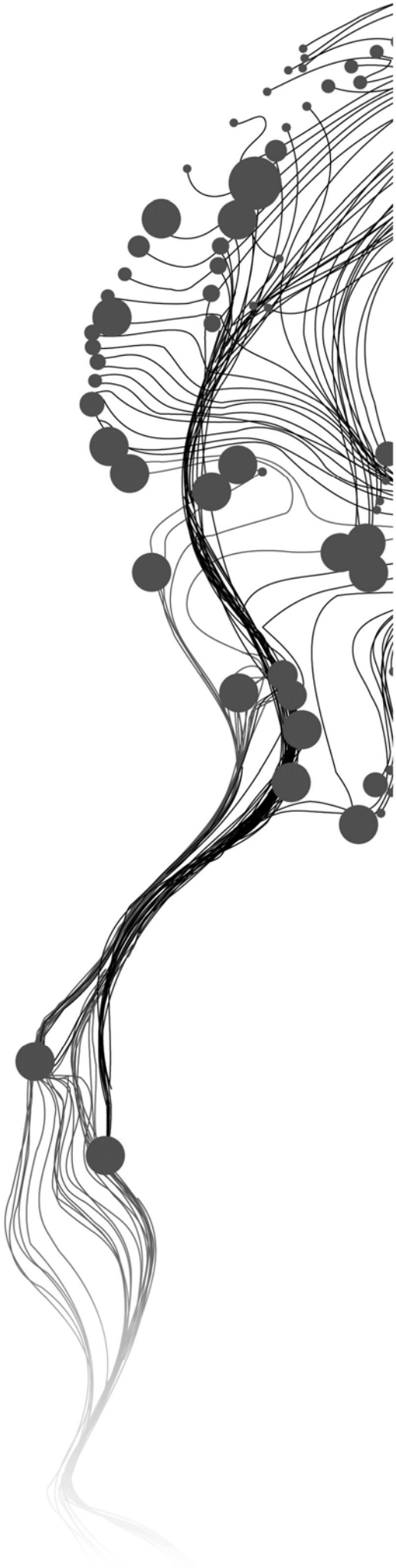
Integration of Stream Sediment Geochemical and Airborne Gamma-ray Data for Surficial Lithologic Mapping using Clustering Methods

HUSIN SETIA NUGRAHA

March, 2011

SUPERVISORS:

Dr. E.J. M. Carranza
Dr. M. van der Meijde



Integration of Stream Sediment Geochemical and Airborne Gamma-ray Data for Surficial Lithologic Mapping using Clustering Methods

HUSIN SETIA NUGRAHA
Enschede, The Netherlands, March, 2011

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Applied Earth Sciences

SUPERVISORS:
Dr. E.J. M. Carranza
Dr. M. van der Meijde

THESIS ASSESSMENT BOARD:
Prof. Dr. F.D. van der Meer (Chair)
Dr. D.G. Rossiter (External Examiner, ITC)

ABSTRACT

In surficial lithologic mapping, geologists use remotely sensed data prior to fieldwork, however, the utility of these datasets are limited due to vegetation cover. Thus, the use of other sources of information about chemical and physical properties of rocks such as geochemical data (e.g., from stream sediment samples) and airborne geophysical data (e.g., radiometric data) becomes important. In this study, two clustering algorithms, partition around medoids (PAM) and model-based clustering (Mclust) were performed in stream sediment geochemical (SSG) and airborne-gamma-ray (AGR) data as well as in SSG and AGR together to map surficial lithologies in vegetation-covered areas in Central Part of British Columbia Province-Canada. Prior to clustering two approaches, conventional and compositional (CoDa), were applied to SSG and AGR data in order to study the influences of closure problems within the data. In SSG data analysis, clustering was applied using all 13 elements and selected nine elements. In addition, two types of data integration was done SSG all element and AGR (Reference Data I); and SSG selected element and AGR (Reference Data II). Overall accuracy and kappa coefficient was computed for the results and, two references were used to assess accuracy of the classification which is simplified existing lithological map (Reference Data I) and the lithological map based on the interpretation of airborne magnetic data (Reference Data II).

The study results reveal that Mclust and PAM clustering could be alternatives techniques for classifying stream sediment geochemical data to help lithological mapping in an area with limited information. The images of their results depict pattern similarities to the existing litholigical map. In addition, the assessments of the results show moderate accuracy up to 51% and 0.41 for overall accuracy and kappa coefficient, respectively. Furthermore, for a large homogeneous lithology, the producer's accuracy is quite high up to 80%. Moreover, in general, the application of CoDa approach in data preparation to both SSG and AGR data do not produce better accuracy than conventional approach. The assessments show the differences between the application of CoDa and conventional approach reach up to 10% and 0.1 for overall accuracy and kappa coefficient, respectively.

The integration of SSG and AGR data produces better results than those using both SSG and AGR data separately. The percentage accuracies of integration data compare to their separated data increase quite significant up to 17% and 0.15 for overall accuracy and kappa coefficient, respectively. In addition, Mclust produces better classifications for lithilogical mapping relatively to PAM clustering base on both qualitative and quantitative assessments. Qualitatively, from visual evaluation, the patterns of Mclust results are more similar to lithological patterns in the existing lithological map than PAM clustering. Quantitatively, assessments results in each separated data (SSG or AGR) show up to 5% and 0.7 differences for overall accuracy and kappa coefficient, respectively, whereas for the integrated data (SSG and AGR) produces non-significant difference results (1% and 0% differences for overall accuracy and kappa coefficient, respectively). Therefore, Mclust could be applied to integrate and classify SSG and AGR data for lithological mapping in regional scale.

Keywords: *closure problems, compositional data (CoDa), partition around medoids (PAM), model-based clustering (Mclust), airborne magnetic, British Columbia Canada*

ACKNOWLEDGEMENTS

Firstly, I am indeed grateful to Allah SWT, the Almighty God, for the sustenance, strength and the successful completion of my MSc program. I would like to thanks to the Dutch Government and NESO Indonesia for the scholarship (STUNED) and also many thank to the Director of Geothermal, Sugiharto Harsoprayitno for giving me a permission to pursue the study.

My profound gratitude goes to my supervisors, Dr. E.J.M. Carranza and Dr. M. van der Meijde for their technical guidance, constructive comments, critical reading and directions. I want to thank the Applied Earth Science (AES) Course Director, Drs T.M. Loran and the Chairman of the department of Earth System Analysis, Prof. Dr. F.D.Van der Meer for their efforts in ensuring a successful running of the MSc program. My thanks are also for Dr. F.J.A. van Ruitenbeek, Drs. J.B. de Smeth for their company and valuable instructions during my study in Earth Resource Exploration stream.

Undoubtedly, my unreserved gratitude goes to my lovely wife, Rosmayanti Haerani and my handsome son, Muhammad Fathir Putra Nugraha as well as my parents, Engkos Kosasih and Euis Warni, for their patience and unbroken link during my MSc Program. I cannot but appreciate my colleagues, Abigail June Agus, Maruvoko Elisamia Msechu, Woinshet Taye Tessema and Engdawork Admassu Bahru for the discussion and checking my drafts. Last but not least, many thanks to the Indonesian in AES department; Novita Hendrastuti, Rana Wiratama and Syams Nashrullah Suprijatna for sharing the tears and the joyful.

To all my instructors and colleagues in the department of Applied Earth Science, I say thank you.

Husin Setia Nugraha

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

A lithological map provides both bedrock and surface geology information, which is important in many disciplines for various purposes such as natural resources exploration, geohazard management and city planning. A lithological map is one of the most crucial information in order to discover natural resources, e.g., hydrocarbon (oil, natural gas and coal), mineral and groundwater. In geohazard management, lithological maps are becoming important inputs in modelling of geohazards such as landslides and flooding to determine risk and safe zones. In civil engineering, a lithological map has great importance in many activities, e.g., excavation of road cuts. For city planners, it is an advantage to have a lithological map in order to determine settlement areas (Lisle, 2004). In mineral exploration, lithological maps with both bedrock and surficial geology information are used to understand geological process such as mineralization to support exploration activities (Smith, 1996).

Besides the conventional method of field work to map lithology, analyses of remote sensing data such as spaceborne spectral imagery and airborne geophysics have become important methods of lithological mapping because of their advantage to cover large and inaccessible areas compared to fieldwork. The capability of remote sensing data to provide synoptic views of large areas is great importance in lithological mapping at regional to district scales because they allow geologists to obtain lithological information even before going on fieldwork. Remote sensing data have also been used widely to update existing lithological maps because they provide lithological information in unvisited places that significantly outnumber the places that can be visited during fieldwork. However, integration of fieldwork and remote sensing data is even more important in lithological mapping.

1.2. Previous works

Lithological mapping, in areas of various climates, has been approach in various ways. In arid and semi-arid areas, lithological mapping has made use of optical remote sensing data, e.g., Landsat TM (Alberti et al., 1993; An et al., 1995) and ASTER (Gomez et al., 2005; Ninomiya et al., 2005; Rowan and Mars, 2003). Other advanced optical remote sensing data such as airborne hyperspectral images are now also being employed for lithological mapping (Bedini, 2009; Rowan et al., 2004). In tropical areas, where cloud cover and vegetation significantly hinder spectral remote sensing, airborne geophysical data are more useful then multispectral or hyperspectral data for lithological mapping (An et al., 1995; Graham and Bonham-Carter, 1993; Martelet et al., 2006).

In terms of fieldwork data, geochemical data from various sampling data, aside from lithological observations have been exploited to recognize lithologies using various statistical techniques. For example, Kerr and Davenport (1990) used composite variables derived from multivariate analysis of multi-element lake sediment and water geochemical data to reveal spatial patterns related to bedrock geology in Labrador-Canada. Shepherd et al. (1987) applied a non-hierarchical ***k***-means clustering technique to soil geochemical data to reveal subtle spatial patterns that were useful in lithological mapping of the very poorly exposed basic-ultrabasic Lizard Complex South-West England. Cocker (1999) analyzed alkali elements in stream sediment samples to assist regional lithologic mapping in Georgia. Stendal (1978) and (Bellehumeur et al., 1994) have demonstrated the usefulness of heavy minerals in stream sediments to assist mapping of bedrock geology. Rantitsch (2000) demonstrated the application of fuzzy-c clustering of stream sediment geochemical data to separate four different lithologies in a geologically complex area of the Eastern Alps (Austria). Recently, Ranasinghe et al. (2009) have shown the capability of stream sediment geochemical data for describing upstream regional and local-scale lithological changes in complex high-grade metamorphic terrains in Sri Langka.

1.3. Research problems

Surficial lithologic mapping in vegetation-covered areas is not simple task for geologists. The situation is worse when, in those areas, only limited outcrops of rocks exist. For lithological mapping in those areas, geologists usually have to derive optimum prior information from available remote sensing data before going on fieldwork. Nevertheless, the use of satellite spectral images will be limited because of vegetation cover. Therefore, in addition to field data, the use of surficial geochemical data (e.g., from stream sediment samples) and airborne geophysical data (e.g., radiometric data) becomes important sources of information about the chemical and physical properties in those areas.

Stream sediment and airborne gamma-ray data contain geochemical properties; thus, it will be advantageous to integrate information from these data sets. Stream sediment data are point data with irregular pattern of sample locations and non-uniform sampling density because the samples are taken by following rivers. Stream sediment data usually contain concentrations of many elements. In contrast, airborne gamma-ray data contain concentrations of only three elements but these data have regular sampling pattern and uniform sampling density. Consequently, when these two types of data are integrated, the strength of one data type compensates the weakness of the other. For example, the multiple elements in stream sediment data compensate for the only three elements in airborne gamma-ray data. In addition, the high sampling density of airborne gamma-ray would result in integrated data with higher spatial resolution than the stream sediment data.

However, a problem that arises when integrating airborne gamma-ray and stream sediment data is in representing point data of stream sediment into continuous data because stream sediment samples represent only materials within catchment basins of every sampling site. Some authors tried to find appropriate technique for representing stream sediment geochemical data. Bonham-Carter et al. (1987); Carranza and Hale (1997) and (Spadoni et al., 2004) applied catchment basin approach to represent stream sediment data. This approach considers that stream sediment samples represent several sources and processes within catchment basins of every sampling site. The sources and processes include minerals of bedrock, minerals formed during weathering, minerals typical of mineralization, and anthropogenic substances (Howarth, 1984; Naseem et al., 2002). Robinson et al. (2004) demonstrated the use of inverse distance weighting (IDW) and kriging interpolation in order to observe regional-scale spatial variation of stream sediment and water geochemical data in New England (USA). Recently, Carranza (2010) explained that representing stream sediment geochemical data as discrete or continuous landscapes depend on mapping scale. For regional scale (e.g., 1:100.000 or smaller), representing stream sediment geochemical data as continuous landscapes by interpolation technique is plausible because its purpose to delineate anomalous areas for further investigations at higher scales could be achieved, whereas representing the data as discrete landscapes such as sample catchment basins could be both tedious and impractical.

Other problems that might rise in using geochemical data such as from stream sediments or airborne gamma ray data are related to "closure" that is inherent in compositional data such concentration of elements. Compositional data are characterized by its relative contained information because the data are ratio values (e.g., expressed as ppm, %, etc.) but not absolute values. Other characteristics of compositional data are that they always have positive values and the sums of the element data per sample are constrained to a constant value ($\boldsymbol{k}$) such as 100 wt% or 1,000,000 ppm. Therefore, compositional data always have limited range between 0 and $\boldsymbol{k}$ (Pawlowsky-Glahn and Egozcue, 2006). One of the problems, which might is caused by this closure property of geochemical data, is the skewed data distribution which means not following a normal distribution. Direct application of statistical techniques to the non-normally distributed data could produce improper results because many statistics techniques rely on the assumption of normal data distribution. Moreover, data transformation such as logarithmic transformation is a common technique in order to solve this problem. However, according to Filzmoser et al. (2009a), conventional data transformations such as logarithmic transformation do not solve problems associated with closure property of compositional data. Furthermore, Pawlowsky-Glahn and Egozcue (2006) explained that closure-related problems also produce less or no significant information in geologic sense when multivariate techniques such as principal components analysis are applied. Other problem associated with closure is untrue correlation among compositional variables, which is caused by the ratio values that are contained to a constant sum in compositional data.

1.4. Research objectives

The main objective of the research is to map surficial lithologies in vegetation-covered areas to assist field work preparation for lithological mapping by integrating stream sediment geochemical data and airborne gamma-ray data in regional scale. The following sub-objectives are composed in order to achieve the main objective:

- To quantify the significance of compositional data approach application in stream sediment geochemical and gamma ray data for surficial lithologic mapping;
- To perform clustering methods in stream sediment geochemical and airborne gamma-ray data for mapping the lithologies;
- To perform clustering methods for integrating stream sediment geochemical and airborne gamma-ray data as applied to surficial lithologic mapping.

The present research used clustering methods in order to integrate stream sediment geochemical data and airborne gamma-ray data. These methods were chosen because they are unsupervised and, thus, are appropriate in areas where no or little a-priori information about the objects to be mapped is available. Moreover, clustering methods are independent of grid size and, thus, are more robust to the influence of significant spatial resolution differences such as between stream sediment and airborne gamma-ray data. The two clustering algorithms used in this research are Model-based clustering (Mclust) and Partition Around Medoids (PAM), representing respectively model-based and distance-based clustering. In distance-based clustering, cluster members are determined by calculating the distances between the samples. In model-based clustering, clusters are determined by selecting an appropriate model for the data. Furthermore, both of those clustering techniques were chosen because their algorithms are robust to existence of outliers in data (Gan et al., 2007; Kaufman and Rousseeuw, 2005). These two methods were also applied to stream sediment geochemical data and airborne gamma-ray data in order to investigate the difference in the performance with respect to these two types of data.

1.5. Research questions

The research attempted to answer the following questions:

- Could stream sediment geochemical data be used to assist lithological mapping in vegetation-covered areas where no or little a-priori information about underlying rock units is available?
- Does the application of compositional data (CoDa) analysis to stream sediment geochemical data and airborne gamma-ray data produce better results than conventional methods?
- Is clustering results using integrated data of stream sediment geochemistry and airborne gamma-ray produce better results than those using stream sediment geochemical or airborne gamma-ray data separately?
- Which clustering technique – Mclust or PAM – gives better result for surficial lithologic mapping based on surficial geochemical datasets?

1.6. Study area

1.6.1. Location

The study area is situated at regional district of Bulkley-Nechako in Northern-Central of British Columbia province (*figure 1-1*). The area was chosen due to its characteristics and data availability that appropriate with the objectives of the research such as vegetation-covered areas (Delong, 1996). Regarding to data availability, besides input data such as stream sediment geochemical and airborne gamma-ray data, reliable geologic map for validation is also available. In addition, the dominant landform of this regional district is the Nechako plateau. The areas consist of Bulkley Valley, the northern part of the Nechako District, and the Omineca District, including portions of the Hazelton Mountains and Omineca Mountains in the west and north of the regional district, respectively (http://en.wikipedia.org/wiki/Regional_District_of_Bulkley-Nechako). The study area bounded by geographic coordinates (372750 mW, 6095500 mN) and (423500 mW, 6134500 mN) and covers an area of ~ 2,000 km^2.

1.6.2. Geology

The area is dominantly underlain by the Quesnel Terrane or Quesnelia. Two groups of rocks form this terrane, the Takla Group at the northern part and the Nicola Group at the southern part. The terrane is intruded by the northwest-elongate Hogem batholiths. The Takla Group consists of sedimentary units of Late Triassic in age. This group is overlain by volcanic, pyroclastic, and epiclastic rocks; and intruded by early a Jurassic pluton. Augite phyric rocks are dominant with plagioclase and hornblende (Nelson et al., 1992; Nelson, 1991). Takla Group volcanics are unusually K-rich and alkalic (Delong, 1996).

Nelson (1991) divided the Takla Group into four interfingering formations, the Rainbow Creek, Inzana Lake, Witch Lake and Chuchi Lake Formations. In stratigraphy, Rainbow Creek is the lowest unit overlain by the Inzana Lake, Witch Lake and Chuci Lake Formation, in upward sequence. The Rainbow Creek Formation is comprised of dark grey to black slates or phyllites with interbeded quartz-rich siltstone and sandstone. The Inzana Lake Formation consists of epiclastic and sedimentary rocks with minor pyroclastic rocks. The Witch Lake Formation is dominated by an augite porphyry suite, which was produced from explosive intermediate volcanism. The Chuchi Lake Formation is made up of volcanic rocks with andesitic to latite-andesite composition. The phenocryst assemblage of these volcanic rocks is dominated by plagioclase with variable amounts of augite and hornblende (Delong, 1996; Nelson, 1991).

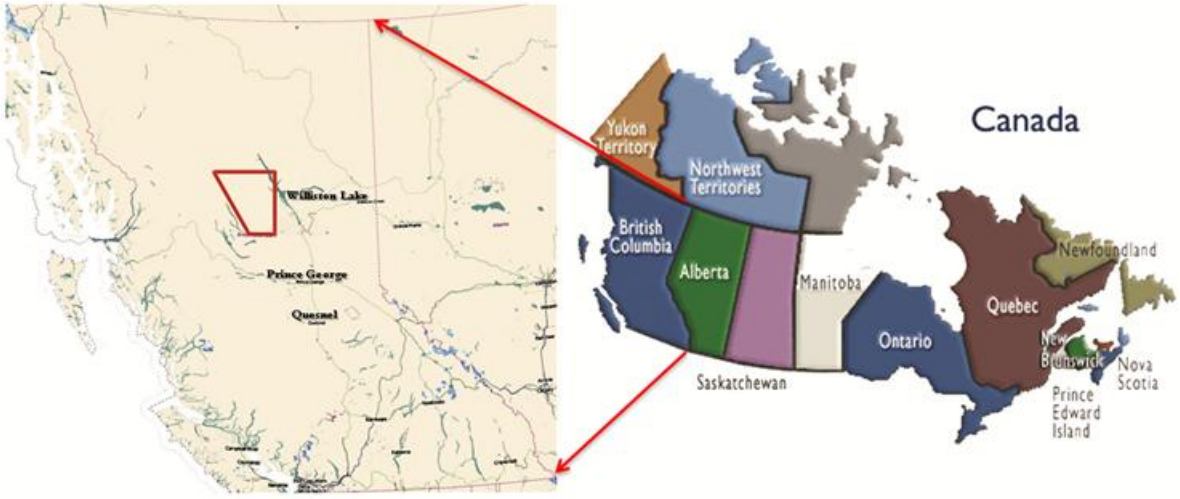


Figure 1-1 Location of study area (in red polgyon) in the northern central part of the British Colombia Province

Figure 1-2 is simplified lithologic map, which was used for validation of results of this study. The geologic map from Massey et al. (2005a, b, c,d) were simplified base on regional geologic map from Nelson (1991). The lithological units were divided according to age group. For sedimentary rocks, the lithological units were divided into four lithological units based on age (from Proterozoic to Quaternary). Small lithological units such as ultramafic rocks and metamorphic rocks were merged with the larger lithological unit wherein they lie. Intrusive rocks comprise two formations, which are the Chuchi Syenite and Klawli Pluton Formations. Two formations, the Chuchi Lake Succession Formation and Witch Lake Formation, comprise volcanic rocks. Therefore, there are six lithological units for validation which are Intrusive Rocks, Volcanic Rocks and four sedimentary rocks units. The four sedimentary rocks units are Sedimentary Rocks 1 which comprises of sedimentary rocks from Jurassic to Quaternary, Sedimentary Rocks 2 that is contained sedimentary rocks from Triassic to Jurassic and small parts of metamorphic rocks, Sedimentary Rocks 3 which are constituted by sedimentary rock from Ordovician to Jurassic and Sedimentary Rocks 4 that comprises sedimentary rock from Proterozoic to Ordovician and small parts metamorphic rocks. The subset maps from the map shown in *figure 1-2* that were used for validation of the results of the study are shown in *appendix 1-1*.

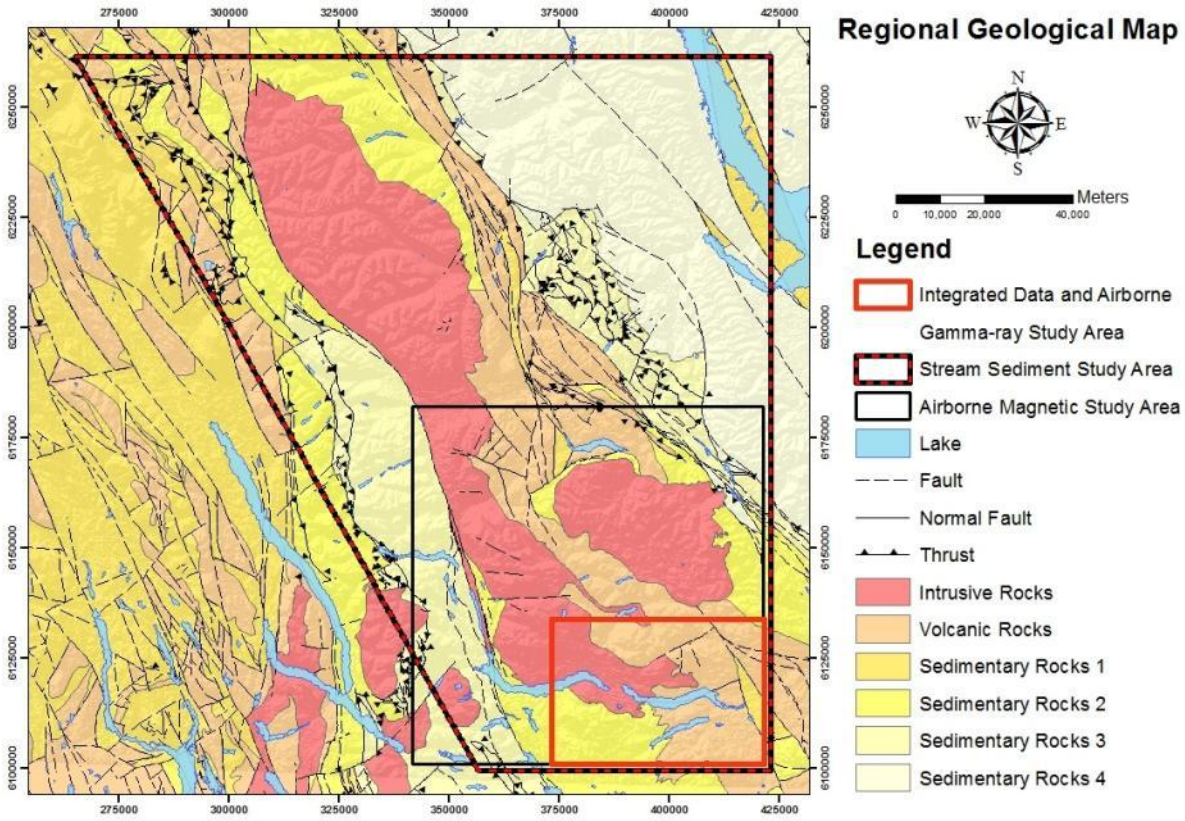


Figure 1-2 Simplified geological map used for validation of results (modified from Massey et al., 2005a, b, c, d)

The study area lies between two regional-scale northwest-trending fault systems (*figure 1-3*). The Pinchi Fault system lies at the western part and the Manson-McLeod Faults at eastern part belong to the Northern Rocky Mountain Trench Fault system (Nelson et al., 1992; Nelson, 1991).

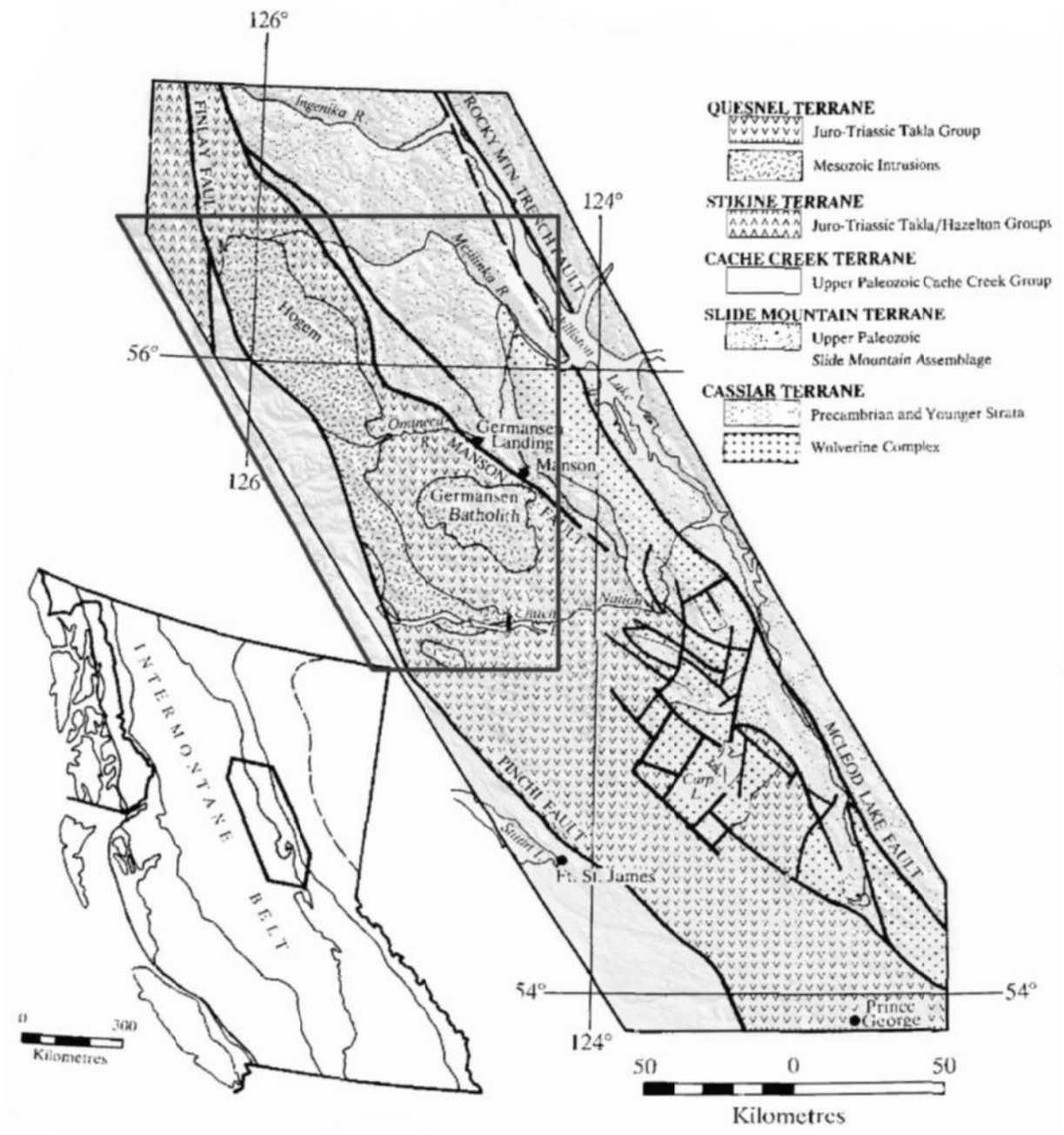


Figure 1-3 Regional structures in the study area (Nelson et al., 1992; Nelson, 1991)

1.7. Thesis Outline

This thesis consists of six chapters, which are the Introduction, Stream Sediment Geochemical Data Analysis, Airborne Gamma-ray Data Analysis, Airborne Magnetic Data Analysis, Integrated Analysis of Stream Sediment Geochemical and Airborne Gamma-ray Datasets, Conclusions and Recommendations.

2. STREAM SEDIMENT GEOCHEMICAL DATA ANALYSIS

2.1. Introduction

Analyzing stream sediment geochemical data in this chapter aims to achieve two purposes. The first is performing unsupervised classification, in this case using clustering methods, to stream sediment geochemical data in order to help lithological mapping. This analysis is based on the assumption that in the area there is little or no a-priori information about the underlying rocks. Two clustering algorithms, Model-based clustering (Mclust) and Partition Around Medoids (PAM), were applied. The second aim is to investigate the application of compositional data (CoDa) approach to the data.

2.2. Description of stream sediment geochemical datasets

Stream sediment geochemical data used here were collected by Geological Survey of British Columbia during a National Geochemical Reconnaissance Program of Canada (NGR) that began in 1975. The data were sampled from the first and/or second order streams, producing average density about a sample per 13 km2. The samples were taken from active part of stream channel with two-thirds of the sample paper bag was filled with silt or fine sand. In the laboratory, the samples were air dried at temperature below 40°C and sieved using a minus 80-mesh (177 μm) screen. The samples were analyzed for base and precious metals, pathfinder elements and rare earth elements by instrumental neutron activation analysis (INAA) and inductively coupled plasma mass spectrometry (ICP-MS). For quality control, control reference and blind duplicate samples were inserted into each block of twenty stream sediment samples (Jackaman and Balfour, 2008).

The stream sediment geochemical data were downloaded from the Geoscience Data Repository of Natural Resources Canada website (http://gdrdap.agg.nrcan.gc.ca/geodap/home/Default.aspx?lang=e), then subset to the research area. The data consist of sixteen elements (Zn, Cu, Pb, Ni, Co, Ag, Mn, Fe, Mo, Hg, Sb, As, Ba, Ce, Cr and Rb) with 2,478 sampling points including 284 duplicate samples. The concentrations of elements were measured in ppm, except for Fe in percentage whereas Ag and Hg were measured in ppb. Duplicate samples were used to analyze data quality and were excluded from statistical analysis.

2.3. Methodology

Methodology is divided into five stages which are, data quality assessment, data preparation, clustering, post-clustering and clustering assessment (*figure 2-1*). Two techniques were applied to assess the quality of the stream sediment data in order to select elements to be used in the next analysis. In data preparation, two approaches were performed, conventional and CoDa approach. In conventional approach, some data processes were applied as suggested by Reimann et al. (2008) before applying multivariate analysis for getting comparability (equality) of the variance whereas the aim of the processes in the CoDa approach is to transform the data into appropriate feature space. It is because when the data are treated as CoDa, they lie on different feature space thus certain transformations are needed before applying multivariate analysis (Filzmoser et al., 2009a, b). The feature space for CoDa, so-called *Simplex* (***S***), accommodates all CoDa characteristics as explained in *section 1.3*. Furthermore, two clustering algorithms were used to classify stream sediment geochemical data. Some post-clustering processes such as rasterizing, reclassification and filtering were conducted before the calculation of overall accuracy and kappa coefficient.

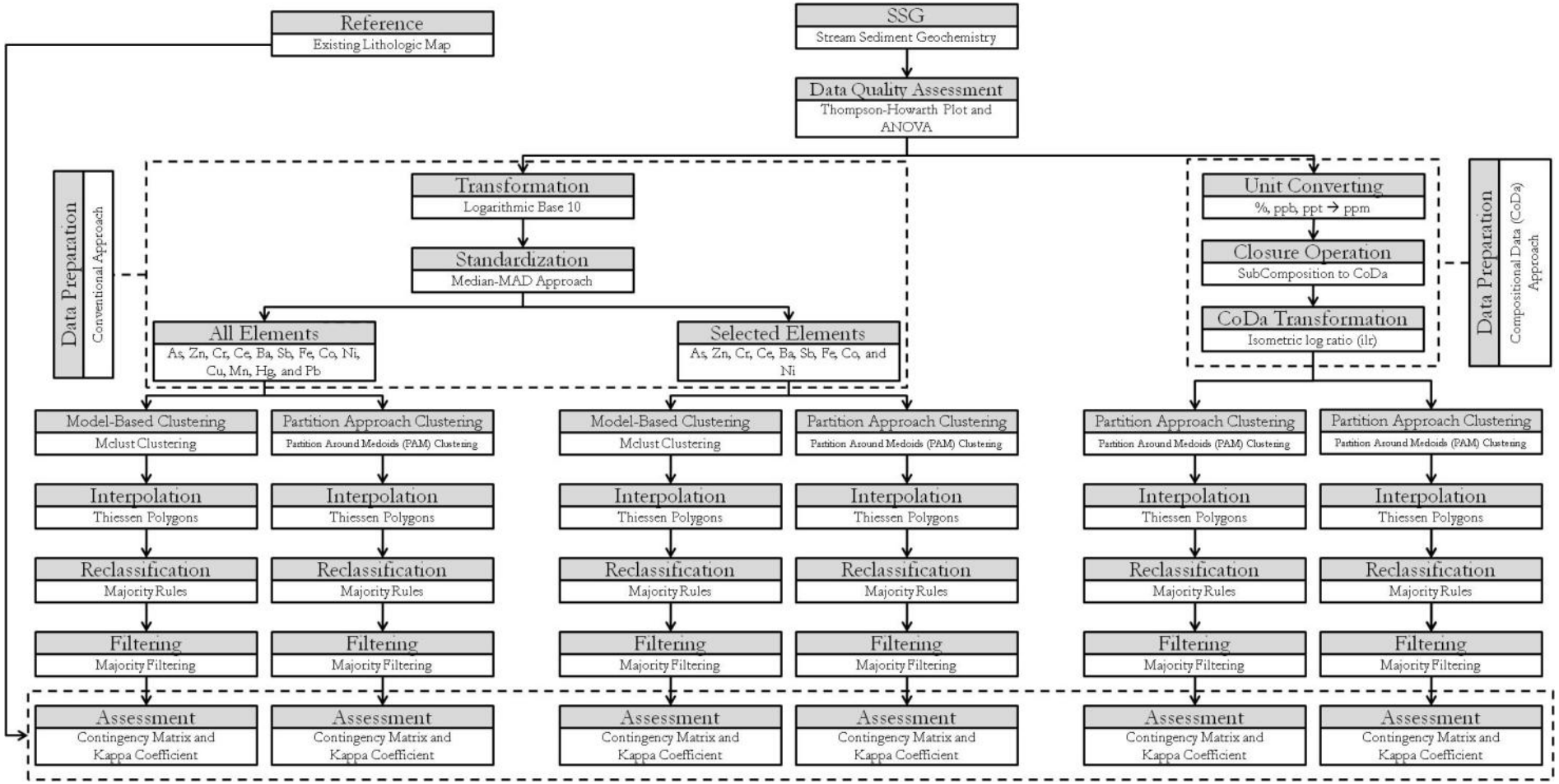


Figure 2-1 Flow chart of methodology to map lithology using stream sediment geochemical data by applying two types of clustering algorithm, Model-based clustering (Mclust) and Partition Around Medoids (PAM) clustering with two approaches in the data preparation stage, conventional and compositional data (CoDa).

2.3.1. Data quality assessments

The quality of the stream sediment geochemical data was assessed by two methods, Thompson-Howarth plot and analysis of variance (ANOVA). The *rgr* package of *R* was used for making the Thompson-Howarth graph and for ANOVA calculation (Garrett, 2010).

2.3.1.1. Precision analysis

Precision is the degree of closeness between test results obtained under certain standards. It denotes a distribution of random errors (Reimann et al., 2008). Thompson (1983) defined precision (*P*) as:

$$P = \frac{2S_c}{c} \times 100\% \qquad \text{Equation 2-1}$$

where S_c and C are, respectively the standard deviation and the mean of data from duplicate samples. The precision of duplicate samples can be analyzed graphically using the Thompson-Howart plot. The data are plotted by first calculating the means and absolute differences of duplicate analyses. The absolute differences are plotted as a function of the mean concentrations (Howarth, 1983).

