DEVELOPMENT AND APPLICATION OF STRUCTURAL EQUATION MODELING METHOD FOR GEOCHEMICAL DATA ANALYSIS

JIANGTAO LIU

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN EARTH AND SPACE SCIENCE AND ENGINEERING
YORK UNIVERSITY
TORONTO, ONTARIO

October 2015

©JIANGTAO LIU, 2015

Abstract

A new Structural Equation Modeling (SEM) approach was proposed and the corresponding algorithms were designed and implemented for model estimation and evaluation in this research. By way of contrast to traditional SEM methods which focus on confirmatory analysis, the new SEM approach is mainly designed for exploratory analysis, which has plenty of applications in geoscience data processing and interpretation.

In order to generate an initial model for the new SEM analysis, a constrained variable clustering method was proposed based on a new index representing a type of conditional correlation, which was defined and calculated through SEM. Differently from the conventional conditional correlation coefficient, the new index was designed for measuring the quantity/percentage of the variance existing in two variables related to a response variable, rather than the level of independency of the two variables conditioned by a response variable. It can be used in Principal Component Analysis (PCA) and Factor Analysis (FA) for extracting factors restricted by a response variable. Thereby, these PCA and FA can be considered as constrained PCA and FA.

The programs designed for the new SEM are model parameters estimation, conditional correlation coefficient calculation, clustering analysis, and the SEM-based Weights of Evidence (WofE) modeling. The new SEM technology was applied to a lake sediment geochemical dataset to assist

for identification of multiple geochemical factors related to gold mineralization in a study area located in Southern Nova Scotia, Canada. The model was further applied in conjunction with the WofE method to integrate geochemical and geological information in mapping mineral potential in the same study area. The results showed that the application of the new SEM method could reduce the effect of the conditional dependency of the evidences involved in WofE.

Acknowledgements

I would like to express my special appreciation and thanks to my supervisors Dr. Qiuming Cheng and Dr. Jian-guo Wang, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless.

I would also like to thank Dr. Gunho Sohn, and Dr. Costas Armenakis for serving as my supervisor committee members. I want to thank you for your invaluable comments at each year's research evaluation. I would extend my sincere thanks to the examining committees, Dr. Baoxin Hu, Dr. Eric Grunsky, Dr. Jarvis Gary, and Dr. Dong Liang, for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions. I would like to take this opportunity to thank the Graduate Program Assistant and Administrative Assistant of ESSE at York University Mrs. Marcia Gaynor and Mrs. Paola Panaro, for your kindly help in the past six years. I would like to thank my friends in GIS research lab, Siping Xu, Xitao Xing, Wenlei Wang, Jie Zhao, Mangen Li，Gaifang Wang, Jingwei Gao, Deyi Xu, Quanming Liu, Yongzhi Wang and Guoxiong Cheng, for the happy time with all of your guys.

At the end I would like express special appreciation to my wife Huili Peng and my child Pengzheng Liu, without their trust in past years, this thesis would not have been possible.

Definition of Notation

x	Independent measurement variable
y	Dependent measurement variable
ξ	Latent exogenous variable
η	Latent endogenous variable
δ	Modeling error associating the latent exogenous variables
ε	Modeling error associating the latent endogenous variables
ζ	Modeling error associating the structural model
X	Vector of independent measurement variables (x)
Y	Vector of dependent measurement variables (y)
Ξ	Vector of latent exogenous variables (ξ)
H	Vector of latent endogenous variables (η)
Δ, E, Z	Vectors of the modeling error terms (δ, ε, ζ)
Matrix	
B, Γ	Coefficients (β, γ) in structural model
M	Relationships (μ) between the observed variables and the latent variables
Probability	
$P(A)$	Probability of event A
$P(A \cap B)$	Probability that of events A and B

$P(A \cup B)$	Probability that of events A or B
$P(A \mid B)$	Probability of event A given event B occurred
$var(x)$	Variance of random variable x
σ^2	Variance of a random variable
σ_x	Standard deviation of random variable x
$cov(x,y)$	Covariance of random variables x and y
$R_{y(x_i,x_j)}$	A new conditional correlation between x_i and x_j under the restriction of y.
$\rho_{x,y}$	Correlation coefficient of variables x and y
$\rho_{y.x_i,x_j}$	Multiple correlation coefficient between y and $\{ x_i , x_j \}$
$P_{x,y}$	Direct effect of x to y
λ	Eigenvalue
$R^2_{(x_i,x_j)}$	The goodness of fit between x_i and x_j
Operators	
$\sum$	Summation - sum of all values in range of series
$\sum\sum$	Double summation
Others	
SEM	Structural equation modeling
PLS-SEM	Partial least square SEM
CB-SEM	Covariance based SEM
MLR	Multiple linear regression

Table of Used Geochemical Elements

Ag	Silver
As	Arsenic
Au	Gold
Cu	Copper
F	Fluorine
Li	Lithium
Nb	Niobium
Pb	Lead
Rb	Rubidium
Sb	Antimony
Sn	Tin
Th	Thorium
Ti	Titanium
W	Tungsten
Zn	Zinc
Zr	Zirconium

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Statistics in geoscience

With the development of detection technology and the support of geographic information systems (GIS), the field of geo-data processing becomes more and more popular (Ali et al., 2007; Atekwana and Slater, 2009; Campo, 2012; Hart and Martinez, 2006; Jensen, 2009; Madden and Julian, 1994; Minasny et al., 2008; Mouillot et al., 2014; Nykiforuk and Flaman, 2009; Rao et al., 2008; Rollinson, 2014; Selva et al., 2014; Weng, 2014; Wielicki et al., 1996). In recent years, the volume of geo-data from multiple sources (e.g. real-time flood data, surface- and ground-water data; and information related to natural hazards, etc.) has increased rapidly. Also, the modern web technologies are making the utilization of geological and geospatial data increasingly global, accessible and instantaneous. Moreover, global energy and mineral crises, abnormal climate, and natural hazards etc. compel the geoscientists to provide more detailed and timely useful information from massive databases. Undoubtedly, it is a big challenge to extract and fuse information for useful applications in geoscience, which can be achieved using computer hardware (e.g. cloud technology), geoinformatics software (e.g., ArcGIS) and statistical methods. By considering the importance of extracting useful information from massive databases, the geo-data processing technology was listed as one of the future directions in solving the global challenges during 2007-2017 by the U.S. Geological Survey (GSurvey,

2007).

Mathematical methods have been playing a pivotal role in geo-data processing for several decades. Obviously, most of the techniques in quantitative geology are involving statistical approaches. Interest in areal or block averages for ore reserves in the mining industry led to the development of geo-statistical analyses in the 1950s, which aimed to provide quantitative descriptions of natural variables distributed in space and/or time. The development and proliferation of powerful personal computers have aided the widespread distribution of statistical software and sharable data over the internet through organizations such as the International Association for Mathematical Geosciences (IAMG). Some mathematical techniques have become standard practices in some geo-data processing. For example, the Principal Component Analysis (PCA) method can be used for extraction of geochemical factors (Cheng et al., 2011; Wang and Cheng, 2008), the Weight of Evidence (WofE) method can be used for mineral exploration (Agterberg, 1989; Agterberg and Bonham-Carter, 1990; Bandeen-roche et al., 1997; Bonham-Carter et al., 1988; Bonham-Carter et al., 1989; Bonham-Carter, 1994), and the Concentration Areal – Spectral Areal (C-A or S-A) methods have applied to detection of geological anomalies (Cheng et al., 2000; Cheng, 1999, 2007a, 2007b, 2012a, 2012b, 2014). The applied statistics is especially important in the petroleum and mineral industry, where it becomes instrumental in identification of anomalous mineralization and thus, provides more precise target areas for exploration.

The availability and abundance of data present both opportunities and challenges to scholars,

practitioners, and the governmental agencies. Although high volume of geological data is now readily available, there is a dearth of professionals utilizing their analytical skills to understand and extract useful information from the data. Geo-data analysis requires rigorous scientific approaches, which rely on knowledge of statistics, measurements, logic, theory, experience, and situational context (Harris et al., 2015; Gao et al., 2014; de Caritat and Grunsky, 2013; Grunsky et al., 2014; Savinykh and Tsvetkov, 2014; Wathne et al., 1996). Therefore, it is crucial to extract useful, accurate and timely information from geo-data sets that use multiple sources (e.g. GIS and Remote Sensing) and levels (e.g. multi-level resolution and completeness data). For example, the WofE method is one of the most popular methods for information fusion in mineral exploration (Agterberg, 1989; Agterberg and Bonham-Carter, 1990; Agterberg and Cheng, 2002; Bonham-Carter et al., 1988; Cheng and Agterberg, 1999). With this method, an evidence is considered as a dependent variable when being extracted from a multi-source data (e.g. geochemical element variables, geophysics variables, and geologic features). It is considered as an independent variable in calculating of the posterior probability of mineralization. Because the evidences are estimated without considering the conditional independence (CI), it makes them hard to meet the CI requirement when they are adopted in the calculation of the posterior probability. In order to solve this problem, there are several methods that have been proposed for testing the CI (Bonham-Carter et al, 1989; Agterberg, 1992; Bonham-Carter, 1994; Agterberg and Cheng, 2002) and reducing the effect of it in mineral prediction (Bonham-Carter,1994; Journel, 2002; Krishnan et al. 2004; Polyakova and Journel, 2007; Cheng, 2008, 2015; Deng, 2009, 2010a, b; Zhang et al., 2009; Agterberg, 2011; Schaeben, 2012).

Structural Equation Modeling (SEM) can be defined as a class of methodologies that seeks to represent hypotheses about the means, variances, and covariances of observed data in terms of a smaller number of 'structural' parameters defined by a hypothesized underlying conceptual or theoretical model (Kaplan, 2008). It provides a statistical approach to testing hypotheses about relations among observed and latent variables (Hoyle, 1995), which has been widely applied in social and behavioral sciences. This method may be applied to address the multi-level, and multi-model problems in geo-data processing. The definition of SEM was articulated by geneticist Sewall Wright (1921), economist Trygve Haavelmo (1943) and cognitive scientist Herbert A. Simon (1977), and formally defined by Judea Pearl (2000) using a calculus of counterfactuals. As shown in ***Table** 1.1*, SEM is considered as a second generation statistical technology (Fornell, 1987; Lohmöller, 1989; Hair Jr et al., 2013).

***Table** 1.1 Organization of multivariate methods (Hair Jr et al., 2013)*

	Primarily Exploratory	**Primarily Confirmatory**
1st generation	Cluster analysis Exploratory factor analysis Multidimensional scaling	Analysis of Variance Logistic regression Multiple regression
2nd generation	PLS-SEM	CB-SEM, including Confirmatory factor analysis

In recent years, proliferation of computer hardware and software along with more user-friendly interfaces has enabled SEM to become more and more popular. Its theory and statistical properties have well been developed and found plenty of applications across many disciplines including education, psychology, sociology, and environmental epidemiology (Browne and

Arminger, 1995; Muthen, 1984; Sánchez et al., 2005; Yuan and Bentler, 2000; Yuan and Bentler, 1997). In mainstream statistical journals, SEM theory has also appeared under the terms “mean and covariance structures” and “latent variable models” (Bandeen-roche et al., 1997; Jöreskog, 1970, 1978; Lee and Shi, 2001; Sammel and Ryan, 1996; Yuan and Bentler, 1997). The advantages of SEM over first generation statistical technologies can be characterized as follows (Hair Jr et al., 2013):

1. *SEM allows making use of several indicator variables per construct simultaneously, which leads to more valid conclusions on the construct level.* Using other methods of analysis would often result in less clear conclusions, and/or would require several separate analyses.

2. *SEM allows modeling and testing complex patterns of relationships, including a multitude of hypotheses simultaneously as a whole (including mean structures and group comparisons)*.

3. *SEM provides a confirmatory approach for complex models.* For hypothesis, simple statistical procedures usually provide tests on the basis of explained variance in single criterion variables. This may not be suitable for evaluating complex models containing a multitude of variables and relationships. In contrast, SEM allows to test complex models for their compatibility with the data in their entirety, and allows to test specific assumptions about parameters for their compatibility with the data. This allows for global assessment, local assessment and exploratory suggestions for potential model improvements (modification indices).

Geochemical data is typically reported as compositions, in the form of some proportions such as weight percents, parts per million, etc., subject to a constant sum (e.g. 100%, 1,000,000 ppm). The relations of elements in components are different with which in real sample space, i.e. in correlation analysis, the compositional data may result a type of spurious correlations among components if they are processed as unconstrained vector, which was pointed out by Karl Pearson in his 1897 paper (Pearson, 1896) firstly. For overcoming the problems of compositional data in geo-data processing, Aitchison (1982, 1984) introduced proper representations of a composition in order to have all the relevant information contained in a set of log-ratios. The additive log-ratio (ALR) (Aitchison, 1986) is one of the methods proposed for compositional data transformation, in which one part of compositions is chosen as the common denominator of all the ratios. In order to overcome the inconvenience of ALR which depends on the permutation of the components, Aitchison (1986) introduced a centered log-ratio transformation (CLR) to represent a D-part composition using D CLR-coefficients. The CLR transform does not result in an orthonormal space. Thus standard parametric modelling procedures cannot be applied. Egozcue et al. (2003) proposed the isometric log-ratio transformation (ILR) to work with orthonormal bases and their corresponding coordinates. More information about compositional data analysis can be found in recent publications (Pawlowsky-Glahn et al., 2015; Pawlowsky-Glahn and Buccianti, 2011).

1.2 Motivation

The motivation to explore the use of SEM in geo-data processing mainly comes from the followings:

1) *A structural equation model, as a combination of a measurement model to define latent variables using one or more observed variables, and a structural regression model to link latent variables together, is found on several multivariate statistical analysis methods, e.g. Factor analysis, PCA, Multi-linear regression, Path analysis, Latent variable analysis, which are very popular in geoscience data processing.*

SEM is a combination of many multivariate statistical models, each of which can be considered as a special case of SEM. For example, in SEM, while the measurement model is analogous to factor analysis, the structural model may be considered analogous to multi-linear regression. Since both the factor analysis and multiple linear regression methods are widely used in the geo-data analysis, the SEM method may be more suitable for cases where the former methods can be applied.

Also, SEM is a popular concept in many disciplines. As a fact, since the year of 2000, hundreds of papers relating to SEM have been published. The statistics showed that there were relative fewer papers related to SEM in the geosciences (***Fig** 1.1*) from 2000 to 2009 (McArdle and Kadlec, 2013), although it has widely been applied in social sciences, arts and humanities. It should be noted that SEM is a class of methodologies, part of this concept has been discussed

and applied in geo-data analysis (Harris et al., 2015), but the SEM discussed in current research is a narrowly defined method which include at least one measurement model and one structural model. For applying SEM in geoscience data analysis, one of the main difficulties is that a predefined model is required in SEM calculation, but it is usually hard to precisely define before analysis.

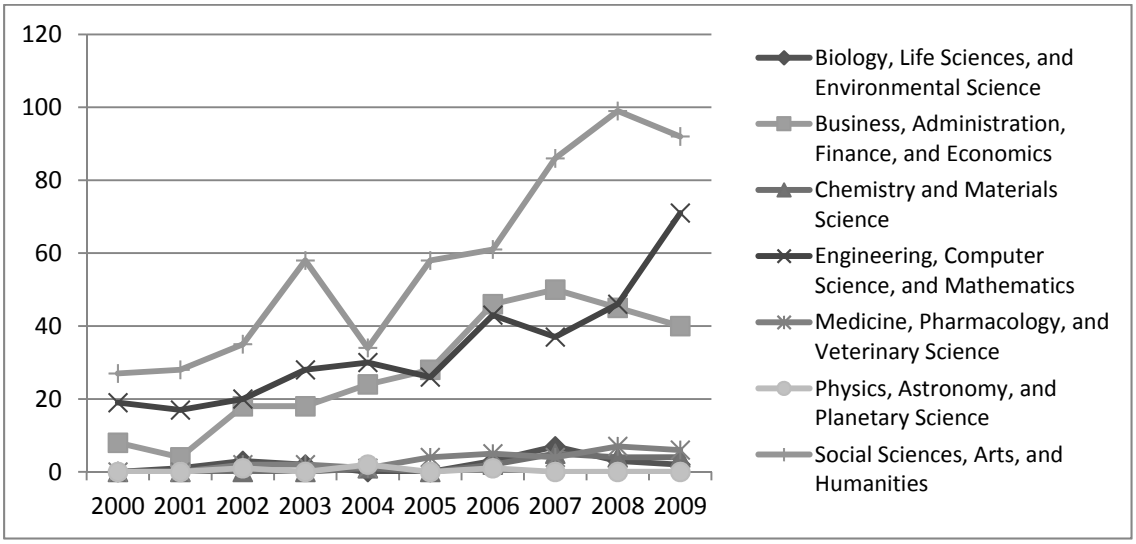


***Fig** 1.1 A Chart of SEM publications from 2000 - 2009.(McArdle and Kadlec, 2013)*

2) *In order to handle multi-level and multi-process problems to address the challenges in geo-data processing, SEM might be potentially adoptable due to its capability of combining concept and process.*

The major challenge of combining concept with process is not only about creating a model to express the idea but also estimating a set of parameters to match the designed model. Taking the method of ordinary weights of evidence (WofE) as an example, the conventional WofE

integrates multiple evidence layers that are of conditional independence from each other with respect of a point event. An evidence can be considered as a latent variable that cannot be measured/observed directly but extracted from raw data. The process of constructing evidences is analogous to factor analysis. However, it is considered as an independent variable when used in WofE for estimating the posterior probability of point event. This process is analogous to logistic regression. The traditional WofE has been implemented through two separate modeling processes (***Fig*** *1.2*): extracting the evidences from geochemical, remote sensing, and geophysical data; and then combining the evidences by a

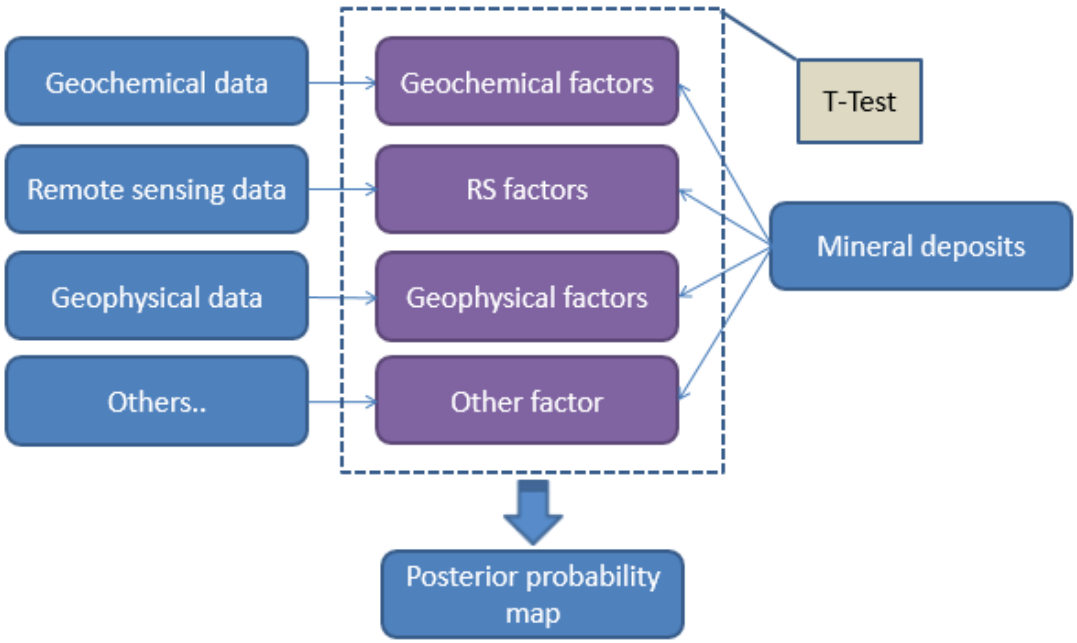


Fig *1.2 The traditional WofE calculation process in geo-data integration for mineral prediction*

logistic model. Since this method estimates the evidences based on the rules that express the

main information in source data, the extracted evidences are hard to meet the CI assumption of WofE method. There are several ways developed in the literature for solving the problem. In the current work it explores alterative solution to partially solve the problem by creating an SEM model to combine the factor analysis and regression, and further estimate the parameters with a global optimum function.

3) Application of *SEM in geoscience is an interdisciplinary approach and has great potential for future research in the field.*

As two popular methods in geo-data processing, PCA and FA can be considered as two special cases of SEM, which are processes to extract several "latent variables" through a measurement model. At the same time, SEM has been extended many subjects already, For example, similar as random forest and artificial neural network, SEM tree can be used for a decision (Oztekin et al., 2011; Brandmaier et al., 2013), which provides a data-driven but theory-constraint search in model space. Similar as the Dempster-Shafer, some of the SEM's extension can be used for information integration too (Punniyamoorthy et al., 2011; Steinberg, 2009), i.e. the proposed SEM based WofE method in current research.

The broad successful SEM applications in social science indicate that SEM is not only about establishing a mathematic model in data processing, but also about developing a software tool, which includes to design the model in graphics, output calculation results in tables and graphics and manage projects. For the same reason, the application of SEM in geosciences will depend on both of the correct, efficient mathematic model and the software which is in line with the

practice of geoscientist. This demands for a good understanding in statistics, geoscience and computer science as well, which definitely increases the challenge of this research, but on the other hand, it provides more avenue for innovation because its interdisciplinary nature.

1.3 Objectives

The overall goal of current research is to expand the application of SEM into geo-data processing and analysis. Based on a systematic study of the SEM concept, model, software and application, the feasibility of this method in geo-data processing will be tested and an efficient software tool for geoscience research will be provided.

The research of this proposal includes the following THREE objectives:

1) *Apply SEM in geo-data processing and analysis to address the problems in multi-level models with latent variables,* i.e. the implementation of WofE in mineral exploration. The successful application of SEM in mineral prediction may provide not only a new tool for geo-data processing, but also a new concept for information extraction and integration. The core idea in the SEM approach is to use a global optimum target function instead of gradually optimal methods to estimate a geological model involving multiple levels and processes.

2) *Compose and develop a new SEM for geo-data analysis and accordingly design the algorithm and program for implementation*: The second objective of this research is to propose the modeling methods and algorithms using C# and R programming languages, which includes

algorithm design (regression, factor analysis, and clustering), GIS function and user interface, etc. The proposed SEM method is designed as a software package in R, which allows efficient analysis of geo-data.

3) *Validate the new method and software through case studies.* The newly designed method and the developed software tools will be validated through a geochemical dataset obtained from 671 lake sediment samples. This process would include the identification of geochemical factors in regression to gold mineralization endogenous variables, and the integration of geochemical factors and geological factors for mineral potential mapping in the southwestern Nova Scotia, Canada.

In order to achieve the objectives, the following tasks have been conducted:

1) *Exploration of the challenges of SEM in geoscience applications*: Although SEM has many advantages over the traditional statistical techniques and great potential in a broad range of the applied scientific research, its application in geological data processing comes with some statistical and interpretational challenges, primarily due to the inherent nature of geoscience data (e.g., the problem in model identification / parameter estimation based on fuzzy model). A multitude of parameters (path coefficients, factor loadings, variances, etc.) corresponding to various hypotheses are estimated simultaneously so that the empirical relationships between the observed variables can be reproduced by the model in a robust manner. This is possible when the empirical data can provide adequate and correct information to estimate all these parameters. However, in most cases, geoscience data is a combination of "cascading" and/or "nesting",

further complicated by "masking" and "swamping" of under sampled processes. Therefore, part of this research will be to devise a method to extract an optimal geological process/model from a series of candidate processes/models using SEM.

2) *Generation of an initial model*: An initial model is required for a SEM application both for exploratory and confirmatory analysis. It may not be a big problem for some confirmatory analysis if the target of the research is to test some hypothesis, where the researchers usually have a model beforehand and then collect data related to their model, which is usually tested by some parameters such as the goodness-of-fit. But for some analysis in geoscience (e.g., mineral prediction, oil pipe route designation, urban planning, etc.), the relationships between different variables in a preliminary model is not clear prior to the analysis. For example, in mineral exploration, there exists some knowledge of the relationships between the ore-mineralization controlling factors. However, the detailed relationships among most of the factors are not obvious. Therefore, “generating the initial model” becomes the first problem which needs to be solved in order to apply SEM for geo-data processing.

3) *Evaluation of the new model parameters*: There have been two types of SEM: Covariance-based SEM (CB-SEM) and Partial Least Squares-based SEM (PLS-SEM) developed for different applications (Hair Jr et al., 2013). CB-SEM is usually applied to model testing or confirmatory analysis. In mineral exploration, this means creating a model based on previous studies, conclusions and testing whether the model hypothesis is true. Hoyle and Panter (1995) recommended some indexes about overall model fit (e.g. unadjusted chi-square) for hypothesis testing. The PLS-SEM method is mostly applied in the exploration analysis using an undecided model, where a regression of mineral-related targets can be specified to extract some latent

variables, which may subsequently be used in mapping the probability of mineral occurrence. The challenges faced in model estimation and evaluation chiefly come from the special nature of geo-data. For example, the geochemical data are influenced by combination of multiple sources and geological processes. Therefore, the variance and covariance of the data was decided by the source/ process providing the largest information, which maybe not the one we are interested in. However, in the CB-SEM method, the evaluation of the SEM model is based on Chi-square and some fitting indexes, all of which are calculated from the variance/covariance matrix. This method directly affects the final calculation if the variance/covariance does not represent the required information. A modified PLS-SEM as proposed in this research, may remove redundant information and extract respondent latent variables for a specified target.

4) *Application of SEM concept into WofE for mineral prediction*: The weight of evidence (WofE) is an artificial intelligent quantitative method based on the application of Bayes' rule. The method, originally designed for a non-spatial application in medical diagnosis (Agterberg, 1989), is one of the most popular models using Bayes' theory of conditional probability to quantify spatial association between evidence layers (or geological factors) and known mineral occurrences (Agterberg, 1989; Bonham-Carter, 1994; Carranza, 2004; Cassard et al., 2008; Porwal et al., 2010). Besides, WofE is applied to map landslide sensitivity evaluation (e.g., Lee and Choi, 2004; Neuhäuser and Terhorst, 2007; Regmi et al., 2010; Cervi el al., 2010), and ecology mapping (e.g., Romero-Calcerrada and Luque, 2006; Cho et al., 2008; Romero-Calcerrada et al., 2010; Gorney et al., 2011).

Condition of independence (CI) is one of the most important problems within the ordinary WofE (Agterberg and Bonham-Carter, 1990). It has been a topic of research for the last few years and the problem of CI has been solved with various theoretical and practical solutions. Journel (2002) and Krishnan et al. (2004) put forward a new geostatistical model: Tau model, which attempted to address the restriction of CI. This has led to a number of weighted and stepwise modified models for WofE. Polyakova and Journel (2007) suggested the new Nu model as an alternate of the Tau model, which involved an extra parameter to measure the data interaction. Some of the limitations of the weights of evidence, Tau and Nu models are discussed in Schaeben (2012). Agterberg (2011) proposed a modified WofE model to estimate the weights for adjusting the dependency of evidences which applies logistic regression. The regression coefficients resulted from the logistic regression could be used as Tau weights to modify the ordinary weights of evidence. Zhang et al. (2009) proposed a similar approach to estimate the Tau weights using ordinary linear regression in association with the posterior logits resulting from weights of evidence. Several modified WofE methods were also developed towards significant reduction of the CI's effect (Deng, 2009, 2010a, b; Cheng, 2008). A new solution to overcome the CI problem was proposed by Cheng (2015) on sequential overplay of evidences accounting the dependency of the evidences using a new model - BoostWofE, based on ad boosting algorithm. All above solutions for solving CI problem is based on predetermined evidences.

In this research, a SEM-based WofE model is proposed to extract evidences with less effect of conditional independence, which in turn can improve the accuracy of the posterior probability of WofE.

1.4 Outline

Chapter 1 provides a general introduction to current research, which includes the research motivation, objectives and contributions.

Chapter 2 presents supportive materials for the background information related to SEM including its history and major research progresses in recent years. This chapter provides a more detailed introduction of the SEM algorithm focused on estimation of model parameters. Two methods, the Covariance Based SEM (CB-SEM) and Partial Least Square SEM (PLS-SEM), are discussed.

Chapter 3 provides the description of the data (for case study and validation) and the software for algorithm design. The attributes of the data, distribution of variables, and methods for normalizing data along with the geological background and previous research conducted in the study area are also discussed in this chapter.

Chapter 4 proposes a new SEM method, which combines the principles of cluster and regression analysis. The proposed SEM in Chapter 4 is applied to the extraction of three gold mineral related factors based on the data set in Chapter 3. Moreover, the method for parameter estimation and generation of an initial model is introduced, which involves calculation of optima based on a global target function.

Chapter 5 discusses an application of the SEM concept in WofE to reduce its conditional

independence (CI) problem in mineral potential mapping. It can be considered as an SEM application in geo-data integration in addition to the extraction of geo-factors.

Chapter 6 and 7 introduce a “supervised” variable clustering method based on a new index to solve the problem of creating an initial model. The new index is a conditional correlation coefficient of two variables under the restriction of regression to a response variable. The new variable clustering is proposed based on Clustering around Latent Variable (CLV) method, which includes the hierarchical and partial clustering algorithms. There are two differences between the new method and CLV. Firstly, the distance between two variables is defined as the new index proposed in Chapter 6, rather than their covariance. Secondly, the centroid of each cluster is a prediction for a response variable from the variables in each cluster, rather than the first principal component. Their differences are discussed in detail in Chapter 7. A computer program was designed for calculating the new index and solving the clusters in Chapter 6 and 7.

Chapter 8 concludes the research, highlights the contributions and points out the remaining challenges and tasks for future research.

Chapter 2 Structural Equation Modelling: General Considerations

2.1 Introduction

Structural equation modeling (SEM) is a statistical technique for testing and estimating causal relations using a combination of statistical data and qualitative causal assumptions (Pearl, 2000). With the development of the advent of computer science and engineering, particularly in recent years with the widespread access to many more method due to the user-friendly interfaces with technology-enabled knowledge, the application of SEM has been expanded dramatically. The theory and statistical properties of SEM are well developed but are scattered throughout several fields of research, particularly in education, psychology, sociology, and environmental epidemiology (Browne and Arminger, 1995; Muthen, 1984; Sánchez et al., 2005; Yuan and Bentler, 2000; Yuan and Bentler, 1997). SEM theory has also appeared in mainstream statistical journals through the terminology of mean and covariance structures and latent variable models (Bandeen-roche et al., 1997; Jöreskog, 1970; Lee and Shi, 2001; McArdle and Kadlec, 2013; Sammel and Ryan, 1996; Yuan and Bentler, 1997).

SEM is considered as a second-generation statistical technique and enables researchers to incorporate unobservable variables measured indirectly by indicator variables (Hair Jr et al., 2013). It also facilitates to account for measurement error in the observed variables (Chin, 1998)

and can be considered “more as an idea than a technique”. In the current research, it is applied as a concept rather than a tool. The success of SEM requires clear specifications about the initial model. The model hypothesis must be clearly outlined, which forms the basis of the calculations and estimations. Therefore, the rationale behind so many SEM applications for data analysis was questioned. The SEM’s ability to draw path diagrams is not a strong rationale. It was widely accepted in behavioral science research for the following reasons (McArdle and Kadlec, 2013):

1) *SEM can examine a priori ideas in real data.* If some ideas come out, which are beyond Analysis of Variance (ANOVA) and the so-called General Linear Model (GLM) framework, and need to be validated, the SEM can provide such a method through SEM estimators, statistical indices, and overall goodness-of-fit indices.
2) *SEM can directly estimate scores of latent variables’ (LV).* Although LVs are not directly measured or measurable, one would like to model them. Thus, the inclusion of LV in a model enhances clarity. Also, it is apparent that the accurate distribution representation of the observed variables may require more complex measurement models than the typical normality assumptions.
3) *SEM can help to select the “true”, “correct”, or at least “adequate” model for a dataset.* An adequate model is based on invariant parameters that are not affected by difference in sampling or occasion. In linear regression, the model which explains the highest variance in the data is not always desirable. On the contrary, a model, which is capable of replicating over several simulations, is more desirable for data analysis. SEM can be desirable for finding such a model from a dataset.

The ability to estimate LV scores under a regression is crucial in geo-data processing, because it can combines the above three advantages. In mineral exploration, usually a series of geological factors (LVs) need to be extracted from the geo-data based on a simple initial model. These LVs obtained from the previous process do not account for the highest explained variance for the dataset, but are the most related to the object in which we are interested (the target variable). The subsequent sections will discuss how to construct such a model and how to estimate the model parameters and LV scores.

2.2 Measurement theory and structural theory

The SEM is considered as an extension of path models, which are diagrams used to visually display the being examined hypotheses and variable relationships (Hair et al., 2011; Hair et al., 2003). An example of a path model is shown in ***Fig** 2.1*.

The constitution of structural equation modeling with latent variables usually embodies two models: the measurement model and the structural model, which are expressed by the following THREE equations (Jöreskog and Sörbom, 1996):

- Measurement model associating the latent exogenous variables (Ξ) and measurement variables (X):

$$X = \Lambda\Xi + \Delta \qquad (2.1)$$

where $X = \begin{bmatrix} x_1 \\ \vdots \\ x_q \end{bmatrix}$, $\Lambda = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \vdots & \ddots & \vdots \\ \lambda_{q1} & \cdots & \lambda_{qm} \end{bmatrix}$, $\Xi = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_m \end{bmatrix}$, $\Delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_q \end{bmatrix}$.

- Measurement model associating the latent endogenous variables (H) and measurement variables Y:

$$Y = MH + E \tag{2.2}$$

where $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}$, $M = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1n} \\ \vdots & \ddots & \vdots \\ \mu_{p1} & \cdots & \mu_{pn} \end{bmatrix}$, $H = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \end{bmatrix}$, $E = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{bmatrix}$

Further, the general structural equation model can be expressed as follows:

$$H = BH + \Gamma\Xi + Z \tag{2.3}$$

where $B = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1n} \\ \vdots & \ldots & \vdots \\ \beta_{n1} & \cdots & \beta_{nn} \end{bmatrix}$, $\Gamma = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nm} \end{bmatrix}$, $Z = \begin{bmatrix} \zeta_1 \\ \vdots \\ \zeta_n \end{bmatrix}$

Herewith, X and Y represent the vector of the observed independent variables and the vector of the observed dependent variables, respectively. Ξ and H are the vectors of latent variables involved in the two measurement models, respectively, corresponding to the factors from the interdependent variables in X and the dependent variables in Y . The $q \times m$ matrix Λ and the $p \times n$ matrix M represent the relationships between the observed variables and the latent variables, typically referred to as the factor loadings. The coefficient matrices B and Γ in the

structure model associated with two latent vectors (Ξ and H) are to be determined. The symbols Δ, E and Z represent the modeling error vectors in ***Eq***. *(2.1)*, *(2.2)* and *(2.3)*, respectively. A confirmatory factor analysis (CFA) can be used to test whether the measurement variables could be represented by a set of factors (latent variables) as in the measurement models.

Measurement theory specifies how the latent variables are measured. In general, there are two different ways to measure the unobservable variables: reflective measurements or formative measurements. For example, as shown in ***Fig*** *2.1*, the constructed variables $\xi_1 - \xi_m$ are modeled using a formative measurement model. Note that the directional arrows are pointing from the indicator variables ($x_1 - x_p$) to the constructed ones (also given as the constructs), which indicates a causal (predictive) relationship in that direction. In the reflective measurement model, the directions of the arrows are going from the constructs to the indicator variable, which indicate the assumption that the constructs cause the measurements (covariation) of the indicator variables.

The meaning of structural model can be defined by different ways. In Hair Jr et al.(2013), structural model is defined as several linear models which shows how the latent variables are related to each other. The location and sequence of different sub model can be constructed based on theory or the researcher's experience and knowledge. The variables which are on the left side of and on the right side of the path model are independent variables and dependent variable, respectively. Just being similar as a linear regression, variables on the left are shown as sequentially preceding and predicting the variables on the right. However, being different from

a single linear regression model, variables may also serve as both the independent and dependent variable. When latent variables serve only as independent variables, they are called exogenous latent variables ($\xi_1 - \xi_m$). When latent variables serve only as dependent variables or both independent and dependent variables, ***Fig** 2.1* it has only one dependent variable η_1, they are called endogenous latent variables. Any latent variable that has only single-headed arrow going out of it is an exogenous latent variable. In contrast, an endogenous latent variable can have either single-headed arrows going both into and out of them (η_1).

2.3 SEM in geo-data processing

As mentioned in previous sections, SEM is a concept and any actual model should be discussed based on a specific problem. A basic SEM model is here proposed for application in processing and integration of geo-data for the following objectives:

1) How to extract geological factors from raw data under a restriction of regression.
2) How to construct a WofE based on SEM for prediction of the occurrence of mineral deposits from several patterns.

The model structure shown in ***Fig** 2.1* involves one level of measurement model and also one level of structural model. More levels of measurement model and structural model occur for general SEM modelling. The latent exogenous vector Ξ consists of m latent variables ($\xi_1, \xi_2, .., \xi_m$), (drawn as blue ellipses in ***Fig** 2.1*). The latent endogenous vector (H) includes only one latent variable η_1 measured by the observed variable y_1 (shown as a red ellipse in ***Fig***

2.1). The measurement variables are represented by the blue rectangles in ***Fig*** *2.1*.

The basic model includes a number of independent observed variables x_i and one dependent variable *y,* which are related through several latent variables. One application of this model is to extract the ore-control factors from geo-data. For instance, if x_i represents a number of

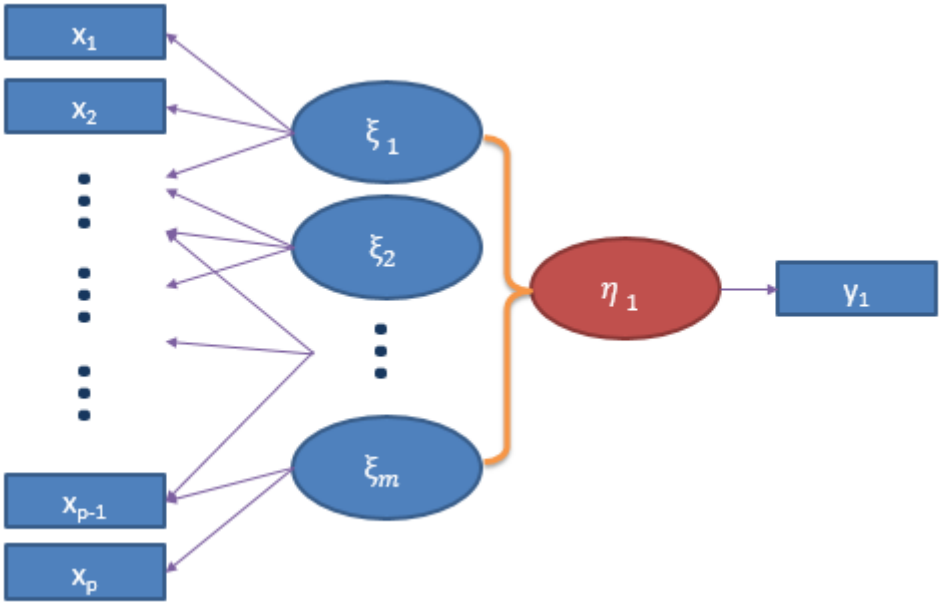


Fig *2.1 A flowchart showing a simple structural equation model consisting of one level of structure model and one level of measurement model. Rectangle symbols represent observed variables; blue eclipses for latent exogenous variables and red eclipse for latent endogenous variables.*

geochemical elements and y_1 represents one variable related with mineralization, the mineralization related factors in these elements can be estimated by the vector ξ of latent variables, and then mapped through their scores. Another application in geo-data processing is to classify the independent variables x_i (i =1, 2, …) under the restriction of dependent variable *y*, the membership of the measured variable can be represented by the structural model and the

measurement model and evaluated by the overall goodness-of-fit of the model.

2.4 CB-SEM and PLS-SEM

As mentioned in Chapter 1, two main approaches for parameter estimation in SEM are CB-SEM and PLS-SEM (Anderson and Gerbing, 1988; Hair et al., 2011; Hair et al., 2012; Hair et al., 2006; Hendry, 1976; James and Singh, 1978; McDonald, 1977; Mehta and Swamy, 1978; Ramsey, 1978; Sargan, 1978; Zellner, 1978). PLS-SEM is one of the commonly used algorithms originally developed by Wold (1966) on the basis of PLS path models and further developed by others (e.g. Hair et al., 2011; Hair et al., 2006). In PLS path models, the explained variances of the endogenous latent variables are maximized by estimating partial model relationships through iterative ordinary least squares (OLS) regression. PLS-SEM is mainly applied in exploratory analysis rather than confirmatory analysis (Hair Jr et al., 2013; Hair et al., 2011; Hair et al., 2006). The estimation procedure for PLS-SEM is an OLS regression-based method while the maximum likelihood (ML) estimation procedure is for CB-SEM. PLS-SEM uses available data to estimate the path relationships in the model by minimizing the errors (i.e. residual variance) of the endogenous constructs. In other words, PLS-SEM estimates the coefficients (i.e. path model relationships) by maximizing the R-square values of the target endogenous constructs. This specific feature achieves the objective of PLS-SEM: prediction. PLS-SEM is, therefore, the preferred method when the research objective is to develop a theory and explain variance (prediction of constructs). For this reason, PLS-SEM is regarded as a variance-based approach to SEM.

Partial least square (PLS) regression is a regression-based approach that explores the linear relationships between multiple independent variables and a single or multiple dependent variables. It differs from the regular regression and constructs composite factors from both of the multiple independent and the dependent variables by means of PCA. PLS regression is particularly useful in predicting a set of dependent variables from a large set of independent variables. It originated in the social sciences (Wold, 1966) but became popular first in chemometrics (Geladi and Kowalski, 1986) and in sensory evaluation (Martens and Naes, 1992). But PLS regression is also becoming a tool of choice in the social sciences as a multivariate technique for non-experimental and experimental data processing (Mcintosh et al., 1996). It was first presented as an algorithm akin to the power method but was rapidly interpreted in a statistical framework (Höskuldsson, 1988; Frank and Friedman, 1993; Helland, 1990).

Note that PLS-SEM is similar but not equivalent to PLS regression. PLS regression is a technique that generalizes and combines features from principal component analysis and multiple regressions. PLS-SEM, on the other hand, relies on a pre-specified network of relationships between the constructs as well as between constructs and their measurements. More details about PLS-SEM and PLS can be found in Mateos-Aparicio (2011).

2.5 Remarks

On the ground of the general overview of the SEM, a basic SEM structure is proposed for geo-data processing and analysis, which is a SEM with one level measurement model and one level

structural model. After such a simple model, the measurement model is described in detail as a factor analysis and the structural model is described as a regression model. Although the CB-SEM and PLS-SEM techniques are introduced, PLS-SEM is selected as the algorithm to estimate model parameters in ***Fig*** *2.1*.

Chapter 3 Dataset and software

3.1 Geological background of study area

The study area, located in Western Meguma Terrain of Nova Scotia, Canada (***Fig** 3.1*), covers about 25,000 km^2, and mainly consists of Cambro-Ordovician low-middle grade metamorphosed sedimentary rocks and a suite of aluminous Devonian granitoid intrusions (Ryan and Ramsay, 1997; Sangster, 1990). The metamorphosed sedimentary strata of the Meguma Group include two rock formations: the lower sand-dominated flysch Goldenville Formation and the upper shaly flysch Halifax Formation. Both of them were deformed during the Devonian granitoid intrusion emplacement resulting in northeast-southwest trending folds (Kontak et al., 1998).

The South Mountain Batholith (SMB), which is a complex of multiple intrusions, occupies nearly one-third of the whole study area. Abundant ***Sn***, ***W***, ***U*** and ***Au*** mineralization and mineral deposits have been found in this area. While the ***Sn***, ***W*** and ***U*** mineralization occurs mainly inside the SMB and in the contact zones between the complex intrusion and the metamorphic sedimentary rocks, the ***Au*** deposits occur mainly in the Meguma Group, especially around the Goldenville and Halifax Contact (GHC) zones (Chatterjee, 1983).

Studies of known ***Au*** deposits and their regional geological environment have shown that these deposits are turbidite-hosted ***Au*** deposits (Mawer, 1986; Ryan and Ramsay, 1997). The major

mineralization-related geological features described by previous researchers included GHC, northeast-southwest trending anticline axes and northeast-southwest trending shear zones (Kontak et al., 1990; Kontak and Kerrich, 1997; Ryan and Ramsay, 1997; Sangster, 1990). Litho-geochemical analyses have shown that ***As***, as a main path-finder element of ***Au***, has strong but complex relationships with ***Au*** mineralization. For example, ***Au*** and ***As*** are highly correlated in alteration zones related to ***Au*** mineralization controlled by fracture zones or faults and within the GHC, but not in all gold-bearing quartz veins (Crocket et al., 1986; Kerswill, 1988; Zentilli et al., 1985).

The first regional geochemical survey sampling of the center-lake bottom sediments in Meguma Terrain took place in 1977-1978 and about 4000 samples were collected. The samples were air dried, disaggregated in a ball mill, and sieved to obtain a 20 g portion of the 200 mesh fraction (Rogers et al., 1985). The collection and quality control methods were described by Garrett et al. (1980). The sampling density was about 1 per 5 km^2 (Rogers et al., 1987). In 1985, 2950 of the original samples were reanalyzed for ***Cu***, ***Pb***, ***Zn***, ***Ag***, ***Li***, ***Rb***, ***Nb***, ***Ti***, ***Sn***, ***Zr***, ***Th***, ***Sb***, ***As***, ***W*** and ***Au***. ***As*** and ***Au*** were detected by the instrumental neutron activity method with a detection limit of 1 ppm and 5ppb, respectively (Rogers et al., 1987). The study area includes 1948 of the 2950 samples and 1312 of these samples had ***Au*** values below the detection limit of 5 ppb; 48 samples had ***As*** values below 1 ppm detection limit. The geochemical data from lake sediment samples have been intensively studied by geochemists who have worked in the area, not only due to the fact that high values of ***Au***, ***Sn*** and ***W*** partially correspond to known mineral deposits and occurrences, but also due to the possibility that specific association of elements may reflect

the main rock units. For example, while ***F***, ***Li***, ***Nb***, ***Rb***, and ***Sn*** may reflect existence of granitoid rocks, ***Sb***, ***As***, ***Au*** and ***W*** indicate the occurrence of metamorphosed sedimentary rocks (Bonham-Carter et al., 1988; Dunn et al., 1991; Rogers et al., 1990; Rogers et al., 1987). Distances between the anomalies in lake sediments and their sources were studied in the vicinity of the East Kemptville ***Sn*** deposit and surrounding lake basins (Rogers and Garrett, 1987), which indicated that these have elemental associations similar to those in the bedrock lithogeochemistry. Distances between anomalies in the lake sediments and their sources normally range from several hundreds of meters to several kilometers. The glacial till translation distance in Meguma Terrain ranges from 100 to 1500 m (Graves and Finck, 1988). For regional geochemical research with 1km resolution, the influence of glaciation may not be significant. The bed geologic units are shown in ***Fig*** *3.1* and *3.2*.

