The proposed workflow is applied to determine the NC. The real data used in this chapter come from the Government of the Northwest Territories (Falck et al., 2012). The missing data are eliminated. The data consist of 8500 observations and 46 variables. Some of the variables contain over 1000 BDL data as shown in the previous chapter. Since the NC in high-dimensional data cannot be visually examined, the workflow is of great importance.

Some may argue there are too few data for the 46 dimensional space, also known as "the curse of dimensionality" (Bellman, 1966; McLachlan, 2004; Taylor, 1993). For an M dimensional space, the number of data should be around the magnitude of 10^M to make meaningful statistical inference. It is necessary for mean and variance, but it wouldn't be a problem for cluster analysis. In Fig.4.27, there are only 10 data in a 2D space but the two clusters are still distinguishable. In cluster analysis,

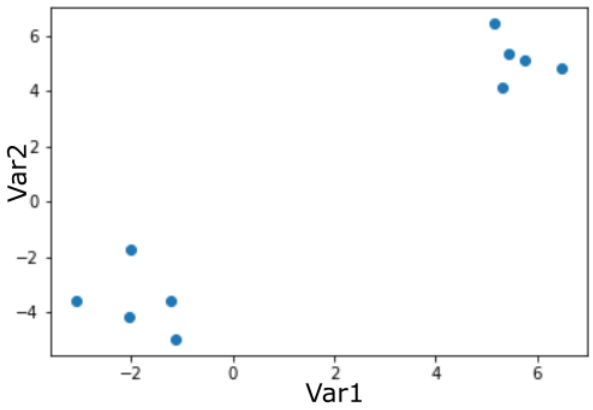


Figure 4.27: Ten samples on a 2d space still give two clusters.

the proximities between data are more important than the number of data.

The Hopkins statistic for the 46 dimensional data is 0.74, indicating there could be clusters existence. The data are transformed using transformation with correctness rate higher than 0.9 in the previous section. Fig.4.28 shows the gap statistic and silhouette coefficients on k-means clustering results. The gap statistic and silhouette coefficient show different trends. The optimal NC is between 6 to 10 for gap statistic, and different transformations give different plots. The uncertainty indicates the NC provided by the gap statistic may not be reliable. Also, the gap statistic should only be used as complement to the silhouette coefficient. All silhouette coefficient plots indicate the optimal NC is 2, and the value at 2 is much higher than other NC, so 2 clusters are used for the further analysis. The cross-validations on the four transformation are all above 0.9. It indicates k-means separates the two clusters in the real data well.

It is difficult to visualize the 46 dimensional space to validate the two clusters, but the high-dimensional data can be projected to an easily visualized 2D plane. If there are two obvious clusters on such a plane, we can conclude that there are at least two clusters in the real data. Since the workflow has eliminated possibilities of other NC, the two clusters on a 2D plane should be sufficient to validate the two clusters existence in the high-dimensional space.

Algorithm 2 explains the procedure of finding such planes. Real data are projected to a 2D plane, and k-means using 2 clusters are conducted on the 2D data, obtaining the corresponding silhouette coefficient. The process iterates thousands of times, and the 2D plane with the highest silhouette coefficient is returned. While there are an infinite number of 2D planes, the algorithm only needs to repeat enough times to find a certain plane that shows the clusters. The procedure of effectively sampling is explained in Algorithm 3. Imagine there are 3 clusters in a 3D space. To view the most separable clusters on a 2D plane, the plane should be the one created by the 3 cluster centroids. When there are only 2 clusters, planes are sampled parallelly to the two cluster centroids. The same logic can be applied to higher dimensional spaces. When there are more than 3 clusters,

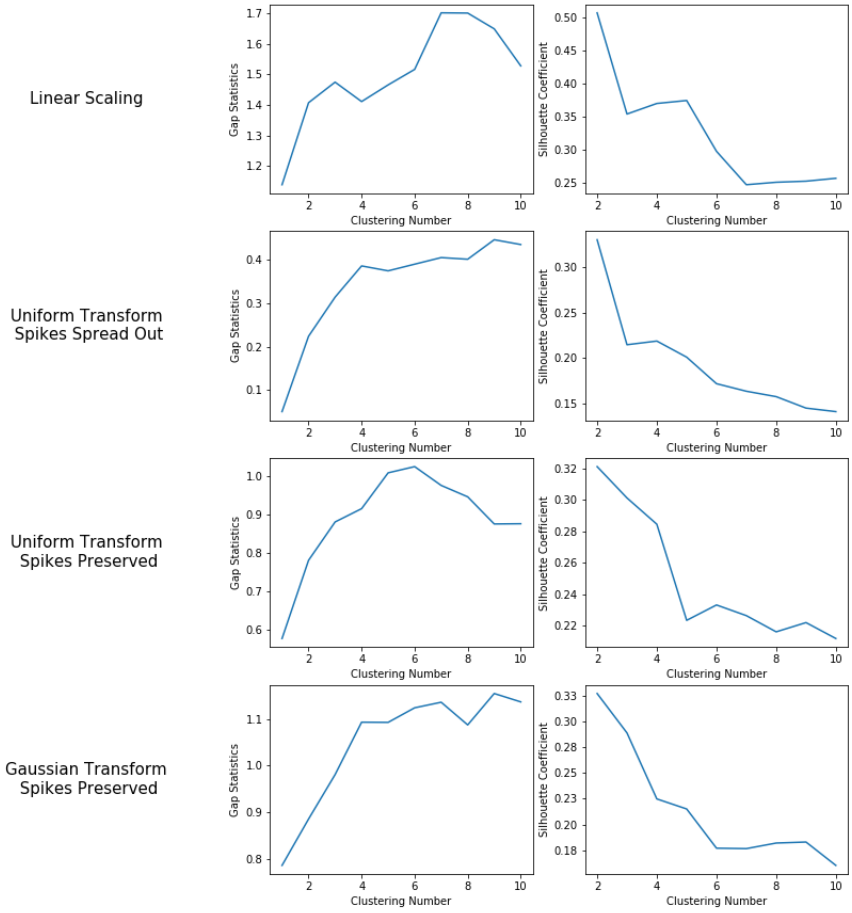


Figure 4.28: Gap statistic and silhouette coefficient results for different transform methods using k-means.

3 cluster centers are randomly sampled to generate a plane. The algorithm samples around the cluster centroids for more flexibility.

The purpose of the algorithms is not to find the plane that gives the highest silhouette coefficient. It only needs to find one plane that can show the two clusters of the projected 2D data. After applying the algorithm on differently transformed data, Fig.4.29 shows the 2D projected data that reveal 2 clusters. The coordinates do not have to be aligned with any variables. The two clusters in the left two figures are closely connected with some extent of overlap. They have elongated shape and can be distinguished by different orientations. The two clusters in the two right figures have elliptical shapes, and they can be distinguished by different sizes and cluster centers. The four transformed data all indicate 2 clusters on a 2D plane. Now, we can conclude that there can be two clusters in

Algorithm 2 Find 2D plane

$silht\_score = 0$
$N = 0$
Use the workflow on the high-dimensional data to find the optimal number of clusters n_cls
repeat
  $N = N + 1$
  Simulate a 2D plane in the high-dimensional space
  Project the high-dimensional data to the plane, obtaining the 2D projection data $data\_proj$
  Calculate the silhouette coefficient $silht\_hyper$ on $data\_proj$ using n_cls
  if $silht\_hyper > silht\_score$ **then**
    $silht\_score = silht\_hyper$
    $data\_save = data\_proj$
  end if
until $N > 10000$
return $silht\_score, data\_save$

Algorithm 3 Simulate 2D plane

Find the number of clusters K and the corresponding centroids $C_i(i = 1, \cdots, K)$
for i in range(K) **do**
  Sample a point P_i around cluster centroids C_i
end for
if N==2 **then**
  return A plane parallel to P_1 and P_2
else
  Sample 3 points from P_K
  return A plane created from the 3 points
end if

the high-dimensional data.

4.5 Conclusion

This chapter covers several aspects of cluster analysis on data with spikes. First, the workflow of determining the optimal NC is introduced. The workflow consists of several statistic tools. Hopkins statistic examines if there is any cluster in data. The silhouette coefficient and gap statistic find the optimal NC by plotting the value against a range of NC. The optimal NC should have the highest value of the measurements. Prediction strength validates if the chosen NC is reliable by applying the clustering on testing data. This chapter also examines the compatibility of the workflow with different transformations and clustering methods. The two clustering methods are k-means and GMM. Provided the true labels of synthetic data, correctness rate measures the proportion of data truly belong to the same cluster, and it is used to evaluate the performance of the clustering results. For the synthetic data, the suitable transformations are linear rescaling, uniform transform with spikes spread and preserved, and Gaussian transform with spikes preserved. The appropriate clustering method is k-means. These transformations and k-means clustering are applied on real high-dimensional data with many BDL spikes. The results show 2 clusters in all four differ-

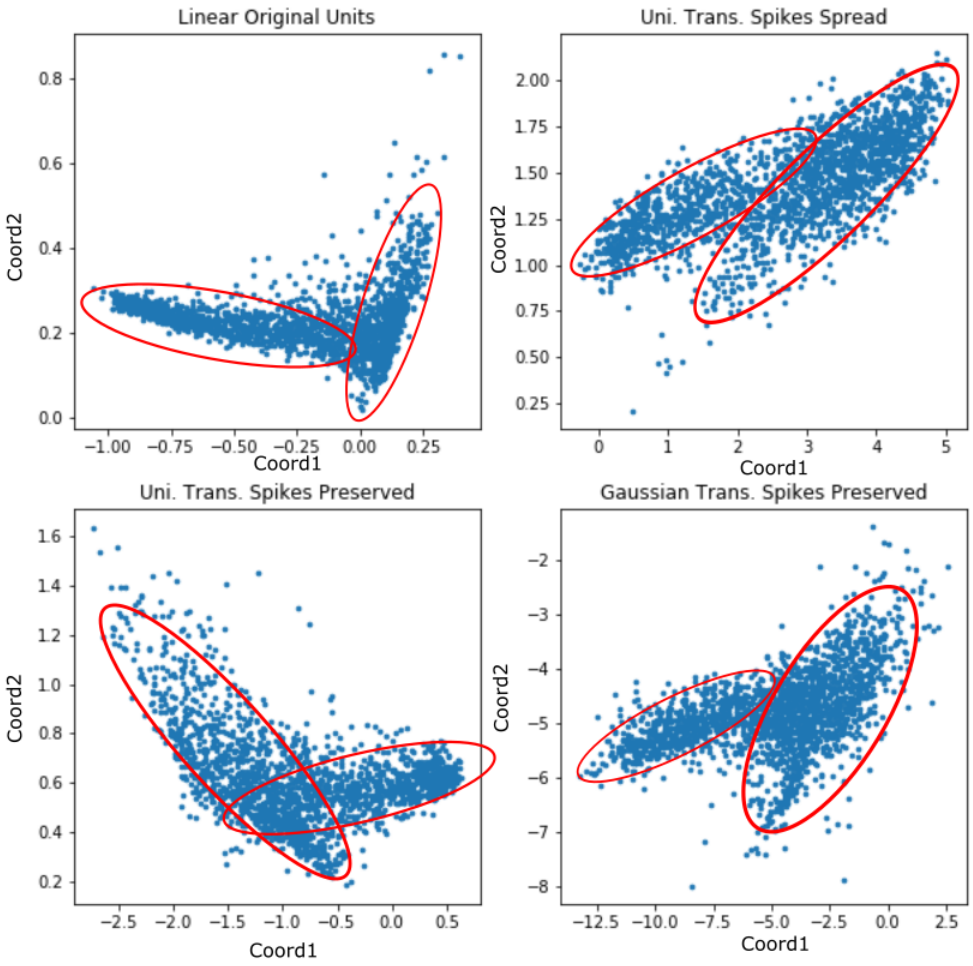


Figure 4.29: The planes illustrating two clusters in real data.

ently transformed data. To validate the results visually, high-dimensional data are projected to 2D planes, using the proposed algorithm to sample efficiently. All four figures indicate there are two clusters in the 2D projected data.

Chapter 5

Ensemble clustering and classification

Different from conventional multivariate data, geostatistical data have two components: multivariate and spatial. In traditional clustering methods, the multivariate data are clustered into continuous groups according to their multivariate properties but the clusters or groups do not have any spatial control. This is problematic for geostatistical modeling. Previous research used ensemble clustering to reach an optimum state between multivariate and spatial continuity, but the performance relies on the subjective choice of thresholds. In this chapter, an alternative workflow is proposed. The ensemble clustering method is still applied to generate multiple sets of multivariate labels. Given the clustering labels, classification on the spatial data is conducted considering both the multivariate and spatial continuity. The multivariate and spatial aspects are incorporated in an objective function. The importance of each term is adjusted through a spatial weight term. The classification starts from a random assignment of domains. Each data is iterated through all possible domains, and the domain label giving the minimum objective function value is preserved. The reassignment is conducted multiple times until it converges. The robustness of the workflow is demonstrated. The effect of spatial weight is assessed through multivariate geostatistical modeling. Practitioners can choose different spatial weight and number of domains, and make their own decision considering geological knowledge.