The model of Thompson-Howarth plot is used to test the quality of the data. Data input for the model consist of relative standard deviation (RSD) and the percentile value. RSD of the population is percentage of the ratio of standard deviation to mean value. The value of 5% of RSD is equivalent to 10% precision at two times of standard deviation in *equation 2-1*. The percentile line is associated with a half-normal distribution. For example, the 95th-percentile line means that there are 5% of odd values in the population.

2.3.1.2. Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) was employed to assess data quality in term of data precision. ANOVA test can be used to explain data variances in a data population. Variance in data population is defined as the average of the square of a measure of deviation of the values to their mean (*equation 2-2*).

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \qquad \text{Equation 2-2}$$

where σ is the variance, μ is the mean, X_i is individual measurements of element concentrations and n is the number of duplicate samples (Swan and Sandilands, 1995). Deviation of the values from their mean is caused by random variations in the sample population; procedural errors, and inhomogeneity of the samples. However, the variances due to procedural errors and heteogeneity of the samples are not easy to separate, thus total variance could be written as the sum of sample variance (σ_g^2), and procedural variance (σ_p^2).

$$\sigma_t^2 = \sigma_p^2 + \sigma_g^2 \qquad \text{Equation 2-3}$$

Since duplicate samples are a subset of the sample population, the above equation can be written as:

$$MS_t = MS_p + MS_g \qquad \text{Equation 2-4}$$

where *MS* is the mean of squares, which is calculated using the following formula:

$$SS_t = \sum_{1}^{n} X^2 - \frac{\left(\sum_1^n X\right)^2}{n} \qquad \text{Equation 2-5}$$

$$SS_g = \frac{\sum_i^i \left(\sum_1^j X\right)^2}{j} - \frac{\left(\sum_1^n X\right)^2}{n} \qquad \text{Equation 2-6}$$

$$SS_p = SS_t - SS_g \quad \text{Equation 2-7}$$

$$MS_g = \frac{SS_g}{(i-1)} \quad \text{Equation 2-8}$$

$$MS_p = \frac{SS_p}{i(j-1)} \quad \text{Equation 2-9}$$

where:
SS_t = total sum of squares
SS_g = sum of squares due to geochemical variance
SS_p = sum of squares due to procedural variance
MS_g = mean of squares of geochemical variance
MS_p = mean of squares of procedural variance
X = individual measurements of element concentrations
i = number of samples with duplicate measurements
j = number of measurement within each group
$n = ij$ = total number of measurements

A F-test is used to estimate the significance of the geochemical data variance. A F-value, calculated from a ratio of MS_g to MS_p, is compared with a critical F-value at certain significance level for particular number of degrees of freedom. If the F-value is less than the critical F_c, it means that procedural error is too large such that the measurements obtained are not sufficient to denote geochemical variations within an area. Therefore, in order to get a clear pattern in natural geochemical variance, the procedural error, MS_p should not exceed 20% of the total variance (Ramsey et al., 1992).

2.3.2. Data preparation

2.3.2.1. Missing data imputation

Sample points with missing values of element concentrations need pre-treatment before any statistics method could be applied. In multivariate analysis, samples with missing information could not simply be removed as in univariate analysis. Removing sample points will cause loss of available measurements for analysis. Thus, filling in the missing values with appropriate values in sample points or imputing the data is becoming an alternative way than just deleting the sample points.

A ***k***-nearest neighbours (***k***-nn) imputation method was applied to fill up missing values for several sample points in the stream sediment geochemical dataset. The ***k***-nn imputation method use Euclidean distance to find ***k*** number of the nearest neighbours points of the element containing a missing value among observation elements and to replace missing value by using available element variable information of the neighbours (Hron et al., 2010; Troyanskaya et al., 2001). Troyanskaya et al. (2001) suggested applying logarithmic transformation to the data before applying the method to overcome sensitivity to outliers because of the use of Euclidean distance. In CoDa, the same principal of ***k***-nn imputation is also applied. However, CoDa have its own feature space, *Simplex* (***S***), thus Aitchinson distance is preferred to be used than Euclidean distance due to differences of feature space (Hron et al., 2010). The *impute* package in *R* from Hastie et al. (2010) was employed to impute the missing values for conventional approach whereas *robComposition* package for compositional data (Templ et al., 2010).

2.3.2.2. Conventional approach

Univariate statistical analysis

Univariate statistical analysis was used to investigate data distribution due to necessity in symmetric shape in data distribution and comparable magnitude value range for multivariate analysis including cluster analysis (Reimann et al., 2008; Templ et al., 2008). Significant differences in range value could cause spurious pattern due to dominance of particular components. Thus, when data are not following the

above conditions, it is important to apply some processes such as data transformation before applying multivariate analysis.

Data transformation and standardization

Data transformation and standardization are applied when the data are not showing a symmetrical shape of data distribution with comparable magnitude value range. Transformation is chosen based on skewness that is nearest to zero. The zero skewness value usually shows symmetric shape of data distribution even it is not necessary. Standardization using Median Absolute Deviation (MAD) and median, which was developed by Yusta et al. (1998), *equation 2-10*, was employed in order to make comparable data range. This type of standardization is preferred to be employed than standardization using mean and standard deviation because its robustness to existence of outlier data (Carranza, 2008; Reimann et al., 2008). The formula is as shown below:

$$Z_{ij} = \frac{X_{ij} - median_j}{MAD_j} \qquad \text{Equation 2-10}$$

where

$$MAD = median[|X_i - median(X_i)|] \qquad \text{Equation 2-11}$$

X = measurements of element concentrations
I = sample number
J = element number

2.3.2.3. Compositional data approach

In the CoDa approach, two transformations were applied to the data, which are closure operation and isometric log ratio (*ilr*) transformation, after all data were converted into the same unit. The closure operation was applied because not all geochemical element concentrations were measured in the sample, thus the data are considered as sub-compositions of stream sediment geochemical data. The sub-composition is obtained when not all elements of the samples are measured or only particular elements are interesting to be analyzed. The closure operation is needed to make sub-composition data as the 'close' data, thus the operation for compositional data can be applied to the data. If the full composition of stream sediment geochemical data is formulated as $x=C[x_1, x_2, ..., x_D]$ in $\boldsymbol{S^D}$ then the data as a group of part is defined by a set of r subscripts; let $R = (i_1, i_2, ..., i_r)$ be such a set, pointing out the parts x_{i1}, x_{i2}, ..., x_{ir}. The $\boldsymbol{R}$-subcomposition of $\boldsymbol{x}$ is defined as composition in $\boldsymbol{S^r}$.

$$sub(x; R) = C[x_1, x_2, x_3, \dots, x_n] \qquad \text{Equation 2-12}$$

where the closure only affects the $\boldsymbol{r}$ parts in the $\boldsymbol{R}$-group with $\boldsymbol{C}$ is the closure operation. Thus, the Equation 2-12 can be written as:

$$sub(x; R) = C[x_1, x_2, \dots, x_n]$$
$$= \left[\frac{x_1 .k}{\sum_{i=1}^{n} x_i}, \frac{x_2 .k}{\sum_{i=1}^{n} x_i}, \dots, \frac{x_n .k}{\sum_{i=1}^{n} x_i}\right] \qquad \text{Equation 2-13}$$

with $\boldsymbol{k}$ is a constant, e.g., $\boldsymbol{k}$=1 if the data are fractions or $\boldsymbol{k}$=100 if the data are percentage (Filzmoser et al., 2009a; Pawlowsky-Glahn and Egozcue, 2006). After the closure operation was applied to the data, the *ilr* transformation was performed. This transformation makes the geometry of feature space of CoDa (simplex, $\boldsymbol{S}$) the same as that of Euclidean feature space. Thus, the distance among points in CoDa after ilr transformation is the same as Euclidean distance; therefore, multivariate analysis, such as cluster analysis, could be applied directly. The formula for *ilr* transformation is as follow:

$$z_i = \sqrt{\frac{i}{i+1}} \log \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}} \qquad \text{for } i = 1,2,\dots,n-1 \qquad \text{Equation 2-14}$$

with $z=(z_1, z_2,..., z_{n-1})$ is result of ilr transformation (Egozcue et al., 2003). All processing for CoDa approach was performed using *compositions* which is the *R* package developed by Van den Boogaart and Tolosana-Delgado (2008) and Van den Boogaart et al. (2008).

2.3.3. Clustering methods

The main purpose of clustering is to find patterns such as groupings in characteristics or behaviours within observation datasets. In this chapter, observation data are measured as element concentrations of stream sediment geochemistry. The data, based on their characteristics, are classified into groups/clusters/classes based on particular similarity/dissimilarity criteria. The aim of clustering algorithms is to minimize the dissimilarity objects within a group. Consequently, objects with a high degree of similarity are classified into the same cluster.

Furthermore, according to similarity/dissimilarity criteria, clustering algorithms could be divided into two approaches, distance-based and model-based approaches (Gan et al., 2007; Reimann et al., 2008). The distance-based approach clustering algorithm determines cluster members by calculating the distances between the samples; whereas the model-based by selecting appropriate model for the data as shown in (*appendix 2-1*).

A common type of distance measurement used in a clustering method is Euclidean distance, as formulated in *equation 2-15*.

$$d(x,y) = \left(\sum_{j=i}^{d} (x_j - y_j)^2\right)^{\frac{1}{2}} \qquad \text{Equation 2-15}$$

where; x and y are two data point, $x = (x_1, x_2, ..., x_n)$, $y = (y_1, y_2, ..., y_n)$ and $i=(1,2,...,n)$.

PAM clustering and Mclust, representing distance-based and model-based approaches, respectively, were applied to the data. Both of these clustering techniques were chosen because of their robustness technique to outlier data (Gan et al., 2007; Kaufman and Rousseeuw, 2005). In addition, Templ et al., (2008) stated that partitioning method such as PAM performs better than hierarchy method for large data and the results from model-based clustering are more reliable and interpretable. Therefore, these two techniques were chosen for the purpose of comparing the performance of distance- and model-based clustering algorithms in classifying the multivariate geochemical to assist lithological mapping.

2.3.3.1. Constraints in using clustering methods

Some problems which arise in using heuristic clustering algorithm such as PAM clustering are in determining "correct" cluster number and selecting "appropriate" components to be included. In determining correct cluster number, for distance-based, Templ et al. (2008) suggested to use plot for sum of squares of ratio distance within and between clusters versus cluster numbers (ratio plot). The cluster number is determined at the point where there is abrupt change in trend. However, sometimes the plot is showing no or several optimum indicators points. Especially for PAM clustering, Kaufman and Rousseeuw (2005) suggested to use Silhouette Coefficient (SC) value, which is defined as the maximum average silhouette width for entire data set. Cluster number is selected at which the SC value reaches a maximum value. The maximum value depicts that the cluster reach probable the most natural classification. The small value of SC describes that the data are not well separated but ambiguous in several clusters. In addition, as a reference, the subjective guidance to interpret SC value could be used (*appendix 2-2)*. Furthermore, in selecting components to be included, dendogram and principal component analysis could be used; however, their application to an area where little or no *a-priori* knowledge is available is not a trivial task and more subjective depend on the experiences. However, in model-based clustering, those two constrains could be resolved automatically. Algorithms from Raftery and Dean (2006) and Raftery (2009) could used to select components to be included in clustering; whereas the

optimum cluster number is determined automatically in Mclust algorithm. The selection by the algorithm is according to Bayesian Information Criterion (BIC).

2.3.3.2. Model-based Clustering (Mclust)

The Mclust algorithm optimizes the fit of the shape between the data and the models. The algorithm chooses cluster shape models and assigns memberships of individual samples into particular clusters. A cluster is describes by density of multivariate normal distribution with a particular mean and covariance. For this purpose, the Expectation Maximisation (EM) algorithm is used. This algorithm is applied to several clusters and with several sets covariance matrices of the clusters. The best model with certain cluster number was determined by the highest BIC value (Fraley and Raftery, 2002; Fraley and Raftery, 2006; Reimann et al., 2008; Templ et al., 2008).

Gan et al., (2007) divided model-based clustering algorithm into three main steps. First is initializing the EM algorithm using the partitions from model-based agglomerative hierarchical clustering. Then, the parameters are estimated using the EM algorithm. The last step is choosing the model and the number of clusters according to the BIC (Fraley and Raftery, 2002; Gan et al., 2007). The models of Mclust can be seen in *appendix 2-1*. *R* package from Fraley and Raftery (2002; 2006), *mclust* package, was employed to transformed data both for conventional and CoDa approach.

2.3.3.3. Partition Around Medoids (PAM) clustering

The aim of PAM algorithm is to minimize sum average distances to the cluster medians. These medians are representative objects which represent the structure of the data. These medians are so-called medoids of the cluster. The first step in the algorithm is to set number of medoids (***k***) then ***k*** clusters are constructed by assigning each object of the dataset to the nearest medoids. The nearest criterion is determined base on Euclidean distance or Manhattan distance. In this research, Euclidean distance was employed.

According to Kaufman and Rousseeuw (2005), the algorithm of PAM clustering is as follow. Let set of objects is denoted as $X = \{x_1, x_2, ..., x_n\}$ and the dissimilarity between objects x_i and x_j denoted by $d(i,j)$. The algorithm consist two steps. First is selecting of objects as medoids in cluster: y_i is defined as binary variable (1 or 0). The value of y_i will equal to 1 if the object x_i $(i=1,2,..., n)$ is selected as a medoids. Second step is to assign each object $\mathrm{x_j}$ to one of the selected medoid. The value of z_{ij} is also binary value (0 or 1). The z_{ij} has value of 1 if and only if the object $\mathrm{x_j}$ is assigned to cluster of which x_i is the medoid.

$$\text{minimize } \sum_{i=1}^{n}\sum_{j=1}^{n} d(i,j)\, z_{ij} \qquad \text{Equation 2-15}$$

subject to

$$\sum_{i=1}^{n} z_{ij} = 1, \qquad j = 1,2,\dots,n$$

$$z_{ij} \le y_i, \qquad i,j = 1,2,\dots,n$$

$$\sum_{i=1}^{n} y_i = k, \quad k = \text{number of cluster}$$

$$y_i, z_{ij} \in \{0,1\}, \qquad i,j = 1,2,\dots,n$$

thus the dissimilarity of an object *j* and its medoid is as following

$$\sum_{i=1}^{n} d(i,j) z_{i,j} \qquad \text{Equation 2-17}$$

because all objects must be assigned, the total dissimilarity can be written as in *equation 2-17*. The PAM clustering was performed using *cluster* package in *R* statistic software (Maechler, 2005).

2.3.4. Post clustering

2.3.4.1. Interpolation and rasterizing

Interpolation and rasterization are the next steps after clustering in order to assess the quality of classification. Thiessen polygon was performed to interpolate unsample areas. Then polygons were converted to raster image with particular grid size. The grid size was determined by using *Equation 2-18* proposed by Hengl (2006).

$$p = 0.25 \,.\, \sqrt{\frac{A}{N}} \qquad \text{Equation 2-18}$$

where A is study area in m^2 and N is the total number of observations/samples. The formula is suitable for random or clustered distribution sample points such as stream sediment sample points.

2.3.4.2. Reclassifications and filtering

Reclassification using majority rules, the same that developed by Lang et al. (2008) for labelling the classification results, was applied to clustering results in order to make number of cluster the same as number of reference classes (the existing lithological map). First, all clusters were assigned into category which the highest pixel number of the cluster lies on. If there are two or clusters having the same categories, the cluster with the highest pixel numbers among them was selected as 'key' cluster. If the same cluster was selected as key cluster, the second highest was taken as key cluster and so on. In the case where there is no key cluster in category, a cluster with the highest pixel number within that category was assigned as key cluster. The final step is joining non-key cluster into key cluster with the same category.

Majority filtering was applied to the images of clustering results as a suggested filtering technique by Lillesand and Kiefer (2000) to "smooth" classified data. The purpose of clustering is to eliminate a single or small cluster; thus, the final results only show the dominant cluster that presumably is the correct classification. The filtering employs a moving window of 3x3 pixels. The window is moved over the image like a moving kernel. At every position, the filtering will change the identity of centre pixel in a moving window to majority class when its class is not the majority one. In the case where there is no majority class; the identity of the centre pixel still remains.

2.3.5. Assessments

Two types of assessment, overall accuracy percentage from an error matrix or a contingency table and kappa coefficient, were calculated from the results of clustering compared to simplified existing lithological map (*appendix 1-1 (a)*). The meaning of accuracy in this research is a coincidence percentage between clustering results and the existing geological maps. A cross-validation on the field is needed in order to indicate a "real" accuracy for such method (Schetselaar, 2000).

2.3.5.1. Error matrix

Error matrix is one of the most common ways to express accuracy of the classification. Error matrix shows relationship between known reference data and the corresponding cluster/class of clustering results. An existing lithological map was used as the reference data for the research. The matrices have the same rows and columns as the numbers of categories as lithological features in existing lithological map (Lillesand and Kiefer, 2000).

In an error matrix, two type of accuracy, overall and producer's accuracy, were taken to assess the accuracy of the clustering. Overall accuracy is a ratio of total pixels number correctly classified to total pixels number, whereas producer's accuracy is a ratio between the numbers of correctly classified pixels by the total pixel number in each category. The values describe how well the pixels are classified using clustering method.

2.3.5.2. Kappa coefficient

Kappa coefficient could be used to indicate the level of the percentage correct values in an error matrix caused by true agreement or only chance agreement. *Equation 2-19* is used to calculated kappa coefficient:

$$k = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+} - x_{+1})}{N^2 - \sum_{i=1}^{r}(x_{i+} - x_{+i})} \qquad \text{Equation 2-19}$$

where

r = number of rows in the error matrix
x_{ii} = the number of observation in row i and column i (on the major diagonal)
x_{i+} = total of observations in row i (shown as marginal total to right of matrix)
x_{+i} = total of observations in row i (shown as marginal total at bottom of matrix)
N = total number of observations included in matrix.

The kappa coefficient ranges from 0 to 1. One indicates true agreements and zero indicates chance agreements.

2.4. Results and discussion

2.4.1. Stream sediment geochemical data reliability

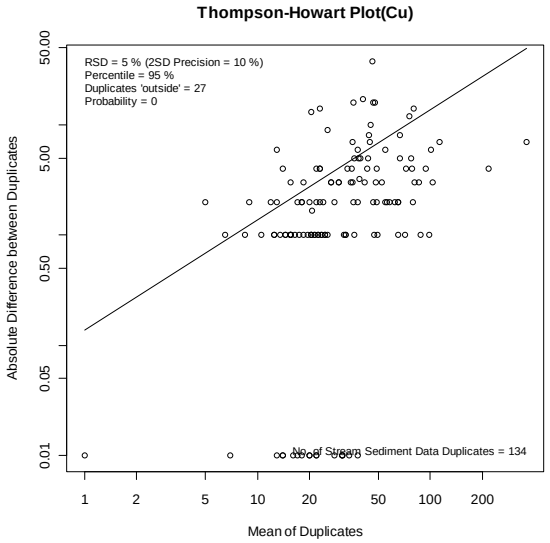


Figure 2-2 Cu data lie on Thompson-Howarth Plot (a model for 10% precision and 95% significant level)

Figure 2-2 shows Thompson-Howarth plot, a precision model, which contains information of the numbers of 'outlier' data in duplicate samples for Cu. The value of outlier data lies outside the data range of the precision model. The data are represented by points that fall to above/left of the line. The points represent a pair of duplicate samples whereas the line represent percentile of the half normal distribution (Garrett and Grunsky, 2003; Reimann et al., 2008). In *figure 2-2* shows Howarth-Thompson plot for Cu with 5% of RSD value or 10% precision and 95th-tile line. Thompson-Howarth plot for others elements could be seen at *appendix 2-3*. In addition, *table 2-1* is summary of percentage that fall below the 95th-tile line. Thus, as illustration for Cu, the table shows that the value is 80%, it means 20% of the data are lying outside the data range of the model for 10% precision and 5% significant level.

The results from ANOVA, *table 2-2*, shows that the *F*-values for all elements are greater than F_c (=1.33). The F_c was determined for 5% in confidence level, 133 and 134 for numerator and denominator degree of freedom, respectively. In addition, regarding the percentage of procedural error value, Zn, Cu, Pb, Ni, Co, Mn, Fe, Hg, Sb, As, Ba, Ce, and Cr have satisfactory results, whereas Ag, Mo and Rb have poor results because their percentages are more than 20%. The Howarth-Thompson plot, in this case, could not be used to determine the reliability of stream sediment geochemical data because all elements have precisions of less than 10%. Thus the reliability of the data was determined by ANOVA test results. Consequently, from the results, there are three elements (Ag, Mo, Rb) were excluded for the next analysis. It is because their procedural error is too large thus their variance could not represent geochemical variance in nature.

Table 2-1 Summary of geochemical data quality assessment based on a Thompson-Howarth Plot (a model for 10% precision and 95% significant level)

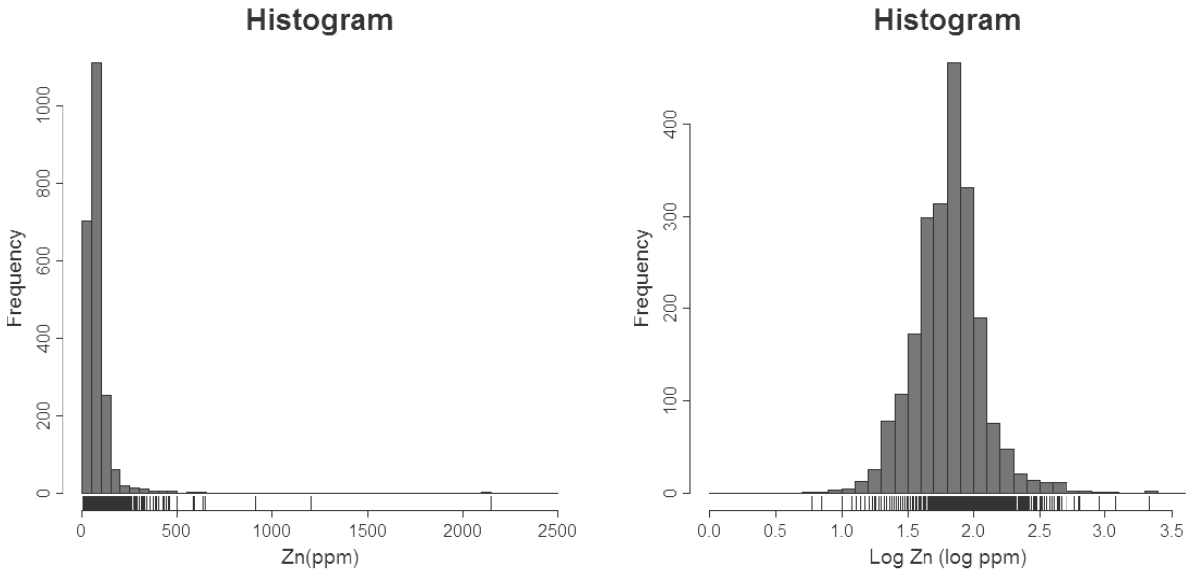
Elements	% Point Plotted Below 95^{th} Tile Line
Zn	86%
Cu	80%
Pb	57%
Ni	84%
Co	77%
Ag	86%
Mn	74%
Fe	83%
Mo	75%
Hg	51%
Sb	62%
As	42%
Ba	61%
Ce	56%
Cr	52%
Rb	51%

Table 2-2 Summary of geochemical data quality assessment based on ANOVA test

Elements	MS_g	MS_p	F - value
Zn	96.8%	3.2%	30.3
Cu	97.5%	2.5%	39.0
Pb	95.1%	4.9%	19.4
Ni	98.7%	1.3%	75.9
Co	96.0%	4.0%	24.0
Ag	61.9%	38.1%	1.6
Mn	93.3%	6.7%	13.9
Fe	96.0%	4.0%	24.0
Mo	72.2%	27.8%	2.6
Hg	86.5%	13.5%	6.4
Sb	90.4%	9.6%	9.4
As	88.2%	11.8%	7.5
Ba	89.9%	10.1%	8.9
Ce	95.1%	4.9%	19.4
Cr	94.8%	5.2%	18.2
Rb	80.5%	19.5%	4.1

2.4.2. Data distribution

Histogram show that data distribution of Zn is positively skewed (*figure 2-3(a)*). The same trend happens for others elements of stream sediment geochemistry (*appendix 2-4*). It means most of the data have low values and distribution is not symmetric. Therefore, base-10 logarithmic transformation was applied to the data as suggested by Reimann et al. (2005) that stream sediment geochemical data usually follow logarithmic normal distributions. Results of the transformation could be seen in *figure 2-4(b)*, showing that more symmetric and less skewed data distribution. Histogram for logarithmic transformation for others element could be seen in *appendix 2-4*. Significant differences of skewness value before and after transformation could be seen at the last column in *table 2-3*.



(a) (b)

Figure 2-3 Histogram for Zn, (a) graphs for raw data without transformation, (b) graphs for base-10 logarithmic transformation data

Table 2-3 Univariate statistics for stream sediment geochemical raw data and skewness of base-10 logarithmic transformed data

Elements	Units	Detection Limit	Missing Value	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.	Skewness	
										No transf.	Log
Zn	ppm	0.1	0.50%	5.0	45.0	66.0	78.5	89.0	2,150.0	12.90	0.40
Cu	ppm	0.01	0.50%	1.00	20.00	32.00	45.04	54.25	1,170.0	7.67	-0.08
Pb	ppm	0.01	0.05%	1.00	2.00	4.00	6.21	7.00	530.0	23.65	0.37
Ni	ppm	0.10	0.05%	1.00	14.00	23.00	37.44	33.00	1,310.0	9.17	0.27
Co	ppm	0.1	0.05%	1.0	8.0	11.0	12.4	15.0	217.0	7.44	-0.62
Mn	ppm	1.00	0.05%	24.00	315.80	545.50	710.50	850.50	18,700.0	9.01	0.10
Fe	%	0.01	0.05%	0.05	1.80	2.50	2.58	3.20	18.0	2.41	-0.93
Hg	ppb	5.00	0.05%	10.00	30.00	50.00	112.60	80.00	100,000.0	49.51	0.72
Sb	ppm	0.02	3.90%	0.10	0.60	1.00	1.23	1.50	26.4	8.45	-0.44
As	ppm	0.10	3.90%	0.50	4.00	7.70	11.49	13.00	438.0	10.95	-0.46
Ba	ppm	0.5	3.90%	50.0	660.0	946.0	1,046.0	1,300.0	29,700.0	16.72	-0.55
Ce	ppm	0.50	4.90%	3.00	34.00	46.00	69.73	73.00	937.0	3.81	0.45
Cr	ppm	5	4.90%	5	81	140	223	240	5,610.0	7.27	-0.38

2.4.3. Clustered Images

The algorithm from Raftery and Dean (2006) and Raftery (2009) has selected nine elements from 13 elements to be included in Mclust to get an optimum cluster number. The nine elements (As, Zn, Cr, Ce, Ba, Sb, Fe, Co, Ni) will be called Selected Elements. These selected elements were also employed in PAM clustering to test whether the clustering using selected elements gives better results than using all elements.

Different model types and cluster numbers are obtained from Mclust for different datasets and transformation. In conventional approach to data preparation, Model types fitting the data are a VEV-model with six clusters and a VVV-model with eight clusters for all elements and selected elements, respectively. A VVV-model with five clusters fits the data in the CoDa approach. The VEV-model represents clusters with ellipsoidal distribution, equal shape and varying volume, shape, and orientation; whereas the VVV-model represents clusters with ellipsoidal, varying volume, shape, and orientation (Fraley and Raftery, 2006). More model types can be seen in *appendix 2-1*.

Cluster number (***k***) from 3 to 15 was tried as an input in PAM clustering in order to get an optimum cluster number which shown by the highest SC value. However, SC values for all datasets do not show an optimum SC value rather than relatively constant. The SC values are steady in range of 0.10 to 0.12. According to SC subjective interpretation table from Kaufman and Rousseeuw (2005), SC value below or

equal to 0.25 is indicating no substantial structure in the data. Therefore, the same cluster number as an optimum cluster numbers from Mclust for the same dataset was used as an input in PAM clustering.

Figure 2-5 shows the results of clustering using Mclust and PAM clustering with two different data preparation approach, conventional and CoDa. In general, the results show prominent cluster patterns for three features, one feature lies on the centre and elongates from north-west to south-east, other lies on the eastern part and the rest lies on the southern part. Moreover, according to existing lithological map (*see appendix 1-1*), these features might be related to three lithologies which are Intrusive Rocks, Volcanic Rocks and Sedimentary Rocks 4, respectively.

Furthermore, *figure 2-6* shows final results after post clustering processes (interpolation, rasterizing, reclassification, and filtering) were performed. In rasterizing, a 500 m grid size was selected as a result of compromising among proper grid size according to sample density and memory of ***R*** application. Whereas according to *equation 2-18* by using density sample (1 sample per 10 km^2) as an input, 250 m is obtained as a proper grid size. In addition, generally, the patterns in these final results have the same patterns as the clustering results.

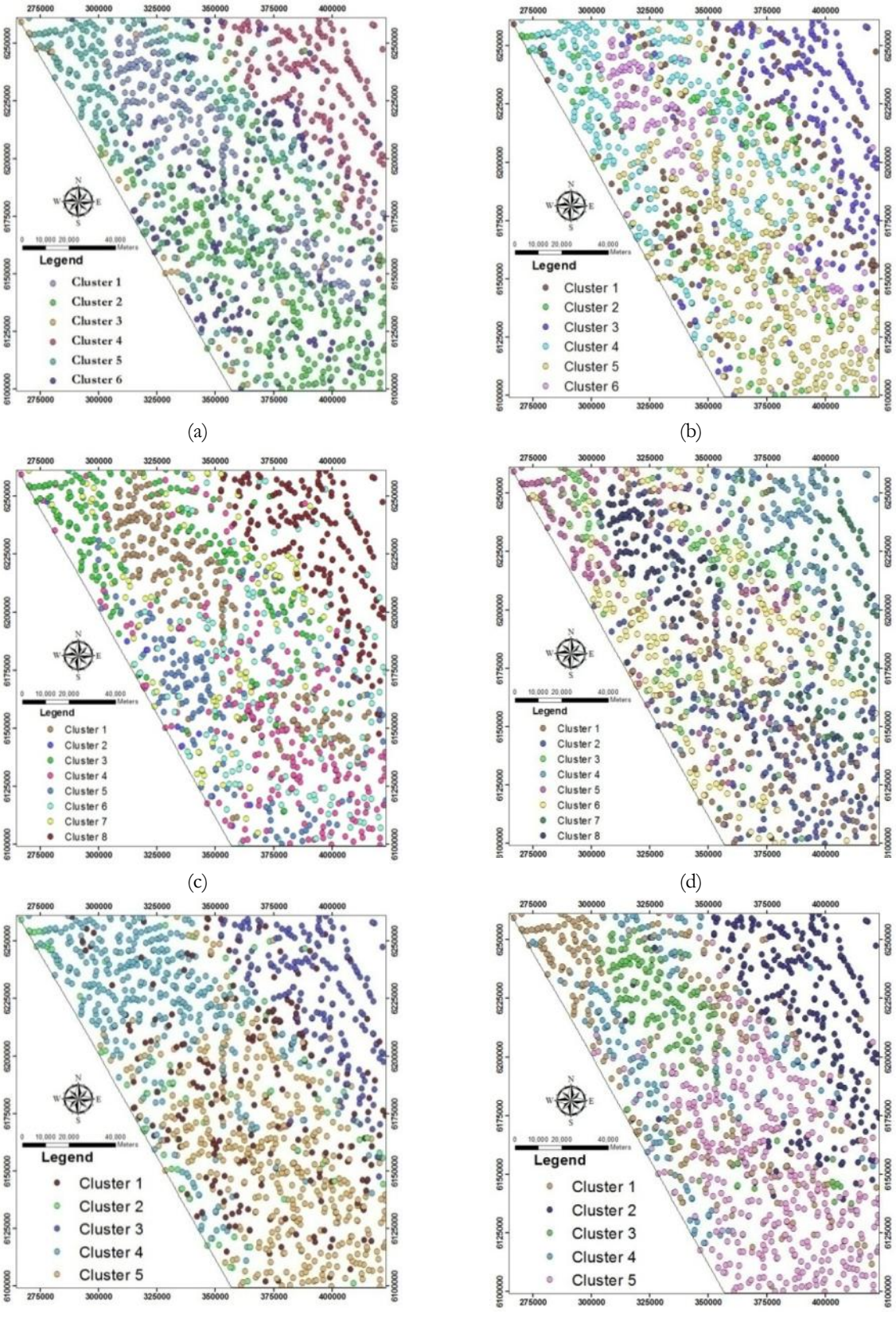


Figure 2-4 Clustering results of stream sediment geochemical data for different clustering techniques and different approaches in data preparation, (a) Mclust with conventional approach for all elements, (b) PAM clustering conventional approach for all elements, (c) Mclust with conventional approach for selected elements, (d) PAM clustering with conventional approach for selected elements, (e) Mclust with CoDa approach, (f) PAM clustering with CoDa approach

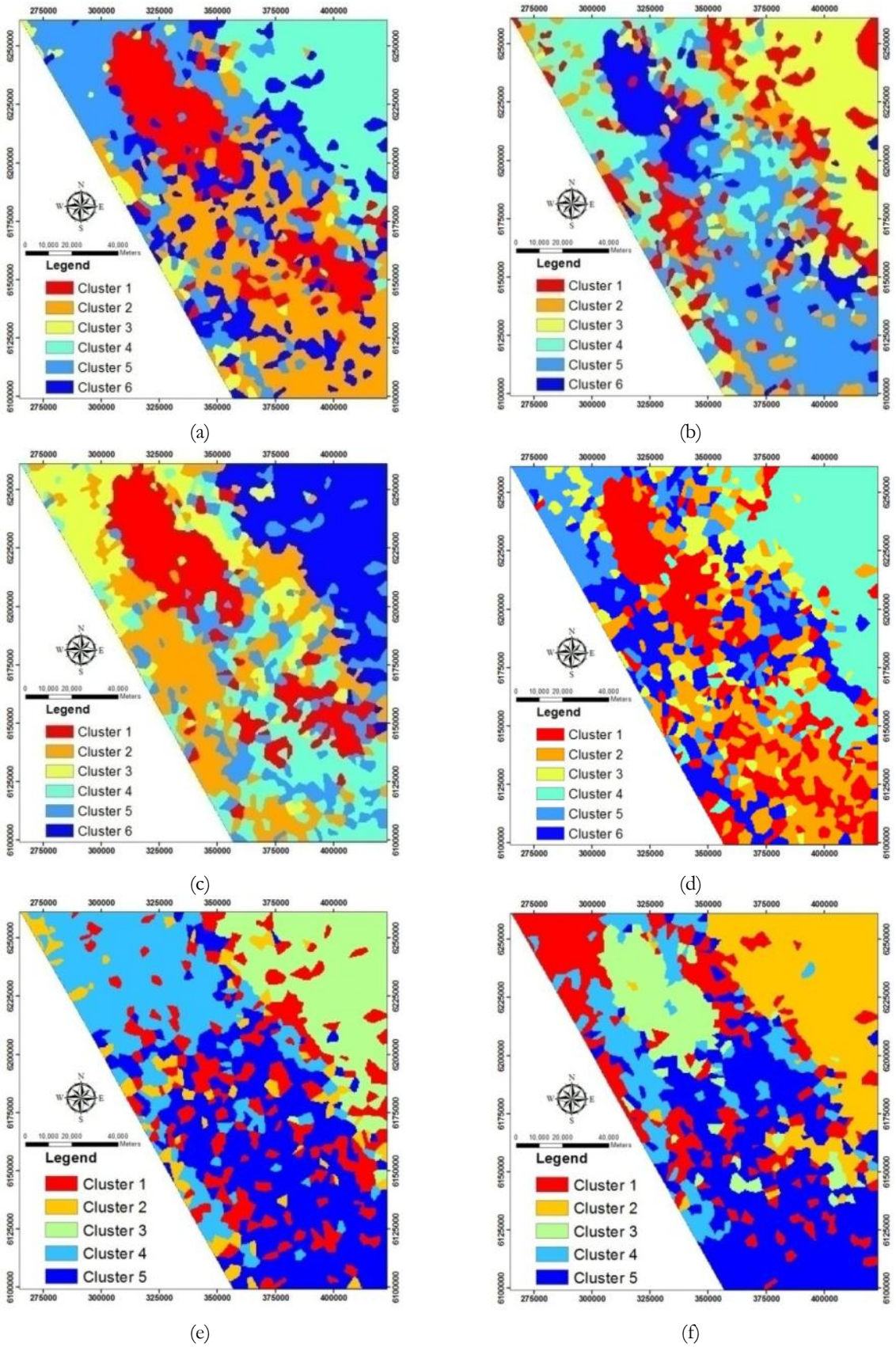


Figure 2-5 Results after post clustering (rasterizing, reclassification and filtering) of stream sediment geochemical data for different clustering techniques and different approaches in data preparation, (a) Mclust with conventional approach for all elements, (b) PAM clustering conventional approach for all elements, (c) Mclust with conventional approach for selected elements, (d) PAM clustering with conventional approach for selected elements, (e) Mclust with CoDa approach, (f) PAM clustering with CoDa approach

2.4.4. Quantitative data quality

Producer's accuracy table shows that Mclust accuracy is higher than that of PAM clustering in most lithologies (*figure 2-7*). Lithologies A, B and F have moderate to high accuracies for most clustering results. These results are consistent with visual interpretation of clustering results that show three prominent features that are related to Volcanic Rocks, Intrusive Rocks and Sedimentary Rocks 4. For Lithology C, the Sedimentary Rocks 1, the accuracy is low for most clustering results. This might be because this lithology lies beneath Williston Lake and is surrounded by another and more abundant Sedimentary Rocks 4 thus its geochemical signature was covered.