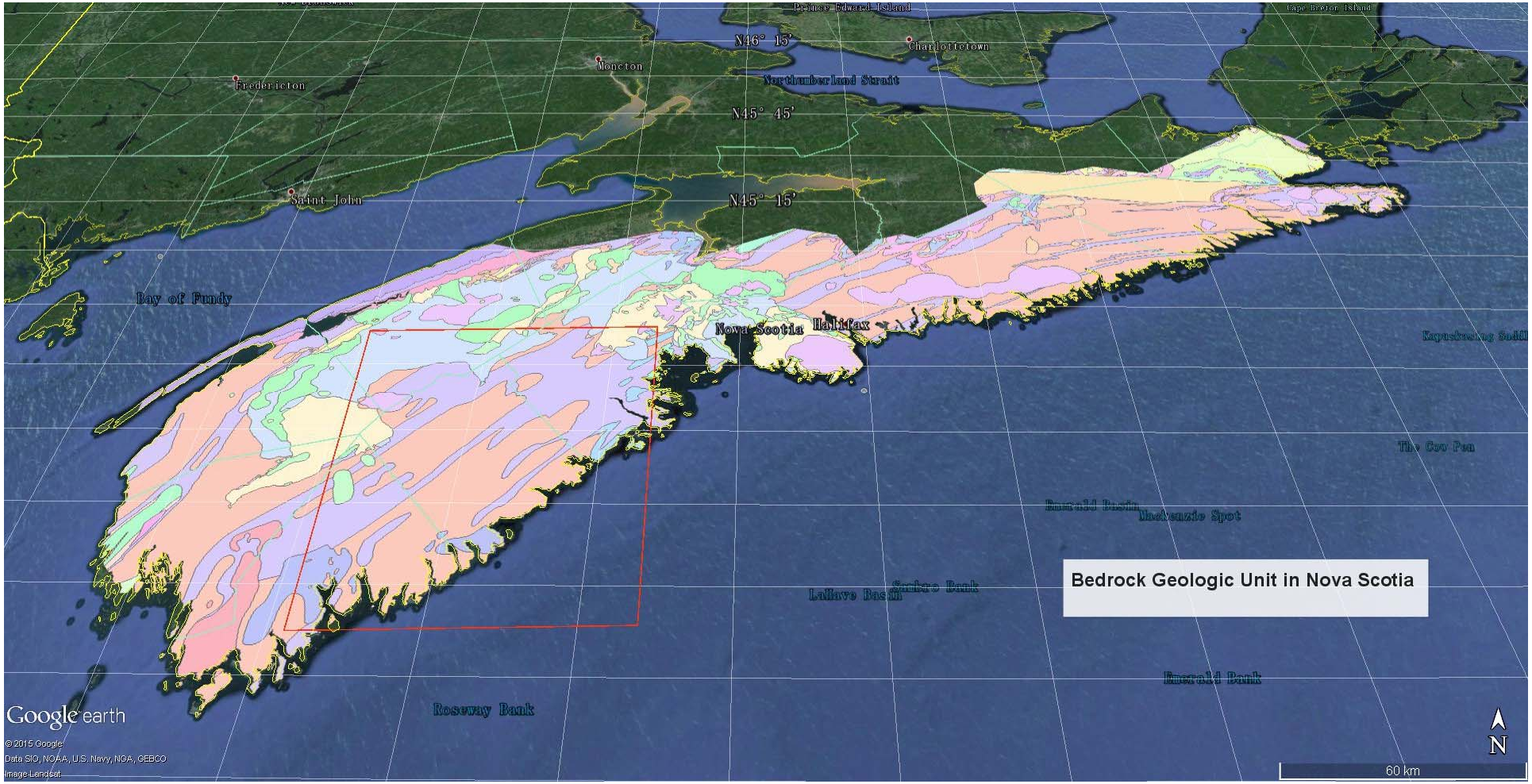


***Fig** 3.1 The study area in Google earth, the colorful area is the DP ME 132, Version 2, 2006, Regional Lake Sediment Geochemical Survey by the Nova Scotia Department of Natural Resources over Southern Nova Scotia, 1977-1978. The red frame is the study area.*

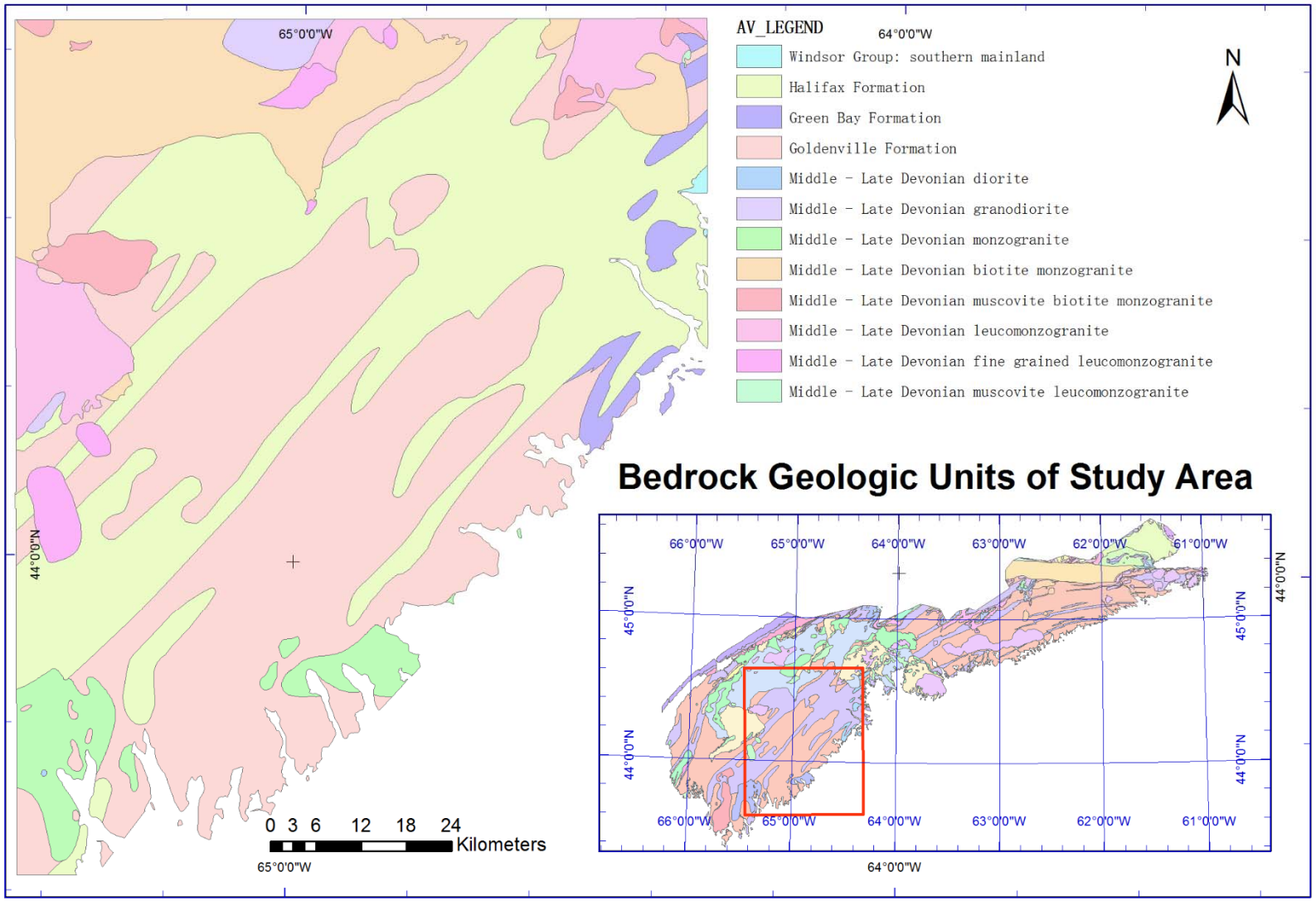


***Fig** 3.2 Bed geologic units of study area.*

***Table** 3.1 Correlation coefficient between 16 elements in four formations: Goldenville(G), Halifax(H), Granite and Granodiorite (GG), Gneiss and Schist (GS);*

Pb	All	.26														
	G	.22														
	H	.23														
	GG	-.05														
	GS	-.17														
Zn	All	.67	.39													
	G	.54	.41													
	H	.77	.32													
	GG	.31	.21													
	GS	.81	-.16													
Ag	All	.01	.02	.01												
	G	.14	.09	.09												
	H	.05	.13	.03												
	GG	.00	.00	.02												
	GS	-.23	.26	-.36												
F	All	.35	.26	.42	-.01											
	G	.41	.29	.43	-.07											
	H	.30	.20	.39	.07											
	GG	.13	.17	.28	-.03											
	GS	-.27	-.14	-.07	-.46											
Li	All	.41	.39	.58	-.01	.65										
	G	.35	.37	.51	-.07	.79										
	Halifax	.27	.31	.54	.06	.46										
	GG	.18	.37	.50	.00	.53										
	GS	-.38	-.21	-.49	-.15	.33										
Nb	All	.22	.21	.21	.00	.37	.53									
	G	.25	.22	.15	-.04	.60	.66									
	H	.10	.16	.17	.00	.23	.38									
	GG	-.04	.07	-.04	.02	.11	.29									
	GS	.42	.16	.38	-.38	-.27	-.27									
Rb	All	.11	.32	.25	-.03	.53	.72	.43								
	G	.12	.28	.20	-.10	.75	.75	.63								
	H	-.06	.20	.16	-.02	.32	.68	.33								
	GG	-.03	.41	.21	-.02	.43	.77	.25								
	GS	-.15	.11	-.45	.06	-.04	.86	-.10								
Sn	All	.01	.17	.14	.00	.18	.22	.05	.34							
	G	-.02	.19	.17	-.01	.22	.20	.15	.33							
	H	.01	.18	.17	.11	.20	.26	.01	.33							
	GG	-.03	.14	.04	-.03	.12	.25	-.08	.35							
	GS	.57	.20	.32	-.17	-.18	.21	.66	.51							
Zr	All	.02	.23	.12	-.02	.30	.47	.39	.65	.28						
	G	.05	.24	.25	-.04	.57	.63	.57	.72	.38						
	H	-.17	.13	-.05	-.05	.11	.36	.30	.66	.23						
	GG	.05	.32	.20	-.01	.34	.63	.26	.68	.30						
	GS	-.15	.11	-.50	.27	-.21	.76	-.10	.96	.49						
Ti	All	.27	.32	.38	-.02	.49	.71	.56	.60	.22	.67					
	G	.20	.37	.39	-.05	.77	.84	.70	.82	.33	.77					
	H	.05	.21	.18	-.03	.25	.52	.44	.54	.19	.65					
	GG	.24	.21	.39	.00	.39	.66	.32	.41	.13	.60					
	GS	.69	.21	.31	.11	-.22	-.15	.00	.21	.49	.27					
Au	All	.07	.05	.05	.01	.06	.12	.03	.09	.07	.04	.06				
	G	.05	-.01	.06	-.05	.11	.13	.00	.14	.15	.06	.03				
	H	-.01	.10	.00	.04	-.09	-.05	.06	-.06	.00	.01	.04				
	GG	.01	.05	-.10	.06	-.01	.00	-.02	.05	-.04	.01	.02				
	GS	.57	.04	.53	-.27	-.56	-.65	.79	-.45	.29	-.40	.11				
Sb	All	.09	.12	.10	.01	.04	.06	.12	.01	.00	-.03	.05	.10			
	G	.23	.19	.24	.10	.22	.13	.08	.17	.15	.06	.16	.23			
	H	-.03	.05	.01	.03	-.02	-.03	.15	-.09	-.06	-.13	-.08	.12			
	GG	.03	.23	.15	.03	.07	.17	-.01	.08	.06	.13	.22	-.10			
	GS	-.09	.68	.18	.21	-.16	-.25	-.08	-.11	-.08	-.18	-.05	.02			
As	All	.35	.20	.47	.02	.23	.18	.02	.12	.08	.01	.14	.14	.08		
	G	.08	.11	.27	.08	.30	.07	.00	.20	.13	.06	.16	.17	.25		
	H	.60	.23	.63	.14	.16	.18	-.06	-.02	.06	-.11	-.01	.02	-.02		
	GG	.22	.15	.52	.00	.24	.35	-.15	.21	.12	.12	.19	-.05	-.01		
	GS	.93	-.02	.83	-.35	-.12	-.44	.32	-.25	.44	-.28	.73	.52	.04		
Th	All	.51	.35	.65	.01	.55	.73	.45	.57	.19	.53	.62	.07	.11	.30	
	G	.48	.33	.71	-.02	.71	.79	.51	.59	.23	.62	.69	.05	.19	.16	
	H	.49	.27	.60	.07	.41	.58	.30	.45	.16	.34	.41	.03	.07	.40	
	GG	.13	.24	.33	.04	.39	.67	.38	.61	.19	.77	.64	-.02	-.04	.27	
	GS	.07	-.04	-.39	.05	.03	.74	-.10	.90	.54	.88	.43	-.39	-.35	-.06	
W	All	.03	-.01	.02	-.01	.20	.04	.04	.19	.11	.02	.11	.04	.10	.45	.03
	G	-.01	-.05	-.05	-.03	.33	-.01	.04	.32	.19	.02	.14	.04	.30	.57	.00
	H	.32	.11	.37	-.05	.14	.18	-.04	.09	.08	.04	.07	-.03	-.05	.33	.13
	GG	.07	.13	.29	.00	.20	.24	.07	.18	.07	.10	.23	-.09	.19	.34	.19
	GS	.43	-.49	.58	-.37	-.04	.26	-.10	.17	.18	.08	.17	.01	-.03	.44	.06
		Cu	Pb	Zn	Ag	F	Li	Nb	Rb	Sn	Zr	Ti	Au	Sb	As	Th

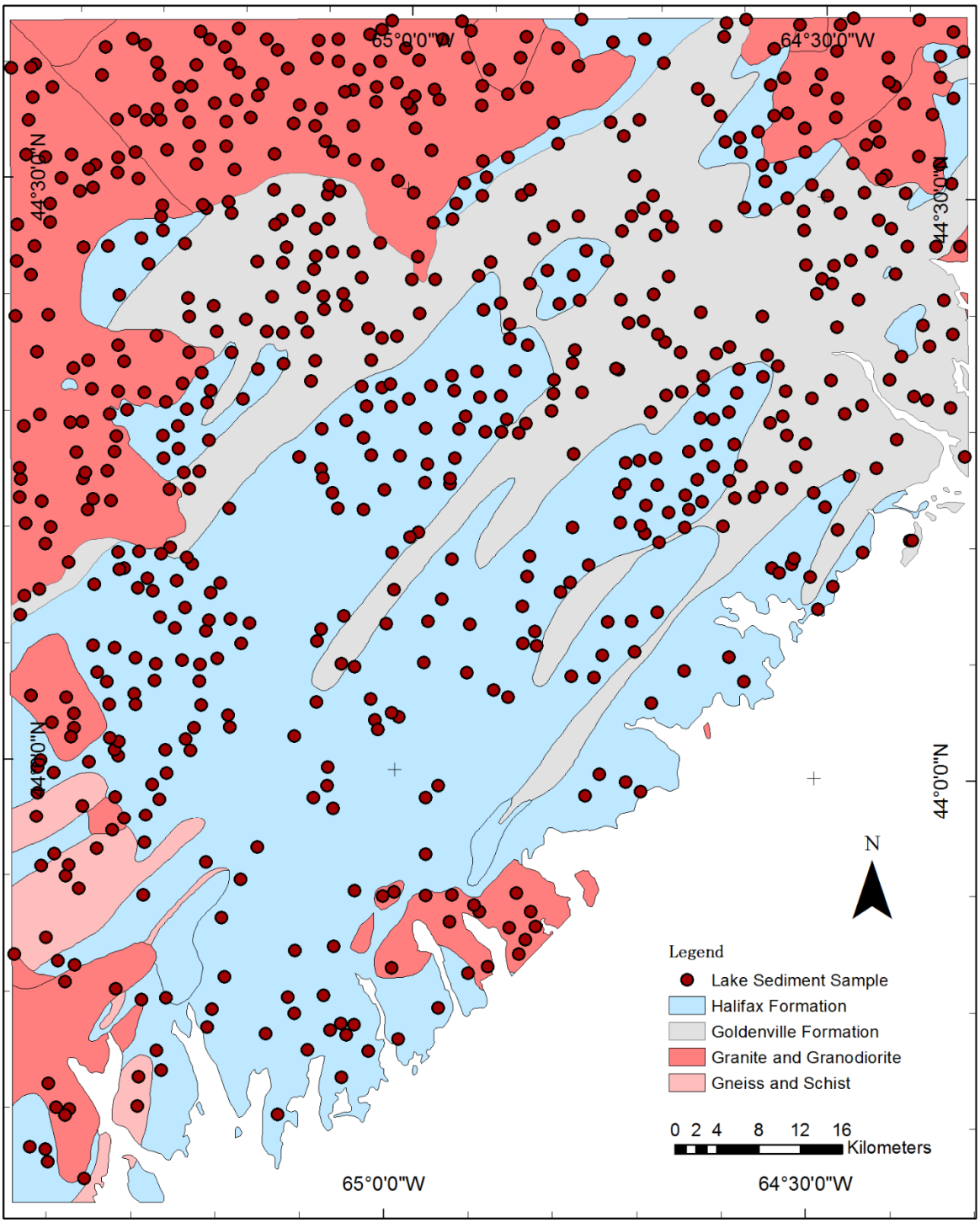


***Fig** 3.3 Lake sediment sample location in study area*

3.2 Dataset and transformation

A basic understanding of their statistics is essential for investigating the application of the geochemical dataset used in this research. The samples are grouped into five categories according to the areal geologies: (i) all samples (671), (ii) samples in the Goldenville formation (166), (iii) samples in the Halifax formation (282), (iv) samples in Granite and Granodiorite (GG) rock-type (214), and (v) samples in Gneiss and Schist (GS) rock-type (9). The correlation coefficients between the 16 elements selected for the research in different bedrock geological units in study area are shown in ***Table*** *3.1*. Since the goal of this research is to find the factors controlling ***Au*** mineralization, the focus tends to be on those elements that are strongly related to ***Au***. From the correlation coefficients of ***Au*** with other 15 elements (***Fig*** *3.5*), it is clear that ***Au*** has the highest correlation (0.14) with ***As***. In the other categories, the ***Au-As*** correlation coefficients are: 0.17 (Goldenville), 0.02 (Halifax), -0.05 (GG), and 0.52 (GS). The distribution map of the ***As*** (***Fig*** *3.8A*) and ***Au*** (***Fig*** *3.8B*) and other statistical information about the current dataset including the maximum, minimum, mean and stand deviation of samples are given in ***Table*** *3.2*. The log-transformed ***As*** and ***Au*** are further mapped in ***Figs*** *3.8C* and *3.8D*, respectively. It can be seen in ***Fig*** *3.8B* that gold mineral occurrences do not occur in the areas with high ***Au*** concentration values which might be due to: 1) the low accuracy of the data may not show strong correlation, and 2) the occurrence of mineral deposits may not correspond to high ***Au*** concentration values in lake sediments in some locations in the current study area.

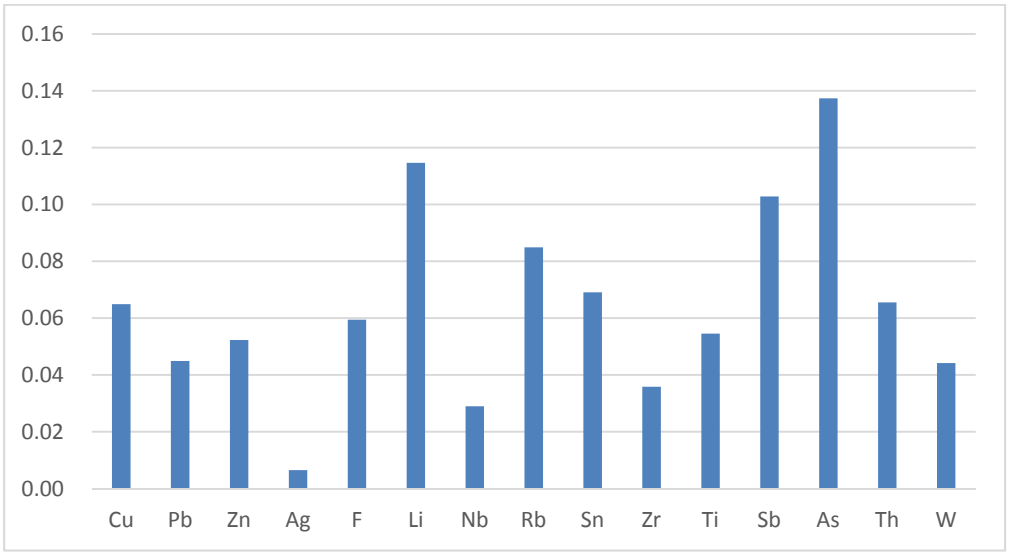


***Fig** 3.4 Correlation coefficients between Au and other geo-chemical elements, As has the highest correlation with Au.*

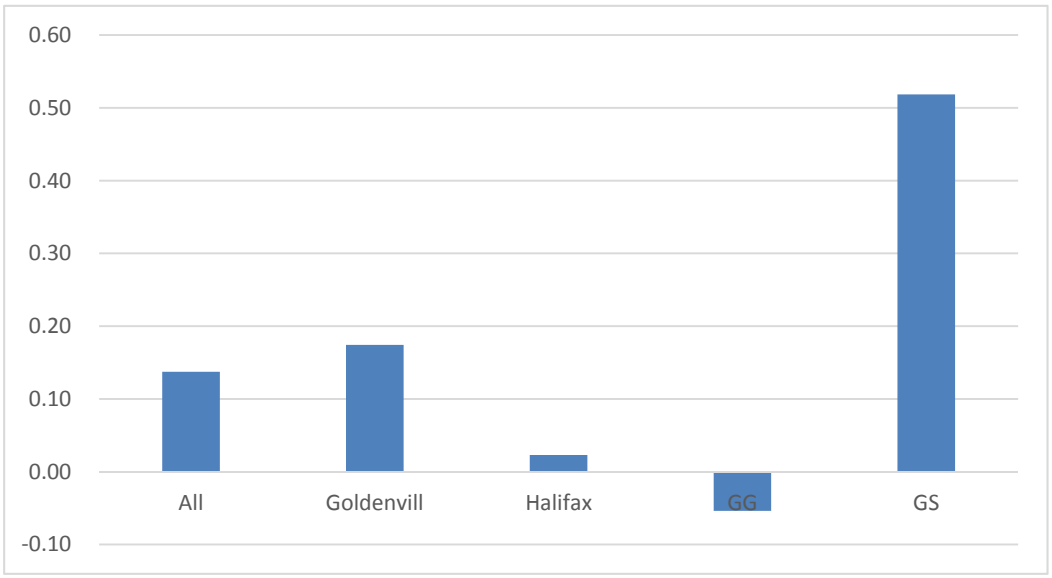


***Fig** 3.5 Correlation coefficients between As and Au in different rock units: the count of samples in each formation are: Goldenville – 166, Halifax – 282, Granite and Granodiorite (GG) – 214, Gneiss and Schist (GS) – 9;*

***Table** 3.2 Statistical analysis of geochemical data*

	Unit	Minimum	Maximum	Mean	Std-Dev
Ag	ppm	0.1	22.6	0.23	0.92
As	ppm	0.5	372	10.35	25.88
Au	ppb	0	130	3.76	6.02
Cu	ppm	1	75	10.36	7.11
F	ppm	20	990	107.05	78.44
Li	ppm	1	63	8.28	8.04
Nb	ppm	0.5	16	1.83	2.25
Pb	ppm	1	81	12.77	8.42
Rb	ppm	2	263	39.05	33.01
Sb	ppm	0.1	5.3	0.43	0.82
Sn	ppm	1	10	2.18	1.85
Th	ppm	0.3	15.3	3.07	1.89
Ti	ppm	0	1.3	0.17	0.12
W	ppm	0.3	434	1.55	16.77
Zn	ppm	6	296	48.24	38.85
Zr	ppm	7	406	84.44	62.68

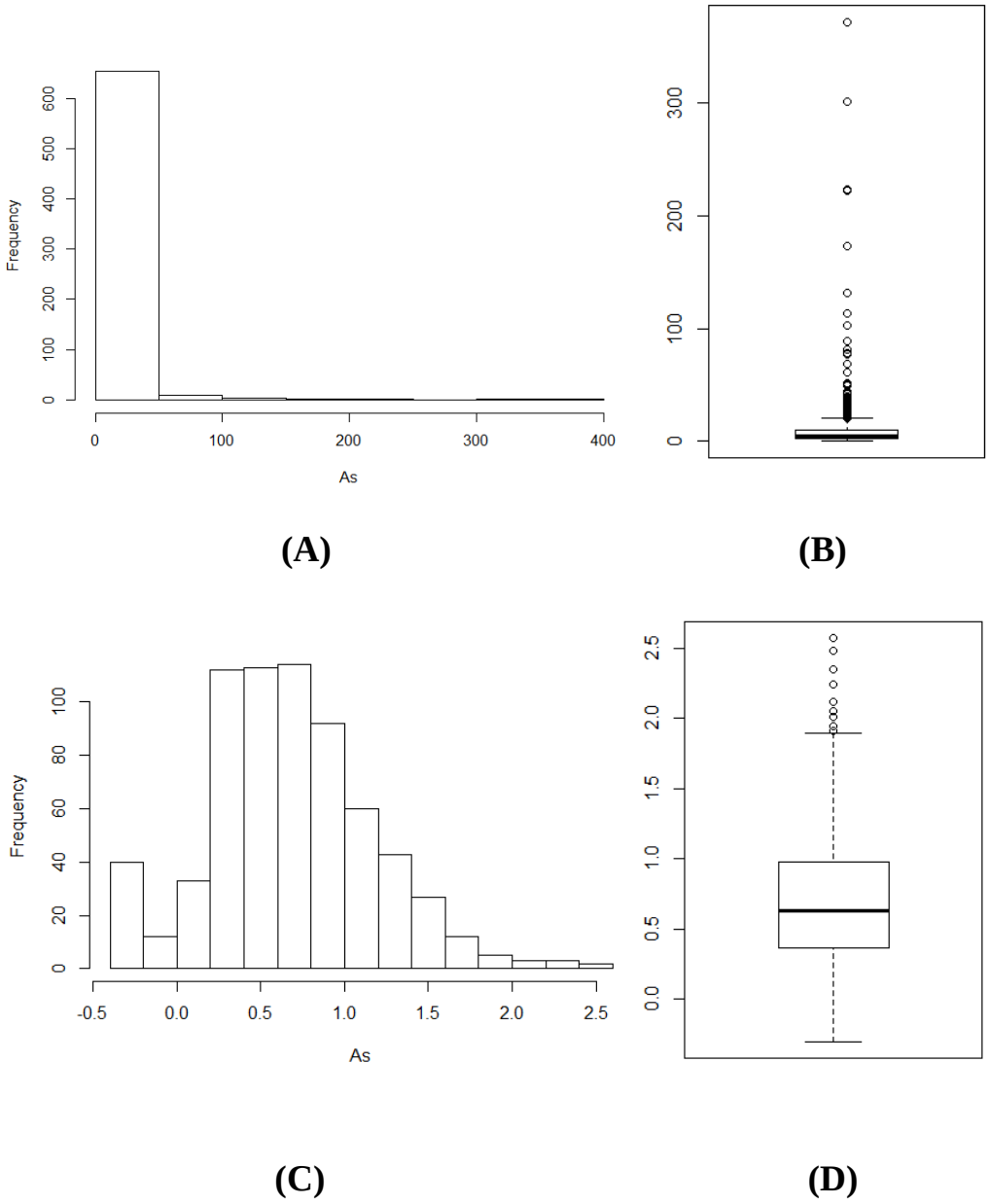
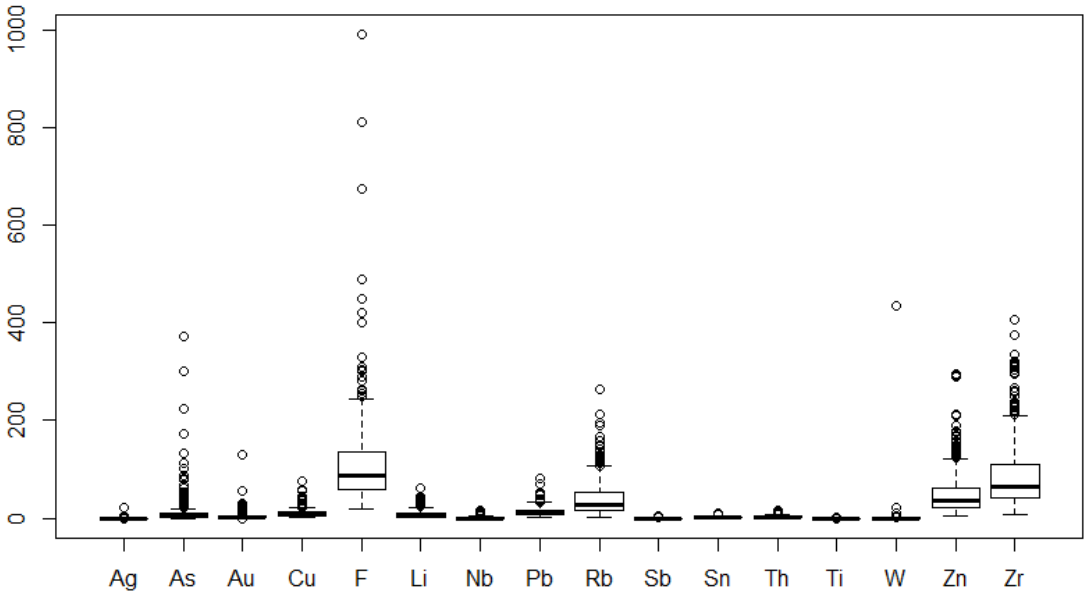
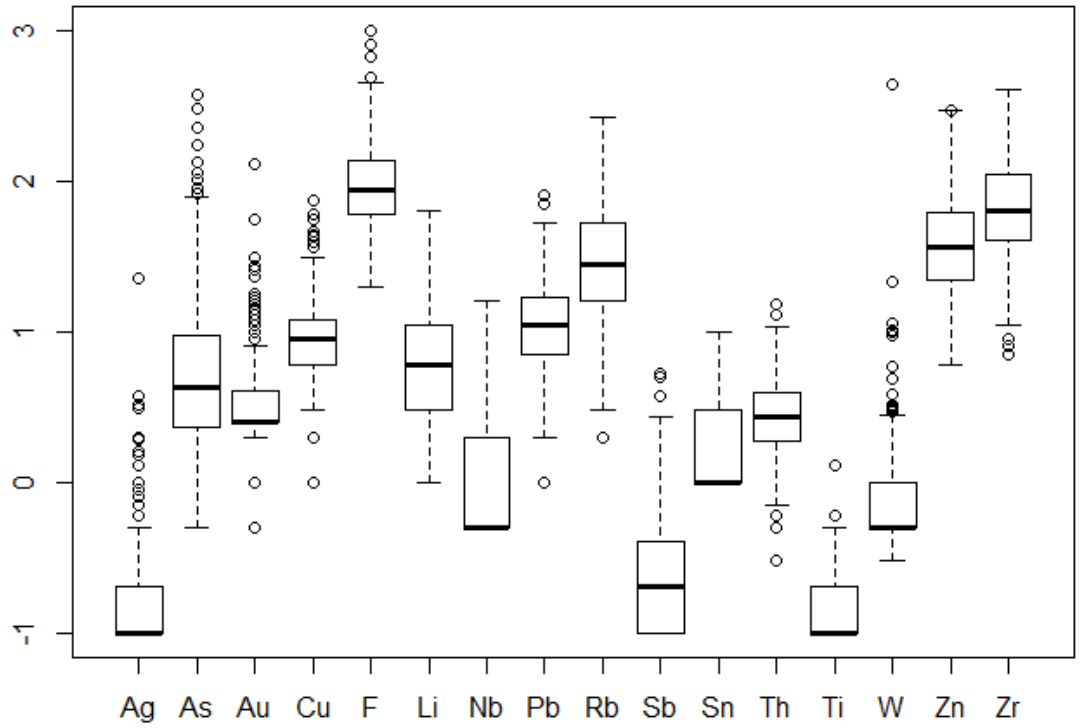


***Fig** 3.6 Histograms and boxplots for the raw As data (A: Histograms, B: boxplots), log-transformed As data (C: Histograms, D: boxplots)*



(A)



(B)

***Fig** 3.7 Histogram of 15 ore elements: (A) raw data, (B) log-transformed data.*

From the histograms and boxplots of the original concentration data and their log-transformed values for ***As*** (***Figs*** *3.6*) and all the 16 elements (***Figs*** *3.7*), it can be observed that the log-transformed data exhibit more symmetrical quantile distributions, while the elements show right-skewed distributions for the original data. Thereby, the log-transformed data can reduce the effects of outliers to certain extent. Under the consideration of the potential effect of compositional data and skewedness of distribution of geochemical data on multivariate statistical analysis in Euclidean space (Garrett et al., 1980; Reimann and Filzmoser, 2000; Vistelius, 1960), several transformations can be applied to the data prior to their use in multivariate analysis, e.g., (not limited to) additive log-ratio transformation (ALR), centered log-ratio transformation (CLR) and isometric log-ratio transformation (ILR) (Aitchison, 1986; Egozcue et al., 2003). For simplification of comparisons of the new SEM approach with other traditional statistical methods, the calculations and analyses are based on a log-transformed dataset in this research of Chapter 4, 6 and 7. The results based on centered log-ratio transformed geochemical data are attached in Chapter 6 and 7. The effect of compositional data on SEM would be studied in the future work.

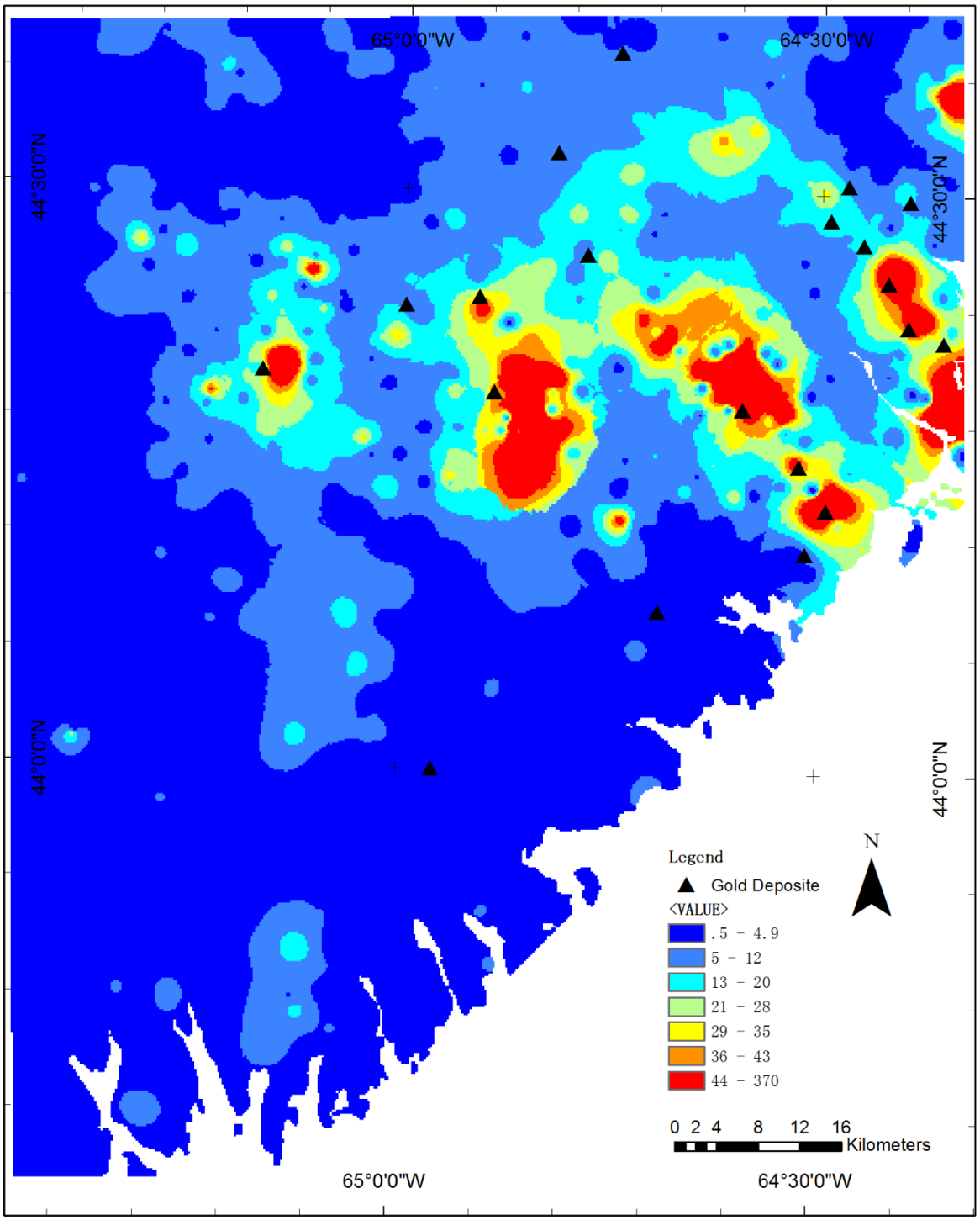


***Fig** 3.8A Spatial distribution of As in study area, unit: ppm, which interpreted from lake sediment samples through IDW method, the map is classified by interval of 1 standard deviation.*

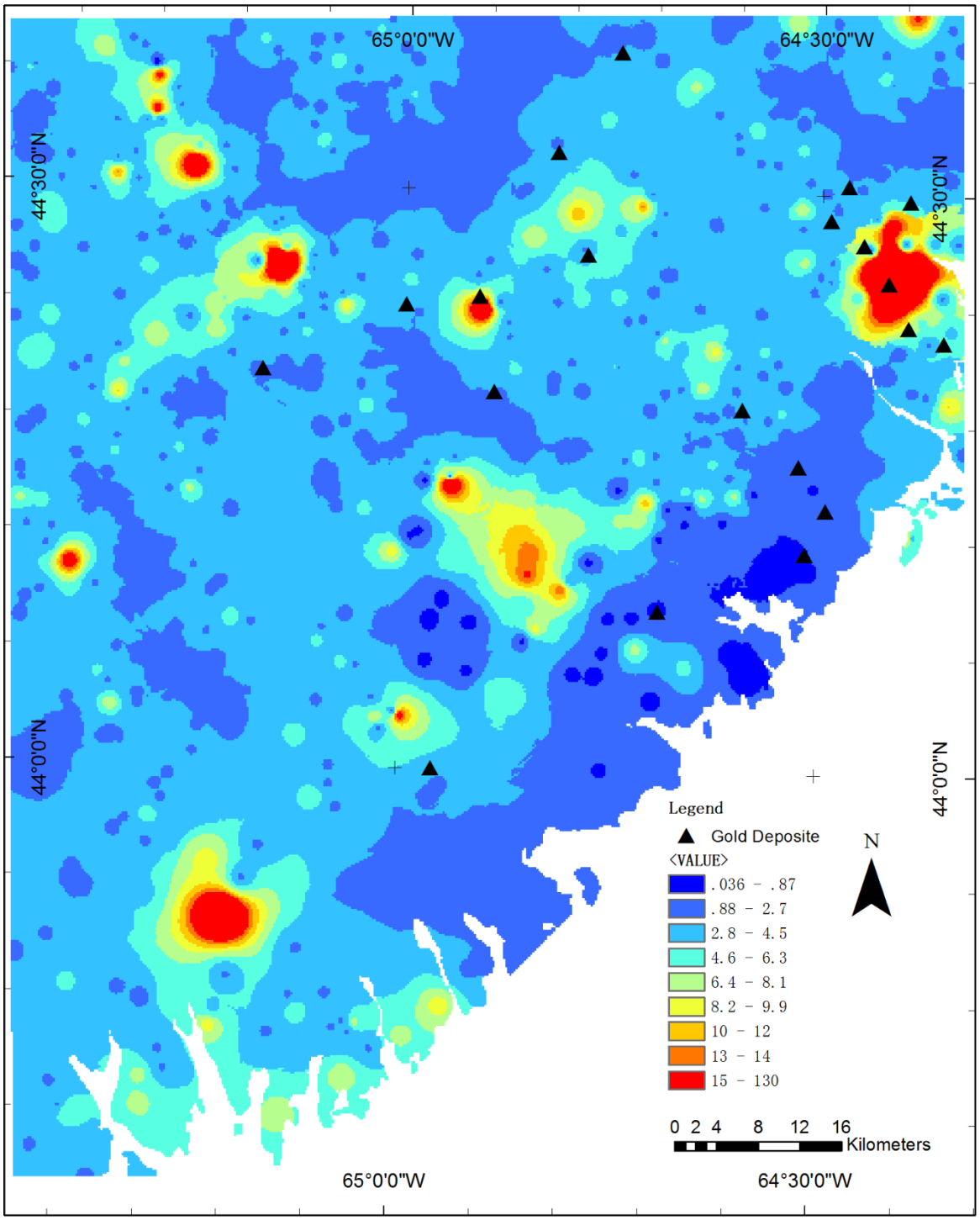


***Fig** 3.8B Spatial distribution of Au in study area, unit: ppm, which interpreted from lake sediment samples through IDW method, the map is classified by interval of 1 standard deviation.*

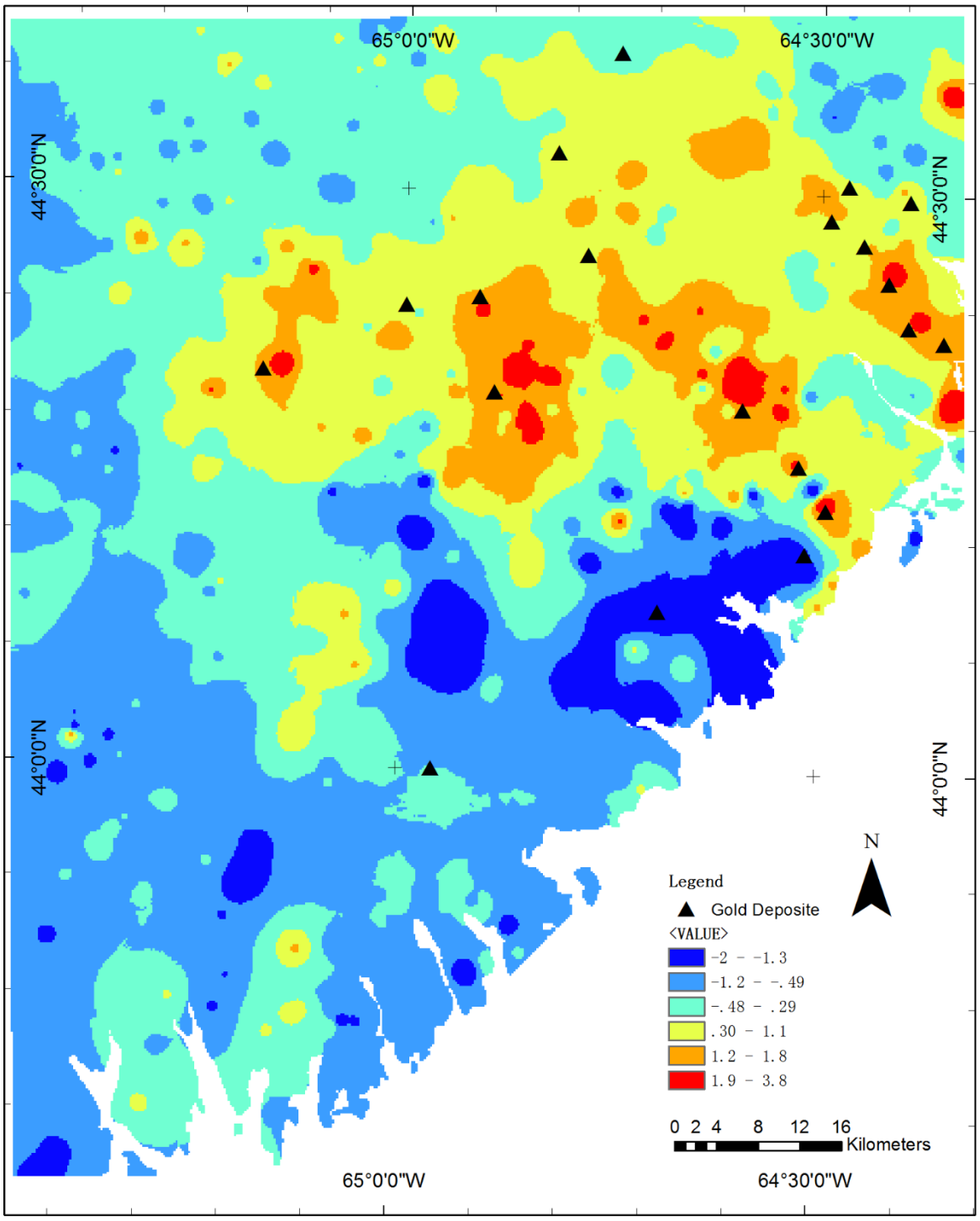


***Fig** 3.8C Spatial distribution of $Log_{10}(As)$ in study area, which interpreted from lake sediment samples through IDW method, the map is classified by interval of 1 standard deviation.*

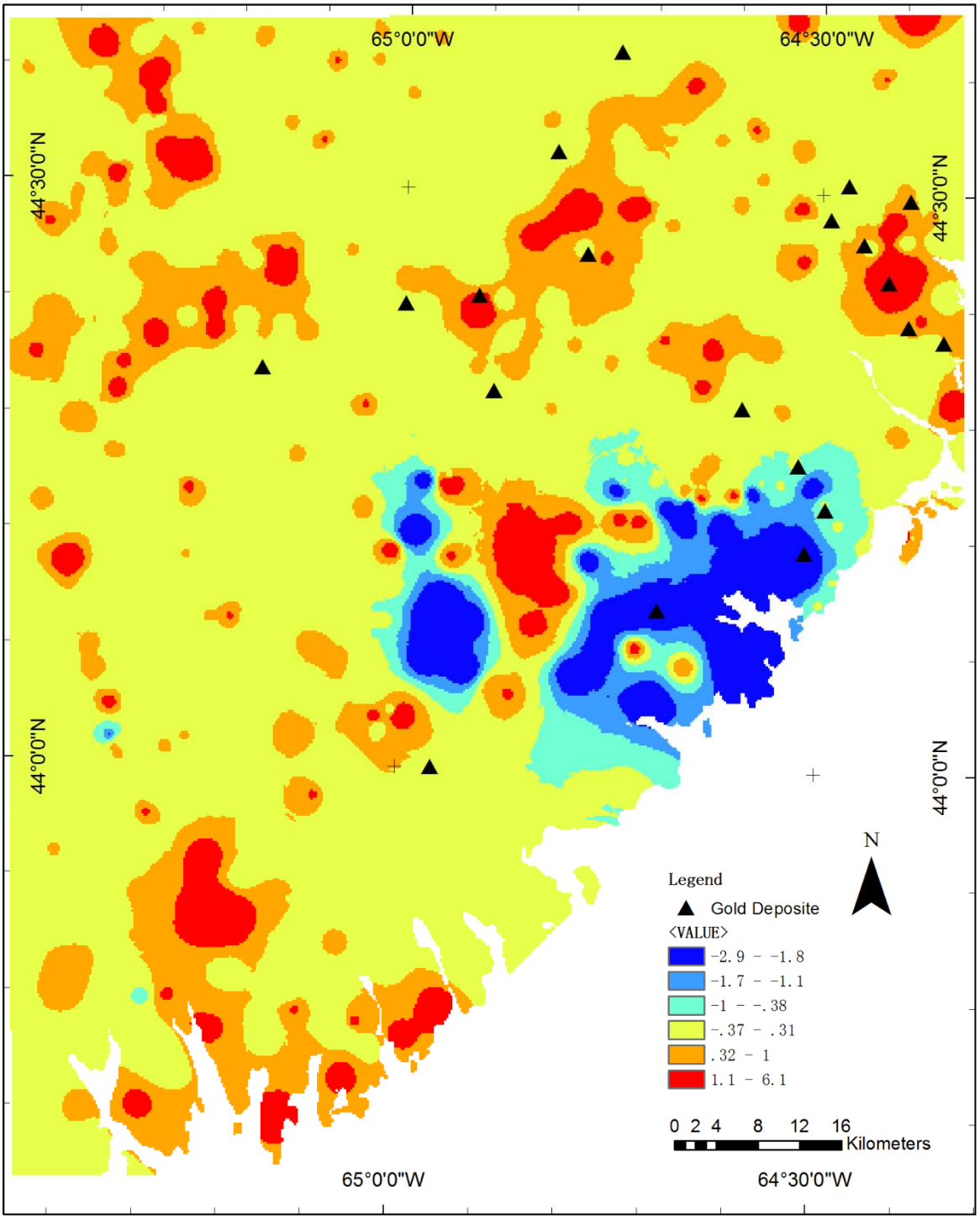


***Fig** 3.8D Spatial distribution of $Log_{10}(Au)$ in study area, which interpreted from lake sediment samples through IDW method, the map is classified by interval of 1 standard deviation.*

3.3 GIS software and development environment

One of the goals of this research is to design algorithms for conducting experiments, which includes statistical models and spatial analysis in GIS. The statistical models will be achieved using R (Ihaka and Gentleman, 1996; Team, 2012) and the required spatial analysis will be designed based on MapWinGIS (Ames, 2007; Ames et al., 2007).

R, a powerful language for statistical computation, is the product of a public domain (GNU) project, which is considered as a different implementation of the commercially available S language developed at the Bell Laboratories (now Lucent Technologies). R is a freeware and has plenty of resources available for research, which makes it very popular in the academic and research communities. Although it is similar to some programming packages such as MATLAB, R is more user-friendly than other programming languages such as C++ and FORTRAN. In addition to the basic functions, several specific libraries for data analysis are made available in R, e.g., the random sampling library "*sampling*" (Tillé and Matei, 2009), statistical library "*stats*" (Sinnwell et al., 2007), and database management library "*rredis*" (Lewis et al., 2014).

MapWinGIS is an open source GIS software with its application programming interface (API) distributed under the Mozilla Public License (MPL), built upon the Microsoft .NET Framework. It has been maintained by an active group of international developers who regularly release updates and bug fixes through the MapWindow.org website. It has been adopted by the United

States Environmental Protection Agency as the core GIS platform for its BASINS watershed analysis and modeling software which is used by environmental professionals at all levels of U.S. government and internationally in 2005. In general, MapWinGIS (MapWindow) is a mapping tool and a GIS modeling system in a redistributable open source form for its simplicity of use and for running on the most popular Microsoft Windows (Ames et al., 2007).

For this research, *sampling*, *stats*, and *rredis* will be used for the Monte Carlo simulation, solving model parameters through an optimum method "*optim*", and the database management, respectively.

Chapter 4 Identification of geochemical factors in regression to mineralization endogenous variables using SEM

4.1 Introduction

This chapter will propose a new SEM model for geochemical factors extraction, which is considered as a factor analysis for a geochemical dataset under the restriction of a multiple regression to a response element.

SEM is a class of multivariate statistical models that allow complex modeling of relational structures between independent and dependent variables. It combines factor analysis (FA) and multiple linear regressions (MLR) (Ullman and Bentler, 2003), calculates the factor loadings on latent variables or factors and the regression coefficients of latent variables with respect to dependent variables using a group of equations. The FA in the SEM is referred as to the observed model whilst the path analysis or MLR is referred as to the structural model. Therefore, SEM is different from the ordinary PCA or the exploratory FA (EFA) that are commonly used in geochemical data processing. The latter determine the orthogonal components with ranking of variances. But these components are calculated on the basis of the interrelationships of variables involved and they are usually not associated only with a particular objective variable of interest. The former determines the loadings of the pre-assigned factors according to their

association with external dependent variables. PCA provides a type of solution for latent variables based on orthogonal transformation of independent variables, but not always related to specific purposes in applications, while SEM in current research attempts to provide latent variables for different applications through setting response variables for specific applications. SEM is also different from MLR, since it involves the latent variables in structural model as independent variables rather than the original explanatory variables are used in MLR. CB-SEM and PLS-SEM are two main algorithmic types in SEM applications, which estimate the parameters through the covariance matrix and PLS path models, respectively. More discussions about PLS-SEM and CB-SEM are referred Wold (1982, 1985) and Lohmöller (1989).

To author's knowledge SEM has never been applied in geochemical data processing for mineral exploration previously. One of the main reasons for this situation might be due to the fundamental drawbacks of the existing SEM models, which are the requirement for a predetermined structural model and the incapability to generate and refine structural model.

A new SEM method is here proposed based on PLS-SEM, which combines the principles of cluster analysis and regression analysis. Thus, the new mathematical model can not only generate factors to form a structure model, but also ensure the optimum relationship to the objective dependent variables. Besides the introduction of the new SEM mathematical model, its applicability in geochemical data processing is also validated by a case study in terms of the factor identification for gold mineralization in Southern Nova Scotia, Canada. The case study uses the concentration values of 16 elements from 671 lake sediment samples collected in the

study area. For comparison purpose, all three methods: MLR, EFA and SEM were applied to analyze the same dataset. In the implementation of MLR and the new SEM model, the element ***As*** was utilized as the dependent and objective variable and other 15 elements were used as independent variables. The loadings and regression coefficients of the latent variables with respect to ***As*** were analyzed and compared.

4.2 The methods of parameter estimation

4.2.1 The PLS‐SEM Algorithm

The structural or the "inner" model in the PLS‐SEM context describes the relationships (paths) among the latent constructs. Since the PLS‐SEM allows only recursive relationships in the structural model, the structural paths between latent constructs are always unidirectional. Such paths could either be exogenous or endogenous. While the exogenous constructs are used to describe latent constructs without having any structural path relationships pointing at them (analogous to factor analysis), the endogenous constructs describe latent target constructs that are explained by other constructs via structural modeling relationships (analogous to multiple regression analysis).

The measurement or "outer" model, includes unidirectional predictive relationships between each latent construct and its associated observed indicators. Since multiple relations are not allowed in PLS-SEM, indicator variables are associated with only a single latent construct.

PLS-SEM can handle both of the formative and reflective measurement models. The associated coefficients for these paths are called "outer loadings" in PLS-SEM.

***Fig** 4.1* shows an example of a simple SEM model with one endogenous (dependent) latent construct η_1 and two exogenous (independent) latent constructs ξ_1 and ξ_2 (oval shapes). In the exogenous construct, each latent construct is measured with two formative indicator variables shown by rectangles (x_1 to x_4), with arrows pointing toward the constructs. In contrast, in the endogenous construct, the latent variable η_1 is measured with three reflective indicator variables (y_1 to y_3) with arrows pointing away from the construct. Most theoretical constructs (especially formative ones) will be measured by six or seven indicator variables, but our example includes fewer indicators to facilitate understanding the concept.

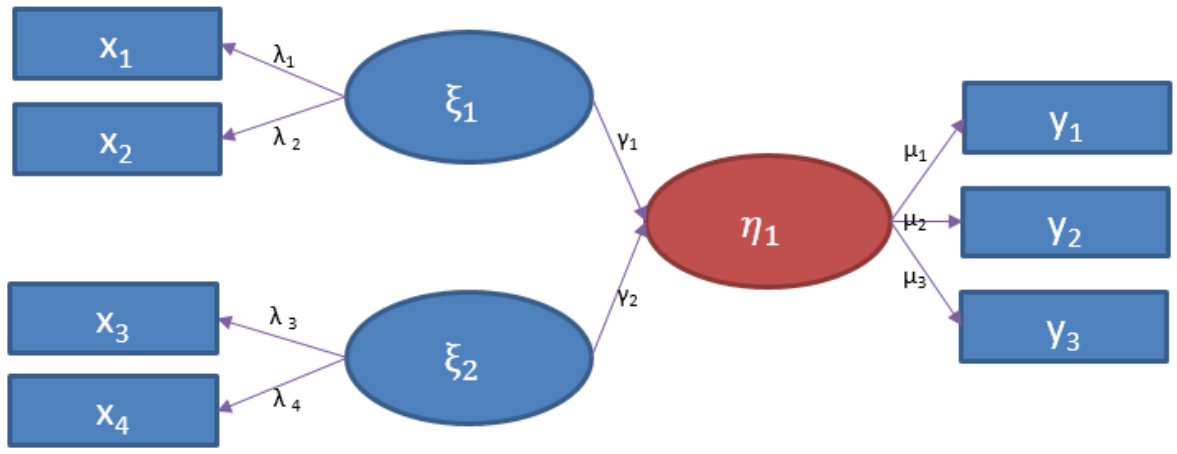


***Fig** 4.1 An example of PLS-SEM.*

The basic PLS-SEM algorithm follows a two-stage approach (Lohmöller, 1989). While the

scores of the latent constructs are estimated (***Table** 4.1*) in the first stage, the second stage estimates the final outer weights and loadings as well as the path coefficients of the structural model. The path modeling procedure is referred to as "partial" because the iterative PLS‑SEM algorithm estimates the coefficients for the partial ordinary least squares regression models in the measurement and structural models. More specifically, when a formative measurement model is assumed, a multiple regression model is estimated with the latent construct as the dependent variable and the assigned indicators as the independent variables (computation of outer weights). In contrast, when a reflective measurement model is assumed, the regression model includes single regressions with each indicator individually being the dependent variable, whereas a latent construct is always an independent variable (computation of outer loadings). When the structural model relationships are calculated, each endogenous latent construct represents the dependent variable with its latent construct antecedents as independent variables in a partial regression model. All partial regression models are estimated by the iterative procedures of the PLS‑SEM algorithm.

The first stage involves four steps. In *Step 1*, outer proxies of the latent construct scores are computed as the linear combinations of the values of all (standardized) indicators associated with a particular latent construct. For example, values of x_1 and x_2 are used to compute the proxy score for the latent construct ξ_1. Later iterations use the estimated coefficients of the paths (e.g., λ_1 and λ_2 for ξ_1, whereby λ represents the outer weight or loading coefficient) between the latent constructs and the indicator variables from *Step 4* of stage one. For the initial iteration, any combination of indicators can serve as a proxy for the latent construct. PLS‑SEM software

programs, such as SmartPLS (Ringle et al., 2005), use a uniform value of 1 as an initial value for each of the outer weights (λ_1 to λ_4, μ_1 to μ_3).

In *Step 2*, the PLS‑SEM computes the proxies for the structural model relationships (γ_1 and γ_2). Several different weighting schemes are available to estimate these proxies. This method develops latent construct scores that maximize the final R^2 value of the endogenous latent constructs (Lohmöller 1989). In *Step 3*, the inner proxies of the latent construct scores (ξ_1, ξ_2, and η_1) are calculated as the linear combinations of their respective adjacent latent construct outer proxies (from *Step 1*) using the previously determined (*Step 2*) inner weights. Finally, in *Step 4*, the outer weights (λ_1 to λ_4, μ_1 to μ_3) are calculated in two different ways, depending on the type of measurement model represented by each construct. If a construct is measured reflectively, then the correlations between the inner proxy of each latent construct and its indicator variables are applied (outer loadings). If a construct is measured formatively, then regression weights (i.e., outer weights) are applied that are the result of the ordinary least squares regression of each latent construct's inner proxy on its indicator variables.

The four steps in *Stage 1* are repeated until the change in the sum of outer weights between two iterations drops below a predetermined limit. A threshold value of 10–5 is recommended to ensure convergence of the algorithm and computational parsimony. If the algorithm converges in *Step 4* of *Stage 1*, then the final outer weights are used to compute the final latent construct scores in *Stage 2*, which are used to perform the ordinary least squares regressions for each construct to determine the path coefficients.