5.1 Introduction

5.1.1 Motivation

In exploratory data analysis (EDA), clustering techniques group large data into smaller groups, making further analysis more precise. When clustering geostatistical data, the incompatibility between the units and the properties of multivariate space and spatial coordinates can be problematic. Most clustering techniques ensure the continuity in multivariate space but the corresponding spatial domains are scattered, and this causes difficulties in geostatistical modeling (M. J. Pyrcz & Deutsch, 2014). For example, the scattered domains increase the difficulty of variogram inference and prediction of what domain label prevails at an unsampled location. In this chapter, labels refer to multivariate labels and domains refer to spatial labels.

In Fig.5.1, the bivariate data is clustered using k-means (Krishna & Murty, 1999). Note XCOO represents x coordinates, and YCOO represents y coordinates. Although the bivariate space is grouped into two continuous clusters, the spatial domains are scattered. In Fig.5.2, the k-means clustering is conducted on the spatial data although this is not recommended, because the cluster-

ing on spatial data cannot accommodate the complex spatial shapes of geological features. While the spatial continuity is assured, the multivariate clustering result is scattered. In these examples, there is a trade-off between the continuity of multivariate clusters and spatial domains (They are the same labels, but shown in different space).

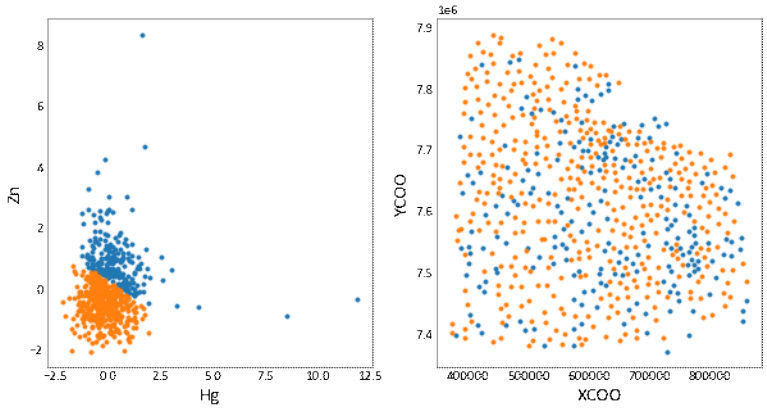


Figure 5.1: k-means clustering results on 2D multivariate data. Left represents the multivariate labels. Right is the domain distribution.

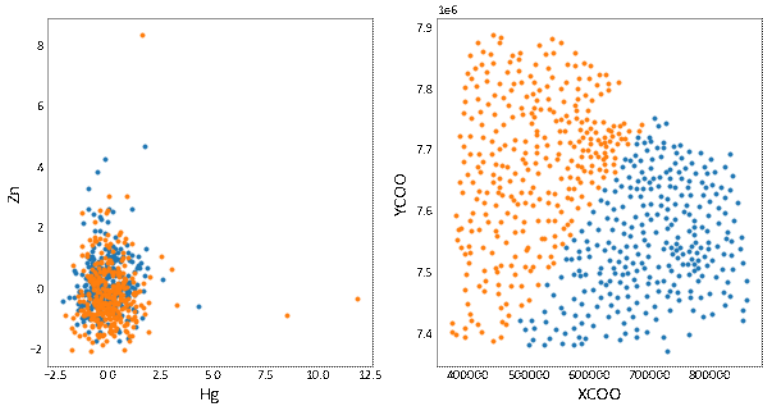


Figure 5.2: k-means clustering results on 2D spatial data. Left represents the multivariate labels. Right is the domain distribution.

To address the trade-off problem, Martin (2019) proposes a clustering method considering the spatial continuity and combined with ensemble clustering. This work extends Martin's work into an optimization framework. First, ensemble clustering is introduced. Ensemble clustering assembles multiple clustering techniques together to obtain an optimal clustering result (Fred & Jain, 2005). The individual clusterings used in the ensemble clustering are also called weak clustering (the term weak and individual clustering are used interchangeably). As shown in Fig.5.3, individual k-means clusterings (weak clusterings) do not group the clusters well, but the merged ensemble clustering result looks correct.

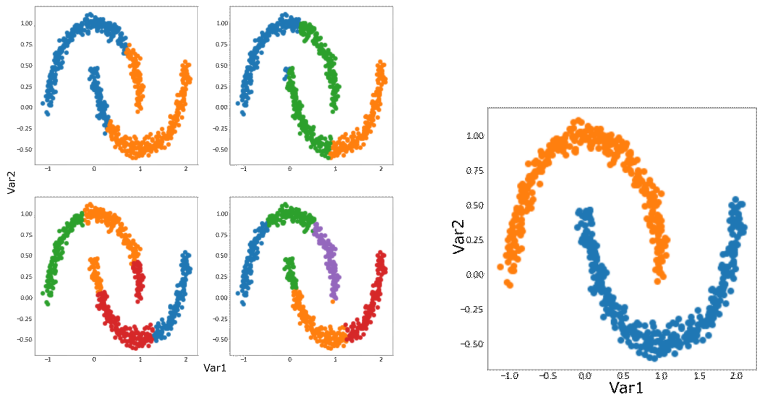


Figure 5.3: An illustration of ensemble clustering method. Left four plots are individual clustering results. Right plot is the merged ensemble clustering result.

The weak clusterings are generated similar to samples in random forests (Svetnik et al., 2003). Each individual clustering uses different clustering techniques, set of data, and number of clusters. These parameters do not have to be correct, but they are required to be various enough to generate independent and identically distributed samples. From the ensemble of these weak clusterings, a similarity matrix is obtained, and this matrix is used in hierarchical clustering (Johnson, 1967) to determine the final clustering results.

In Martin (2019), the individual clustering technique considers both the spatial domains and multivariate clusters. The performance of each individual result is measured quantitatively. The spatial continuity is measured using entropy (Shannon, 2001) and the multivariate continuity is obtained using within cluster sum of squares (WCSS). High entropy represents scattered domains and high WCSS represents scattered clusters. These two measurements are negatively correlated. As observed from the previous figures (Fig.5.1 and Fig.5.2), when the spatial entropy for a clustering result is high, the WCSS is low, and vice versa (Fig. 5.4). Reasonable clustering results should have low entropy and WCSS. To obtain the optimal ensemble clustering result, the entropy and WCSS of each individual clustering in the ensemble is calculated and only the individual clusterings with entropy and WCSS value below specific thresholds are used for the ensemble merging.

Although Martin's method provides reasonable results, subjective choices of the thresholds determine the quality of final labels. When deciding the thresholds, practitioners need to examine dozens of individual results. However, there are thousands of results that could be merged, so the quality of the ensemble result is uncertain and subjective.

5.1.2 Hierarchical clustering

Hierarchical clustering is a classic and robust clustering method in addition to k-means and Gaussian mixture model (GMM). It is used for merging individual clusterings in ensemble clustering.

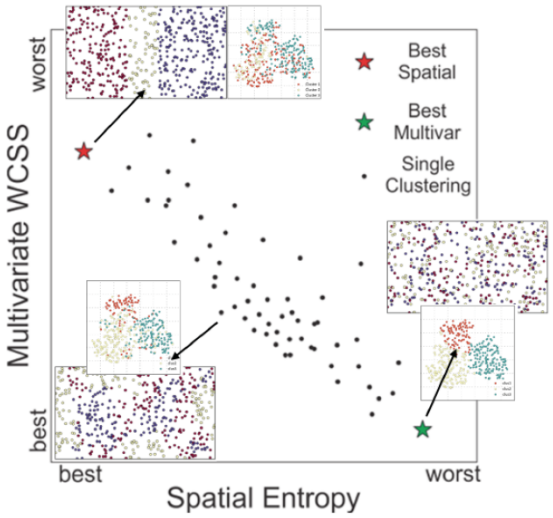


Figure 5.4: The WCSS and entropy are negatively correlated (Martin, 2019).

The main difference of hierarchical clustering from the other two is the clusters do not have centroids, so the resulting clusters can be in any shapes. This feature is useful when the shapes of clusters are uncertain. There are two types of hierarchical clustering: agglomerative and divisive. The first one is also referred to as "bottom up" approach and the latter one "top-down" approach (**?**). Here, the "bottom up" approach is considered. This approach treats each data as its own cluster at the start, and merges the closest pair at each step. The process goes iteratively until all data are merged into one group.

Similar to k-means, the distances between data are the key to hierarchical clustering results. Most commonly, Euclidean distance or Manhattan distance are used. In The intermediate steps, two groups are to be clustered together. Since each group contains many data, there are several ways to determine the distance between groups. The distance between groups is also called linkage criteria. Common linkage criteria include complete-linkage, single-linkage and average linkage (Fig.5.5). As observed from the figure, complete-linkage considers the maximum distance of data as the distance between groups, single-linkage considers the minimum distance of data as the distance between groups, and average-linkage considers the average distance of data as the distance between groups. There is also Ward's criterion (Murtagh & Legendre, 2014; Ward Jr, 1963). At each merge step, Ward's method merges the two groups which lead to the minimum group variance, and it provides similar results to average-linkage. In ensemble clustering, Euclidean distance and average-linkage are applied.

The last step of hierarchical clustering is to determine the number of clusters. This can be achieved from observing the dendrogram. Fig.5.6 is an example using only 20 data. The x axis shows the data index, and the y axis is the distance between each group. Each data is shown as leaf nodes and they are clustered into one group when the distance reaches the maximum. The smaller the distance between groups, the earlier they are grouped together. From the figure, there

are three major groups. To reveal these major groups, 3 clusters are chosen and the corresponding distance is at 1.1. The number of clusters can be chosen directly, or the group distance is set and the corresponding number of clusters can be easily found. In this chapter, the number of clusters is chosen directly.

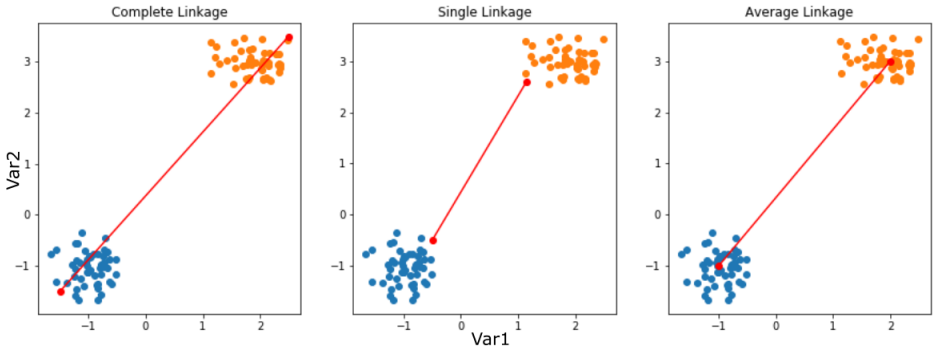


Figure 5.5: An illustration of three linkage method.

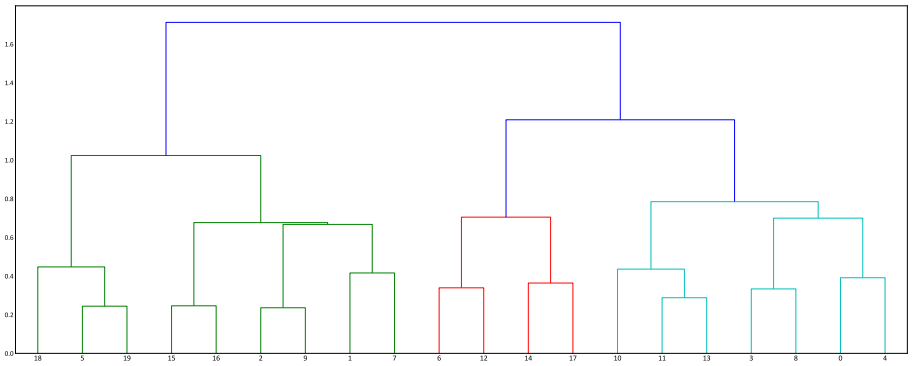


Figure 5.6: An example of dendrogram using 20 data. x axis is the index of data. y axis is the distance between data.

5.1.3 Proposed workflow

A novel workflow is introduced in this chapter. The traditional ensemble clustering is conducted first. The individual clustering techniques used is k-means where the parameters are sampled from a set of choices. The purpose is to generate independent and identically distributed random samples. The less correlated the individual cluster results are, the more robust the final ensemble results.

The next step is to use the ensemble clustering results as inputs for classification. An objective function in the classification considers both spatial and multivariate continuity, and their relative importance is adjusted through a spatial weight parameter. The inputs of the classification can be

multiple clustering results with different number of clusters. At the beginning, the domains are assigned randomly, and the objective function is calculated. Then each data are resampled through all possible domains. If the new domain improves the objective function, it is preserved. The process keeps iterating until the algorithm converges. Since there is a possibility of the algorithm converging to a local minima, multiple initial states are generated and the best result is kept. The number of domains and spatial weight are used as hyper-parameters. Practitioners can generate a matrix of results given a range of number of domains and spatial weight, then choose the optimal one considering the geological understanding of the domains. The robustness of the workflow is checked, and the effect of spatial weight is demonstrated through geostatistical modeling. With high spatial weight, the modeling simulates more connected values, while with low spatial weight, the modeling simulates more random disconnected values.