According to *table 2-4,* Mclust is better than PAM clustering when clustering using both all elements and selected elements. Clustering number, as input in PAM clustering, might be one of the factors why the accuracy of PAM clustering is lower compared to Mclust. As explained in the previous paragraph, the SC values for PAM clustering are not indicating an optimum number of clusters. In addition, the table also describes that the clustering using selected elements produces a slightly higher accuracy than using all the elements. However, the Kappa coefficient for PAM clustering using all elements is little higher than clustering using selected elements. Thus, it is hard to say that clustering using selected elements is better than using all elements. Furthermore, in regard to data preparation, CoDa approach resulted in lower accuracy for Mclust but higher for PAM clustering compared to the conventional approach. Furthermore, Mclust is more appropriate using conventional approach, whereas for PAM clustering, CoDa is preferable to conventional approach. However, according to the Kappa coefficient value for PAM clustering, there is no significant difference in the value between using conventional and CoDa approach in data preparation. This suggests that the data are likely not too much affected by the closure problem. It might be because the average of the total concentrations of stream sediment geochemistry is less than 5%. The closure problem arises when total concentrations of elements approaches ***k***, in this case 100%.

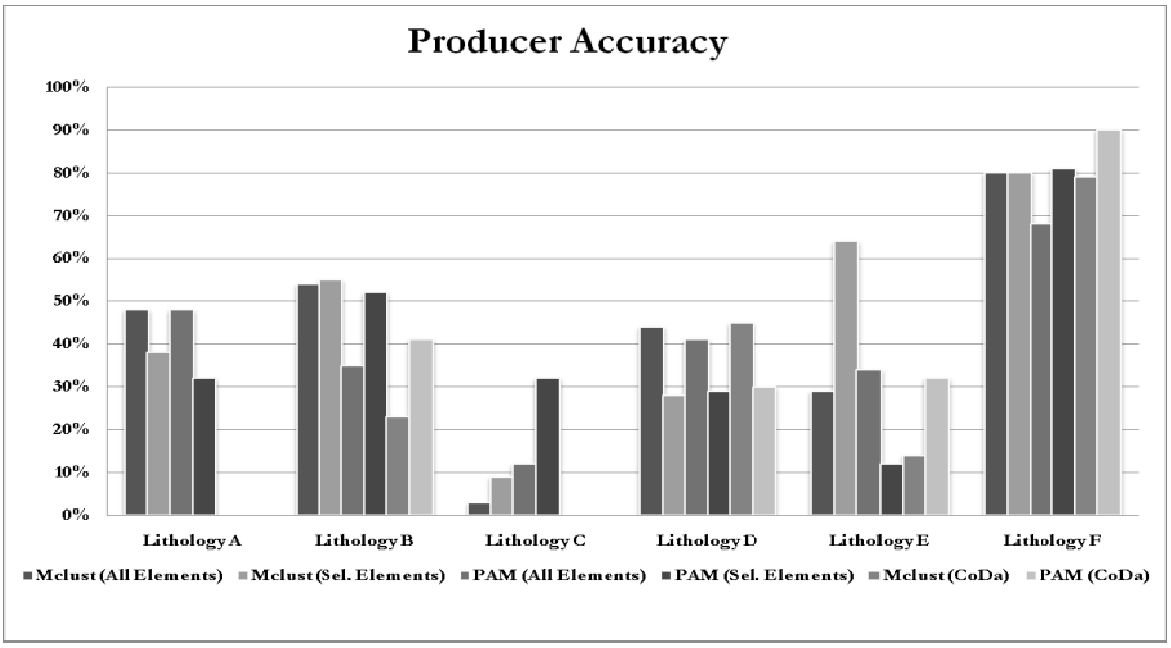


Figure 2-6 Producer's accuracy diagram from assessment of clustering results for stream sediment geochemical data (for error matrices see *appendix 2-5*); Lithology A, Volcanic Rocks ; Lithology B, Intrusive Rocks ; Lithology C, Sedimentary Rocks 1; Lithology D, Sedimentary Rocks 2; Lithology E, Sedimentary Rocks 3; Lithology F, Sedimentary Rocks 4 (for explanations about the lithology at the study area see *section 1.6.2*)

Table 2-4 Data quality of clustering results for stream sediment geochemical data

Clustering	Overall Accuracy	Kappa Coefficient
Mclust (All Elements)	50%	0.39
Mclust (Sel. Elements)	51%	0.41
PAM (All Elements)	44%	0.32
PAM (Sel. Elements)	43%	0.31
Mclust (CoDa)	43%	0.27
PAM (CoDa)	50%	0.37

2.5. Concluding remarks

- Conventional approach in data preparation stage is plausible to be employed for stream sediment geochemical data with low total elements concentration before the application such clustering methods.
- Clustering using selected elements in stream sediment geochemical data produces non-significant accuracies. The accuracies are increasing only 1% and and 0.02 for overall accuracy and kappa coefficient, respectively.
- The application of Mclust to stream sediment geochemical data classification produces better accuracies than the application of PAM clustering in most assessment including the producer's accuracy for most lithologies (see *figure 2-6* and *table 2-4*).

3. AIRBORNE GAMMA-RAY DATA ANALYSIS

3.1. Introduction

Similar to *chapter 2*, in this chapter two applications were tested which are the application of different clustering techniques and the application of compositional data (CoDa) approach. Like in the previous chapter, two clustering algorithms, model-based clustering, Mclust and distance-based, PAM (Partition Around Medoids), were performed. Furthermore, conventional and CoDa approach were applied in the data preparation stage.

3.2. Descriptions of the Airborne Gamma-ray Dataset

The airborne radiometric data used in this research were downloaded from the National Gamma-Ray Spectrometry Program (NATGAM) Data Base from the Geological Survey of Canada website (Natural Resources Canada, 2010b). Data on concentrations of radioelements (K, eTh, and eU) are available in 250 m grid size. K concentration is measured in percentage (%), whereas eTh and U are in parts per million (ppm). The data are the result of several hundreds of airborne surveys of the Geological Survey of Canada for over 30 years since 1970. The aircraft flew along a pattern of parallel flight lines at 120 m terrain clearance with line spacing 200-500 m. The aircrafts flew at speed around 190 km/h. The data were acquired by sampling every 1 second interval, which equivalent around 60 m on the ground (Natural Resources Canada, 2010b).

3.3. Methodology

The methodology described in this chapter could be divided into several stages such as data preparation, clustering, post clustering and assessments. In data preparation stage, two approaches, conventional and CoDa approach, were applied for comparison. Furthermore, two types of clustering techniques were applied for data classification, Mclust and PAM clustering. The post-clustering comprises reclassification and filtering. The assessment is the last stage, which was employed after post-clustering reclassification was conducted. The flow chart of the methodology can be seen in *figure 3-1*.

3.3.1. Data preparation

Univariate graph such as histogram was used to investigate element data distribution. For examining spatial distribution, a visualization of element concentration data to an image is quite useful to get an impression of the spatial data distribution. In addition, square-root transformation was chosen to be applied according to the smallest skewness value (Reimann et al., 2008; Weisberg, 2005). For standardization, median-MAD (median absolute deviation) standardization was applied to transformed data (*equation 2-10*). Transformation and standardization were conducted to get a symmetric data distribution shape and comparable range values as suggested by Reimann et al. (2008) and Templ et al. (2008).

In the CoDa approach, several processes were applied in data preparation stage such as data transformation. The first process was converting the data into the same unit of measurement. In this case, K concentrations were converted form percentage into ppm. The three radioelement concentration data from airborne gamma-ray surveys are considered as sub-composition data; thus, they are subjected to the closure operation using *equation 2-13*. Then, the isometric log ratio (*ilr*) transformation, *equation 2-4*, was also applied before the applications of cluster analysis algorithms.

3.3.2. Clustering

Two types of clustering algorithm, Mclust and PAM clustering, were applied to the data for classification. An optimum cluster number for PAM clustering is determined by the highest value of SC (or Silhouette Coefficient); whereas for Mclust, the number is determined by the algorithm based on the highest BIC (or Bayesian Information Criterion) value (*see section 2.2.3.1*).

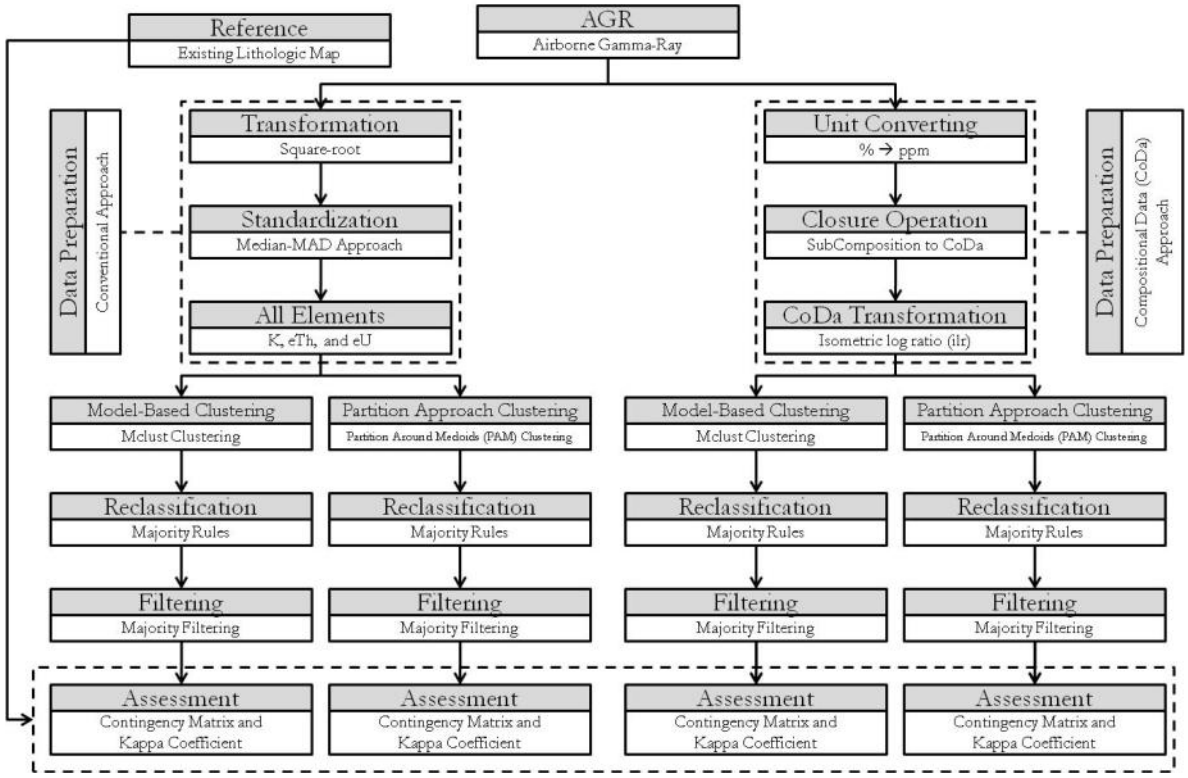


Figure 3-1 Flow chart of methodology to map lithology using airborne gamma-ray datasets by applying two types of clustering algorithm, Model-based Clustering (Mclust) and Partition Around Medoids (PAM) clustering with two approach in data preparation stage, conventional and compositional data (CoDa)

3.3.3. Post clustering

Reclassification and filtering were conducted in sequence after clustering algorithm was applied to the data. Reclassification is needed to make cluster number the same as the reference features for validation whereas the purpose of filtering is to eliminate small or individual cluster. Reclassification using majority rules and majority filtering were performed before the assessments as explained in *chapter 2*.

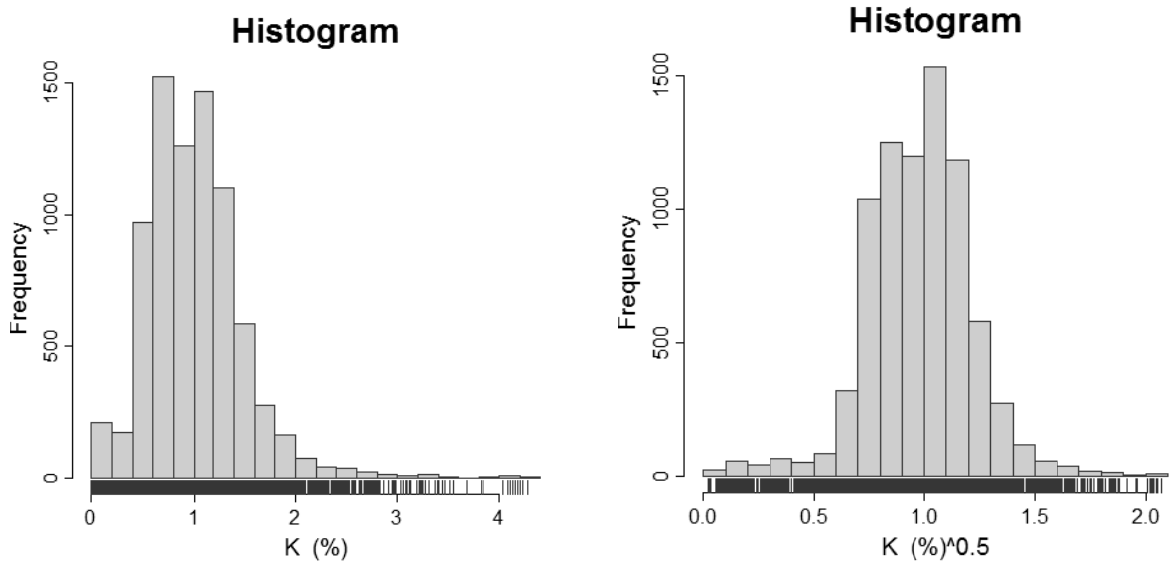
3.3.4. Assessments

Error matrix and kappa coefficient were calculated to assess quality of classification results from Mclust and PAM clustering. The clustering results were compared to reference data which is a simplified existing lithological map (*appendix 1-2*). From the error matrix, the overall accuracy and producer's accuracy was calculated to assess the classification outcome. Furthermore, the kappa coefficient, as an indicator to estimate true agreement of classification from the percentages of an error matrix, was also calculated (Lillesand and Kiefer, 2000).

3.4. Results and discussions

3.4.1. Data distribution of airborne gamma-ray elements

Histogram in *figure 3-2 (a)* show that data distributions thereof K data is positively right skewed. *Appendix 3-1,* data distribution for eTh and eU, show the same event as that of K. The event describes that low concentration of the elements is more dominant than high concentration. Consequently, the data distributions have asymmetric shape. Therefore, square-root transformation was performed to the data in order to get more symmetric data distribution. The transformation was chosen base on the smallest skewness values. The transformation reduced skewness values more than eight times in average compared to skewness values before transformation (*table 3-1*). *Figure 3-2 (b)* and *appendix 3-1* visualize that after the transformation, the data distribution is more symmetric.

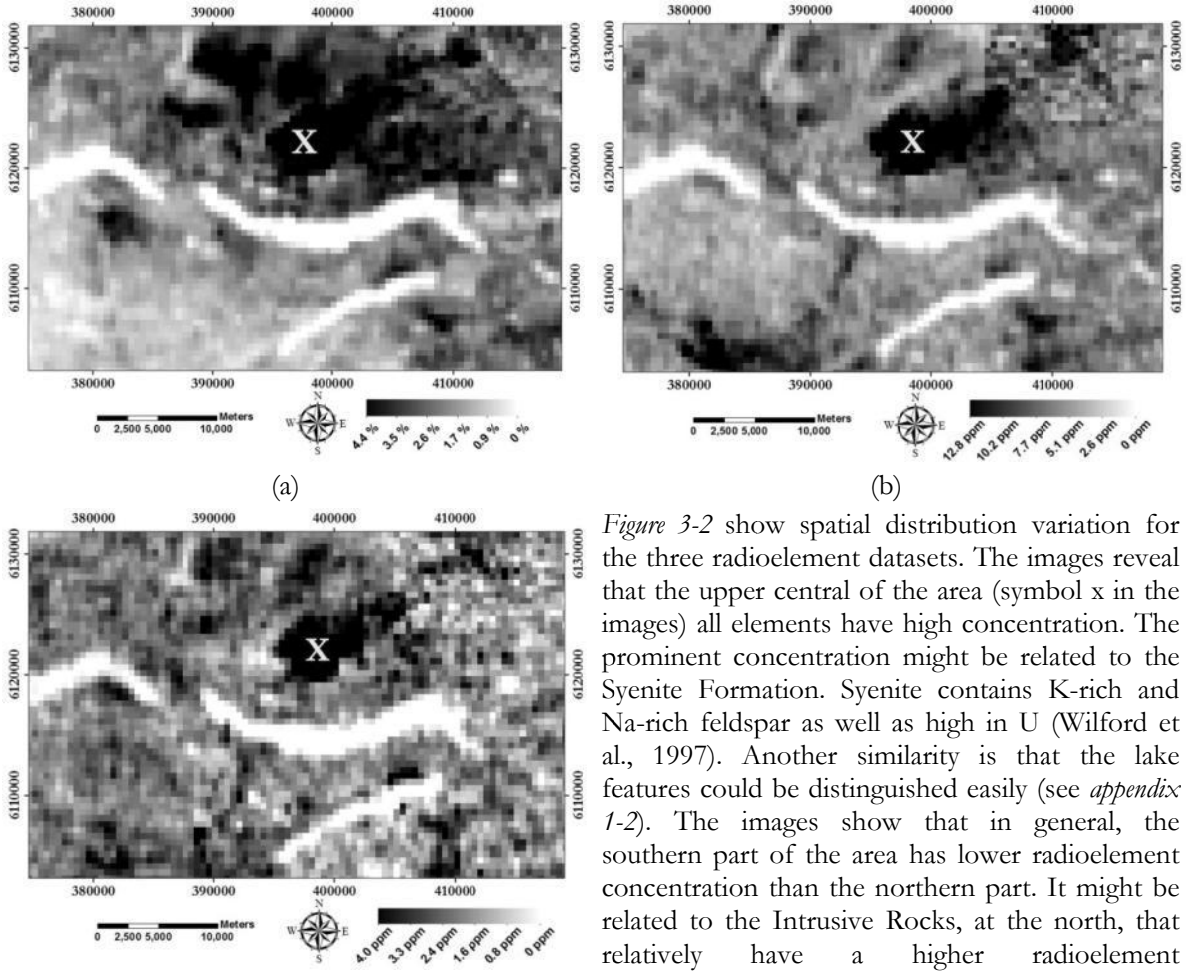


(a) (b)

Figure 3-2 Histogram for potassium (K), a) histogram for raw data, b) histogram after square-root transformation

Table 3-1 Univariate statistics summary for airborne gamma-ray raw data and its skewness of square-root transformed data

Elements	Units	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.	Skewness	
								No Transf.	square-root Transf.
K	%	0.00	0.68	0.98	1.01	1.26	4.28	1.28	-0.14
eTh	ppm	0.00	1.26	1.54	1.62	1.86	12.70	5.21	1.18
eU	ppm	0.00	0.55	0.68	0.68	0.79	4.14	3.20	0.25



(a) (b)

(c)

Figure 3-3 Spatial data concentration distribution of airborne gamma-ray elements (a) K (%), (b) eTh (ppm) and (c) eU (ppm)

Figure 3-2 show spatial distribution variation for the three radioelement datasets. The images reveal that the upper central of the area (symbol x in the images) all elements have high concentration. The prominent concentration might be related to the Syenite Formation. Syenite contains K-rich and Na-rich feldspar as well as high in U (Wilford et al., 1997). Another similarity is that the lake features could be distinguished easily (see *appendix 1-2*). The images show that in general, the southern part of the area has lower radioelement concentration than the northern part. It might be related to the Intrusive Rocks, at the north, that relatively have a higher radioelement concentration to sedimentary rock, at the south. Furthermore, in general, Th and U have a similar spatial concentration distribution than with K. It is because Th and U have converse characteristic with K such as in weathering and mineralization. Nevertheless, spatial distribution of U is more erratic as can be seen at eastern part. The heterogeneous spatial distribution of U is understandable because this element is more easily influenced by weathering and mineralization than K and Th (Dickson and Scott, 1997; Wilford et al., 1997).

3.4.2. Clustered images

Application of Mclust to radiolement data based on different approaches in data preparation produces different models (*figure 3-4(a)* and *figure 3-4(c)*): a VVV-model with nine clusters and a VEV-model with five clusters are the models obtained after application of conventional and CoDa data preparation approaches, respectively. The VVV-model represents clusters with ellipsoidal, varying volume, shape, and orientation in features spaces; whereas the VEV-model represents clusters with ellipsoidal distribution, equal shape and varying volume, shape, and orientation (Fraley and Raftery, 2006). More model types can be seen in *appendix 2-1*.

SC values using a cluster number (k) from 3 to 15, as an input in PAM clustering, are ranging from 0.21-0.30. The highest numbers obtained at k =3, however the cluster result is too general when it was visualized. Therefore, the same cluster numbers as an optimum cluster numbers from Mclust, were selected to get more clusters as shown in *figure 3-4(b)* and *figure 3-4(d)*.

Clustering result images for CoDa approach (*figure 3-4(c)* and *figure 3-4(d)*) have more isolated and small clusters than those of conventional transformation (*figure 3-4(a)* and *figure 3-4(b)*). The small clusters could be seen at the north-eastern part of the area. The images could not identify the area, where all

radioelements have high concentrations as clearly seen in *figure 3-3*, and the lake features. However, an area at the southwest is more homogeneous in the results of CoDa approach than those in the results based on the conventional approach.

After post clustering processes were conducted, all images show that a consistency in identifying some particular features. For example, the images could give perceptions about the boundaries between sedimentary rocks and volcanic rocks as well as sedimentary rocks and intrusive rocks. Moreover, the area at the centre of the area, which contains high concentrations of K, eTh and eU, which might be related to occurrence of Chuchi Syenite Formation, could be detected in most images. This area might be. However, the algorithms failed to differentiate between volcanic rocks and intrusive rocks at north western and eastern part respectively (*figure 3-4*). The above results are consistent with the references that explained about the confusion in differentiating among volcanic rocks, intrusive rocks. This could be due to count levels of volcanic rocks is much lower than those of the intrusive units (Graham and Bonham-Carter, 1993).

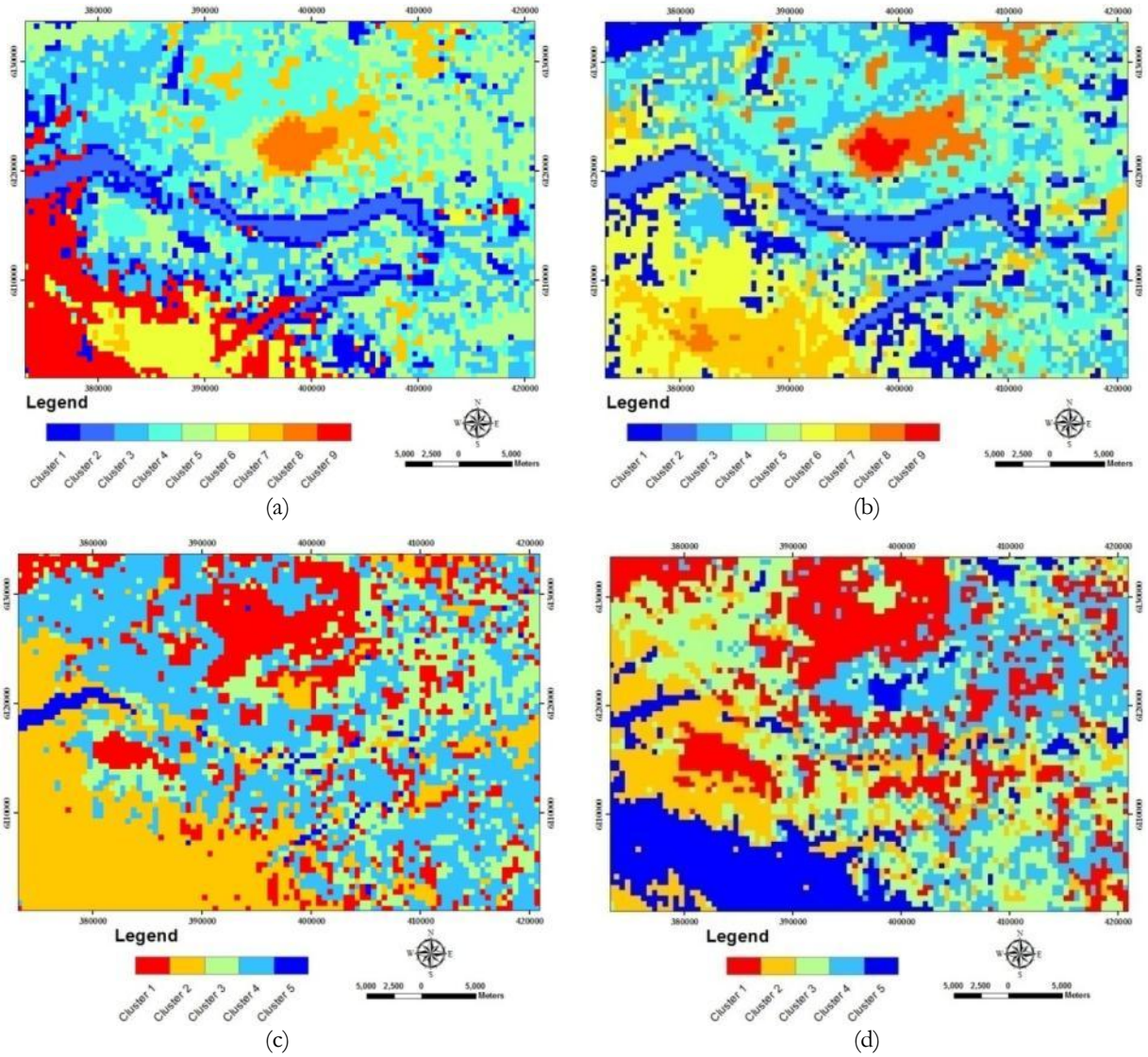


Figure 3-4 Clustering results of airborne gamma-ray data for different clustering techniques and different approaches in data preparation, (a) Mclust with conventional approach, (b) PAM clustering conventional approach, (c) Mclust with compositional data approach, (d) PAM with compositional data approach

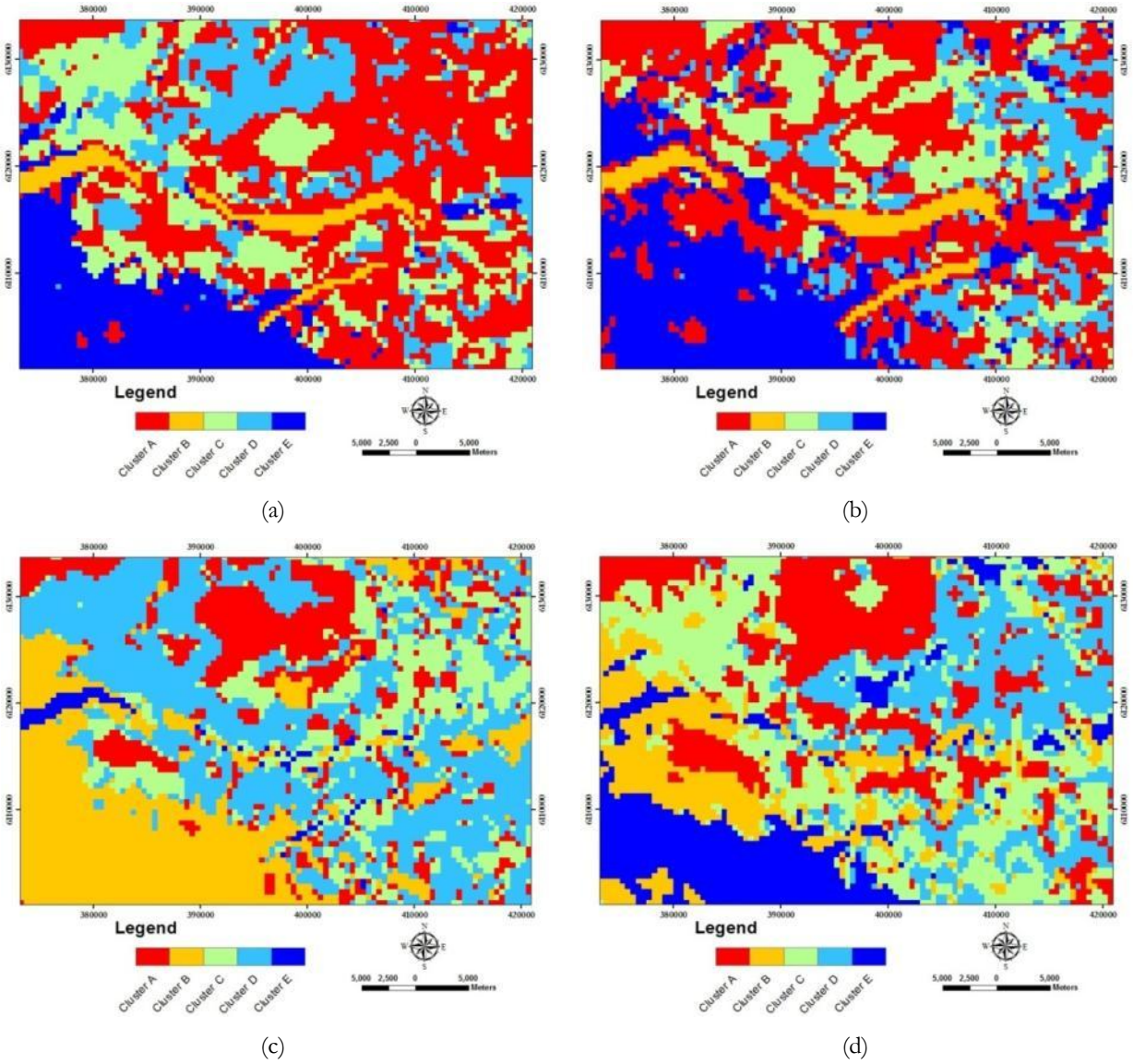


Figure 3-5 Results after reclassification and filtering of airborne gamma-ray data for different clustering techniques and different approaches in data preparation, (a) Mclust with conventional approach, (b) PAM clustering conventional approach, (c) Mclust with compositional data approach, (d) PAM with compositional data approach

3.4.3. Quality of the classification

Figure 3-6 show that producer's accuracy of Mclust with conventional approach is the best among the others clustering algorithms because its consistency. The results have a moderate accuracy for Feature B (Intrusive Rocks) and Feature C (Sedimentary 1), in which others clustering have a low accuracy. The relative low accuracy for Feature B is due to confusion between Volcanic and Intrusive Rocks as explained in previous paragraph. For Feature C, this might be due to the fact that its areas are small and surrounded by Volcanic Rocks thus its radiometric signature become unclear.

Mclust accuracies are higher about 15% than those of PAM clustering as shown in *table 3-2*. The producer's accuracy of Mclust is relatively higher for most lithologies than that of PAM clustering. Its low accuracy due to the algorithm does not find a substantial structure in the data. When no a substantial structure finds, the results is becoming less representative to be use for explaining the processes/phenomena in the area where the data are taken.

Regarding the data preparation approach, CoDa approach does not produce better classification results compared to the conventional one (*table 3-2*). Classification accuracy based on CoDa is lower than that based on the conventional one with more 15% and 25% differences for overall accuracy and kappa

coefficient, respectively. In addition, *figure 3-6* a diagram of producer's accuracies, shows that the conventional approach produces better results than CoDa. It might be due to the fact that total concentrations of K, eTh, and eU are only less than 5%, as explained by Pawlowsky-Glahn and Egozcue (2006) that closure problems arise when the sum of data per sample approaches the constant k (k = 100% or k = 1,000,000 ppm in this case). Therefore, CoDa approach likely is not appropriate to be applied in this type of data.

The moderate to low level accuracy of clustering results might be influenced by input data and the field condition. The input data are not primary data but were already in grid format, and we know that gridded data are interpolated data. Consequently, interpolation errors contribute to the total error of the assessment. According to field conditions, the area is heavily weathered and covered by a thick layer of sand and gravel left behind by glaciers (Geoscience BC, 2010). Consequently, the boundaries from clustering results have a low accuracy compared to the reference existing lithological map.

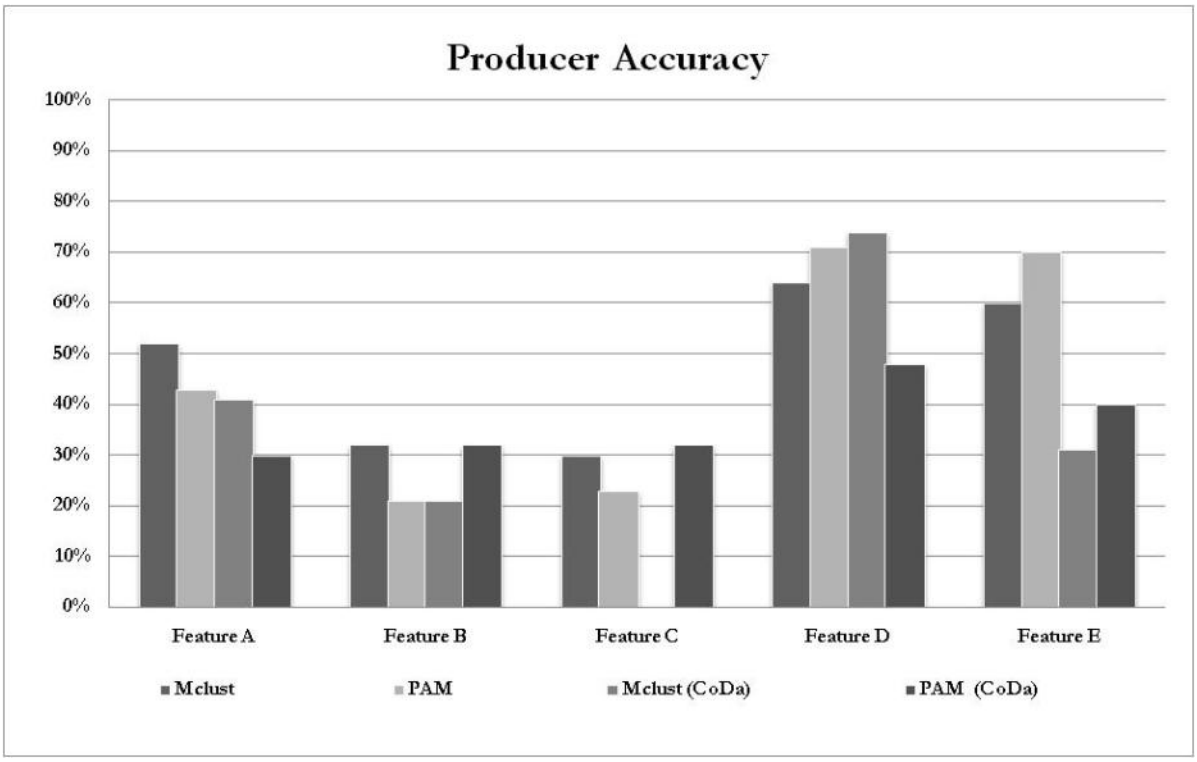


Figure 3-6 Producer's accuracy diagram from assessment of clustering results for airborne gamma-ray data (for error matrices see *appendix 3-2*); Feature A, Volcanic Rocks; Feature B, Intrusive Rocks; Feature C, Sedimentary Rocks 1; Feature D, Sedimentary Rocks 2; Feature E, Lake (for explanations about the lithology at the study area see *section 1.6.2*)

Table 3-2 Data quality of clustering results for airborne gamma-ray data

Clustering	Overall Accuracy	Kappa Coefficient
Mclust	50%	0.36
PAM	45%	0.30
Mclust (CoDa)	42%	0.26
PAM (CoDa)	35%	0.17

3.5. Concluding remarks

- The clustering method could be used to recognize lithologies with distinctive characteristics and/or large homogenous areas. Lithological units with distinctive characteristics, e.g. high concentration in gamma-ray elements such as Chuchi Syenite formation, are prominent in almost all clustering algorithm. Furthermore, a large homogenous area such as Sedimentary Rock 2 also could be identified (see *figure 3-6*).
- CoDa approach in data preparation likely is not a proper technique to be applied in airborne gamma-ray that shown by its lower accuracies (see *table 3-2*).This might be because the closure problems are insignificant in the data.
- Classification qualities of Mclust to airborne gamma-ray data are better than these of PAM clustering with the differences up to 7% and 0.09 for overall accuracy and kappa coefficient, respectively.

4. AIRBORNE MAGNETIC DATA ANALYSIS

4.1. Introduction

In this chapter, edge detection methods – rotation-variant template matching (RTM) and clustering-based for edge detection (CED) – were applied to identify lithology boundaries from airborne magnetic data. The methods were chosen rather than region-based segmentation due to characteristics of magnetic data. Magnetic data have non uniform value rather than gradual value even within one lithology. Thus edge detection is more suitable because the methods detect abrupt change in the values not their similarity such in region-based segmentation. Furthermore, before application of these techniques, two common transformations in analyzing magnetic data – reduce-to-pole (RTP) and analytic signal (AS) – were applied as data preparation methods. The results were used to interpret lithological boundaries. The interpretations are to be used in *chapter 5* for assessing classification accuracy.