***Table** 4.1 Stages and steps in calculating the basic PLS-SEM algorithm*

Stage One: Iterative estimation of latent construct scores

$$\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \mu_1 = \mu_2 = \mu_3 = 1$$

Step 1: Outer approximation of latent construct scores (the scores of ξ_1, ξ_2, and η_1 are computed based on the manifest variables' scores and the outer coefficients)

$$\xi_1 = \lambda_1 x_1 + \lambda_2 x_2, \qquad \xi_2 = \lambda_3 x_3 + \lambda_4 x_4, \qquad \eta_1 = \mu_1 y_1 + \mu_2 y_2 + \mu_3 y_3$$

Step 2: Estimation of proxies for structural model relationships between latent constructs (γ_1 and γ_2)

$$\max_{a_1, a_2 \in \mathbf{R}} [R^2(\eta_1, a_1\xi_1 + a_2\xi_2)] = R^2(\eta_1, \gamma_1\xi_1 + \gamma_2\xi_2)$$

Step 3: Inner approximation of latent construct scores (based on scores for ξ_1, ξ_2, and η_1 from Step 1 and proxies for structural model relationships, γ_1 and γ_2, from Step 2)

$$\eta_1' = \gamma_1\xi_1 + \gamma_2\xi_2,; \xi_1' = {(\eta_1 - \gamma_2\xi_2)}/{\gamma_1}; \xi_2' = {(\eta_1 - \gamma_1\xi_1)}/{\gamma_2}; (\gamma_1, \gamma_2 \neq 0)$$

$$\eta_1 = \eta_1'; \ \xi_1 = \xi_1'; \ \xi_2 = \xi_2'$$

Step 4: Estimation of proxies for coefficients in the measurement models (the relationships between indicator variables and latent constructs with scores from Step 3; λ_1 to λ_4, μ_1 to μ_3)

$$\max_{a_1, a_2, a_3 \in \mathbf{R}} [R^2(\eta_1, \mathrm{a}_1 y_1 + \mathrm{a}_2 y_2 + \mathrm{a}_3 y_3)] = R^2(\eta_1, \mu_1' y_1 + \mu_2' y_2 + \mu_3' y_3)$$

$$\max_{a_1, a_2 \in \mathbf{R}} [R^2(\xi_1, \mathrm{a}_1 x_1 + \mathrm{a}_2 x_2)] = R^2(\eta_1, \lambda_1' x_1 + \lambda_2' x_2)$$

$$\max_{a_1, a_2 \in \mathbf{R}} [R^2(\xi_2, \mathrm{a}_1 x_3 + \mathrm{a}_2 x_4)] = R^2(\eta_1, \lambda_3' x_3 + \lambda_4' x_2)$$

$$\lambda_1 = \lambda_1'; \ \lambda_2 = \lambda_2'; \ \lambda_3 = \lambda_3'; \lambda_4 = \lambda_4'; \ \mu_1 = \mu_1'; \ \mu_2 = \mu_2'; \ \mu_3 = \mu_3'$$

Stage Two: Final estimates of coefficients (outer weights and loadings, structural model relationships) are determined using the ordinary least squares method for each partial regression in the PLS‑SEM model.

If $(\lambda_1 - \lambda_1')^2 + (\lambda_2 - \lambda_2')^2 + (\lambda_3 - \lambda_3')^2 + (\lambda_4 - \lambda_4')^2 + (\mu_1 - \mu_1')^2 + (\mu_2 - \mu_2')^2 + (\mu_3 - \mu_3')^2 \leq \delta$, δ *is a predefined positive value*.

Then estimate γ_1 and γ_2 and output λ_1, λ_2, λ_3, λ_4, μ_1, μ_2, μ_3.

4.2.2 A new algorithm based on PLS-SEM

There are two reasons why a new algorithm was proposed for calculating the model parameters in ***Fig*** *2.1*.

First, *unlike the model shown in Fig 2.1, the traditional PLS-SEM method requires that the relationships between the latent variables in the structural model should be "recursive"*. Since the exogenous latent variable η_1 is measured by only one indicator variable y_1, the adoption of the traditional algorithm to solve for model parameters would become equivalent to a two-stage regression modelling. While the first stage involves a regression model between the indicator variables in each group and y_1 and using the prediction as latent variable $\xi_1 - \xi_m$, in the second stage of the model, $\xi_1 - \xi_m$ is regressed to y_1 again. This is equivalent to a direct multiple regression model of x_1 - x_p with to y_1. Therefore, the PLS-SEM method requires the structural model to be "recursive", which is a precondition of traditional iteration algorithm (***Table 4.1***).

Second, a*lthough the traditional PLS-SEM algorithm requires unique association of indicator variables with a latent variables, sometimes the indicators are associated with multiple latent variables*. However, in an exploratory model, it often desires to have the least restrictions for an initial model and the indicator variables may have multiple associations with the latent variables.

Prior to proposing a new method for parameter estimation, basic rules for a "good" model need

to be outlined as below:

Rule one: the extracted latent variables should be as independent as possible. This comes from the concept of factor analysis. If the measurement model is considered in analogy to a factor model, the extracted latent variables should represent independent factors to reflect different geological processes.

*Rule two: the extracted latent variables should be related to exogenous latent variable (η_1 in **Fig** 2.1)*. Because the proposed exogenous latent variable include only one indicator (y_1), the extracted latent variables should be related to y_1 in ***Fig*** *2.1*. That is, the correlation of each extracted latent variable with target y_1 should be as strong as possible.

In general, the proposed algorithm is considerably a type of factor analysis under the restriction of a regression to a target, or multiple regression beyond a series of un-decided independent variables (latent variables).

To describe the two rules mentioned above, a goal function, for example in ***Fig*** *2.1*, can be proposed as follows:

$$F = \frac{\sum_{i=1}^{m} d(\xi_i,\eta_1)}{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} d(\xi_i,\xi_j)} \qquad (4.1)$$

wherein, $d(\xi_i,\xi_j)$ represents the distance between centroids of two group variables ξ_i and ξ_j and $d(\xi_i,\eta_1)$ represents the distance between latent variables ξ_i and η_1; ξ and η are defined

in ***Eq*** *2.1* and *2.2* in Chapter 2. If $R^2\left(\xi_i,\xi_j\right)$ is a correlation of determination between ξ_i and ξ_j, $d(\xi_i,\xi_j)$ is defined as $d\left(\xi_i,\xi_j\right)=1-R^2\left(\xi_i,\xi_j\right)$. Since the score of ξ depends on the coefficients in the measurement model and the value of target function depends on the score of ξ, F can be minimalized through changing the coefficients in measure model. The proposed method is an optimization of the target function while the optimum function in R has been introduced in Chapter 3.

The process can be described as ***Fig*** *4.2*. The initial input of the algorithm is the coefficients in the measurement models, which is a group of random numbers with the range of -1 to 1. With the defined target function, the coefficients in the measurement models can be calculated through an optima function in R language. In order to reduce the effect of the initial value in optima function, the calculation will be repeated until the outer coefficients reach the requirement of algorithm.

For the model showed in ***Fig*** *2.1*, latent variable η_1 includes only one observation variable y_1, so the objective function for the model in ***Fig*** *2.1* can be expressed by ***Eq*** *4.2* as well, which will be applied in actual calculation. Otherwise, if η_1 includes more than one variables (e.g. y_1, y_2 and y_3 in ***Fig*** *4.1*), η_1 can be represented by the first component through PCA method.

$$F=\frac{\sum_{i=1}^{m} d(\xi_i,y_1)}{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} d(\xi_i,\xi_j)} \qquad (4.2)$$

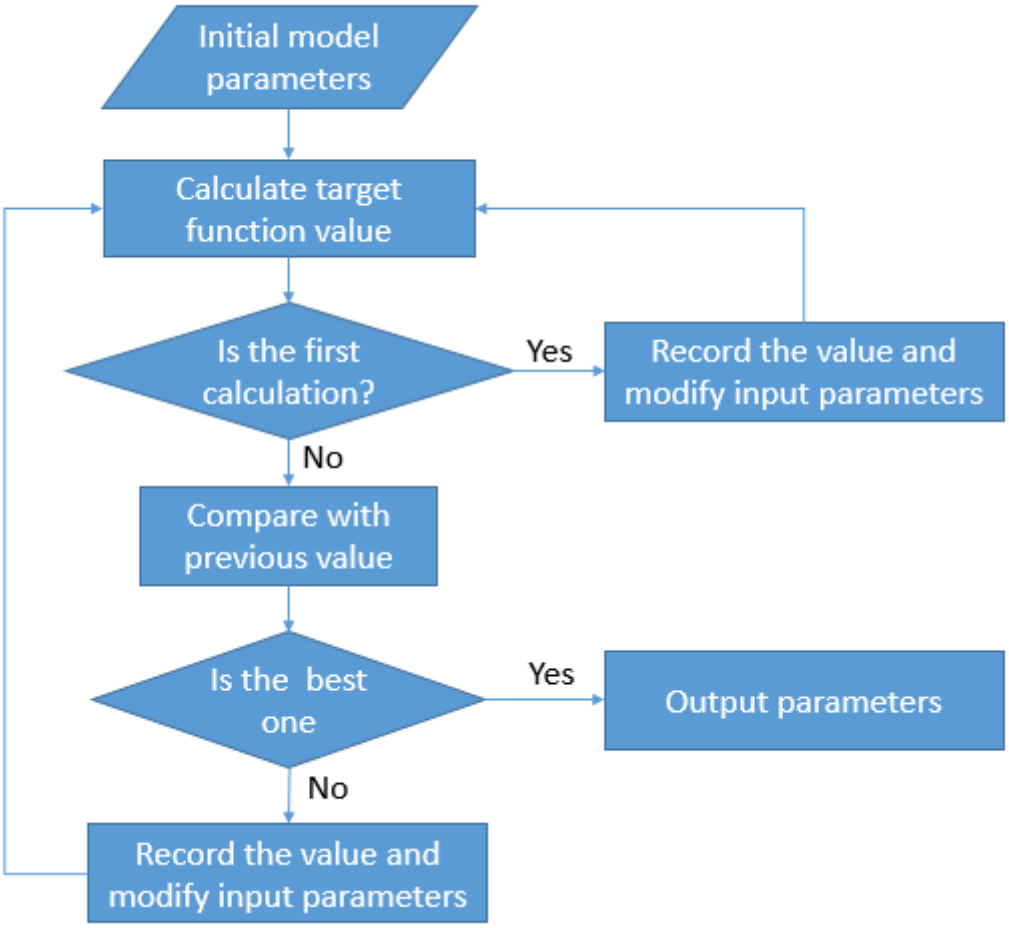


***Fig** 4.2 A new method for estimating PLS-SEM parameters.*

4.3 MLR and FA

Multiple linear regression analysis is used to predict the value of one or more responses from a set of predictors and also to estimate the linear association between the predictors and responses. The structural model in SEM can be considered as a MLR. The model calculated in ***Fig** 4.3* has one dependent variable (Var4) and 3 independent variables (Var1, Var2 and Var3).

Factor analysis was originally developed for the case of one common factor by Spearman (1904), and then later to the case of multiple factors generalized by Thurstone (1947) and others (Mulaik, 2009; Comrey and Lee, 2013; Harman, 1976; Iacobucci, 1994; Kim and Mueller, 1978a, b; Lawley and Maxwell, 1967), and then adopted to extract "controlling processes" in geochemistry by geoscientist . It can be used to explore a large data set for hidden multivariate data structures. An advantage of it in regional geochemical data processing is to reduce the data dimensions with a minimal loss of information, usually used to reveal unrecognized multivariate structures in the data that may be indicative of certain geochemical processes, or, of hidden mineral deposits in exploration geochemistry (Reimann et al., 2002). The extracted latent variables can be thought as the factors, which represent the main information of correspond observing variables, because the measurement model of SEM is a type of factor analysis.

Since the specified SEM in this research involves both of the MLR and CFA (or EFA) models, a MLR model will be created between 15 elements and ***As*** and a factor analysis is introduced separately, which corresponds to the structural model and measurement model. And then the predicted score maps of ***As*** are obtained from MLR based on 15 other elements, MLR based on 3 factors (calculated from EFA), and SEM, respectively.

The parameters of MLR and EFA are calculated using the functions in R language library.

4.4 Case study

In order to apply the SEM introduced in Section 2.3 to extract factors (combination of elements) that characterize gold mineralization in the study area, all 16 geochemical elements, whose statistic information together with the and geology background of study area has been introduced in Section 3.1 and 3.2, were chosen as variables to create a SEM.

4.4.1 Construction and refinement of structure equation model

In the SEM analysis, a predefined structural model is needed for parameter estimation and hypothesis examination. However, for exploratory analysis and data mining purpose, a structural model is needed to order to apply the SEM to the data. The challenge is that there does not exist mathematical method for model construction in the current SEM.

In this research, a new method has been proposed to assist the model construction in SEM that involves one level measurement and structural models. This method uses a random sampling technology to classify the measurement variables $x_1, x_2, ..., x_p$ into certain distinct groups under the conditions that the ratio of the distance between these groups and an objective variable such as η_1 over the distance between these groups themselves is minimized. The distance between a group of measurement variables and an objective variable is defined as the regression error between the variables in the group and an objective variable while the distance between two groups is computed as the Euclidean distance of two group centers (as vectors).

In order to form a structure model for SEM, only one representative variable from each group will be used as the characteristic variables to define the latent variables to be included in the structure model. The characteristic variables will be embodied exclusively in only one latent variable. For example, the variable in each group with the largest regression coefficient with respect to target variable y_1 can be treated as the main variable for each group. Except the main variables, the remaining variables in X will be embraced in all latent variables. By using such a method, an initial model is constructed for subsequent calculation.

Regarding the case study of the factor extraction in characterization of gold mineralization in the study area, all 15 geochemical elements were chosen as inputting variables for the classification with ***As*** as the dependent variable y_1 *(**Eq** 4.3)*. There were two main reasons to choose ***As*** instead of ***Au*** as the dependent variable for the classification. Firstly, ***As*** was highly correlated with ***Au*** in the gold mineralization in the area (Agterberg et al., 1990; Xu and Cheng, 2001). Secondly, the concentration values of ***Au*** from some samples were below the detection limit. The results obtained using the classification method introduced in ***Eq** 4.2* are shown in ***Table** 4.2*. The regression coefficients in each group are shown in ***Fig** 4.3*.

***Table** 4.2 The results obtained by the classification with **As** as the objective variable*

Group	Elements
Group 1	***Au***, ***Cu***, ***Sb***, ***Th***, ***Zr***
Group 2	***F***, ***Li***, ***Nb***, ***Pb***, ***Rb***, ***Ti***, ***Zn***
Group 3	***Ag***, ***Sn***, ***W***

As can be seen in ***Fig*** *4.3*, three groups of elements classified according to the linear regression relationship to ***As*** may represent three distinct but ***As*** associated geochemical factors. The first group indicates a factor of ***Au***, ***Cu*** and ***Th*** dominated which may imply ***Au*** and ***Cu*** mineralization. The second group is mainly associated with ***Zn***, ***Pb*** and ***F*** which may imply ***Pb*** and ***Zn*** mineralization. The third group may represent ***Ag*** and ***W*** mineralization. These results demonstrated that the arsenic may be involved in multiple mineralization processes in the study area. To further comprehensively evaluate all elements related to the three groups one dominate element from each group were chosen to form a SEM model. For example, ***Cu***, ***Zn*** and ***W*** were selected from these three groups and used as unique variables in forming each latent variable in the structural model. Three latent variables were defined in such that each latent variable include one of the three chosen elements ***Cu***, ***Zn*** or ***W*** and the remaining elements. The arsenic served as the measured variable of endogenous latent variable. The SEM model is shown in ***Fig*** *4.4*.

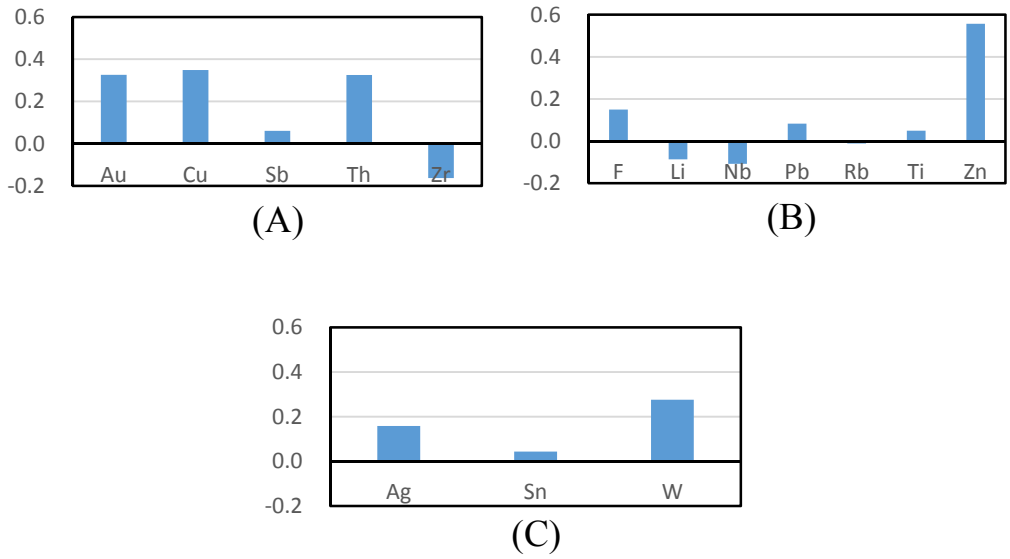


Fig *4.3 Regression coefficients of elements in each groups with* ***As*** *as dependent variable. (A) –(C) Groups 1-3, respectively. The elements with the largest value in each group are: Cu (Group1); Zn (Group2) and W (Group3). The analysis is based on a log-transformed dataset.*

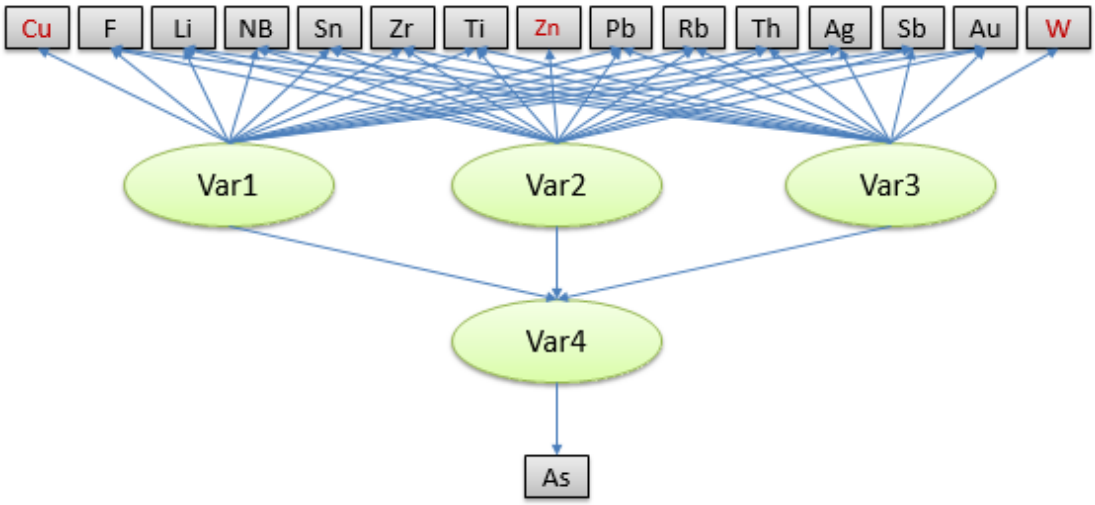


Fig *4.4 The SEM map for multi-geochemical elements and arsenic. The distinct element of each group is marked in red color.*

4.4.2 The results

The values for all 39 parameters involved in the SEM model inclusive of 3 regression coefficients and 36 loadings of the three latent variables (Var1, Var2 and Var3) are shown in ***Tables*** *4.3*, respectively. The loadings on the three latent variables showed that the first latent variable was dominated by elements ***Au***, ***Cu*** and ***Th*** positively and ***Li***, ***Nb*** and ***Zr*** negatively; the second latent variable dominated by ***Zn***, ***Pb*** and ***Au*** positively and ***F*** and ***Sb*** negatively; and the third latent variable dominated by ***Au***, ***F***, ***Pb***, ***Sb***, ***Th***, ***Ti*** and ***W*** positively and ***Nb*** and ***Zr*** negatively. These results had some similarities as those obtained by the classification method (***Eq*** *4.3*) but there were notable differences. For example, on one hand, the main elements classified in each group still remained as main elements on the corresponding latent variables according to the loadings of the elements. On the other hand, the associations of elements with the latent variables were different from those classified in each group. For example, ***Au*** shows significant loadings on all three latent variables whereas only in first group classified by the classification method (***Eq*** *4.3*). The results obtained by the SEM were more reasonable and all three latent variables were associated with ***Au*** that implied the three latent variables representing the gold mineralization associated geochemical factors. The regression coefficients obtained for three latent variables were 1.59, 2.03 and 1.20, respectively, whose significance were found at the same level with t-values= 9.96. The results indicated that all three latent variables were statistically correlated with the dependent variable ***As***.

***Table** 4.3 Regression coefficient from MLR and SEM.*

	MLR	Var1	Var2	Var3
Ag	0.068	0.034	-0.004	0.019
Au	0.317	0.120	0.023	0.066
Cu	0.141	0.089	N/A	N/A
F	0.077	0.028	-0.026	0.071
Li	-0.119	-0.031	-0.052	0.030
Nb	-0.116	-0.052	-0.002	-0.025
Pb	0.080	0.024	-0.008	0.048
Rb	0.097	-0.011	0.045	0.019
Sb	-0.007	0.009	-0.031	0.035
Sn	0.001	-0.007	0.001	0.009
Th	0.168	0.060	-0.015	0.087
Ti	0.026	0.002	-0.008	0.032
W	0.105	N/A	N/A	0.088
Zn	0.409	N/A	0.202	N/A
Zr	-0.166	-0.072	0.006	-0.053

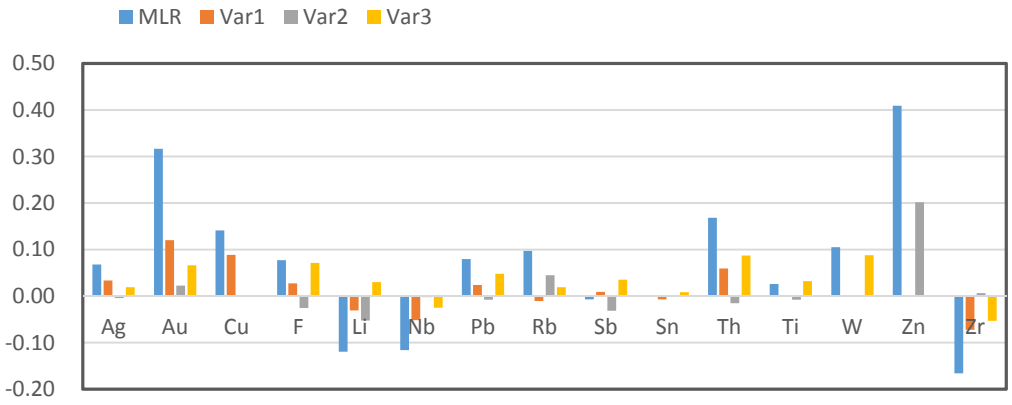


***Fig** 4.5 The regression coefficients in SEM (measurement model) and MLR. The analysis is based on a log-transformed dataset.*

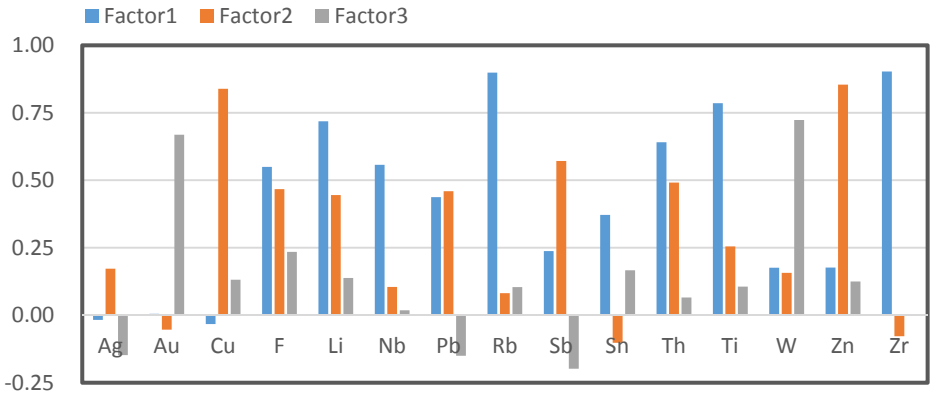


***Fig** 4.6 Loadings on factors obtained by FA. The analysis is based on a log-transformed dataset.*

4.4.3 Comparisons between SEM, MLR and FA

In order to compare the results of SEM with the results from other relevant methods the MLR and FA were implemented to the same data. The MLR was performed in a regression model between the dependent variable (***As***) and 15 interdependent variables (15 elements). The regression coefficients obtained for 15 elements are shown in ***Fig***. *4.5*. The square of the regression correlation coefficient of the MLR was R^2 = 0.57 and the critical value in the regression between prediction obtained from MLR and ***As*** was t = 28.80, which implied a significant correlation between ***As*** and the regression function of 15 elements.

Further, the FA was applied to the same data of all 15 elements (without ***As***). The loadings on the first three factors obtained by FA are shown in ***Fig***. *4.6*.

As shown in ***Fig***. *4.6*, the first factor represented the association of ***Rb***, ***Zn***, ***Ti***, ***Li***, ***Th***, ***Pb*** and ***Sn***. The second component was associated with ***Cu***, ***Zn***, ***Sb***, ***Th***, ***Pb***, ***F*** and ***Li***. The third factor mainly represented the association of ***Au*** and ***W***. The results obtained by FA were very different from the three latent variables obtained by SEM. The loadings of three factors obtained by FA and SEM represented different geochemical factors to imply different geological processes. To further compare the similarities and differences among the results from SEM, MLR and FA, the spatial distribution of scores calculated using these three methods were analyzed and will be given in the next section. The scores on the three factors or latent variables obtained by SEM and FA are shown in ***Figs*** *4.7A-F*. For comparison, the log-transformed values of ***As*** were also mapped (***Fig*** *4.7G*). The results obtained by SEM and FA as shown in ***Fig*** *4.7* are geologically

meaningful. For example, the Goldenville formation showed the elevated values of ***As***. The scores on all three latent variables obtained by SEM and scores on the second factor by FA showed high values in and around Goldenville formation. In addition, the high score values on the maps showed spatial association with the location of gold mineral deposits and mineral occurrences which may imply that the anomalous values of scores on these factors were associated with gold mineralization. The observed values of ***As*** are plotted against the scores on three factors calculated by FA and three latent variables by SEM in ***Fig*** *4.8*. The correlation coefficients calculated on the basis of these plots were generally low although the R^2 values obtained between ***As*** value and latent variables were generally higher than those obtained between ***As*** value and factors by FA. For example, the correlation coefficients calculated between ***As*** and the three factors obtained by FA and three latent variables by SEM are shown ***Fig*** *4.8*. The results were R^2 = 0.20(factor 1), 0.20(factor 2), 0.09(factor 3) and 0.37 for latent variables 1, 2 and 3. All of the T-test values are more than 8. The correlation coefficients also indicated that the correlation between latent variables and ***As*** were generally stronger than those between three factors and ***As***. The angles between each pair of latent variables (as vectors) were also calculated, about 62 degrees for all pairs which implied these latent variables to have some degree of non-correlation. In order to further compare the predictability of ***As*** by linear regression based on the three factors of FA, the three latent variables of SEM and all 15 elements, three MLRs were created. Further, ***Figs*** *4.8A-C* show the predicted values of ***As*** based on FA factors, latent variables and 15 elements, respectively. The corresponding correlation coefficients between the observed values of ***As*** and three predicted values of ***As*** were 0.49, 0.57 and 0.57 (t-values = 24, 28, 28) (***Figs*** *4.9A – C*). The results in ***Fig*** *4.9* indicated the similar

trends of predicted values of ***As*** by the three methods although the results obtained by SEM method and by the all elements give larger multiple correlation coefficients between the predicted values and the observed values of ***As***. The results indicated that the multiple correlation coefficients obtained by SEM method and by the all element are the same which is slightly higher than that obtained by FA method. ***Fig*** *4.10* shows the distributions of regressed values of ***As*** with all elements or latent variables obtained by two methods: SEM and FA. If the prediction for ***As*** through a multivariate linear regression based on 15 elements represents a factor with maximum correlation with ***As*** (***Fig*** *4.10C,* the correlation coefficient with ***As*** is 0.75), the prediction of ***As*** through the latent variables (***Fig*** *4.10B*) should include all variance related with ***As*** in 15 elements (the correlation coefficient with ***As*** is 0.75/0.75), but the prediction through top three principal components may only represent partial correlation (the correlation coefficient with ***As*** is 0.70/0.75). The regressions based on SEM latent variables and all 15 elements gave the same coefficient implying that the prediction of ***As*** by the three latent variables generated by SEM reached the same result as a global MLR model based on all elements.

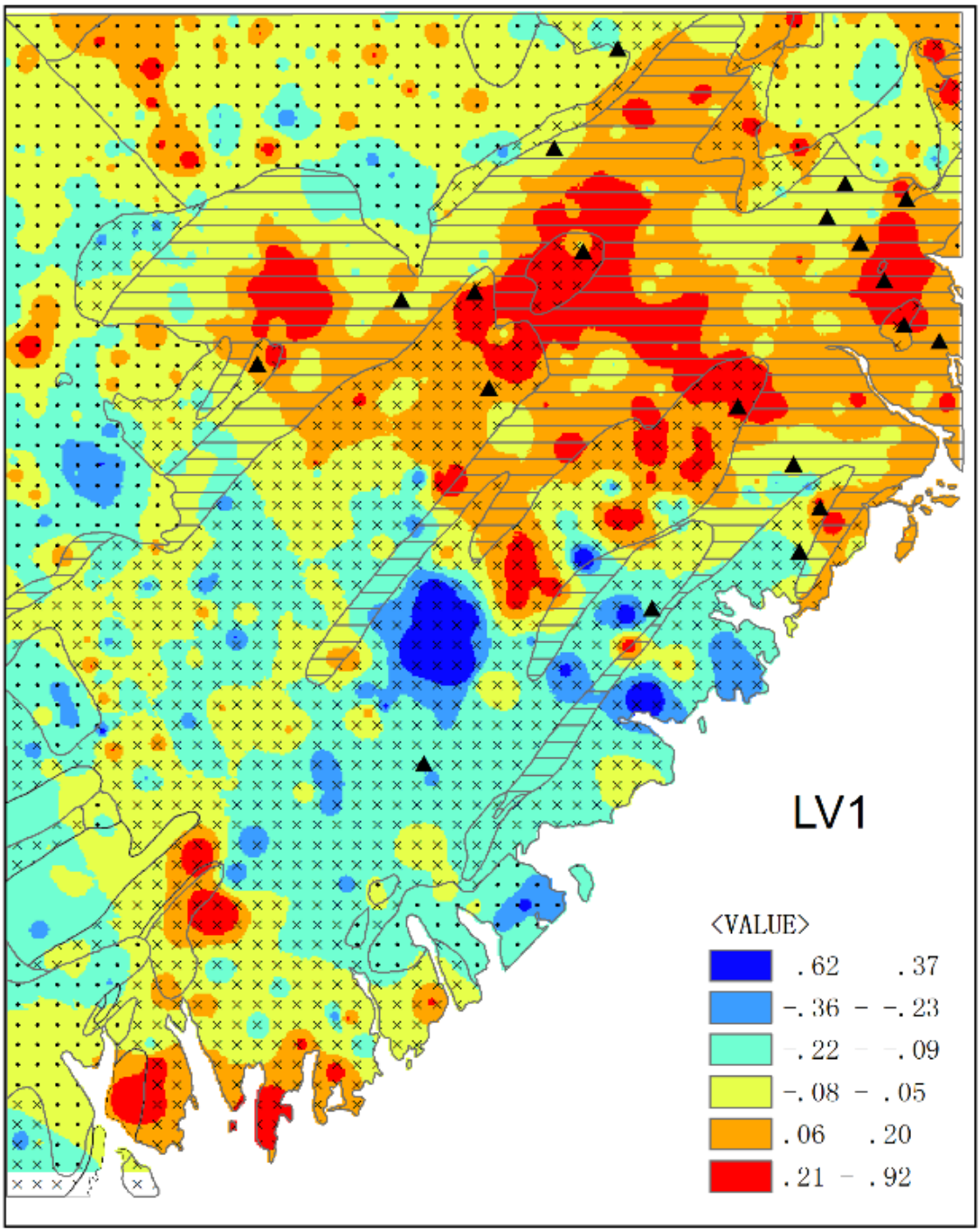


***Fig** 4.7A Scores of the first latent variable obtained from SEM.*

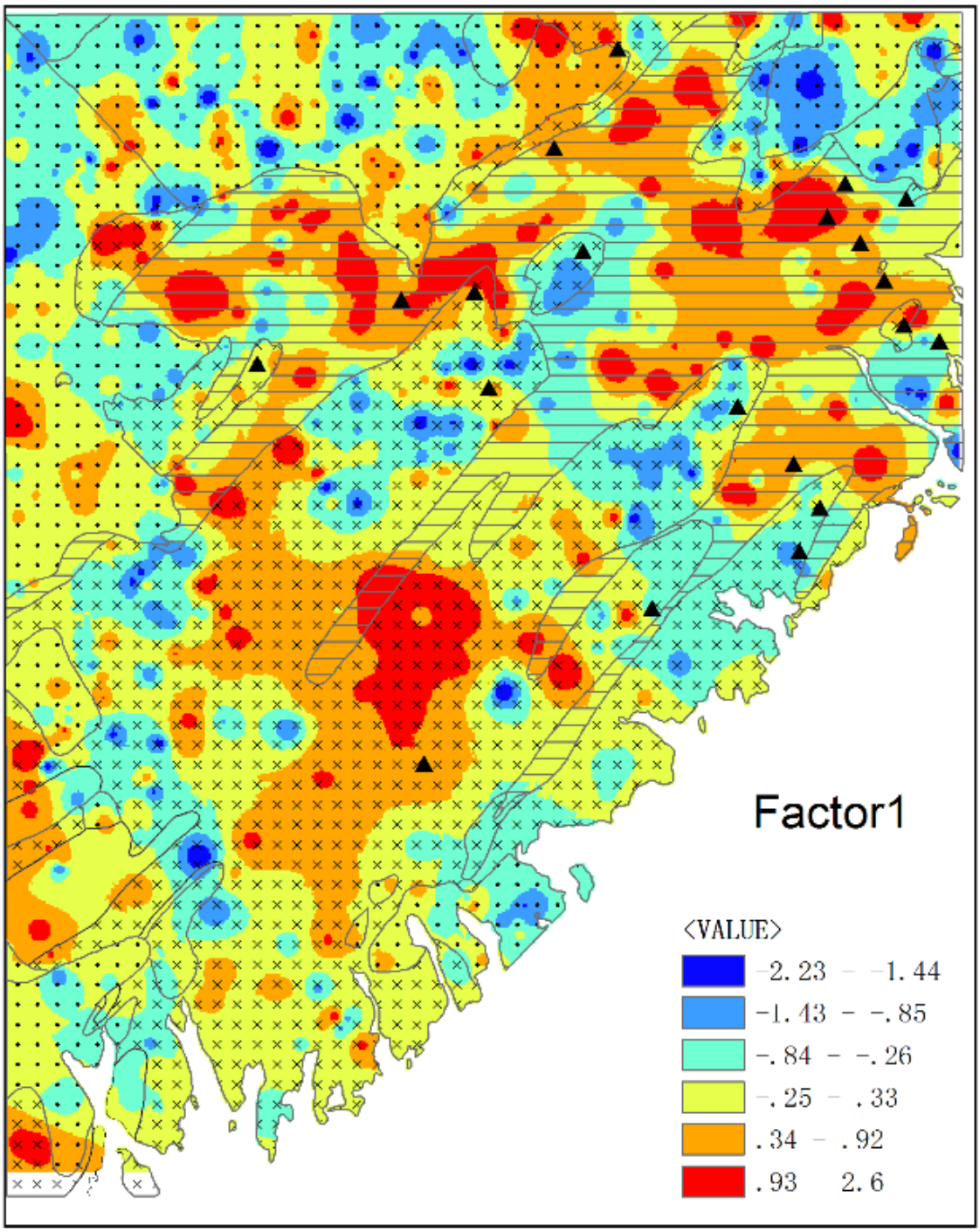


***Fig** 4.7B Scores of the first factor obtained from FA.*

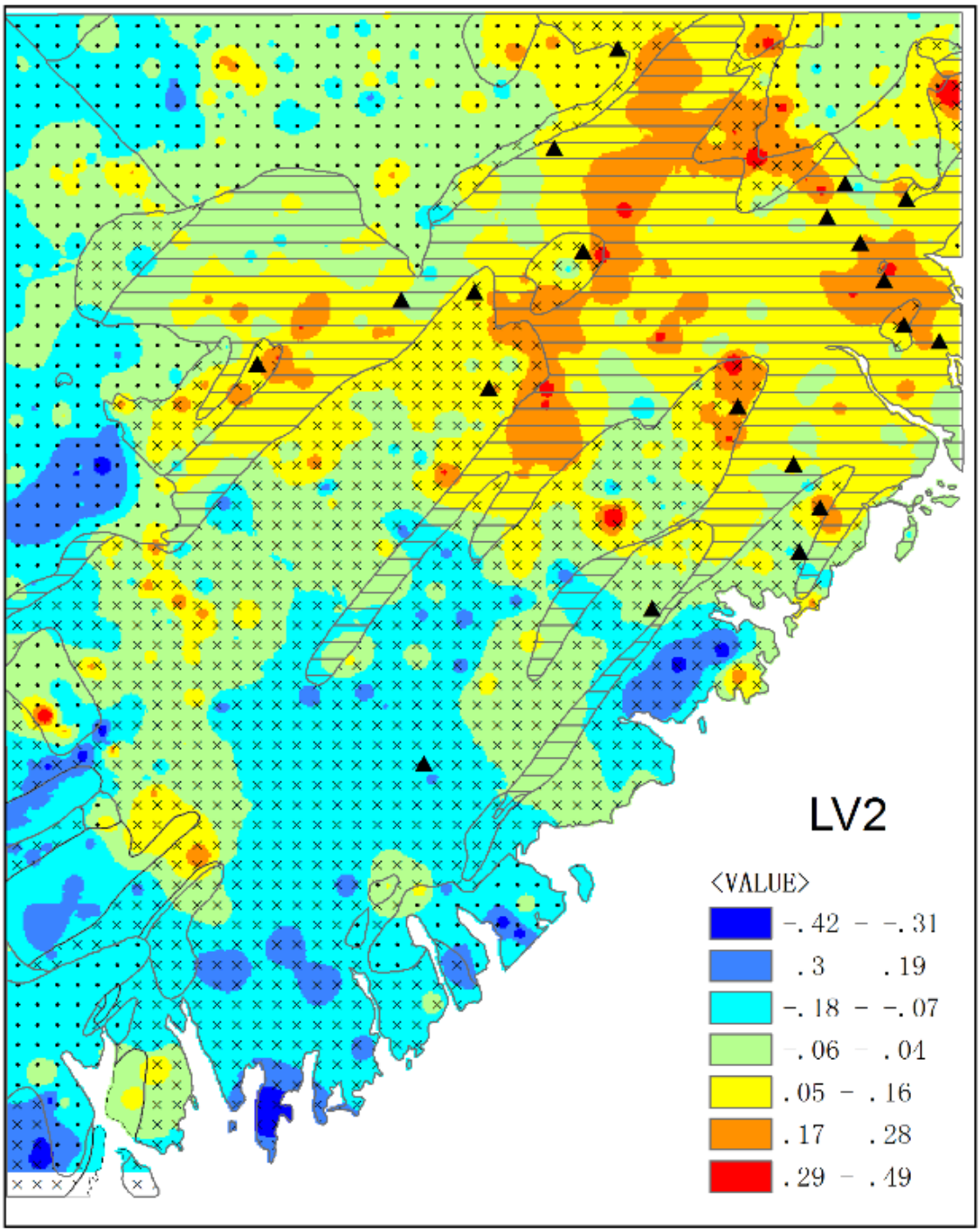


***Fig** 4.7C Scores of the second latent variable obtained from SEM.*

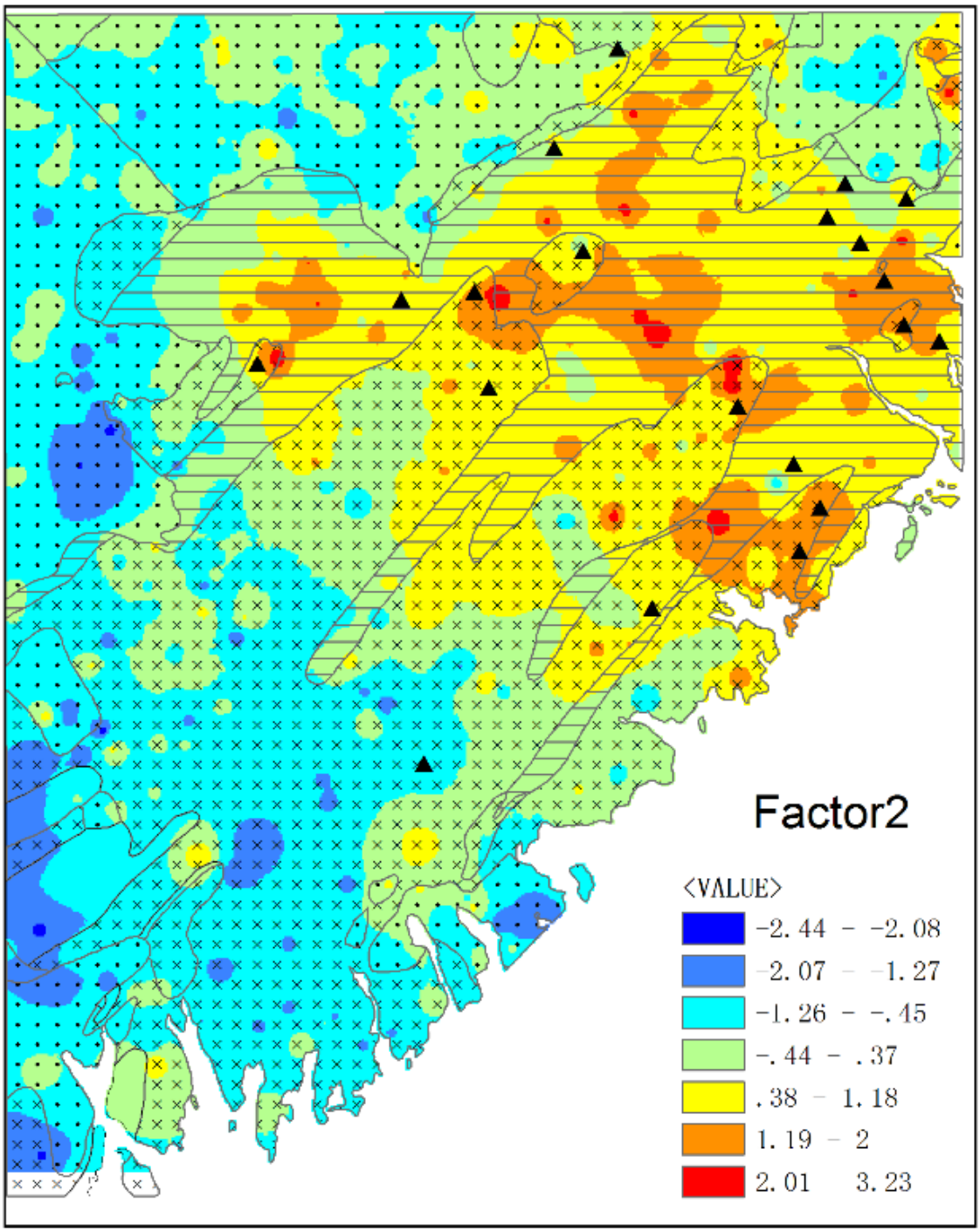


***Fig** 4.7D Scores of the second factor obtained from FA.*

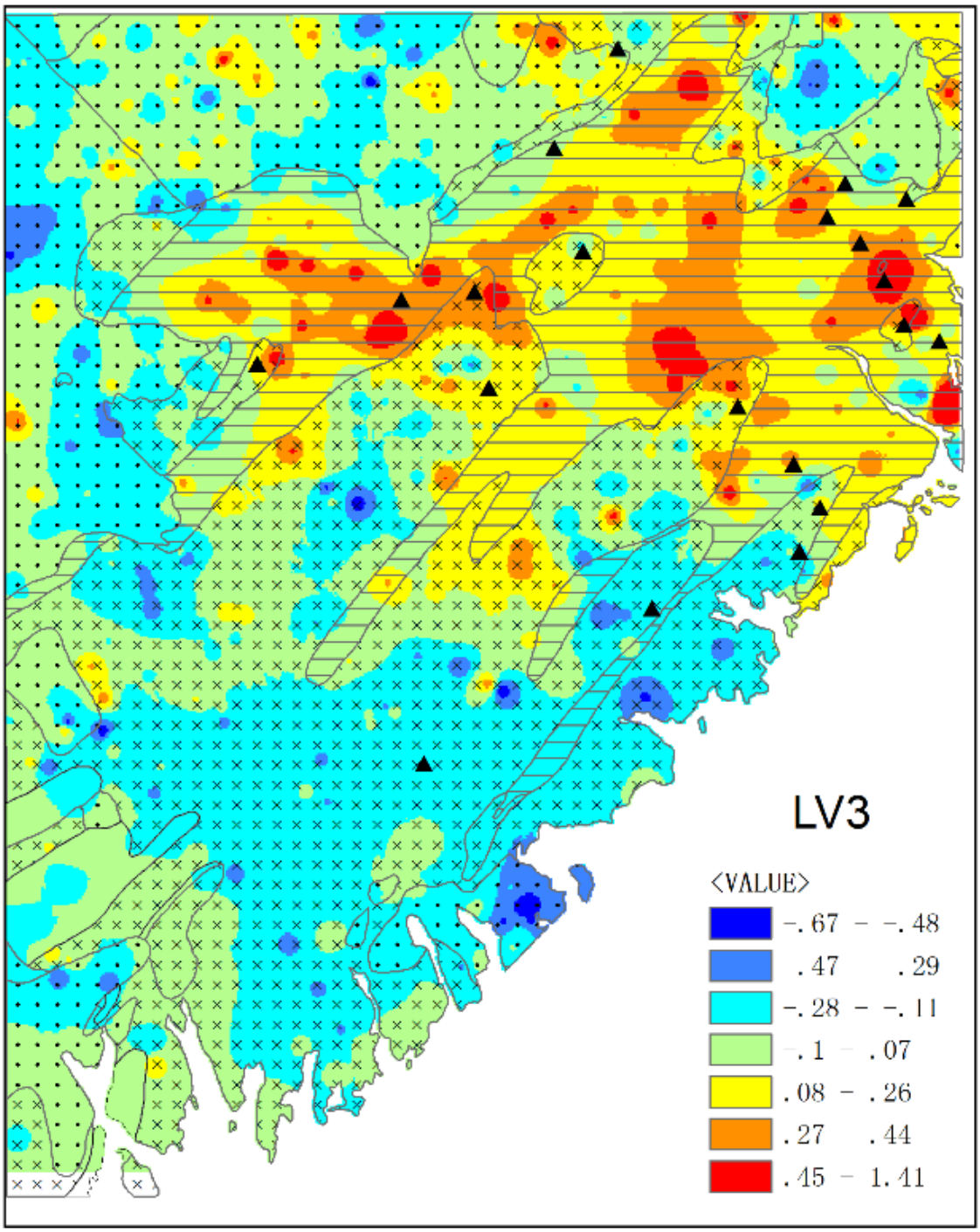


***Fig** 4.7E Scores of the third latent variable obtained from SEM.*

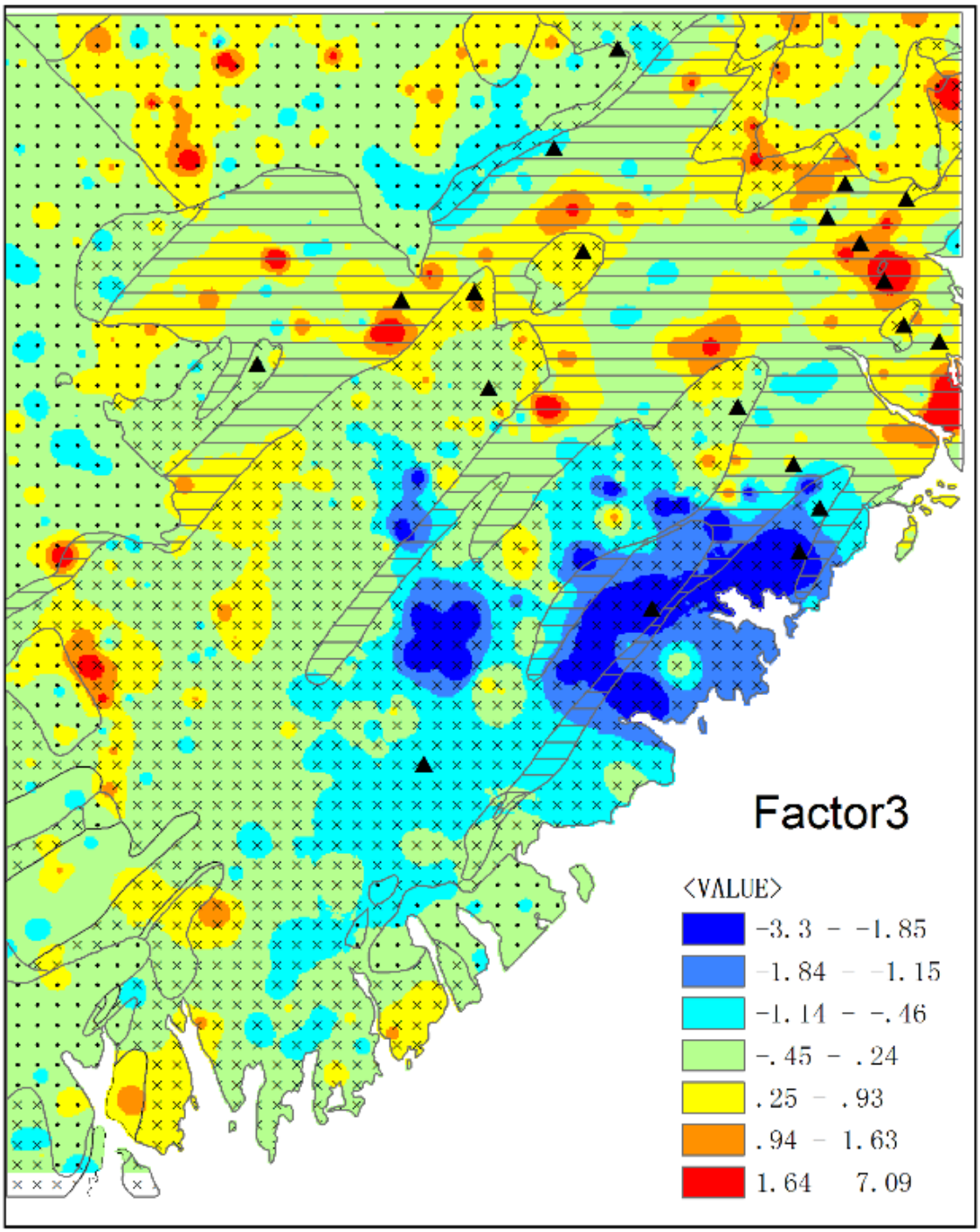


***Fig** 4.7F Scores of the third factor obtained from FA.*

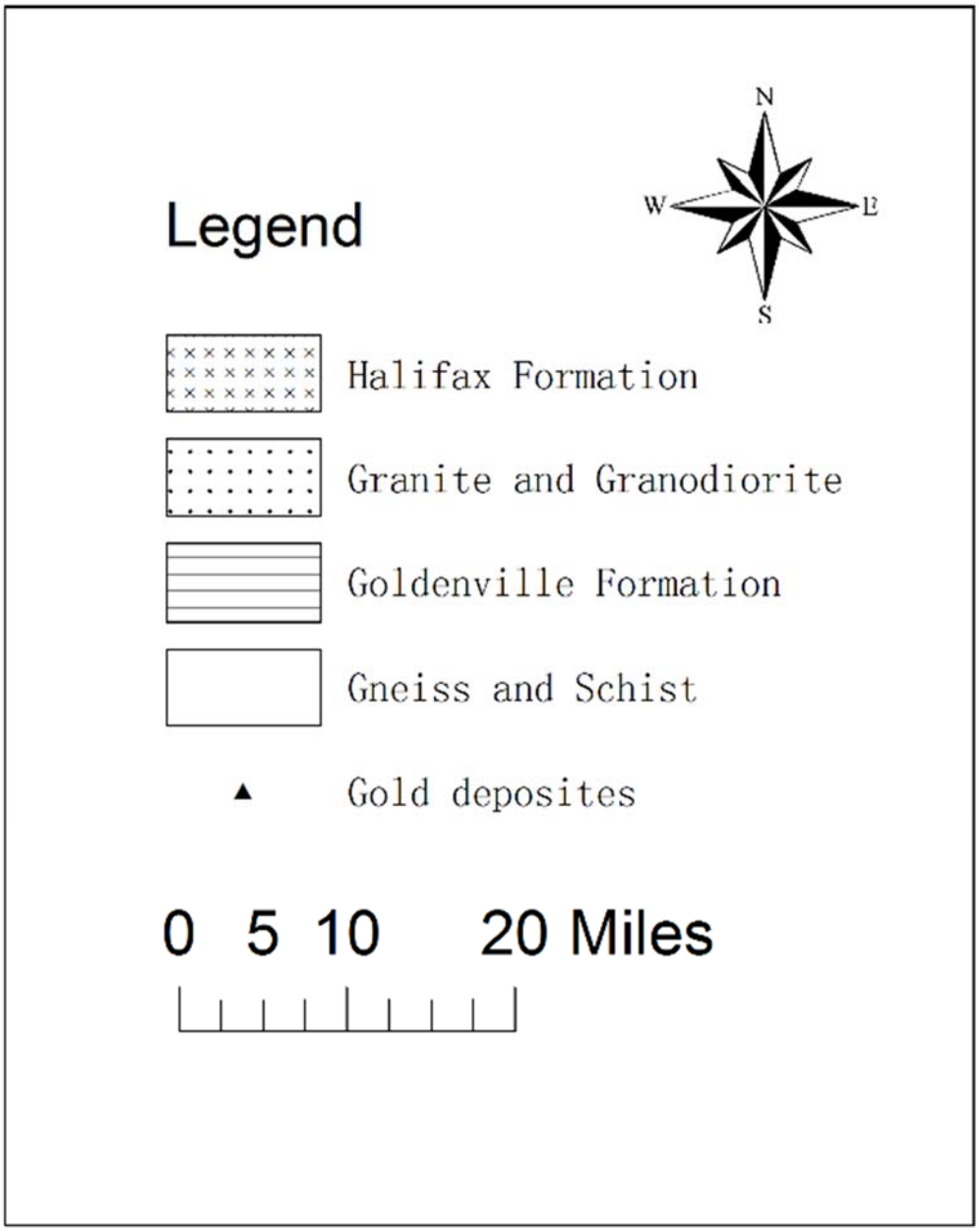


***Fig** 4.7G Legend, north arrow and scale bar for Figs 4.7A-F.*

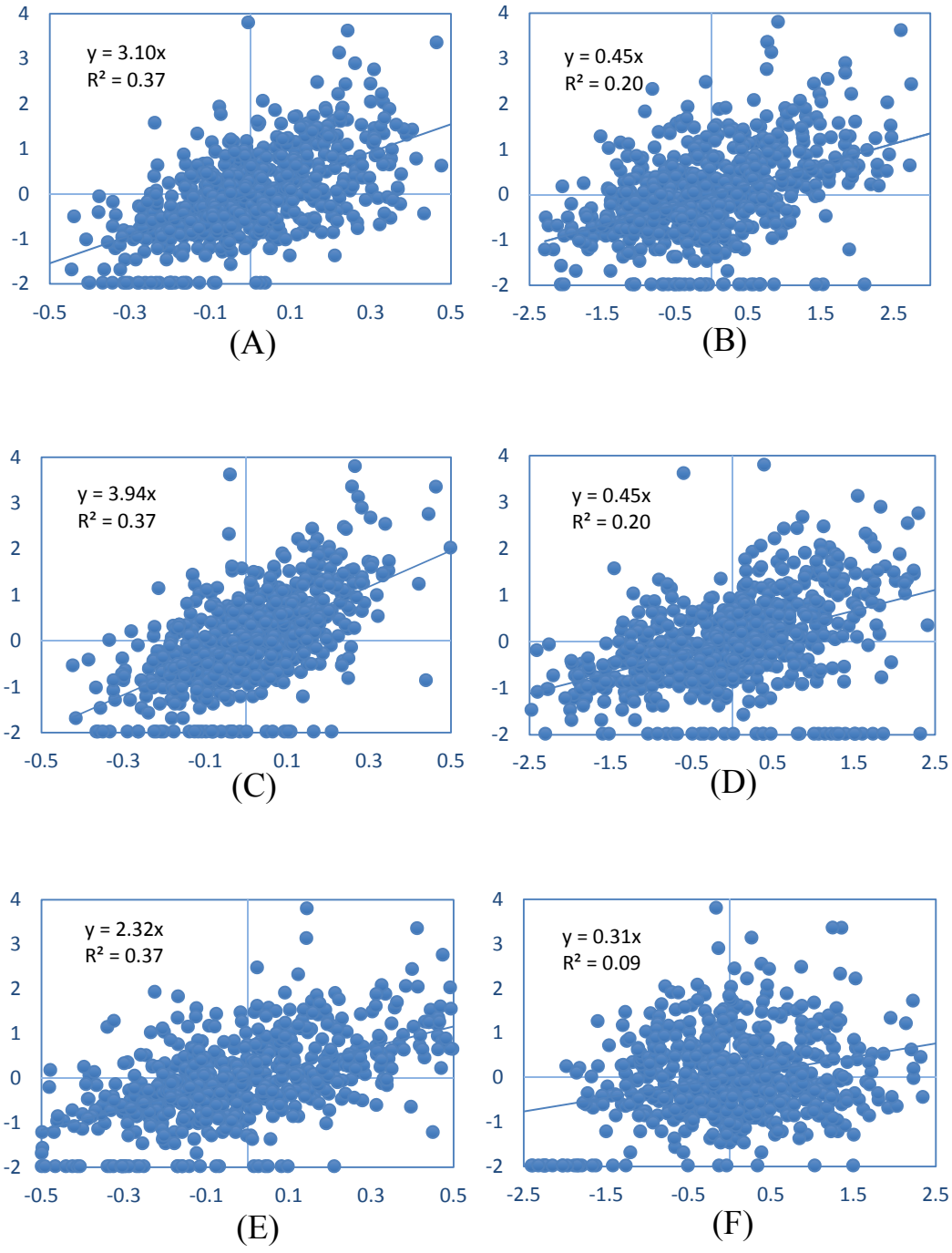


Fig *4.8 Relationships between the calculated variables and As. (A), (C), (E): latent variable 1, latent variable 2 and latent variable 3 with As, (B), (D) and (F): factor 1, factor 2 and factor 3 with As.*