5.2 Ensemble clustering

The proposed workflow addresses the trade-off between spatial and multivariate continuity in two steps. The first step is to only consider the multivariate continuity, generating multiple clustering labels. These labels are used as inputs in the second step, in which the importance of spatial continuity is controlled by spatial weight. To obtain an optimal spatial and multivariate continuity, the quality of the inputs clustering labels are important. In this section, the procedure to obtain clustering labels using ensemble clustering is explained, and the superior results quality compared with k-means is demonstrated.

As explained above, ensemble clustering merges multiple individual clustering results to obtain a better one. The individual clustering used here is k-means with a set of different parameters. To sample independent realizations, each k-means realization samples 80% of the data with replacement and the number of clusters range from 10 to 25. 100 realizations of k-means results are used for one realization of ensemble clustering. The similarity matrix used for the merging step is generated by calculating the data's pairwise occurrence in the same group.

Suppose there are N data $\mathbf{z}(\mathbf{u}_i),\ i = 1, \cdots, N$, and individual clusterings are conducted M times on $\mathbf{z}(\mathbf{u})$. Each individual clustering $k = 1, \cdots, M$, has its own data samples and they are denoted as $\mathbf{z}_k(\mathbf{u})$. The clustering label for $\mathbf{z}(\mathbf{u}_i)$ is denoted as $y_k(\mathbf{u}_i)$. If $\mathbf{z}(\mathbf{u}_i)$ is not sampled in $\mathbf{z}_k(\mathbf{u})$, $y_k(\mathbf{u}_i) = NaN$. S is a $N \times N$ matrix representing the pairwise similarity. To calculate the similarity between the $\mathbf{z}(\mathbf{u}_i)$ and $\mathbf{z}(\mathbf{u}_j)$:

$$S_{ij} = \frac{\sum_{k=1}^{M} 1\{y_k(\mathbf{u}_i) = y_k(\mathbf{u}_j)\}}{\sum_{k=1}^{M} 1\{y_k(\mathbf{u}_i) \neq NaN \ \& \ y_k(\mathbf{u}_j) \neq NaN\}}, i,\ j = 1, \cdots, N \tag{5.1}$$

where $1\{True\} = 1$ and $1\{False\} = 0$. The similarity between data i and j is the proportion of their sharing the same label throughout their co-occurrence in M iterations. If two data are grouped together most of the time, they possibly belong to the same cluster. For example, in 100 realizations,

$\mathbf{z}(\mathbf{u}_i)$ and $\mathbf{z}(\mathbf{u}_j)$ are present together 80 times and are grouped together 60 times. Then S_{ij} is 0.75.

After obtaining the similarity matrix from Eq.(5.1), the distance matrix of the data is simply calculated as

$$M_{ij} = 1 - S_{ij}, \;\; i, \; j = 1, \cdots, N$$

The real data used in the ensemble clustering comes from "Kola Ecogeochemistry Project" (Filzmoser, Garrett, & Reimann, 2005; Reimann, 2005; Reimann, Filzmoser, & Garrett, 2005). The site is famous for its rich mineral deposits. The geochemical data consist of 618 data points and 26 variables. The data are standardized before clustering. The distance matrix (Fig.5.7) is used in hierarchical clustering to determine the ensemble clustering results. As observed from the figure, the distance is equal to 0 for diagonal elements, as the the distance to a data itself is 0. The larger the distance between data, the less likely they belong to the same cluster. The next step is to use it as the predefined distance matrix in hierarchical clustering. The advantage of hierarchical clustering is that the number of clusters does not change the clustering mechanism and the optimal number of clusters can be observed from dendrogram. As shown in Fig.5.8, closer data are grouped earlier. The small clusters grouped last on the left of the figure have long distance to the rest of the data, indicating they could be outliers. Average-linkage merging criterion is used here.

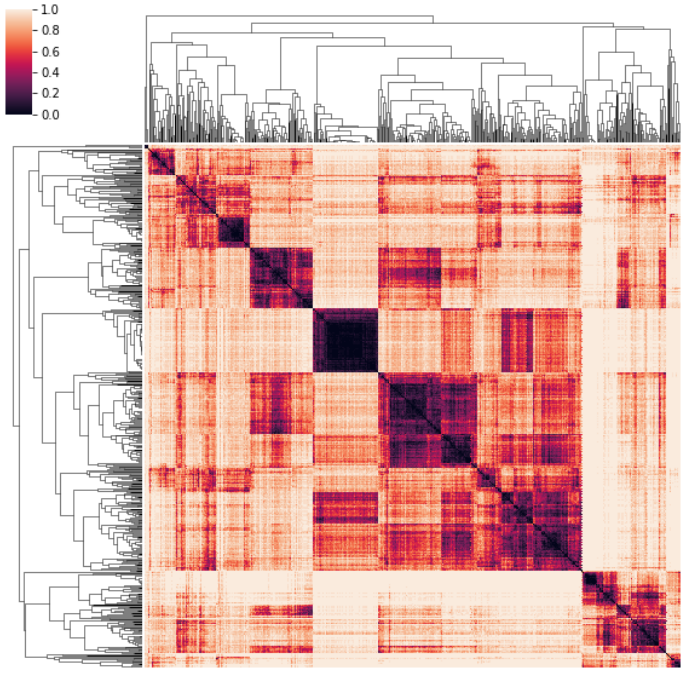


Figure 5.7: The distance matrix calculated from the ensemble clustering method.

Fig.5.9 and Fig.5.10 are the results of the ensemble clustering and k-means clustering respectively. The number of clusters ranges from 3 to 8. The silhouette coefficient is used to evaluate the performance of the clustering results. The higher the value, the better the data are grouped in multivariate space. Considering the general higher silhouette coefficient, the ensemble method

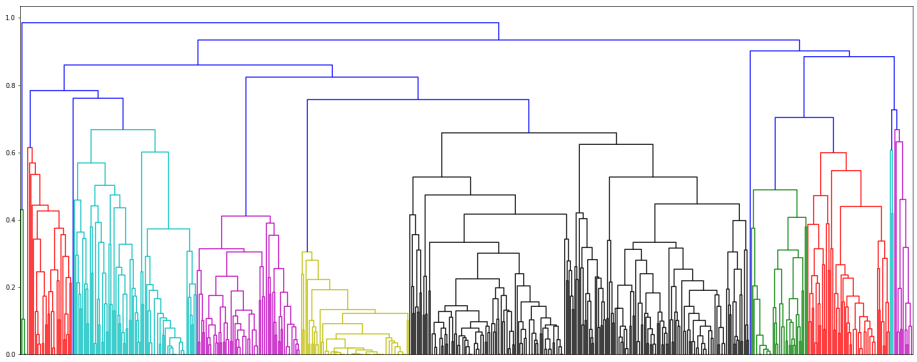


Figure 5.8: The dendrogram calculated from distance matrix. Each node on x axis represents a data point. y axis represents the data distance.

outperforms k-means. It is worth noting that in the ensemble clustering, when the number of clusters is low, the clustering technique mainly isolates outliers. In contrast, k-means groups outliers with main clusters, which can be the reason for its lower silhouette coefficient. The performance is only evaluated in the multivariate space and the spatial plot is only used for displaying the labels distribution.

Figure 5.9: The result of ensemble clustering on the real data. x and y axises represent location. Different colors represent different groups.

Figure 5.10: The result of k-means clustering on the real data. x and y axises represent location. Different colors represent different groups.

5.3 Classification

One advantage of the proposed workflow is that there can be multiple clustering labels used as inputs for the classification. Since there are some randomness in ensemble clustering, multiple ensemble labels can be generated, and the reasonable ones are inputs. The difficult problem of finding the best clustering label is avoided at the first step.

5.3.1 Objective function

Clustering ensures the continuity in multivariate space, but it does not make sense to conduct clustering on spatial data due to the complexity of geological shapes, a method is needed to ensure the spatial continuity of domains. Another issue is the clustering results may identify many small outlier groups. In the case of 8 clusters in Fig.5.9, there are only 2 to 3 main clusters. Finding major clusters is more important than identifying small outlier groups for the problem. The domain classification mitigates this problem.

Like other classification methods, the domain classification needs an objective function to indicate approaching an optimum. Since there is a trade-off, a hyper-parameter (W_{sp} for spatial weight) to adjust the importance of the spatial continuity is needed. Also the number of domains needs to be specified. The general format of the objective function is as follows:

$$O(\mathbf{d}, \mathbf{y}) = (1 - W_{sp}) \cdot M(\mathbf{d}, \mathbf{y}) + W_{sp} \cdot S(\mathbf{d}) \tag{5.2}$$

where $O(\mathbf{d}, \mathbf{y})$ is the objective function value, $M(\mathbf{d}, \mathbf{y})$ is the multivariate entropy, $S(\mathbf{d})$ is the spatial entropy, $\mathbf{d}$ represents the spatial domains and $\mathbf{y}$ represents the clustering labels. The objective

function has the following features: when the number of domains is fixed, with W_{sp} equal to 0, complete multivariate continuity is ensured. When W_{sp} increases, the domains are redistributed and deviate from the labels. When W_{sp} is close to 1, complete spatial continuity is ensured. The classified domains should give the minimum value of the objective function in each circumstance.

Now, the details of each term in Eq.(5.2) are explained. The multivariate entropy measures the discrepancy between multivariate labels and spatial domains. Suppose the number of domains is D, and the number of clusters is C. In domain i ($i = 1, \cdots, D$), the probability of finding label j ($j = 1, \cdots, C$), P_{ij}, is calculated as the number of data with label j in domain i, $N(\mathbf{y}_{ji})$), divided by the number of data in domain i, $N(\mathbf{d}_i)$.

$$P_{ij} = N(\mathbf{y}_{ji})/N(\mathbf{d}_i) \tag{5.3}$$

The entropy of domain i is calculated as:

$$E_i = -\sum_{j=1}^{C} P_{ij} \log P_{ij} \tag{5.4}$$

The multivariate entropy of the whole data is the weighted sum of E_i and is calculated as:

$$M(\mathbf{d}, \mathbf{y}) = \frac{\sum_{i=1}^{D} N(\mathbf{d}_i) \cdot E_i}{N} \tag{5.5}$$

where N is the total number of data. When $M(\mathbf{d}, \mathbf{y})$ is low, domains are consistent with labels. When $M(\mathbf{d}, \mathbf{y})$ is high, the labels are distributed randomly within domains. For example, Table.5.1 shows the calculation of the probabilities. There are 3 labels and 2 domains. The probability is calculated within each domain. After obtaining the probabilities, the entropy for domain 1 and 2 are:

$$E_1 = -(0.133 \log 0.133 + 0.2 \log 0.2 + 0.667 \log 0.667) = 0.86$$

$$E_2 = -(0.3 \log 0.3 + 0.5 \log 0.5 + 0.2 \log 0.2) = 1.02$$

The entropy for the whole data is the the weighted average of E_1 and E_2:

$$M = \frac{150 \cdot E_1 + 100 \cdot E_2}{250} = 0.924$$

The distribution of labels in domain 2 is more random, which means the domain and labels are less consistent. When the number of domains is the same as the number of labels and there is only one unique label in each domain, the multivariate entropy is zero, and this represents full multivariate continuity. In practice, there can be P sets of labels used for multivariate entropy calculation, and the final entropy is the average entropy over the P sets.

$$M(\mathbf{d}, \mathbf{y}_1, \cdots, \mathbf{y}_P) = \frac{\sum_{p=1}^{P} M(\mathbf{d}, \mathbf{y}_p)}{P} \tag{5.6}$$

The final objective function is adjusted from Eq.(5.2) to

$$O(\mathbf{d}, \mathbf{y}_1, \cdots, \mathbf{y}_P) = (1 - W_{sp}) \cdot M(\mathbf{d}, \mathbf{y}_1, \cdots, \mathbf{y}_P) + W_{sp} \cdot S(\mathbf{d}) \tag{5.7}$$

	Domain1	Domain2
Label1	20	30
Label2	30	50
Label3	100	20

	Domain1	Domain2
Label1	0.133	0.3
Label2	0.200	0.5
Label3	0.667	0.2

Table 5.1: Example data of calculating multivariate entropy. Left is the number of data within each label and domain. Right is the corresponding probabilities.

where P is the number of available sets of clustering labels.