4.2. Descriptions of the Airborne Magnetic Datasets

Airborne magnetic data (residual total magnetic intensity) were obtained from the Canadian Aeromagnetic Database maintained by the Geological Survey of Canada (Natural Resources Canada, 2010a). The data were obtained from the airborne surveys using the flight-line orientations that were approximately perpendicular to the regional geological strike in constant line spacing. The aircraft flew with 305 m of terrain clearance and 800 m of flight line spacing. The data were available in grid form with 200 m grid size with coordinate system in WGS 1984 UTM zone 10 N and datum of WGS 1984 datum. The residual total magnetic intensity was obtained by subtracting from the observed total field the International Geomagnetic Reference Field (IGRF) model (Natural Resources Canada, 2010a).

4.3. Methodology

Two techniques of edge detection were applied to the airborne magnetic data in order to identify boundaries between lithological units. Rotation-variant template matching (RTM) algorithm developed by (Van der Werff et al., 2007) and clustering-based for edge detection (CED) algorithm developed by (Dinh et al., 2009) were applied to the airborne magnetic data after some data preparation processes were performed to these data (*figure 4-1*). RTM algorithm was applied to a subtraction image whereas CED was applied to the horizontal derivative image. It is because the two algorithms have a difference in defining the edge. RTM defines the edge as an area among the areas which have at least two contrasting characteristics e.g. spectral, texture whereas CED define as an area where there are abrupt changes of the characteristics. RTM was applied using ENVI software with *ESA Tools* add-ons which developed by Department of Earth System Analysis-ITC, whereas ***k***-means clustering, the last part of CED algorithm, was applied using *R* statistic software with *stats* package. In addition, the rest of the processes such as in data preparation were conducted using *Oasis Montaj 7.2.1*

4.3.1. Data preparation

For comparison, two transformations – reduce to pole (RTP) and analytic signal (AS) – were applied to the airborne magnetic data before the application of the edge detection techniques. As it is shown in *figure 4-1*, after RTP or AS transformation, upward continuation (UC) was applied to each image. The last process in preparation stage is raster/grid subtraction of RTP and AS image by their UC. The results of their subtraction are so-called RTP subtraction (RTPSub) and AS subtraction images (ASSub), respectively. Subtraction images are another approach to separate between information derived from depth surface and near-surface. As we know that UC transformation tries to attenuate high-frequency signals than low-frequency signal. The high frequency is mostly derived from shallow source (Milligan and Gunn, 1997). Therefore, subtraction images describe more shallow surface anomalies signal.

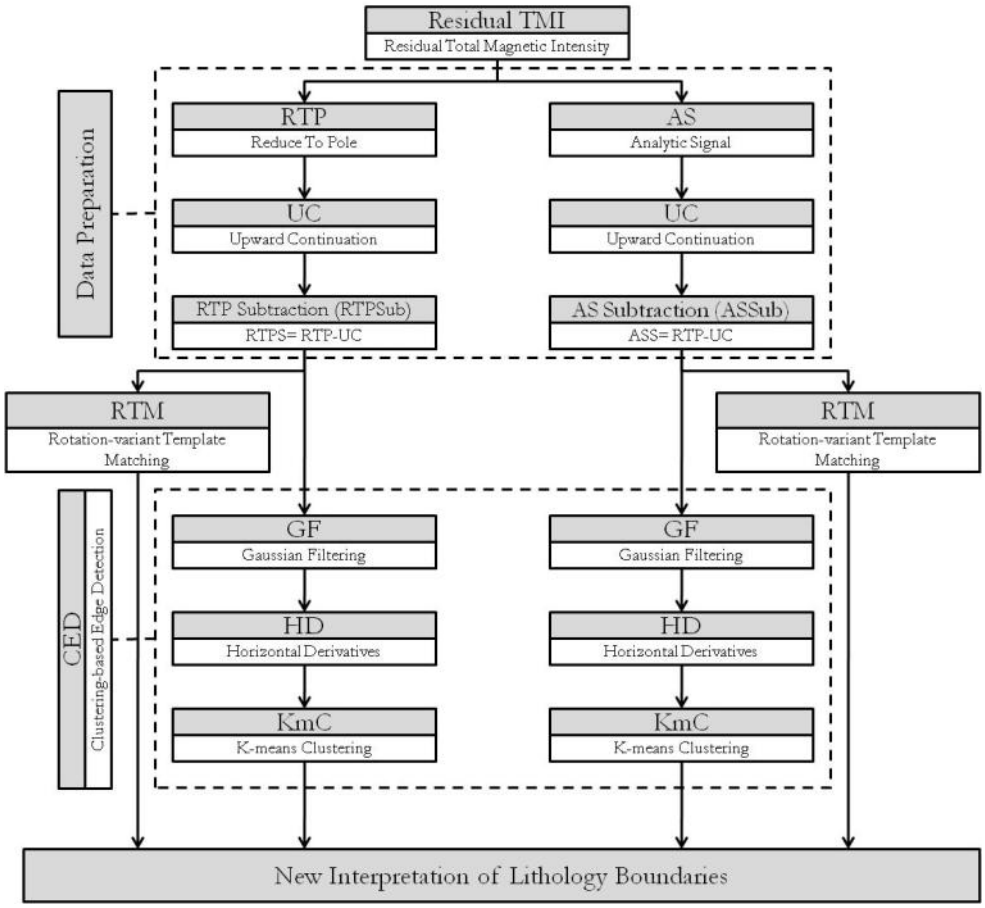


Figure 4-1 Flow chart of the methodology with two different transformations, RTP and AS; and techniques, RTM and CED

4.3.1.1. Reduce To Pole (RTP)

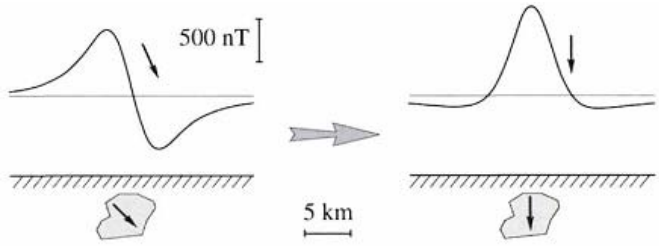


Figure 4-2 A magnetic anomaly profile and its relation to geological features, before and after being reduced to the pole (Blakely, 1996)

Milligan and Gunn (1997) defined reduce-to-pole (RTP) as the process to convert the magnetic field from a magnetic at particular latitude to the field at a magnetic pole, where the inducing field is vertical. At a magnetic latitude, magnetic anomalies due to induction have asymmetric form (skewed form) related to their sources, but when the inducing field is vertical the induced anomalies are directly over their sources as shown in *figure 4-2 (Blakely, 1996)*.

4.3.1.2. Analytic Signal (AS)

AS is a linear combination of first derivatives of magnetic anomaly in horizontal and vertical directions (Roest et al., 1992). The AS has a form over causative bodies that depend on the locations of the bodies but not on their directions of magnetization. This means that all bodies with the same geometry have the same analytic signal. The analytic signal peaks over the edges of source bodies of positive and negative anomalies and thus can be used to determine edge locations (Milligan and Gunn, 1997).

4.3.1.3. Upward Continuation (UC)

Upward continuation transforms the potential field measures on one surface to the field that would be measured on another surface farther from all sources. This transformation attenuates anomalies with

respect to wavelength; the shorter the wavelength the greater the attenuation. This upward continuation tends to emphasize anomalies caused by deep sources at and to reduce of anomalies caused by shallow sources (Milligan and Gunn, 1997).

4.3.1.4. Subtraction Image

Subtraction image describes signal of magnetic anomaly from shallow sources because it is a result of subtraction between responses signals from all-depth sources anomalies and these of deep-source anomalies. Responses signals from all-depth sources anomalies represent by transformed of residual total magnetic intensity (RTP and AS image) whereas responses signals from deep-source anomalies represent by UC images. Therefore, there are two subtraction images, one is RTP transformed subtraction image (RTPSub) and another is AS transformed subtraction image (ASSub).

4.3.2. Rotation-variant Template Matching (RTM)

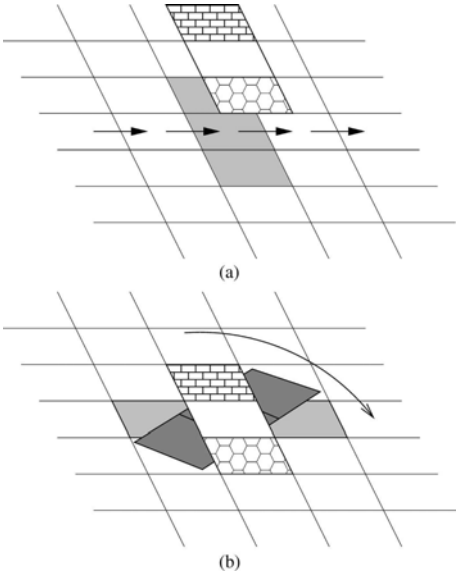


Figure 4-3 Template is matched by (a) moving it over an image and (b) changing the template orientation at every position by 45° increments up to a total of eight orientations (Van der Werff et al., 2007)

RTM algorithm was applied to the subtraction image both for RTP transformation (RTPSub) and AS transformation (ASSub). The first step to apply the method is determining region of interests (ROIs) to assign into template. The ROIs consist two areas considered representatives of different lithologies. Van der Werff et al. (2007) explained that the template is a miniature image which consists of 3x1 pixels. The template contains information about a boundary between two contrast pixels. These two contrast pixels represent two different lithologies. This template is moved over the image like a moving kernel. At every position, the template is rotated, and a statistical fit is calculated for every pose (*Figure 4-3*).

The next steps are choosing one result from 12 results which produced by the algorithm. Then, the threshold value is determined base on natural breaks using Jenk's optimization (Jenks, 1967). The threshold divides the data into edge and non-edge pixel. The algorithm, in principle, tries to minimize the variance within the classes and maximize and maximize the variance between the classes. The natural breaks technique was used when the data is not evenly distributed and not heavily skewed toward one end of the distribution.

4.3.3. Clustering-based Edge Detection (CED)

Clustering-based edge detection (CED) algorithm from Dinh et al. (2009) for edge detection was applied to the airborne magnetic data. The algorithm consists of several steps as follow: first, Gaussian blur convolution was applied to the subtraction image (RTPSub or ASSub) to reduce the noise in order to avoid local anomalies when horizontal derivative was applied. The horizontal derivate was employed to get gradient image in lateral direction. The step was needed due to the algorithm is looking for sudden change of value. The next step is applying ***k***-means clustering to the gradient images. Dinh et al. (2009) suggested cluster number four to eight to be tried as an input in k-means clustering. Whereas two clusters is not recommended because the pixel values tend become non-edge cluster thus there is possibility to loss edge clusters. The next step is distinguishing clusters into edge and non-edge cluster base on pixel number of cluster members. Two clusters with the highest pixel number were classified into non-edge cluster and the rest were into edge cluster.

4.4. Results and discussion

Figure 4-4 and *figure 4-5* show the results of data preparation for RTP and AS transformation part, respectively. *Figure 4-4(a)* is a result of RTP transformation with 20.3° and 74.9° for geomagnetic inclination and declination, respectively. The values of inclination and declination were generated by *Oasis Montaj 7.2.1* with latitude and longitude of the area as well as the time of data acquisition as the inputs. In addition, *figure 4-5(a)* is a result of AS transformation. In both images, the features have the same pattern which is north-west to south-east trending. From both images, the area with high values in RTP is wider than these in AS. This is might be due to remnant magnetization in the area. As explained by Milligan and Gunn (1997), the remnant magnetization could smear the anomaly.

Figure 4-4 (b) and *figure 4-5(b)* are the results of upward continuation at an equivalent plane level of 500 m. Both images have smoother images compare to their parents (*figure 4-4(a)* and *figure 4-5(a)*). It is because UC transformation smooths out short-wavelength anomalies relative to long-wavelength. The same as their parent, UC from RTP have wider area in high value than that from AS for the same reason.

The last two images in *figure 4-4* and *figure 4-5* are subtraction images and horizontal derivatives or horizontal gradient images. These subtraction images depict the anomaly from shallow sources at the area. *Figure 4-4(c*), RTPsub, shows wider area for high value than *figure 4-5(c)*, ASSub. This is because the influence of their parent images, which have the same phenomena, is still remained. Both these images were used as input in RTM algorithm. Furthermore, *figure 4-4(d)* and *figure 4-5(d)* are horizontal derivative of subtract images that show change intensity of subtract images value. The horizontal derivatives of RTP transformation images show more continuous features than that of AS. However, the AS one has clearer separation between the high value area and its surrounded. In addition, CED algorithm was performed to these images.

Figure 4-6 show the results of application edge detection method to airborne magnetic data. *Figure 4-6 (a)* and *figure 4-6 (b)* are the result of Frank Ratio 2 image of RTM algorithm for RTP and AS transformed part, respectively. The images have been classified using natural break optimization of Jenk's algorithm in *ArcGIS 10*. Frank ratio was chosen because the edges are seen the most clearly in the images. The Frank Ratio 2 is a ratio between rotation variance in mean spectral fit and rotation average spectra variance. The rotation variance in mean spectral fit could be used to indicate existence of crisp boundaries. In addition, the rotation average spectra variance have high value when at least one of the spectral signatures is found and forms a crisp boundary (Van der Werff et al.,2007). Thus, the ratio could be used to indicate crisp boundaries.

In general, all results produce the same edge patterns trend, north-west to south-east trend (*figure 4-6*). This trend is the same as general trend of lithologies and lineament in existing lithological map. However, results of CED are better than RTM because RTM creates multiple responds for a single edge for some parts. Between CED for RTP transformed and AS, it is difficult to assess which one is better. Nevertheless, the image of AS is more reliable when input data, *figure 4-4 (a)* and *figure 4-5 (a)*, are included as consideration. Therefore, this image was chosen for updating the lithological map.

The edge pixels can be distinguished as lithological edge and non-lithological edge that reflect lithological boundaries and lineament signatures, respectively (*figure 4-7)*. The lithological edge is defined as the most outside pixels of clustered edge pixels thus edge pixels within cluster and individual edge pixels are classified to non-lithological edge. From *figure 4-7*, clustered edge pixels, thus as well as the lithological edges, are only related with the particular intrusive rocks, the Hogem Batholiths. Therefore, the lithological could be used to update of Hogem Batholith boundaries as shown in *figure 4-7 (a)*. The clustered edge pixel in Hogem Batholith might be related to the complex processes of its formation. As explained by Nelson et al. (1992), Hogem Batholith occurred in at least three phases that each create complex features. Thus, to describe its complexity, Hogem Batholith is called as an intrusive complex.

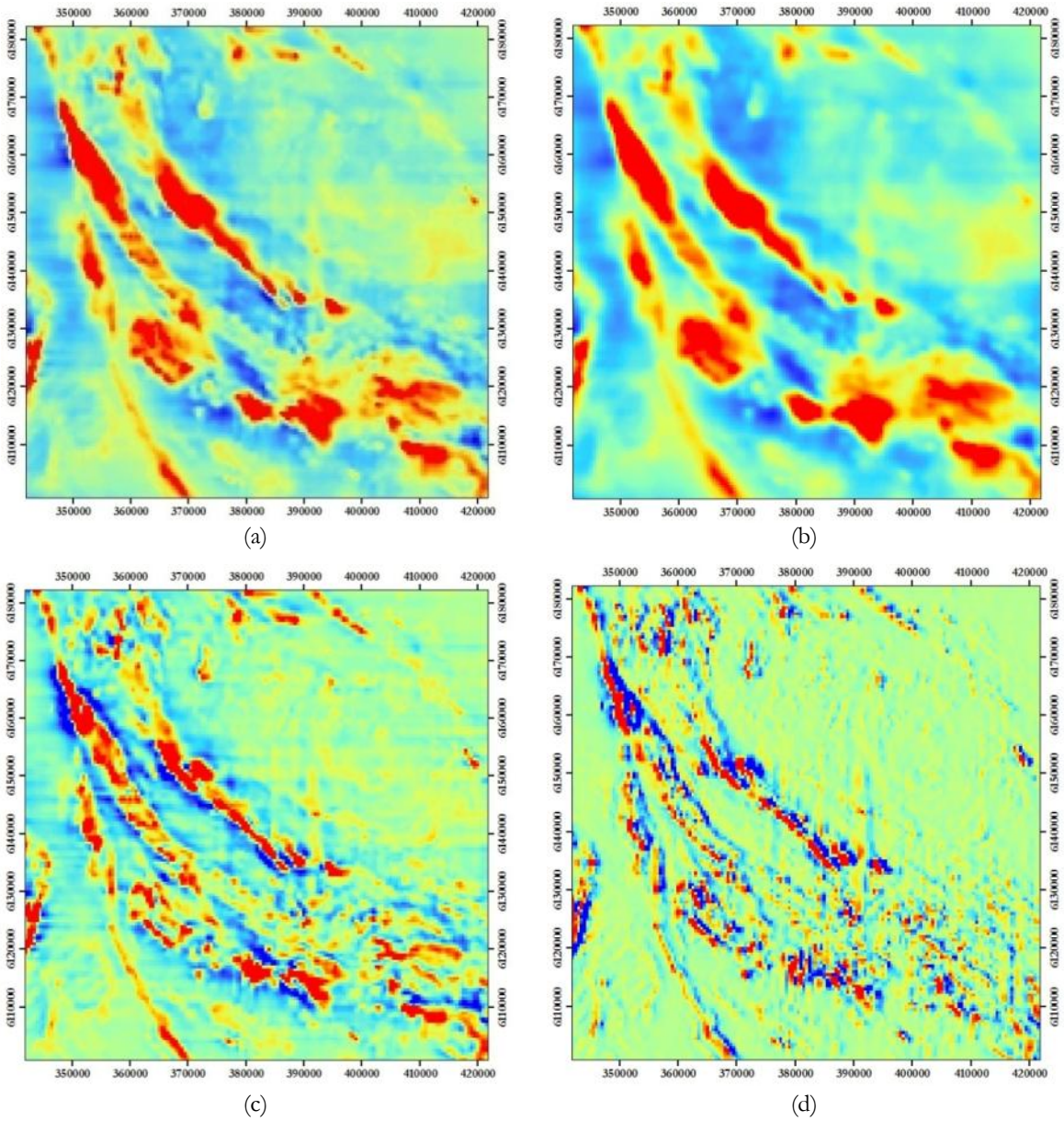


Figure 4-4 Results of processing of airborne magnetic data; (a) reduced-to-pole (RTP), I=20.3°, D=74.9°; (b) upward continuation of RTP at 500 m (UC); (c) RTP subtraction image (RTPSub), RTPS=RTP-UC; (d) horizontal derivative of RTP subtraction image

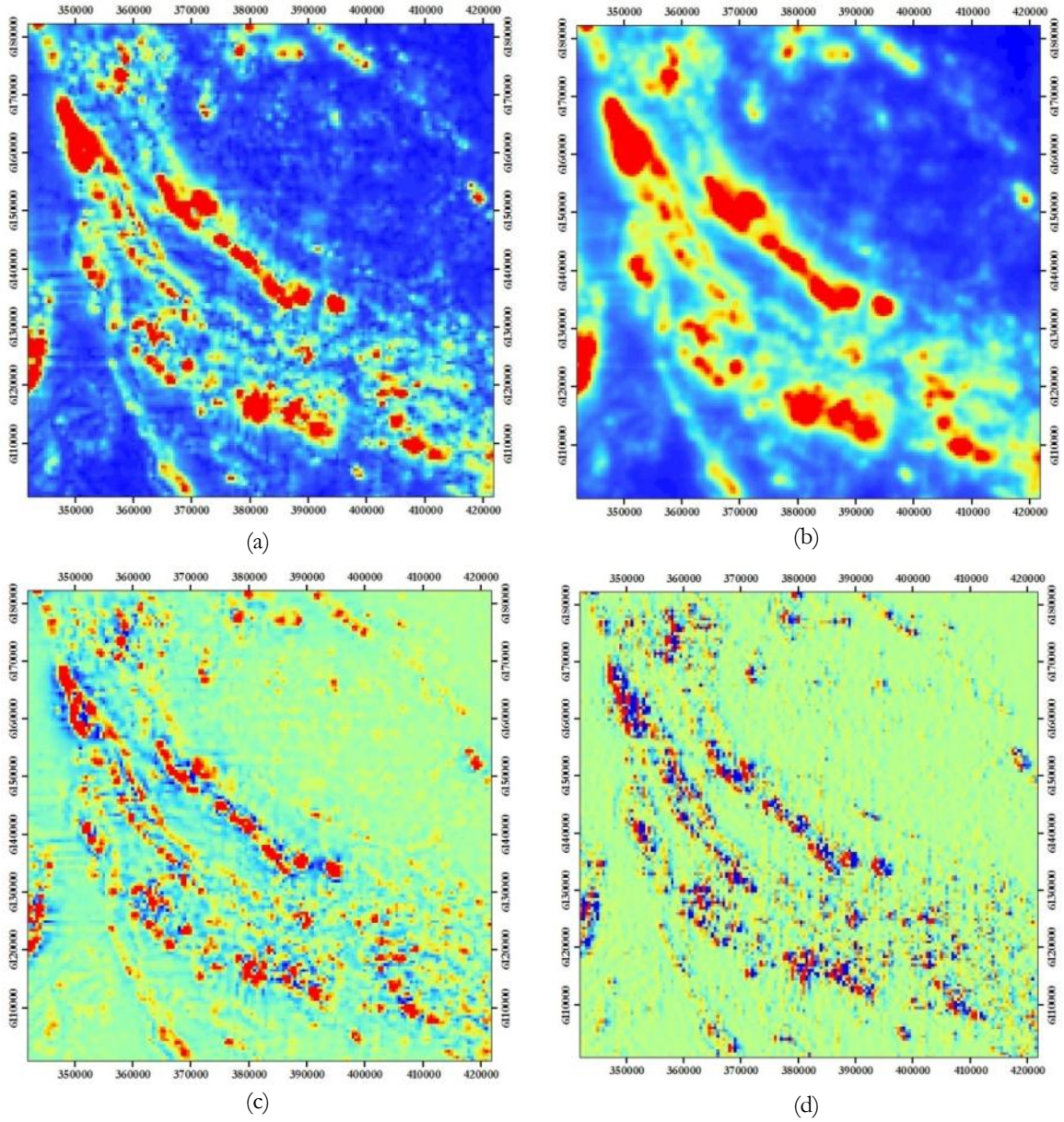


Figure 4-5 Results of processing of airborne magnetic data; (a) analytic signal (AS), (b) upward continuation of AS at 500 m (UC), (c) AS subtraction image (ASSub), ASS=AS-UC, (d) horizontal derivative of AS subtraction image

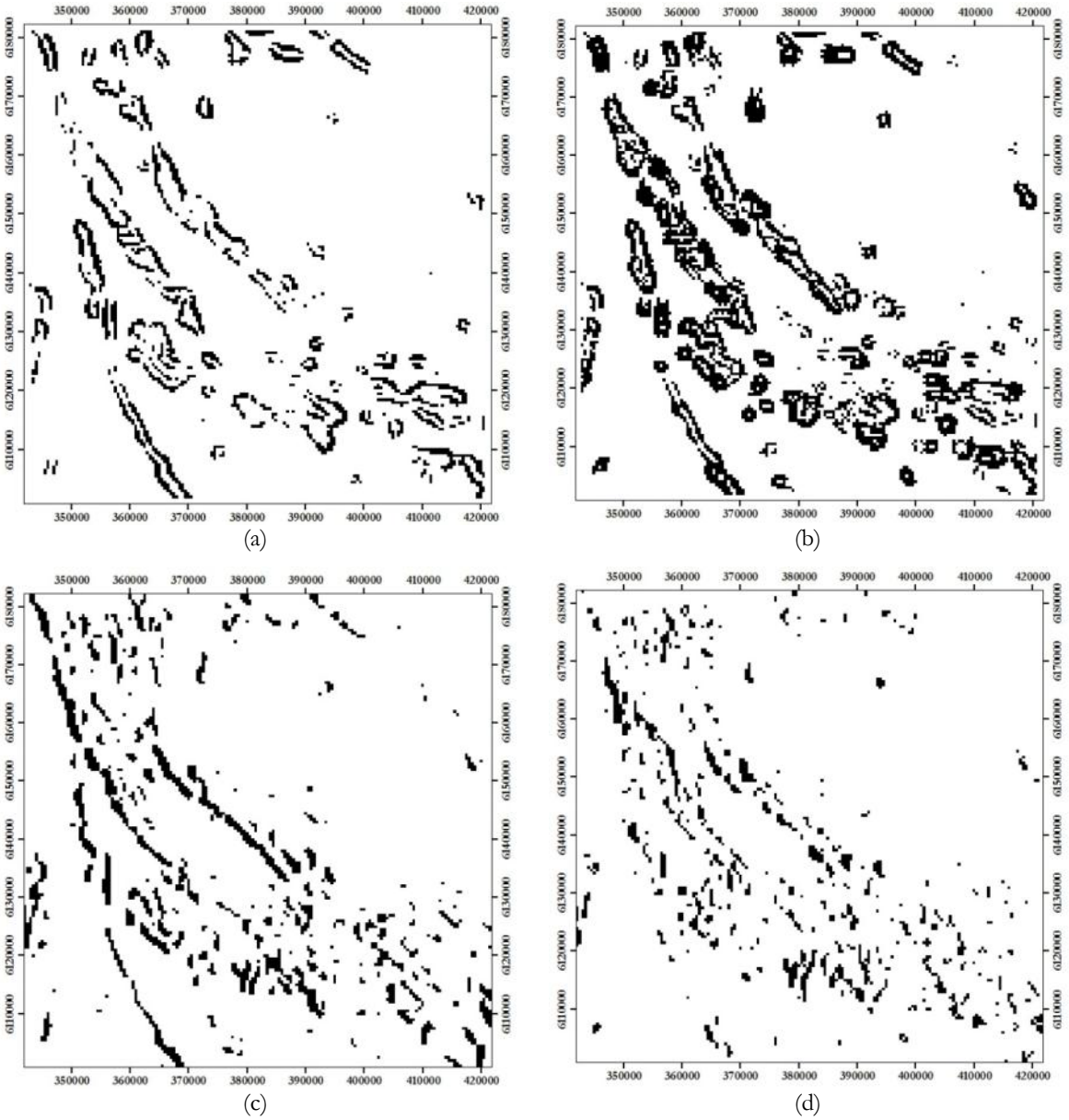
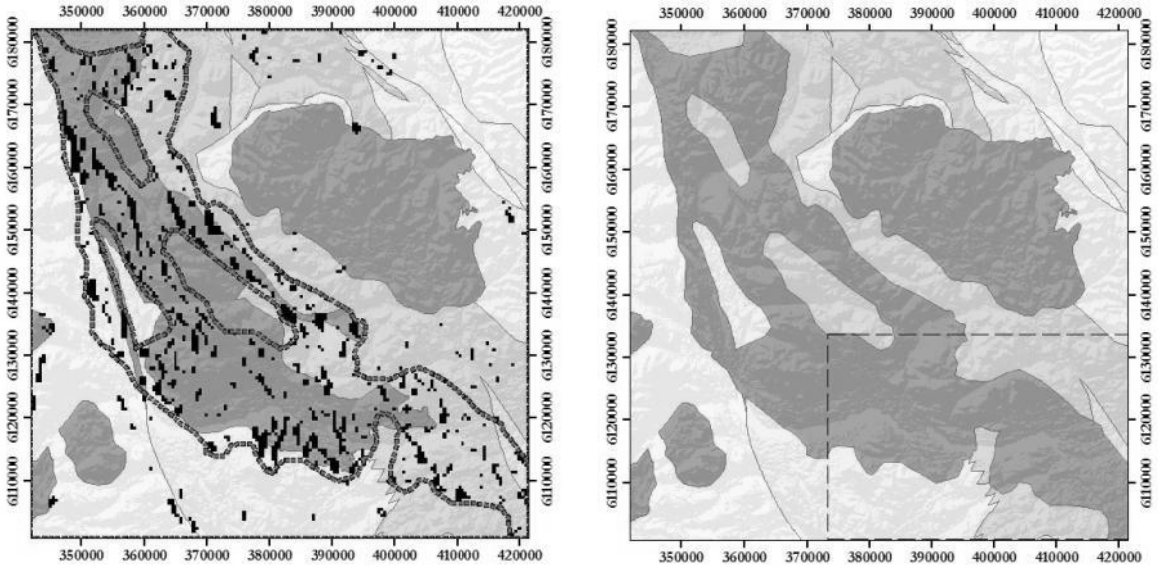


Figure 4-6 Images of edge detection technique results, (a) result for RTM technique with RTP transformation, (b) result for RTM technique with AS transformation, (c) result for CED technique with RTP transformation, (d) result for CED technique with AS transformation



(a) (b)

Figure 4-7 Interpretation of clustering-based for edge detection (CED) result base on analytic signal (AS) transformed horizontal derivatives images, (a) an existing lithological map is underlying to the interpretation result; dash-thick-lines are boundaries of the new interpretation, (b) the final results of new interpretation lithological map; dash-line box is the area for assessing classification accuracy in *chapter 5*

4.5. Concluding remarks

- The application of AS to residual total magnetic intensity data in data preparation before the application of edge detection methods is more reliable than RTP due the fact that AS is more robust to the existence of remnant magnetization.
- CED could become an alternative unsupervised technique for classifying airborne magnetic data to indicate the edge features such as lithological boundaries and lineaments.

5. STREAM SEDIMENT GEOCHEMICAL AND GEOPHYSICAL DATA INTEGRATION

5.1. Introduction

In this chapter, stream sediment geochemical (SSG) data and airborne gamma-ray (AGR) data were integrated by using them as inputs in two different clustering methods – model-based clustering (Mclust) and partition around medoids (PAM) – for unsupervised lithological classification. In the assessment stage, not only the existing lithological map was used to validate the clustering results but also the lithological map based on the interpretation of airborne magnetic data in *chapter 4*. In addition, accuracies of clustering the SSG data (*chapter 2*) and the AGR data (*chapter 3*) were compared to the accuracies of clustering the SSG and AGR data together.

5.2. Datasets for integration study

The SSG data consist of 2,194 sampling points for 13 elements, from which nine elements have been selected using the BIC (Bayesian Information Criterion) in order to get the optimum number of clusters (see *section 2.3.3.1*). The 13 elements are As, Zn, Cr, Ce, Ba, Sb, Fe, Co, Ni, Cu, Mn, Hg, and Pb; the first nine elements were the ones selected using the BIC. The stream sediment sampling density varied from about 1 sample per 8 km^2 to 1 sample per 13 km^2 (Jackaman and Balfour, 2008).

Figure 5-1 shows the study area for the integration of SSG and AGR data. This area is the same as study area for the AGR data only (see *figure 1-2* and *chapter 3*). For SSG data preparation, all the stream sediment sample data (see *figure 2-4* for location map) were used in data processing such as interpolation. Then, in integration phase the processed SSG data were subset to the study area for the analysis described in this chapter (*figure 5-1*). Furthermore, two lithological maps were used for validation of the results. One map (*figure 5-1(a)*) is the existing geological map that that was used for validating the clustering results using the AGR data (see *chapter 3*). The other map (*figure 5-1(b)*) is the result of analysis and interpretation of the airborne magnetic data (*see figure 4-7 (b)*).

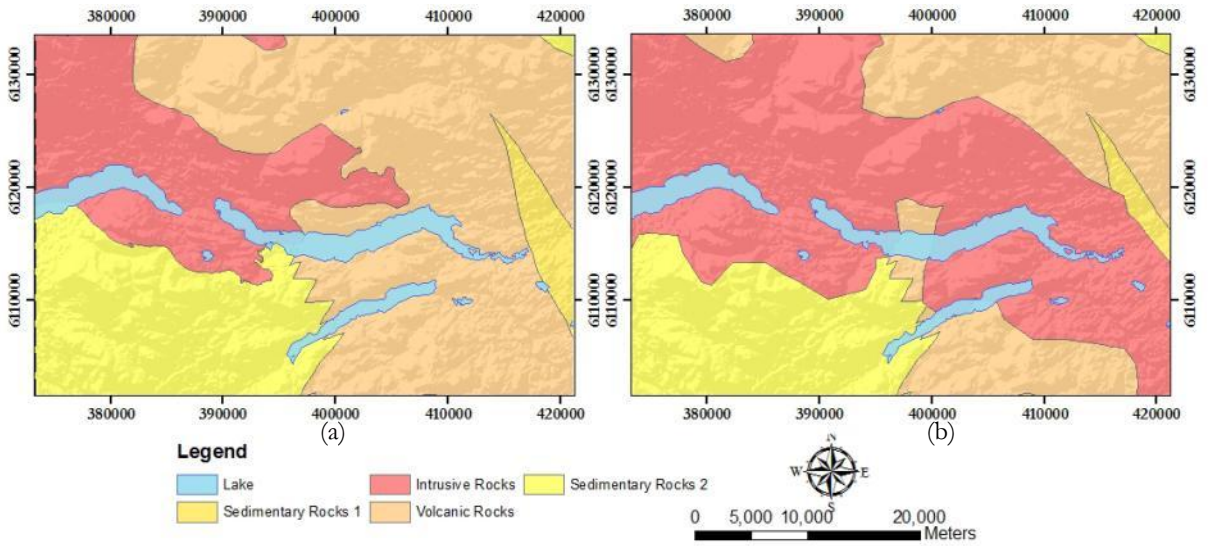


Figure 5-1 The study area and maps used for validation of results: (a) existing lithological map; (b) lithological map based on the interpretation of airborne magnetic data, see *figure 4-7 (b)* (modified from Massey et al. (2005a, b, c, d))

5.3. Methodology

The integration of SSG and AGR data consist of four stages, namely: data preparation; data integration using clustering method; post clustering; and assessment (*figure 5-2*). Data preparation comprises of data transformation and standardization of SSG data, and then their interpolation in order to convert the point SSG data into a continuous data layer like the AGR data. The interpolation of the SSG data was performed using *ArcGIS 10*, whereas variogram analysis before interpolation and clustering analyses were executed using *R* statistical software. Like in *chapter 2* and *chapter 3*, two clustering algorithms – Mclust and PAM – were performed to integrate the SSG and AGR datasets. Post clustering processes such as reclassification and filtering were employed before the assessments. From the error matrix, overall and producer's accuracy and kappa coefficient were calculated to assess the classification results. The assessments were conducted using the two reference maps shown in *figure 5-1*.

5.3.1. Data preparation

To investigate the spatial data structure of the geochemical data for each element, variogram analyses were performed (*appendix 5-1*). The variogram models and the point geochemical data were used to interpolate unsampled areas using universal kriging. As explained by Carranza (2010), interpolation is a plausible technique to make continuous geochemical landscapes from stream sediment geochemical data in regional scale. Furthermore, Robinson et al. (2004) stated that universal kriging is an appropriate technique to interpolate stream sediment geochemical data in regional scale.

Data transformation and standardization were then applied to the SSG data before clustering techniques were applied. Base-10 logarithmic and square root transformation were applied to the SSG and AGR data, respectively, to get symmetric shape in data distribution. The transformations were chosen according to the skewness values of individual datasets as explained in *chapter 2* and *chapter 3*. In addition, the standardization using median-MAD (*equation 2-10*) was applied to both datasets to make the range of the data comparable.

5.3.2. Clustering

Mclust and PAM clustering methods were applied to stacked raster images of SSG and AGR data. As experiments besides using all 13 elements in the SSG data, clustering was also applied to the nine selected elements in the SSG data. Thus, there are two data integration results: Integrated Data I is from combination of SSG all elements with AGR elements and Integrated Data II is from combination of SSG selected elements with AGR elements.

5.3.3. Post clustering

For the post clustering stage, reclassification and filtering were applied to images of clustering results. The images were reclassified into five classes representing five features as in the reference maps (*figure 5-1*). Furthermore, before the assessments, majority filtering was applied to the reclassified images.

5.3.4. Assessments

Two lithological maps, *figure 5-1(a)* and *figure 5-1(b)*, were used to validate the images resulting from the clustering analyses. The map in *figure 5-1(a)*, the existing lithological map, was designated as Reference Data I, whereas the map in *figure 5-1(b)*, derived from the interpretation of airborne magnetic data in *chapter 4*, was designated as Reference Data II. This latter reference data were used to test if interpreted lithological boundaries from analysis of airborne magnetic are consistent with results of clustering the SSG and AGR data together. If assessment results using Reference Data II are better than those using Reference Data I, it means that the lithological boundaries interpreted from the airborne magnetic data might be better than those portrayed in the existing lithological map. Furthermore, overall and producer's accuracy from error matrix and kappa coefficients were used to assess the quality of the classification results.