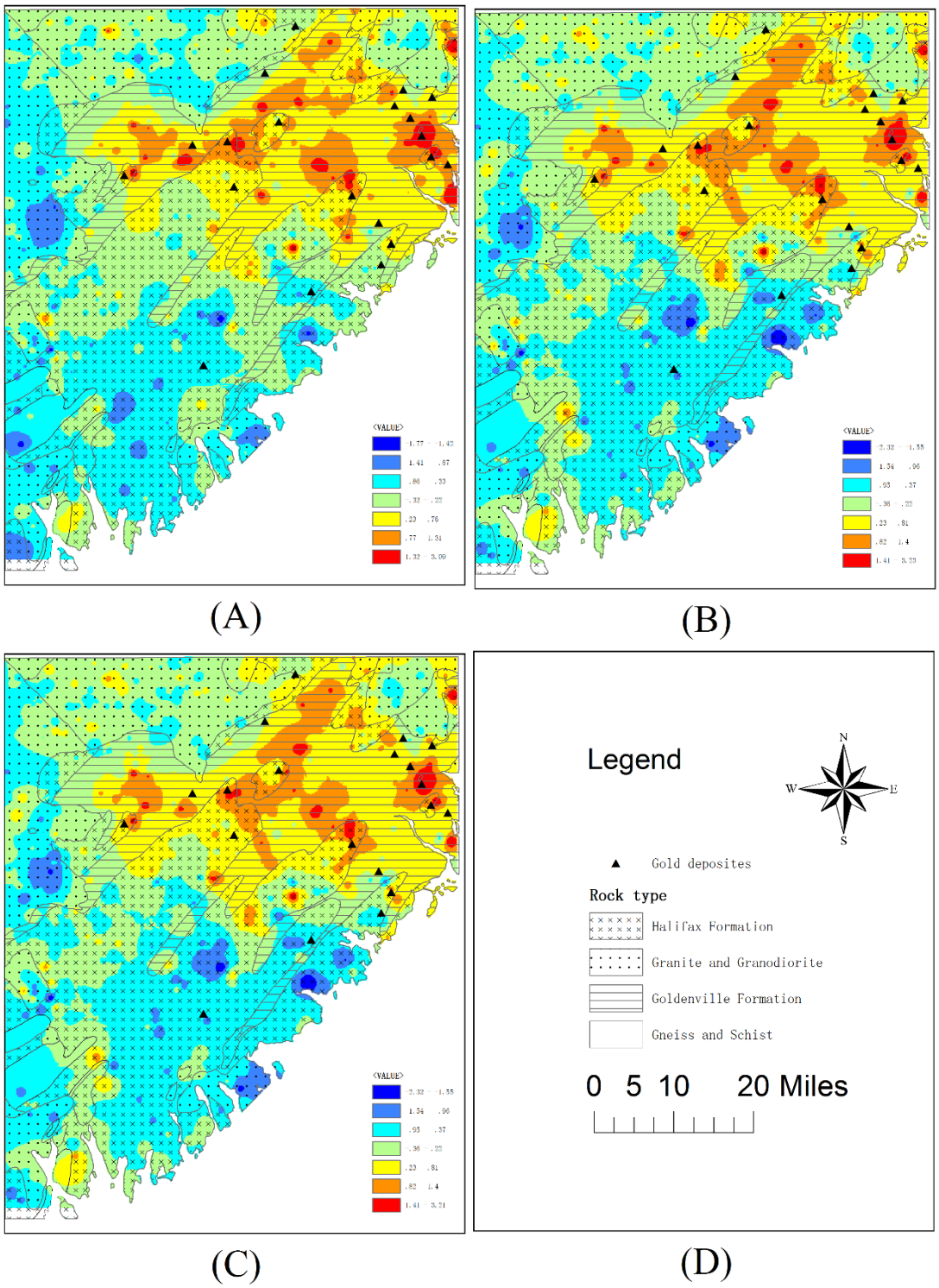
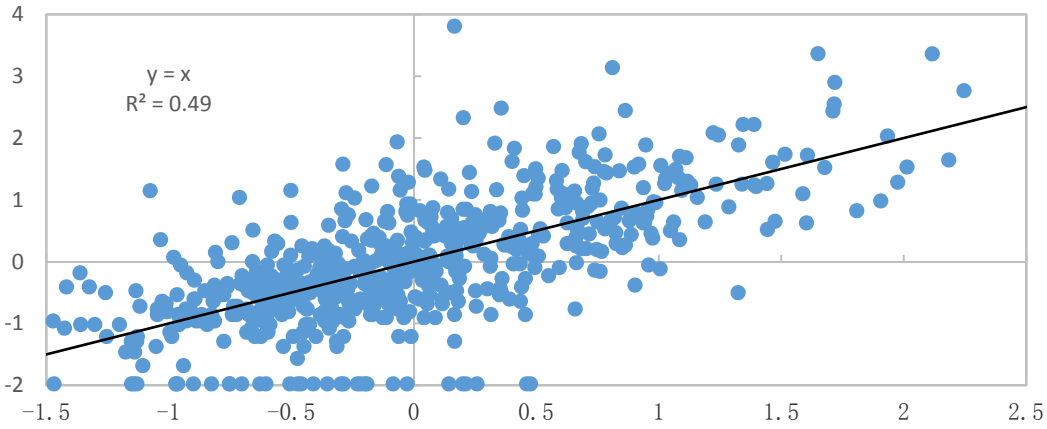
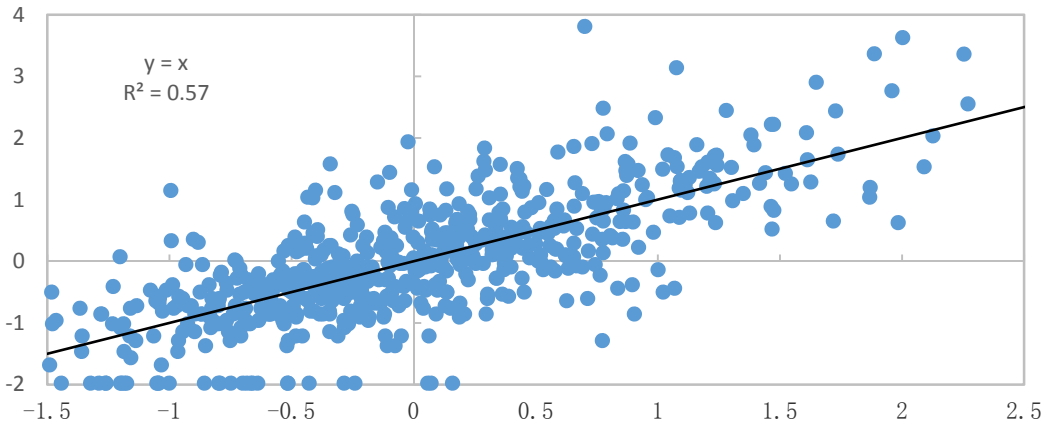


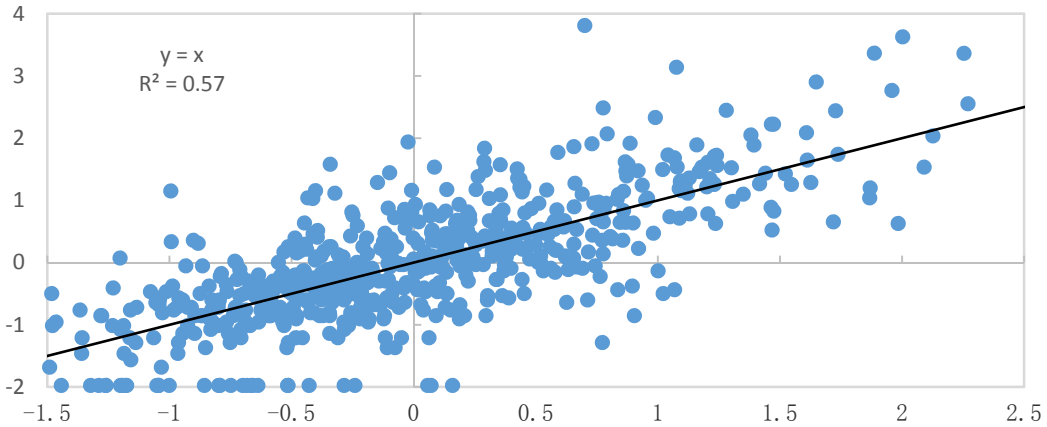
***Fig** 4.9 The estimated values for As by three linear regression models: (A) three factors obtained by FA as independent variables; (B) three latent variables obtained by SEM as independent variables and (C) transformed values of 15 elements as independent variables.*



(A)



(B)



(C)

Fig *4.10 The observed value and predicted values of As by three different methods: (A) factors obtained by FA; (B) three latent variables by SEM and (C) 15 elements.*

4.5 Discussion and conclusions

This chapter provided the first attempt to apply the SEM technique to extract mineralization associated geochemical factors and has proposed a solution for SEM to be used as an exploratory tool in geochemical data processing. With the new method, an initial model could be created by a new cluster analysis under the constraint of an objective variable. Then, a more general SEM model could be constructed and the regression coefficients involved in SEM could be estimated. A case study was conducted to validate the method using the concentration values of 16 geochemical elements from 671 lake sediment samples. Three latent variables or factors were obtained to characterize gold mineralization in the study area. The three latent variables were characterized by three groups of elements (***Au***, ***Cu*** and ***Th***), (***Zn***, ***Pb*** and ***Au***) and (***Au***, ***F***, ***Pb***, ***Sb***, ***Th***, ***Ti*** and ***W***) which may imply three types of gold mineralization associated processes. At the same time, the angles between each pair of latent variables (as vectors) were 62 degrees. If 90 degrees mean independence with each other, i.e. the factors from FA, it implied that these latent variables have had certain degree of non-correlation. The comparisons between SEM and other methods such as MLR and FA demonstrated that the results obtained by SEM were different from those obtained either by MLR or FA. Unlike FA, the new SEM gave the factors with ***As*** as a constraint, in other words, SEM gave ***As*** associated factors whereas FA generated the factors without the external constraint. Therefore, the results obtained by FA may or may not be associated with any objective variables. From this respect, SEM could be considered as the external variable constrained FA. The comparison between SEM and MLR indicated that instead of global regression as MLR between ***As*** and all other elements, the SEM created a two-

step regression: regression between ***As*** and latent variables and regression between latent variables and independent variables. The latter could be considered as a decomposed regression of MLR.

Chapter 5 A modified WofE method based on SEM concept

5.1 Introduction

In this Chapter, the SEM concept is utilized to improve the effect of the conditional independence (CI) in applications of the Weight of Evidence (WofE) method for mineral prediction.

WofE is an artificial intelligent quantitative method based on Bayes' rule, predicts the presence or absence of events by the integration of the given information, which can be seen as evidence layers. As discussed in Schaeben (2014) referring to Markov random fields and log-linear models, the WofE method is a special case of logistic regression model, and was originally designed for a non-spatial application in medical diagnosis (Agterberg, 1989). The WofE was first introduced into mineral potential mapping by Bonham-Carter et al (1989) and Agterberg et al. (1989; 1990).

Currently, WofE, as one of the most popular models using Bayes' theory of conditional probability, is being utilized to quantify spatial association between the evidence layers (or geological factors) and the known mineral occurrences (Agterberg, 1989; Bonham-Carter, 1994; Carranza, 2004; Cassard et al., 2008; Cheng, 2008; Porwal et al., 2010). The WofE models are also used to evaluate landslide sensitivity (e.g., Cervi et al., 2010; Lee et al., 2004; Neuhäuser and Terhorst, 2007) and ecology mapping (e.g., Gorney et al., 2011; Romero-Calcerrada et al.,

2010; Romero-Calcerrada and Luque, 2006).

Since the CI is crucial in terms of the performance of a WofE model in mineral exploration, many studies have focused on testing the CI. For example, Bonham-Carter et al (1989) and Agterberg (1992) applied pairwise G2 and X2 to the CI test as Bonham-Carter (1994) proposed an informal rule that the sum of posterior probabilities exceeding the observed number of deposits by 15% resulted in the failure of the CI test; and Agterberg and Cheng (2002) developed an "omnibus test" (or A-C test) for CI test.

Another research area has focused on how to minimize the effect of the CI. For example, Bonham-Carter (1994) used the derived variables through PCA instead of the initial evidence layers to decrease the significance of CI in WofE modeling. Journel (2002) and Krishnan et al. (2004) put forward a new geostatistical model: Tau model, which attempted to address the restriction of CI. This has led to a number of weighted and stepwise modified models for WofE. Polyakova and Journel (2007) suggested the new Nu model as an alternate of the Tau model, which involved an extra parameter to measure the data interaction. Some of the limitations of the weights of evidence, Tau and Nu models are discussed in Schaeben (2012). Agterberg (2011) proposed a modified WofE model to estimate the weights for adjusting the dependency of evidences which applies logistic regression. The regression coefficients resulted from the logistic regression could be used as Tau weights to modify the ordinary weights of evidence. Zhang et al. (2009) proposed a similar approach to estimate the Tau weights using ordinary linear regression in association with the posterior logits resulting from weights of evidence.

Several modified WofE methods were also developed towards significant reduction of the CI's effect (Deng, 2009, 2010a, b and Cheng, 2008). A new solution to overcome the CI problem was proposed by Cheng (2015) on sequential overplay of evidences accounting the dependency of the evidences using a new model BoostWofE based on ad boosting algorithm. All above solutions for solving CI problem is based on predetermined evidences.

The approach proposed here is a type of modified WofE based on SEM concept. The evidences in WofE are considered as latent variables and undetermined before the calculation of posterior probabilities. The process of combining all evidences to a posterior probability map is implemented using weights of evidence method based on a logistic model with mineral deposits. After the construction of a SEM model in this method, the latent variables could be extracted from the original data and the model parameters could be calculated under the restriction of a target function. The target function is designed to test both of the goodness in the CI and the logistic regression. The estimation method is a type of optimum algorithm under a specified target function. In a case study, the new method was applied to construct a posterior probability map for the occurrence of mineral deposits by combining the evidences previously defined in Cheng (2008).

5.2 WofE model for mineral potential mapping

5.2.1 Mathematical model

WofE was originally developed for a non-spatial application in medical diagnosis, in which the

evidence consisted of a set of symptoms and the hypothesis was of the type "this patient has disease x". It was adapted in the late 1980s for mineral potential mapping with GIS. In this situation, the evidence consists of a set of exploration maps, and the hypothesis is "this location is favorable for occurrence of deposit type x". Weights are estimated from the measured association between known mineral occurrences and the values on the maps to be used as predictors. The hypothesis is then repeatedly evaluated for all possible locations on the map using the calculated weights, producing a mineral potential map in which the evidence from several map layers is combined (Bonham-Carter, 1994).

Assume that a series of the known binary maps are to be used prediction of mineral potential of a particular type in a particular region, and further, that the locations of a number of mineral deposits, or occurrences, are known. While the occurrences may be treated as points, the binary predictor maps can be considered as evidences. The desired end-products are output maps to show the probability of the occurrence and the associated uncertainty of the probability estimates. For complete of the explanation the following section will be reintroduced using the similar notation as used in many other papers such as Bonham-Carter (1994).

If the study area is divided into unit cells with a fixed area of u km^2, and the total area is t km^2, the total number of unit cells is T = t/u in the study area. If there are D unit cells containing an occurrence, which is equal to the number of occurrences, if u is small enough (i.e. one occurrence per cell), then the prior probability possessed by a unit cell, chosen at random for containing an occurrence, is $\mathrm{P}(D) = D/T$, and as the odds by

$$O(D) = \frac{P(D)}{1-P(D)} = \frac{D}{T-D} \qquad (5.1)$$

With the *j*-th binary predictor map (*j*=1, 2, …, *n*), the area of pattern present in terms of unit cells is $B_j = b_j/u$, wherein b_j is the area in km^2; the area $\overline{B}_j$, in which the pattern is not present, is equal to T-B_j unless some of the region is unknown with respect to the *j*-th map. The areas of over map between the known occurrences and the j-th binary pattern are $B_j \cap D$, $\overline{B}_j \cap D$, $B_j \cap \overline{D}$ and $\overline{B}_j \cap \overline{D}$. The conditional probability for choosing a cell with an occurrence, given that the cell contains pattern B_j, is

$$P(D|B_j) = \frac{B_j \cap D}{B_j} \qquad (5.2A)$$

Similarly, more conditional probabilities can be defined as follows:

$$P(\overline{D}|B_j) = \frac{B_j \cap \overline{D}}{B_j} \qquad (5.2B)$$

$$P(D|\overline{B}_j) = \frac{\overline{B}_j \cap D}{\overline{B}_j} \qquad (5.2C)$$

$$P(\overline{D}|\overline{B}_j) = \frac{\overline{B}_j \cap \overline{D}}{\overline{B}_j} \qquad (5.2D)$$

But according to Bayes' rule

$$\mathrm{P}\left(D|\mathrm{B}_j\right) = \frac{\mathrm{P}\left(\mathrm{B}_j|D\right)P(D)}{P\left(\mathrm{B}_j\right)} \quad (5.3\mathrm{A})$$

$$\mathrm{P}\left(D|\overline{\mathrm{B}}_j\right) = \frac{\mathrm{P}\left(\overline{\mathrm{B}}_j|D\right)P(D)}{P\left(\overline{\mathrm{B}}_j\right)} \quad (5.3\mathrm{B})$$

So, if the weights for pattern j are defined as

$$\log_e O\left(D|\mathrm{B}_j\right) = W_j^+ + \log_e O\left(D\right) \quad (5.4\mathrm{A})$$

$$\log_e O\left(D|\overline{\mathrm{B}}_j\right) = W_j^- + \log_e O\left(D\right) \quad (5.4\mathrm{B})$$

the positive weight for the presence of B_j is:

$$W_j^+ = ln\frac{P(B_j|D)}{P(B_j|\overline{D})} \quad (5.5\mathrm{A})$$

And the negative weight for $\overline{\mathrm{B}_J}$ is:

$$W_j^- = ln\frac{P(\overline{B_j}|D)}{P(\overline{B_j}|\overline{D})} \quad (5.5\mathrm{B})$$

The approximate variances of the weights can be obtained from:

$$\sigma^2\left(W_j^+\right) = \frac{1}{n(B_jD)} + \frac{1}{n(B_j\overline{D})} \quad (5.6\mathrm{A})$$

$$\sigma^2(W_j^-) = \frac{1}{n(\overline{B_j}D)} + \frac{1}{n(\overline{B_j}\overline{D})} \quad (5.6B)$$

wherein $n(B_jD)$, $n(B_j\overline{D})$, $n(\overline{B_j}D)$, $n(\overline{B_j}\overline{D})$ stand for the area of B_j and $\overline{B_j}$ in D and $\overline{D}$, respectively. Let have two binary predictor patterns, B_j, j = 1,2, from probability theory, one has

$$P(D|B_1B_2) = P(B_2|DB_1)P(B_1|D)P(D) \quad (5.7)$$

If B_1 and B_2 are conditionally independent with respect to the mineral occurrence points, then:

$$P(B_2|DB_1) = P(B_2|D) \quad (5.8)$$

Thus,

$$P(D|B_1B_2) = P(B_2|D)P(B_1|D)P(D) \quad (5.9)$$

It can readily give:

$$\log_e O(D|B_1B_2) = W_1^+ + W_2^+ + \log_e O(D) \quad (5.10A)$$

$$\log_e O(D|B_1\overline{B}_2) = W_1^+ + W_2^- + \log_e O(D) \quad (5.10B)$$

$$\log_e O\,(D|\overline{\mathrm{B}}_1\mathrm{B}_2) = W_1^- + W_2^+ + \log_e O\,(D) \quad (5.10C)$$

$$\log_e O\,(D|\overline{\mathrm{B}}_1\overline{\mathrm{B}}_2) = W_1^- + W_2^- + \log_e O\,(D) \quad (5.10D)$$

Similarly, if more binary predictor maps are used, they can be added provided that they are also conditionally independent with respect to the mineral occurrence points. In general, with the binary predictor maps, B_j, j= 1, 2, …, n , the log posterior odds are:

$$\log_e O\left(D|B_1^k \cap B_2^k \cap B_3^k \ldots B_n^k\right) = \sum_{j=1}^{n} W_j^k + \log_e O\,(D) \quad (5.11)$$

wherein the superscript k refers to the presence or the absence of the binary pattern and

$$W_j^k = \begin{cases} W_j^+ & for\ j-th\ pattern\ present \\ W_j^- & for\ j-th\ pattern\ absent \\ 0 & for\ no\ data \end{cases} \quad (5.12)$$

The posterior probability is calculated as

$$P = O/(1+O) \quad (5.13)$$

For each predicted map, the contrast

$$\mathrm{C} = \mathrm{W}^+ - \mathrm{W}^- \quad (5.14)$$

gives the useful measure of the correlation with the mineral occurrence points. The weights W^+ and W^- have the opposite signs, except that both become zero and C becomes zero, if a map pattern has a distribution spatially independent of the points. It is a convenient measure for the strength of the spatial correlation between a point pattern and the map layer (Agterberg, 1989; Bonham-Carter et al., 1988). For a positive spatial association, C will have positive values while C would take on negative values in a similar range for a negative association. Except in the special case of C = 0, W^+ will always be of opposite sign with that of W^- (Bonham-Carter, 1994). The variance of the contrast C is:

$$\sigma^2(C) = \frac{1}{n(B_j D)} + \frac{1}{n(B_j \bar{D})} + \frac{1}{n(\bar{B_j} D)} + \frac{1}{n(\bar{B_j} \bar{D})} \quad (5.15)$$

Besides the value of contrast C, a "studentized" C is calculated, as the ratio of contrast C and its standard deviation $\sigma(C)$, to test the hypothesis that C=0 in the following sections (Bonham-Carter et al., 1989). A value greater than 1.96 indicated that the hypothesis can be rejected at $\alpha = 0.05$.

5.2.2 Issue under the conditional independence in WofE

The CI is a strong assumption in the WofE models for mineral prediction, whereby all predicator patterns become conditionally uncorrelated. If D presents the occurrence of deposits, B_1 is the first pattern and B_2 is the second one, the events B_1 and B_2 are independent under the condition of the event D if

$$P(B_1 \cap B_2|D) = P(B_1|D)P(B_2|D) \qquad (5.16)$$

$$P(B_1 \cap B_2|\overline{D}) = P(B_1|\overline{D})P(B_2|\overline{D}) \qquad (5.17)$$

wherein, $P(B_1 \cap B_2|D)$ is the occurrence probability of event B_1 and B_2 given that event D has occurred, $P(B_1|D)$ are the conditional probability of event B_1 given that event D has occurred and $P(B_2|D)$ are the conditional probability of event B_2 given that event D has occurred. This condition is strong CI which can be weakened by letting the ratios of the above equations are identical (Cheng 2008). More discussion about the conditional independence in the logistic model can be found in Journel (2002).

There have been several methods developed for testing of the CI of patterns, which included Contingency Table Tests (Bishop et al., 1975), Overall or "Omnibus" Test and new "Omnibus" Test (Agterberg and Cheng, 2002; Kemp et al., 1999) and so on. The current research adopted an overall test method to estimate the independence of evidences. Other methods for reducing the effect of CI are refereed to Cheng (2015).

The end product of WofE modeling is a posterior probability map. If p binary patterns are considered and there are no data missing, the unit cells with the same posterior probability form classes that belong to one of the 2^p possible "unique" conditions. Let us assume that T represents the sum of the posterior probabilities for all unit cells in the study area, ideally should be equal to the total number of deposits n. However, in practice, T may exceed n and one can assume that T > n due to lack of conditional independence of map layers. This is the rationale of the

overall or the so-called “omnibus test” for the conditional independence. However, it is argued that T should not exceed n by more than 15% (Bonham-Carter, 1994), for example.

For most of the applications in mineral exploration, the input evidences are usually hard to meet the CI requirement. But the CI effect can be reduced by many methods including the weighted weights of evidence (Agterberg, 2011) and the boost weights of evidence (Cheng, 2015). Three methods were mainly proposed in the literature for reducing the CI in weights of evidence (Cheng 2008; 2015): i). redefining evidence by combining evidences such as using PCA to form new evidences which improve the CI; ii). Correcting the results of WofE generated with CI effect; and iii). Modifying the weights of evidence so that they are not affected by CI even if it is exist. Here, the author proposes to modify the definition of inputting evidences so that they may show less CI effect on the final results, which falls in the first group and was realized differently. For example, for a given pattern, multiple binary maps can be constructed with different cut-off values. Theoretically all of these binary maps can be used as evidences to calculate a posterior probability maps. The key is how to decide which binary map should be used. To demonstrate this idea a simple situation is discussed. For example, ***Fig** 5.1* shows an input re-classified map, which can be transformed into binary patterns by different cut-off values, of which two of them will be selected as the inputs for WofE evidence. The gold deposits are shown as stars in maps. The comparison of the estimated number of the mineral deposits with the observed number of the deposits may reveal the CI effect in WofE method.

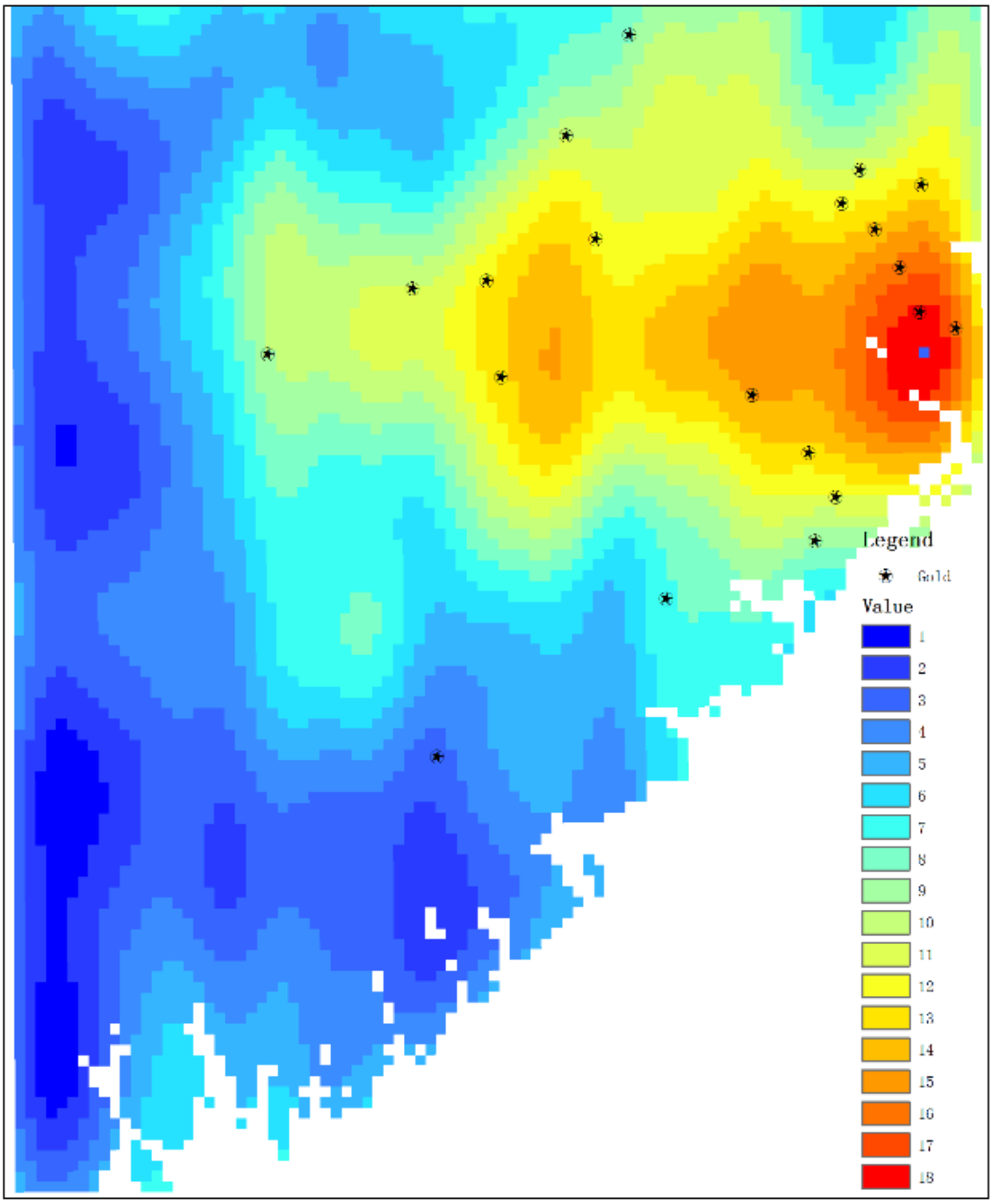


***Fig** 5.1 A re-classified layer with 18 values.*

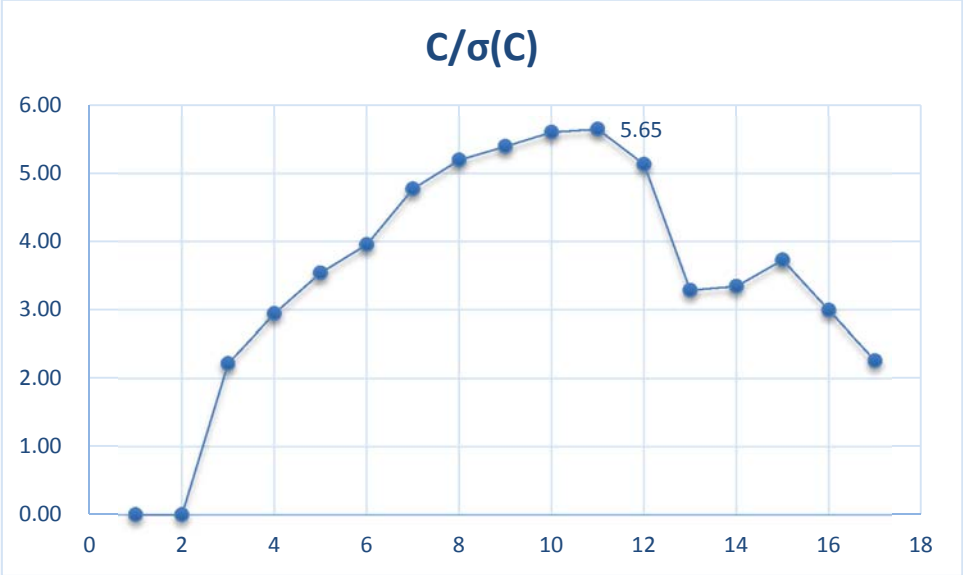


***Fig** 5.2 T-test value(C/ σ (C)) of evidences in Fig 5.1A, X-Axis is cut-off value, Y-Axis is the corresponding t-test value. The maximum t-test value is 5.65, which cut-off value is 11.*

***Table** 5.1 Estimated deposits number (T) under different combinations and the ratio of T-N over N, N is the observed deposits number.*

Group	**9_13**	**10_13**	**11_13**	**9_12**	**10_12**	**12_13**	**13_13**	**9_11**
T	30.08	30.97	32.68	33.37	34.58	34.58	34.93	36.52
(T-N)/N	0.50	0.55	0.63	0.67	0.73	0.73	0.75	0.83
Group	**11_12**	**10_11**	**9_10**	**12_12**	**11_11**	**10_10**	**9_9**	
T	36.87	38.06	39.30	39.45	40.97	41.13	42.69	
(T-N)/N	0.84	0.90	0.97	0.97	1.05	1.06	1.13	

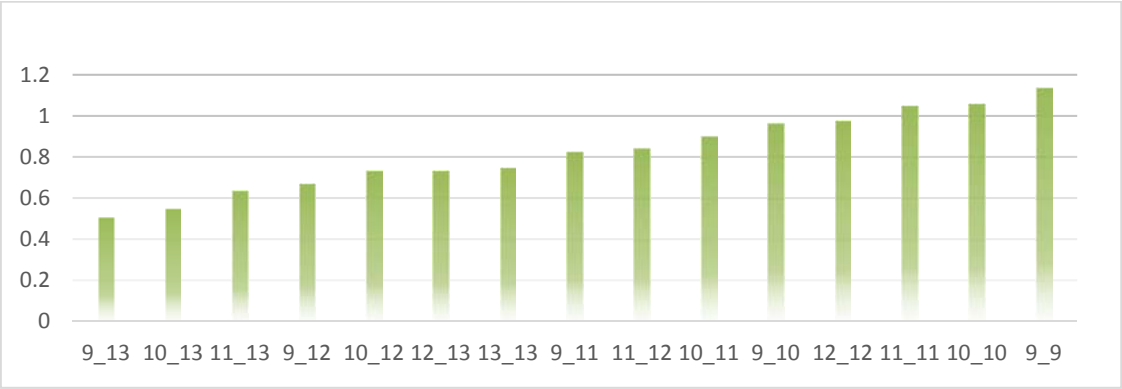


***Fig** 5.3 The relation between the number of estimated deposits and the combination index.*

The traditional WofE method usually adopts the pattern with the highest t-value as the input evidence. There are 5 patterns selected (***Fig*** *5.1*) for input evidences, which have the *t*-test values around the highest one (***Fig*** *5.2*), with the corresponding cut-off value from 9 to 13. If two of them are selected as the input evidences in the WofE calculation, the index of combinations is constructed by the corresponding cut-off values, i.e. the combination with index 9_9 means that the input are two patterns which cut-off values are 9 and 9, respectively. The number of possible combinations is $C_5^2 = 15$. The ratio of estimated number of the deposits (T) and the observed number (N), ((T-N)/N), calculated through different combinations, are shown in ***Table*** *5.1*. The combinations are sorted by (T-N)/N (***Fig*** *5.3*) in the ascend order. The results in ***Fig*** *5.3* indicate that the estimation number and the ratio of the over estimation get the minimum value at the index of 9_13 and the maximum value at the index of 9_9. The pattern index is from 9 to 13. Therefore, 9_13 is the combination with the most of the different patterns and the 9_9 is the combination with the least of the different patterns. This implies that the over estimation can be reduced through changing the difference of input patterns even these patterns represent the same geo-factor. The minimum ratio of the over estimation in ***Table*** *5.1* is over 50%, over the acceptable level 15% (Bonham-Carter, 1994). Because all of these patterns come from the same re-classified data, the results showed that all cases were unacceptable. However, the fact that the change of the binary pattern combinations has greatly affected the result demonstrates that the formation of new binary patterns through pattern modification might possibly reduce the overall value of (T-N)/N.

5.3 The SEM based WofE

5.3.1 SEM construction for WofE

The result in Section 5.2.2 indicated that the ratio of the over estimation can be reduced through modifying the combination of patterns. This provides a new method to reduce the CI effect. The next task is how to find such a combination which can satisfy the CI requirement as much as possible. This section will focus on how to construct a SEM model for solving the above problem.

The measurement model and the structural model need to be initiated before the SEM construction, which can be defined from the concept of traditional WofE method. A traditional WofE model in mineral exploration is shown in ***Fig** 5.4*, in which, the individual evidences can be extracted from the reclassified data through the t-test method and then used to calculate posterior occurrence probability of mineral deposits through Bayes' rule. Since the evidences are extracted from original data rather than observed directly, they can be considered as latent variables and the corresponding process can be considered as measurement model.

Furthermore, the process of calculating the posterior probability map can be considered as a structural model because it is based on a logistic model between the input evidences and the occurrence probability of mineral deposits. In such a structural model, the independent variables are the evidences generated from measurement model, and the dependent variable is the occurrence probability of mineral deposits in different zones. For example, the study area

is divided into different zones (i.e. S_1, S_2, …, S_i) by different estimated posterior probabilities, a zone is the unit with the same posterior probability. Also, the probability of deposits occurrence can be calculated directly through N_i/A_i, where N_i is the observed deposits number in S_i, and A_i is unit number of S_i. A regression can be created between the evidences and the observed probabilities. The new WofE model is shown in ***Fig*** *5.5*.

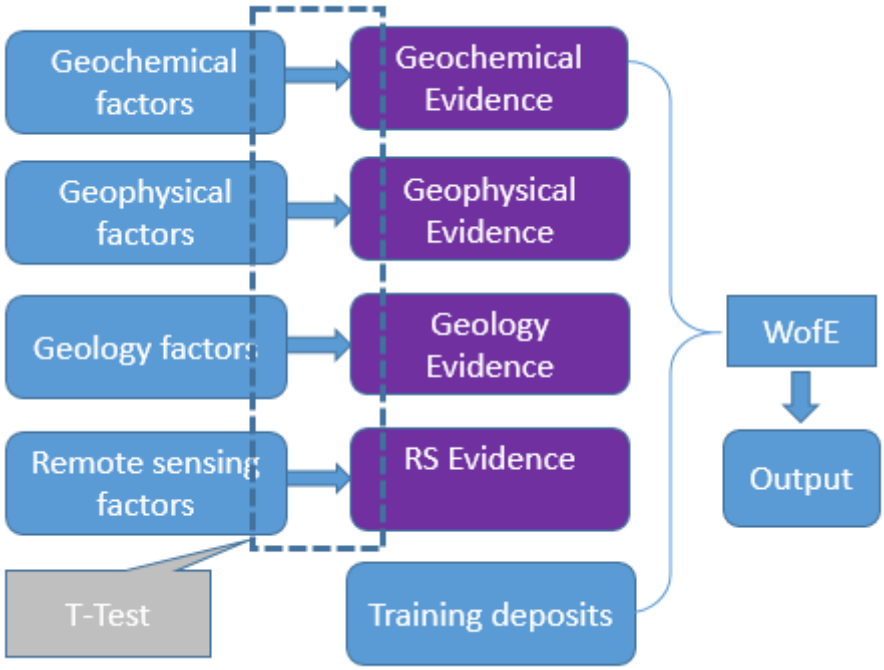


Fig *5.4 The traditional WofE calculation process in geo-information integration for mineral exploration. The blue rectangle represents observed variables and the purple rectangle represents evidences for WofE calculation.*

There are TWO main differences between the new and traditional WofE approaches:

1) *Besides the locations of the mineral deposits, the evidences in traditional WofE are determined by the related data sources, while the evidences in new WofE method are*

determined by the corresponding data sources and the other evidences.

2) *The evidences in traditional WofE method are calculated individually, while they are calculated together in the new WofE method.*

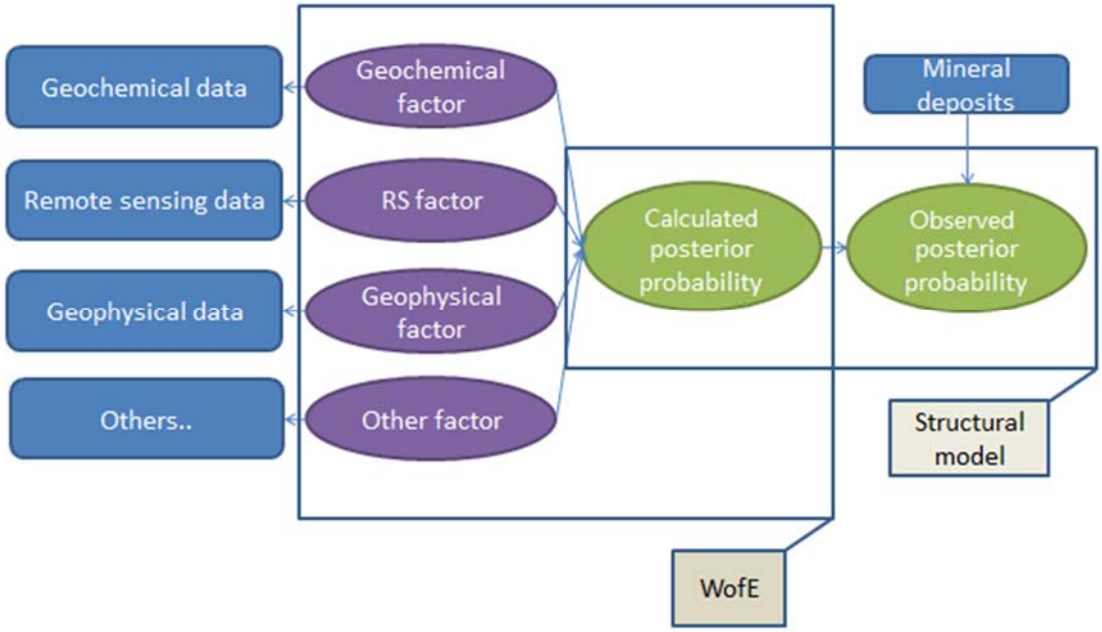


***Fig** 5.5 SEM of WofE method in mineral exploration. Blue rectangle represents observed variables; purple ellipse represents exogenous latent variables, green ellipse represents endogenous latent variable.*

5.3.2 Target function

The target function is designed to ensure the following conditions:

1) *The extracted evidences are as independent as possible from each other under the condition*

of occurrence of deposits.

2) *The distribution of the calculated posterior probability is as close to the spatial distribution of the observed deposits as possible*.

The first condition ensures that the predicted number and the observed number of deposits should be as close as possible. About the second one, a regression between the predicted and the observed probabilities in different posterior probability zones is adopted. In this way, the study area is divided into a maximum of 2^n zones, wherein n is the number of the input evidences. The number of deposits in each zone is calculated through an accumulated posterior probability which can be compared with the observed number of deposits. A regression can be constructed between the observed number and the estimated number of deposits. The best estimation can be determined in terms of the maximum R^2 of the regression.

Based the above principles, the target function is defined as:

$$F = \frac{(T-N)^2/min(T-N)^2}{R^2/max(R^2)} \qquad (5.18)$$

wherein T is the number of the predicted deposits, N is the number of the observed deposits, R is their correlation coefficient in each zone, $min(T-N)^2$ means the best prediction in deposits, and $max(R^2)$ means the best prediction in correlation with observations. The $(T-N)^2/min(T-N)^2$ and $R^2/max(R^2)$ are the standardizations for $(T-N)^2$ and R^2, respectively. The target function would reach a balance between the criteria of predicted number of the deposits and the predicted correlation with observations.

5.3.3 Parameter estimation

For the current model, the inputs are the reclassified data (grid map) and deposits. The solutions of the parameters include the cut-off values of input patterns and the posterior probability map of mineral occurrence. For example, there are four evidences (latent variables), four reclassified layers (independent observed variables) and one mineral deposit layer (dependent observed variables) shown in ***Fig*** *5.5*. After the calculation, the model should output the posterior probability map and four corresponding cut-off values adopted in each pattern.

The algorithm for parameter estimation is the same as the one introduced in Chapter 4, which performs an optimum calculation based on a target function through adjusting the input evidence. The algorithm is programed in *R* language and implemented aided by MapWinGIS Dynamic Link Library (DLL).

5.4 Case study

5.4.1 Reclassified geo-data used in case study

There are four evidences adopted in this case study. Besides the layer shown in ***Fig*** *5.1*, there are three other layers shown in ***Figs*** *5.6A, 5.6B* and *5.6C*. The t-test values of these layers with the different cut-off values are shown in ***Fig*** *5.7*, correspondently. ***Fig*** *5.6A* is a multiple ring buffer result of the contact boundary between Halifax formation and Goldenville formation, which ring's distance is 1500 meters. ***Fig*** *5.6C* is a multiple ring buffer result of the fold axis

in study area, which ring's distance is 1500 meters too. ***Fig** 5.1* and ***Fig**.5.6B* are the background and anomalies of geochemical field in study area, respectively. They are the Fourier decomposition of the first principal component of 16 geochemical elements. The basic geological background was introduced in Chapter 3, refer to Cheng (2008) for more details about these evidences.