The calculation of the spatial entropy is relatively simpler. We search in a local window, calculate the entropy of proportions of different domains, and average over all data. Fig.5.11 shows such a search window. The window is centered in a data point with a radius of 60 km. For data point $\mathbf{z}(\mathbf{u}_q)$ $q = 1, \cdots, N$, the spatial entropy $S(d_q)$ is calculated as

$$S(d_q) = -\sum_{k=1}^{K} p(k) \log p(k), \ q = 1, \cdots, N \tag{5.8}$$

where $p(k)$ is the proportion of domain k and K is the available domains. A value of $p(k) = 0$ would contribute 0 to the entropy (at the limit). The average $S(\mathbf{d})$ over all data is simply calculated as

$$S(\mathbf{d}) = \frac{\sum_{q=1}^{N} S(d_q)}{N}, \ \ q = 1, \cdots, N$$

The more scattered the spatial labels are, the higher $S(\mathbf{d})$ is, and this is not desirable. The optimal domain distribution should give the lowest possible $O(\mathbf{d}, \mathbf{y}_1, \cdots, \mathbf{y}_P)$.

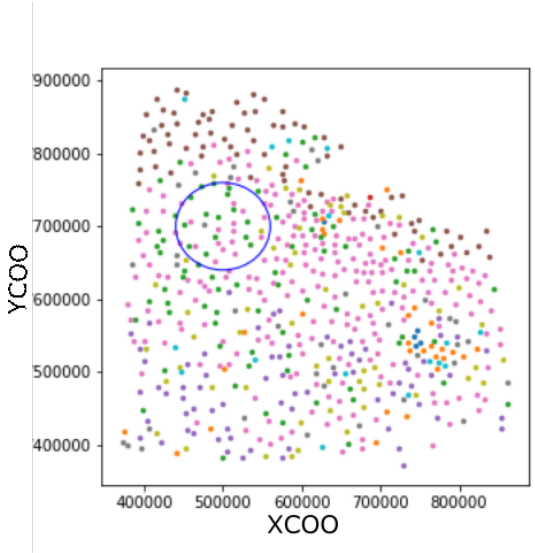


Figure 5.11: An illustration of a local search window. The window is marked as a blue circle.

5.3.2 Classification process

With the objective function established, it is used to classify spatial domains. The spatial weight and the number of domains need to be specified. Suppose there are 3 domains. The spatial domains are randomly assigned at the beginning. Domains $\mathbf{d}$ are uniformly sampled from 1, 2 and 3, and the initial measurement O_{init} from Eq.(5.7) is obtained. Then the domains are resampled one by one in a random order. If O decreases compared with the previous state, the new domain label is preserved, otherwise, dismissed. All data are visited once in each iteration, and be revisited in a different order in the next iteration until the algorithm converges. As shown in Algorithm 4, the objective function is recalculated every time when data domains are changed. If there are 3 domains and 600 data, in each iteration, the objective function is calculated 1800 times. The maximum iteration is set to 10. In practice, the algorithm converges much faster, because it can settle in a local minima. To overcome this problem, the algorithm is run multiple times, which is equivalent to initiating multiple beginning states, and the domains giving the lowest O are preserved.

Algorithm 4 Classification of the domains using the objective function.

Input: spatial weight W_{sp}, number of domains D and C sets of clustering labels $\mathbf{y}_1, \cdots, \mathbf{y}_C$,
The original order data is $z(\mathbf{u})$
Random assign the domains obtaining initial domains $\mathbf{d}_{init}$
$O_{init} = (1 - W_{sp}) \cdot M(\mathbf{d}_{init}, \mathbf{y}_1, \cdots, \mathbf{y}_C) + W_{sp} \cdot S(\mathbf{d}_{init})$
The current $O_{curr} = O_{init}$
$N_{count} = 0$
repeat
  $N_{count} += 1$
  Random order the data $z(\mathbf{u})$ obtaining $R(\mathbf{u})$
  for data i in $R(\mathbf{u})$ **do**
    for domain j in D **do**
      Assign domain j to data i, calculate temporary O_t
      if $O_t < O_{curr}$ **then**
        Keep j as the domain of data i
        $O_{curr} = O_t$
      else
        Change the domain of data i to the original one
      end if
    end for
  end for
until $N_{count} > 10$

With the classification procedure explained, its robustness needs to be validated. When $W_{sp} = 0$, the algorithm only considers the multivariate continuity. If there is only one set of clustering label input and the number of domains is set equal to the number of clusters, even starting from a random distribution of domains, a robust algorithm should return a domain distribution identical to the clustering labels. They may have different label names, but the spatial distribution should be the same. Fig.5.12 shows the resulting domains in this situation. The domains are identical to the input clustering labels. Although the colors are different, within a same domain, there is only one type

of clustering label, and this is verified by the 0 multivariate entropy.

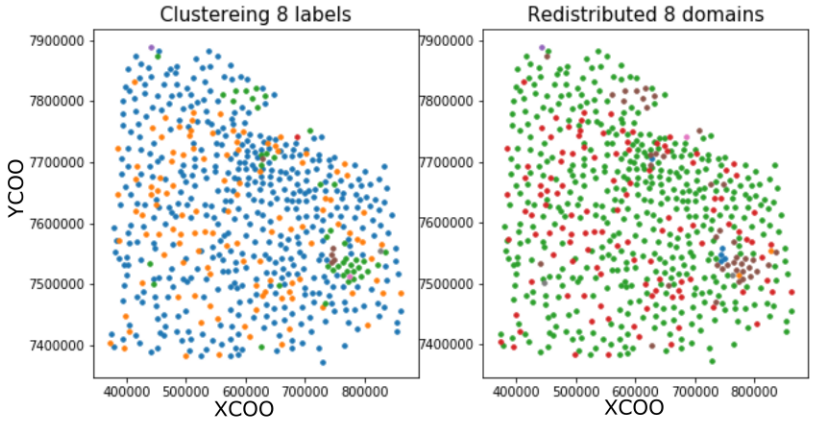


Figure 5.12: The classification of the domains when spatial weight is set to 0. Left is the input clustering labels. Right is the classified domains.

The algorithm is used to classify clustering labels with the spatial weight W_{sp} increasing. The spatial continuity increases and the multivariate continuity decreases. Fig.5.13 shows the input clustering labels for the algorithm. The multivariate data used for the ensemble clustering is standardized, which gives them 0 mean and variance of 1. This can decrease the influence of highly skewed distributions. The labels are obtained from multiple realizations of ensemble clustering with different numbers of clusters. The 6 inputs are denoted as $\mathbf{y}_1, \cdots, \mathbf{y}_6$. When calculating the objective function, O has the form of $O(\mathbf{d}, \mathbf{y}_1, \cdots, \mathbf{y}_6)$. The number of clusters ranges from 8 to 14, from which some are tiny outlier groups. When considering the number of domains, only the major groups are considered. Here, the number of domains is set from 3 to 5. Also, W_{sp} is another important hyper-parameter. Multiple W_{sp} values are tested and the change of spatial continuity is demonstrated.

Fig.5.14 shows the domains obtained from the input clusters in Fig.5.13. W_{sp} ranges from 0 to 1, showing the process of the domain classification emphasizing more on spatial continuity. In each small figure, MV represents multivariate entropy, and SP represents spatial entropy. The lower the entropy value, the higher the corresponding continuity. In each row (when the number of domains is fixed), as W_{sp} increases, the multivariate continuity decreases and spatial continuity increases, which is the desired performance when the algorithm is designed. In each column (when W_{sp} is fixed), when the number of domains increases, the multivariate continuity increases and the spatial continuity decreases. This is anticipated as more domains group data finer in multivariate space, and lead to less continuous domains. With W_{sp} lower than 0.25, the domains are fairly scattered, while with W_{sp} larger than 0.75, the domains are too continuous. In practice, W_{sp} can be set between 0.25 and 0.75, and be fine tuned.

When W_{sp} is zero, classified domains can be viewed as the averaged clustering labels. In some

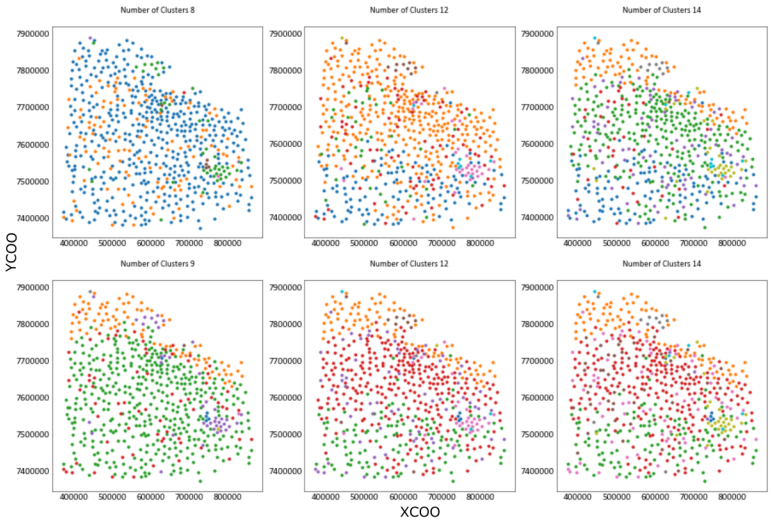


Figure 5.13: 6 sets of clustering labels obtained from ensemble clustering.

high W_{sp} figures, the actual number of domains are smaller than the defined value, because when the number of domains decreases, the continuity of domains increases. Since the domains are assigned randomly at the beginning, the final results have high uncertainty when W_{sp} is close to 1, because no clustering information can put constraints on the classification. In practice, clustering labels are considered and W_{sp} is set in a reasonable range, which stabilize the algorithm. Practitioners can generate their own matrix of domains and make decisions considering external geological knowledge.

5.4 Statistic Validation

The proposed workflow is conducted on data for the purpose of dividing them into groups with distinguishable features, and this can be verified by testing the within group variance (Kasim & Raudenbush, 1998). The relative size of domains is another measurement of the classification performance. These quantitative measurements can also be used to determine appropriate hyperparameters.

The total variance of data represents how scattered they are distributed. When the data are grouped into smaller clusters, there are within group variance and between group variance. If the groups are well clustered, the variance within each cluster should be small, and the corresponding between group variance should be large. It also means the differences between data within the same group are small and the differences between data in different groups are large.

Suppose there are N data, and they are classified into K groups. The following equation regard-

Figure 5.14: The matrix of domains, given multiple W_{sp} and number of domains.

ing variances holds:

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2=\frac{1}{N}\sum_{j=1}^{K}N_j(\bar{x_j}-\bar{x})^2+\frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N_j}(x_{ij}-\bar{x_j})^2 \tag{5.9}$$

where N_j is the number of data in group j, $\bar{x}$ is the grand mean over N data and $\bar{x_j}$ is the mean of group j. The first term is referred to as the total variance, the second term between group variance and the third term within group variance. The total variance is a constant when data is fixed. When data are classified into different groups, the latter two terms change. When dealing with multivariate data, the data are standardized in each dimension, the variances are calculated in each dimension separately, and the average variances over all dimensions are inputs for Eq.(5.9). For example, there are data with 5 variables. The standardized data have 0 mean and standard deviation of 1 in every dimension. Then the total variance is calculated as the average variance over 5 variables. In this case, the total variance is always 1. When data are clustered, in one of the clusters, the within group variances are 0.5, 0.6, 0.7, 0.8 and 0.9 for 5 dimensions respectively. The average within group variance 0.7 is used in Eq.(5.9).

Table.5.2 shows the within group variance of the results obtained from the domain classification. The within group variance may not have a dramatic decrease because the variances are the average over 26 variables. Some variables may not be very informative to help with the classification. Suppose an extreme case where half dimensions have within group variance of 0 and the

	Wsp 0.0	Wsp 0.25	Wsp 0.5	Wsp 0.75	Wsp 1.0
3 domains	0.913521	0.917744	0.851015	0.927525	0.982765
4 domains	0.882553	0.875699	0.798630	0.935520	0.933935
5 domains	0.762482	0.751216	0.846246	0.916249	0.919460

Table 5.2: The within group variance of the domains obtained from the classification.

	Wsp 0.0	Wsp 0.25	Wsp 0.5	Wsp 0.75	Wsp 1.0
3 domains	1.080367	1.063830	0.660792	0.874103	0.675675
4 domains	1.269141	1.271895	1.219986	1.257460	0.850766
5 domains	1.281053	1.413696	1.570471	1.200598	1.454580

Table 5.3: The entropy measurements of domain sizes obtained from classification.

other half have within group variance of 1, the resulting univariate within group variance is 0.5, so a 0.2 decrease of the average variance can be significant. When W_{sp} is around 0.5, the within group variance is smaller which indicates the data are better grouped. From the table, the optimal choice of hyper-parameters can be 5 domains with W_{sp} equal to 0.25. The results can be improved if the W_{sp} is finer tuned.

Another aspect to evaluate the performance of the classification is the domain size. The domain sizes should be similar, but there can also be cases where some small domains are very different from the rest. In general, similar size domains can represent well grouped data. To quantify this, the entropy of the domain probabilities is measured. For K domains,

$$E = -\sum_{i=1}^{K} P_i \log P_i \quad i = 1, \cdots, K \tag{5.10}$$

where P_i is the proportion of data grouped into domain i. From the equation, if the entropy is large, the domain sizes are similar. If the entropy is small, one of the domains dominates. Table.5.3 shows the entropy value for domains in Fig.5.14. These values are consistent with the distribution in the figure. 3 domains with W_{sp} equal to 0.5 leads to one large domain. The corresponding entropy is only 0.66. When there are 5 domains and W_{sp} is equal to 0.5, the domains have similar size and the entropy is 1.57. The entropy combined with the within group variance can provide information about distinguishable groups with similar size. Since low within group variance and high entropy value are preferred, simply dividing the values elementwisely in Table.5.2 by Table.5.3 gives the desired measurement. The merged results is shown in Table.5.4, in which lower values represent better classification results. From the table, choosing 5 domains with W_{sp} between 0.25 and 0.5 can result in reasonable results.