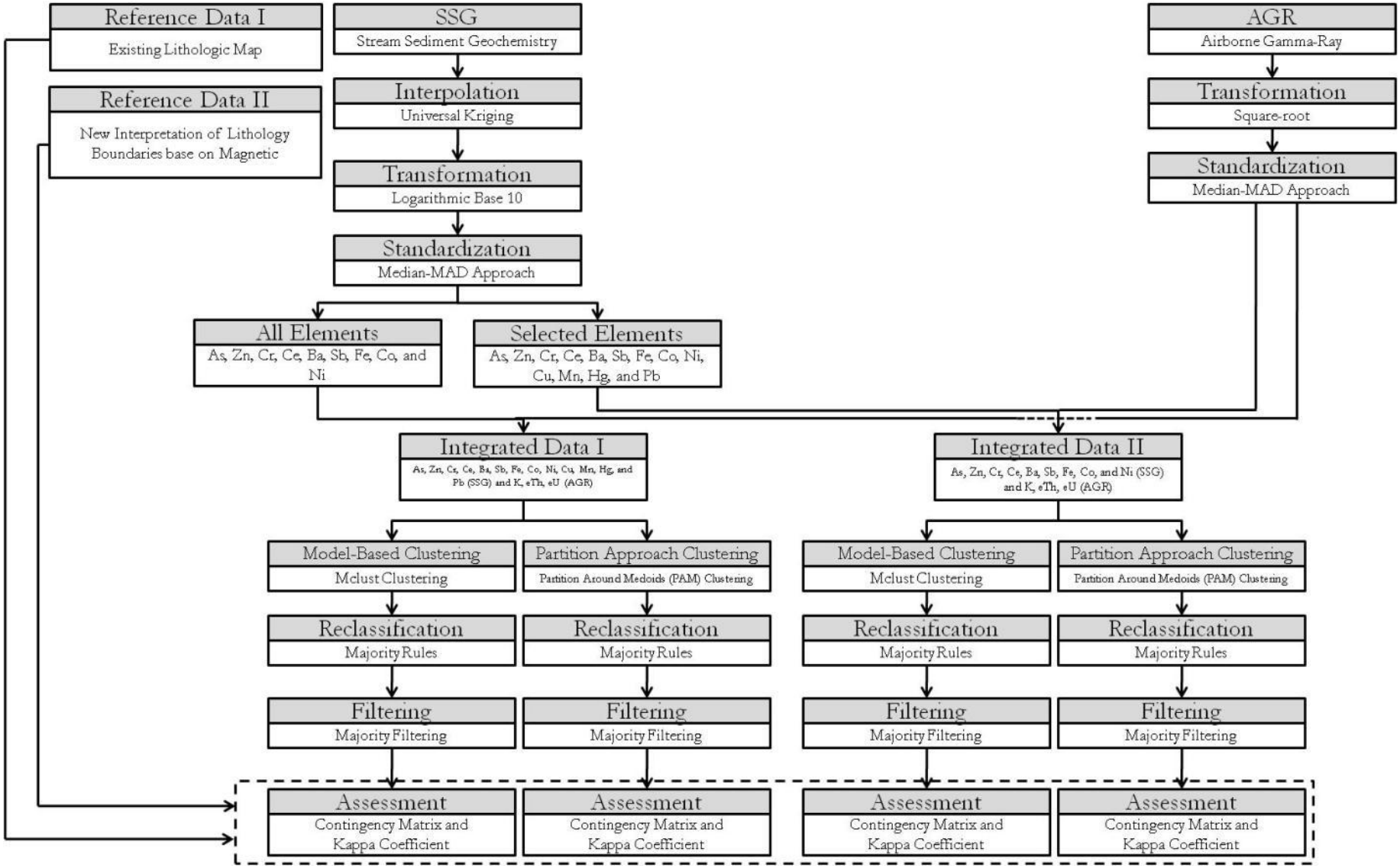


Figure 5-2 Flow chart of the methodology to integrate stream sediment geochemical and airborne gamma-ray data using two types of clustering algorithm, Model-based Clustering (Mclust) and Partition Around Medoids (PAM) clustering.

5.4. Results and discussion

5.4.1. Interpolated images

Figure 5-3 and *table 5-1*, respectively, show the variogram model for Zn and summary of variogram components for 13 elements in the SSG dataset. A model was chosen based on the smallest root mean square values from three variogram models which had been tried, exponential, gaussian and spherical. The variogram model show spatial data structure of the element at the study area. All the elements data fit with the exponential model that means the data have high variance change within small distance. The range values, the last column of *table 5-1*, show maximum distance at which the samples still have spatial correlation. Therefore, the models are reliable to be used due to the fact the smallest distance between the samples at the study area are less than 500 m. The variogram models for the other elements in the SSG dataset are given in *appendix 5-1*. In addition, *figure 5-4* shows the spatial distribution of Zn as a result of universal kriging interpolation; whereas maps of spatial distributions of the other elements are shown in *appendix 5-2*.

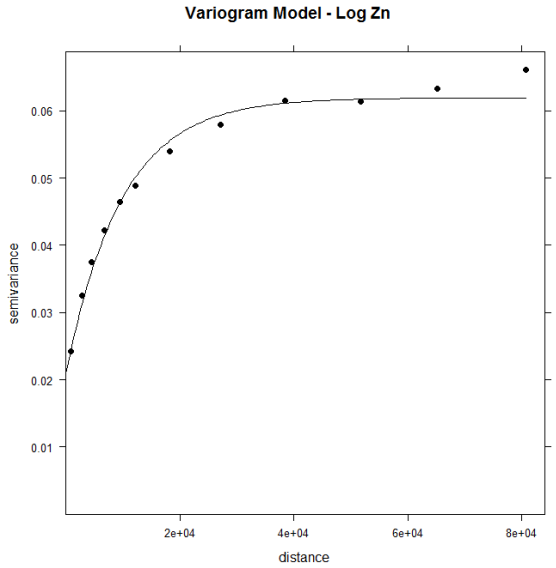


Figure 5-3 Exponential variogram model for logarithmic base-10 transformed Zn data

Table 5-1 Summary of variogram components of individual elements in the SSG dataset

Element	Model	Nugget	Partial Sill	Range
Zn	Exponential	0.02	0.04	9,791
Cu	Exponential	0.02	0.10	13,795
Pb	Exponential	0.05	0.09	15,692
Ni	Exponential	0.03	0.16	15,181
Co	Exponential	0.02	0.05	8,623
Ba	Exponential	0.03	0.04	8,988
Mn	Exponential	0.04	0.05	5,706
Fe	Exponential	0.02	0.03	9,188
Ce	Exponential	0.03	0.07	23,845
Cr	Exponential	0.05	0.15	14,717
Hg	Exponential	0.04	0.09	18,299
Sb	Exponential	0.03	0.09	14,135
As	Exponential	0.06	0.15	12,860

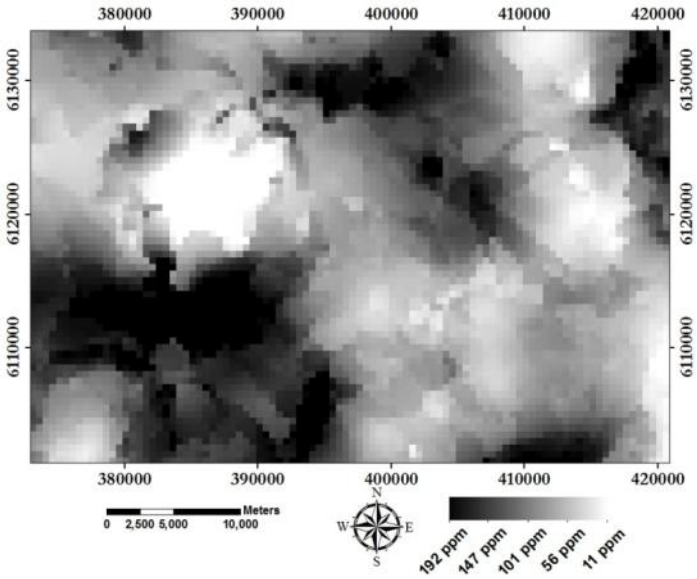


Figure 5-4 Image of spatial distribution of Zn as a result of universal kriging

5.4.2. Clustered images

Figure 5-5(a) is an image of Mclust results using SSG and AGR data together (Integrated Data I). The best model for the data base on BIC is an EEV-model with nine clusters. This model means that the nine clusters have equal volumes and shapes but vary in orientations of ellipsoidal distributions in feature space. In addition, the image depicts several homogenous clusters with clear separation between them.

For PAM clustering, the optimum cluster number for Mclust for the same data was taken as an input. It is because values of SC (Silhouette Coefficient; see *section 2.3.3.1*) in PAM clustering for cluster number (k) from 3 to 15 did not show an optimum cluster number but rather fluctuate at a constant range between 1.3 to 1.5. According to Kaufman and Rousseeuw (2005) when the SC value is below 0.25, it means that there is no substantial structure in the data as shown in *appendix 2-2*. When there is no substantial in the data, the data using this technique become less reliable to be used to explain the processes/phenomena. Furthermore, *figure 5-5(b*), the result of PAM clustering using Integrated Data, shows several clusters with clear boundaries among them and several individual clusters within them. However, the PAM clustering image shows that lake boundaries could be detected quite well as can be seen in *figure 5-5(b)*.

Figure 5-5(c) and *figure 5-5(d)* are the images resulting from clustering using the nine selected elements in the SSG data and the AGR data (Integrated Data II). These images show similar patterns as their corresponding resulting from Integrated Data I. For Mclust, the same model as for Integrated Data I, an EEV-model with nine clusters, was also obtained. In addition, PAM clustering used the same cluster number from the Mclust result as an input in the process. The number was taken due to the difficulties in determining an optimum cluster number based on SC.

Comparing images resulting from Mclust, *figure 5-5(a)* and *figure 5-5(c)*, with those resulting from PAM clustering, *figure 5-5(b)* and *figure 5-5(d)*, reveals that the clusters of Mclust are larger and more homogeneous than those of PAM clusters. Furthermore, it is apparent that Mclust is a better technique than PAM clustering in distinguishing large features such as both sedimentary rocks. However, PAM clustering could detect small features, such as lakes, better.

Figure 5-6 and *figure 5-7*, which are results of reclassification and filtering, demonstrate that patterns of particular clusters are similar to patterns of particular feature in reference data. Both figures have similar patterns, except for Intrusive Rocks. It is due to the facts that *figure 5-7* using Reference Data II as a basis for reclassification. As explained in previous section, the Reference data II data is a lithological map base on airborne magnetic data analysis which updating new boundary for Intrusive Rocks. Furthermore, for example, in *figure 5-6(a)*, pattern of Cluster D is similar to that of the Sedimentary Rocks 2 in south-western parts of the areas. This pattern is found also in *figure 5-6(c)*, and *figure 5-6(d),* whereas for *figure 5-6(b)* the pattern is somewhat elongated to the east. Besides the pattern of Cluster D, in *figure 5-6(a),* the pattern of Cluster A is also similar to that of Intrusive Rocks in the existing lithological map. Moreover, among all images, *figure 5-5 (a)* has the clearest boundaries of lithological units, except for the lake features. Like in *figure 5-6*, in *figure 5-7*, the sedimentary rocks unit could also be recognized. The same as *figure 5-6(a)*, Mclust results for Integrated Data I, *figure 5-7(a)* could identify the Instrusive Rocks and give the most obvious boundaries among lithological units.

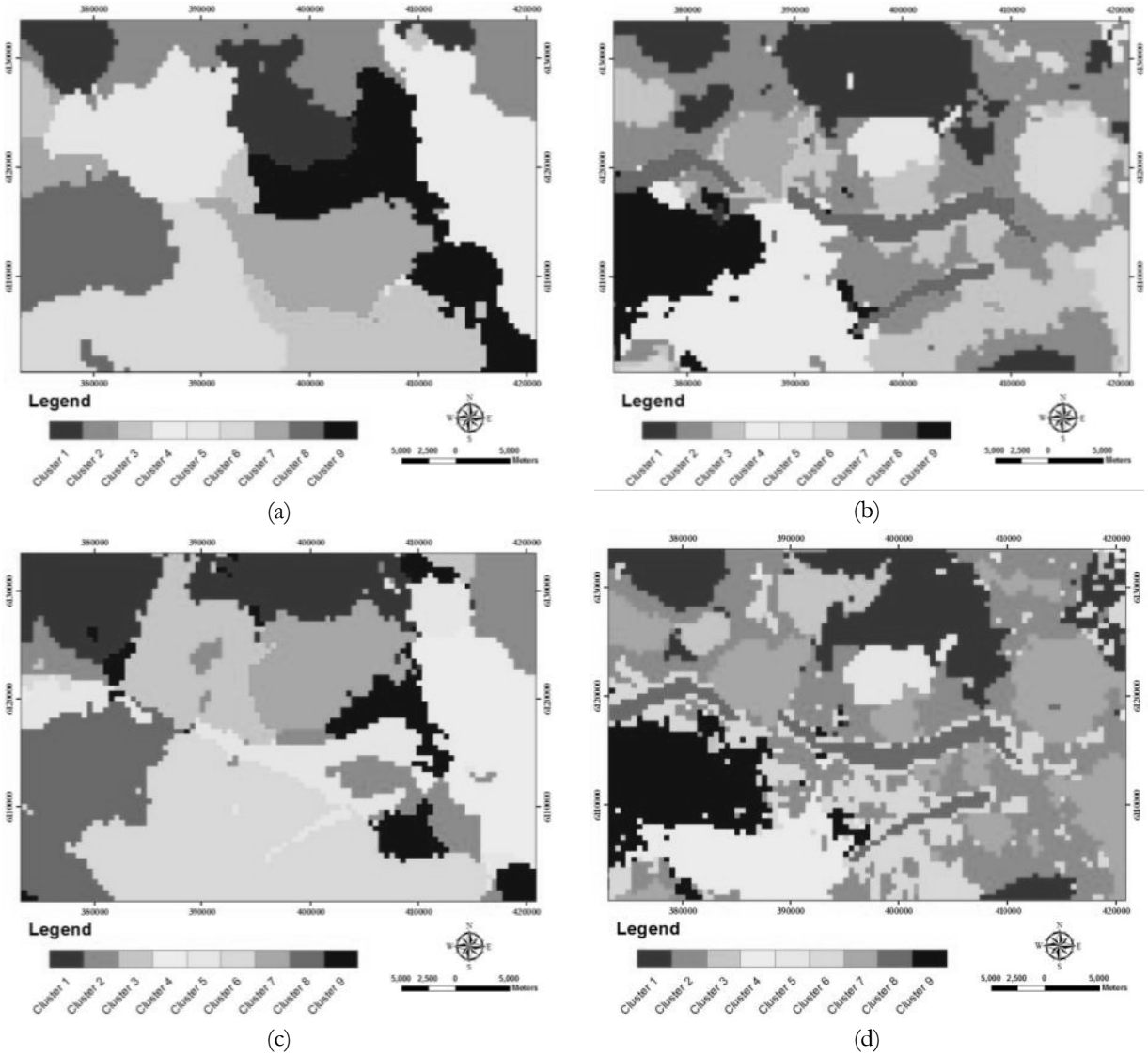


Figure 5-5 Clustering results image, (a) Mclust for all elements of stream sediment geochemical data and gamma-ray data integration, (b) PAM clustering for all element of stream sediment geochemical data and gamma-ray data integration, (c) Mclust for nine selected element of stream sediment geochemical and gamma-ray data integration, (d) PAM clustering for nine selected element of stream sediment geochemical and gamma-ray data integration

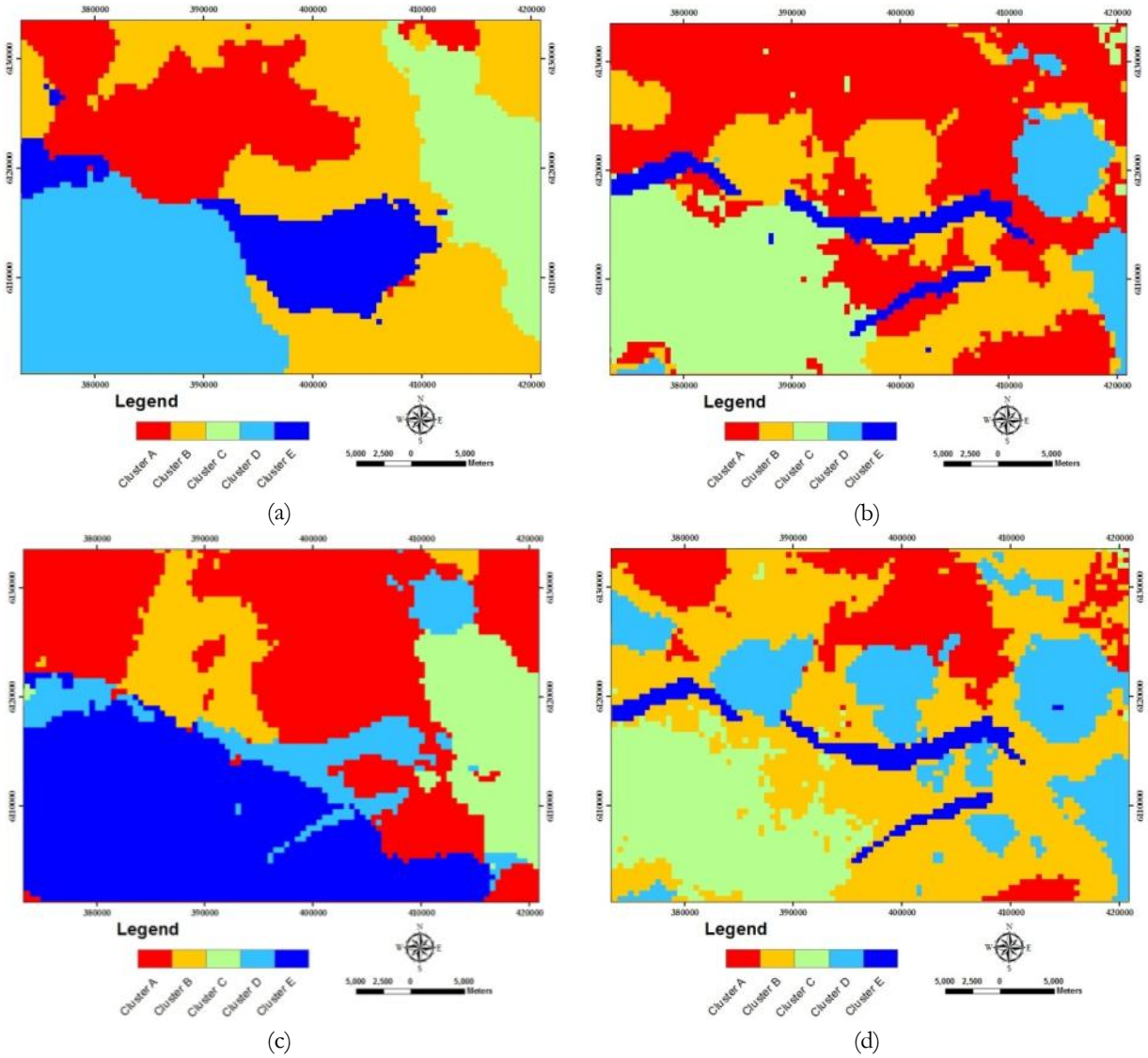


Figure 5-6 Images after reclassification and filtering using existing lithological map (Reference Data I), (a) Mclust for all element of stream sediment geochemical and gamma-ray data integration, (b) PAM clustering for all element of stream sediment geochemical and gamma-ray data integration, (c) Mclust for selected nine elements of stream sediment geochemical and gamma-ray data integration, (d) PAM clustering for nine selected element of stream sediment geochemical and gamma-ray data integration

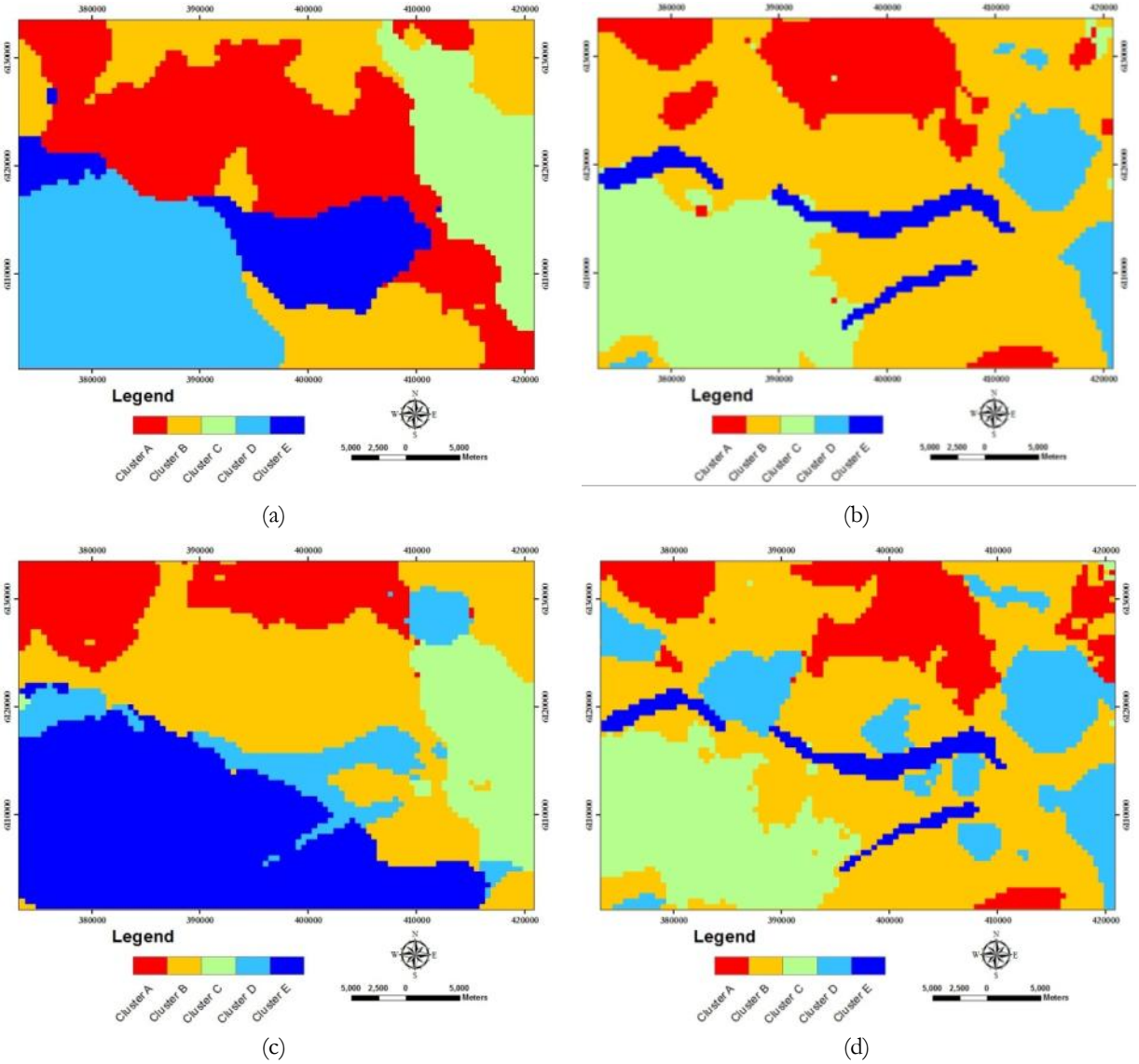


Figure 5-7 Images after reclassification and filtering using lithological map based on the interpretation of airborne magnetic data in *chapter 4* (Reference Data I), (a) Mclust for all element of stream sediment geochemical and gamma-ray data integration, (b) PAM clustering for all element of stream sediment geochemical and gamma-ray data integration, (c) Mclust for selected nine elements of stream sediment geochemical and gamma-ray data integration, (d) PAM clustering for selected nine elements of stream sediment geochemical and gamma-ray data integration.

5.4.3. Quality of classification

Regarding producer's accuracy, consistent results, in terms of differences among lithological units, are obtained as displayed in *figure 5-8*. For assessment using Reference I, the consistency could be observed such as for Features A, D, and E. The differences are roughly 10% between the highest and the lowest percentage. Producer's accuracy for Feature A is roughly the same at 50-60% whereas producer's accuracy for Features units D and E are higher at 70-80%. In addition, producer's accuracy for Features B and C is variable depend on clustering type and data preparation.

Table 5-2 shows that assessment results vary from 51% to 62% and from 0.3 to 0.45 for overall accuracy and kappa coefficient, respectively. The lowest accuracy of 51% is obtained in PAM clustering of Integrated Data II with respect to both Reference Data I and II. The lowest accuracy, even is still in moderate level, might be due to its low SC value which describe that the data are less representative to depict the process in the area. The highest accuracy is 62% when Mclust clustering was applied into Integrated Data I with respect to Reference Data I as shown in *table 5-2 (a)*. Similar to overall accuracy, kappa coefficient is worse when Integrated Data II was used with respect to both Reference Data I and

Reference Data II. In addition, *table 5-2* also shows that the assessments using Reference Data I are better than using Reference Data II.

According to assessment results, Mclust and PAM clustering have similar accuracy values. However visually, Mclust images are better than PAM clustering images. It is because the cluster patterns from Mclust have similar general patterns as in the existing lithological map except for the lakes features. Thus, for the next research, it is suggested to integrate satellite imagery such as Landsat or Aster imagery in order to identify water bodies such as lakes features. In addition, Mclust result images have more homogeneous cluster areas than PAM clustering images. It might be because in PAM clustering an optimum cluster number was not achieved.

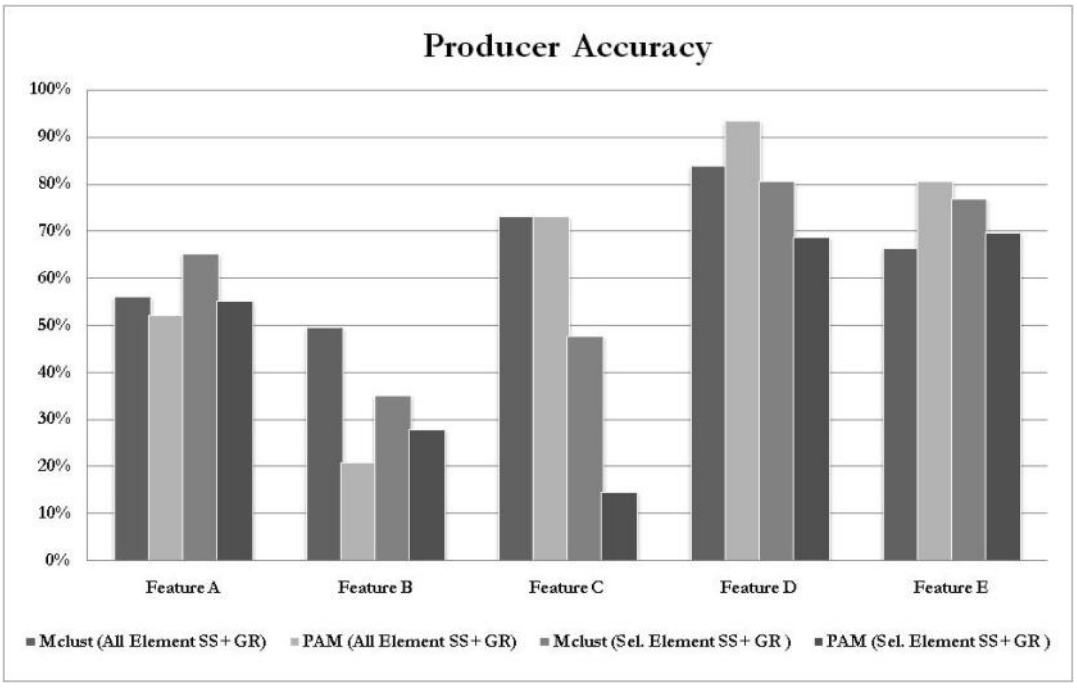
5.4.4. Comparison of individual data

Producer's accuracies of integrated data with respect to Reference Data I tend to be similar to the results of using AGR data (*chapter 3*). It might be due to the influence of sample density and regularity pattern of AGR data. AGR data have denser sampling density than SSG data, the former have density about 1 sample per 0.25 km^2 whereas the latter have density of 1 sample per 13 km^2. In addition, the pattern of the sample points may also influence accuracy in interpolation results. Furthermore, in interpolating SSG data, it was difficult to find appropriate variogram models when only using SSG data within the integrated/AGR study area; thus, data within the whole stream sediment study area were employed (*chapter 2*).

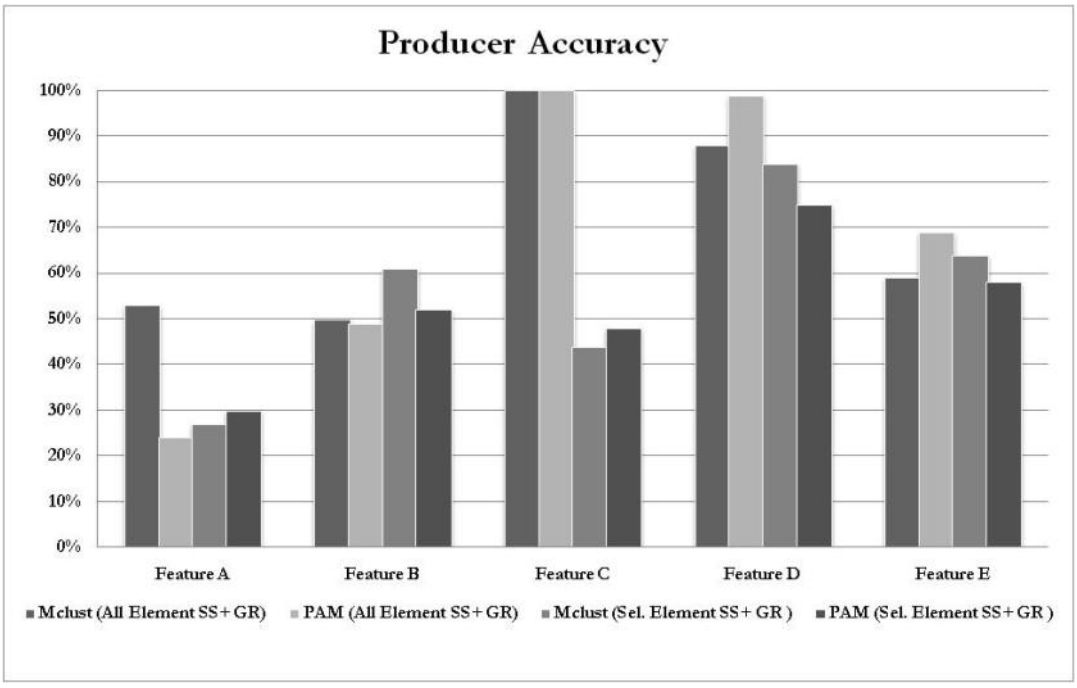
Producer's accuracies of integrated data with respect to Reference Data II are not similar to those of only either SSG data (*chapter 2*) or AGR data (*chapter 3*). Thus, the interpretation of lithological boundaries based on airborne magnetic data (*chapter 4*) might not be better than the existing lithological map. It is due to the fact that the similarity signatures between intrusive and volcanic rocks in magnetic data whereas in the field, geologist could differentiate between them. However, there is similarity of pattern for the intrusive lithological unit when comparing between the lithological map based on airborne magnetic data, *figure 5-1 (c)* and the clustering result base on CoDa for AGR, *figure 3-6 (c)* and *figure 3-7(d)*. This result needs further investigation whether there is real correlation or it is just due to a chance.

Overall accuracies of clustering the SSG and AGR data together are higher than these of SSG and AGR data. However, their better results are inconsistent with respect to producer's accuracy except for Mclust using combination of SSG all 13 elements with AGR data (Integrated Data I). In addition, when the accuracies of SSG data is compared to these of AGR data, Intrusive Rock could be well detected in SSG data whereas Volcanic Rocks could be well distinguished in AGR data. For other features, accuracy of both data shows almost the same results. Furthermore, producer's accuracies of Mclust using Integrated Data I have relative a higher or the same as these of SSG and AGR data. The selection elements of SSG data are not produce better results, as shown in *chapter 2,* when they are combined with AGR data. It might be because structures in SSG data only are different from these in the SSG and AGR data together.

Accuracy of clustering results using the integrated data is dependent on spatial data structure. It is because values in the integrated datasets that were used inputs for clustering are interpolated data. The spatial data structure will influence the selection of appropriate interpolation technique. In the case when the data have no spatial structure, interpolation technique such as kriging is not an appropriate technique.



(a)



(b)

Figure 5-8 Producer's accuracy diagram from assessment of clustering results (for error matrices see *appendix 5-3 and 5-4*), (a) an assessment using existing lithological map for the reference (Reference Data I), (b) an assessment using map of the interpretation of airborne magnetic data (Reference Data II); Feature A, Volcanic Rocks; Feature B, Intrusive Rocks; Feature C, Sedimentary Rocks 1; Feature, Sedimentary Rocks 2; Feature E, Lake (for explanations about the lithology at the study area see *section 1.6.2*)

Table 5-2 Comparison of clustering results assessment based on stream sediment geochemical data only, airborne gamma-ray only, and integration of both datasets, (a)accuracy using combination of stream sediment geochemical (SSG) all 13 elements with airborne gamma-ray (AGR) elements (Integrated Data I), (b) accuracy using combination of SSG selected nine elements with AGR elements; *) an assessment using existing lithological map for the reference (Reference Data I), **) an assessment using map of the interpretation of airborne magnetic data (Reference Data II)

(a)

Assesement	Mclust				PAM			
	SSG	AGR	(SSG[1)]+AGR)[*)]	(SSG+AGR)[**)]	SSG	AGR	SSG+AGR[*)]	SSG+AGR[**)]
Overall Accuracy	50%	50%	62%	59%	44%	45%	61%	57%
Kappa Coefficient	0.39	0.36	0.45	0.37	0.32	0.30	0.45	0.35

(b)

Assesement	Mclust				PAM			
	SSG	AGR	(SSG+AGR)[*)]	(SSG[2)]+AGR)[**)]	SSG	AGR	(SSG+AGR)[*)]	(SSG+AGR)[**)]
Overall Accuracy	51%	50%	56%	54%	43%	45%	51%	51%
Kappa Coefficient	0.41	0.36	0.37	0.30	0.31	0.30	0.30	0.26

5.5. Conclusion remarks

- Based on the overall accuracy and kappa coefficient, clustering results of integrated data (stream sediment geochemistry and airborne gamma-ray) are higher than those of individual data.
- Mclust results are better than PAM clustering results. It is shown by the general pattern similarity of clustering image to existing lithological map.
- Mclust technique could be used as an alternative technique in order to integrate stream sediment geochemical and airborne gamma ray data for mapping the lithology in regional scale.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusions

Base on the research results, the research questions could be answered as follows:

Could stream sediment geochemical data (SSG) be used to assist lithological mapping in vegetation-covered areas where no or little a-priori information about underlying rock units is available?

Mclust and PAM clustering could be alternatives techniques for classifying stream sediment geochemical data to help lithological mapping in area with limited information. The images of their results depict pattern similarities to the existing litholigical map. In addition, the assessments of the results show moderate accuracy up to 51% and 0.41 for overall accuracy and kappa coefficient, respectively. In addition, for a large homogeneous lithology, the producer's accuracy is quite high up to 80%.

Does the application of compositional data (CoDa) analysis to SSG data and airborne gamma-ray (AGR) data produce better results than conventional methods?

In general, the application of CoDa approach in data preparation to both SSG and AGR data do not produce better accuracy than conventional approach. The assessments show the differences between the application of CoDa and conventional approach reach up to 10% and 0.1 for overall accuracy and kappa coefficient, respectively.

Is clustering results using integrated data of SSG and AGR produce better results than those using stream sediment geochemical or airborne gamma-ray data separately?

The integrated data of stream sediment geochemistry and airborne gamma-ray produce better results than those using stream sediment geochemical or airborne gamma-ray data separately. The percentage accuracy increments of integration data compare to their separated data are quite significant up to 17% and 0.15 for overall accuracy and kappa coefficient, respectively.

Which clustering techniques, between Mclust and PAM clustering, give better result for surficial lithologic mapping?

Mclust produces better classifications for lithological mapping relatively to PAM clustering base on both qualitative and quantitative assessments. Qualitatively, from visual evaluation, the patterns of Mclust results are more similar to lithological patterns in the existing lithological map than PAM clustering. Quantitatively, assessments results show up to 5% and 0.7 difference for overall accuracy and kappa coefficient, respectively, in each separated data (SSG or AGR) whereas for integrated data (SSG and AGR) produces non-significant difference in the assessments (1% and 0% differences for overall accuracy and kappa coefficient, respectively).

Therefore, Mclust could be applied to integrate and classify SSG and AGR data for lithological mapping in regional scale.

6.2. Recommendations

It is suggested to integrate more data such as elevation data e.g. SRTM image, and/or satellite images e.g. ASTER and Landsat image, to consider other factors than only chemical factor. In addition, it is also suggested to test application of Mclust for larger scale such as for district scale. Therefore, in order to be able in analyzing larger data (a larger area and/or more type of data and/or a smaller grid size images) for lithological mapping, integrating *R* application with GIS software such as *ArcGIS* is recommended because the GIS software has a strength in handling large datasets but limited in statistical calculation and analysis. This limitation can be compensated by using *R*.

LIST OF REFERENCES

Alberti, A., Alessandro, V., Pieruccini, U., and Pranzini, E., 1993, Landsat TM data processing for lithological discrimination in the Caraculo area (Namibe Province, SW Angola): Journal of African Earth Sciences (and the Middle East), v. 17, p. 261-274.

An, P., Chung, C.F., and Rencz, A.N., 1995, Digital lithology mapping from airborne geophysical and remote sensing data in the Melville Peninsula, Northern Canada, using a neural network approach: Remote Sensing of Environment, v. 53, p. 76-84.

Bedini, E., 2009, Mapping lithology of the Sarfartoq Carbonatite Complex, Southern West Greenland, using hymap imaging spectrometer data: Remote Sensing of Environment, v. 113, p. 1208-1219.

Bellehumeur, C., Marcotte, D., and Jébrak, M., 1994, Multi-element relationships and spatial structures of regional geochemical data from stream sediments, Southwestern Quebec, Canada: Journal of Geochemical Exploration, v. 51, p. 11-35.

Blakely, R.J., 1996, Potential theory in gravity and magnetic applications: Cambridge, Cambridge University Press, 441 p.

Bonham-Carter, G.F., Rogers, P.J., and Ellwood, D.J., 1987, Catchment basin analysis applied to surficial geochemical data, Cobequid Highlands, Nova Scotia: Journal of Geochemical Exploration, v. 29, p. 259-278.