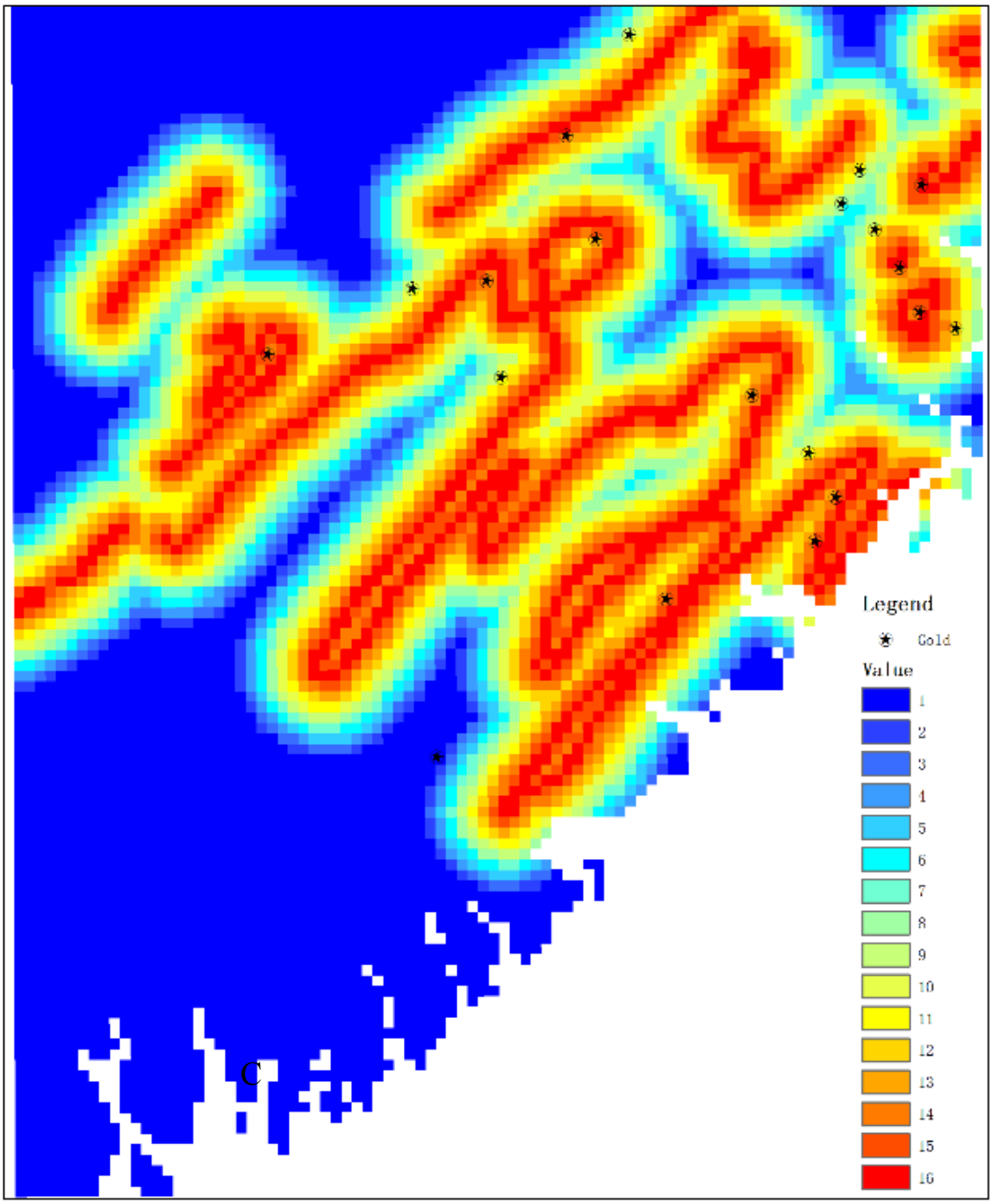


***Fig** 5.6A Evidence used in WofE calculation: Halifax formation and Gordenvill formation boundary buffer, each ring distance is 1.5 KM.*

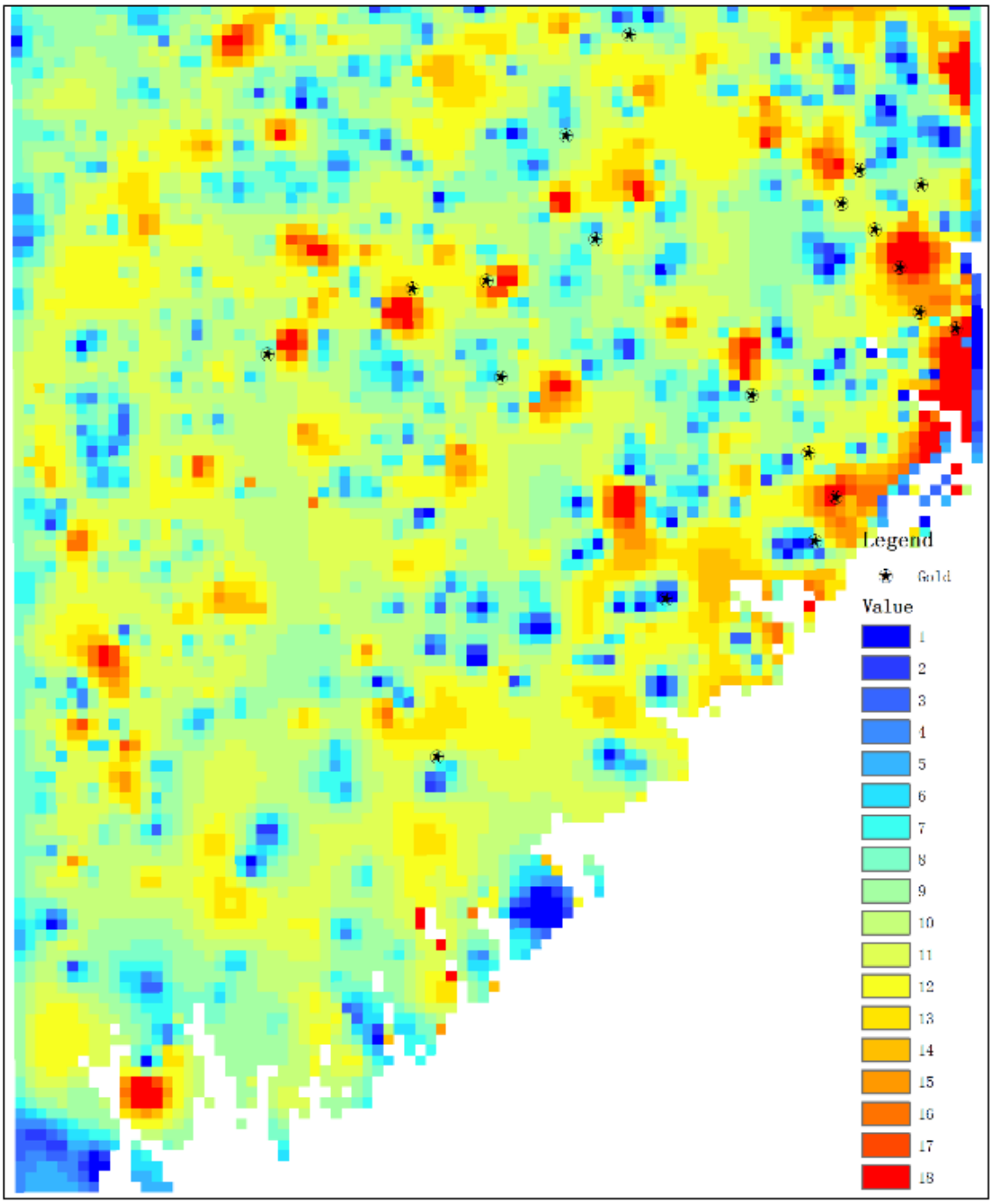


***Fig** 5.6B Evidence used in WofE calculation: anomalies which extracted from the first component of 16 geochemical elements by Fourier method.*

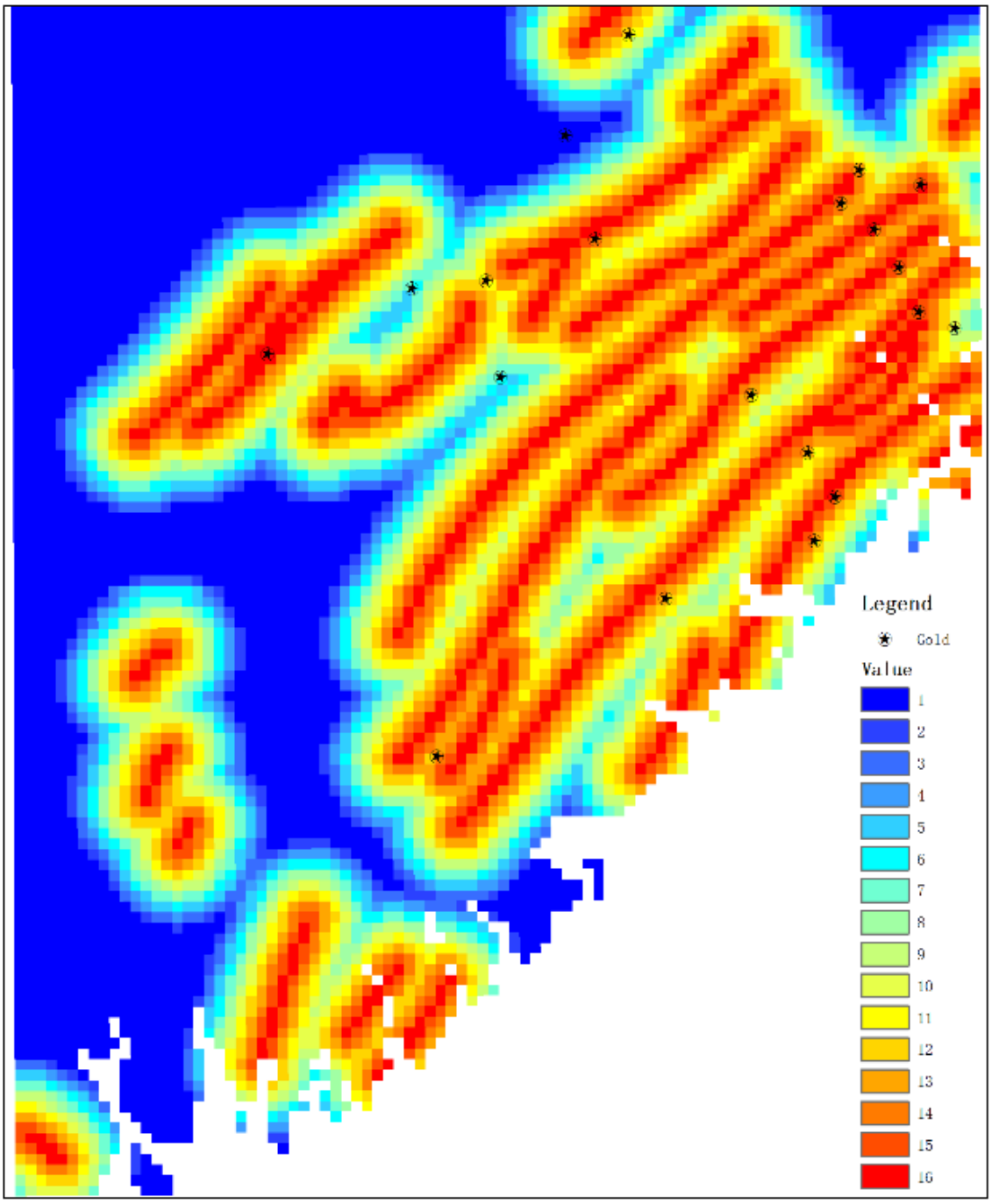


***Fig** 5.6C Evidence used in WofE calculation: NE-SW direction fold buffer, each buffer distance is 1.5 KM.*

5.4.2 T-value of evidences and the input patterns

The t-values obtained between the four layers of the evidences and the deposits are shown in ***Fig*** *5.7*. The highest values in each of the four maps (***Fig*** *5.1A,* ***Figs*** *5.6A, 5.6B* and *5.6C*) is located at cut-off value 11 (***Fig*** *5.7A*), 9 (***Fig*** *5.7B*), 15 (***Fig*** *5.7C*) and 11(***Fig*** *5.7D*), respectively. Assume that 5 cut-off values closed to the optimal t-values can be acceptable. The cut-off value ranges of these four evidences could be 8-12, 7-11, 13-17 and 9-13, respectively. In the new method, one from each group will be adopted as the cut-off value for transforming the reclassified input map to a binary pattern as the evidence. Through this way, the measurement model can be constructed.

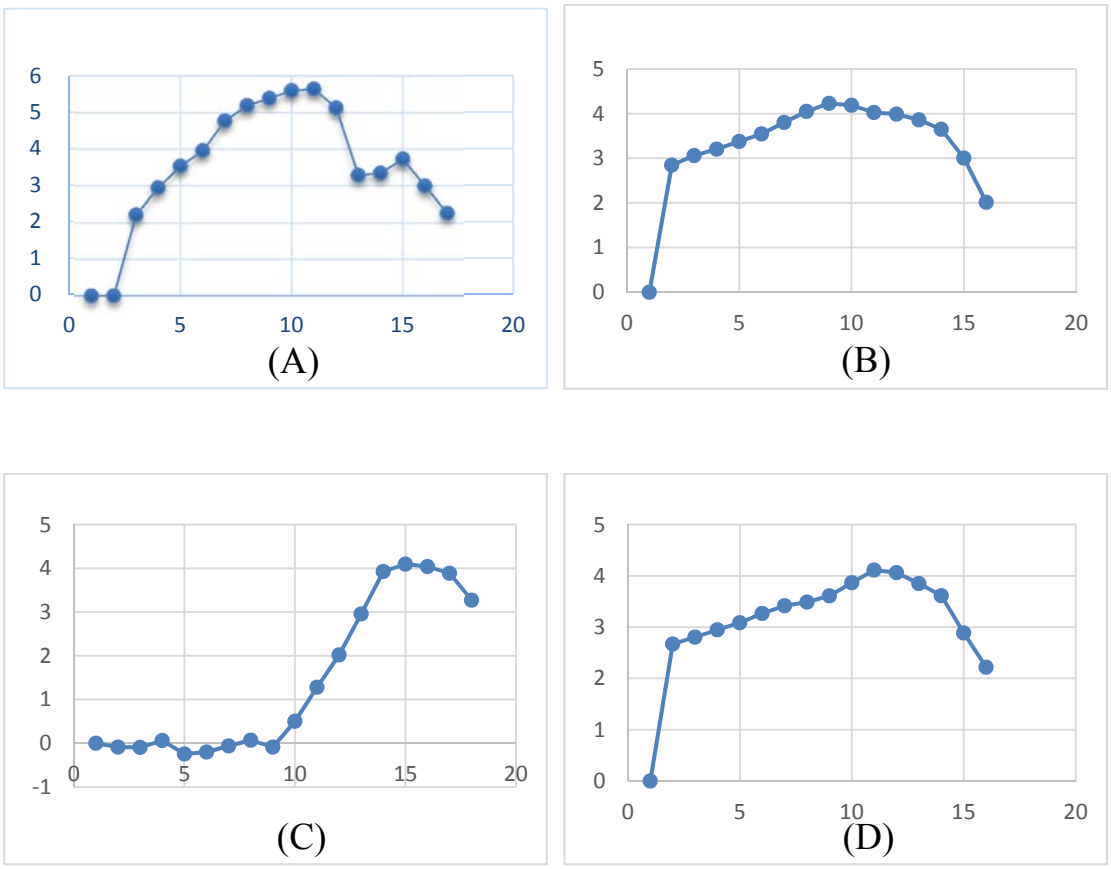


***Fig** 5.7 T-test values of the evidences, A: t-test value list of map 1(Fig 5.1A); B: t-test value list of map 2(Figs 5.6A); C: t-test value list of map 3(Figs 5.6B); D: t-test value list of map 4(Figs 5.6C).*

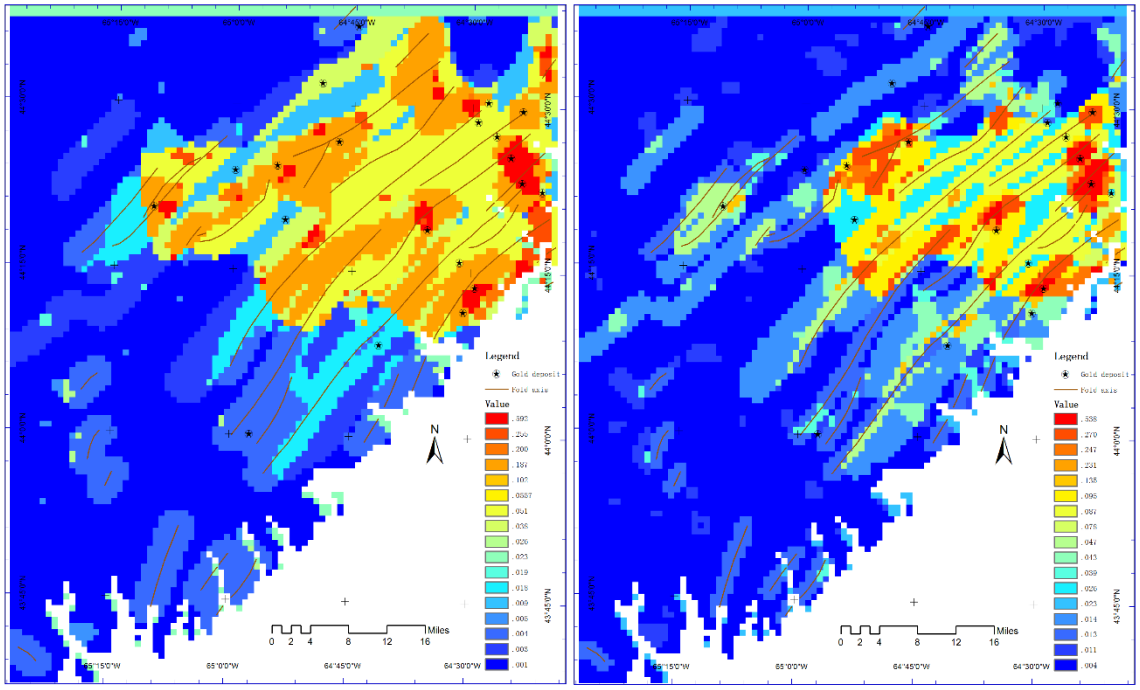
5.4.3 Results and interpretation

In order to validate the new method, two posterior probability maps of mineral occurrence are estimated through the original WofE method and the new WofE method. The cut-off value adopted in the traditional and the new WofE methods for the reclassified input maps are (12, 9, 15, 11) and (13, 12, 13, 14), respectively. The posterior probability maps obtained by using the

two sets of binary maps are shown in ***Figs*** *5.8A* and *5.8B*. The weights, the contrasts and their standard deviations for each evidence using the traditional and the new methods are shown in ***Tables*** *5.2* and *5.3*. The predicted number of the deposits (T) from original and the new WofE methods give T/N = 1.8, 1.5, respectively. These results indicate that the four new binary patterns defined by the new method generated an improved result than the result delivered by the original method, although both of the cases still showed the strong effect of the CI which needs to be corrected using other methods.

Table *5.2 Weights, contrasts and their standard deviations for predictor maps in Fig 5.8A*

	W+	σ.W+)	W-	σ.W-)	C	σ(C)	C/σ(C)
Evidence1	0.32	0.16	-0.36	0.02	0.67	0.17	4.03
Evidence2	0.57	0.15	-0.30	0.03	0.87	0.15	5.65
Evidence3	0.22	0.15	-0.10	0.03	0.31	0.16	2.02
Evidence4	0.33	0.15	-0.26	0.03	0.60	0.16	3.86



(A) **(B)**

Fig *5.8 The posterior probability map of deposits occurrence estimated from (A) traditional WofE and (B) new method, the cut-off value used for evidences are (12,9,15,11) and (13,12,13,14), respectively.*

Table *5.3 Weights, contrasts and their standard deviations for predictor maps in Fig 5.8B*

	W^+	$\sigma(W^+)$	W^-	$\sigma(W^-)$	C	$\sigma(C)$	$C/\sigma(C)$
Evidence 1	0.30	0.22	-0.61	0.02	0.92	0.22	4.23
Evidence 2	0.57	0.15	-0.30	0.03	0.87	0.15	5.65
Evidence 3	0.72	0.19	-0.08	0.06	0.80	0.20	4.10
Evidence 4	0.30	0.18	-0.47	0.02	0.76	0.19	4.11

Tables *5.4* and *5.5* show the statistical results of the posterior probability maps in ***Figs*** *5.8A and 5.8B,* respectively. The first column ("PP") gives the posterior probability values in the individual zones. The number of units with the same posterior probability is given in second column ("Area"). In the third column ("AA"), the accumulated unit numbers are ordered from high posterior probability to low while the predicted number of deposits in each zone is listed in the fourth column ("Pre"). Here each deposit is represented by $3 \times 3 = 9$ units. Therefore, the deposit number for each zone can be calculated as $\mathrm{PP} \times \mathrm{AA} / 9$, wherein PP is the posterior probability and AA is the accumulated unit number. For example, in the first row of ***Table*** *5.4*, the PP value is 0.6483, the number of unit AA is 99. So, the predicted deposits number is given by $0.65 \times 99/9 \approx 7.13$. The very last three columns list the observed number of deposits in each zone ("Obv"), the accumulated number of predicted deposits ("A_Pre") and the accumulated observed ("A_Obs"), respectively.

The regressions between the accumulated numbers of the deposits predicted and observed in different zones from two methods are shown in ***Fig*** *5.9*. The R^2 is 0.90 and 0.93, respectively. The R^2 from the new method is larger than the traditional one, which means the prediction from the former better than the latter in terms of the spatial correlation with the deposits in study area.

The previous result indicated that the prediction based on the new patterns obtained by the new WofE was better than that by the traditional method in this case study both in the prediction of the deposit number and the spatial correlation ship with the observed deposits.

The ***Fig*** *5.10A*, *B* and C give the frequencies of R^2 (regression between the predicted deposits

and the observed deposits), number of the predicted deposits and the value of target function calculated through different input evidence combinations. ***Fig*** *5.11* is the map between the above number of the predicted deposits and the R^2.

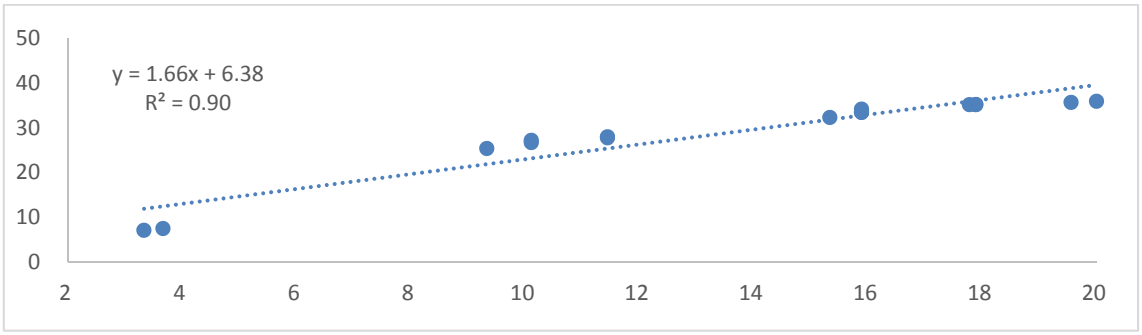
The frequency histograms with the random samples in different inputs are plotted here to characterize the stable optimum calculation. In ***Fig*** *5.10*, the distribution of the target function values (***Fig*** *5.10C*) looks very likely normal than others (***Fig*** *5.10A* and *5.10B*), which implies that the results based on the target function is more stable than other two indexes. The asymmetrical distribution of ***Fig*** *5.10C* means the calculation may be convergence with a smaller sampling number than other two histograms. Moreover, ***Fig*** *5.11* shows that the good predictions for the deposit numbers (closed to the number of the observed deposits, whose points were located around x-axis) do not always have the good R^2 (closed to 1, whose points were located away y-axis), and the correlation between the R^2 and the prediction of the deposits is very weak (0.01). Therefore, it is necessary for the designed target function to be constrained by these two criteria.

***Table** 5.4 The statistical result of posterio probability map in Fig 5.8A*

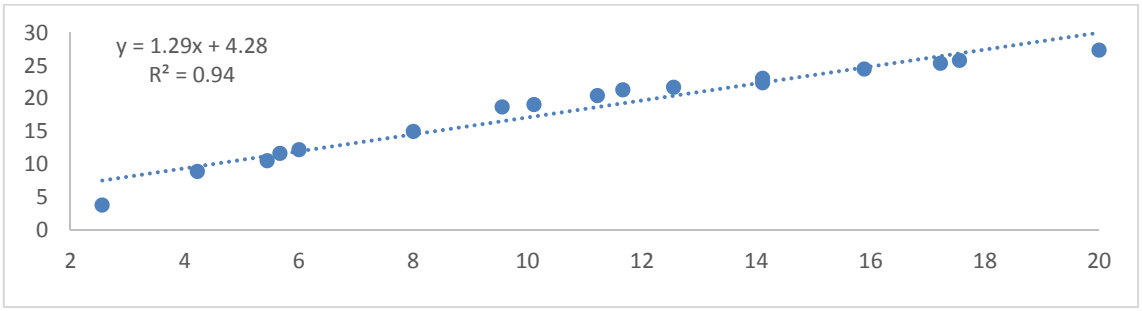
PP	Area	AA	Pre	Obs	A_Pre	A_Obs
0.6483	99	99	7.13	3.33	7.13	3.33
0.2405	14	113	0.37	0.33	7.51	3.67
0.2254	714	827	17.88	5.67	25.38	9.33
0.1988	60	887	1.33	0.78	26.71	10.11
0.1824	24	911	0.49	0	27.2	10.11
0.0476	108	1019	0.57	1.33	27.77	11.44
0.0409	889	1908	0.22	0	27.98	11.44
0.0377	48	1956	4.33	3.89	32.32	15.33
0.034	1035	2991	1.11	0.56	33.43	15.89
0.0292	294	3285	0.08	0	33.51	15.89
0.023	24	3309	0.68	0	34.19	15.89
0.0067	265	3574	0.97	1.89	35.16	17.78
0.006	1312	4886	0.01	0.11	35.17	17.89
0.0051	9	4895	0.02	0	35.19	17.89
0.0047	41	4936	0.47	1.67	35.66	19.56
0.0008	2884	7820	0.26	0.44	35.92	20

***Table** 5.5 The statistical result of posterio probability map in Fig 5.8B*

PP	Area	AA	Pre	Obs	A_Pre	A_Obs
0.5376	64	64	3.82	2.56	3.82	2.56
0.2699	170	234	5.10	1.67	8.92	4.22
0.2474	59	293	1.62	1.22	10.54	5.44
0.2309	44	337	1.13	0.22	11.67	5.67
0.1353	36	373	0.54	0.33	12.21	6.00
0.0946	265	638	2.79	2.00	15.00	8.00
0.0871	383	1021	3.71	1.56	18.71	9.56
0.0782	43	1064	0.37	0.56	19.08	10.11
0.0474	257	1321	1.35	1.11	20.44	11.22
0.0424	184	1505	0.87	0.44	21.30	11.67
0.0388	94	1599	0.41	0.89	21.71	12.56
0.0263	234	1833	0.68	1.56	22.39	14.11
0.0230	265	2098	0.68	0.00	23.07	14.11
0.0139	915	3013	1.41	1.78	24.48	15.89
0.0127	611	3624	0.86	1.33	25.34	17.22
0.0113	358	3982	0.45	0.33	25.79	17.56
0.0036	3838	7820	1.54	2.44	27.33	20.00



(A)



(B)

Fig *5.9 The regression between the predicted deposits and observed deposits. The x is accumulated observed deposits number (A_Obs in Table 5.4 and 5.5) and y is accumulated predicted (A_Pre in Table 5.4 and 5.5).A: the regression for traditional WofE method; B: the regression for new method.*

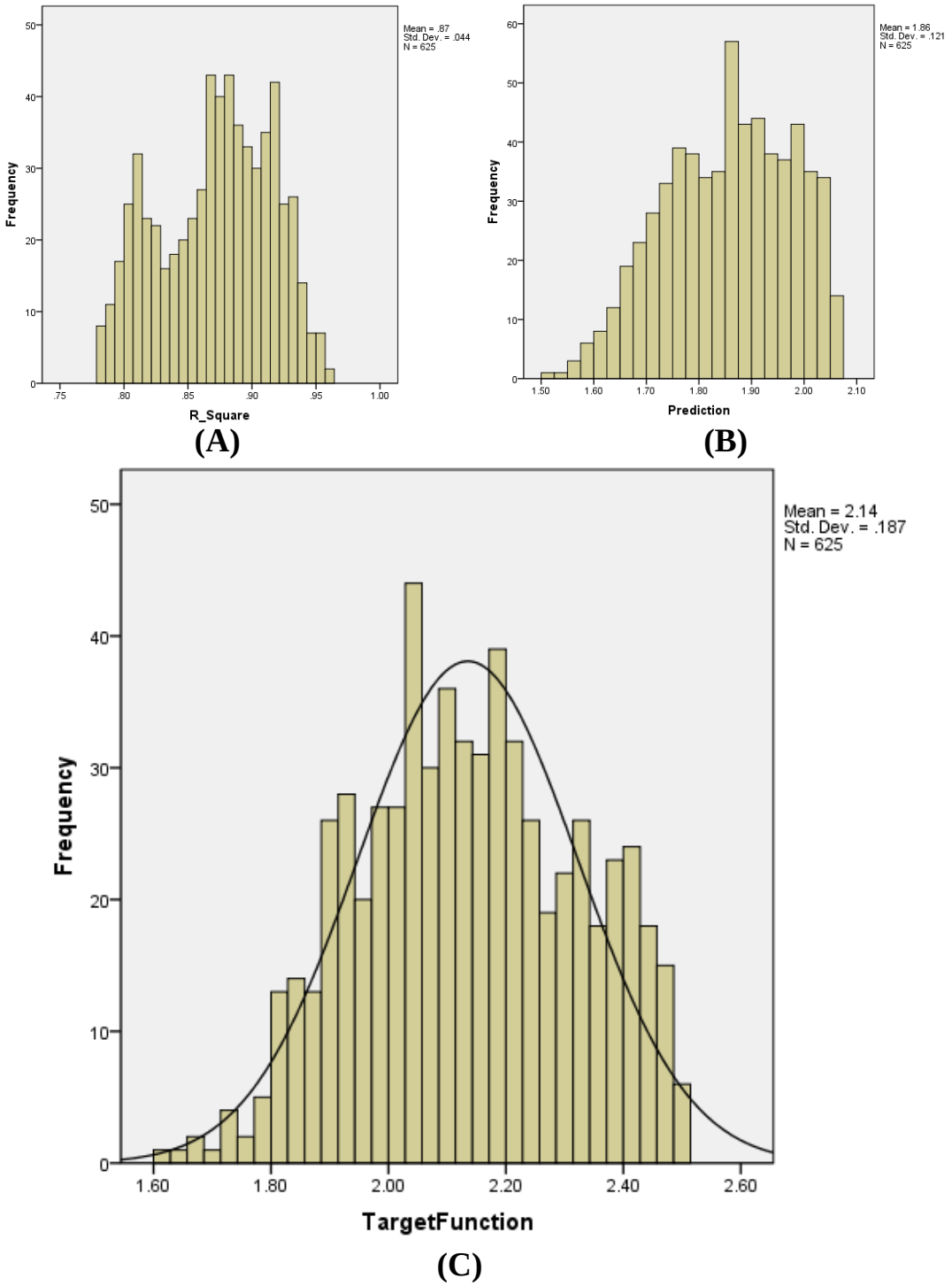


***Fig** 5.10 A: the frequency of R square, B: the frequency of over estimation ratio (T/N), C: the frequency of target function value, which calculated through different evidence combinations.*

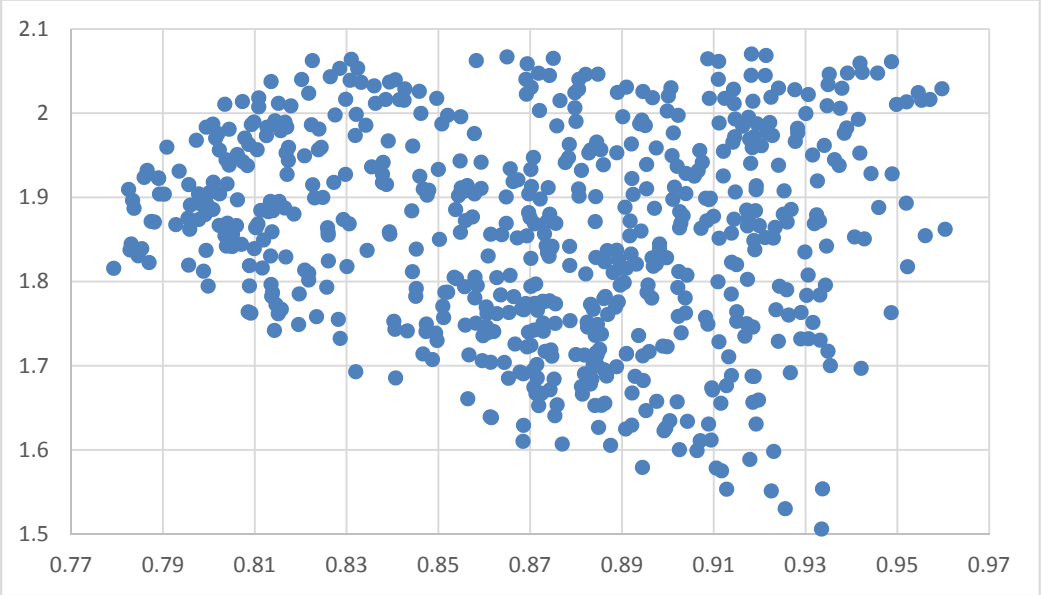


***Fig** 5.11 The relationship between the over estimation ratio (T/N) and R square. Horizontal axis: R square of regression between cumulated predicted and observed deposits number in different zones. Vertical axis: over estimation ratio T/N, where T and N are the numbers of predicted and observed deposits, respectively.*

5.5 Discussion and conclusions

This Chapter extended the application of the new SEM technique in WofE method to reduce the effect of the CI in mineral exploration. In this way, the traditional WofE method (***Fig** 5.4*) has been transformed from a step-by-step calculation to a globe assessment restricted by a target function. In the traditional WofE method, each process of evidence extraction is independent of others so that the extracted evidences may hardly pass the CI test, but the processes in the new WofE method are calculated under the same target function control, the goodness of the extracted evidences is determined by both the correlation with the deposits and with other evidences, rather than the correlation with mineral in original method. The traditional WofE can ensure each evidence used in calculation is the most correlated one (with the highest t-test

value) with the existed mineral deposits among its group. However, the new method would find out an evidence which can cooperate with other ones and the combination of the evidences has a better prediction than other combinations. In other words, similar to a game with cards, the combination of the best players is not necessarily the best combination to win. The principle of the new method is to find the best combination rather than the best players in individual groups.

The case study, which included four evidences, indicated that the overestimation ratio in the new WofE method is decreased 37.5% from the traditional WofE method. The prediction number from the traditional WofE and the new WofE are 1.8 and 1.5 times of observed one, respectively. At the same time, the goodness of fit between the calculated posterior probability and the observed posterior probability are 0.90 and 0.93 for the traditional WofE and the new WofE, respectively.

It should be noted that the new method cannot solve the CI problem exhaustively. For example, the over estimation ratio was reduced from 80% to 50% in current research, but it still could not pass the CI assumption test. Therefore, it could be better if the new method was used with other weighted WofE method together when the extracted patters through new method cannot fully match the CI requirement. This would be the future work of the proposed method.

Chapter 6 A constrained geochemical variable classification method based on conditional correlation coefficient

6.1 Introduction

As an important data analysis method, clustering is broadly used in geochemical data processing (Castillo-Muñoz and Howarth, 1976; Gustavsson and Bjorklund, 1976; Hanesch et al., 2001; Howarth, 1973; Ji et al., 2007; Kramar, 1995; Rantitsch, 2000; Templ et al., 2008; Vriend et al., 1988; Xie et al., 2004). The principal aim of this method is to split multivariate observations into a number of meaningful, multivariate, and homogeneous groups. Several procedures for variable clustering, mainly based on similarity (or dissimilarity) of measurements between variables, have been developed by researchers (Qannari et al., 1997; Qannari et al., 1998; etc.). Moreover, the specific clustering methods based on PCA such as the clustering around latent variables (CLV) (Vigneau and Qannari, 2003) and the diametrical clustering (Dhillon et al., 2003) have been proposed. Additional classification methods based on sparse PCA (Jolliffe et al., 2003; Zou et al., 2006), sparse partial least square (PLS) (Chun and Keleş, 2010; Lê Cao et al., 2008), mixture models using factor analysis (Subedi et al., 2013), and CLV under the constraint of a specified variable (Chen and Vigneau, 2014) were also discussed.

Geochemical element clustering is usually used for analyzing the relationship between geochemical samples or elements to establish multivariate geochemical background patterns. Element clustering is also used for pollution identification (Hanesch et al., 2001) and for

investigation of relationships between regional geochemical patterns and ore deposits (Xie et al., 2004). The number of groups and the membership of the samples/elements are usually obtained by analyzing the association between the samples/elements, which are measured as some types of distances. The groups and their sample memberships can then be used in geological interpretation. Problems and possibilities in data transformation to use such methods have been reviewed by Templ et al. (2008).

In this Chapter, a new constrained variable clustering method is proposed for creating an initial SEM model based on a newly designed index, which measures the association of two variables when applied in a regression model as independent variables. The new index is defined and calculated through a structural equation model (SEM).

The geochemical dataset to be analyzed included the geochemical concentrations of 16 elements (***Ag***, ***As***, ***Au***, ***Cu***, ***F***, ***Li***, ***Nb***, ***Pb***, ***Rb***, ***Sb***, ***Sn***, ***Th***, ***Ti***, ***W***, ***Zn***, ***Zr)*** abstained from 671 lake sediment samples from Southern Nova Scotia area (Rogers et al., 1987). The purpose of the data analysis was to extract geochemical factors with respect to ***Au***. Due to a large number of the samples with ***Au*** value below detection limit, the highly correlated element ***As*** was used as the response variable and the remaining 15 elements were analyzed by the new conditional correlation coefficient. To illustrate the implementation of the new cluster method with the new index and to compare the results with other existing indexes, three representative elements (viz., ***Au***, ***Cu***, ***Rb***) were analyzed in details. Further, PCA and hierarchical clustering with the divisive analysis algorithm (DIANA) (Kaufman and Rousseeuw, 2009) with matrices defined by the

correlation coefficient and by the new index were applied to the 15 geochemical elements.

6.2 Method

6.2.1 Association of two variables in regression to the dependent variable

The association of two variables (x_1, x_2) adopted in traditional clustering (i.e. CLV) methods is the correlation coefficient $\rho(x_1, x_2)$:

$$\rho(x_1, x_2) = \frac{Cov(x_1,x_2)}{\sigma(x_1)\times\sigma(x_2)} \qquad (6.1)$$

wherein $Cov(x_1, x_2)$ is the covariance between x_1 and x_2, $\sigma(x_1)$ and $\sigma(x_2)$ are the standard deviations of x_1 and x_2, respectively. The high value of $\mathrm{abs}(\rho(x_1, x_2))$ indicates a strong relationship between x_1 and x_2. However, the association of two variables with large correlation coefficient could be small if it is measured in regression to a response variable (y). For example, if x_1 is uncorrelated with y, for example, $\rho_{x1,y} = 0$, the conditional correlation coefficient of x_1 and x_2 in regression to y must be 0. A structural equation model (SEM) can be imported to describe the association of x_1 and x_2 under the given response variable (y) and to calculate the conditional correlation coefficient (***Fig** 6.1*). In ***Fig** 6.1*, the latent variable (x_{latent}) represents such a factor in x_1 and x_2 and its effect on y ($\rho_{y,x_{latent}}$) is equal to the combined effect of x_1 on y and x_{latent} ($\rho_{x_1,x_{latent}} \times \rho_{x_1,y}$). Similarly, the effect of the x_{latent} on y should be equal to the

combined effects of x_2 on y and x_{latent} ($\rho_{x_2,x_{latent}} \times \rho_{x_2,y}$). The above relationships can be expressed in equation (6.2). The parameters in ***Fig*** *6.1* are explained in ***Table*** *6.1*.

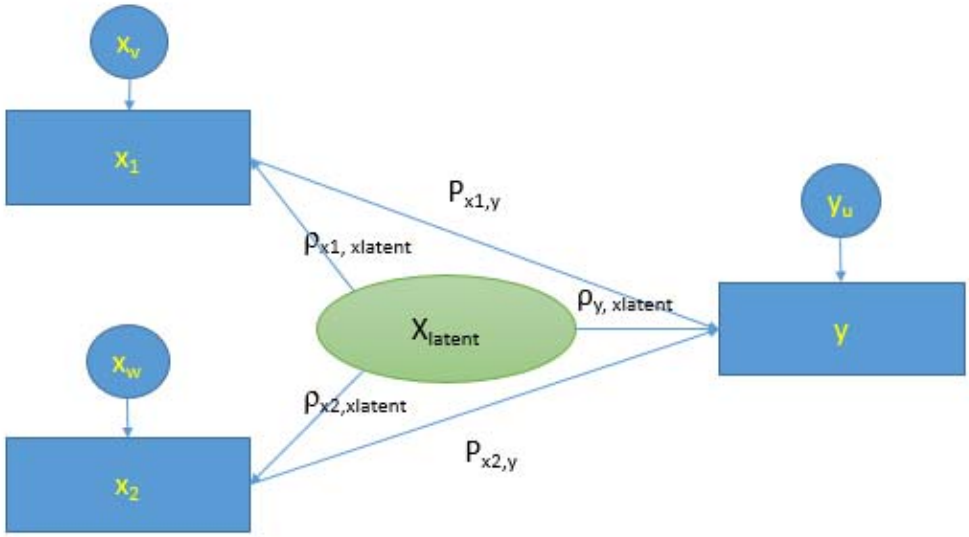


Fig *6.1 A structural equation model between observed variables* x_1 *and* x_2 *and a response variable y. The standardized association of* x_1 *and* x_2 *in regression to y is the ratio of* $\rho_{y,x_{latent}}$ *and* $\rho_{y.x_1,x_2}$*, where* $\rho_{y.x_1,x_2}$ *is the multiple correlation coefficient of* (x_1, x_2) *with y*

Let $x_{latent} = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 y$ and all of the variables (x_1, x_2, y, x_{latent}) be Z-standardized. Thus, their variancec and standard deviations become to 1, in addition, the covariances are equal to their correlation coefficients.

The relationship of x_{latent} with x_1, x_2 and y is defined as:

$$\rho_{x_1,x_{latent}} \times \rho_{x_1,y} = \rho_{x_2,x_{latent}} \times \rho_{x_2,y} = \rho_{y,x_{latent}} \qquad (6.2)$$

Table 6.1 Definition of parameters in Fig 6.1

Parameter	Definition
x_1, x_2, y	Observed quantitative variables
x_{latent}	Latent variable
y_u, x_v, x_w	Random disturbances
$\rho_{12}, \rho_{1y}, \rho_{2y}$	Correlation coefficient between x_1 *and* x_2, x_1 *and* y, x_2 *and* y
$P_{x1,y}$	Direct effect of x_1 to y
$P_{x2,y}$	Direct effect of x_2 to y
$\rho_{x1,xlatent}$	Loading of x_{latent} on x_1/Correlation coefficient between x_1 *and* x_{latent}
$\rho_{x2,xlatent}$	Loading of x_{latent} on x_2/Correlation coefficient between x_2 *and* x_{latent}
$\rho_{y,xlatent}$	Loading of x_{latent} on y/Correlation coefficient between y *and* x_{latent}

From ***Eq*** *(6.1)*, the correlation coefficient between x_{latent} and x_i (i = 1, 2) is expressed by:

$$\rho_{x_i,x_{latent}} = \frac{Cov(x_i,x_{latent})}{\sigma(x_i)\times\sigma(x_{latent})} = Cov(x_i, x_{latent}) \quad (6.3)$$

wherein $Cov(x_i, x_{latent})$ is the covariance between x_i and x_{latent}, and can be expressed as follows:

$$Cov(x_i, x_{latent}) = Cov(x_i, \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 y)$$

$$= \lambda_1 Cov(x_i, x_1) + \lambda_2 Cov(x_i, x_2) + \lambda_3 Cov(x_i, y)$$

$$= \lambda_1 \rho_{x_i,x_1} + \lambda_2 \rho_{x_i,x_2} + \lambda_3 \rho_{x_i,y} \quad (6.4)$$

And then, $\rho_{x_i,x_{latent}}$ is further given as

$$\rho_{x_i,x_{latent}} = \left(\lambda_1\rho_{1i} + \lambda_2\rho_{2i} + \lambda_3\rho_{yi}\right) \quad (6.5)$$

wherein ρ_{1i}, ρ_{2i}, ρ_{yi} are the correlation coefficients of x_i with x_1, x_2, and y, respectively. According to ***Eqs*** *(6.2)* to *(6.5)*, one can derive the following relations:

$$\rho_{x_1,x_{latent}} = \left(\lambda_1 + \rho_{12}\,\lambda_2 + \rho_{1y}\lambda_3\right) \quad (6.6)$$

$$\rho_{x_2,x_{latent}} = \left(\rho_{12}\lambda_1 + \lambda_2 + \rho_{2y}\lambda_3\right) \quad (6.7)$$

$$\rho_{y,x_{latent}} = \left(\rho_{1y}\lambda_1 + \rho_{2y}\lambda_2 + \lambda_3\right) \quad (6.8)$$

The standard deviation of x_{latent} ($\sigma_{x_{latent}}$) is:

$$\sigma(x_{latent}) = \sigma(\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 y)$$

$$= \sqrt{{\lambda_1}^2 + {\lambda_2}^2 + {\lambda_3}^2 + 2\rho_{12}\lambda_1\lambda_2 + 2\rho_{1y}\lambda_1\lambda_3 + 2\rho_{2y}\lambda_2\lambda_3} = 1 \quad (6.9)$$

wherein $\lambda_1, \lambda_2, \lambda_3$ can be estimated through ***Eq*** *(6.2)* and *(6.9)*.

$\rho_{y,x_{latent}}$ can be used to quantitatively measure the association between x_1 and x_2, when they are applied in regression with respect to *y*. The standardized $\rho_{y,x_{latent}}$ is given by:

$$\mathrm{R_y}(\mathrm{x_1},\mathrm{x_2}) = \left|\frac{\rho_{y,\mathrm{x_{latent}}}}{\rho_{\,y.\mathrm{x_1},\mathrm{x_2}}}\right| \qquad (6.10)$$

wherein $\rho_{\,y.\mathrm{x_1},\mathrm{x_2}}$ is the multiple correlation coefficient of $\{x_1, x_2\}$ with *y*.

The new index has the following characteristics:

1. $0 \leq \mathrm{R} \leq 1$, *large R means strong similarity between* x_1 *and* x_2, *for example, let* $x_1 = x_2$, then $x_{latent} = x_1 = x_2$, $\mathrm{R} = 1$.
2. If $x_1 = y$ or $x_2 = y$, then $x_{latent} = x_2 \; or \; x_1$, $\rho_{x_{latent},y} = \rho_{x_1,x_2}$, $\rho_{(x_1,x_2),y} = 1$, $R = \rho_{x_1,x_2}$.
3. If $\rho_{x_1,y} = 0$ or $\rho_{x_2,y} = 0$, then x_{latent} cannot be estimated, and $\rho_{x_{latent},y} = 0$, then $R = 0$.
4. $\mathrm{R_y}(\mathrm{x_1},\mathrm{x_2}) = \mathrm{R_y}(\mathrm{x_2},\mathrm{x_1})$
5. $\rho_{x_1,y} = \rho_{y.x_1,x_{latent}}$, $\rho_{x_2,y} = \rho_{y.x_2,x_{latent}}$, $\rho_{y.x_1,x_2} = \rho_{y.x_1,x_2,x_{latent}}$, where $\rho_{y.x_1,x_{latent}}$, $\rho_{y.x_2,x_{latent}}$ and $\rho_{y.x_1,x_2,x_{latent}}$ are multiple correlation coefficients between $\{x_1, x_{latent}\}$, $\{x_2, x_{latent}\}$ and $\{x_1, x_2, x_{latent}\}$ with *y*, respectively.

The above properties indicate that the new index *R* is a symmetrical statistic measure of the relative conditional correlation between x_1 and x_2 under the regression to variable y. This index will be applied to analyze the data from 15 elements in lake sediment samples with ***As*** as the dependent variable as explained in the next section.

6.3 Case study

6.3.1 The dataset

The geochemical data used in this research is geochemical lake sediment data and was acquired from Southwestern Nova Scotia, Canada. The data included 671 samples with the concentration values of 16 elements: ***Ag***, ***As***, ***Au***, ***Cu***, ***F***, ***Li***, ***Nb***, ***Pb***, ***Rb***, ***Sb***, ***Sn***, ***Th***, ***Ti***, ***W***, ***Zn*** and ***Zr*** (Rogers and Garrett, 1987). The detailed information about this dataset was introduced in Chapter 3 and additional information about it and the study area was also made available in literature (Bonham-Carter et al., 1988; Dunn et al., 1991; Rogers et al., 1990; Rogers et al., 1987).

In order to extract geologic factors related to gold mineralization based on the clusters, ***As*** was selected as the domain element in the classification instead of ***Au*** mainly for two reasons: firstly, ***As*** was highly correlated with ***Au*** in the gold mineralization in the area (Agterberg et al., 1990; Xu, 2001; Xu and Cheng, 2001) and secondly, the concentration values of ***Au*** from some samples were below the detection limit. The previous research shows that there is a particularly marked coherence between arsenic and gold in practically all types of gold deposits (Boyle and Jonasson, 1973) and the multiple sources of arsenic is more conductive to extract independent factors corresponding different sources. The multiple sources of ***As*** can represent several geology processes/factors. Some of them may be related with the gold mineralization but others may be not. The latent variables derived by constraints of ***As*** from multiple elements may provide information for distinguishing multiple sources of ***As***.

To normalize and standardize the variables, the log-transformation and the Z-standardization were applied to the dataset. The correlation coefficient matrix based on the transformed data is shown in ***Table** 6.2*, while the standardized new distance index (*R*) matrix is shown in ***Table** 6.3*. To explain the difference between these two indexes, the matrices of new index ($\rho_{x_{latent},y}$) and multiple correlation coefficient of two elements with ***As*** ($\rho_{y.x_1,x_2}$) were also calculated (***Tables** 6.4* and *6.5*).

6.3.2 The difference between two indexes for *Au*, *Cu* and *Rb*

Four parameters with the clustering method were compared and contrasted for meaningful interpretation: i). Standardized covariance (ρ_{x_1,x_2}), ii). Standardized new index ($R_y(x_1, x_2)$) between x_1 and x_2, iii). New index ($\rho_{x_{latent},y}$), iv). Multiple correlation coefficient of $\{x_1, x_2\}$ to y ($\rho_{((x_1,x_2),y)}$). The elements ***Au***, ***Cu*** and ***Rb*** were selected for further discussion because they represent particular types of relationships with other elements. For example, although ***Au*** is a very important geochemical element for gold mineral exploration, it has relatively low correlation coefficients with most of the other elements (except for ***As***). Therefore, the importance of ***Au*** cannot be reflected in clustering through the existing correlation coefficient index. However, the new index may enhance the relationships of ***Au*** with other elements under the conditional of ***As***. On the contrary, ***Rb*** is weakly related to ***Ag***, ***Au***, ***As***, ***Cu***, and ***Zn*** but strongly related with other elements, whereas ***Cu*** is related to ***Ag***, ***Au***, ***As***, ***Pb***, and ***Zn.*** How the new index exhibits different relationships between ***Au***, ***Cu*** and ***Rb*** and other elements was checked and will be further explained below.

1) *The relationship of gold with other elements*

The relationship of ***Au*** with other elements is shown in ***Fig*** *6.2* as a radar graph. There are four indexes mapped in ***Fig*** *6.2*: 1. Multiple correlation coefficient between {***Au***, x_i} and ***As*** ($\rho_{As.Au,x_i}$, blue line); 2. New index between ***Au*** and x_i under the restriction of ***As*** ($\rho_{x_{latent},As}$, red line); 3. Correlation coefficient between ***Au*** and x_i (ρ_{Au,x_i}, purple line); and 4. Standardized new index between ***Au*** and x_i under the restriction of ***As*** ($R_{As}(Au, x_i)$, green line). If the area of the index polygon represents the total effect of ***Au*** in clustering, it is observed that the standardized new index (green line) is greatly enhanced relative to the correlation coefficient (purple line). On the unit circle representing the correlation coefficient, all the elements except ***As*** and ***Au*** are located inside the first circle of 0.2. However, with the standardized new index, all elements except ***Nb*** and ***Zr*** are located outside the first circle. The multiple correlation coefficient of {***Au, Zn***} and {***Au, Cu***} with ***As*** (blue line) is larger than the related multiple correlation coefficients of the other elements with ***As***. However, large values only at 0.6 and 0.8 for the standardized new index indicate that most of the ***As***-related information in ***Au*** is different from that in ***Zn*** and ***Cu***. The most related element is ***W***, both with the correlation coefficient and the standardized new index. In ***Eq*** *(6.11)*, a new parameter $\delta_{(x_0,x_i),y}$ is introduced to measure the degree of importance of x_0 with other elements both using the correlation coefficient (ρ_{x_0,x_i}) and the standardized new index $R_y(x_0, x_i)$. In this case, y is ***As***, x_0 is ***Au*** and x_i is an element other than ***As*** and ***Au***. A positive value of $\delta_y(x_0, x_i)$ means that the relationship between x_0 and x_i in the standardized new index is enhanced than the correlation coefficient and reduced otherwise.

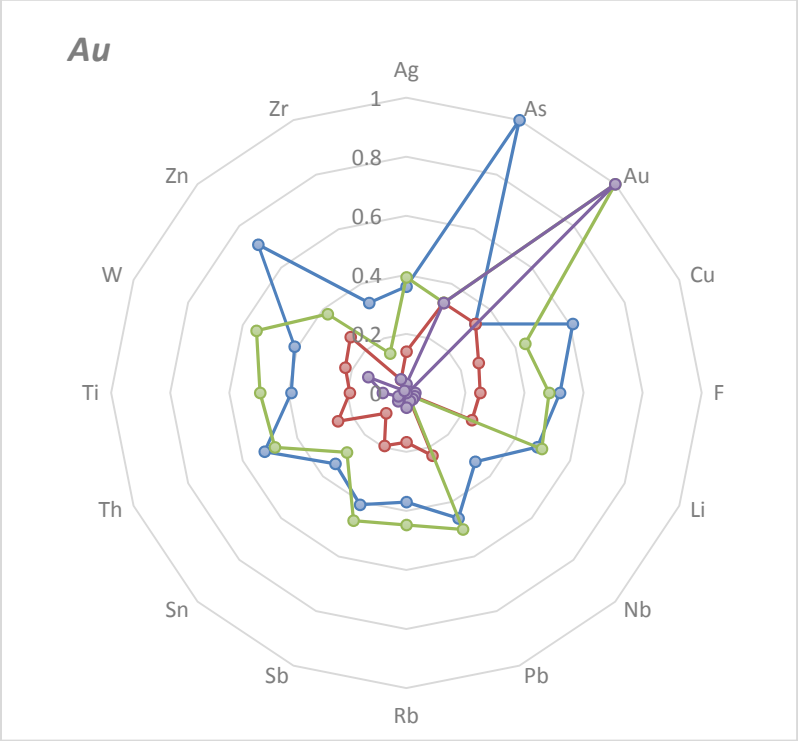


Fig *6.2 Relationships of Au with other elements in* $\rho_{As.Au,x_i}$ *(blue line),* $\rho_{x_{latent},As}$ *(red line),* ρ_{Au,x_i} *(purple line), and* $R_{As}(Au,x_i)$ *(green line), where* x_{latent} *was defined in Fig 6.1 and* $R_{As}(Au,x_i)$ *was defined in Eq 6.10.*

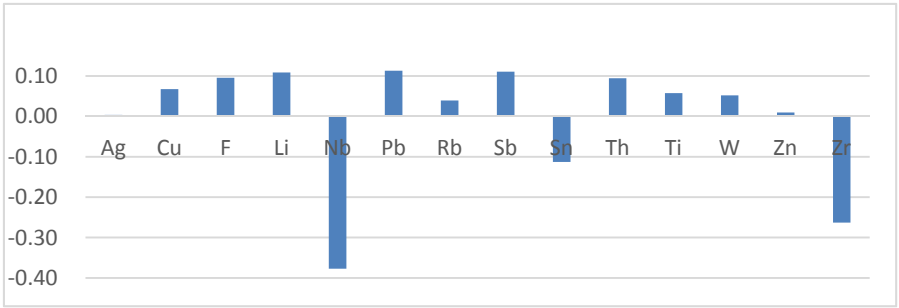


Fig *6.3* $\delta_{As}(Au,x_i)$*: Change in the relationships of Au with other elements in the standardized new index* $R_{As}(Au,x)$ *and correlation coefficient* ρ_{Au,x_i}*;*

The relationship change between ***Au*** (x_0) and element x_i is shown in ***Fig*** *6.3*.

$$\delta_y(x_0, x_i) = \frac{1}{n}\sum_{j=j\neq i}^{n}\left[R_y(x_0, x_i) - R_y(x_0, x_j) - abs\left(\rho_{x_0, x_i}\right) + abs\left(\rho_{x_0, x_j}\right)\right] \quad (6.11)$$

wherein the x_i represents the i^{th} element except y and x_0.