	Wsp 0.0	Wsp 0.25	Wsp 0.5	Wsp 0.75	Wsp 1.0
3 domains	0.845565	0.862679	1.287870	1.061116	1.454494
4 domains	0.695394	0.688499	0.654622	0.743976	1.097758
5 domains	0.595200	0.531384	0.538849	0.763161	0.632113

Table 5.4: The merged measurement of the domains performance.

	Al	Si	Mg
Al	1.000000	0.348090	0.046925
Si	0.348090	1.000000	0.004131
Mg	0.046925	0.004131	1.000000

Table 5.5: The correlation matrix of three variables.

5.5 Flow simulation

In this section, the effects of the spatial weight W_{sp} on geostatistical modeling are demonstrated. The workflow includes simulating a gridded multivariate model in the region of Fig.5.11, relating permeability models to multivariate models, and running flow simulation on the permeability models. 100 realizations of multivariate models are simulated for each W_{sp}. Several realizations are plotted for visual checking. The flow simulation is used to assess the results of 100 realizations. If the multivariate models are different, the resulting permeability models are different, and this is reflected in highly non-linear sensitive response variables such as breakthrough time.

5.5.1 Data preparation

The first step is to choose the variables to be used for multivariate modeling. The variables should be as uncorrelated as possible to assess the influence of different variables. The performance of multivariate modeling is also affected by the number of available data. Since there are 618 data, using 3 variables is appropriate. From Fig.5.15, there are some variables strongly correlated such as Cr and As. Modeling these correlated variables do not provide much extra information, so Al, Si and Mg are chosen to be the variables for the multivariate modeling. Table.5.5 shows the correlation of the three variables. Mg is not correlated with Al and Si. Al and Si are barely correlated. Fig.5.16 shows the scatter plots of data. From the figure, the data are more clustered in low value regions and generally skewed. There are no extreme outliers. Note the data are standardized, which is the reason for the negative values. Although the negative value has no physical meaning, it does not influence the flow simulation as the permeability models are generated from the relative values of the data.

The next step is to obtain domain labels. The domain labels are generated using the procedure demonstrated in Section 5.3. The clustering inputs are shown in Fig.5.17. To demonstrate the effect of different spatial weight, the spatial weight is chosen to be 0.0 and 0.7. The reason for not choosing

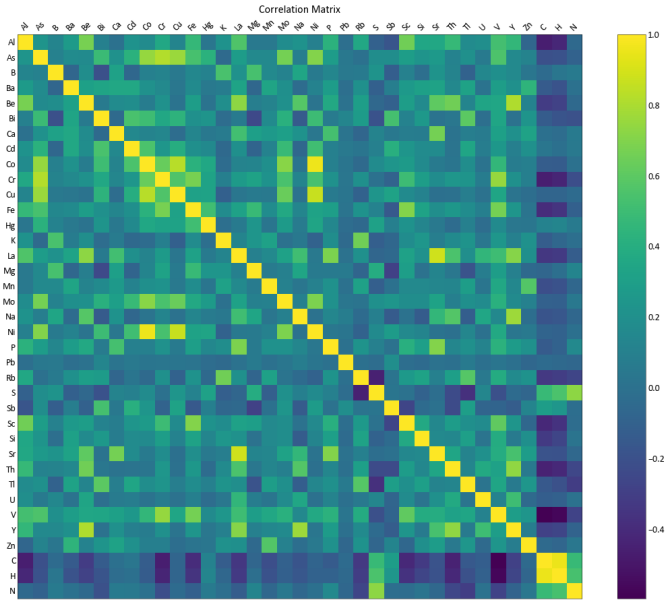


Figure 5.15: The correlation matrix of the data.

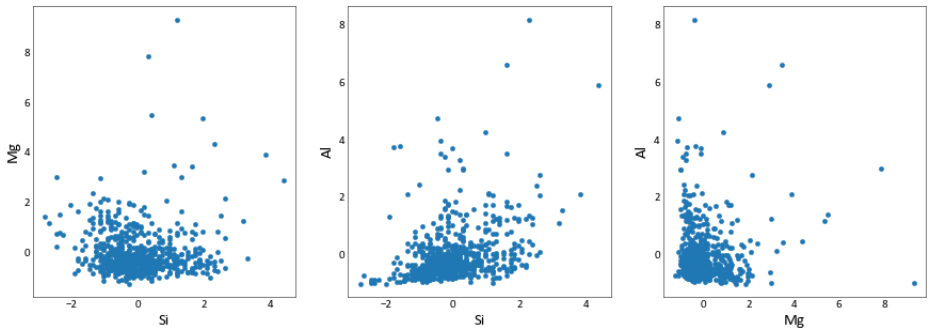


Figure 5.16: The 2D scatter plots of the multivariate data.

W_{sp} larger than 0.7 is that the full spatial continuity ignores the clustering inputs and the final domains may have artificial errors. The number of domains is set to be 3. Fig.5.18 shows the results of the domain classification and the location maps of the variables. As observed from the figure, the domain distribution for $W_{sp} = 0.0$ is scattered as only multivariate continuity is considered. Domain distribution of $W_{sp} = 0.7$ is more continuous as expected. Note the domain names can be different in each case. For example, domain 1 in $W_{sp} = 0.0$ is referred to as domain 0 in $W_{sp} = 0.7$. The data distributions are shown in the second row. For Al , high value data cluster in south east region. For Si , the southern region has general higher value than the northern half. On the contrary, Mg has higher values in the northern part.

Fig.5.19 is the reflection of the domain labels in multivariate space, in which the distributions are consistent with the statement of this chapter. Spatially scattered domains have continuous multivariate clusters, while spatial continuous domains have scattered multivariate clusters. Fig.5.20 illustrates 3 realizations of the domain models for each W_{sp}. The simulated domain layouts are consistent with the domain labels. The grid size is 50×50. Data in Fig.5.18 are modeled independently within each domain, and merged together based on their domain models in Fig.5.20. For example, data labeled as domain 0 are used as inputs for multivariate modeling, and the results are only kept where the grid cell is labeled 0.

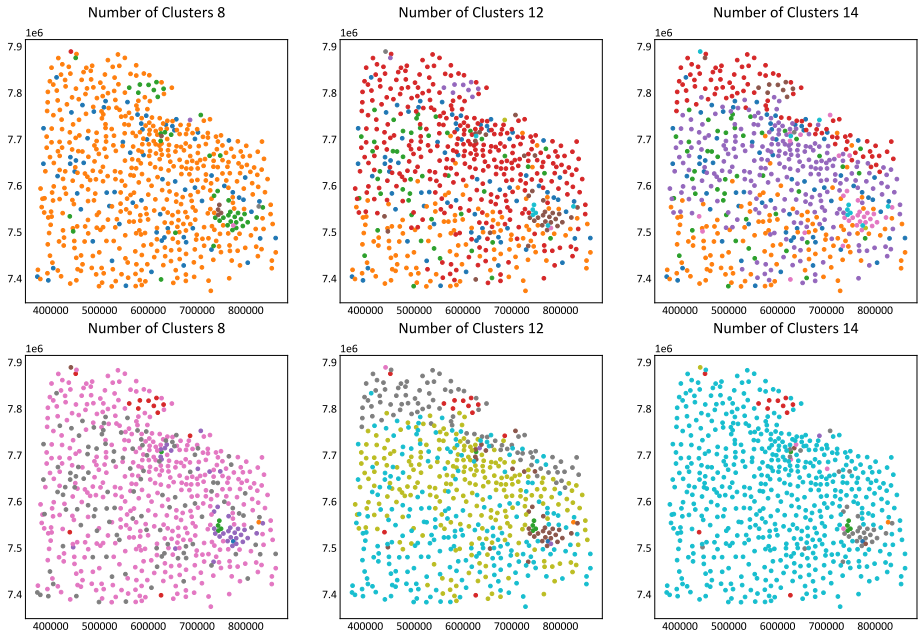


Figure 5.17: The cluster labels used as inputs for the domain classification.

5.5.2 Multivariate modeling

When conducting multivariate modeling, it is important to model the variables efficiently and retain their multivariate shape. Projection pursuit multivariate transform (PPMT) is used for this purpose (Barnett, Manchuk, & Deutsch, 2014; Barnett, Manchuk, Deutsch, et al., 2016). The idea is to use a series of methods to transform the multivariate data to an identical multi-Gaussian shape, model them independently and back-transform, returning the original multivariate relations. The transformation methods include linear decorrelation such as principal component analysis (Abdi & Williams, 2010) and min/max autocorrelation factors (Vargas-Guzmán & Dimitrakopoulos, 2003). These transformations start with sphered data (normal scored and with a correlation of 0). Correla-

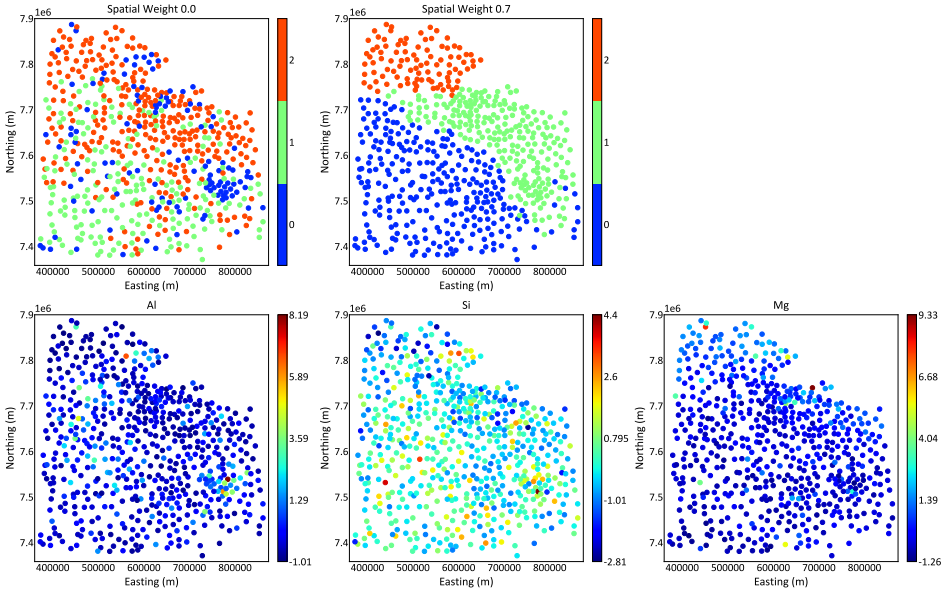


Figure 5.18: The domain labels and the location map of the three variables.

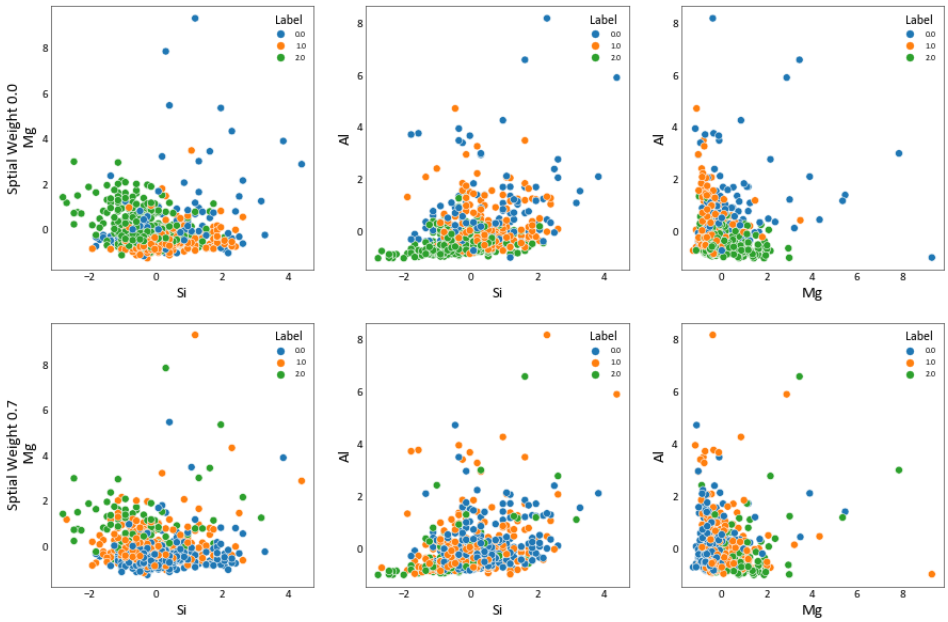


Figure 5.19: The domain labels in multivariate space. Upper row for $W_{sp} = 0.0$. Lower row for $W_{sp} = 0.7$.