Carranza, E.J.M., 2008, Geochemical anomaly and mineral prospectivity mapping in GIS: Amsterdam, Elsevier, 366 p.

Carranza, E.J.M., 2010, Mapping of anomalies in continuous and discrete fields of stream sediment geochemical landscapes: Geochemistry-Exploration Environment Analysis, v. 10, p. 171-187.

Carranza, E.J.M., and Hale, M., 1997, A catchment basin approach to the analysis of reconnaissance geochemical-geological data from Albay Province, Philippines: Journal of Geochemical Exploration, v. 60, p. 157-171.

Cocker, M.D., 1999, Geochemical mapping in Georgia, USA: a tool for environmental studies, geologic mapping and mineral exploration: Journal of Geochemical Exploration, v. 67, p. 345-360.

Delong, R.C., 1996, Geology, alteration, mineralization and metal zonation of the Mt. Milligan Porphyry Copper-gold Deposits: Vancouver, The University of British Columbia.

Dickson, B.L., and Scott, K.M., 1997, Interpretation of aerial gamma-ray surveys-adding the geochemical factor: Journal of Geology and Geophysics, v. 17, p. 187-200.

Dinh, V., Leitner, R., Paclik, P., and Duin, R., 2009, A clustering based method for edge detection in hyperspectral images, *in* Salberg, A.-B., Hardeberg, J., and Jenssen, R., eds., Image Analysis, Volume 5575: Lecture Notes in Computer Science, Springer Berlin / Heidelberg, p. 580-587.

Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C., 2003, Isometric logratio transformations for compositional data analysis: Mathematical Geology, v. 35, p. 279-300.

Filzmoser, P., Hron, K., and Reimann, C., 2009a, Principal component analysis for compositional data with outliers: Environmetrics, v. 20, p. 621-632.

Filzmoser, P., Hron, K., Reimann, C., and Garrett, R., 2009b, Robust factor analysis for compositional data: Computers & Geosciences, v. 35, p. 1854-1861.

Fraley, C., and Raftery, A.E., 2002, Model-based clustering, discriminant analysis, and density estimation: Journal of the American Statistical Association, v. 97, p. 611-631.

Fraley, C., and Raftery, A.E., 2006, Mclust version 3 for r: normal mixture modeling and model-based clustering, Technical Report No. 504, Department of Statistics, University of Washington (http://cran.r-project.org/web/packages/mclust/index.html).

Gan, G., Ma, C., and Wu, J., 2007, Data clustering: theory, algorithm, and applications, SIAM, Philadelphia, ASA, Alexandria, VA, 466 p.

Garrett, R.G., 2010, rgr: the GSC applied geochemistry EDA package, R package version 1.0.3., (http://gsc.nrcan.gc.ca/dir/index_e.php?id=4961).

Garrett, R.G., and Grunsky, E.C., 2003, S and R functions for the display of Thompson-Howarth plots: Computers & Geosciences, v. 29, p. 239-242.

Geoscience BC, 2010, QUEST: Quesnellia Exploration Strategy, (http://www.geosciencebc.com/s/Quest.asp).

Gomez, C., Delacourt, C., Allemand, P., Ledru, P., and Wackerle, R., 2005, Using ASTER remote sensing data set for geological mapping, in Namibia: Physics and Chemistry of the Earth, Parts A/B/C, v. 30, p. 97-108.

Graham, D.F., and Bonham-Carter, G.F., 1993, Airborne radiometric data - a tool for reconnaissance geological mapping using a GIS: Photogrammetric Engineering and Remote Sensing, v. 59, p. 1243-1249.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G., 2010, impute: imputation for microarray data. R package version 1.22.0 (http://cran.r-project.org/web/packages/impute/index.html).

Hengl, T., 2006, Finding the right pixel size: Computers & Geosciences, v. 32, p. 1283-1298.

Howarth, R.J., 1984, Statistical applications in geochemical prospecting: a survey of recent developments: Journal of Geochemical Exploration, v. 21, p. 41-61.

Hron, K., Templ, M., and Filzmoser, P., 2010, Imputation of missing values for compositional data using classical and robust methods: Computational Statistics & Data Analysis, v. 54, p. 3095-3107.

Jackaman, W., and Balfour, J.S., 2008, QUEST project geochemistry: field surveys and data reanalysis, Central British Columbia (parts of NTS 093A, B, G, H, J, K, N, O), Geoscience BC Summary of Activities 2007; Geoscience BC, Report 2008-1, p. 150.

Jenks, G.F., 1967, The data model concept in statistical mapping, *in* Frenzel, K., ed., International Yearbook of Cartography, Volume 7, Rand McNally & Co, p. 186.

Kaufman, L., and Rousseeuw, P.J., 2005, Finding groups in data: an introduction to cluster analysis: New Jersey, John Wiley & Sons, Inc.

Kerr, A., and Davenport, P.H., 1990, Application of geochemical mapping techniques to a complex Precambrian shield area in Labrador, Canada: Journal of Geochemical Exploration, v. 39, p. 225-247.

Lang, R., Shao, G., Pijanowski, B.C., and Farnsworth, R.L., 2008, Optimizing unsupervised classifications of remotely sensed imagery with a data-assisted labeling approach: Computers & Geosciences, v. 34, p. 1877-1885.

Lillesand, T.M., and Kiefer, R.W., 2000, Remote sensing and image interpretation: New York, etc., John Wiley and Sons, 724 p.

Lisle, R.J., 2004, Geological structures and maps: a practical guide: Amsterdam, etc., Elsevier, 106 p.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. , 2005, Cluster analysis basics and extensions, unpublished (http://cran.r-project.org/web/packages/cluster/index.html).

Martelet, G., Truffert, C., Tourliere, B., Ledru, P., and Perrin, J., 2006, Classifying airborne radiometry data with agglomerative hierarchical clustering: a tool for geological mapping in context of rainforest (French Guiana): International Journal of Applied Earth Observation and Geoinformation, v. 8, p. 208-223.

Massey, N.W.D., MacIntyre, D.G., Desjardins, P.J., and Cooney, R.T., 2005a, Digital Geology Map of British Columbia - Tile NN10 Central B.C., GeoFile 2005-6, B.C. Ministry of Energy and Mines.

Massey, N.W.D., MacIntyre, D.G., Desjardins, P.J., and Cooney, R.T., 2005b, Digital Geology Map of British Columbia: Tile NO10 Northeast B.C, GeoFile 2005-10, B.C. Ministry of Energy and Mines.

Massey, N.W.D., MacIntyre, D.G., Desjardins, P.J., and Cooney, R.T., 2005c, Digital Geology Map of British Columbia:Tile NO9 North Central B.C., GeoFile 2005-9, B.C. Ministry of Energy and Mines.

Massey, N.W.D., MacIntyre, D.G., Haggart, J.W., Desjardins, P.J., Wagner, C.L., and Cooney, R.T., 2005d, Digital Geology Map of British Columbia: Tile NN8-9 North Coast and Queen Charlotte Islands/Haida Gwaii, GeoFile 2005-5, B.C. Ministry of Energy and Mines.

Milligan, P.R., and Gunn, P.J., 1997, Enhancement and presentation of airborne geophysical data: Journal of Geology and Geophysics, v. 17, p. 63-75.

Naseem, S., Sheikh, S.A., Qadeeruddin, M., and Shirin, K., 2002, Geochemical stream sediment survey in Winder Valley, Balochistan, Pakistan: Journal of Geochemical Exploration, v. 76, p. 1-12.

Natural Resources Canada, 2010a, Geoscience data repository, aeromagnetic and electromagnetic data, Natural Resources Canada (http://gdr.nrcan.gc.ca/aeromag/index_e.php).

Natural Resources Canada, 2010b, Geoscience data repository, radioactivity data, Natural Resources Canada (http://gdr.nrcan.gc.ca/gamma/index_e.php).

Nelson, J., Bellefontaine, K., Rees, C., and MacLean, M., 1992, Regional geological mapping in the Nation Lakes Area (93N/2E,7E), Geological Field Work 1991, Paper 1992-1, British Columbia Ministry of Energy, Mines and Petroleum Resources, p. 118.

Nelson, J., Bellefontaine, K.,Green, K., and MacLean, M., 1991, Regional geological mapping near the Mount Milligan Deposit (93N/l,93k/16), Geological Fieldwork 1990, Paper 1991-1, British Columbia Ministry of Energy, Mines and Petroleum Resources.

Ninomiya, Y., Fu, B., and Cudahy, T.J., 2005, Detecting lithology with Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) multispectral thermal infrared "radiance-at-sensor" data: Remote Sensing of Environment, v. 99, p. 127-139.

Pawlowsky-Glahn, V., and Egozcue, J.J., 2006, Compositional data and their analysis: an introduction, *in* Buccianti, A., Mateu-Figueras, G., and Pawlowsky-Glahn, V., eds., Compositional data analysis in the geosciences: from theory to practice: London, The Geological Society, p. 264.

Raftery, A.E., and Dean, N., 2006, Variable selection for model-based clustering: Journal of the American Statistical Association, v. 101, p. 168-178.

Raftery, N.D.a.A.E., 2009, clustvarsel: variable selection for model-based clustering. R package version 1.3. (http://CRAN.R-project.org/package=clustvarsel).

Ramsey, M.H., Thompson, M., and Hale, M., 1992, Objective evaluation of precision requirements for geochemical analysis using robust analysis of variance: Journal of Geochemical Exploration, v. 44, p. 23-36.

Ranasinghe, P.N., Fernando, G., Dissanayake, C.B., Rupasinghe, M.S., and Witter, D.L., 2009, Statistical evaluation of stream sediment geochemistry in interpreting the river catchment of high-grade metamorphic terrains: Journal of Geochemical Exploration, v. 103, p. 97-114.

Rantitsch, G., 2000, Application of fuzzy clusters to quantify lithological background concentrations in stream-sediment geochemistry: Journal of Geochemical Exploration, v. 71, p. 73-82.

Reimann, C., Filzmoser, P., Garrett, R., and Dutter, R., 2008, Statistical data analysis explained: applied environmental statistics with R: West Sussex, John Wiley and Sons, 341 p.

Reimann, C., Filzmoser, P., and Garrett, R.G., 2005, Background and threshold: critical comparison of methods of determination: Science of The Total Environment, v. 346, p. 1-16.

Robinson, G.R., Kapo, K.E., and Grossman, J.N., 2004, Chemistry of stream sediments and surface waters in New England, Volume 2010, p. open-file report 2004-1026.

Roest, W.R., Verhoef, V., and Pilkington, M., 1992, Magnetic interpretation using the 3-D analytic signal: Geophysics, v. 57, p. 116-125.

Rowan, L.C., and Mars, J.C., 2003, Lithologic mapping in the Mountain Pass, California area using Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) data: Remote Sensing of Environment, v. 84, p. 350-366.

Rowan, L.C., Simpson, C.J., and Mars, J.C., 2004, Hyperspectral analysis of the Ultramafic Complex and Adjacent Lithologies at Mordor, NT, Australia: Remote Sensing of Environment, v. 91, p. 419-431.

Schetselaar, E.M., 2000, Integrated analyses of granite - gneiss terrain from field and multisource remotely sensed data : a case study from the Canadian Shield: Enschede, ITC.

Shepherd, A., Harvey, P.K., and Leake, R.C., 1987, The geochemistry of residual soils as an aid to geological mapping: a statistical approach: Journal of Geochemical Exploration, v. 29, p. 317-331.

Smith, R.E., 1996, Regolith research in support of mineral exploration in Australia: Journal of Geochemical Exploration, v. 57, p. 159-173.

Spadoni, M., Cavarretta, G., and Patera, A., 2004, Cartographic techniques for mapping the geochemical data of stream sediments: the "sample catchment basin" approach: Environmental Geology, v. 45.

Stendal, H., 1978, Heavy minerals in stream sediments, Southwest Norway: Journal of Geochemical Exploration, v. 10, p. 91-102.

Swan, R.H., and Sandilands, M., 1995, Introduction to geological data analysis, Blackwell Science.

Templ, M., Filzmoser, P., and Reimann, C., 2008, Cluster analysis applied to regional geochemical data: problems and possibilities: Applied Geochemistry, v. 23, p. 2198-2213.

Templ, M., Hron, K., and Filzmoser, P., 2010, robCompositions: robust estimation for compositional data. R package version 1.4.3 (http://CRAN.R-project.org/package=robCompositions).

Thompson, M., 1983, Control procedures in geochemical analysis, *in* Howarth, R.J., ed., Handbook of Exploration Geochemistry, Volume 2: Amsterdam, Elsevier, p. 437.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B., 2001, Missing value estimation methods for DNA microarrays: Bioinformatics, v. 17, p. 520-525.

Van den Boogaart, K.G., and Tolosana-Delgado, R., 2008, "compositions": a unified R package to analyze compositional data: Computers & Geosciences, v. 34, p. 320-338.

Van den Boogaart, K.G., Tolosana, R., and Bren, M., 2008, compositions:compositional data analysis. R package version 1.01-1 (http://www.stat.boogaart.de/compositions/).

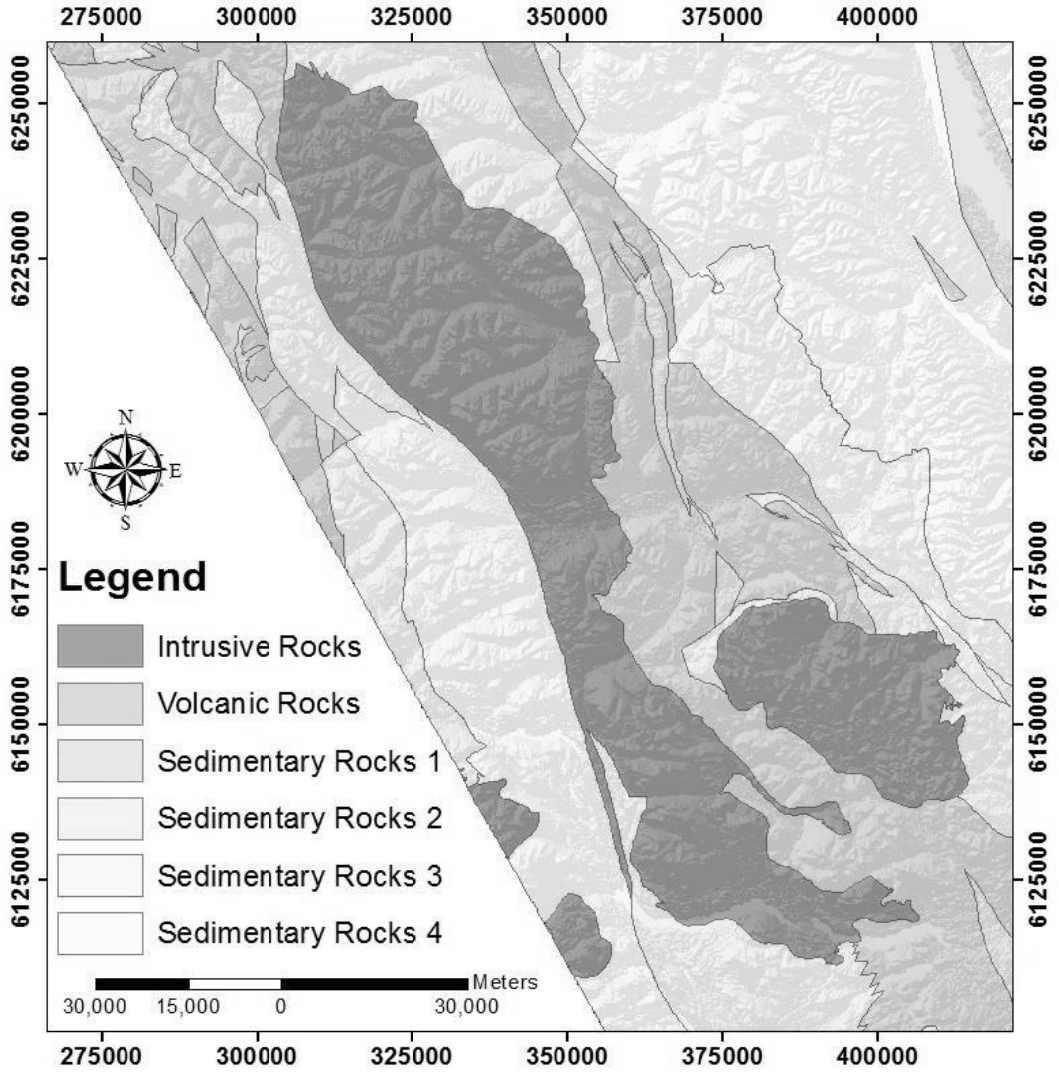
Van der Werff, H., Van Ruitenbeek, H., Van der Meijde, M., Van der Meer, F., Jong, S.d., and Kalubandara, S., 2007, Rotation-variant template matching for supervised hyperspectral boundary detection: IEEE Geoscience and Remote Sensing Letters, v. 4, p. 70-74.

Weisberg, S., 2005, Applied linear regression: New Jersey, John Wiley & Sons, Inc., 310 p.

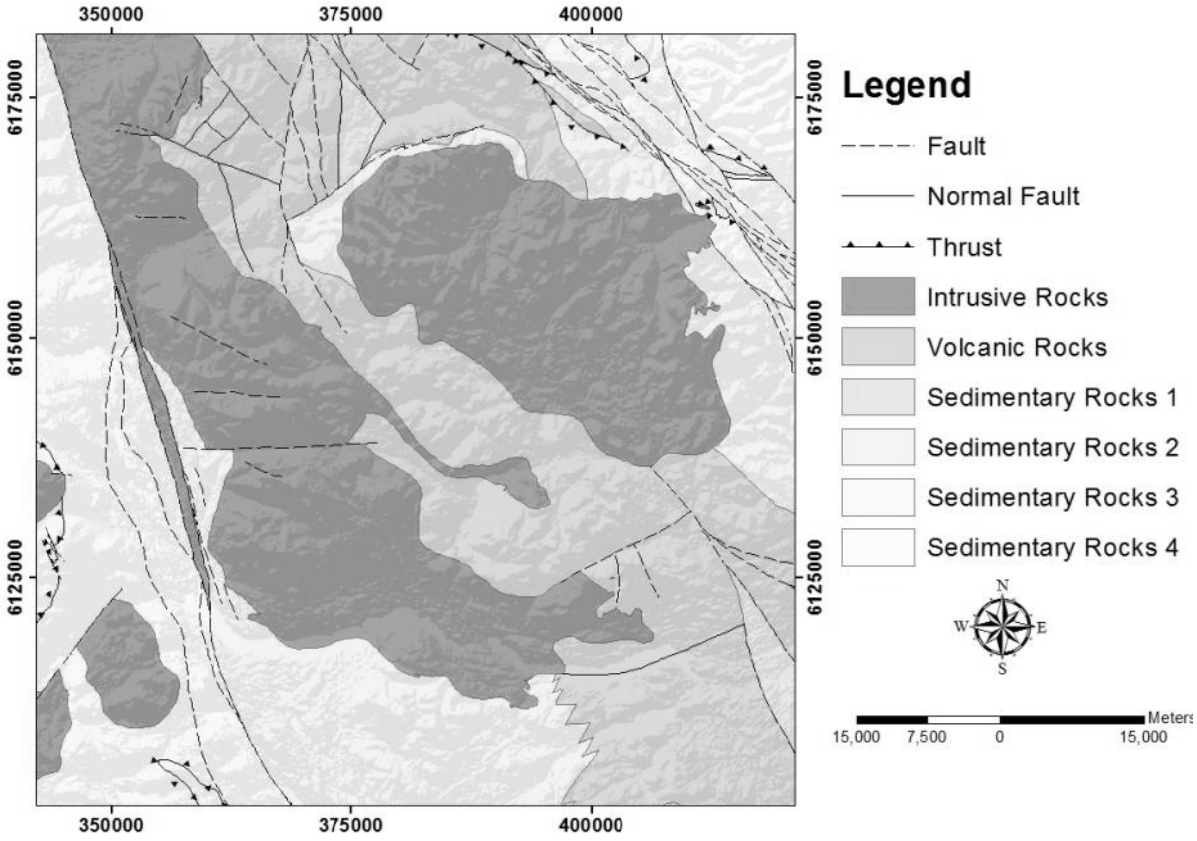
Wilford, J.R., Bierwirth, P.N., and Craig, M.A., 1997, Application of airborne gamma-ray spectrometry in soil/regolith mapping and applied geomorphology: Journal of Geology and Geophysics, v. 17, p. 201-216.

Yusta, I., Velasco, F., and Herrero, J.-M., 1998, Anomaly threshold estimation and data normalization using EDA statistics: application to lithogeochemical exploration in Lower Cretaceous Zn-Pb carbonate-hosted deposits, Northern Spain: Applied Geochemistry, v. 13, p. 421-439.

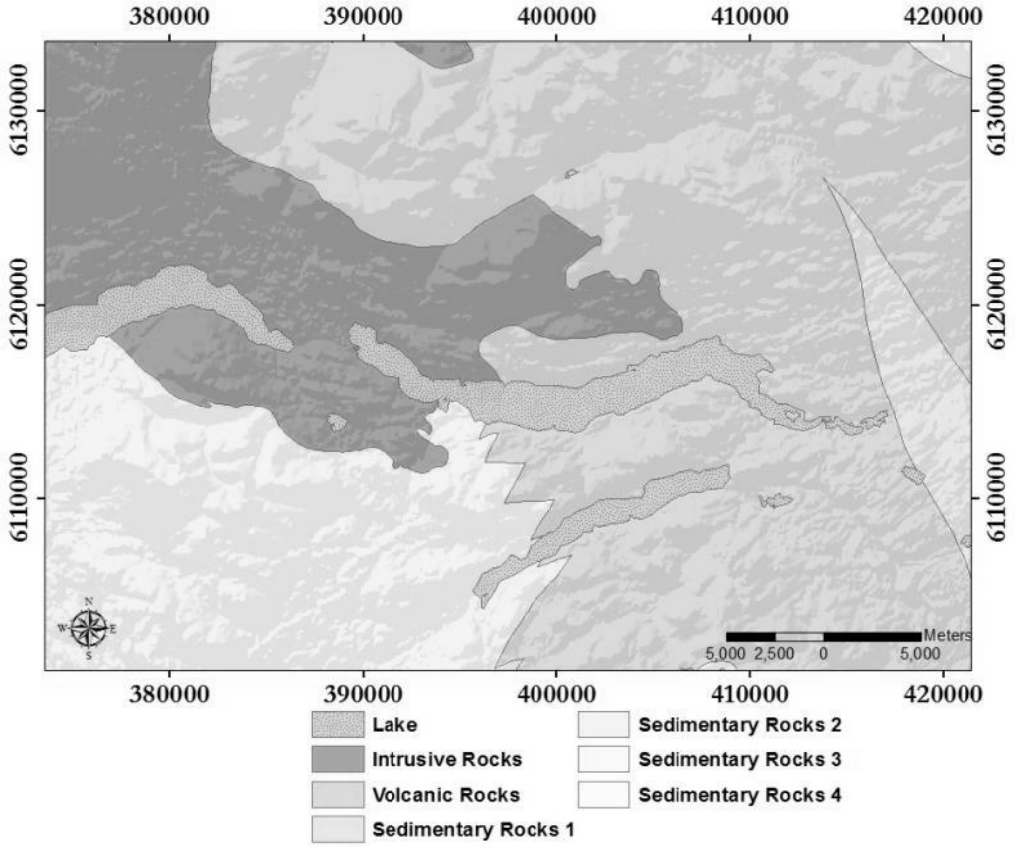
Appendix 1-1 Simplified existing geological map for validation (modified from (Massey et al., 2005a, b, c; Massey et al., 2005d); (a) validation map for stream sediment datasets, (b) validation map for airborne gamma-ray datasets, (c) validation map for airborne magnetic dataset; The lithology at study area comprises of Intrusive Rocks; Volcanic Rocks (volcanic rocks and ultramafic rocks); Sedimentary Rocks 1 (sedimentary rocks from Jurassic to Quaternary); Sedimentary Rocks 2 (sedimentary rocks from Triassic to Jurassic and small parts of metamorphic rocks); Sedimentary Rocks 3 (sedimentary rock from Ordovician to Jurassic); and Sedimentary Rocks 4 (sedimentary rock from Proterozoic to Ordovician and small parts metamorphic rocks).



(a)


350000
375000
400000
6175000
6150000
6125000
Legend
Fault
Normal Fault
Thrust
Intrusive Rocks
Volcanic Rocks
Sedimentary Rocks 1
Sedimentary Rocks 2
Sedimentary Rocks 3
Sedimentary Rocks 4
N
W
E
S
Meters
15,000 7,500 0 15,000

(b)


380000
390000
400000
410000
420000
6130000
6120000
6110000
N
W
E
S
Meters
5,000 2,500 0 5,000
Lake
Intrusive Rocks
Volcanic Rocks
Sedimentary Rocks 1
Sedimentary Rocks 2
Sedimentary Rocks 3
Sedimentary Rocks 4

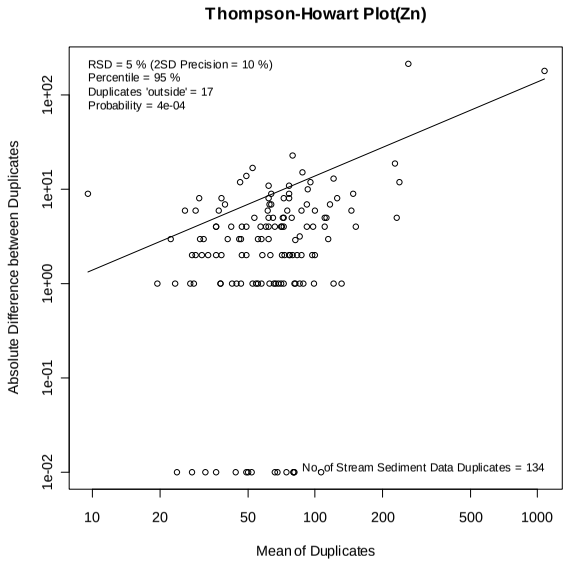
Appendix 2-1 Properties of Mclust models (Fraley and Raftery, 2006)

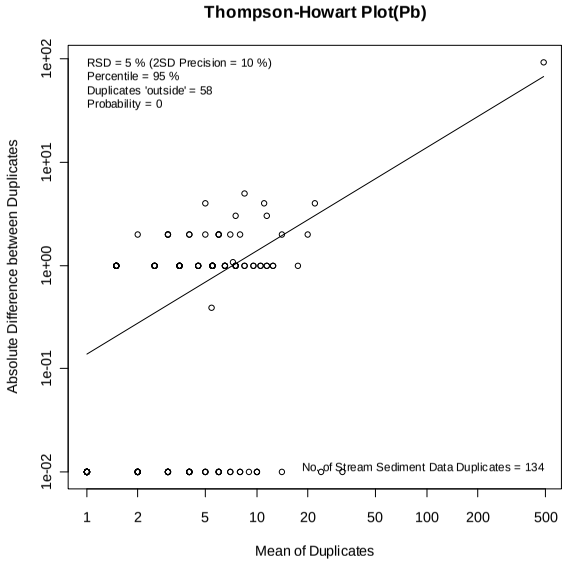
Identifier	Distribution	Volume	Shape	Orientation
EII	spherical	equal	equal	NA
VII	spherical	variable	equal	NA
EEI	diagonal	equal	equal	coordinates axes
VEI	diagonal	variable	equal	coordinates axes
EVI	diagonal	equal	variable	coordinates axes
VVI	diagonal	variable	variable	coordinates axes
EEE	ellipsoidal	equal	equal	equal
EEV	ellipsoidal	equal	equal	variable
VEV	ellipsoidal	variable	equal	variable
VVV	ellipsoidal	variable	variable	variable

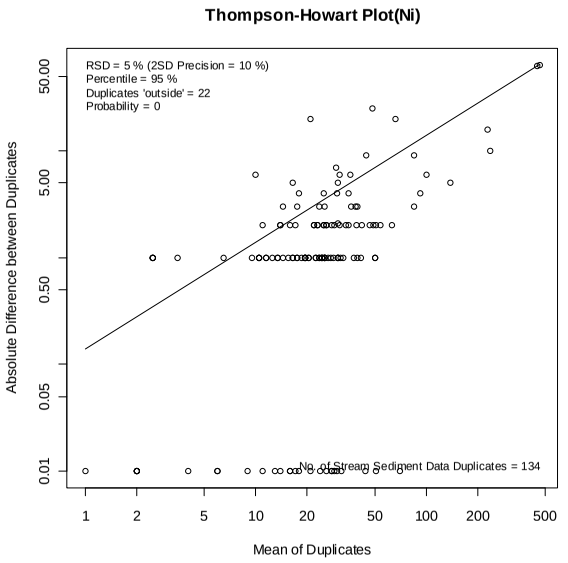
Appendix 2-2 Subjective interpretation of the silhouette coefficient (SC) (Kaufman and Rousseeuw, 2005)

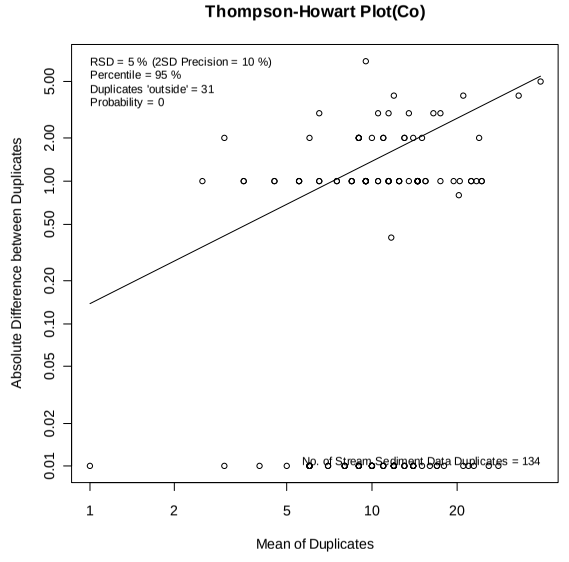
SC	Proposed Interpretation
0.71 - 1.00	a strong structure has been found
0.51 - 0.70	a reasonable structure has been found
0.26 - 0.50	the structure is weak and could be artificial; try additional methods
≤ 0.25	no substantial structure has been found

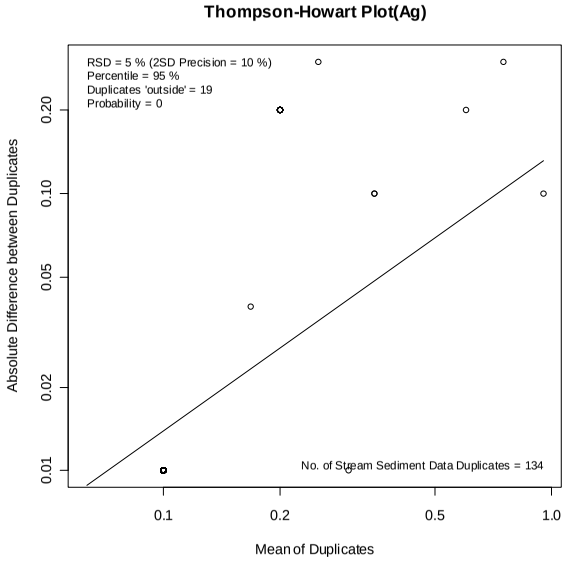
Appendix 2-3 Stream sediment geochemical elements lie on Thompson-Howarth Plot (a model for 10% precision and 95% significant level)

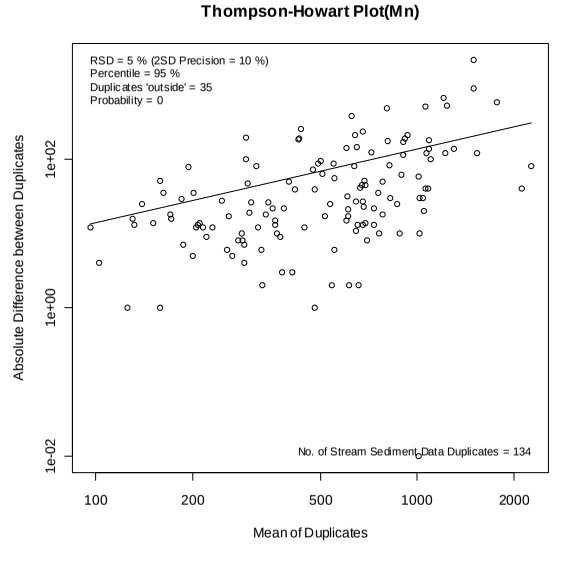


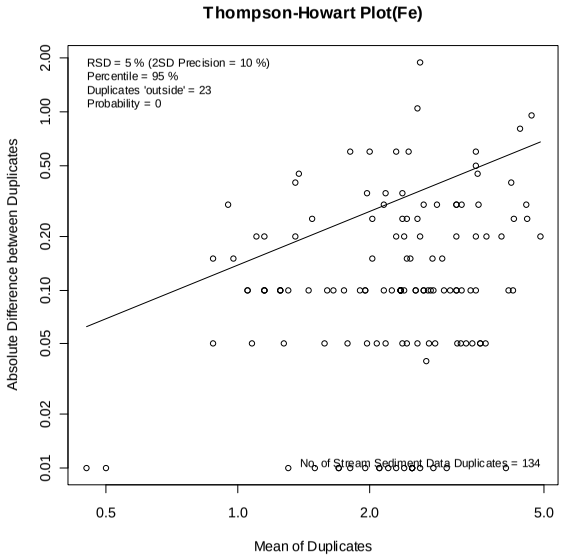


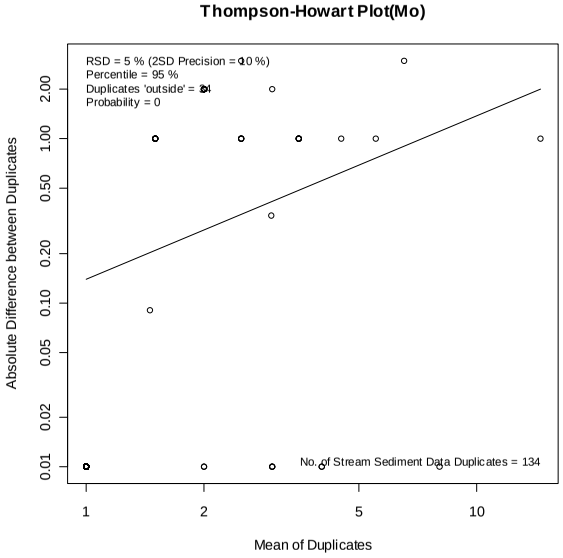


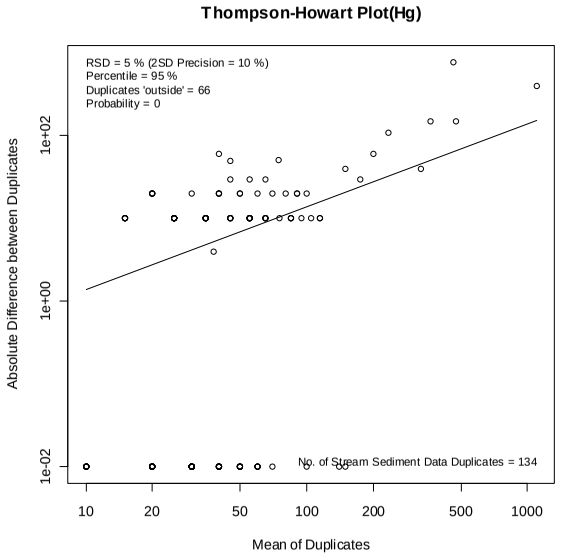


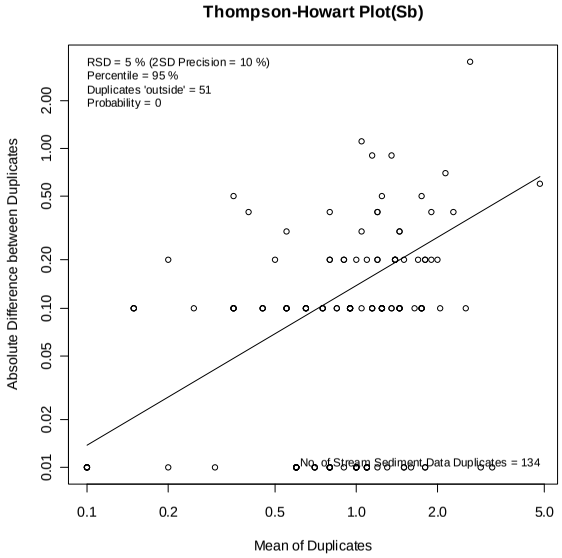


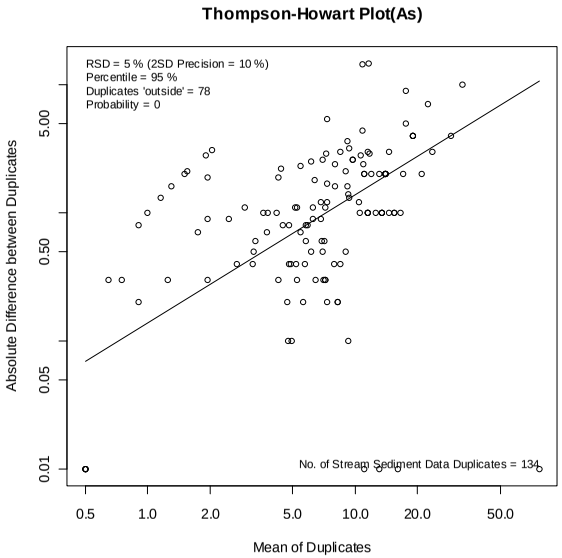


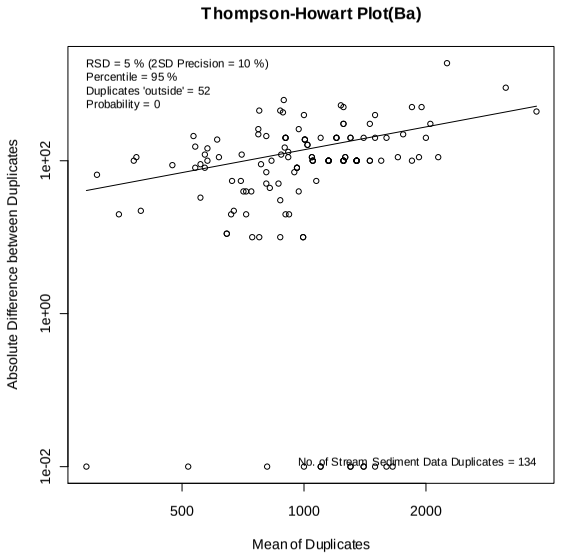

Thompson-Howart Plot(Fe)
RSD = 5 % (2SD Precision = 10 %)
Percentile = 95 %
Duplicates 'outside' = 23
Probability = 0
No. of Stream Sediment Data Duplicates = 134
Absolute Difference between Duplicates
Mean of Duplicates