2) *The relationship of copper with other elements*

The relationship of ***Cu*** with other elements is shown in ***Fig*** *6.4*. There are four indexes mapped in ***Fig*** *6.4*: 1. Multiple correlation coefficient between {***Cu***, x_i} and ***As*** ($\rho_{As.Cu,x_i}$, blue line); 2. New index between ***Cu*** and x_i under the restriction of ***As*** ($\rho_{x_{latent},As}$, red line); 3. Correlation coefficient between ***Cu*** and x_i (ρ_{Cu,x_i}, purple line); and 4. Standardized new index between ***Cu*** and x_i under the restriction of ***As*** ($R_{As}(Cu, x_i)$, green line). Its relationships with other elements using the standardized new index are similar to the one using the correlation coefficient. ***Zn*** and ***Th*** are two most pertinent elements both in the standardized new index and the correlation coefficient, and most of the relationships are strengthened because of their strong relationship with the response variable. However, some differences between two indexes can be noticed in ***Fig*** *6.5*, which is explained by ***Eq*** *(6.11)*. While the relationships of ***Cu*** with ***Au***, ***W***, ***Pb*** and ***F*** are greatly enhanced using the standardized new index, its relationship with ***Zn*** is slightly reduced because of the conservative nature of the correlation coefficient. The relationships of ***Cu*** with ***Nb*** and ***Zr*** are reduced for the same reason discussed in the previous section.

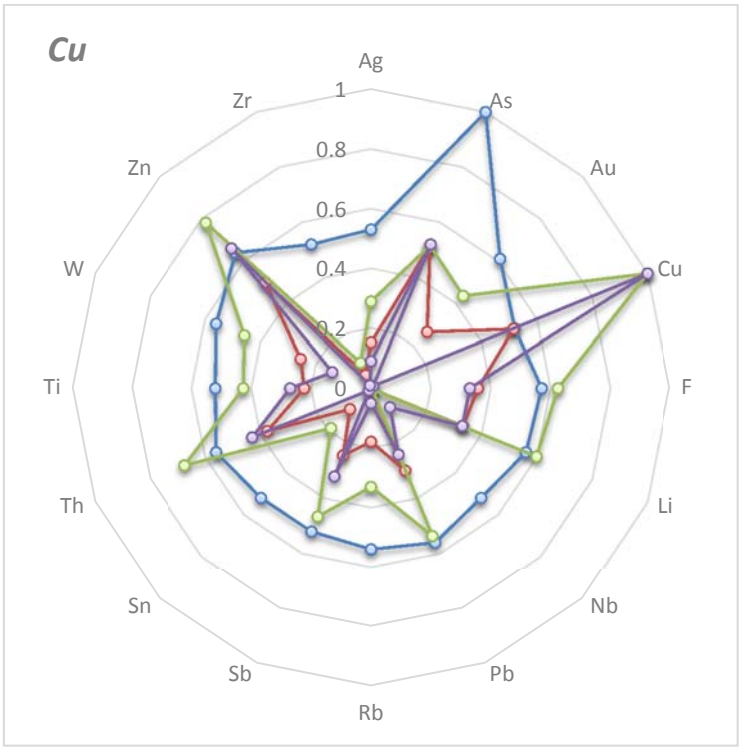


***Fig** 6.4 Relationships of Cu with other elements in $\rho_{As.Cu,x_i}$ (blue line), $\rho_{x_{latent},As}$ (red line), ρ_{Cu,x_i} (purple line), and $R_{As}(Cu,x_i)$ (green line), where x_{latent} was defined in Fig 6.1 and $R_{As}(Cu,x_i)$ was defined in Eq (6.10).*

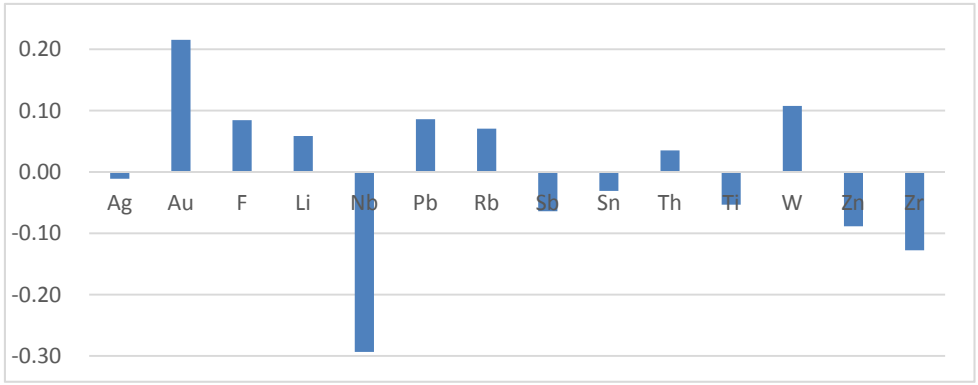


***Fig** 6.5 $\delta_{As}(Cu,x_i)$: Change of relationships of Cu with other elements in the new index $R_{As}(Cu,x)$ and correlation coefficient ρ_{Cu,x_i};*

3) *The relationship of rubidium with other elements*

The relationship of ***Rb*** with other elements is shown in ***Fig*** *6.6*. There are four indexes mapped in ***Fig*** *6.6*: 1. Multiple correlation coefficient between {***Rb***, x_i} and ***As*** ($\rho_{As.Rb,x_i}$, blue line); 2. New index between ***Rb*** and x_i under the restriction of ***As*** ($\rho_{x_{latent},As}$, red line); 3. Correlation coefficient between ***Rb*** and x_i (ρ_{Rb,x_i}, purple line); and 4. Standardized new index between ***Rb*** and x_i under the restriction of ***As*** ($R_{As}(Rb, x_i)$, green line). The difference in relationship of ***Rb*** with other elements in the standardized new index and the correlation coefficient is shown in ***Fig*** *6.7*. Its relationship with ***Zr*** is the strongest one with the correlation coefficient, but small with the standardized new index, mostly because of the poor relationship of ***Zr*** with ***As***. Moreover, using the standardized new index it is observed that while the relationship of ***Rb*** with ***Nb***, ***Li***, ***Th*** and ***F*** are weaker, its relationships with ***Ag***, ***Au***, ***Sb***, ***W*** and ***Cu*** are stronger than using the correlation coefficient. These results indicate that there is a significant change in relationship of ***Rb*** with other elements using the standardized new index.

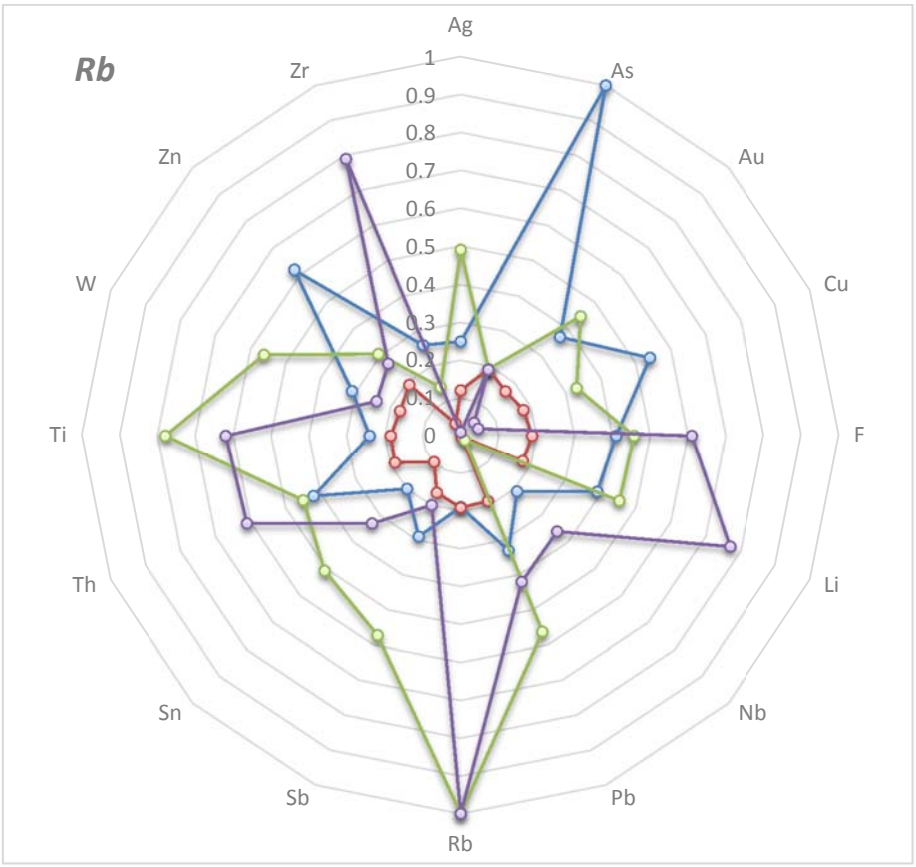


Fig *6.6 Relationships of Rb with other elements in* $\rho_{As.Rb,x}$ *(blue line),* $\rho_{x_{latent},As}$ *(red line),* $\rho_{Rb,x}$ *(purple line), and* $R_{As}(Rb, x_i)$ *(green line), where* x_{latent} *was defined in Fig 6.1 and* $R_{As}(Rb, x_i)$ *was defined in Eq 6.10.*

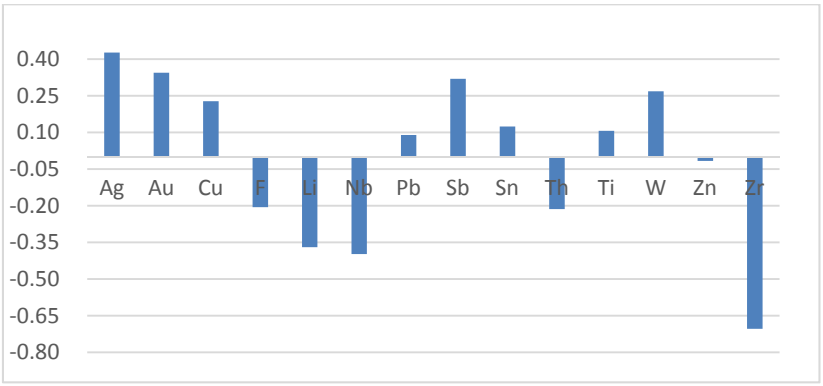


Fig *6.7* $\delta_{As}(Rb, x_i)$*: Change of relationships of Rb with other elements in the standardized new index* $R_{As}(Rb, x)$ *and correlation coefficient* ρ_{Rb,x_i}*.*

***Table** 6.2 Correlation coefficient matrix*

	Ag	***As***	***Au***	***Cu***	***F***	***Li***	***Nb***	***Pb***	***Rb***	***Sb***	***Sn***	***Th***	***Ti***	***W***	***Zn***	***Zr***
Ag	1.00															
As	0.16	1.00														
Au	0.03	0.33	1.00													
Cu	0.09	0.52	0.01	1.00												
F	0.08	0.41	0.03	0.33	1.00											
Li	0.08	0.36	0.03	0.33	0.69	1.00										
Nb	-0.01	0.00	0.03	0.09	0.28	0.35	1.00									
Pb	0.10	0.33	0.03	0.24	0.35	0.43	0.17	1.00								
Rb	0.01	0.19	0.05	0.05	0.61	0.77	0.36	0.42	1.00							
Sb	0.04	0.25	0.00	0.32	0.23	0.28	0.23	0.42	0.20	1.00						
Sn	0.08	0.10	0.04	0.01	0.21	0.25	0.00	0.14	0.33	0.06	1.00					
Th	0.08	0.41	0.03	0.43	0.56	0.65	0.33	0.42	0.61	0.33	0.19	1.00				
Ti	0.04	0.23	0.08	0.27	0.48	0.61	0.52	0.37	0.62	0.32	0.23	0.56	1.00			
W	-0.02	0.28	0.14	0.14	0.29	0.24	0.14	0.11	0.24	0.07	0.16	0.20	0.24	1.00		
Zn	0.12	0.62	-0.01	0.66	0.54	0.59	0.10	0.42	0.27	0.38	0.12	0.51	0.34	0.21	1.00	
Zr	0.00	0.05	0.05	-0.01	0.38	0.55	0.36	0.36	0.79	0.17	0.31	0.58	0.71	0.14	0.10	1.00

Table *6.3 Standardized new index $R_y(x_1,x_2)$ under the restriction of As*

	Ag	***As***	***Au***	***Cu***	***F***	***Li***	***Nb***	***Pb***	***Rb***	***Sb***	***Sn***	***Th***	***Ti***	***W***	***Zn***	***Zr***
Ag	1.00															
As	0.16	1.00														
Au	0.39	0.33	1.00													
Cu	0.29	0.52	0.44	1.00												
F	0.35	0.41	0.48	0.62	1.00											
Li	0.38	0.36	0.50	0.60	0.83	1.00										
Nb	0.02	0.00	0.01	0.01	0.01	0.01	1.00									
Pb	0.41	0.33	0.50	0.54	0.65	0.70	0.01	1.00								
Rb	0.49	0.19	0.45	0.33	0.46	0.45	0.02	0.56	1.00							
Sb	0.46	0.25	0.47	0.47	0.52	0.58	0.01	0.66	0.57	1.00						
Sn	0.49	0.10	0.28	0.19	0.25	0.28	0.03	0.30	0.51	0.36	1.00					
Th	0.35	0.41	0.48	0.68	0.78	0.81	0.01	0.67	0.45	0.55	0.25	1.00				
Ti	0.48	0.23	0.50	0.43	0.56	0.63	0.01	0.62	0.78	0.65	0.42	0.55	1.00			
W	0.41	0.28	0.55	0.46	0.58	0.59	0.01	0.54	0.56	0.52	0.35	0.54	0.60	1.00		
Zn	0.25	0.62	0.38	0.78	0.64	0.58	0.01	0.52	0.31	0.40	0.16	0.64	0.36	0.42	1.00	
Zr	0.28	0.05	0.14	0.09	0.11	0.10	0.06	0.14	0.14	0.19	0.46	0.09	0.14	0.17	0.08	1.00

Table *6.4 New index under the restriction of As*

	Ag	***As***	***Au***	***Cu***	***F***	***Li***	***Nb***	***Pb***	***Rb***	***Sb***	***Sn***	***Th***	***Ti***	***W***	***Zn***	***Zr***
Ag	0.00															
As	0.16	1.00														
Au	0.14	0.33	0.33													
Cu	0.15	0.52	0.26	0.52												
F	0.15	0.41	0.25	0.36	0.41											
Li	0.15	0.36	0.24	0.33	0.35	0.36										
Nb	0.00	0.00	0.00	0.00	0.00	0.00	0.00									
Pb	0.15	0.33	0.23	0.30	0.29	0.29	0.00	0.33								
Rb	0.12	0.19	0.17	0.18	0.19	0.18	0.00	0.19	0.19							
Sb	0.13	0.25	0.19	0.24	0.23	0.23	0.00	0.23	0.16	0.25						
Sn	0.09	0.10	0.10	0.10	0.10	0.10	0.00	0.10	0.10	0.10	0.10					
Th	0.15	0.41	0.25	0.38	0.36	0.35	0.00	0.30	0.19	0.24	0.10	0.41				
Ti	0.13	0.23	0.19	0.22	0.23	0.23	0.00	0.21	0.18	0.19	0.10	0.23	0.23			
W	0.13	-0.28	0.22	0.25	0.26	0.24	0.00	0.22	0.17	0.19	0.10	0.25	0.19	-0.28		
Zn	0.15	0.62	0.27	0.50	0.40	0.36	0.00	0.33	0.19	0.25	0.10	0.40	0.23	0.27	0.62	
Zr	0.05	0.05	0.05	0.05	0.05	0.04	0.00	0.05	0.04	0.05	0.05	0.04	0.04	0.05	0.05	0.05

***Table** 6.5 Multiple correlation coefficient between {x_1,x_2} and As*

	Ag	***As***	***Au***	***Cu***	***F***	***Li***	***Nb***	***Pb***	***Rb***	***Sb***	***Sn***	***Th***	***Ti***	***W***	***Zn***	***Zr***
Ag	0.16															
As	1.00	1.00														
Au	0.36	1.00	0.33													
Cu	0.53	1.00	0.61	0.52												
F	0.42	1.00	0.52	0.57	0.41											
Li	0.39	1.00	0.48	0.56	0.42	0.36										
Nb	0.16	1.00	0.33	0.52	0.42	0.39	0.00									
Pb	0.35	1.00	0.46	0.56	0.45	0.41	0.33	0.33								
Rb	0.25	1.00	0.37	0.54	0.41	0.39	0.21	0.33	0.19							
Sb	0.29	1.00	0.41	0.52	0.44	0.39	0.26	0.35	0.29	0.25						
Sn	0.18	1.00	0.34	0.52	0.41	0.36	0.10	0.33	0.20	0.26	0.10					
Th	0.43	1.00	0.52	0.56	0.46	0.43	0.44	0.45	0.42	0.43	0.41	0.41				
Ti	0.27	1.00	0.39	0.52	0.41	0.36	0.27	0.35	0.24	0.29	0.23	0.41	0.23			
W	0.32	1.00	0.41	0.56	0.44	0.41	0.28	0.41	0.31	0.36	0.29	0.46	0.33	0.28		
Zn	0.63	1.00	0.71	0.64	0.63	0.62	0.63	0.63	0.62	0.62	0.62	0.63	0.62	0.64	0.62	
Zr	0.16	1.00	0.33	0.52	0.42	0.41	0.05	0.34	0.26	0.25	0.10	0.47	0.28	0.28	0.62	0.05

6.4 Main components calculated through different matrixes

The traditional clustering method is an unconstrained method using the correlation coefficient matrix. However, the new index ($\rho_{x_{latent},As}$ *or* $R_y(x_1, x_2)$) is obtained for a constrained clustering. In order to identify the new index, the main components of the element group based on the Eigen decomposition of a matrix are calculated through the correlation coefficient matrix (***Table*** *6.2*), the new index (***Table*** *6.4*) and the standardized new index (***Table*** *6.3*). The Eigen values of decomposition are shown in ***Fig*** *6.8*. The corresponding loadings of the top three main components are shown in ***Fig*** *6.9*. The correlation coefficients of each component with ***As*** are shown in ***Fig*** *6.10*. The spatial distribution of the corresponding components is mapped in ***Fig*** *6.11*. Finally, the difference of the two types of indexes in variable clustering is inspected through the hierarchal clustering with divisive analysis (DIANA) algorithm (Kaufman and Rousseeuw, 2009). The clustering results are shown in ***Fig*** *6.12***.** The dissimilarity matrixes are defined as *1- R* and *1-*R_y respectively; where the *R* and R_y are the correlation coefficient matrix and standardized new index matrix, respectively.

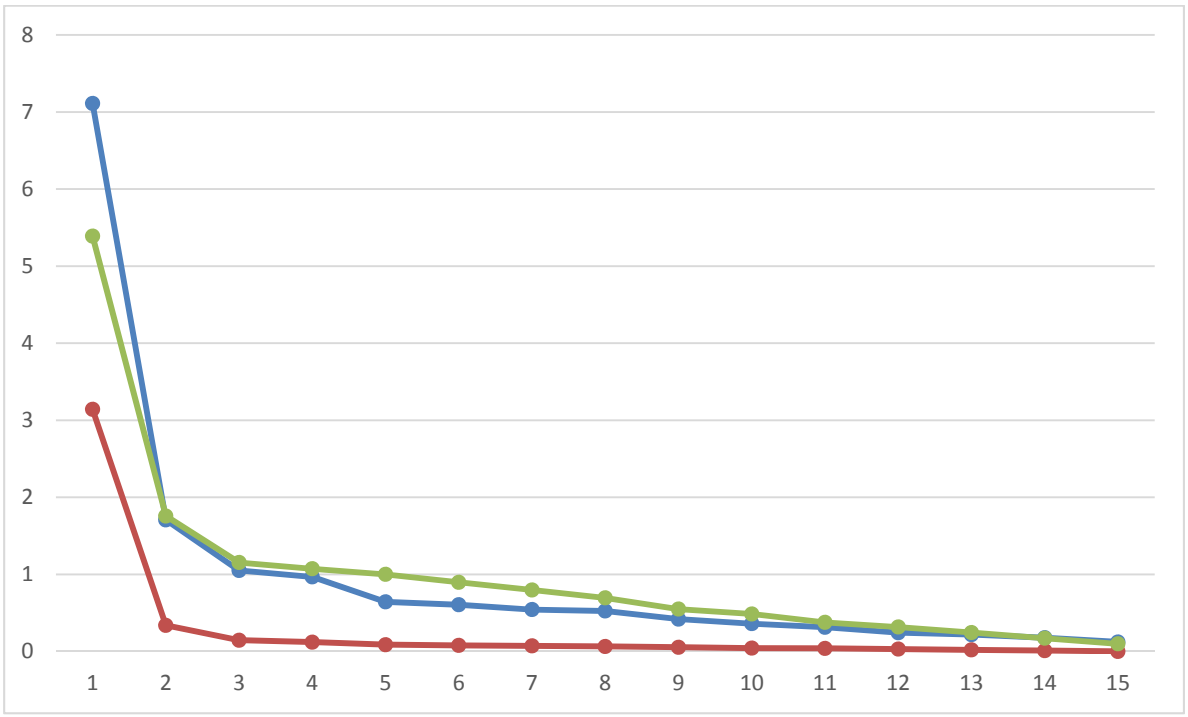


***Fig** 6.8 Eigenvalues of the decomposition of three different matrixes: Green curve: correlation coefficient matrix; Brown curve: new index matrix; Blue curve: standardized new index matrix. Vertical axis: Eigenvalue; parallel axis: index which ordered in descending.*

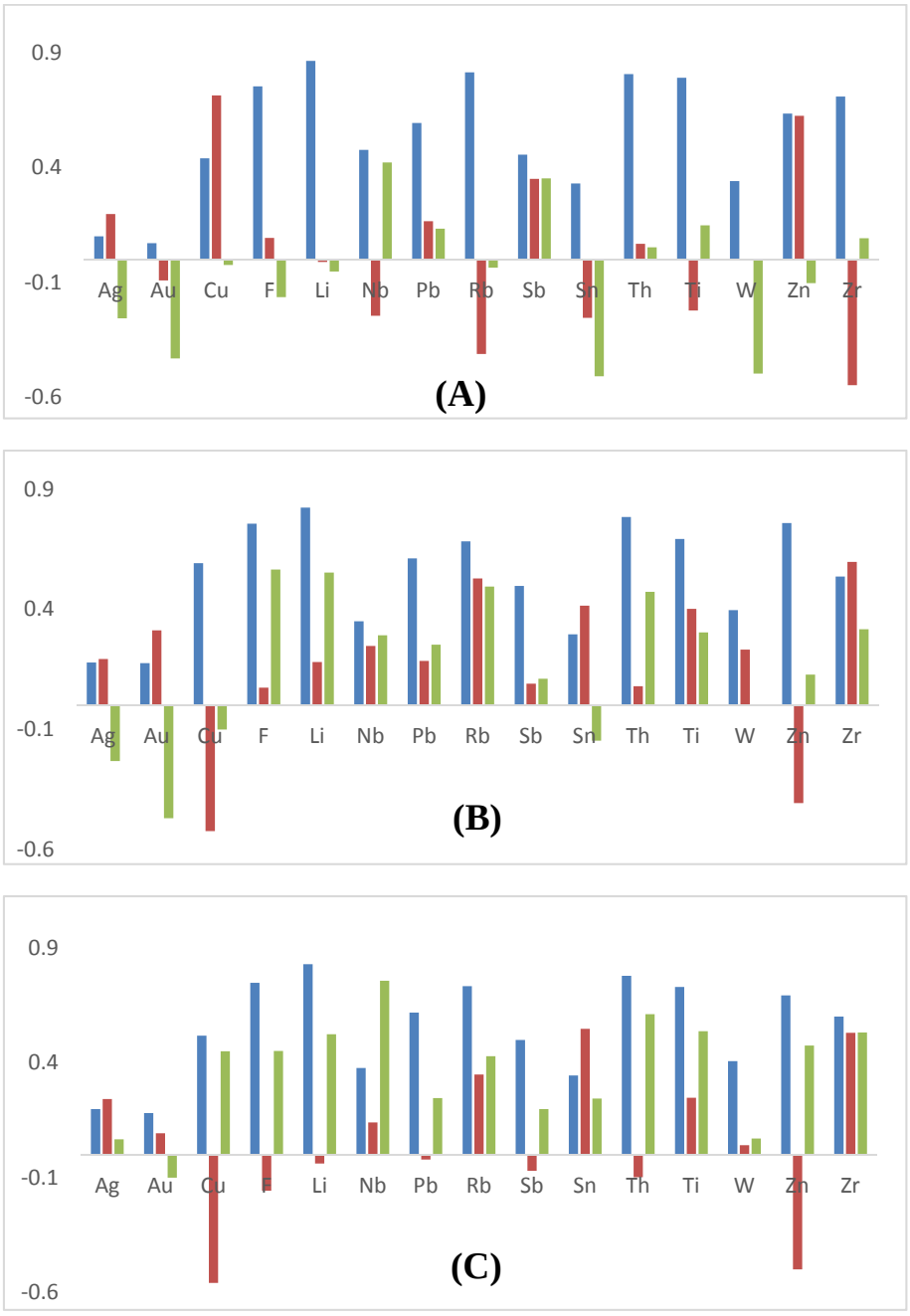


Fig *6.9 Loadings of PC1 (blue bar), PC2 (brown bar) and PC3 (green bar) on 15 elements which calculated through the matrix of: A) correlation coefficient; B) new index; C) standardized new index. Vertical axis: loadings; Parallel axis: elements.*

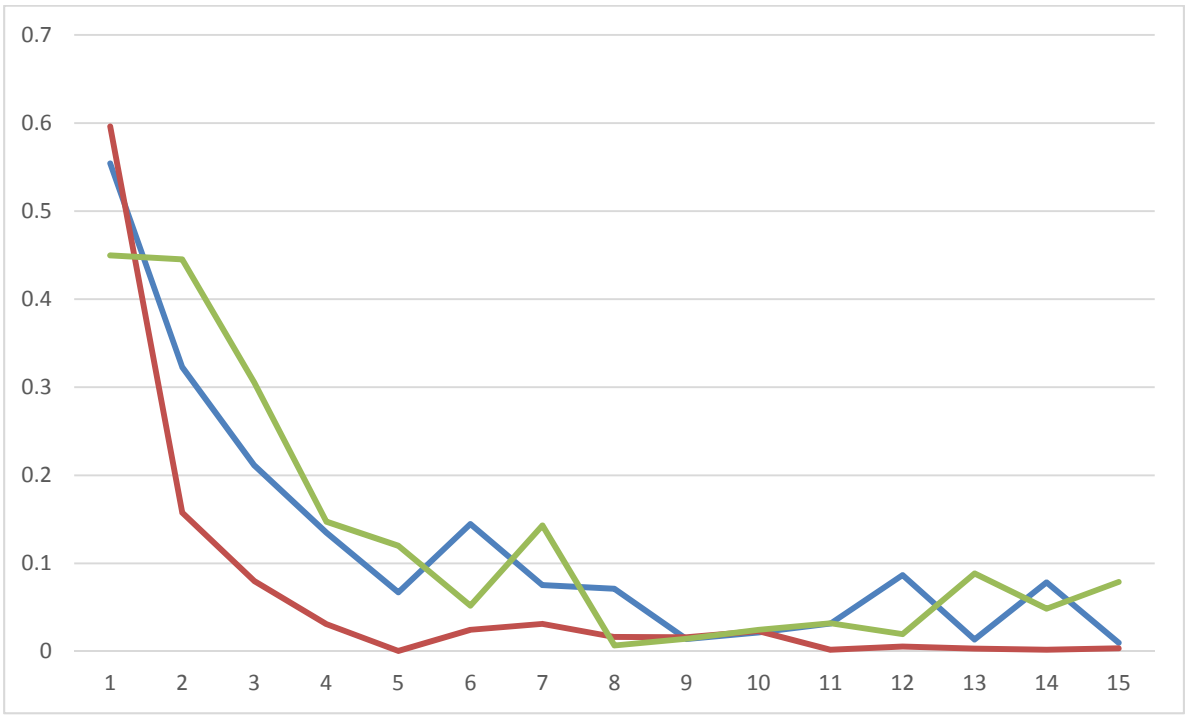


***Fig** 6.10 The absolute value of correlation coefficient of each component with As. vertical axis: correlation coefficient; parallel axis: component index which ordered by Eigen value in descending order. Blue curve: components though standardized new index matrix; Brown curve: components though new index matrix; Green curve: components though standardized new index matrix;*

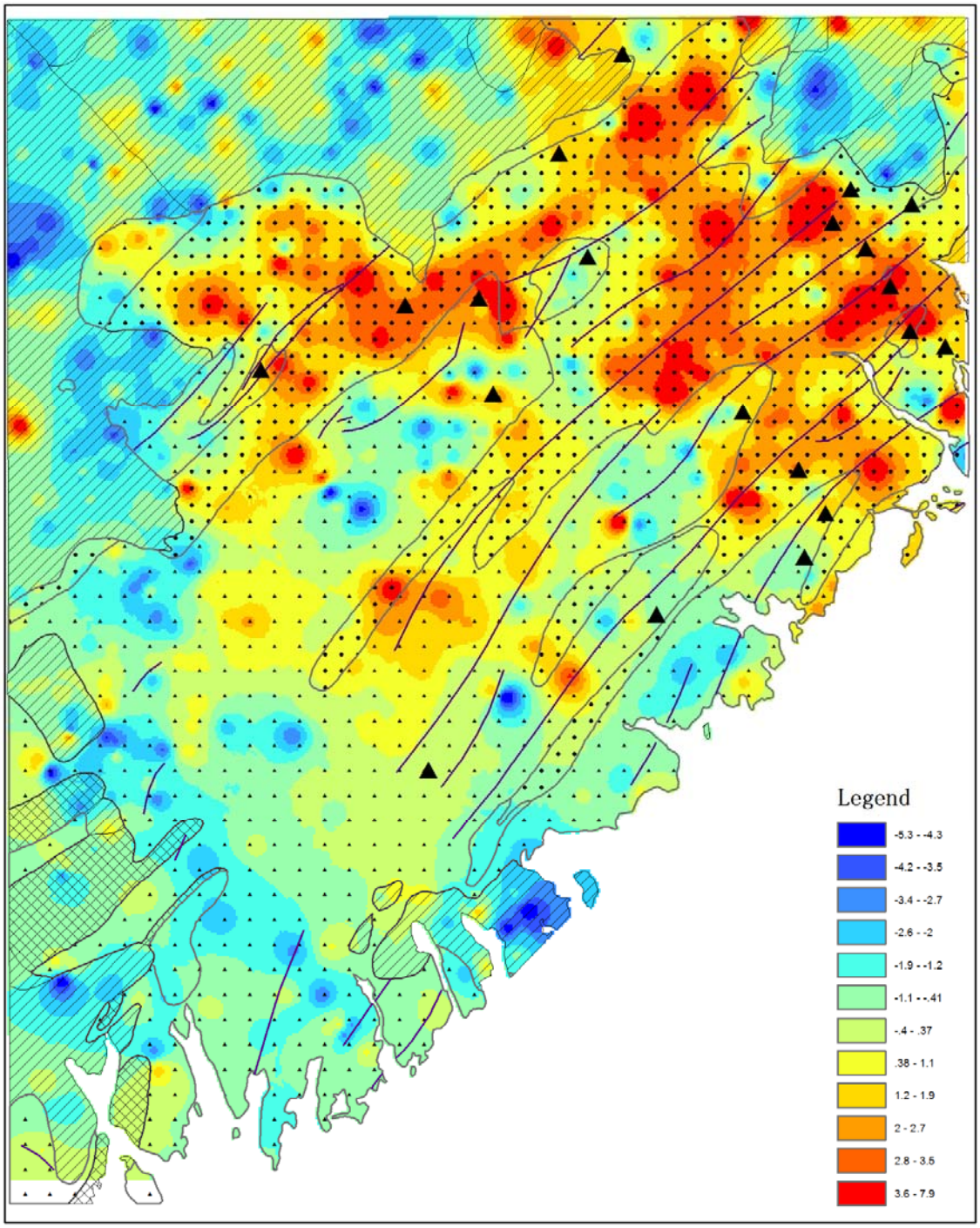


***Fig** 6.11A The first component through matrix of correlation coefficient, other legends are shown in Fig 6.12.*

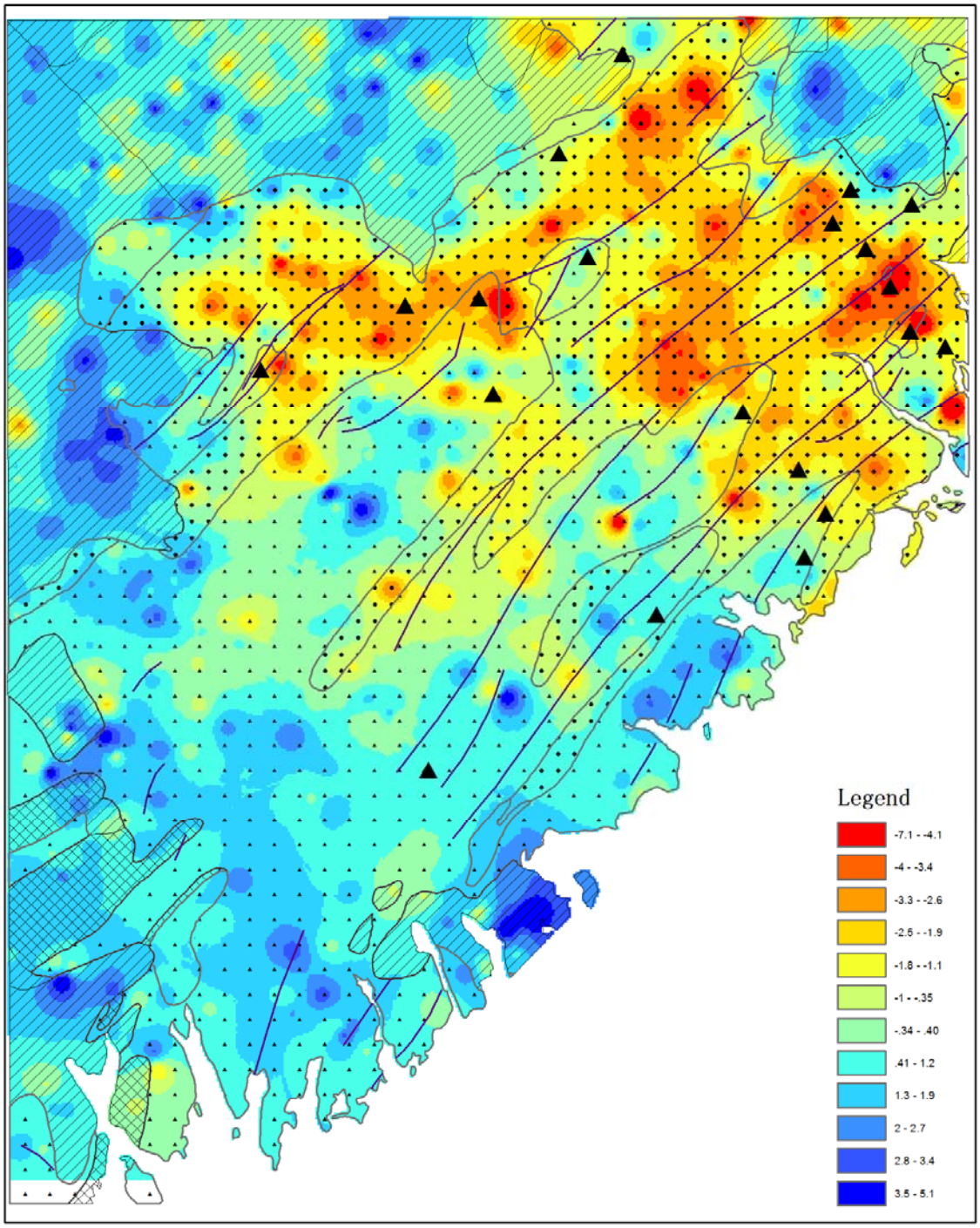


***Fig** 6.11B The first component through matrix of new index, other legends are shown in Fig 6.12..*

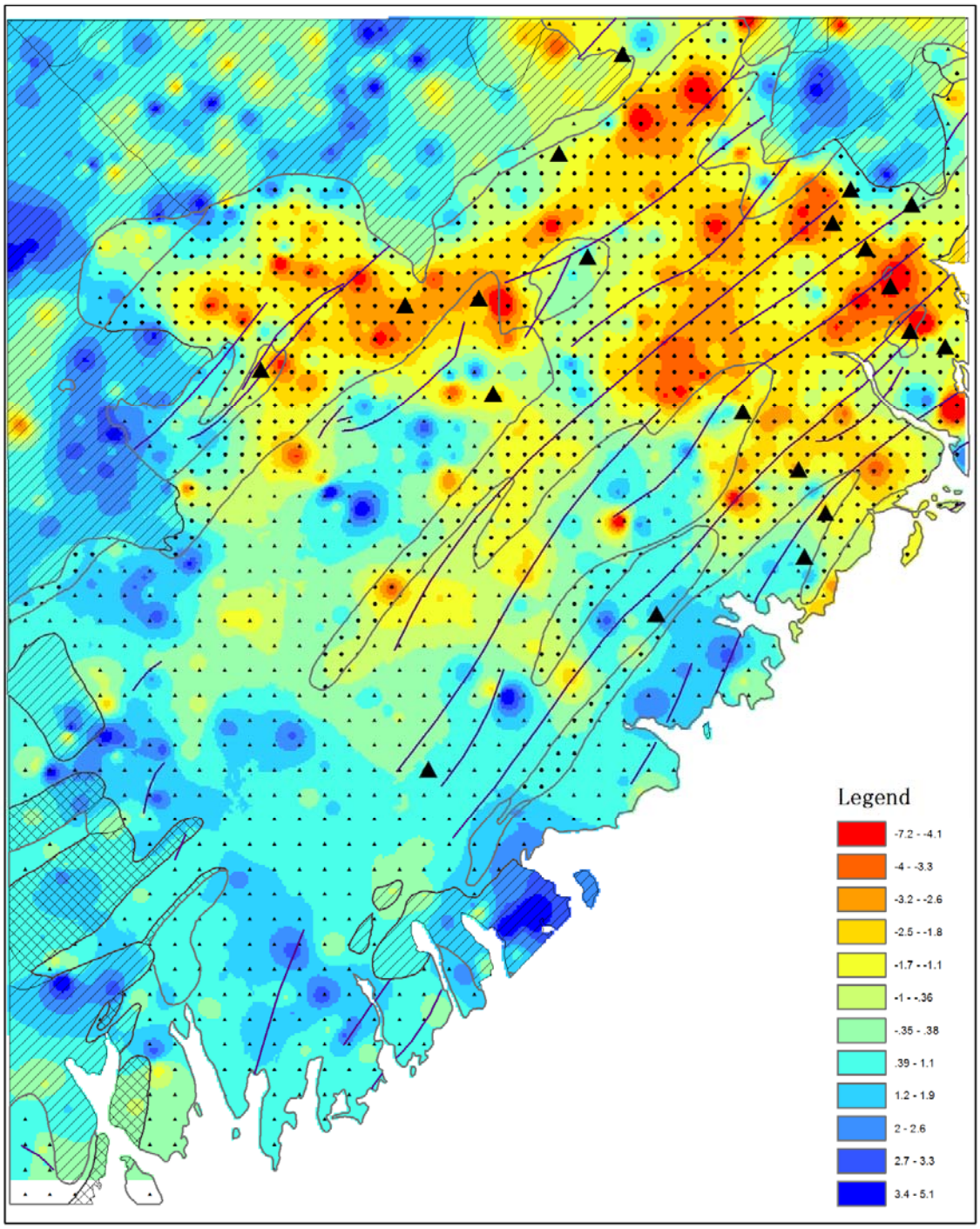


***Fig** 6.11C The first component through matrix of standardized new index, other legends are shown in Fig 6.12.*

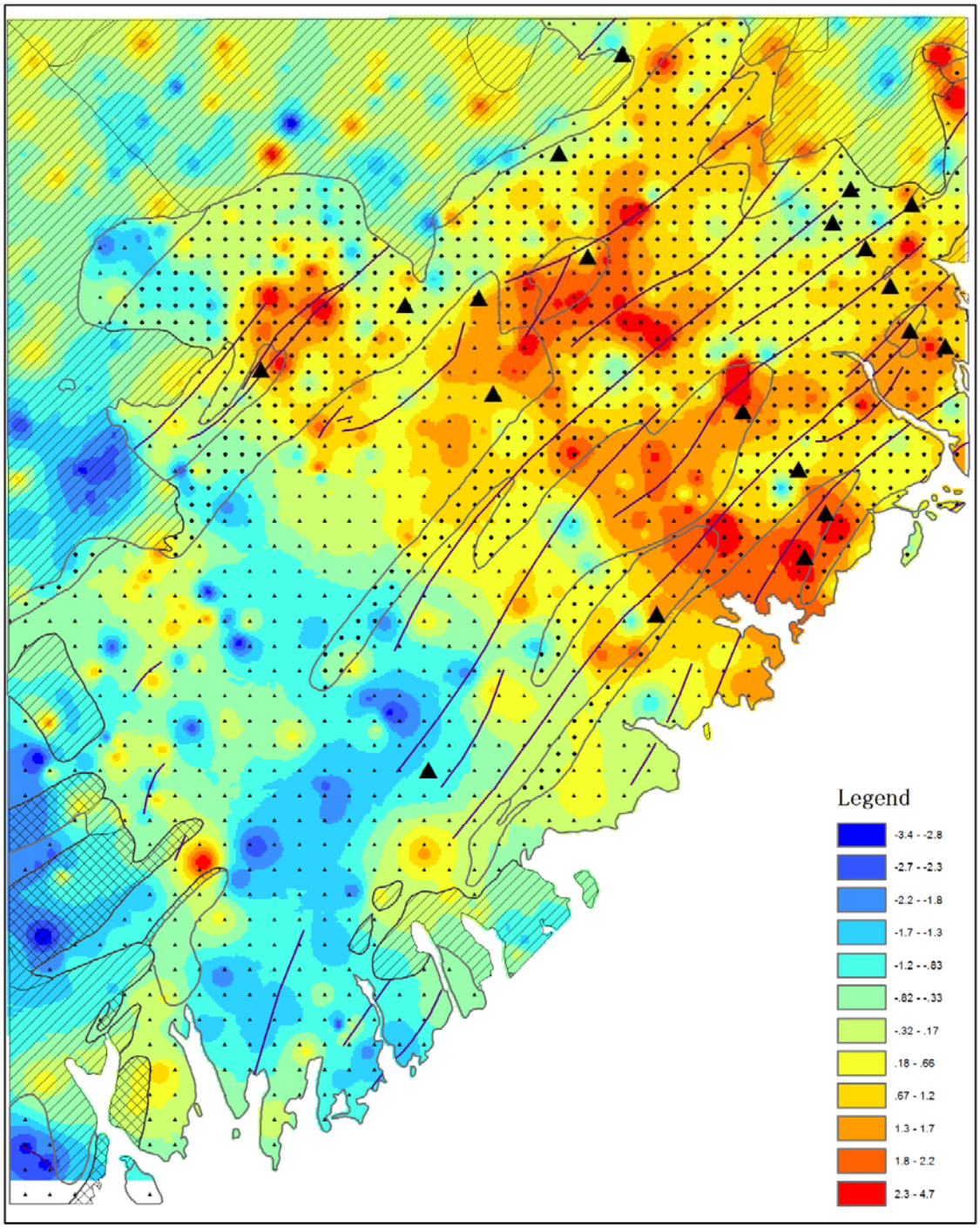


Fig *6.11D The second component through matrix of correlation coefficient. , other legends are shown in Fig 6.12.*