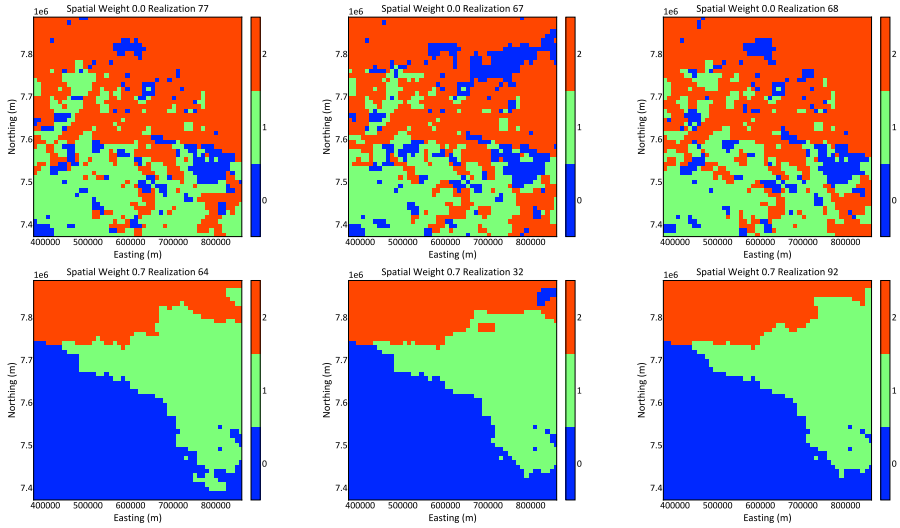


Figure 5.20: Categorical modeling of the domains with grid size 50×50.

tion only describes their linear relation, so the sphered data do not promise a multi-Gaussian shape. If data have multi-Gaussian shape, when projected to any one 1D dimension, they should retain a univariate Gaussian shape, and this 1D dimension does not have to be aligned with the coordinates.

The PPMT method projects the sphered data to a 1D dimension that exhibits the most non-Gaussian shape currently and normal scores the data in that dimension. Then the multivariate data is projected to the second most non-Gaussian 1D dimension and normal scored again. The process is repeated iteratively until the multivariate data exhibits the desired level of multi-Gaussian shape. These 1D transforms are recorded for the back-transformation. Fig.5.21 shows the plots of the multivariate data after PPMT when $W_{sp} = 0.7$, and they are in multi-Gaussian shapes. Note the multivariate models are simulated independently in each domain, so the data are transformed separately based on their domain label. With the data in multi-Gaussian shape, the variables can be modeled independently.

Since the data are modeled within each domain and each variable independently, there are 9 variograms inferred for each W_{sp}. As observed from Fig.5.22 and Fig.5.23, the variograms ranges are longer for Al and Mg. When $W_{sp} = 0.7$, the variograms are relatively more stable, because the variogram inference have more available pairs when domains are more continuous, The major direction is 110°azimuth and the minor direction is 20°azimuth. When the resulting variograms are not stable, omnidirectional search is used. Note the variograms are inferred from the normal score transformed data, not the PPMT transformed data (refer to (Barnett et al., 2014)). Each variogram generates 100 realizations of the univariate models. These univariate models are back-transformed to the original units first and then combined with the domain realizations in Fig.5.20, giving 100

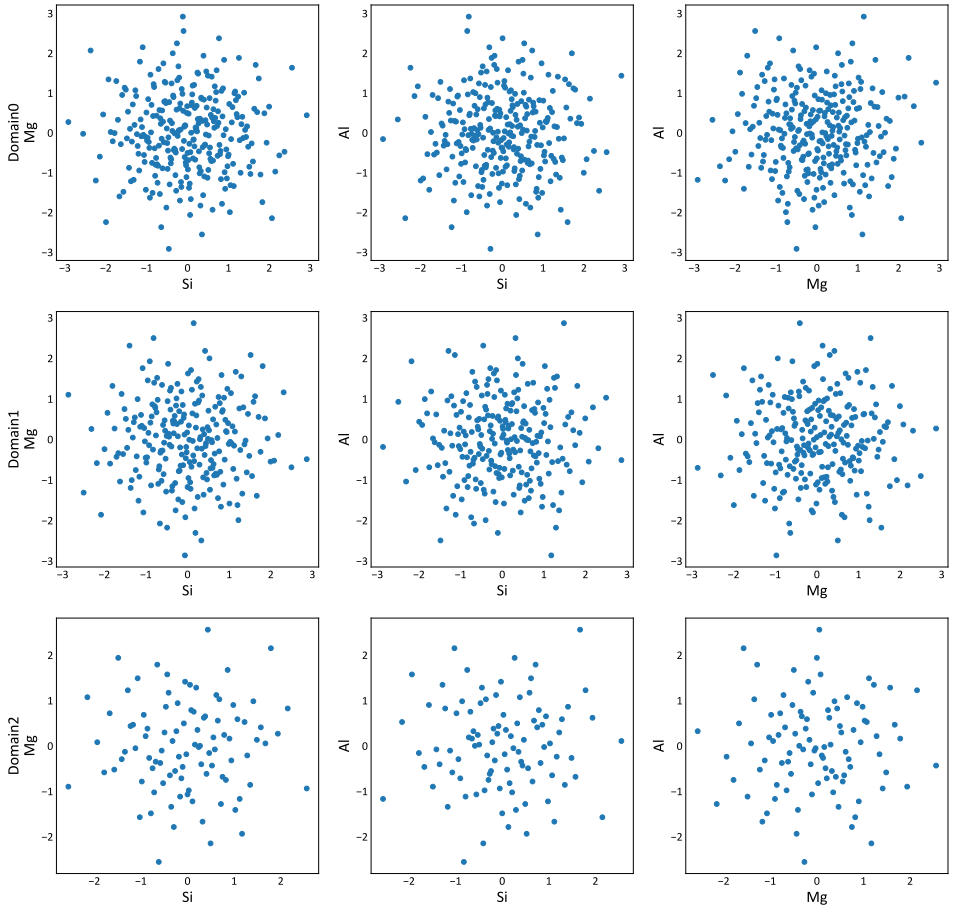


Figure 5.21: The scatter plots of the variables after PPMT. Each row represents the transformed multivariate data in each domain.

realizations in original units for each variable.

The back-transformed and domain merged data are shown in Fig.5.24. The simulated models have similar data distribution as in Fig.5.18. For example, the high value region in the north of the Mg datamap is preserved in the simulations. Although in both cases the simulated realizations retain similar spatial distribution, it is worth noting that the realizations in $W_{sp} = 0.0$ is smoother than those in $W_{sp} = 0.7$. This feature mainly comes from the conditioning data. As observed from Fig.5.19, when $W_{sp} = 0.0$, within each domain the multivariate data share similar values. The conditioning data sharing similar values result in smooth simulated results. Although the domain simulations in Fig.5.20 are scattered, the abrupt changes mostly occur on the domain boundaries. On the contrary, the domain realizations are more continuous in $W_{sp} = 0.7$, but the conditioning

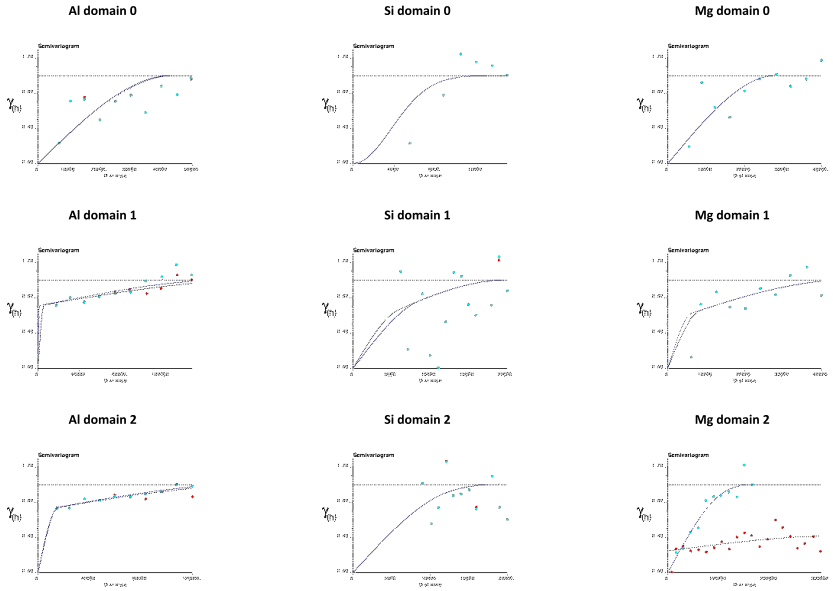


Figure 5.22: The variograms of variables in each domain for $W_{sp} = 0.0$.

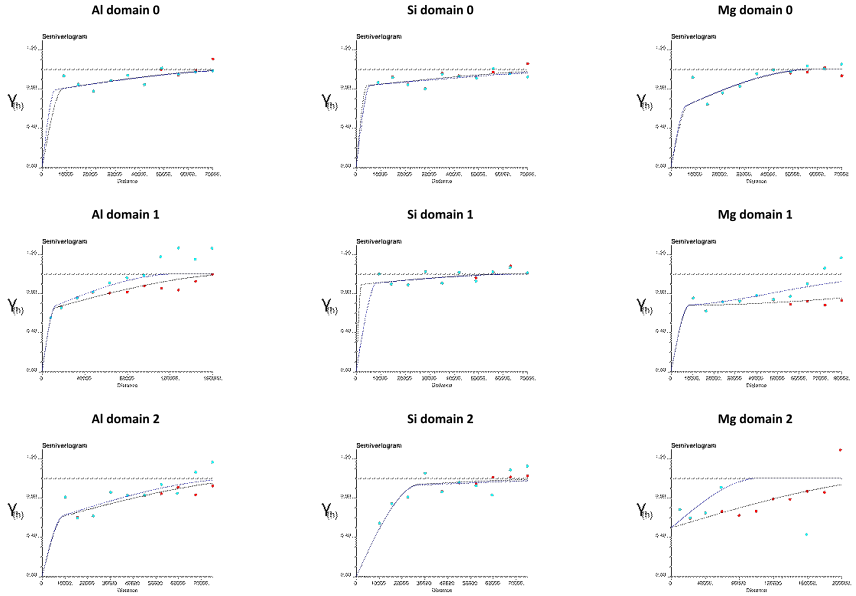


Figure 5.23: The variograms of variables in each domain for $W_{sp} = 0.7$.

data are multivariate scattered, so the resulting multivariate simulations are not as smooth. This feature may be less obvious when the number of domains increases and the domain simulations are more scattered, but their effect on the simulation smoothness may not be as dominant as the conditioning data.

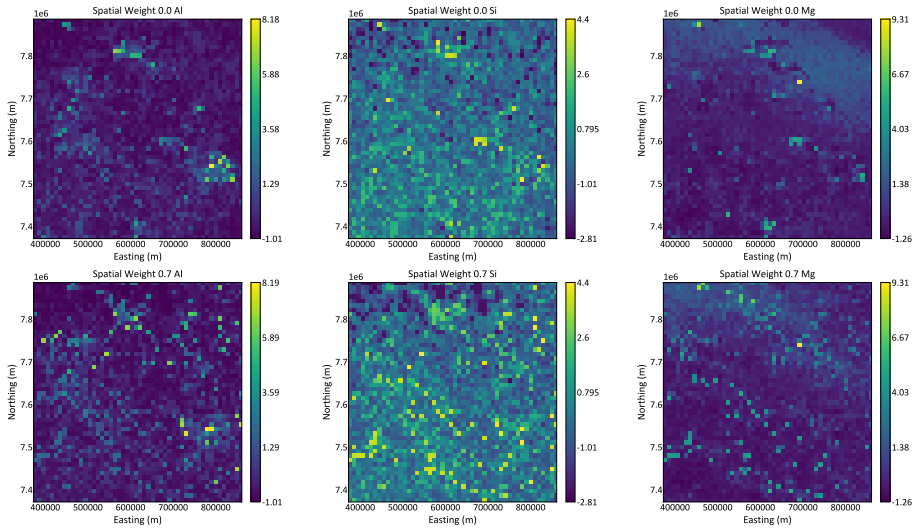


Figure 5.24: One of the realizations of three variables after merging the domain labels. The upper row is for $W_{sp} = 0.0$. The lower row is for $W_{sp} = 0.7$.

Fig.5.25 illustrates the scatter plots of the merged results. The top row is the original data plots (618 data) , the middle row is the plots of one realization of $W_{sp} = 0.0$ results (2500 data), and the bottom row is the plots of the one realization of $W_{sp} = 0.7$ results (2500 data). Both simulation results retain the original multivariate shape of the data. The difference lies in the proportion of low and high values. The $W_{sp} = 0.0$ results have a larger proportion of low values, while the $W_{sp} = 0.7$ results have a larger proportion of high values. This feature may come from the data merging. In the upper row of Fig.5.20, high values are mostly in the blue domain. When merged in the last step of modeling, most of the high values are clipped. While in the $W_{sp} = 0.7$ case, high values are grouped into three domains and easier to be preserved in the merging step.