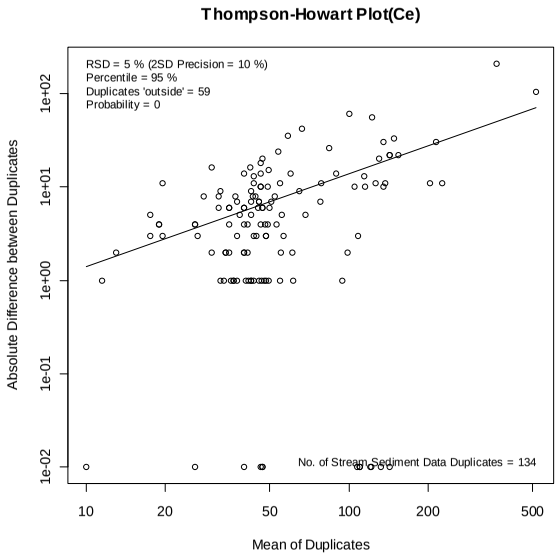

Thompson-Howart Plot(Mo)
RSD = 5 % (2SD Precision = 10 %)
Percentile = 95 %
Probability = 0
No. of Stream Sediment Data Duplicates = 134
Absolute Difference between Duplicates
Mean of Duplicates

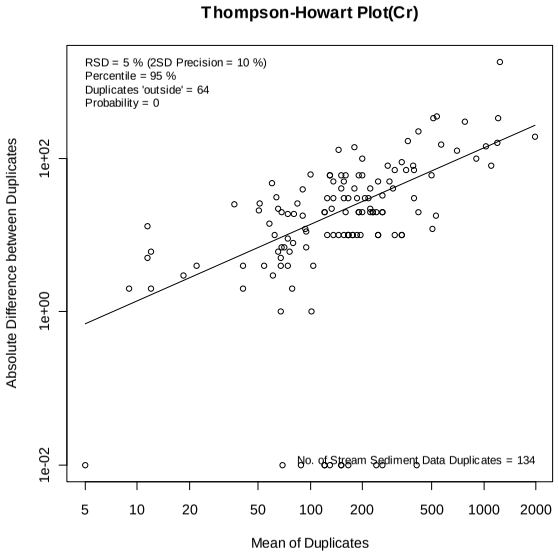

Thompson-Howart Plot(Hg)
RSD = 5 % (2SD Precision = 10 %)
Percentile = 95 %
Duplicates 'outside' = 66
Probability = 0
No. of Stream Sediment Data Duplicates = 134
Absolute Difference between Duplicates
Mean of Duplicates

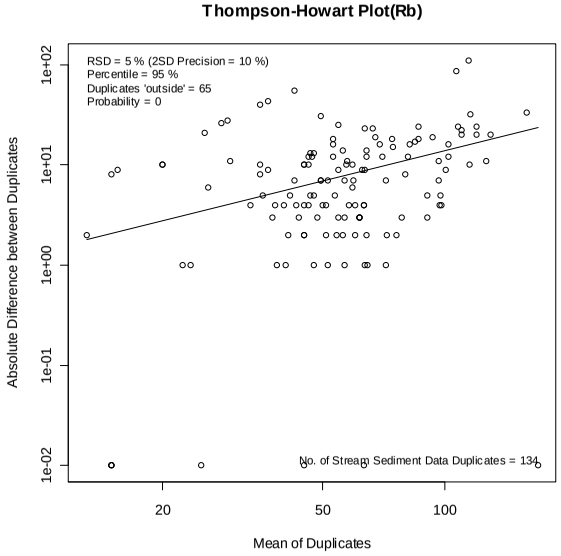

Thompson-Howart Plot(Sb)
RSD = 5 % (2SD Precision = 10 %)
Percentile = 95 %
Duplicates 'outside' = 51
Probability = 0
No. of Stream Sediment Data Duplicates = 134
Absolute Difference between Duplicates
Mean of Duplicates


Thompson-Howart Plot(As)
RSD = 5 % (2SD Precision = 10 %)
Percentile = 95 %
Duplicates 'outside' = 78
Probability = 0
No. of Stream Sediment Data Duplicates = 134
Absolute Difference between Duplicates
Mean of Duplicates

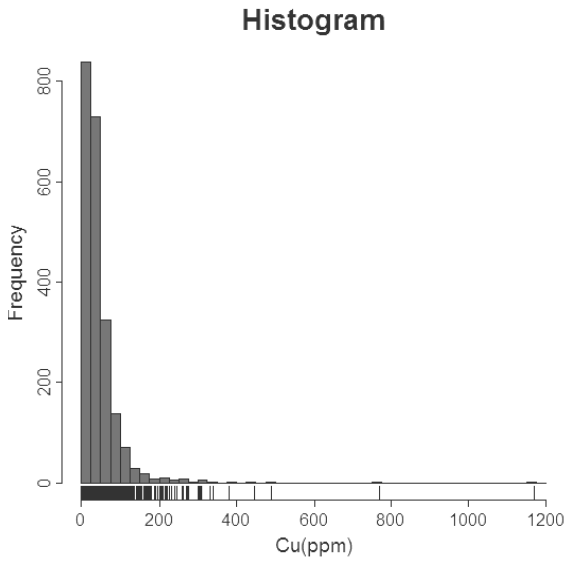

Thompson-Howart Plot(Ba)
RSD = 5 % (2SD Precision = 10 %)
Percentile = 95 %
Duplicates 'outside' = 52
Probability = 0
No. of Stream Sediment Data Duplicates = 134
Absolute Difference between Duplicates
Mean of Duplicates

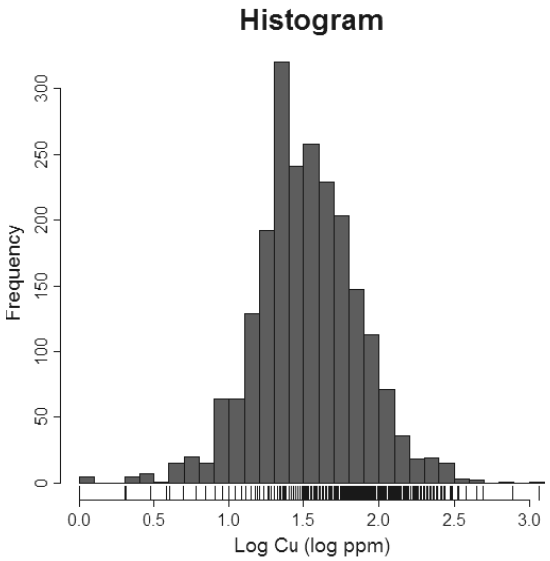

Thompson-Howart Plot(Ce)
RSD = 5 % (2SD Precision = 10 %)
Percentile = 95 %
Duplicates 'outside' = 59
Probability = 0
No. of Stream Sediment Data Duplicates = 134
Absolute Difference between Duplicates
1e-02
1e-01
1e+00
1e+01
1e+02
10
20
50
100
200
500
Mean of Duplicates

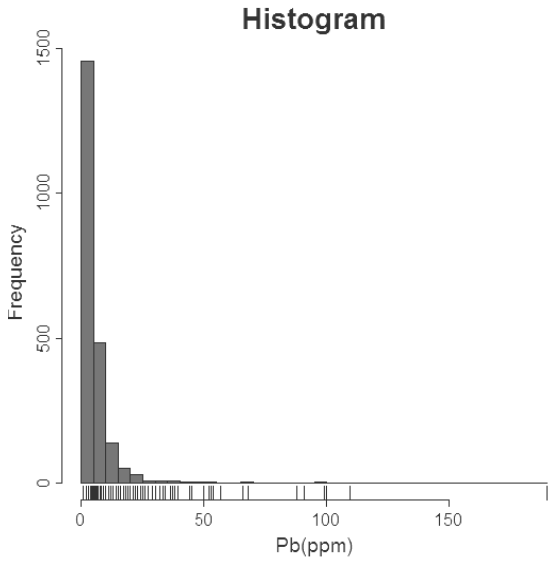

Thompson-Howart Plot(Cr)
RSD = 5 % (2SD Precision = 10 %)
Percentile = 95 %
Duplicates 'outside' = 64
Probability = 0
No. of Stream Sediment Data Duplicates = 134
Absolute Difference between Duplicates
1e-02
1e+00
1e+02
5
10
20
50
100
200
500
1000
2000
Mean of Duplicates

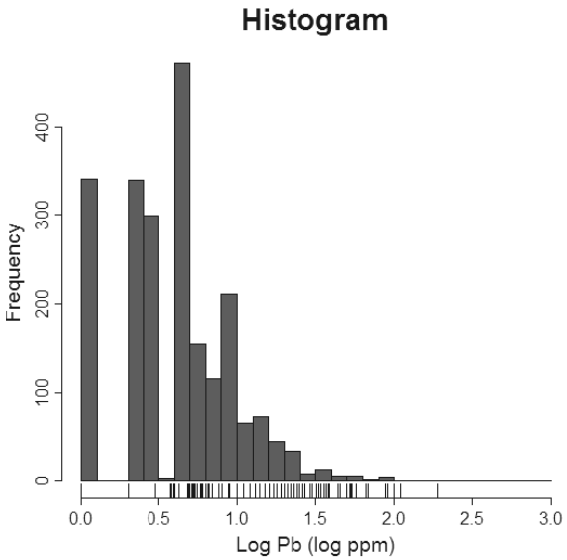

Thompson-Howart Plot(Rb)
RSD = 5 % (2SD Precision = 10 %)
Percentile = 95 %
Duplicates 'outside' = 65
Probability = 0
No. of Stream Sediment Data Duplicates = 134
Absolute Difference between Duplicates
1e-02
1e-01
1e+00
1e+01
1e+02
20
50
100
Mean of Duplicates

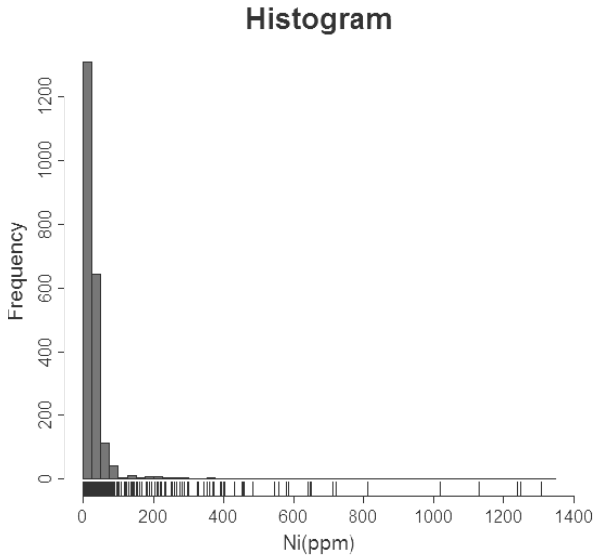
Appendix 2-4 Histogram for stream sediment geochemical elements; graphs for raw data without transformation (left) and graphs for base-10 logarithmic transformation data (right)

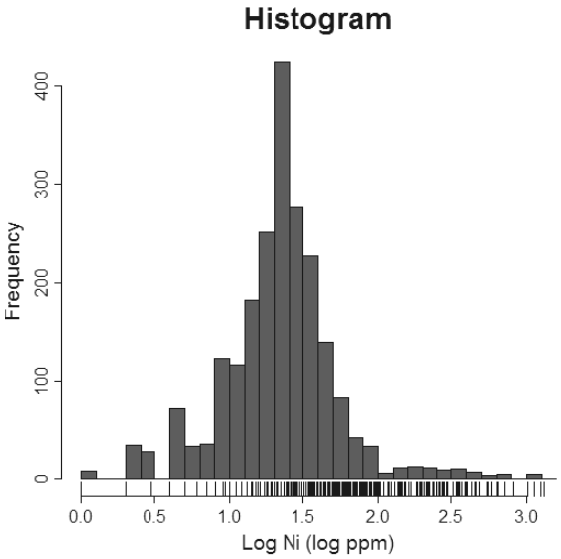


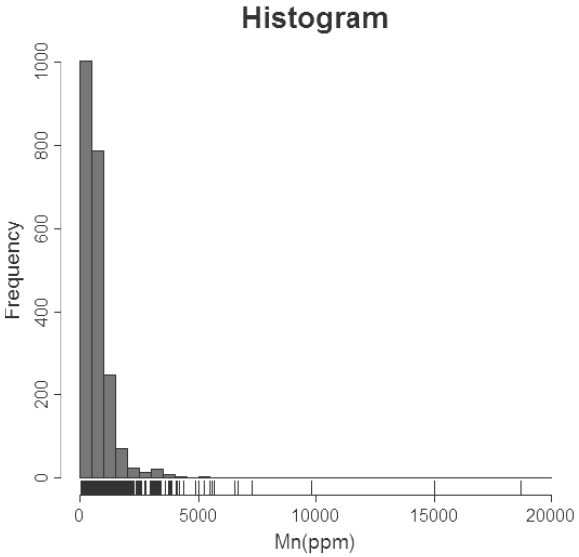


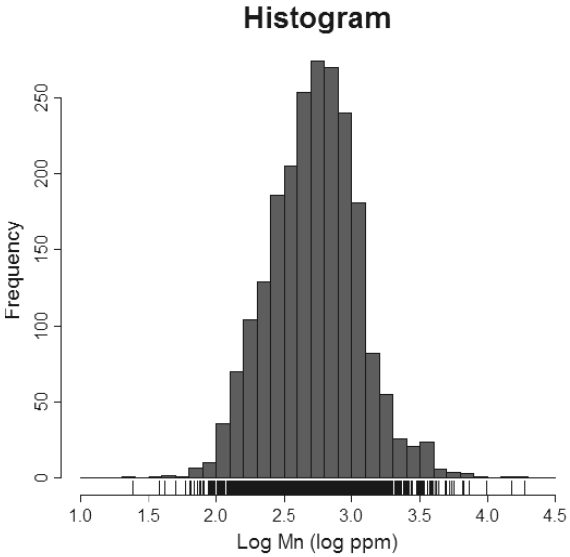


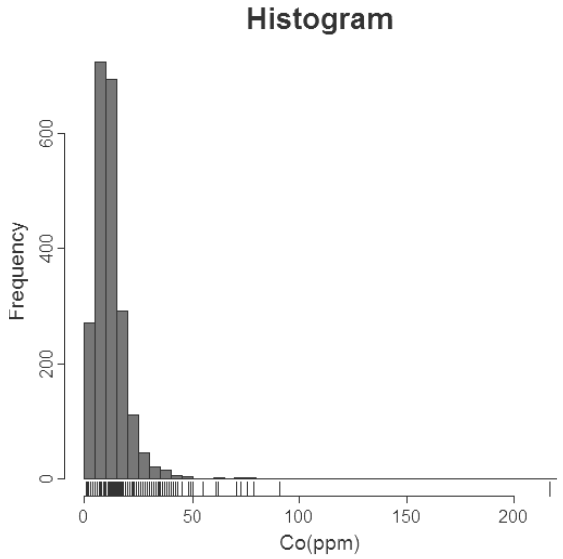


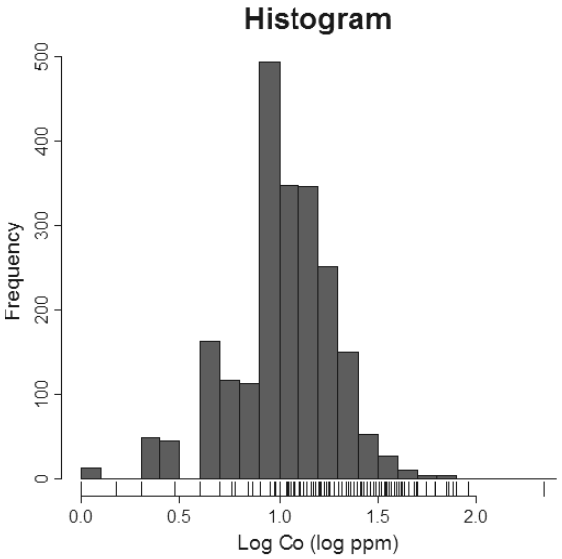


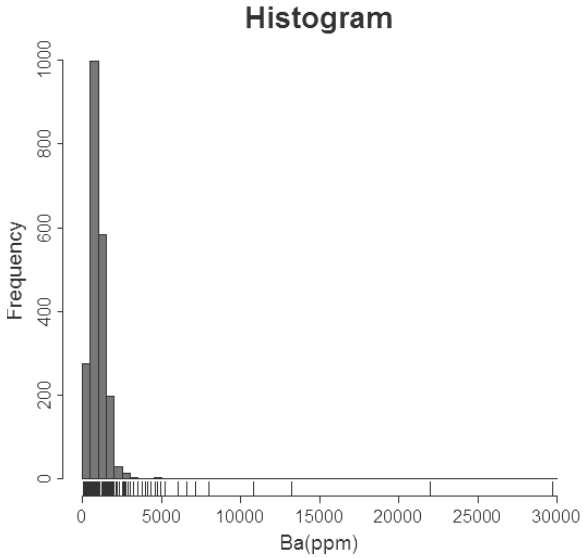


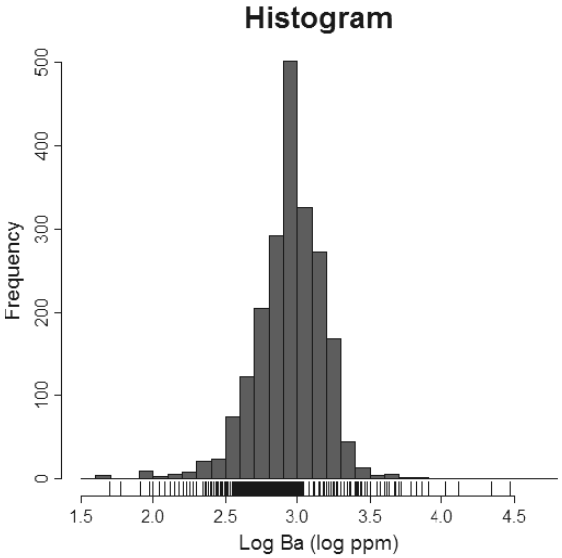

Histogram
Frequency
0
200
400
600
800
1000
0
5000
10000
15000
20000
Mn(ppm)

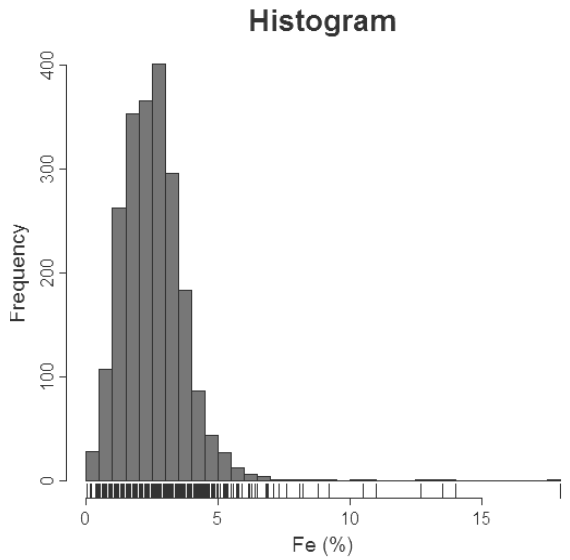

Histogram
Frequency
0
50
100
150
200
250
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Log Mn (log ppm)

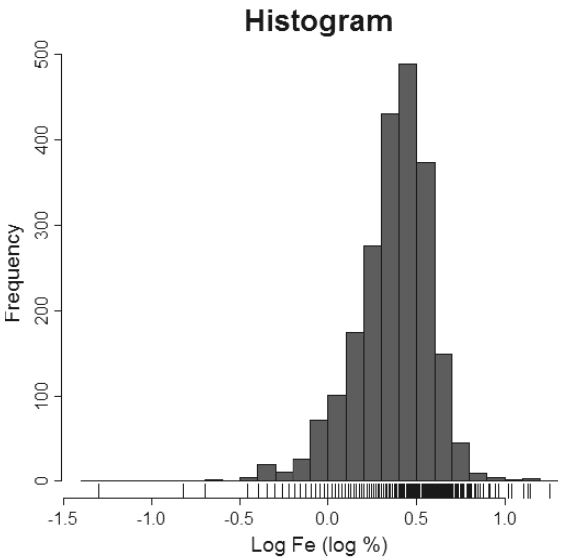

Histogram
Frequency
0
200
400
600
0
50
100
150
200
Co(ppm)

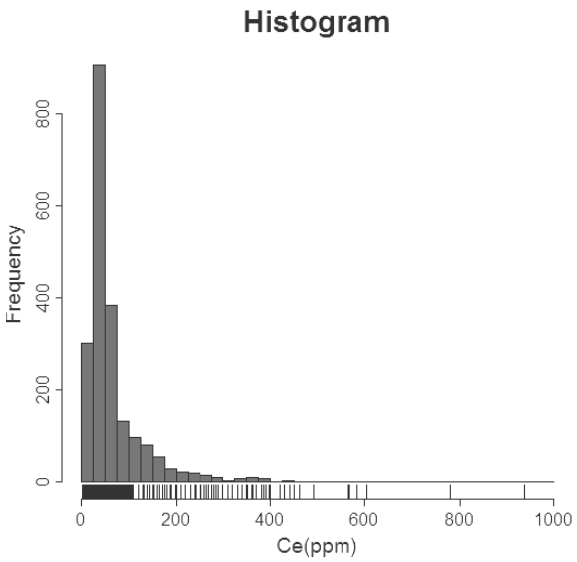

Histogram
Frequency
0
100
200
300
400
500
0.0
0.5
1.0
1.5
2.0
Log Co (log ppm)

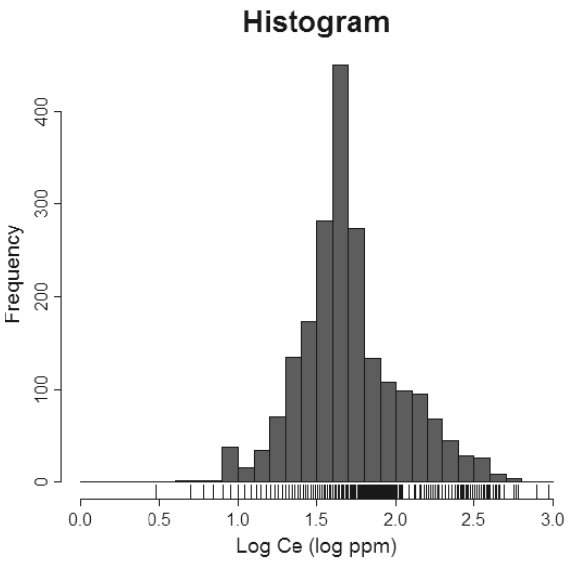

Histogram
Frequency
0
200
400
600
800
1000
0
5000
10000
15000
20000
25000
30000
Ba(ppm)

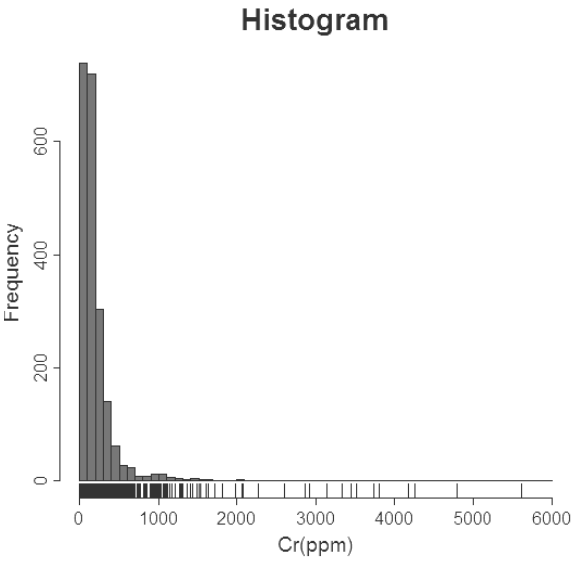

Histogram
Frequency
0
100
200
300
400
500
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Log Ba (log ppm)

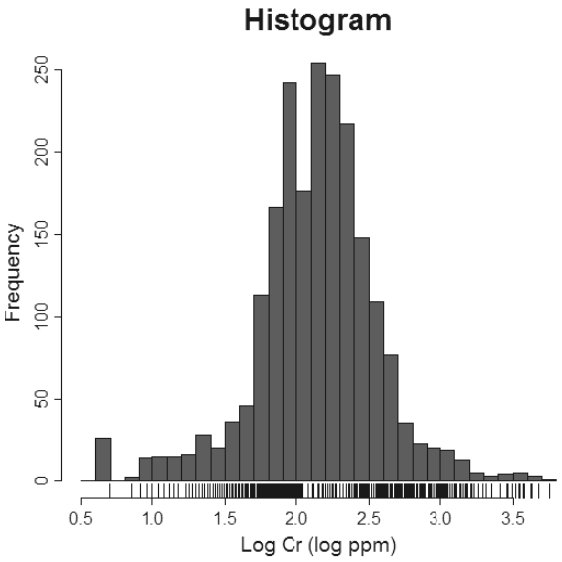

Histogram
Frequency
0
100
200
300
400
0
5
10
15
Fe (%)

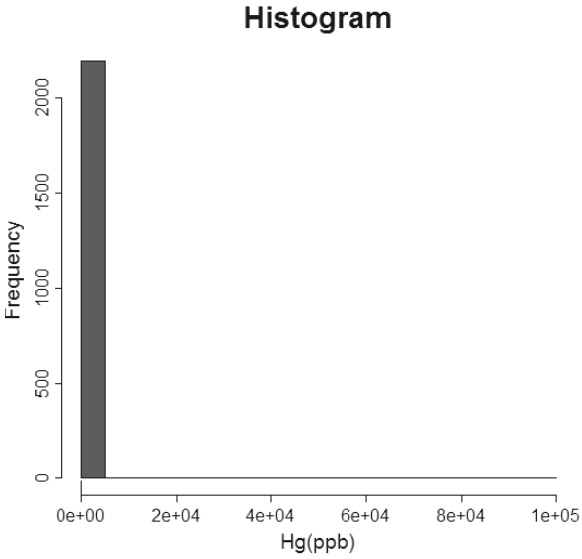

Histogram
Frequency
0
100
200
300
400
500
-1.5
-1.0
-0.5
0.0
0.5
1.0
Log Fe (log %)

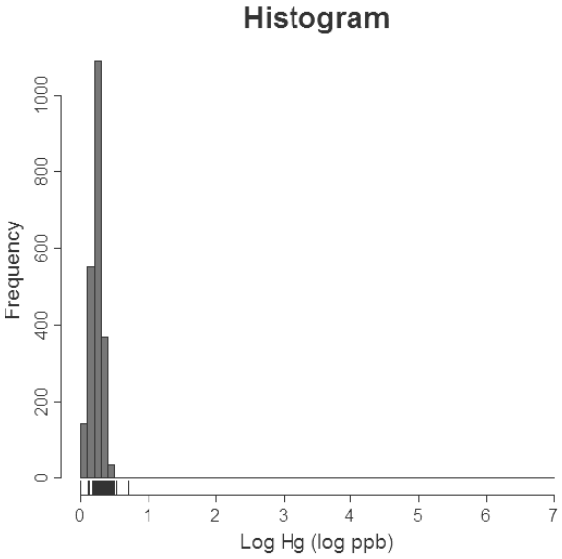

Histogram
Frequency
0
200
400
600
800
0
200
400
600
800
1000
Ce(ppm)

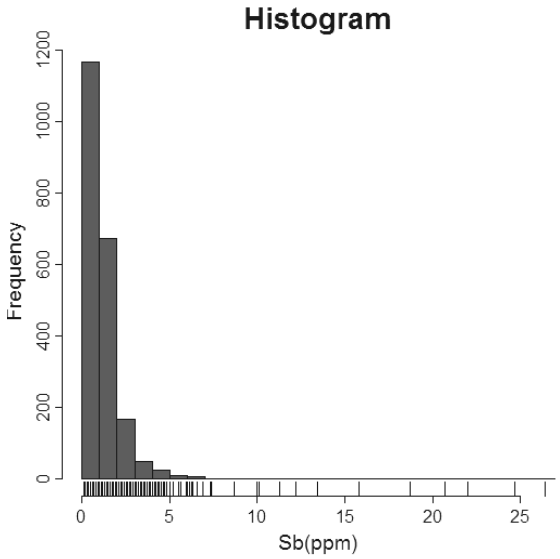

Histogram
Frequency
0
100
200
300
400
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Log Ce (log ppm)

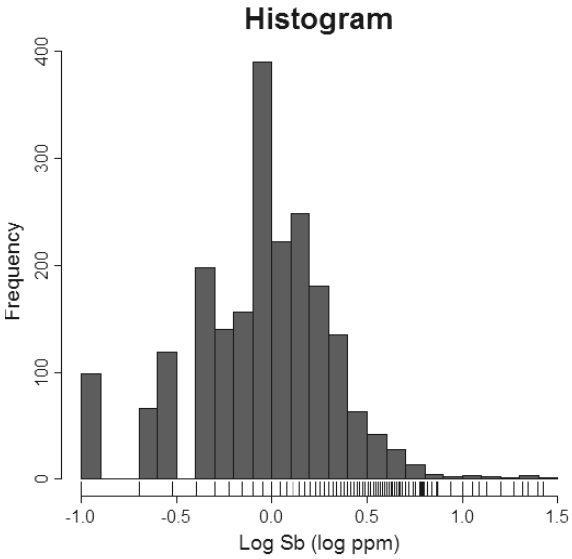

Histogram
Frequency
0
200
400
600
0
1000
2000
3000
4000
5000
6000
Cr(ppm)

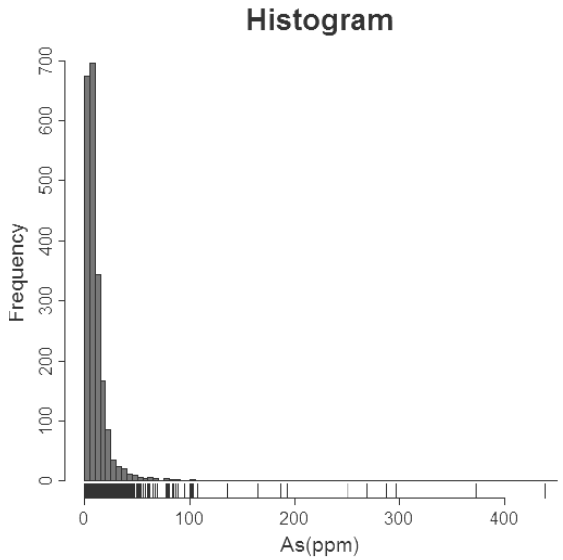

Histogram
Frequency
0
50
100
150
200
250
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Log Cr (log ppm)

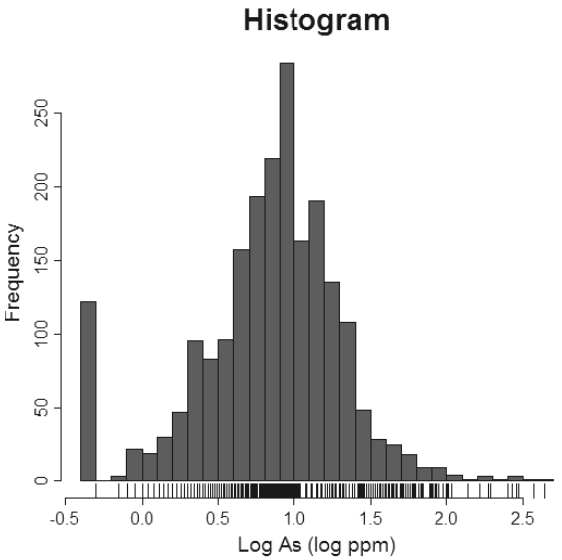

Histogram
Frequency
0
500
1000
1500
2000
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Hg(ppb)


Histogram
Frequency
0
200
400
600
800
1000
0
1
2
3
4
5
6
7
Log Hg (log ppb)


Histogram
Frequency
0
200
400
600
800
1000
1200
0
5
10
15
20
25
Sb(ppm)


Histogram
Frequency
0
100
200
300
400
-1.0
-0.5
0.0
0.5
1.0
1.5
Log Sb (log ppm)


Histogram
Frequency
0
100
200
300
400
500
600
700
0
100
200
300
400
As(ppm)


Histogram
Frequency
0
50
100
150
200
250
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Log As (log ppm)

Appendix 2-5 Error matrices for stream sediment geochemical datasets, (a) Mclust using all elements, (b) Mclust with selected elements, (c) PAM clustering using all elements, (d) PAM clustering using selected elements, (e) Mclust with Compositional Data (CoDa) approach, (f) PAM clustering with Compositional Data (CoDa) approach

Classification Data		Lithology*)						Total
		A	B	C	D	E	F	
Cluster**)	A	9632	923	1	360	861	155	11932
	B	4094	6917	620	262	3184	3623	18700
	C	225	389	136	65	609	700	2124
	D	67	2	1365	11643	123	570	13770
	E	742	3779	1561	421	4836	2734	14073
	F	3063	2294	221	1850	1412	3127	11967
Total		17823	14304	3904	14601	11025	10909	72566

Producer Accuracy		User's Accuracy	
A =	54%	A =	81%
B =	48%	B =	37%
C =	3%	C =	6%
D =	80%	D =	85%
E =	44%	E =	34%
F =	29%	F =	26%
Overall Accuracy =	**50%**		

*) A, Intrusive Rocks ; B,Volcanic Rocks ; C, Sedimentary Rocks 1; D, Sedimentary Rocks 4; E, Sedimentary Rocks 2; F, Sedimentary Rocks 3

**) A, Cluster 1; B, Cluster 2; C, Cluster 3; D, Cluster 4; E, Cluster 5; F, Cluster 6

(a)

Classification Data		Lithology*)						Total
		A	B	C	D	E	F	
Cluster**)	A	9714	290	929	915	1	152	12001
	B	1646	6978	3875	2137	285	718	13639
	C	490	910	3096	3462	1394	248	9600
	D	2795	851	1923	5477	527	138	11711
	E	3122	1382	1082	2313	367	1719	9985
	F	56	498	120	0	1330	11626	13630
Total		17823	10909	11025	14304	3904	14601	72566

Producer Accuracy		User's Accuracy	
A =	55%	A =	81%
B =	64%	B =	45%
C =	28%	C =	32%
D =	38%	D =	47%
E =	9%	E =	4%
F =	80%	F =	85%
Overall Accuracy =	**51%**		

*) A, Intrusive Rocks ; B,Sedimentary Rocks 3; C, Sedimentary Rocks 2; D, Volcanic Rocks; E, Sedimentary Rocks 1; F, Sedimentary Rocks 4.