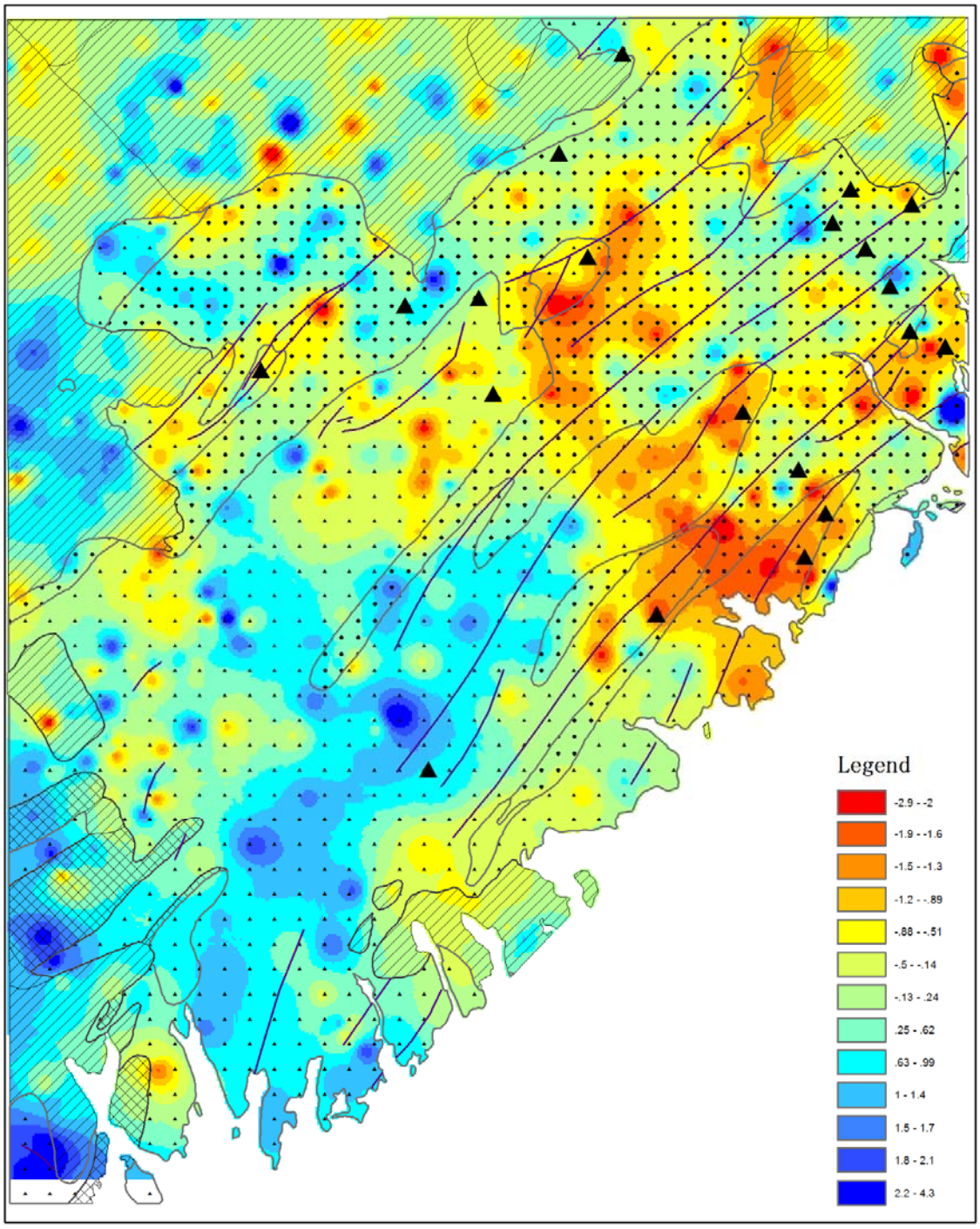


***Fig** 6.11E The second component through matrix of new index, other legends are shown in Fig 6.12.*

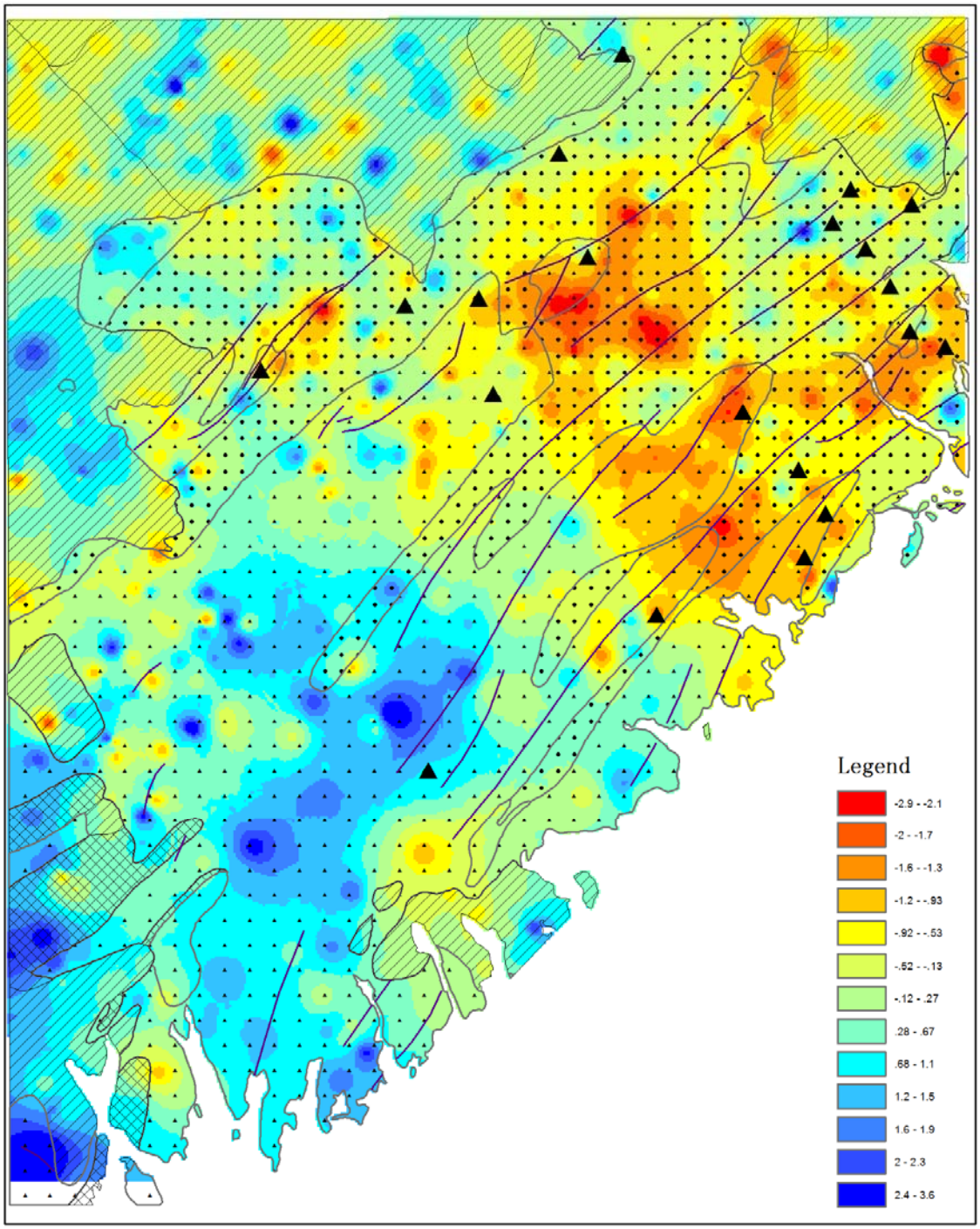


***Fig** 6.11F The second component through matrix of standardized new index, other legends are shown in Fig 6.12.*

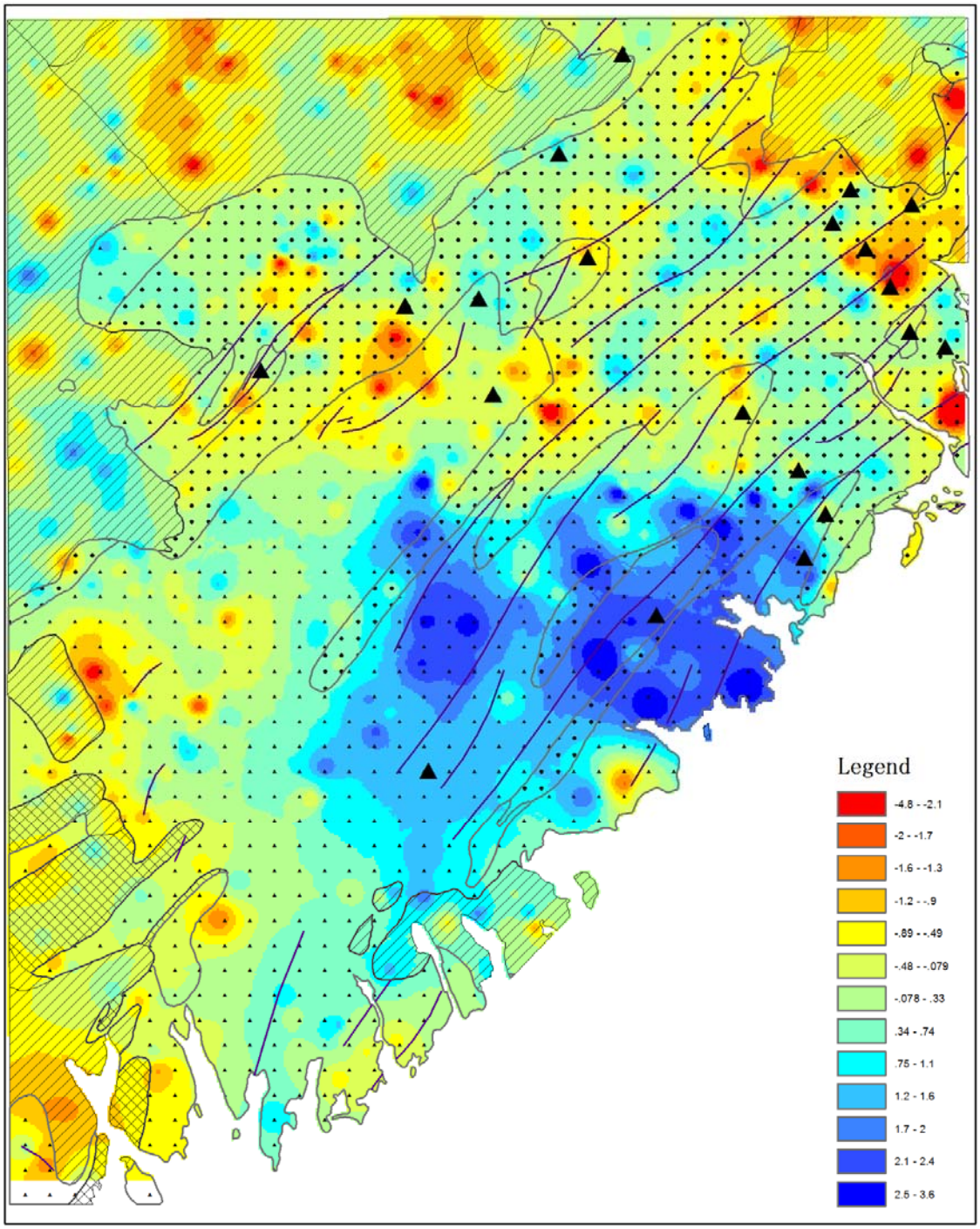


***Fig** 6.11G The third component through matrix of correlation coefficient, other legends are shown in Fig 6.12.*

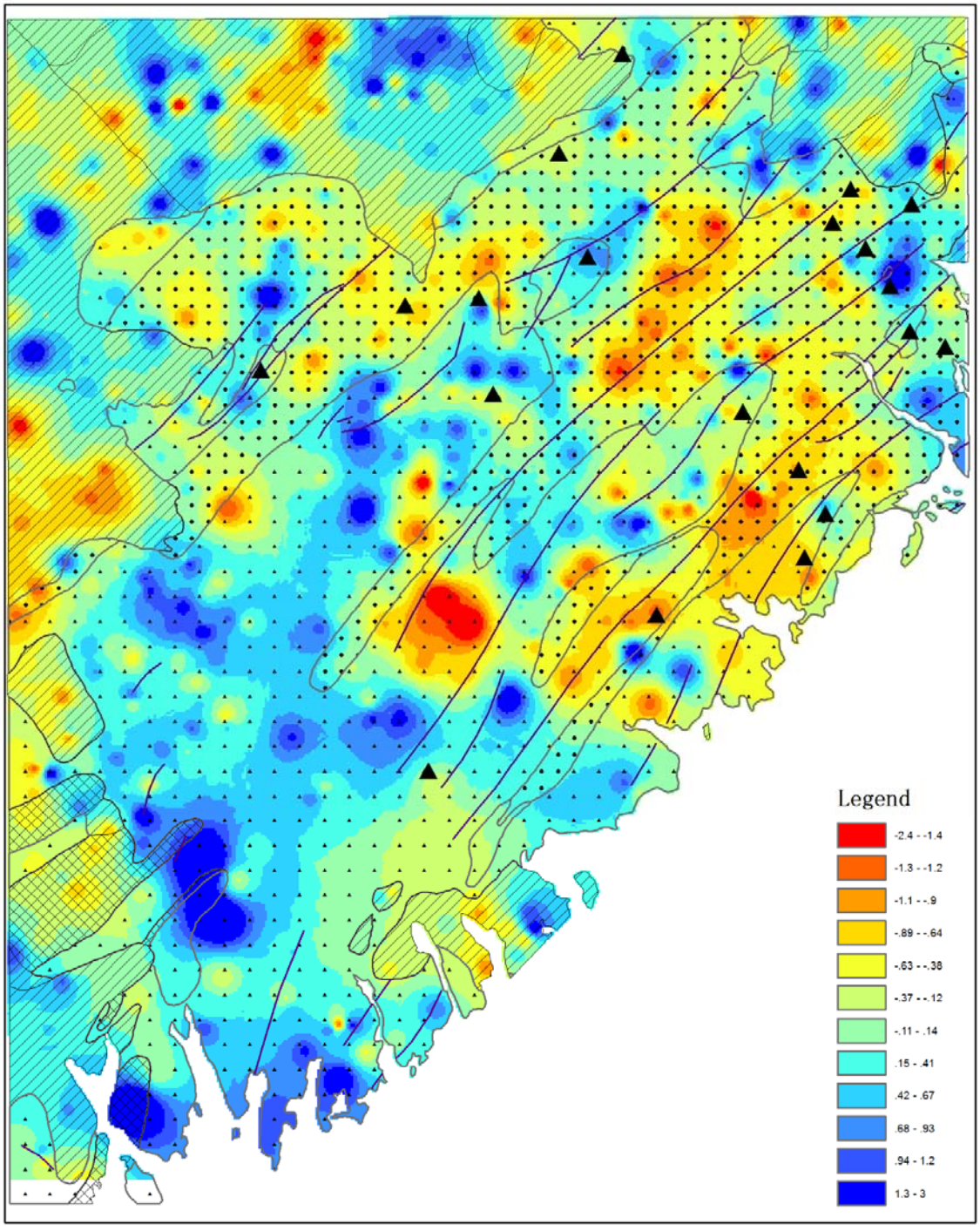


***Fig** 6.11H The third component through matrix of new index, other legends are shown in Fig 6.12.*

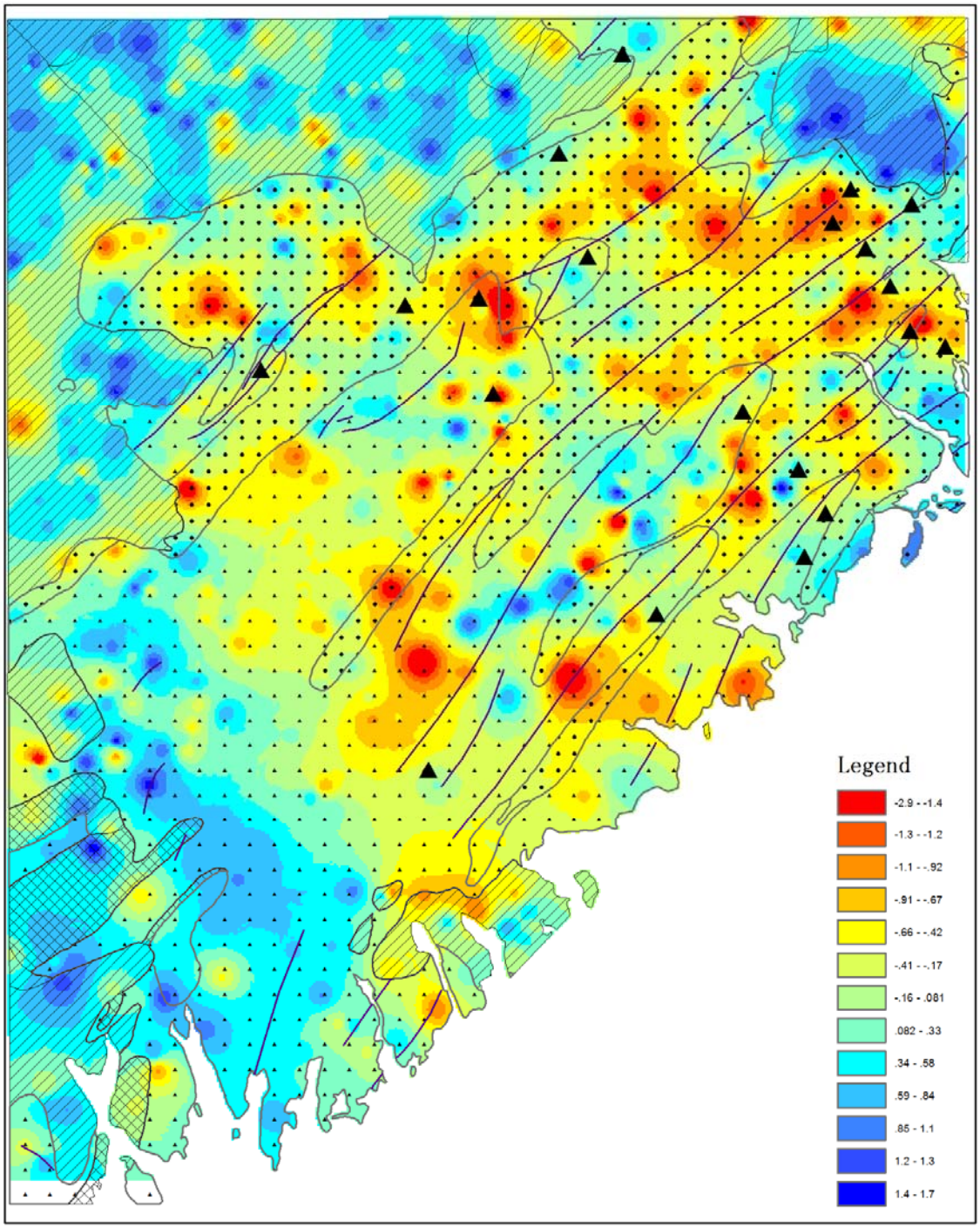


***Fig** 6.11I The third component through matrix of standardized new index, other legends are shown in Fig 6.12.*

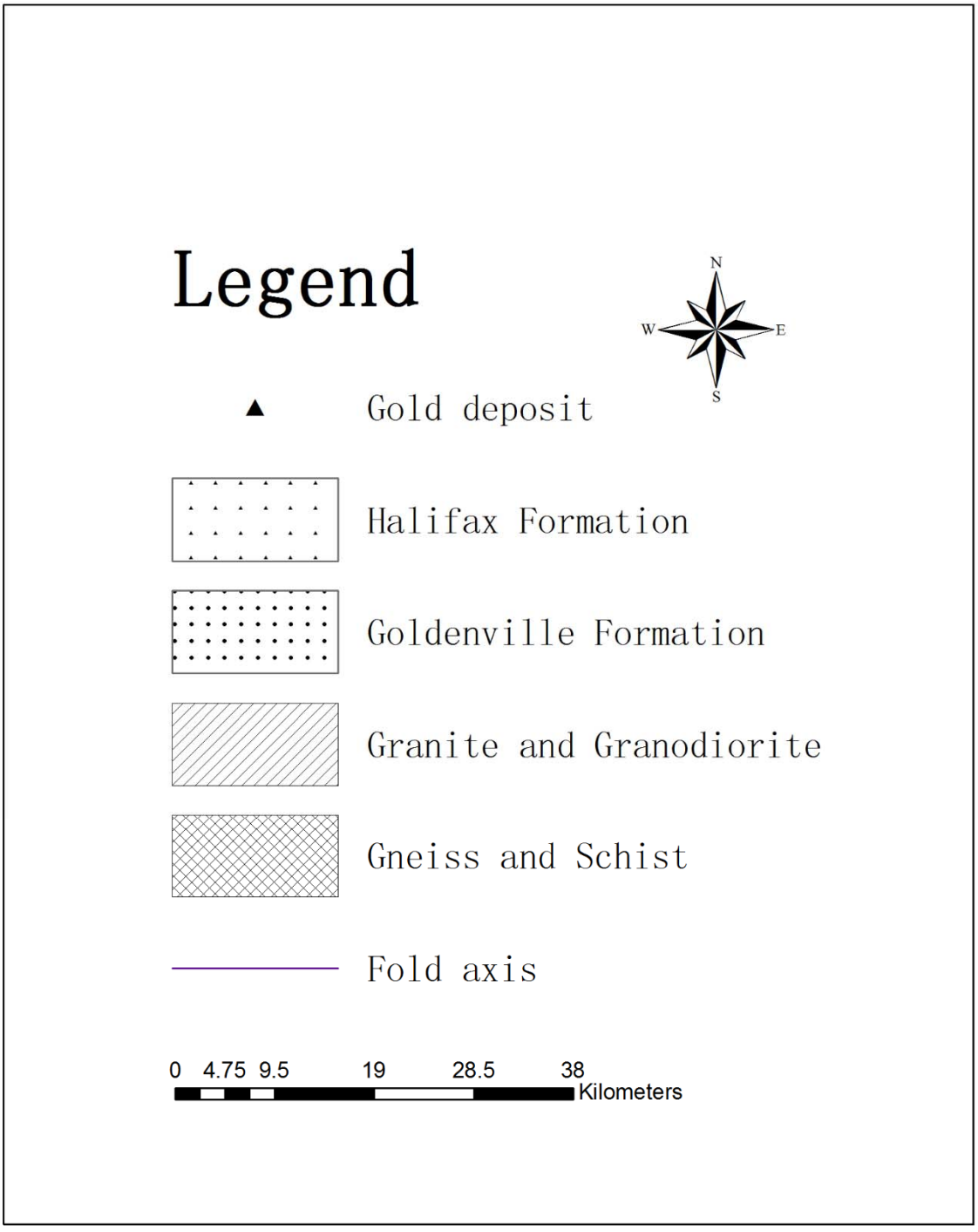
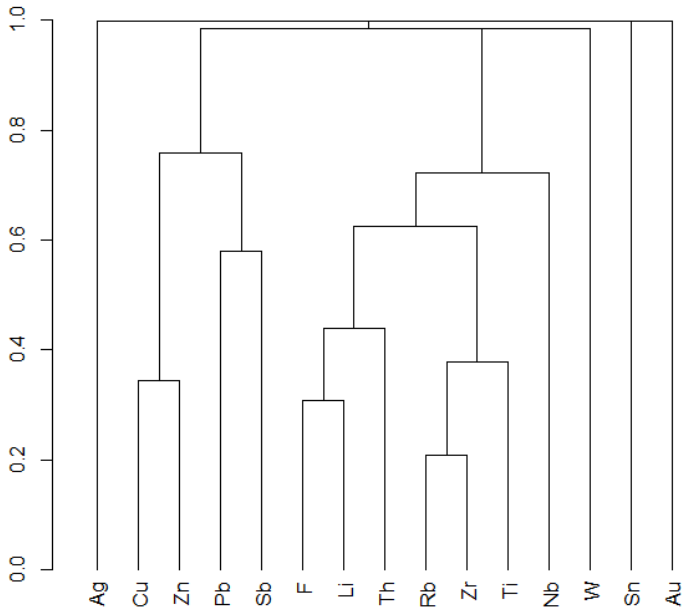
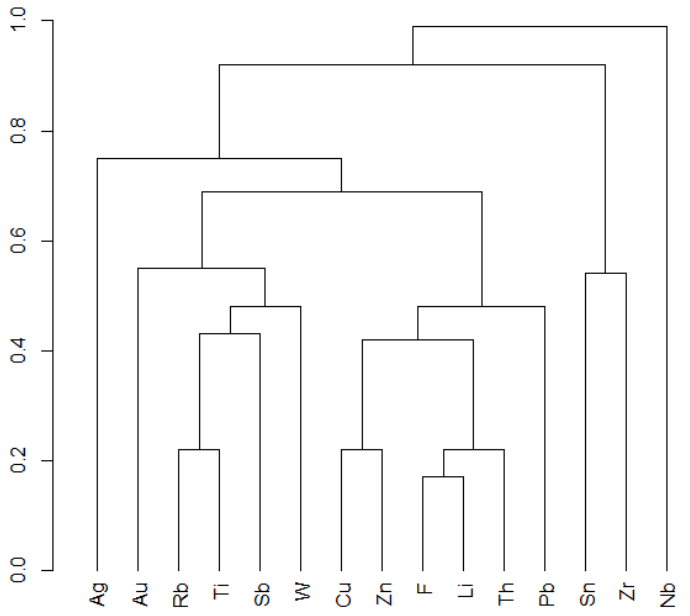


***Fig** 6.12 Legend, north arrow and scale bar for Figs 6.11A-I*



(A)



(B)

Fig *6.13 Hierarchical clustering (DIANA method) results for 15 geochemical elements through (A) based on correlation coefficient matrix (B) based on standardized new index matrix.*

From ***Fig*** *6.9*, the first component from the correlation coefficient included most of the variance existing in the 15 geochemical elements and had a strong relationship with most of the elements. It had a strong relationship with ***As*** (***Fig*** *6.10*). However, the correlation coefficient (0.499) between the first component and ***As*** was at the same level as the one (0.445) between the second component and ***As***, which confirmed that the components derived based on decomposition of the correlation coefficients matrix were ranked by the total variance of the 15 elements, instead of the correlation coefficient with ***As***. On the other hand, from the first five main components which showed some degree of the correlation with ***As*** in the ***Fig*** *6.10*, we observe that not only the first components from the new index and the standardized new index showed the higher correlation coefficients with ***As*** in comparison with the first component obtained based on ordinary correlation coefficient matrix but also the first five components showed descending order on correlation with ***As***. This is reasonable because the components were ordered by the variance based on the correlation coefficient matrix given of ***As***.

The top three components from the new index and the standardized matrix contain similar information both in their loadings on 15 elements and their relationships with ***As*** (***Fig*** *6.10*). Thus, it can be concluded that: i). the standardization did not change the basic attribute of the matrix and ii). the relationship between the new and the standardized indexes is similar as which between covariance and correlation coefficient. While the new index of two elements is a measure of their association with ***As*** explained by both of them, the standardized new index of two elements is the portion of the new index in their multiple correlation coefficient

with ***As***.

The spatial distribution of the top three components in ***Fig*** *6.11* showed that the first component obtained from the three matrixes are similar. The spatial distributions of the second component using the new index and standardized new index are similar. However, they are significantly different when using the correlation coefficient. Spatial distributions of the third component are different in either case.

Two centers are evident in ***Fig*** *6.13A*: {***Cu***, ***Zn***, ***Pb***, ***Sb***} and {***F***, ***Li***, ***Th***, ***Rb***, ***Zr***, ***Ti***, ***Nb***}. It is found that the dendrogram structures are controlled by several strong relationships as outlined in ***Table*** *6.2*, i.e., ***Rb - Zr***, ***F - Li***, and ***Cu - Zn***. However, in this result, the effects of some ***As***-related elements are weakly represented. For example, ***Au***, ***W***, ***Ag*** are left out of the previous clusters and far away from each other.

Three centers are evident in ***Fig*** *6.13B*: {***Au***, ***Rb***, ***Ti***, ***Sb***, ***W***}, {***Cu***, ***Zn***, ***F***, ***Li***, ***Th***, ***Pb***} and {***Sn***, ***Zr***}. The dendogram structures are controlled by several relationships in in ***Table*** *6.3*, i.e., ***Rb - Ti***, ***F - Li***, and ***Cu - Zn***. The strong relationship of ***Rb*** with ***Zr*** in ***Fig*** *6.13A* is absent because of the weak association of ***Zr*** with ***As***. The importance of the ***As***-related elements is improved against which in ***Fig*** *6.13A*, e.g., ***Au, Sb*** and ***W***.

6.5 New index for log-ratio transformed data

Table *6.6* and *6.7* are the correlation coefficient matrix and the standardized new index matrix

calculated through the centralized log-ratio (CLR) transformed data, respectively. The response variable for ***Table*** *6.7* is still ***As***. In ***Table*** *6.6*, ***Zr*** is the element with the strongest correlation with others. The top 4 large correlations of ***Zr*** are 0.6 (***Rb***), -0.53 (***Zn***), -0.50 (***As***), 0.43 (***Cu***). In addition, ***Cu, As, Rb, Li*** have relatively strong correlation with others, i.e. ***Cu~Rb*** (0.56), ***Cu~Zn*** (0.44), ***Li~Rb*** (0.53), ***As~Zn*** (0.38). In the ***Table*** *6.7*, ***Rb***, ***Li***, ***Zr*** are still important elements in terms of the standardized new index, it is the same as their functions in ***Table*** *6.6*, but the functions of ***Au***, ***Ag***, ***Cu*** are improved against which in ***Table*** *6.6* because they have good relationship with response variable ***As***.

Fig *6.14A* and *6.14B* are the hierarchical clustering results for 15 geochemical elements (except ***As***) based on the matrices in ***Table*** *6.6* and *6.7* through DINIA algorithm, respectively. Four centroids can be found in ***Fig*** *6.14A:* {***Au***, ***Li***, ***Ag***}, {***Cu***, ***Rb***, ***Zn***, ***Zr***, ***Ti***}, {***Nb***, ***Sn***}, {***Pb***, ***Sb***, ***Th***, ***W***}. The clustering results are dominated by the strong relationship shows in ***Table*** *6.6*, i.e. ***Cu~Rb, Rb ~Zn***. ***F*** has a relative weak relationship with others in the current result. The dissimilarity in ***Fig*** *6.14B* is more notable than that in ***Fig*** *6.14A*. There are two centroids existed in ***Fig*** *6.14B*, which were dominated by {***Au***, ***Li***, ***Th***} and {***Zr***, ***Zn***, ***Rb***}, respectively.

***Table** 6.6 Correlation coefficient matrix for clr transformed data*

	Ag	***As***	***Au***	***Cu***	***F***	***Li***	***Nb***	***Pb***	***Rb***	***Sb***	***Sn***	***Th***	***Ti***	***W***	***Zn***	***Zr***
Ag	1.00															
As	-0.08	1.00														
Au	0.14	0.09	1.00													
Cu	0.09	0.20	0.01	1.00												
F	-0.10	-0.06	-0.17	-0.04	1.00											
Li	-0.33	-0.07	-0.41	-0.26	0.23	1.00										
Nb	-0.13	-0.37	-0.09	-0.19	-0.09	-0.01	1.00									
Pb	-0.03	-0.06	-0.11	-0.07	-0.13	-0.11	-0.15	1.00								
Rb	-0.30	-0.33	-0.26	-0.56	0.17	0.53	0.05	-0.03	1.00							
Sb	-0.07	-0.08	-0.11	0.09	-0.22	-0.20	0.00	0.18	-0.25	1.00						
Sn	0.09	-0.22	0.06	-0.13	-0.07	-0.19	-0.20	-0.09	0.04	-0.14	1.00					
Th	-0.16	-0.07	-0.23	0.06	0.03	0.12	-0.03	-0.08	0.15	-0.09	-0.16	1.00				
Ti	-0.15	-0.40	-0.08	-0.14	-0.09	0.02	0.30	-0.11	0.17	-0.07	-0.03	-0.01	1.00			
W	-0.02	-0.02	0.14	-0.01	0.00	-0.25	-0.07	-0.17	-0.14	-0.16	0.08	-0.19	-0.08	1.00		
Zn	-0.09	0.38	-0.26	0.44	0.10	0.15	-0.31	0.03	-0.37	0.07	-0.20	0.00	-0.31	-0.12	1.00	
Zr	-0.14	-0.50	-0.08	-0.43	-0.11	0.09	0.12	0.01	0.60	-0.19	0.14	0.22	0.49	-0.13	-0.53	1.00

Table *6.7 Standardized new index $R_y(x_1,x_2)$ under the restriction of As for clr transformed data*

	Ag	***As***	***Au***	***Cu***	***F***	***Li***	***Nb***	***Pb***	***Rb***	***Sb***	***Sn***	***Th***	***Ti***	***W***	***Zn***	***Zr***
Ag	1.00															
As	0.08	1.00														
Au	0.43	0.09	1.00													
Cu	0.33	0.20	0.39	1.00												
F	0.43	0.06	0.52	0.28	1.00											
Li	0.33	0.07	0.65	0.34	0.60	1.00										
Nb	0.20	0.37	0.25	0.47	0.15	0.18	1.00									
Pb	0.45	0.06	0.48	0.27	0.43	0.44	0.14	1.00								
Rb	0.18	0.33	0.28	0.60	0.18	0.18	0.51	0.16	1.00							
Sb	0.46	0.08	0.54	0.31	0.38	0.40	0.20	0.56	0.18	1.00						
Sn	0.35	0.22	0.34	0.56	0.24	0.25	0.34	0.23	0.47	0.28	1.00					
Th	0.42	0.07	0.60	0.31	0.50	0.56	0.19	0.45	0.22	0.45	0.27	1.00				
Ti	0.18	0.40	0.23	0.44	0.14	0.17	0.64	0.13	0.56	0.18	0.40	0.18	1.00			
W	0.27	0.02	0.22	0.12	0.35	0.25	0.06	0.31	0.07	0.26	0.11	0.26	0.06	1.00		
Zn	0.22	0.38	0.19	0.52	0.14	0.16	0.64	0.15	0.67	0.18	0.50	0.19	0.64	0.06	1.00	
Zr	0.15	0.50	0.19	0.40	0.11	0.14	0.51	0.11	0.66	0.13	0.40	0.15	0.70	0.05	0.71	1.00

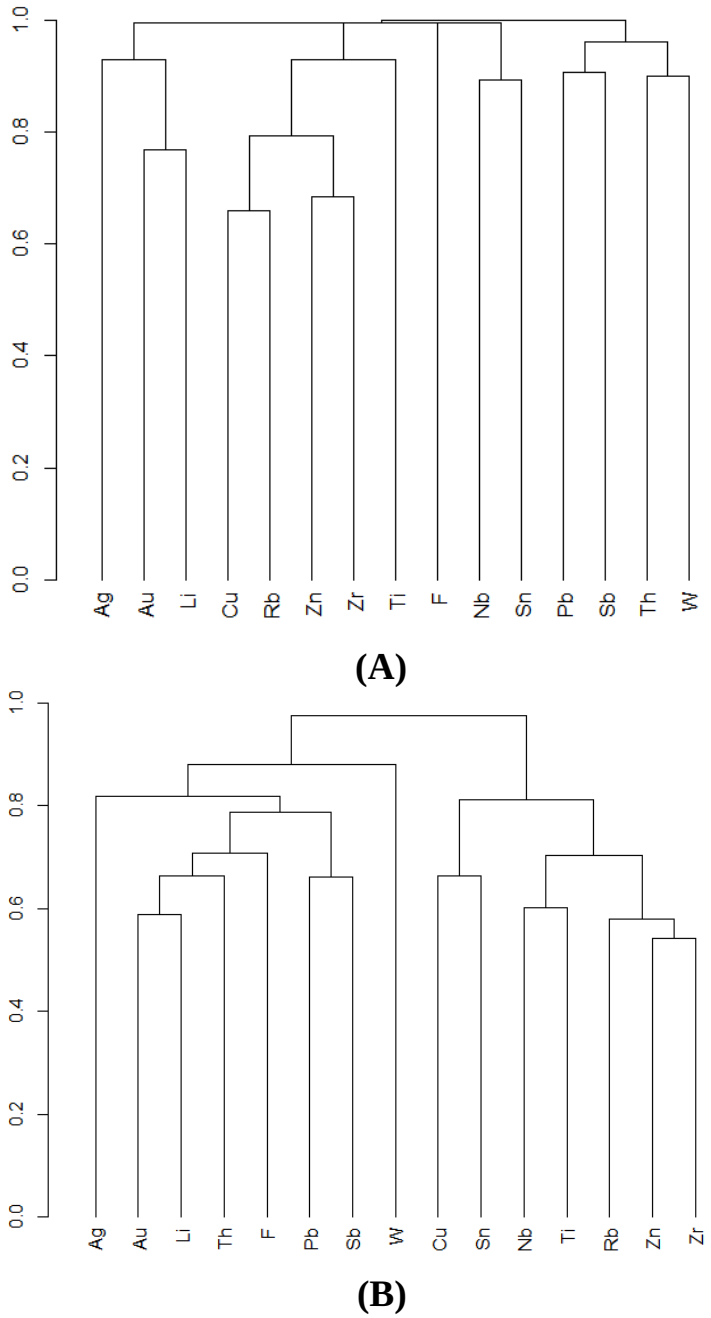


Fig *6.14 Hierarchical clustering (DIANA) results for 15 geochemical elements through (A) based on correlation coefficient matrix (B) based on standardized new index matrix, which data was transformed through clr method.*

6.6 Discussion and conclusions

A new conditional correlation coefficient is defined and calculated through a SEM to describe the relationships between a set of two independent variables and one response variable for measuring the conditional association of two variables when they are applied in a regression as independent variables. The numerical value of the new index is equivalent to the correlation coefficient of a latent variable, a common factor of the two independent variables, with the response variable.

Some properties of the new index were mathematically discussed and validated by case study through the classification of a geochemical dataset including 15 elements (***Cu***, ***Pb***, ***Zn***, ***Ag***, ***F***, ***Li***, ***Nb***, ***Sn***, ***Zr***, ***Ti***, ***Au***, ***Sb***, ***As***, ***Th***, ***W***) constrained by response variable, ***As***. Further the new index was used to form a conditional correlation matrix which was decomposed by PCA except the components were ranked by the overall conditional correlations among 15 elements with ***As*** as response variable. In addition, the comparative clustering results based on correlation coefficient index and the new index showed that the classification enhanced the groups of ***As***-associated elements in For example, ***Au***, ***W***, ***Sb*** are classified in a new group which may represent an important factor indicating mineralization in the study area. This group was not identified in using the ordinary classification method. More application and validation of the new index for restricted classification will be introduced in Chapter 7.

Chapter 7 Response variable constrained clustering algorithm base on the new index

7.1 Introduction

In this chapter, a constrained variable clustering method, which includes hierarchical and partial clustering, is proposed based on the new index and random sampling technology. There are two differences between the new method and traditional unsupervised variable clustering. Firstly, the distance of two variables is defined as the new index proposed in Chapter 6, rather than the covariance among the variables themselves. Secondly, the centroid of each cluster is a prediction for the response variable obtained from independent variables in that cluster rather than the first principal component.

A case study is introduced for validation purpose through the same dataset used in Chapter 6. Two hierarchical clustering dendrograms based on correlation coefficient matrix and the new index matrix will be calculated. Furthermore, three clusters through the proposed clustering method were extracted from the 15 geochemical elements using the new clustering method. This new method was applied to identify the associated factors of gold mineralization in Southern Nova Scotia, Canada.

7.2 Methods

7.2.1 Clustering of variables around latent variables (CLV)

Consider a data matrix X of n observations (samples) evaluated using p quantitative predictors (features), i.e., $X = \{x_1, \ldots, x_p\} = (x_{ij})_{n \times p}$. Let $P_K = (G_1, \ldots, G_K)$ be a partition into K clusters associated with the k components: $c_1, c_2, \ldots, c_k$, respectively. It is expected to result in the clustering solution by maximizing:

$$T = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} Cov^2(x_j, c_k) \qquad (7.1)$$

wherein $\delta_{kj} = 1$ if the j-th variable belongs to the cluster Gk and $\delta_{kj} = 0$, otherwise, $Cov(x_j, c_k)$ represents the covariance between x_j and c_k; c_k is the centroid (latent component) of the k-th cluster, usually defined as the first standardized principal component of X_k (variables belonging to the cluster G_k).

7.2.2 The constrained variable clustering based on the new index

In the new method, based on the above mentioned concepts, Then ***Eq*** *(7.1)* can then be transformed to:

$$T = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} {R_y}^2(x_j, p_k) \qquad (7.2)$$

wherein p_k is a prediction of the response variable (y) by the variables in Gk, R_y is a

conditional correlation coefficient between a latent variable and y (defined and discussed in Chapter 6), which represent the homogeneity of variables x_j and p_k in clustering.

7.2.3 Partial clustering procedures based on new index

A random sampling technique (Monte Carlo simulation) is employed to obtain the clustering solutions. It can also be considered as a type of "expectation maximization" method (Jain and Dubes, 1988). A target function is first created, and all possible classifications that are generated by a random sampling method are then tested. The classification will be set as the final result when the target function (***Eq*** *7.2*) reaches the maximum value.

The algorithm stages are as follows:

Stage 1: *Choose initial model parameters*: initial clusters are generated, the target function (T) is calculated, and then both the cluster and target functions are recorded in a dataset as the best clustering suggestion. The iteration count is set to 1.

Stage 2: *Generate new clusters*: a new cluster is generated through a random sampling process.

Stage 3: *Check duplicity*: check whether the new cluster already exists in the dataset and if the newly generated clusters are not found in the dataset, the target function T are calculated and the clusters recorded to the dataset. Otherwise, return to Stage 2.

Stage 4: *Evaluate the clusters in Stage 3*. Compare the T value in Stage 3 with the existing best suggestion. If the new T value is less than the old value (recorded in the dataset as the best suggestion), set the current clusters and T value as the best suggestion, and increase the iteration counter by 1. Check for the termination of the calculation in the current stage. If yes, terminate the calculation and output the best suggestion; if not, return to Stage 2.

The clustering process can be ended through two possible ways. The first way is to generate a sufficient number of different classifications, which may be expressed as a ratio of the total number of possible classifications. The second way is to stop sampling when the final classification is unchanged after a specified number of samplings. This number of samplings depends on the total number of possible classifications. The random sampling algorithm adopted in this research is taken from the R package *sampling* (Tillé and Matei, 2009). More details about this algorithm and package are available elsewhere (Särndal et al., 1992; Tillé and Matei, 2009). The flow chart is shown in ***Fig*** *7.1*.

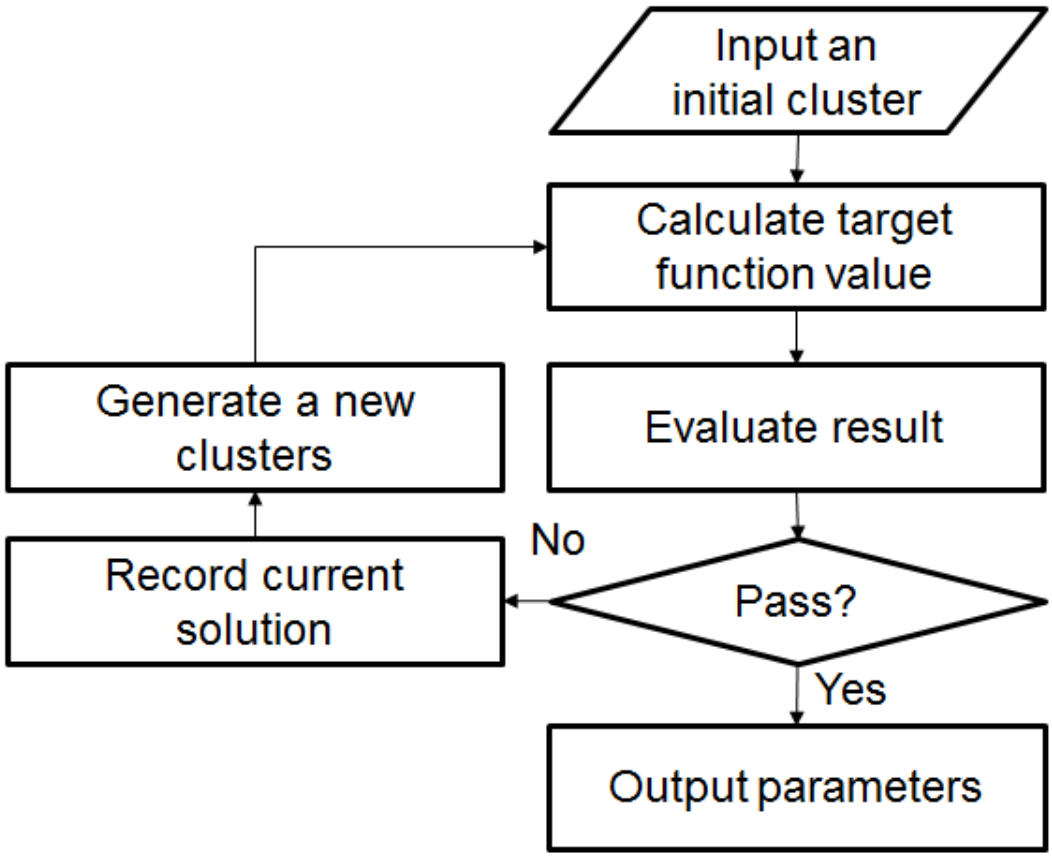


***Fig** 7.1 Flow chart of the new clustering algorithm.*

7.2.4 Hierarchical clustering algorithm based on new index

There are TWO different definitions of dissimilarity based on the new index for hierarchical clustering. The first one (***Eq** 7.3*) has been adopted in Chapter 6 (***Figs** 6. 12*, *6.13*) and the second one (***Eq** 7.4*) will be adopted in current Chapter.

1. The dissimilarity of two variables x_1, x_2 is defined as:

$$\mathrm{d}(x_1, x_2) = 1 - R_y(x_1, x_2) \quad (7.3)$$

wherein $R_y(x_1, x_2)$ is the standardized new index which defined in Chapter 6. The MIN, MAX, Group average can be used for hierarchical algorithm based on the transformed dissimilarity matrix (Maechler et al., 2005). The adopted clustering function is a divisive analysis (DIANA) of the package *cluster* in R (Kaufman and Rousseeuw, 2009).

2. Clusters A and B are chosen such that they possess the smallest dissimilarity d:

$$\mathrm{d}(A, B) = H(A) + H(B) - H(A \cup B) = \lambda_A^1 + \lambda_B^1 - \lambda_{A \cup B}^1 \quad (7.4)$$

This dissimilarity measures the loss of the homogeneity upon merging of the two clusters A and B. This algorithm comes from the concept of a hierarchical CLV proposed by Chavent et al. (2011). λ_A^1 and λ_B^1 are the first Eigen values obtained from the standardized new index matrix in clusters A and B, respectively, and not from the correlation coefficient matrix as in the CLV method.

7.3 Case study

7.3.1 Hierarchical clustering through the covariance and the new index

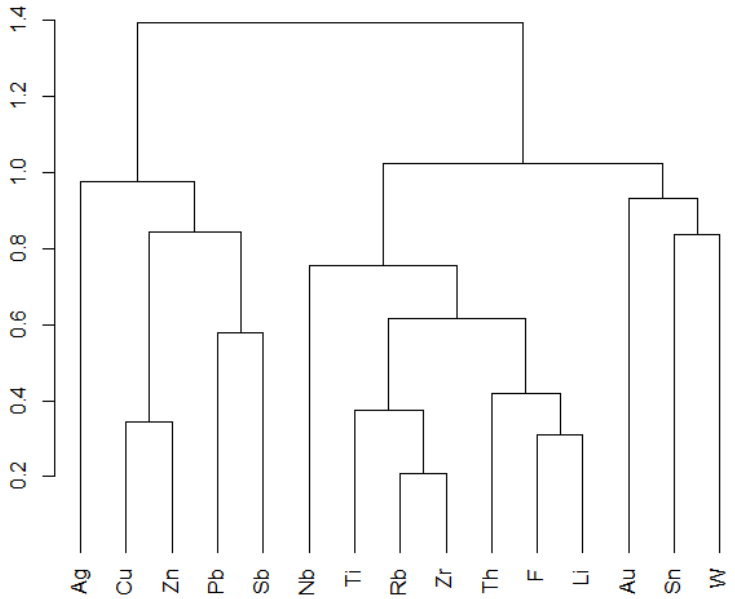
Results from the hierarchical clustering may provide an approximate evaluation of the center and number of classifications from the perspective of information compression and adjacent relationships among variables. The results of hierarchical clustering based on the dissimilarities defined in ***Eq*** *(7.3)* have been discussed in Chapter 6 already. Results based on ***Eq*** *(7.4)* are shown in ***Fig*** *7.2* of which the results in ***Fig*** *7.2A* were through correlation

coefficient matrix and the ones in ***Fig*** *7.2B* through the standardized new index matrix.

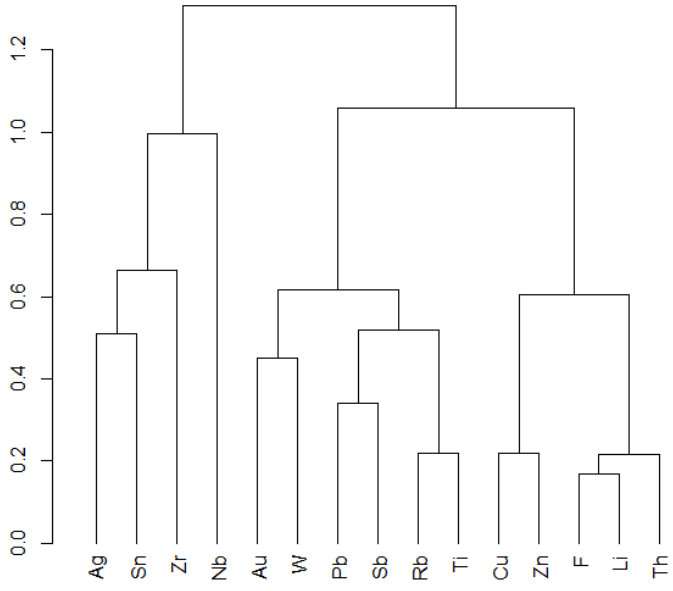
There are three clustering centers in both ***Fig*** *7.2A* ({***Ag, Cu, Zn, Pb, Sb***},{***Nb, Ti, Rb, Zr, Th, F, Li***} and {***Au, Sn, W***}) and ***Fig*** *7.2B* ({***Ag, Sn, Zr, Nb***},{***Au, W, Pb, Sb, Rb, Ti***} and {***Cu, Zn, F, Li, Th***}), which are the results based on the correlation coefficient matrix and the standardized new index matrix, respectively. Comparing with the result in ***Fig*** *6.13* of Chapter 6, the effect of ***As*** related element in result is more obvious in CLV than in DIANA. The ***As*** related elements are far away from the center in the former clusters, but close to the center in the latter, i.e. ***Ag, Au, W***.

The relationships of ***Nb*** with other elements in the new index are almost 0 because of it week relationship with ***As*** (the mean of R_{As} is 0.015), that caused the ***Nb*** is almost no effect on the clustering result in both ***Fig*** *6.13B* of Chapter 6 and ***Fig*** *7.2B* of current Chapter. For the same reason, the effect of ***Zr*** are reduced in both ***Fig*** *6.13B* of Chapter 6 and ***Fig*** *7.2B* of current Chapter.

On the basis of the previous analysis, the number of clusters for the following partial clustering will be set as 3.



(A)



(B)

Fig *7.2 Hierarchical clustering results through CLV algorithm for 15 geochemical elements (A) based on correlation coefficient matrix; (B) based on standardized new index matrix. The height of dendrogram is the dissimilarity of different clusters.*

7.3.2 Partial clustering result

Through the algorithm introduced in Section 7.2.2, the partial clustering result with the new index is (***Ag***, ***Au***, ***Nb***, ***W***), (***Cu***, ***F***, ***Li***, ***Pb***, ***Sb***, ***Th***, ***Zn***, ***Zr***) and (***Rb***, ***Sn***, ***Ti***). It should be noted that the partial clustering output depends on the initial input, therefore, this result only represent a local optimum solution.

Because the centroid of the K^{th} group is the prediction for ***As*** (p_k) through variables in the K^{th} group, the distances/loadings of an element (x_j) to the corresponding centroid (p_k) is $R_{As}(x_j, p_k)$. ***Fig*** *7.3* shows the $R_{As}(x_j, p_k)$ in each group, which is the centroid loadings on elements. The bars in blue for (***Au, W, Ag, Nb***) present the variables in group one, the grey bars for (***Zn, Cu, Th, F, Li, Pb, Sb, Zr***) present group two, and the orange bars for (***Ti, Rb, Sn***) represent group three.

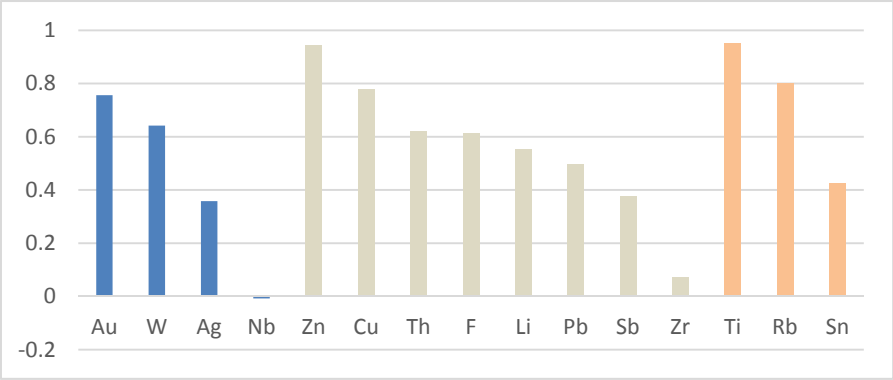


Fig *7.3 Centroid loadings in new index on elements in each cluster; blue, grey and orange bars represent group one, two and three, respectively.*

As can be seen in ***Fig*** *7.3*, ***Au***, ***W*** and ***Ag*** are the main variables in group 1, ***Zn***, ***Cu***, ***Th***, ***F***, ***Li***, ***Pb***, ***Sb*** are the main variables in group 2, and ***Ti***, ***Rb*** and ***Sn*** are the main variables in group 3. The loadings are the homogeneity criterion adopted in current clustering, their level represents the importance of each element in the cluster. From ***Fig*** *7.3* ***one*** can find that ***Au*** and ***W*** can provide the same effect for the clusters as the elements with higher correlation coefficients, i.e. ***F***, ***Li***, ***Th***.

7.3.3 Spatial distribution of cluster centroids

To compare the three clusters, their centroids are mapped (***Fig*** *7.4*) in a geographic information system (GIS) as an interpolated result of samples through the inverse distance weighting (IDW) method. Because a centroid is a prediction for ***As***, the prediction error (***Fig*** *7.5*) of each cluster was mapped at the same time. The legends (north arrow, symbols for gold mineral deposits and geological futures) in ***Figs*** *7.4* and *7.5* are same as which in in ***Fig*** 6.12.

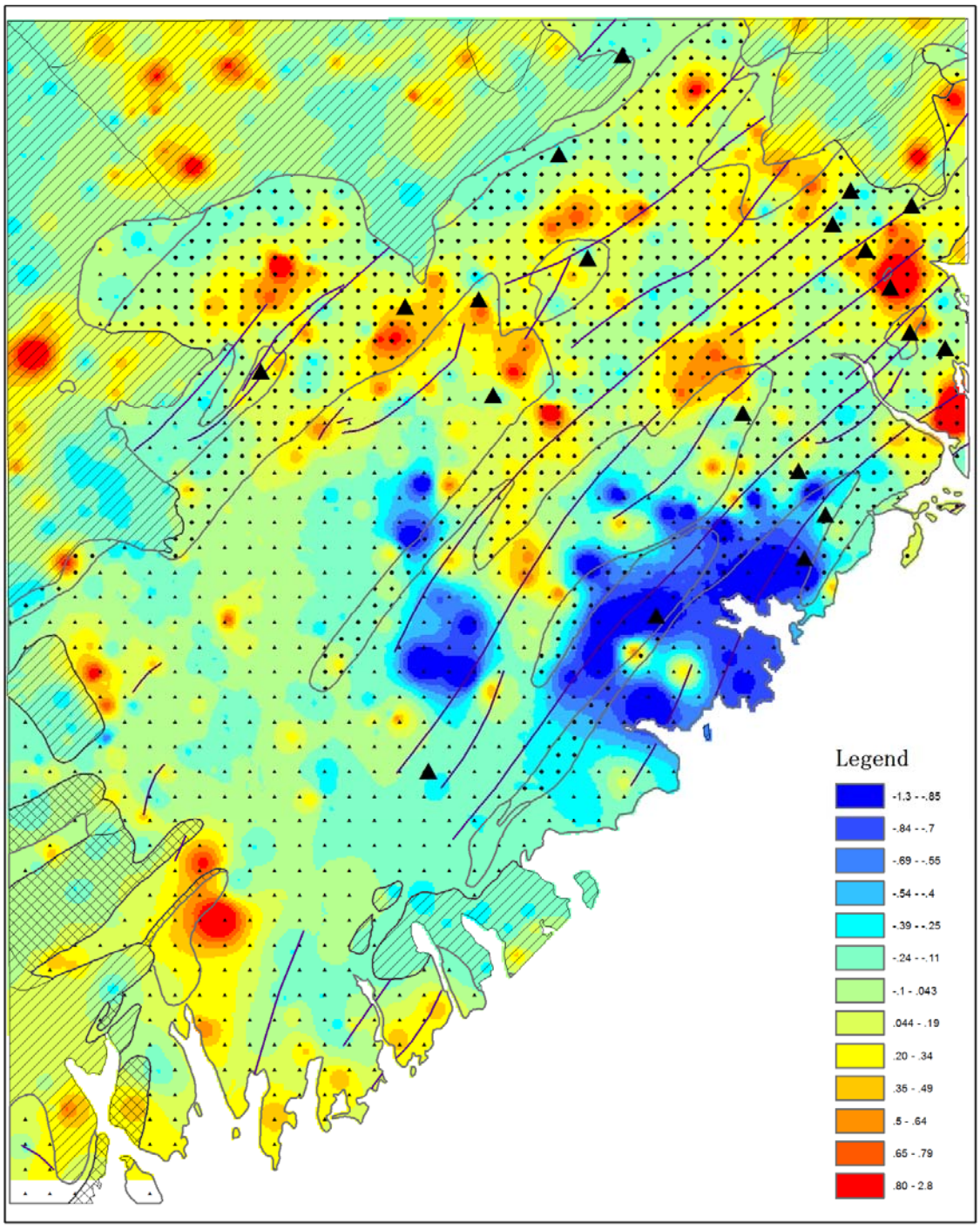


***Fig** 7.4A The centroid of group 1, which is interpreted from 624 samples (total is 671) through the IDW method, the other legends are shown in Fig 6.12.*

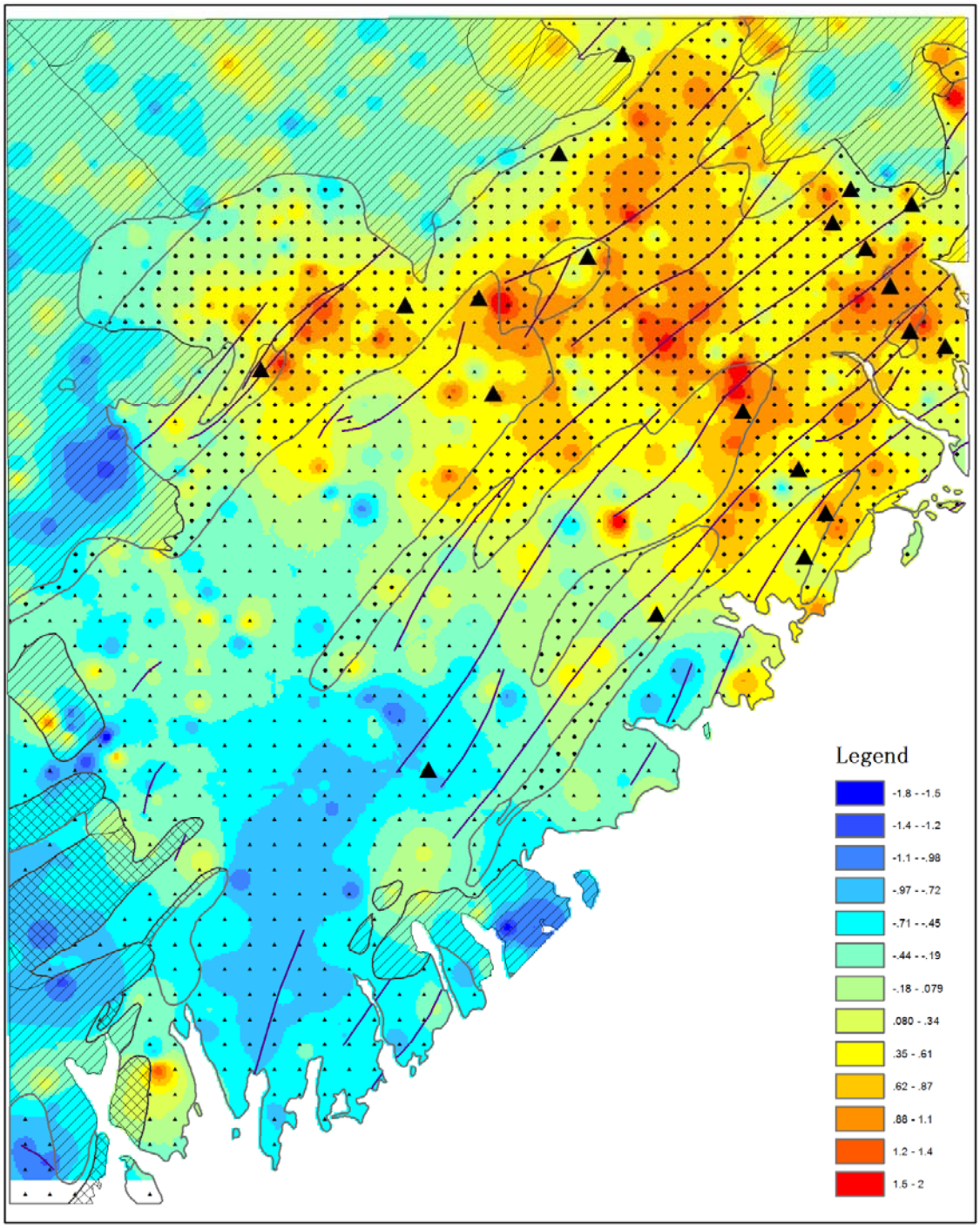


***Fig** 7.4B The centroid of group 2, which is interpreted from 624 samples (total is 671) through the IDW method, the other legends are shown in Fig 6.12*

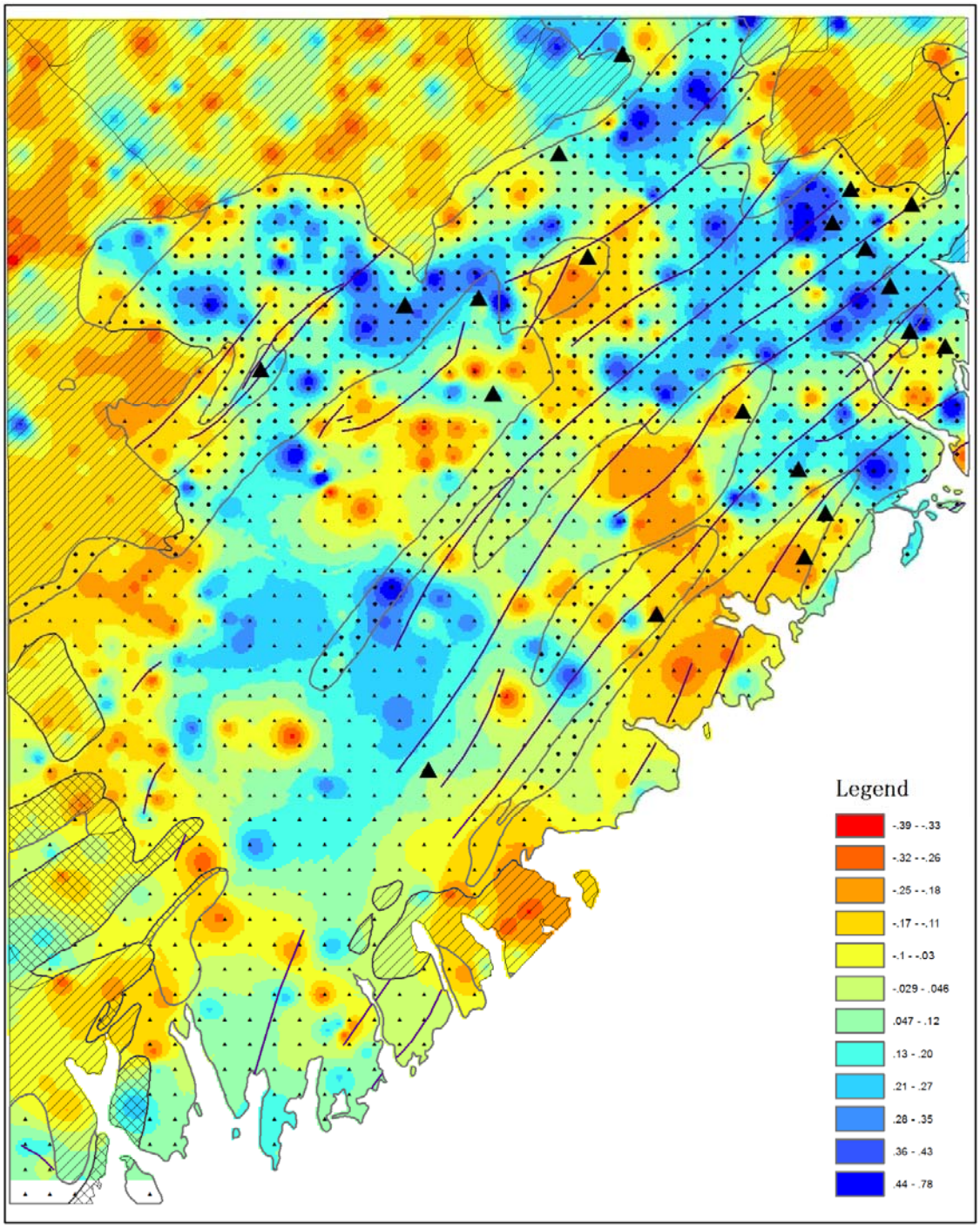


Fig 7.4C The centroid of group 3, which is interpreted from 624 samples (total is 671) through the IDW method, the other legends are shown in Fig 6.12.