5.5.3 Flow Simulation

The observations of Fig.5.24 and Fig.5.25 are based on visual check of several realizations. Flow simulation is used on all realizations, validating the previous observations. First, the multivariate models are converted to univariate permeability models. Since the data are standardized, when adding the variables together, they should contribute similarly to the sum. A new variable D is

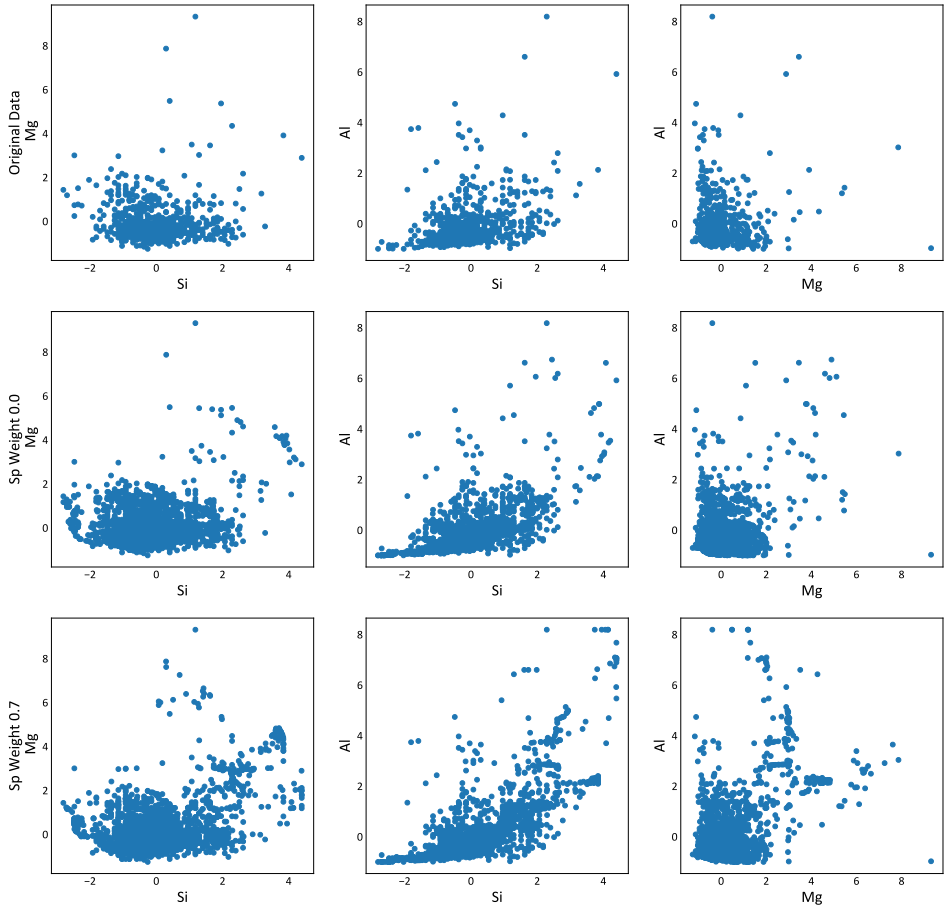


Figure 5.25: The scatter plots of the variables from original data and the realizations of $W_{sp} = 0.0$ and $W_{sp} = 0.7$.

defined to be related to permeability.

$$D(\mathbf{u}_i) = Z_{Al}(\mathbf{u}_i) + Z_{Si}(\mathbf{u}_i) - Z_{Mg}(\mathbf{u}_i) \;\; i = 1, \cdots, N \tag{5.11}$$

where $\mathbf{u}_i$ represents the location i, N is the total available data. For the modeling case, N is equal to 2500. Note the negative Mg is added, because the spatial distribution of Mg values is opposite of Al and Si.

To demonstrate the different low and high values proportion observed in Fig.5.25, $D(\mathbf{u})$ is generated for each case ($D_{0.0}(\mathbf{u})$ and $D_{0.7}(\mathbf{u})$ for $W_{sp} = 0.0$ and $W_{sp} = 0.7$ respectively), and two universal thresholds (T_{high} and T_{low}) are defined for converting $D(\mathbf{u})$ to permeability. When $D(\mathbf{u})$ is above T_{high}, the collocated permeability is set to 10 mD. When $D(\mathbf{u})$ is below T_{low}, the collocated

permeability is set to 0.1 mD. When $D(\mathbf{u})$ is in the middle, the collocated permeability is set to 1 mD. In this case, T_{high} is chosen from the 0.8 quantile of the 100 realizations of $D_{0.0}(\mathbf{u})$ and $D_{0.7}(\mathbf{u})$ (1.1), and T_{low} is chosen from the of 0.2 quantile of the 100 realizations (-1.45). Fig.5.26 shows 3 realizations for each W_{sp}. The realizations do not show significant difference as the multivariate models are generated from the same conditioning data. The low and high value regions does not vary significantly. When converted to permeability, two thresholds group the multivariate data into three categories, making the difference less obvious. The difference between the two W_{sp} generated permeability models needs to be examined through flow simulation. Note the proportion of high, medium and low values are different in each realization.

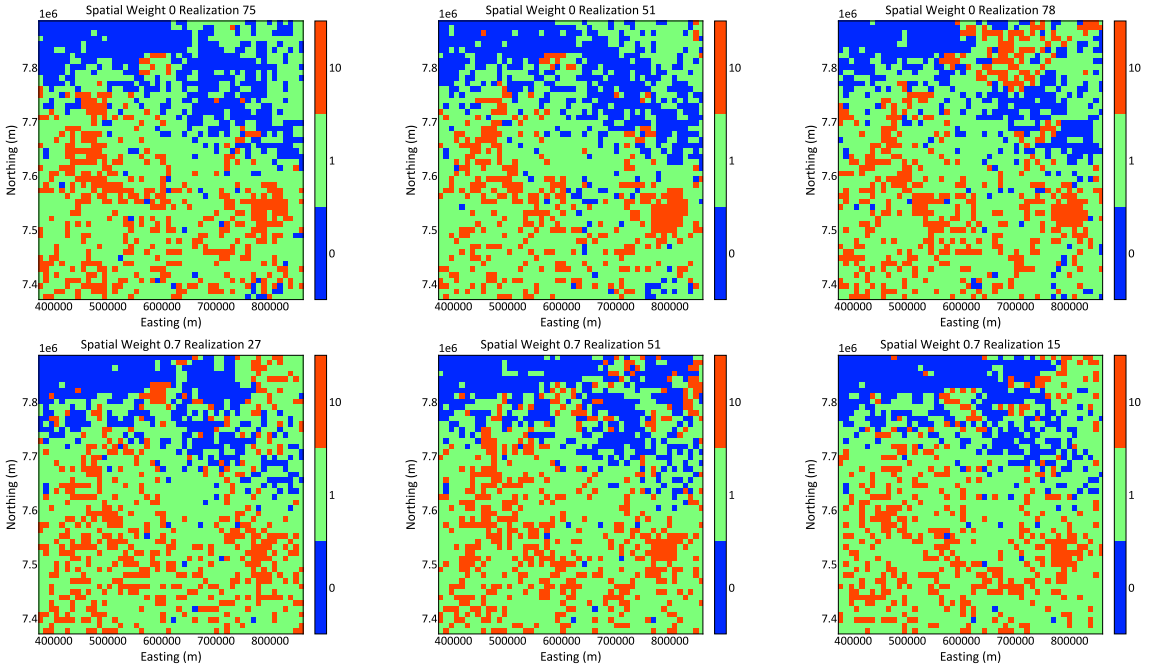


Figure 5.26: The realizations of permeability model using universal thresholds. The upper row for $W_{sp} = 0.0$. The bottom row for $W_{sp} = 0.7$.

With the converted permeability models, a flow simulation can be established. The flow simulation is a simple constant head model, where the left margin has a hydraulic head of 10 m, and the right margin has a hydraulic head of 0 m. The particles flow from left to right. Fig.5.27 shows one realization of the flow paths. There are 100 particles generated for each realization and their breakthrough time to the right margin are recorded, which is determined by the proportion of high and low values. Fig.5.28 shows the breakthrough times of the fast (P_{15}) and slow (P_{85}) particles. They are controlled by the number of 10 mD cells and 0.1 mD cells respectively. In the left figure, when $W_{sp} = 0.7$ most particles arrive before 44 seconds, and when $W_{sp} = 0.0$ most particles arrive after 44 seconds . The faster breakthrough time indicates more high value (10 mD) cells in the permeability models when $W_{sp} = 0.7$. In the right figure, when $W_{sp} = 0.7$, most breakthrough times are earlier than 160 seconds, while when $W_{sp} = 0.0$, the majority of the breakthrough times range from 130 to

200 seconds . In this case, the later breakthrough times indicate there are more low value cells in the permeability models when $W_{sp} = 0.0$. The flow simulation results only demonstrate the effects of different W_{sp}. We do not know which scenario is closer to the correct reference true values.

Since the permeability models are related to multivariate models, the flow simulation validates the observations of the multivariate scatter plots. More spatial continuity simulates a higher proportion of high values, and more multivariate continuity simulates a higher proportion of low values. A W_{sp} value considering both spatial and multivariate continuity should give intermediate results. Most of the geostatistical data share similar multivariate shape as in the real data used here – most data are skewed in low value regions with some high value outliers existing, so the observations can be generalized to other geostatistical data.

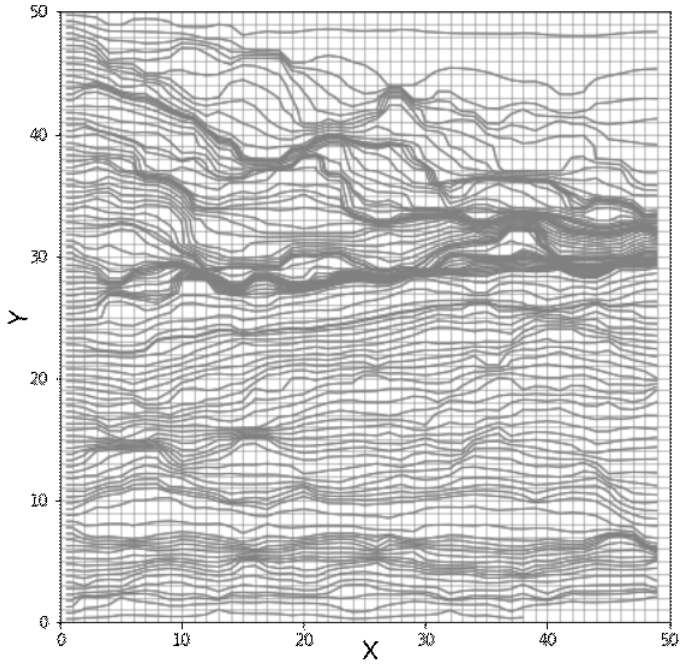


Figure 5.27: One realization of the flow path. The left margin has a hydraulic head of 10 m. The right margin has a hydraulic head of 0 m.

5.6 Conclusion

This chapter demonstrates the trade-off between multivariate and spatial continuity. A novel workflow combining ensemble clustering and classification accommodates the issue. For this dataset, ensemble clustering shows better performance than traditional methods through silhouette coefficient, and clusters multivariate data which are used as inputs for classification. An objective func-

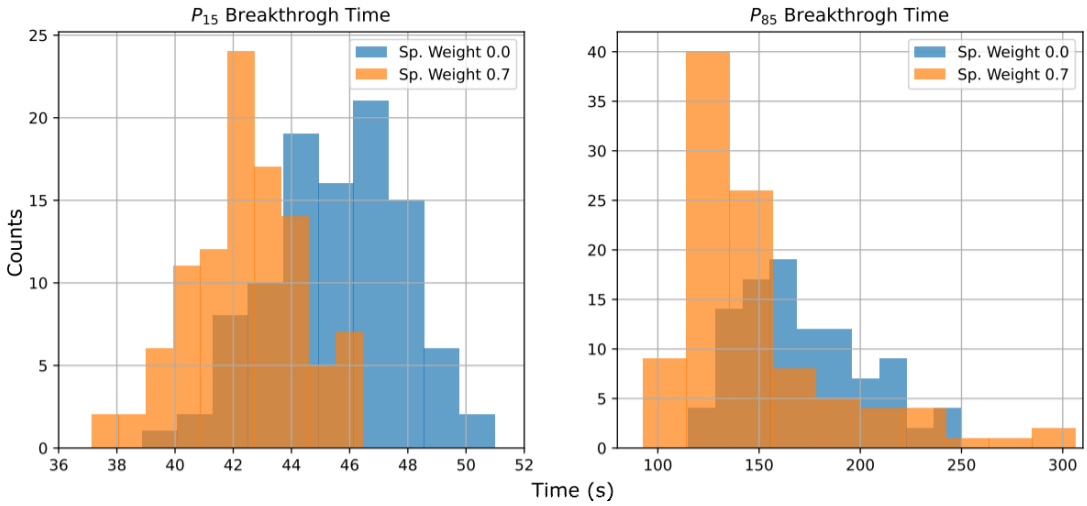


Figure 5.28: The histograms of the arrival time for quantile 0.15 (left) and quantile 0.85 (right) particles of 100 realizations (permeability converted from universal thresholds case). Blue histograms represent $W_{sp} = 0.0$ breakthrough times and orange histograms represent $W_{sp} = 0.7$ breakthrough times.

tion for classification is formulated. The algorithm takes spatial weight and the number of domains as hyper-parameters, and can use multiple clustering labels as inputs. With higher spatial weight, classified spatial domains are more continuous. The variance test validates the effectiveness of the classification method on the demonstrated data by showing the decreased within group variance. Combined with spatial entropy, the variance test also provides a tool to choose the appropriate hyper-parameters. Geostatistical modeling and flow simulation demonstrate the effect of spatial weight in a practical way. PPMT provides an efficient way to simulate multivariate models independently. The two universal thersholds convert the back-transformed multivariate models to univariate permeability. The flow simulation model has constant heads on two edges. Histograms of the early arrival and late arrival particles validate the observations of the scatter plots. For this dataset, when spatial weight is close to zero, higher proportion of low values are simulated. When spatial weight is close to 1, higher proportion of high values are simulated. A reasonable intermediate W_{sp} should give reasonable spatial and multivariate continuity.