**) A, Cluster 1; B, Cluster 2, 5, and 7; C, Cluster 3; D, Cluster 4; E, Cluster 6; F, Cluster 8

(b)

Classification Data		Lithology*)						Total
		A	B	C	D	E	F	
Cluster**)	A	3693	536	3629	1102	707	1047	10714
	B	1531	488	404	1329	1808	1422	6982
	C	983	1154	9997	455	134	1683	14406
	D	2342	1226	264	4551	4417	623	13423
	E	2359	419	207	3115	6913	6897	19910
	F	1	82	100	475	326	6151	7135
Total		10909	3905	14601	11027	14305	17823	72570

Producer Accuracy		User's Accuracy	
A =	34%	A =	34%
B =	12%	B =	7%
C =	68%	C =	69%
D =	41%	D =	34%
E =	48%	E =	35%
F =	35%	F =	86%
Overall Accuracy =	**44%**		

*) A, Sedimentary Rocks 3; B, Sedimentary Rocks 1; C, Sedimentary Rocks 4; D, Sedimentary Rocks 2; E, Volcanic Rocks; F, Intrusive Rocks

**) A, Cluster 1; B, Cluster 2; C, Cluster 3; D, Cluster 4; E, Cluster 5; F, Cluster 6

(c)

Classification Data		Lithology*)						Total
		A	B	C	D	E	F	
Cluster**)	A	9329	2858	1183	668	264	1461	15763
	B	3882	4594	2706	977	289	2370	14818
	C	1070	1481	1300	823	259	1464	6397
	D	2243	110	592	11805	1366	418	16534
	E	381	2563	853	17	1241	2128	7183
	F	918	2699	4275	311	486	3168	11857
Total		17823	14305	10909	14601	3905	11009	72552

Producer Accuracy		User's Accuracy	
A =	52%	A =	59%
B =	32%	B =	31%
C =	12%	C =	20%
D =	81%	D =	71%
E =	32%	E =	17%
F =	29%	F =	27%
Overall Accuracy =	**43%**		

*) A, Intrusive Rocks ; B, Volcanic Rocks ; C, Sedimentary Rocks 3; D, Sedimentary Rocks 4; E, Sedimentary Rocks 1; F, Sedimentary Rocks 2

**) A, Cluster 1 and 8; B, Cluster 2; C, Cluster 3; D, Cluster 4 and 7; E, Cluster 5; F, Cluster 6

(d)

Classification Data		Lithology*)					Total
		A	B	C	D	E	
Cluster**)	A	4084	2649	2232	1686	2251	12902
	B	509	1578	542	749	1402	4780
	C	0	499	11481	99	1309	13388
	D	5905	2746	91	4960	4418	18120
	E	7325	3437	255	3533	8830	23380
Total		17823	10909	14601	11027	18210	72570

Producer Accuracy		User's Accuracy	
A =	23%	A =	32%
B =	14%	B =	33%
C =	79%	C =	86%
D =	45%	D =	27%
E =	48%	E =	38%
Overall Accuracy =	**43%**		

*) A, Intrusive Rocks ; B, Sedimentary Rocks 3; C, Sedimentary Rocks 4; D, Sedimentary Rocks 2; E, Sedimentary Rocks 1 and Volcanic Rocks

**) A, Cluster 1; B, Cluster 2; C, Cluster 3; D, Cluster 4; E, Cluster 5

(e)

Classification Data		Lithology*)					Total
		A	B	C	D	E	
Cluster**)	A	3290	934	1641	2130	6096	14091
	B	310	13190	1509	891	1452	17352
	C	461	20	7232	239	166	8118
	D	3052	210	730	3441	1316	8749
	E	3914	247	6711	4208	9180	24260
Total		11027	14601	17823	10909	18210	72570

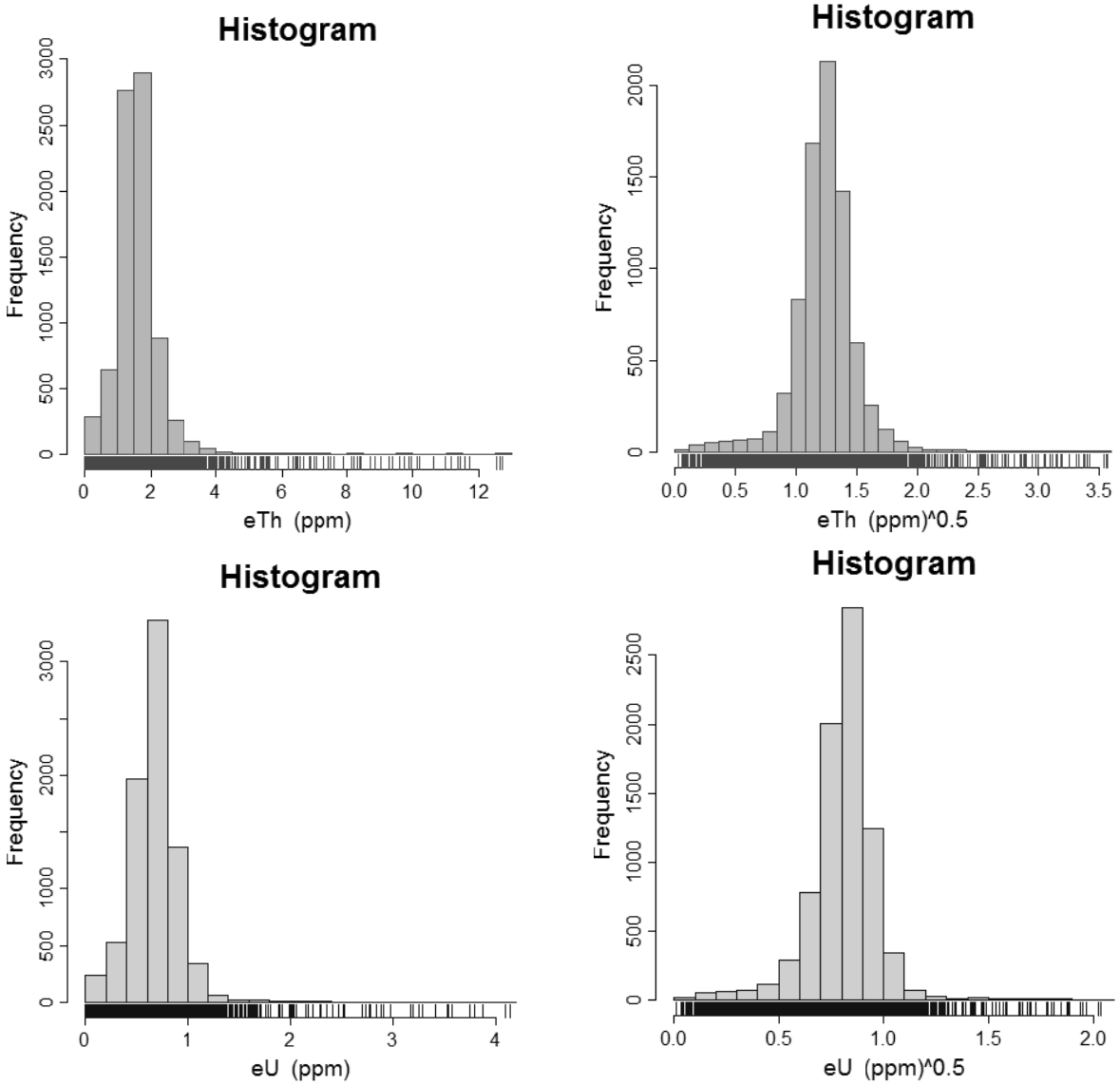
Producer Accuracy		User's Accuracy	
A =	30%	A =	23%
B =	90%	B =	76%
C =	41%	C =	89%
D =	32%	D =	39%
E =	50%	E =	38%
Overall Accuracy =	**50%**		

*) A, Sedimentary Rocks 2; B, Sedimentary Rocks 4; C, Intrusive Rocks; D, Sedimentary Rocks 3; E, Sedimentary Rocks 1 and Volcanic Rocks

**) A, Cluster 1; B, Cluster 2; C, Cluster 3; D, Cluster 4; E, Cluster 5

(f)

Appendix 3-1 Univariate graphs (histogram, box-plot and normal q-q plot) for airborne gamma-ray elements



Appendix 3-2 Error matrices for airborne gamma-ray geochemical datasets, (a)Mclust, (b)PAM clustering (c) Mclust with Compositional Data (CoDa) approach, (e) PAM clusteringwith Compositional Data (CoDa) approach

Classification Data		Feature*) A	B	C	D	E	Total
Cluster**)	A	1722	126	584	144	252	2828
	B	7	223	11	0	6	247
	C	654	17	488	61	238	1458
	D	808	0	294	93	55	1250
	E	139	3	154	14	977	1287
Total		3330	369	1531	312	1528	7070

Producer Accuracy		User's Accuracy	
A =	52%	A =	61%
B =	60%	B =	90%
C =	32%	C =	33%
D =	30%	D =	7%
E =	64%	E =	76%
Overall Accuracy =	**50%**		

*) A, Volcanic Rocks; B, Lake; C, Intrusive Rocks; D, Sedimentary Rocks 1; E, Sedimentary Rocks 2

**) A, Cluster 1,5 and 7; B, Cluster 2; C, Cluster 3 and 8, D, Cluster 4; E, Cluster 6 and 9

(a)

Classification Data		Feature*) A	B	C	D	E	Total
Cluster**)	A	1434	91	722	102	319	2668
	B	25	260	24	0	9	318
	C	752	0	317	104	20	1193
	D	758	12	124	72	93	1059
	E	361	6	344	34	1087	1832
Total		3330	369	1531	312	1528	7070

Producer Accuracy		User's Accuracy	
A =	43%	A =	54%
B =	70%	B =	82%
C =	21%	C =	27%
D =	23%	D =	7%
E =	71%	E =	59%
Overall Accuracy =	**45%**		

*) A, Volcanic Rocks; B, Lake; C, Intrusive Rocks; D, Sedimentary Rocks 1; E, Sedimentary Rocks 2

**) A, Cluster 1,3, and 8; B, Cluster 2; C, Cluster 4 and 9, D, Cluster 5; E, Cluster 6 and

(b)

Classification Data		Feature*) A	B	C	D	E	Total
Cluster**)	A	324	103	16	855	44	1342
	B	369	1133	72	306	38	1918
	C	179	77	115	784	67	1222
	D	658	212	68	1367	162	2467
	E	1	3	98	18	1	121
Total		1531	1528	369	3330	312	7070

Producer Accuracy		User's Accuracy	
A =	21%	A =	24%
B =	74%	B =	59%
C =	31%	C =	9%
D =	41%	D =	55%
E =	0.3%	E =	0.8%
Overall Accuracy =	**42%**		

*) A, Intrusive Rocks; B, Sedimentary Rocks 2; C, Lake; D, Volcanic Rocks; E, Sedimentary Rocks 1

**) A, Cluster 1; B, Cluster 2; C, Cluster 3; D, Cluster 4; E, Cluster 5

(c)

Classification Data		Feature*) A	B	C	D	E	Total
Cluster**)	A	993	54	402	60	75	1584
	B	239	146	258	24	480	1147
	C	976	39	492	111	195	1813
	D	950	61	192	101	48	1352
	E	172	69	187	16	730	1174
Total		3330	369	1531	312	1528	7070

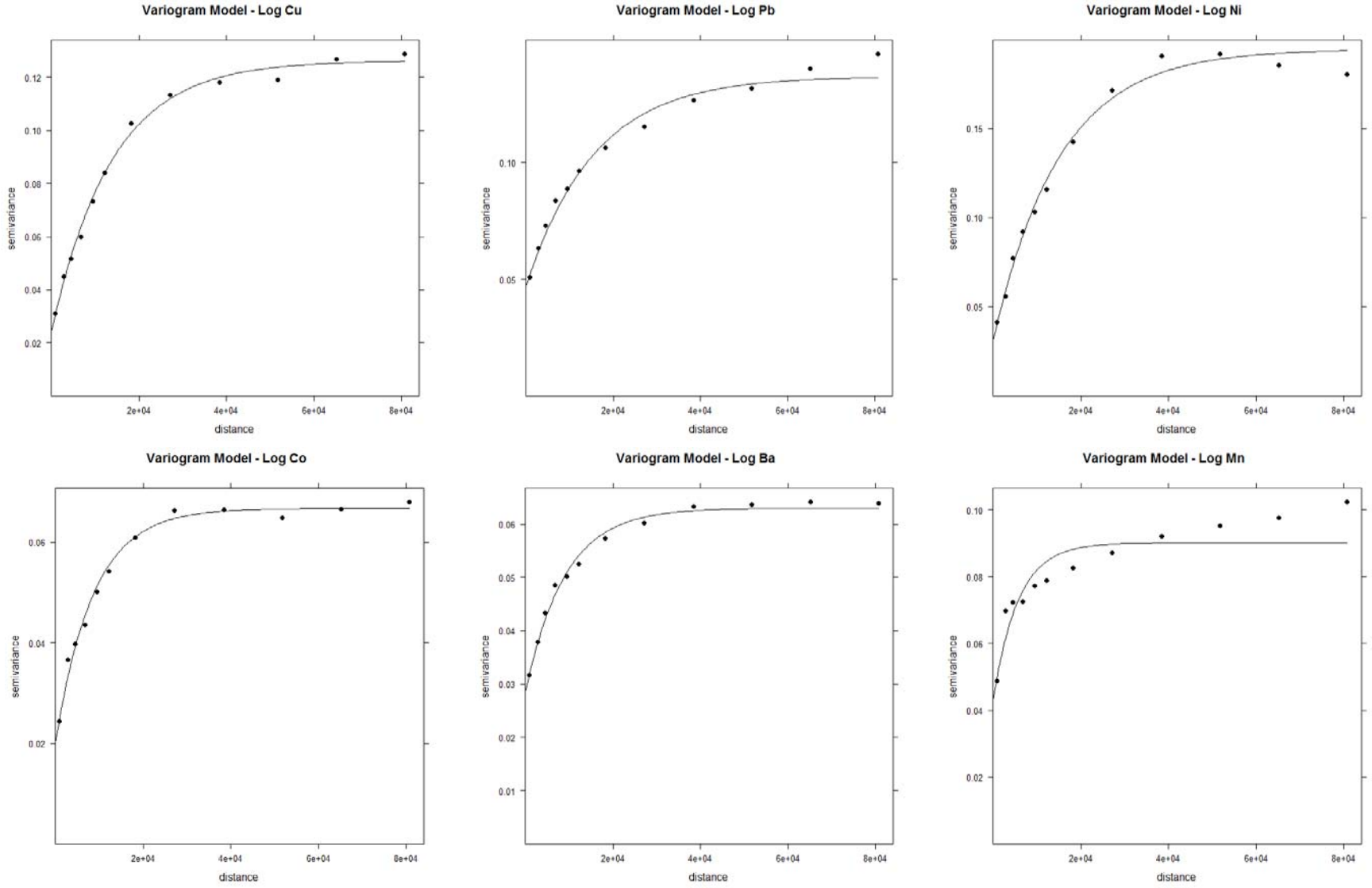
Producer Accuracy		User's Accuracy	
A =	30%	A =	63%
B =	40%	B =	13%
C =	32%	C =	27%
D =	32%	D =	7%
E =	48%	E =	62%
Overall Accuracy =	**35%**		

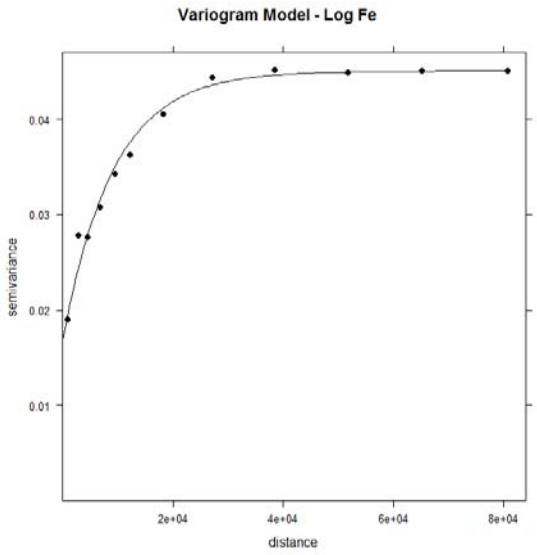
*) A, Volcanic Rocks; B, Lake; C, Intrusive Rocks; D, Sedimentary Rocks 1; E, Sedimentary Rocks 2

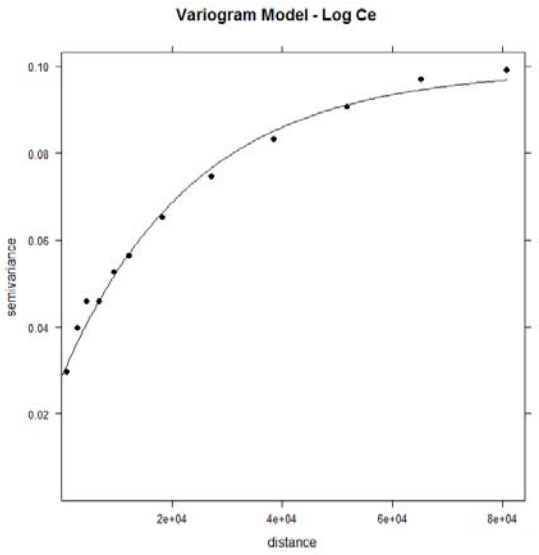
**) A, Cluster 1; B, Cluster 2; C, Cluster 3; D, Cluster 4; E, Cluster 5

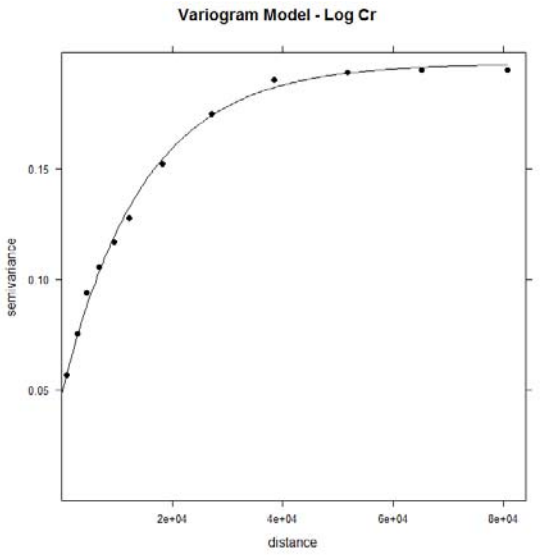
(d)

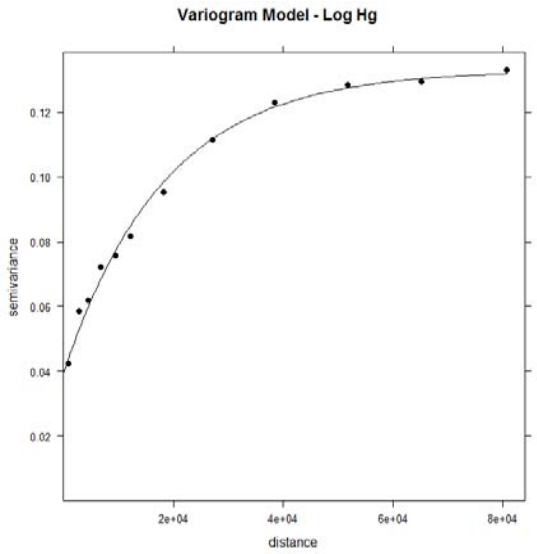
Appendix 5-1 Variogram model for base-10 logarithmic transformed of stream sediment geochemical elements

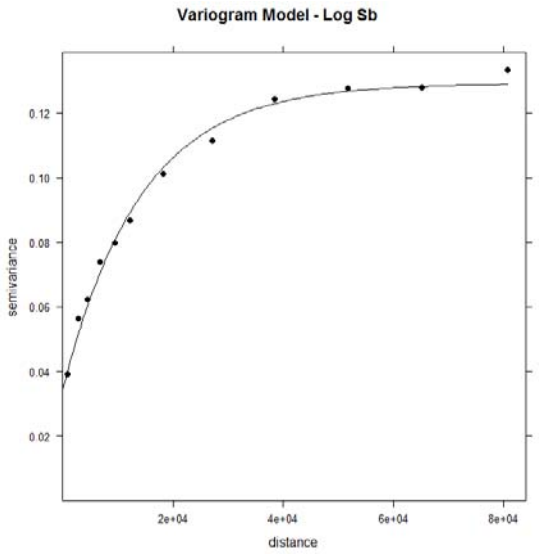


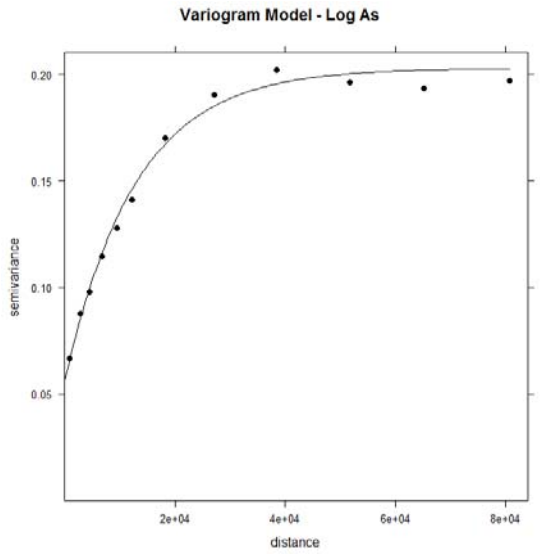

Variogram Model - Log Fe
semivariance
distance
0.01
0.02
0.03
0.04
2e+04
4e+04
6e+04
8e+04


Variogram Model - Log Ce
semivariance
distance
0.02
0.04
0.06
0.08
0.10
2e+04
4e+04
6e+04
8e+04

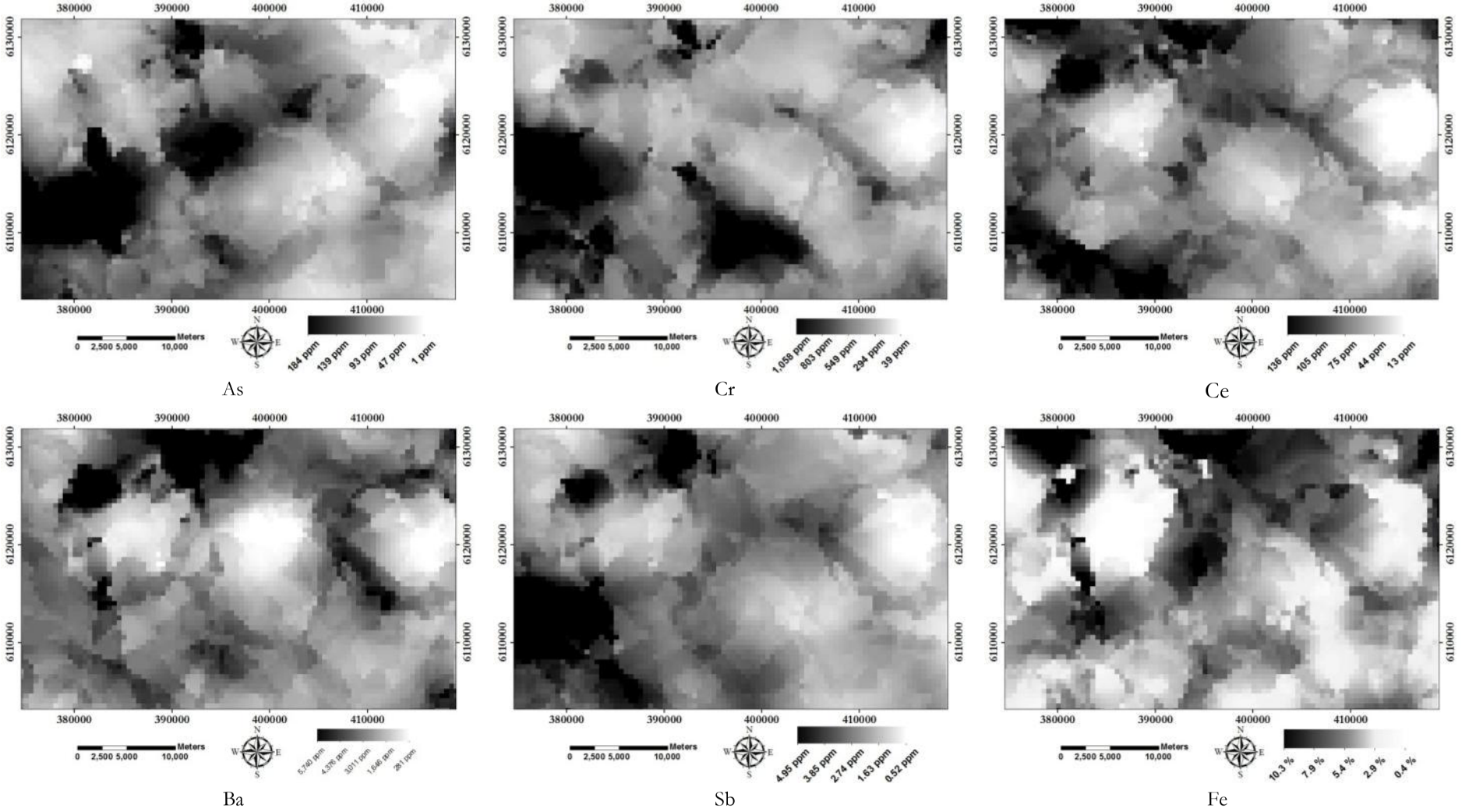

Variogram Model - Log Cr
semivariance
distance
0.05
0.10
0.15
2e+04
4e+04
6e+04
8e+04

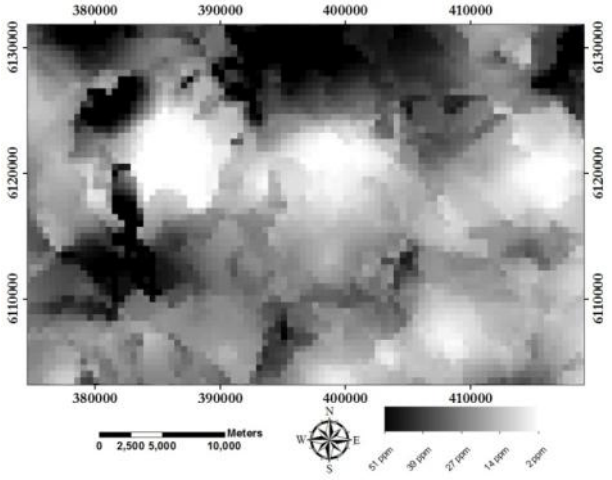

Variogram Model - Log Hg
semivariance
distance
0.02
0.04
0.06
0.08
0.10
0.12
2e+04
4e+04
6e+04
8e+04


Variogram Model - Log Sb
semivariance
distance
0.02
0.04
0.06
0.08
0.10
0.12
2e+04
4e+04
6e+04
8e+04

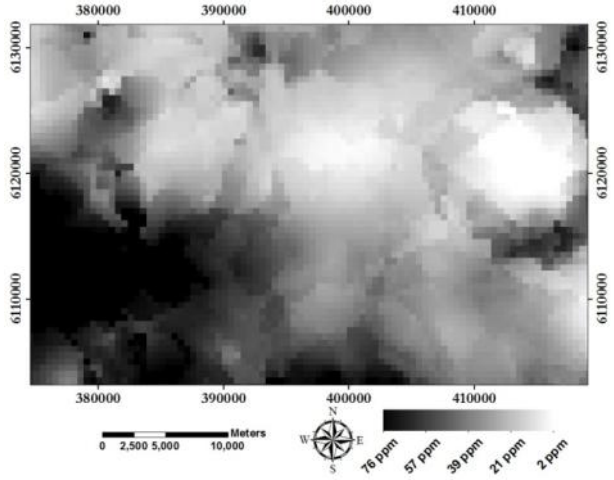

Variogram Model - Log As
semivariance
distance
0.05
0.10
0.15
0.20
2e+04
4e+04
6e+04
8e+04

Appendix 5-2 Spatial data distribution of stream sediment geochemical elements as results of universal kriging

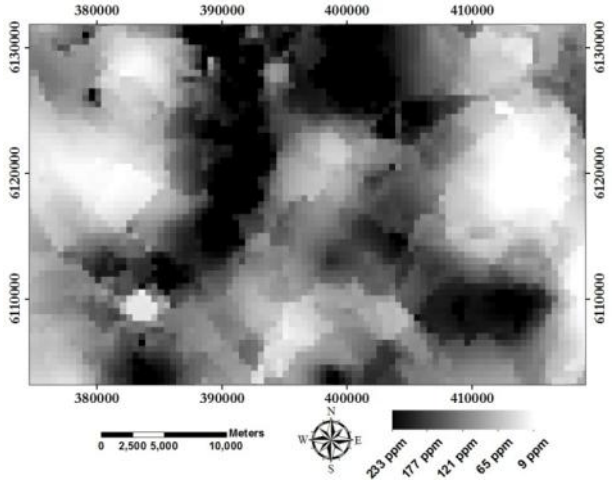




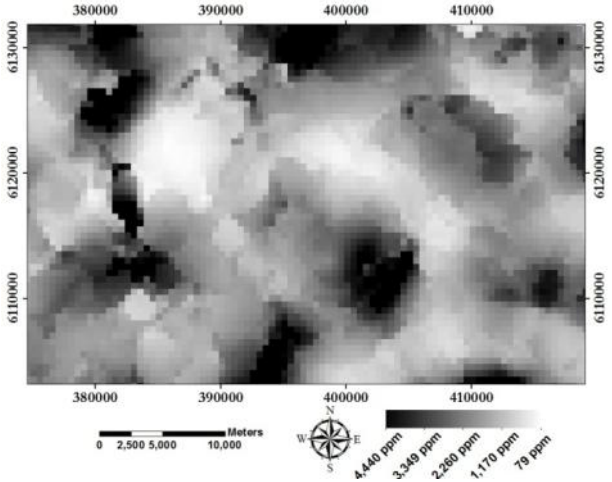
Co



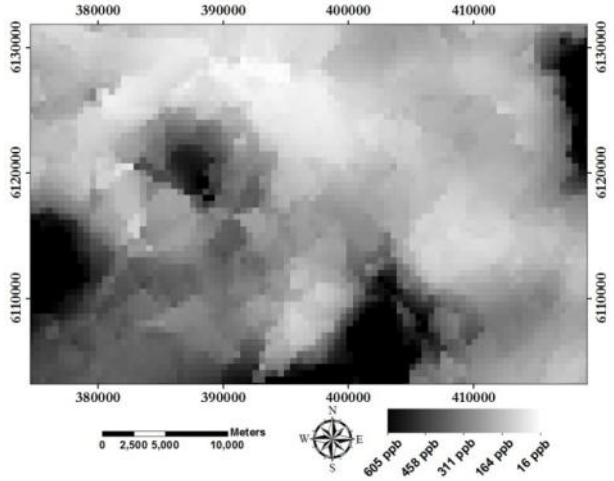
Ni



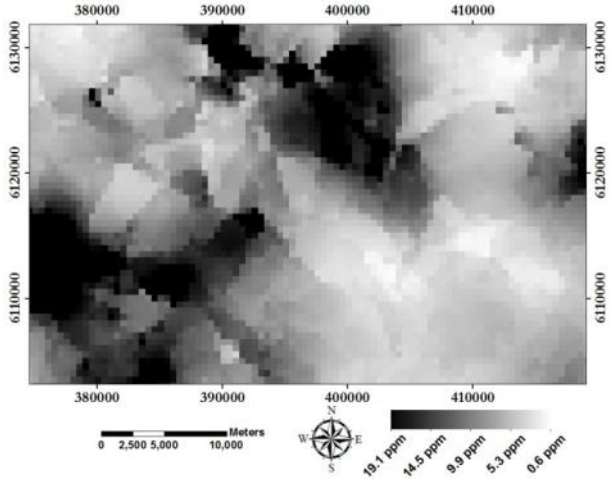
Cu



Mn



Hg



Pb

Appendix 5-3 Error matrices of clustering results using Reference Data I, (a) Mclust of Integrated Data I, (b) Mclust of Integrated Data II, (c) PAM clustering of Integrated Data I, (e) PAM clustering of Integrated Data II

Classification Data		Feature*) A	B	C	D	E	Total
Cluster**)	A	735	433	26	0	45	1239
	B	302	1680	38	123	57	2200
	C	0	584	175	0	10	769
	D	350	5	0	1073	7	1435
	E	92	281	0	83	237	693
Total		1479	2983	239	1279	356	6336

Producer Accuracy		User's Accuracy	
A =	50%	A =	59%
B =	56%	B =	76%
C =	73%	C =	23%
D =	84%	D =	75%
E =	67%	E =	34%
Overall Accuracy =	**62%**		

*) A, Intrusive Rocks ; B,Volcanic Rocks ; C, Sedimentary Rocks 1; D, Sedimentary Rocks 2; E, Lake

**) A, Cluster 1 and 5; B, Cluster 2, 3, and 9; C, Cluster 4; D, Cluster 6 and 8; E, Cluster 7

(a)

Classification Data		Feature*) A	B	C	D	E	Total
Cluster**)	A	1557	724	60	26	74	2441
	B	274	311	4	12	0	601
	C	484	7	175	8	0	674
	D	232	46	0	287	10	575
	E	436	391	0	23	1195	2045
Total		2983	1479	239	356	1279	6336

Producer Accuracy		User's Accuracy	
A =	52%	A =	64%
B =	21%	B =	52%
C =	73%	C =	26%
D =	81%	D =	50%
E =	93%	E =	58%
Overall Accuracy =	**56%**		

*) A, Volcanic Rocks; B, Intrusive Rocks; C, Sedimentary Rocks 1; D, Lake ; E, Sedimentary Rocks 2

**) A, Cluster 1,2,7, and 9; B, Cluster 3; C, Cluster 4; D, Cluster 5; E, Cluster 6 and 8

(b)

Classification Data		Feature*) A	B	C	D	E	Total
Cluster**)	A	1945	641	198	91	32	2907
	B	697	522	28	34	35	1316
	C	24	301	1031	0	12	1368
	D	303	1	16	114	3	437
	E	14	14	6	0	274	308
Total		2983	1479	1279	239	356	6336

Producer Accuracy		User's Accuracy	
A =	65%	A =	67%
B =	35%	B =	40%
C =	81%	C =	75%
D =	48%	D =	26%
E =	77%	E =	89%
Overall Accuracy =	**61%**		

*) A, Volcanic Rocks ; B, Intrusive Rocks; C, Sedimentary Rocks 2; D, Sedimentary Rocks 1; E, Lake

**) A, Cluster 1 and 2; B, Cluster 3, 5 and 7; C, Cluster 4 and 9; D, Cluster 6; E, Cluster 8

(c)

Classification Data		Feature*) A	B	C	D	E	Total
Cluster**)	A	35	710	6	191	0	942
	B	75	1650	368	653	99	2845
	C	0	7	880	214	1	1102
	D	129	604	22	413	8	1176
	E	0	12	3	8	248	271
Total		239	2983	1279	1479	356	6336

Producer Accuracy		User's Accuracy	
A =	15%	A =	4%
B =	55%	B =	58%
C =	69%	C =	80%
D =	28%	D =	35%
E =	70%	E =	92%
Overall Accuracy =	**51%**		

*) A, Sedimentary Rocks 1; B, Volcanic Rocks;C, Sedimentary Rocks 2; D, Intrusive Rocks; E, Lake

**) A, Cluster 1; B, Cluster 2, 3, and 6; C, Cluster 4 and 9; D, Cluster 5 and 7; E, Cluster 8

(d)

Appendix 5-4 Error matrices of clustering results using Reference Data II, (a) Mclust of Integrated Data I, (b) Mclust of Integrated Data II, (c) PAM clustering of Integrated Data I, (e) PAM clustering of Integrated Data II

Classification Data		Feature*)					Total
		A	B	C	D	E	
Cluster**)	A	1566	292	0	0	115	1973
	B	494	771	0	67	15	1347
	C	419	293	48	0	9	769
	D	365	2	0	1030	5	1402
	E	306	102	0	68	208	684
Total		3150	1460	48	1165	352	6175

Producer Accuracy		User's Accuracy	
A =	50%	A =	79%
B =	53%	B =	57%
C =	100%	C =	6%
D =	88%	D =	73%
E =	59%	E =	30%
Overall Accuracy =	**59%**		

*) A, Intrusive Rocks ; B,Volcanic Rocks ; C, Sedimentary Rocks 1; D, Sedimentary Rocks 2; E, Lake

**) A, Cluster 1, 5, and 9; B, Cluster 2, and 3; C, Cluster 4; D, Cluster 6 and 8; E, Cluster 7

(a)

Classification Data		Lithology*)					Total
		A	B	C	D	E	
Cluster**)	A	348	502	0	0	0	850
	B	448	1535	0	72	14	2069
	C	116	498	48	9	0	671
	D	157	169	0	242	3	571
	E	391	446	0	29	1148	2014
Total		1460	3150	48	352	1165	6175

Producer Accuracy		User's Accuracy	
A =	24%	A =	41%
B =	49%	B =	74%
C =	100%	C =	7%
D =	69%	D =	42%
E =	99%	E =	57%
Overall Accuracy =	**54%**		

*) A, Volcanic Rocks; B, Intrusive Rocks; C, Sedimentary Rocks 1; D, Lake ; E, Sedimentary Rocks 2

**) A, Cluster 1; B, Cluster 2, 3, 7 and 9; C, Cluster 4; D, Cluster 5; E, Cluster 6 and 8

(b)

Classification Data		Lithology*)					Total
		A	B	C	D	E	
Cluster**)	A	400	553	6	0	1	960
	B	934	1906	157	26	111	3134
	C	24	317	981	1	13	1336
	D	86	311	16	21	3	437
	E	16	63	5	0	224	308
Total		1460	3150	1165	48	352	6175

Producer Accuracy		User's Accuracy	
A =	27%	A =	42%
B =	61%	B =	61%
C =	84%	C =	73%
D =	44%	D =	5%
E =	64%	E =	73%
Overall Accuracy =	**57%**		

*) A, Volcanic Rocks; B, Intrusive Rocks; C, Sedimentary Rocks 2; D, Sedimentary Rocks 1; E, Lake

**) A, Cluster 1; B, Cluster 2, 3, 5 and 7; C, Cluster 4 and 9; D, Cluster 6; E, Cluster 8

(c)

Classification Data		Lithology*)					Total
		A	B	C	D	E	
Cluster**)	A	435	495	7	0	1	938
	B	817	1644	263	25	129	2878
	C	5	193	871	0	3	1072
	D	191	764	22	23	16	1016
	E	12	54	2	0	203	271
Total		1460	3150	1165	48	352	6175

Producer Accuracy		User's Accuracy	
A =	30%	A =	46%
B =	52%	B =	57%
C =	75%	C =	81%
D =	48%	D =	2%
E =	58%	E =	75%
Overall Accuracy =	**51%**		

*) A, Volcanic Rocks; B, Intrusive Rocks; C, Sedimentary Rocks 2; D, Sedimentary Rocks 1; E, Lake

**) A, Cluster 1; B, Cluster 2, 3, 5 and 6; C, Cluster 4 and 9; D, Cluster 7; E, Cluster 8

(d)+