Chapter 6

Conclusion

Exploratory data analysis (EDA) constitutes an early and important step in geostatistical modeling. Missing data, below detection limit (BDL) data can cause concerns in data transformation and further analysis. Statistical tools examine if the missingness is systematic by comparing the subsets of the collocated variables. If the two subsets are significantly different, the missingness may be systematic. Bivariate analysis on BDL data reveals if the occurrence of BDL spikes are dependent. Expected distribution is compared with the observed distribution using Kolmogorov–Smirnov (KS) test. The more different the two distributions are, the more dependent the BDL occurrence can be. Different transformations of data with spikes can lead to various cluster analysis results. Four suitable transformations are combined with the workflow to find clusters in real data. The inconsistency between multivariate and spatial continuity is addressed by a novel classification method classifies spatial data based on ensemble clustering results.

6.1 Contributions

Missing data come from multiple sources. Even though the missing data are in a few variables, they can cause data locations to be left out in techniques such as principal component analysis (PCA). Chapter 2 uses a data map and combines a permutation test with a KS test to examine the general information of missing data and systematic missingness (missing at random (MAR)) existence respectively. For the data map, it shows the number of missing data and the number of missing variables. Some variables and data locations may be dropped for a complete dataset. The optimal dataset retaining the most data after dropping is highlighted in the data map. For the combined statistical tools, comparing the subsets of non-missing variables conditioning to the missing locations in missing variables helps understand systematic missingness. Combined KS test and permutation test generates a universal measurement p for all variables as an indication of the systematic missingness. The relevance between variables and the missing data size are also taken into consideration for showing the level of systematic missingness.

BDL data form spikes in the data distribution, which causes problems when being quantile transformed. Chapter 3 conducts a univariate and bivariate analysis on BDL data. Univariate analysis provides information about the univariate spikes distributions. It generates a statistic table focusing on the characteristics of the BDL data. Different measurements examine the different types of spikiness in data. Bivariate method provides tools to examine the denpendence of BDL occurrence between variables. It compares the observed distribution with the expected distribution assum-

ing independence of BDL occurrences. The difference is measured using Kullback–Leibler (KL) test. The more different two distributions are, the more dependent of the BDL occurrence in two variables. The measurements of KL test are scaled by the theoretical maximum D, amending the problem of too few combinations of variables when using the absolute D. The scaled D finds 20 combinations of variables in which the BDL occurrences are dependent.

Cluster analysis on data with BDL spikes can be difficult. Different clustering methods can generate different results. Chapter 4 provides methods to handle BDL data in cluster analysis and an algorithm to visually validate the multivariate clustering results. They include a workflow for finding the optimal number of clusters, and comparing the performance of K-means and Gaussian mixture model (GMM) on synthetic data. The compatibilities of different transformations with the workflow and the clustering methods are inspected. Results from synthetic data show k-means is an appropriate clustering method to deal with data with large spikes. Suitable transformations inferred from synthetic data include linear transformation, univariate transformation with spikes spread and preserved, and Gaussian transformation with spikes spread. The appropriate combinations of clustering methods and transformations reveal the number of clusters in the real multivariate data. To validate the results, an algorithm is developed to find a 2D plane that can show the clusters, and the projected multivariate data on such a plane reveal 2 clusters in real data.

Cluster analysis only ensures the continuity in spatial data. The clustering labels are scattered on spatial data. There is a clear trade-off between the continuity between spatial and multivariate continuity. Chapter 5 provides a novel classification tool to find reasonable continuity in both multivaraite and spatial domains. It first examines the better performance of ensemble clustering, which gives more continuous multivariate clusters than individual clustering methods when the spatial continuities are similar. A novel classification method is developed to find optimal domains that ensure reasonable continuity in both multivariate and spatial spaces. Two hyper-parameters include the spatial weight and the number of domains. With higher spatial weight, the classified domains have higher spatial continuity. The proposed classification method also takes multiple clustering labels as inputs, which avoids the problem of choosing the best clustering results. The effect of spatial weight on geostatistic modeling is exmained through flow simulation. The results show in the demonstrated real dataset, high spatial weight generate more high values in modeling, and low spatial weight generates more low values in modeling.

6.2 Limitations and Future Work

Although the tools in this thesis cover multiple aspects of geostatistical EDA, there is future work. The multivariate analysis in BDL only considers the combinations of two variables. It is possible that more dependence of BDL occurrence can be revealed in higher dimensions. This brings up another topic worth investigation. The BDL data in each dimension divide that dimension into 2

parts. When the number of dimensions increase, the subspaces divided by BDL data increase exponentially. An efficient way to calculate the dependency of BDL occurrence in different subspaces can be of great interest. PCA is another common approach in EDA. BDL spikes can change the original variance in data, which causes problems in calculating eigenvalues. Also, the subspace divided by the BDL data can cause problems for calculating representative eigenvectors in PCA. Investigating the influences of BDL spikes in PCA and different transformations to restore correct PCA results can be the future work that helps with further analysis. The novel classification of clustering labels gives practitioners a tool to choose the desired local continuity, but the algorithm becomes unstable when the spatial weight is close to 1, because less local information is used. A method to classify spatial domains considering geological settings is very important. The spatial complexity of geological boundaries makes most of the clustering methods fail to group domains correctly. There can be two general approaches to tackle the problem, converting the geological boundaries into quantitative measurements that are compatible with modern clustering methods, or setting constraints on the proposed classification, which makes the full spatial continuity close to preset domain distributions. The preset domains could come from external geological knowledge.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

Abrevaya, J., & Donald, S. G. (2017). A gmm approach for dealing with missing data on regressors. *Review of Economics and Statistics*, *99*(4), 657–662.

Abubaker, M., & Ashour, W. M. (2013). Efficient data clustering algorithms: improvements over kmeans. *International Journal of Intelligent Systems and Applications*, *5*(3).

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139–160.

Aranganayagi, S., & Thangavel, K. (2007). Clustering categorical data using silhouette coefficient as a relocating measure. In *International conference on computational intelligence and multimedia applications (iccima 2007)* (Vol. 2, pp. 13–17).

Barnett, R. M., & Deutsch, C. V. (2015). Multivariate imputation of unequally sampled geological variables. *Mathematical Geosciences*, *47*(7), 791–817.

Barnett, R. M., Manchuk, J. G., & Deutsch, C. V. (2014). Projection pursuit multivariate transform. *Mathematical Geosciences*, *46*(3), 337–359.

Barnett, R. M., Manchuk, J. G., Deutsch, C. V., et al. (2016). The projection-pursuit multivariate transform for improved continuous variable modeling. *SPE Journal*, *21*(06), 2–010.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131.

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, *44*(1), 108–132.

Chawla, S., & Gionis, A. (2013). k-means–: A unified approach to clustering and outlier detection. In *Proceedings of the 2013 siam international conference on data mining* (pp. 189–197).

d'Alfonso, A., Freitag, B., Klenov, D., & Allen, L. (2010). Atomic-resolution chemical mapping using energy-dispersive x-ray spectroscopy. *Physical Review B*, *81*(10), 100101.

Ding, Q., Han, J., Zhao, X., & Chen, Y. (2015). Missing-data classification with the extended full-dimensional gaussian mixture model: applications to emg-based motion recognition. *IEEE Transactions on Industrial Electronics*, *62*(8), 4994–5005.

Doyen, P., Den Boer, L., Pillet, W., et al. (1996). Seismic porosity mapping in the ekofisk field using a new form of collocated cokriging. In *Spe annual technical conference and exhibition.*

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, *89*(426), 463–475.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Falck, H., Day, S., Pierce, K., Rentmeister, K., Ozyer, C., & Watson, D. (2012). A compilation of heavy

mineral concentrates: results from stream sediment samples collected 2007-2010, mackenzie mountains, nwt. *Northwest Territories Geoscience Office, NWT Open Report*, *1*.

Filzmoser, P., Garrett, R. G., & Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & geosciences*, *31*(5), 579–587.

Fred, A. L., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, *27*(6), 835–850.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Kasim, R. M., & Raudenbush, S. W. (1998). Application of gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, *23*(2), 93–116.

Krishna, K., & Murty, M. N. (1999). Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*(3), 433–439.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Lachheb, I. (2021, July). *lachhebo/pyclustertend: 1.6.0.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.5076249` doi: 10.5281/zenodo.5076249

Lawson, R. G., & Jurs, P. C. (1990). New index for clustering tendency and its application to chemical problems. *Journal of chemical information and computer sciences*, *30*(1), 36–41.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Lücke, J., & Forster, D. (2019). k-means as a variational em approximation of gaussian mixture models. *Pattern Recognition Letters*, *125*, 349–356.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

Martin, R. (2019). Data driven decisions of stationarity for improved numerical modeling in geological environments.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition* (Vol. 544). John Wiley & Sons.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 38). M. Dekker New York.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, *31*(3), 274–295.

Odén, A., Wedel, H., et al. (1975). Arguments for fisher's permutation test. *The Annals of Statistics*, *3*(2), 518–520.

Palarea-Albaladejo, J., & Martin-Fernandez, J. (2013). Values below detection limit in compositional chemical data. *Analytica chimica acta*, *764*, 32–43.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prades, C. F. (2017). *Geostatistics and clustering for geochemical data analysis.* University of Alberta, Edmonton, Canada.

Pyrcz, M., & Deutsch, C. (2018). Transforming data to a gaussian distribution. *Geostatistics Lessons.*. Retrieved from `http://geostatisticslessons.com/lessons/normalscore`

Pyrcz, M. J., & Deutsch, C. V. (2014). *Geostatistical reservoir modeling*. Oxford university press.

Reimann, C. (2005). Geochemical mapping: technique or art? *Geochemistry: Exploration, Environment, Analysis*, *5*(4), 359–370.

Reimann, C., Filzmoser, P., & Garrett, R. G. (2005). Background and threshold: critical comparison of methods of determination. *Science of the total environment*, *346*(1-3), 1–16.

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*.

Richter, S., & Goldberg, S. (2003). Improved techniques for high accuracy isotope ratio measurements of nuclear materials using thermal ionization mass spectrometry. *International Journal of Mass Spectrometry*, *229*(3), 181–197.

Romesburg, C. (2004). *Cluster analysis for researchers*. Lulu. com.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *1*(1), 73–79.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Sander, J., Qin, X., Lu, Z., Niu, N., & Kovarsky, A. (2003). Automatic extraction of clusters from hierarchical clustering representations. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 75–87).

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, *5*(1), 3–55.

Shen, J., Hao, X., Liang, Z., Liu, Y., Wang, W., & Shao, L. (2016). Real-time superpixel segmentation by dbscan clustering algorithm. *IEEE transactions on image processing*, *25*(12), 5933–5942.

Silva, D. S., & Deutsch, C. V. (2018). Multivariate data imputation using gaussian mixture models. *Spatial statistics*, *27*, 74–90.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling.

Journal of chemical information and computer sciences, *43*(6), 1947–1958.

Takeuchi, I., Le, Q., Sears, T., Smola, A., et al. (2006). Nonparametric quantile estimation.

Taylor, C. R. (1993). Dynamic programming and the curses of dimensionality. *Applications of Dynamic Programming to Agricultural Decision Problems.*, 1–10.

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

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511–528.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tukey, J. W., et al. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Vargas-Guzmán, J. A., & Dimitrakopoulos, R. (2003). Computational properties of min/max autocorrelation factors. *Computers & Geosciences*, *29*(6), 715–723.

Verly, G. (1984). *Estimation of spatial point and block distributions: the multigaussian model* (Unpublished doctoral dissertation). Stanford University.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Yamazaki, K., & Watanabe, S. (2003). Singularities in mixture models and upper bounds of stochastic complexity. *Neural networks*, *16*(7), 1029–1038.

Young, I. T. (1977). Proof without prejudice: use of the kolmogorov-smirnov test for the analysis of histograms from flow systems and other sources. *Journal of Histochemistry & Cytochemistry*, *25*(7), 935–941.