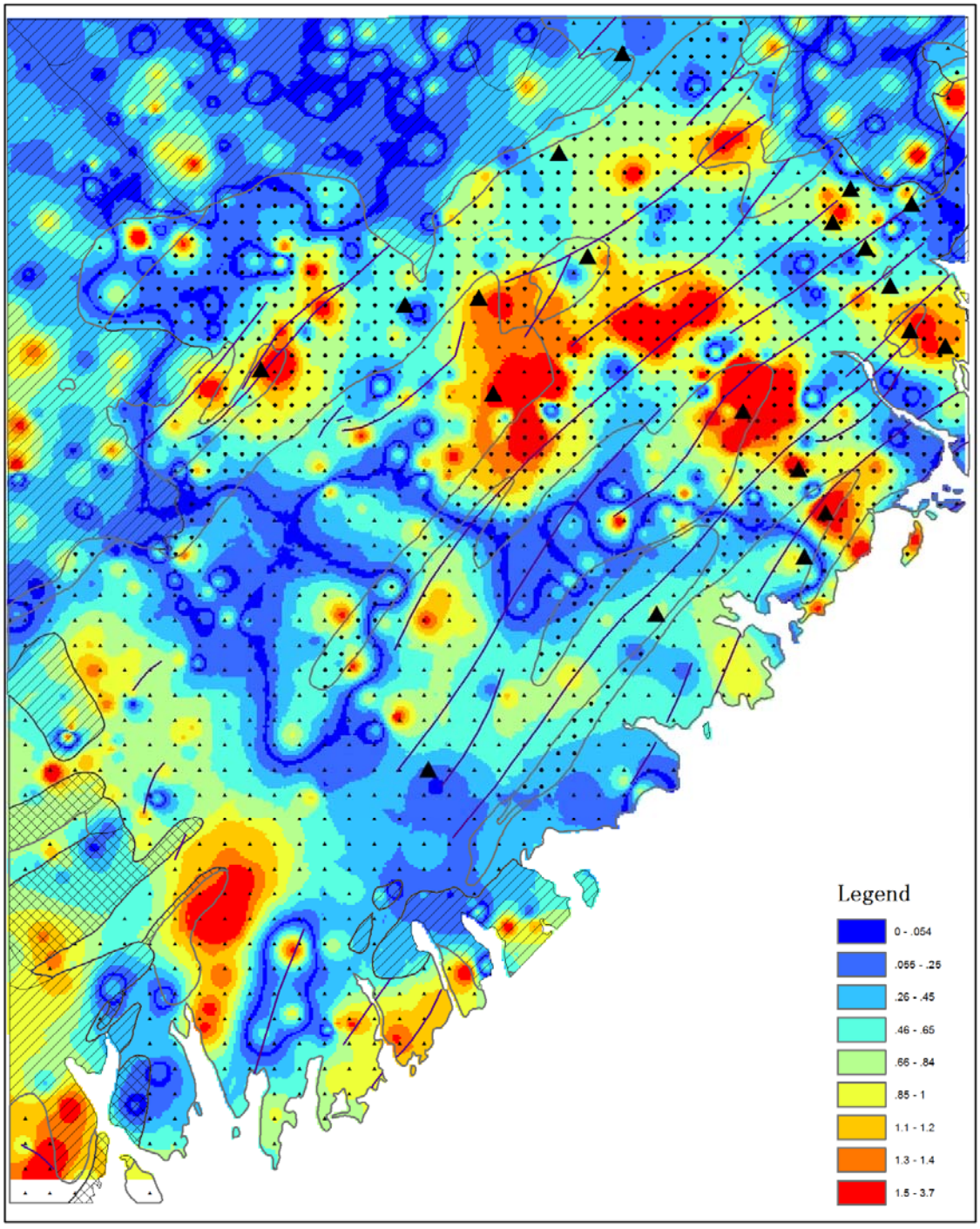


***Fig** 7.5A Prediction error in groups 1, which is interpreted from 624 samples (total is 671) through the IDW method, the other legends are shown in Fig 6.12*

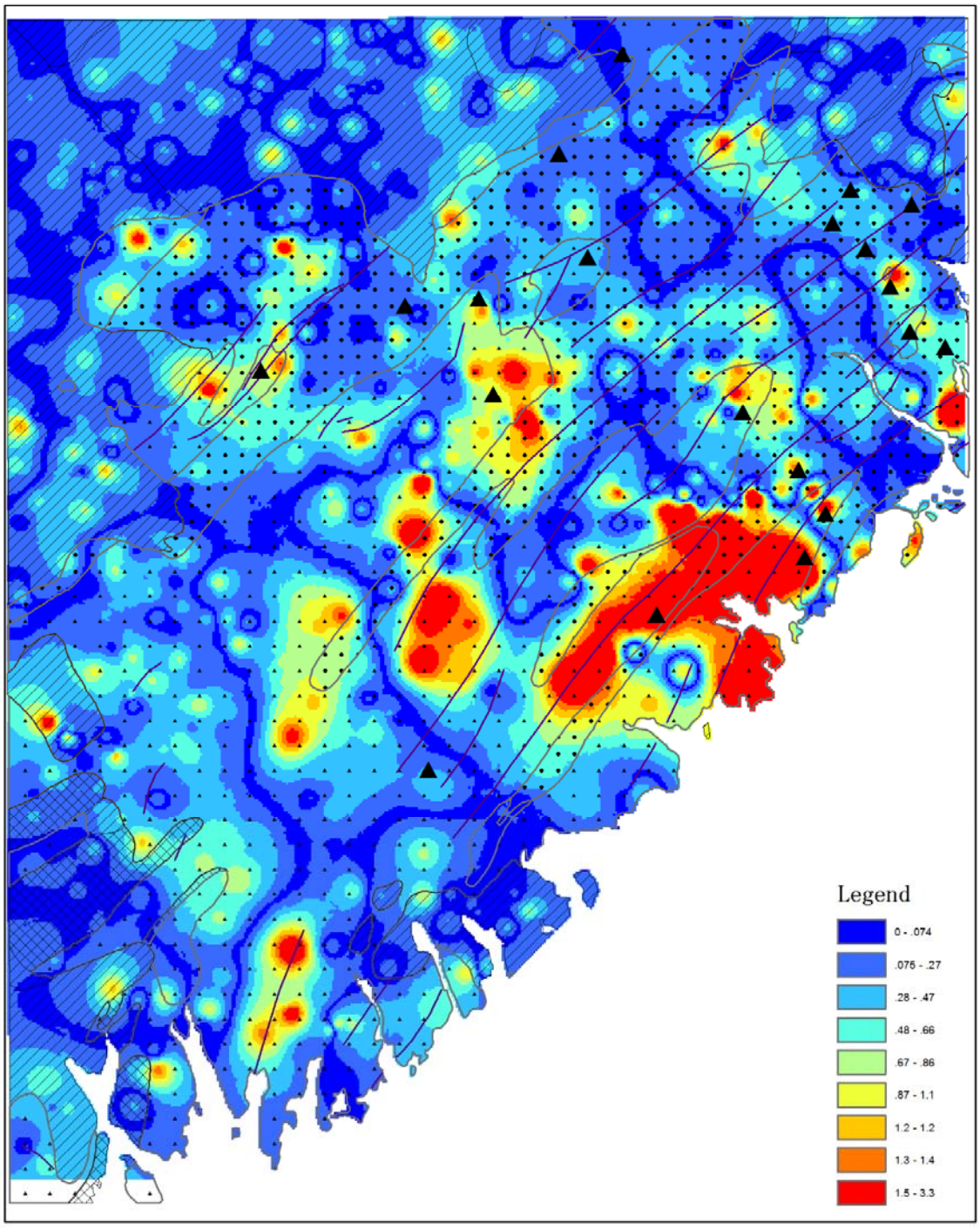


***Fig** 7.5B Prediction error in groups 2, which is interpreted from 624 samples (total is 671) through the IDW method, the other legends are shown in Fig 6.12.*

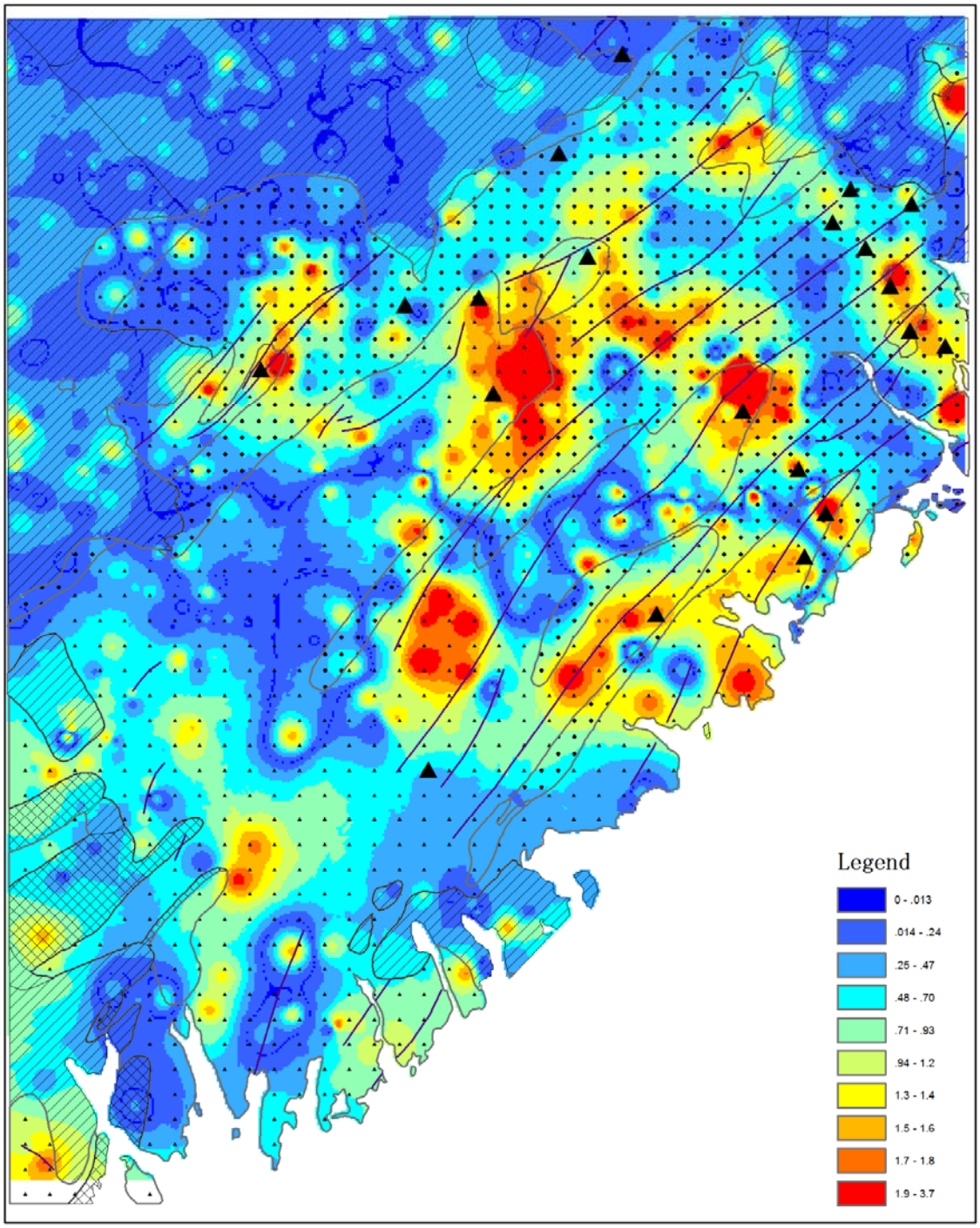


Fig *7.5C Prediction error in groups 2, which is interpreted from 624 samples (total is 671) through the IDW method , the other legends are shown in Fig 6.12.*

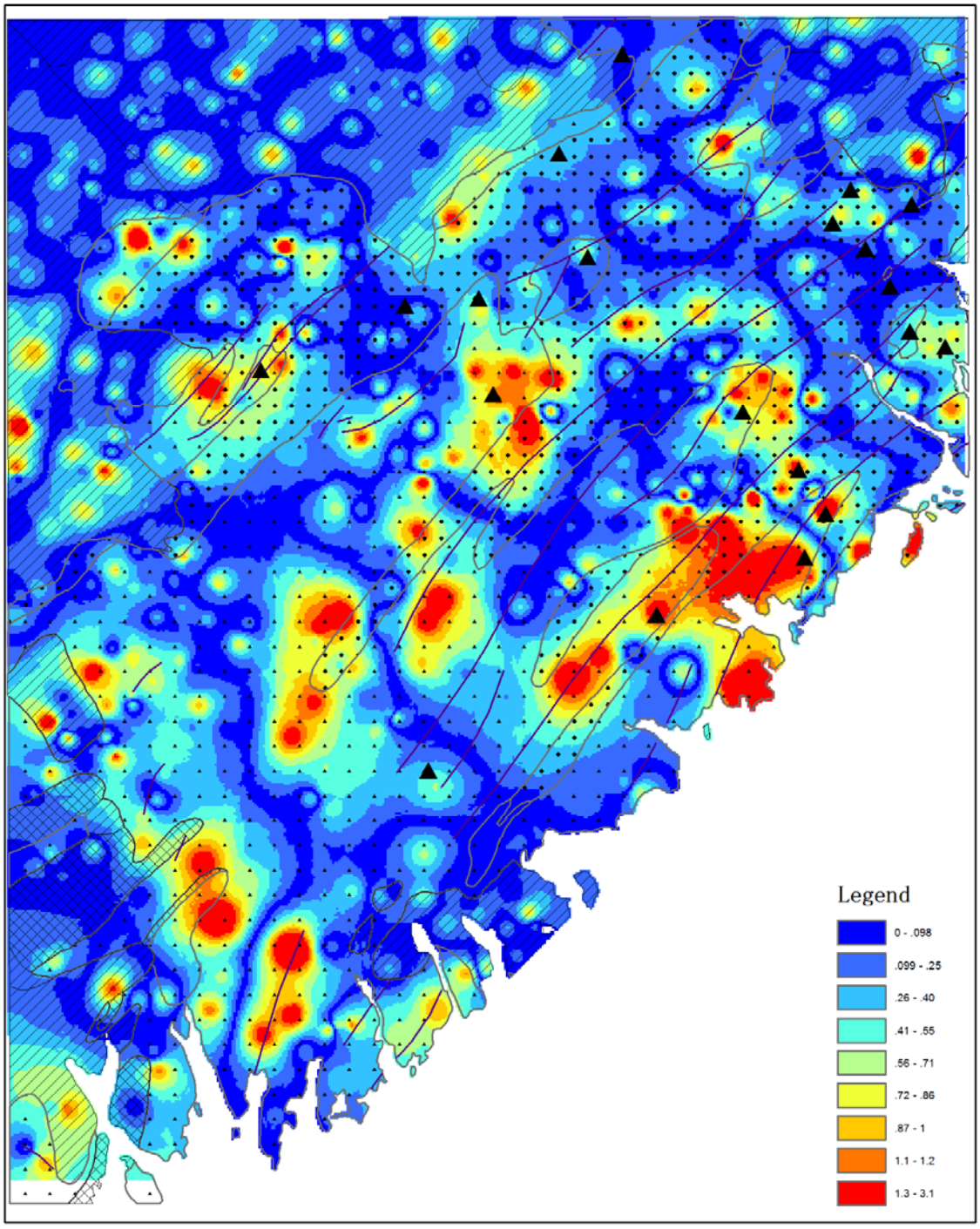


Fig *7.5D Prediction error through all elements, which is interpreted from 624 samples (total is 671) through the IDW method, the other legends are shown in Fig 6.12.*

The clustering result is calculated with respect to the response element ***As***, and the centroids of the clusters represent different factors related to the response variable. In the spatial distribution, these factors should relate to some geologic features. The centroids of clusters 1, 2 and 3 in ***Fig*** *7.4* are named as p_1, p_2 and p_3, respectively, and their prediction errors in ***Fig*** *7.5* are named as e_1, e_2 and e_3, respectively. As shown in ***Fig*** *7.4*, the patterns of p_1, p_2 and p_3 are spatially different. Among them, p_2 includes most of the information about ***As***, Therefore its pattern is similar to that of ***As***, with most of the part in red located within the Goldenville formation. The red part of p_3 relates to granite and granodiorite, and p_1 is related to the boundaries among the Halifax formation, Goldenville formation, and granite and granodiorite. In ***Fig*** *7.5*, the prediction error goes from low to high as color changes from blue to red. Most of the errors in e_1 are observed to be located within the Goldenville formation and the southwest area, but the errors around the boundaries are small; e_2 shows that the prediction of p_2 is good in most of the areas except at the center of the map; and e_3 shows that most of the area with good prediction of p_3 is located within the granite and granodiorite.

7.3.4 The first main component calculated through standardize new index matrix

The centroid from the current clustering method is represented as the prediction for the response element ***As***, which could be considered as the first main component decomposed (Eigen value decomposition) from the standardized new index matrix. In order to validate

this hypothesis, the first main component through the standardized new index matrix is mapped in ***Fig*** *7.6* (north arrow, symbols for gold mineral deposits and geological futures are same as which in ***Fig*** *6.12*). The associations between the predictions for ***As*** from the elements in each group and the first component are plotted in ***Fig*** *7.7*. From the maps, the spatial distribution of the first components are similar as the distribution of predictions. Three regressions are created in plot maps, the corresponding R^2 is 0.95, 0.74 and 0.91. That means that the two types of scores are highly related statistically.

Although the first component of each cluster is strongly related with the prediction of the response variable (***As***) from variables in each cluster, the latter is suggested as the centroid of each cluster rather than the former in the proposed clustering method because of the following two reasons:

1. The information which is related to the response variable included in the first component is less than the one in the prediction.
2. The computational complexity of the algorithm based on the prediction for response variable is less than the one based on the first component.

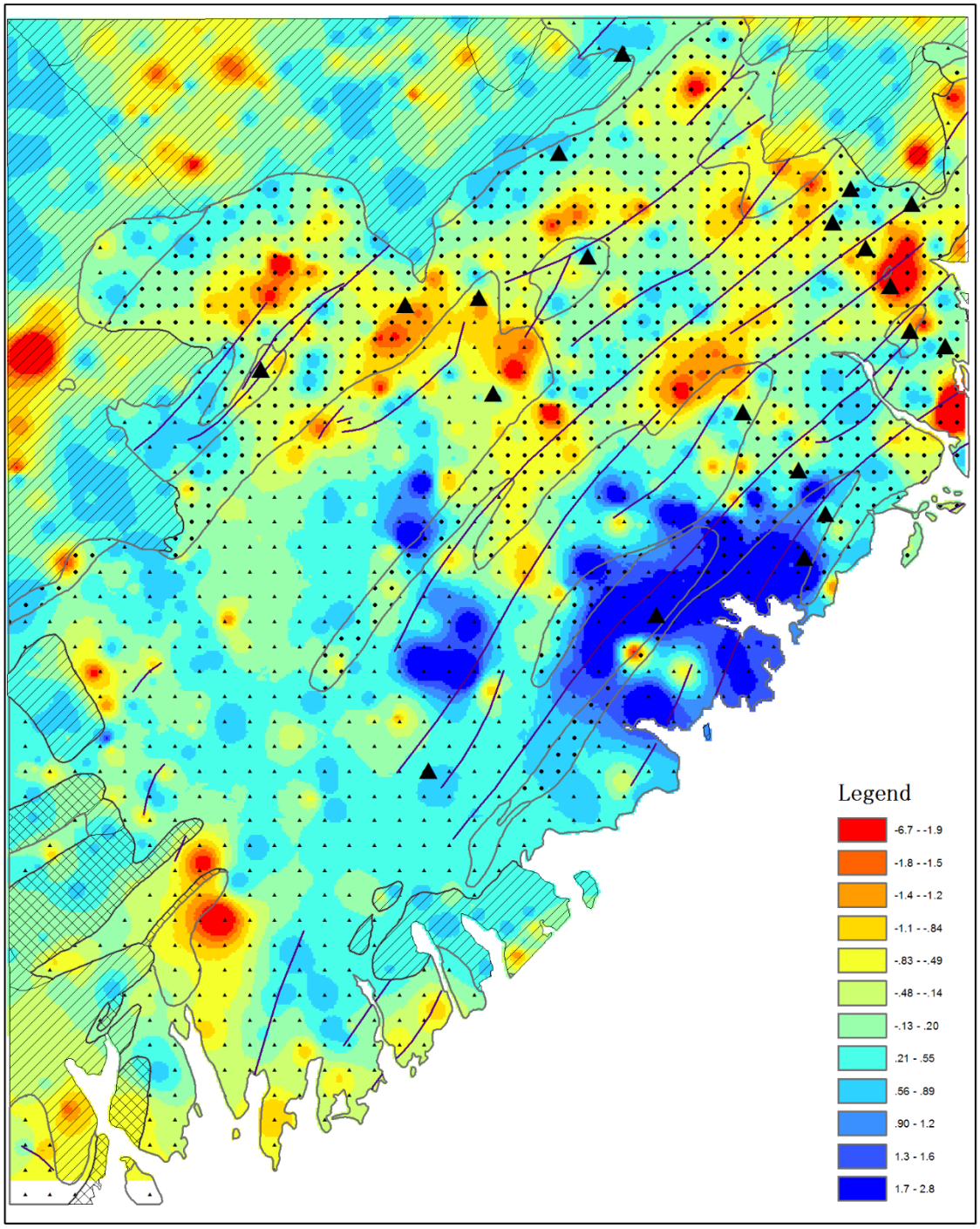


***Fig** 7.6A The first component of group 1 (Fig 7.3) calculated through the standardized new index matrix, the other legends are shown in Fig 6.12.*

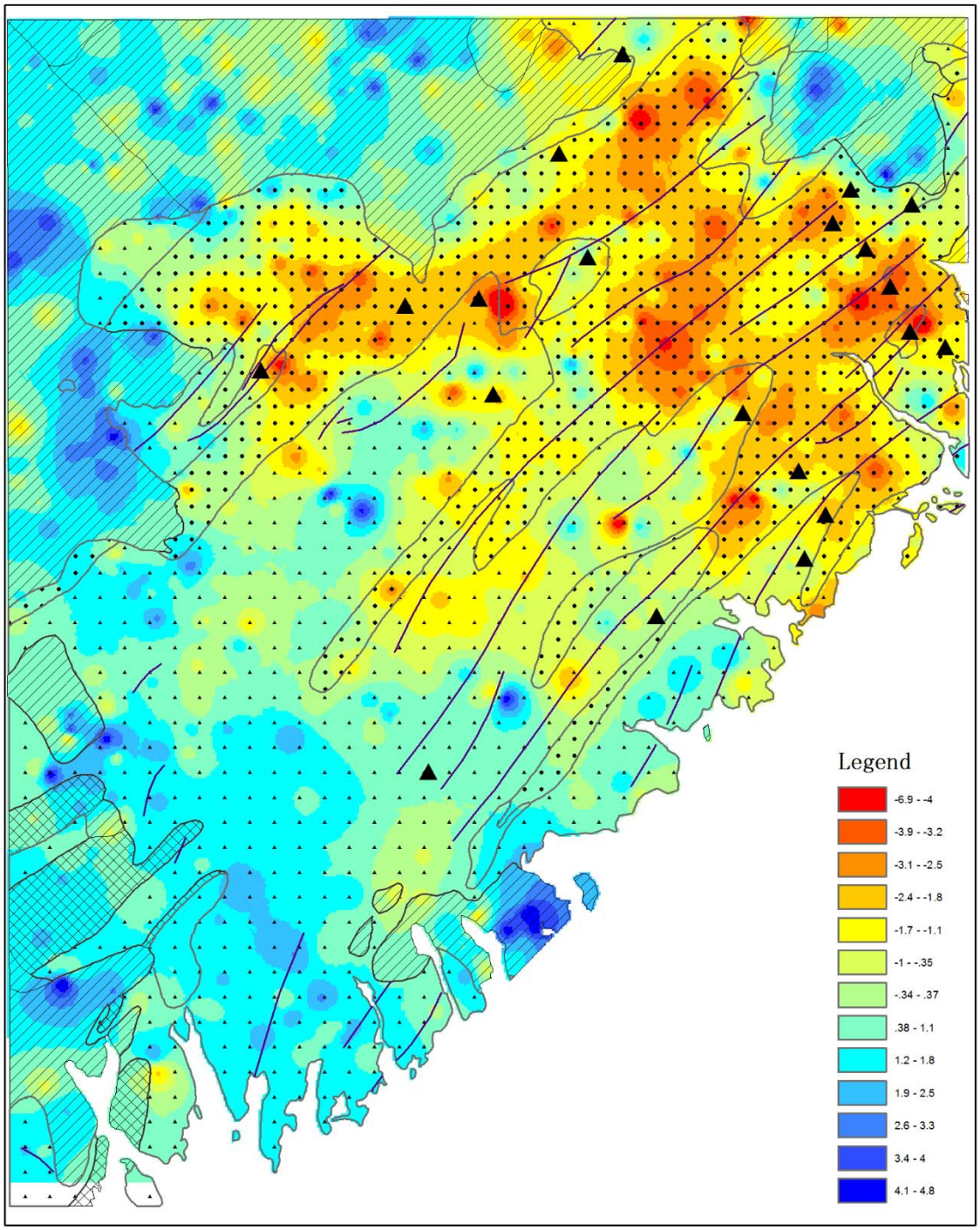


Fig *7.6B The first component of group 2 (Fig 7.3) calculated through the standardized new index matrix, the other legends are shown in Fig 6.12*

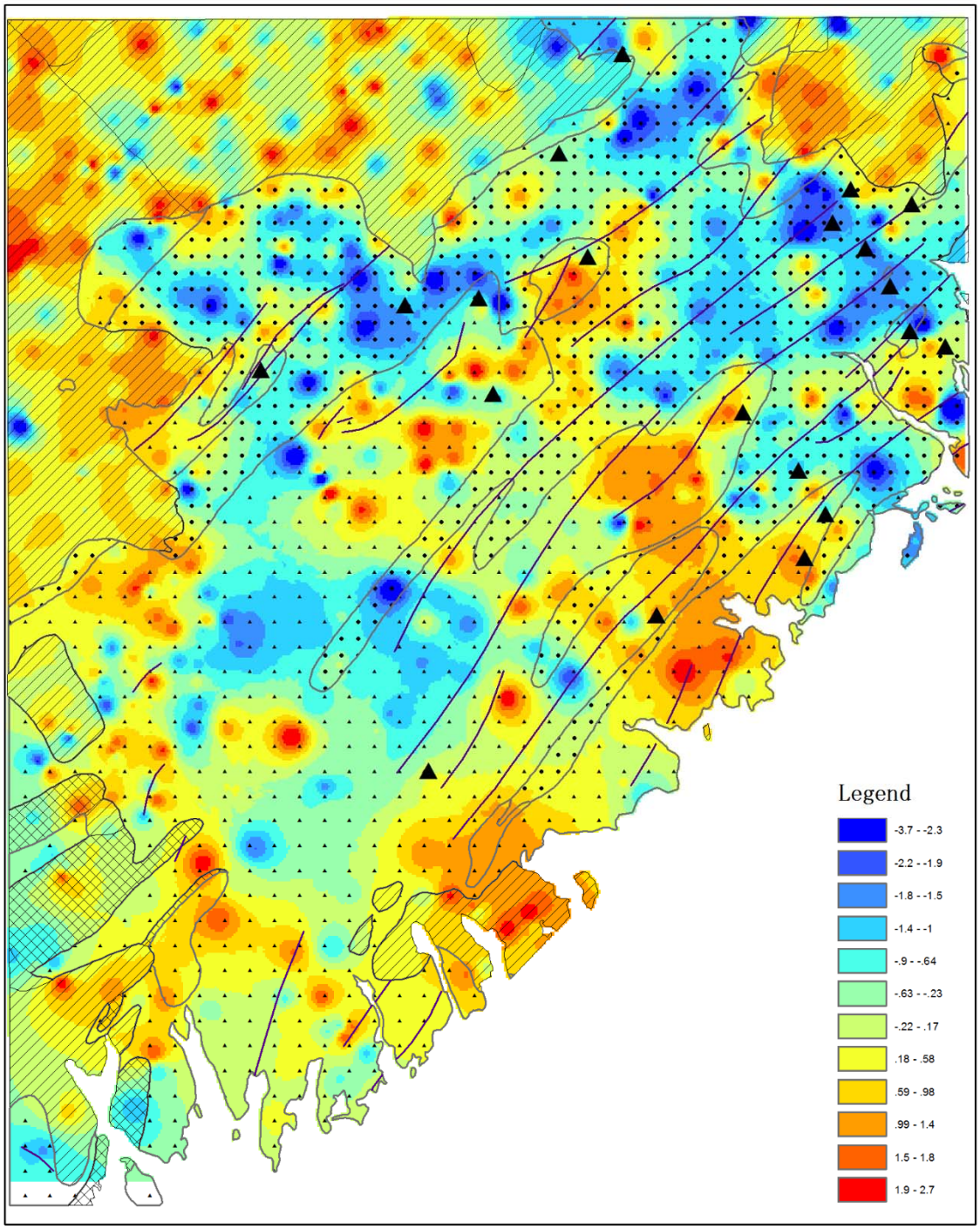


Fig *7.6C The first component of group 3 (Fig 7.3) calculated through the standardized new index matrix, the other legends are shown in Fig 6.12.*

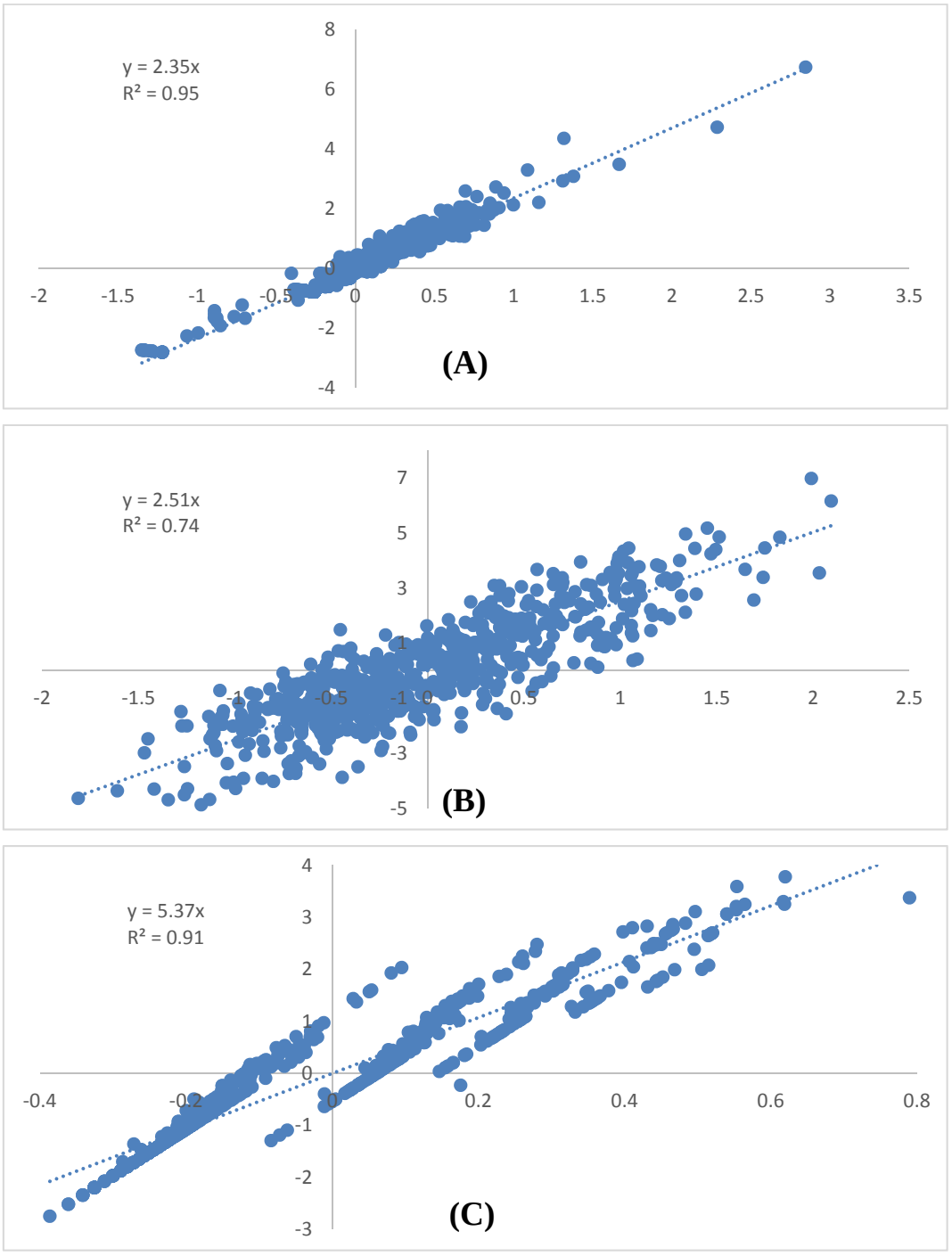
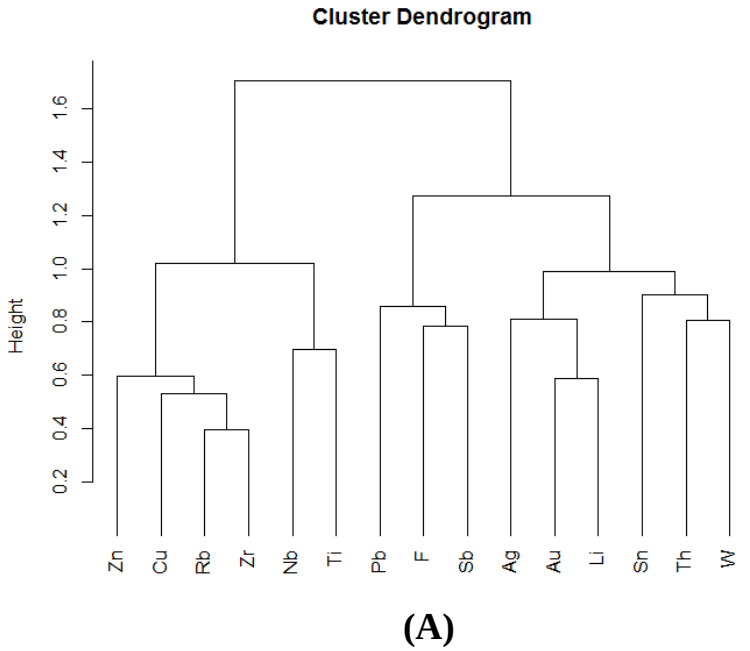


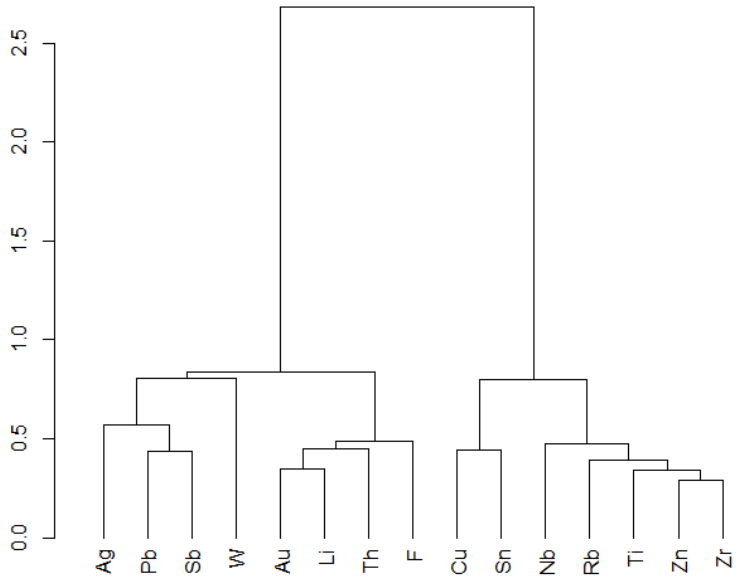
Fig *7.7 The scatter map between (1) the prediction scores for As through the elements in each group and (2) the first component of each group based on the standardized new index matrix. The x axis represent the prediction scores and y axis represent the first main component scores. A, B and C are the maps for group 1, 2 and 3 which shows in Fig.3.*

7.3.5 Clustering for centroid log-ratio transformed data

Fig *7.8A* and *7.8B* are the hierarchical clustering results through CLV method and the constrained method for the CLR transformed dataset, respectively. The result in ***Fig*** *7.8A* is different from the result in ***Fig*** *6.13A*, in which DINIA algorithm was applied and still controlled by the correlation matrix in ***Table*** *6.6*. There are two groups existed in the dendrogram, whose centroids are close to {***Rb***, ***Zr***} and {***Au***, ***Li***}, respectively. Both of the {***Rb***, ***Zr***} and {***Au***, ***Li***} have a strong relationship in correlation. ***Fig*** *7.8B* showed two groups, too, whose centroids are close to {***Au***, ***Li***} and {***Zn***, ***Zr***}. Different with the result in ***Fig*** *7.8A* controlled by the correlation among elements, the elements around the centroid have a strong relationship according to the standardized new index.



(A)



(B)

Fig *7.8 Hierarchical clustering results through CLV algorithm for 15 clr transformed elements through (A) correlation coefficient matrix; (B) standardized new index matrix. The height of dendrogram is the dissimilarity of different clusters.*

7.4 Discussion and conclusions

In this Chapter, a constrained clustering method was proposed based on the new index, which can be considered as a correlation coefficient under the regression to a response variable.

And then a constrained clustering method based on the new index is introduced in the current chapter. Differently from traditional clustering methods, the new method defines the close relationship between two variables under the assumption that they include the same target-related information (large value of the new index), rather than with the high covariance. This new index was adopted to replace the covariance in CLV. Moreover, the centroid was defined as the prediction for a response variable rather than as the principal component in the CLV. In this way, the CLV was transformed into a "supervised" variable clustering method. The clustering result could respond to different factors that are related to a response variable.

The case study on a geochemical element classification demonstrated that the centroid of each group is a meaningful factor that represents a geologic feature. This approach provides a method to extract geochemical factors for a specified object in the current case study, in which three ***As***-related geochemical factors were mapped. The clustering result on the log-ratio transformed dataset shows that the new method is still effective.

There are two types of clustering methods discussed in Chapters 6 and 7: hierarchical and partial clustering method, which have been widely applied in geochemical data clustering. The other techniques commonly used in geochemical clustering, i.e. support vector machines,

random forest, artificial neural network, should be compared with the new index in the future study.

Chapter 8 Conclusions and recommendation for future work

A new SEM approach was proposed and the corresponding algorithms were designed for model estimation and evaluation. An application of the new SEM model in the identification of geochemical factors was successfully introduced and the model was applied to extend the WofE method for the integration of geo-information in mapping mineral potential, which can reduce the CI effect of WofE. A conditional correlation coefficient index was defined and calculated through SEM, which can be used as a new index in PCA, FA and clustering analysis. Based on the defined conditional correlation coefficient, a constrained clustering method was proposed for the variable clustering under a target restriction and the generation of an initial SEM model.

The main achievements of this research in relation to the three major objectives outlined in Chapter 1 can be summarized as follows:

1) *Successful development of a new SEM approach for geo-data processing. The work included: construction of a general SEM model (Chapter 2 & 4); parameter evaluation methods (Chapter 4) and the generation of an initial model (Chapter 6 & 7).*

A new SEM technique, which includes single-level measurement and structural models, was introduced in Chapter 2 & 4 in order to extract mineralization related geochemical factors from the geochemical data under a target restriction. It combines the principles of factor

analysis and regression. Thus, the new mathematical model can generate factors to form model structure, and concurrently ensure the optimum relationships to the objective variables. In order to estimate the SEM model parameters and evaluate the solutions, an optimum method was proposed based on Monte Carlo simulation and random sampling technology, which utilizes the principle of maximization / minimization of a target function. The target function was designed such that the extracted latent variables were mutually less correlated from each other but with the response variable.

A constrained variable clustering method was introduced to generate an initial model in Chapters 6 & 7, which was based on a new index defined for measuring the association of two variables (Chapter 6). It is different from the traditional correlation coefficients used in variable clustering methods, and may provide a method to evaluate the relationship between two variables when they are applied as independent variables in a regression. It is also considered as a conditional correlation coefficient of two variables under the regression to a response variable.

There are two differences between the new constrained method developed here and the traditional unsupervised variable clustering methods. Firstly, the distance between the each pair of variables is defined as the new index as proposed in Chapter 6, rather than the covariance among the variables. Secondly, the centroid of each cluster is a prediction for response variable from the variables in each cluster, rather than the first principal component obtained by PCA from the variables in each cluster.

An application of the new SEM to WofE was introduced in Chapter 5, which may be useful in reducing the CI effect of the WofE method in mineral potential mapping. The advantage

of the new method lies in its ability to define several evidences so that the error between observed and predicted results would be minimized. This treatment may reduce the CI effect of multiple evidence on the posterior probabilities calculated by WofE.

2) *Implementation of the new SEM computer software utility, which includes programs for model parameters estimation; calculation of conditional correlation coefficient, clustering solutions and extraction of evidence and WofE modeling.*

In Chapter 4, a program was designed for the calculation of the new SEM parameters by R language. The program has an iterative structure with an objective to maximize the target function (***Eq** 4.1*). The target function has the model parameters as its input and value of ***Eq** (4.1)* as the output, which determines the goodness of the solutions. The final solution is reached when the target function attains a maximum value. This program can estimate the model parameters only for a pre-defined model structure (***Fig** 4.2*).

In Chapter 5, a program designed for mapping mineral potential based on a new WofE method includes functions to (i) extract the evidence-combinations for input in the WofE model, and (ii) prepare the posterior probability map and provide statistical output of the model.

The first function is the similar to the program described in Chapter 4, which seeks the maximum value of a target function. The evidence for WofE input can be considered as a latent variable, which comes from the reclassified geo-data (i.e. reclassified geochemical data) and controlled by a cut-off or threshold value. Therefore, the output of this function is a group of threshold values obtained through the maximization of a target function (***Eq** 5.18*). The second function provides the WofE model output including the posterior probability of

mineral occurrence, W^+, W^-, contrast and t-value of each evidence.

In Charter 6, a program for solving the conditional correlation coefficient was designed based on ***Eq*** *(6.2)-(6.9)*, which were used for calculating the constrained variable clustering result in Chapter 7.

In Chapter 7, a program was designed for calculating partial and hierarchical variable clustering. While, calculation of the former is based on a random sampling method, calculation of the latter is based on a dissimilarity matrix constructed from a conditional correlation coefficient matrix (***Fig*** *7.2*).

3) *Validation of the proposed method and the developed software utility through case studies.*

The proposed methods and programs have been validated through a case study that used a geochemical dataset of lake sediment samples in southern Nova Scotia, Canada.

The proposed new SEM concept, as introduced in Chapter 4 of this thesis, was applied for identifying geochemical factors associated with gold mineralization. A SEM model consisting of three measurement sub-models and one structural sub-model was created. The calculated results of the new SEM model showed that three geochemical factors are associated with ***As*** and dominated by ***Cu***, ***Zn*** and ***W***, respectively.

For the SEM based WofE method (Chapter 5), the calculation results obtained by this method and by the traditional WofE method indicate that the overestimation for gold deposits in the case study by the new method was reduced 37.5% (from 0.8 times to 0.5 times) and the spatial correlation between the estimation and observation of deposits was increased by 7.4% (from

0.89 to 0.93).

For conditional correlation coefficient described in Chapter 6, the mathematics of the new index is presented and compared with the correlation coefficient based on the same geochemical dataset. The main components of 15 geochemical elements are extracted through correlation coefficients and new index matrix, respectively. The results indicate that the correlation coefficients of the components from the latter with response variable gradually decreases with the Eigen value decline, which has no such clear trends in the components from the former.

The new constrained variable clustering method proposed in Chapter 7 was applied for clustering of the 15 geochemical variables under the restriction of an element (***As***). The partial clustering method using the new index resulted in 3 groups: {***Cu***, ***F***, ***Li***, ***Pb***, ***Sb***, ***Th***, ***Zn***, ***Zr***}, {***Ag***, ***Au***, ***Nb***, ***W***} and {***Rb***, ***Sn***, ***Ti***}. The results indicate that the three clusters of geochemical variables may represent three geological factors related to the Goldenville formation, linear geologic features (e.g., contacts of formation and faults), and granitoid intrusions, respectively.

The current research has demonstrated that the SEM is a relatively new concept in the geosciences and the SEM method may be great potential for application in mineral exploration. Future work is needed to improve the SEM algorithm. Due to the situation that there are no standard software tools available for SEM modeling in the literature, user-friendly software tools are needed in geo-data processing especially for application in

analyzing different types of geo-data types.

The proposed SEM method, using single levels of structural and measurement models, is rather simple for many practical applications. The current algorithm needs to be extended to solve more complex problems such as a geo-model with multiple target variables. Although the SEM method has been tested with in geochemical data processing for mineral exploration, more types of geo-data and case studies are encouraged for method validation.

It should be indicated that the focus of current research is on the development and implementation of the new method, further testing and validation of applications of the proposed method must be encouraged with diverse real applications. The questions to be answered may include but not limited to how to select the response variable and the effect of the compositional data on the performance of the model.

References:

Agterberg, F. P., 1989, Systematic approach to dealing with uncertainty of geoscience information in mineral exploration: Proceedings, Twenty-first Applications of Computers and Operations Research in the Mineral Industry (APCOM) Symposium, Las Vegas, February, p. 165-178.

Agterberg, F. P., 1992, Combining indicator patterns in weights of evidence modeling for resource evaluation: Nonrenewable Resources, v. 1, p. 39-50.

Agterberg, F. P., and G. F. Bonham-Carter, 1990, Deriving weights of evidence from geoscience contour maps for the prediction of discrete events: Proceedings 22nd APCOM Symposium, Berlin, Germany, p. v.2, p. 381-395.

Agterberg, F. P., G. F. Bonham-Carter, and D. F. Wright, 1990, Statistical Pattern Integration for Mineral Exploration, in G. GAáL, and D. F. MERRIAM, eds., Computer Applications in Resource Estimation: Amsterdam, Pergamon, p. 1-21.

Agterberg, F. P., and Q. Cheng, 2002, Conditional independence test for weights-of-evidence modeling: Natural Resources Research, v. 11, p. 249-255.

Agterberg, F. P., 2011, A modified weights-of-evidence method for regional mineral resource estimation: Natural Resources Research, v. 20, p. 95-101.

Aitchison, J., 1982, The statistical analysis of compositional data: Journal of the Royal Statistical Society. Series B (Methodological), p. 139-177.

Aitchison, J., 1984, The statistical analysis of geochemical compositions. Mathematical Geology, 16(6), p. 531-564.

Aitchison, J., 1986, The statistical analysis of compositional data, Springer. Chapman and Hall, London, U.K., 416 pp.

Ali, K., Q. Cheng, and Z. Chen, 2007, Multifractal power spectrum and singularity analysis for modelling stream sediment geochemical distribution patterns to identify anomalies related to gold mineralization in Yunnan Province, South China: Geochemistry: Exploration, Environment, Analysis, v. 7, p. 293-301.

Ames, D. P., 2007, MapWinGIS Reference Manual: A function guide for the free MapWindow GIS ActiveX map component. Idaho State Univ., Idaho Falls, ID, 194 pp.

Ames, D. P., C. Michaelis, and T. Dunsford, 2007, Introducing the MapWindow GIS project: OSGeo Journal, 2. Available from http://www.osgeo.org/journal. [Accessed 17 April 2008].

Anderson, J. C., and D. W. Gerbing, 1988, Structural equation modeling in practice: A review and recommended two-step approach: Psychological bulletin, v. 103(3), p. 411-423.

Atekwana, E. A., and L. D. Slater, 2009, Biogeophysics: A new frontier in Earth science research: Reviews of Geophysics, v. 47.

Bandeen-roche, K., D. L. Miglioretti, S. L. Zeger, and P. J. Rathouz, 1997, Latent Variable Regression for Multiple Discrete Outcomes: Journal of the American Statistical Association, v. 92, p. 1375-1386.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland, 1975, Discrete multivariate analysis: theory and practice: MIT Press, Cambridge, MA, 587 pp.

Bonham-Carter, G. F., F. P. Agterberg, and D. F. Wright, 1988, Integration of geological datasets for gold exploration in Nova Scotia: Digital Geologic and Geographic Information Systems, p. 15-23.

Bonham-Carter, G. F., F. P. Agterberg, and D. F. Wright, 1989, Weights of evidence modelling: a new approach to mapping mineral potential, in F. P. Agterberg, and G. F. Bonham-Carter, eds., Statistical applications in the earth sciences, Energy, Mines and Resources Canada, p. 171-183.

Bonham-Carter, G. F., 1994, Geographic Information Systems for geoscientists: Pergamon, Oxford, 398 pp.

Boyle, R. W., and I. R. Jonasson, 1973, The geochemistry of arsenic and its use as an indicator element in geochemical prospecting: Journal of Geochemical Exploration, v. 2, p. 251-

296.

Brandmaier, A. M., T. von Oertzen, J. J. McArdle, and U. Lindenberger, 2013, Structural equation model trees, Psychological methods, v. 18, p. 71.

Browne, M. W., and G. Arminger, 1995, Specification and estimation of mean-and covariance-structure models, Handbook of statistical modeling for the social and behavioral sciences, Springer, p. 185-249.

Campo, A. G. D., 2012, GIS in environmental assessment: a review of current issues and future needs: Journal of Environmental Assessment Policy and Management, v. 14, 1250007.

Carranza, E. J. M., 2004, Weights of evidence modeling of mineral potential: a case study using small number of prospects, Abra, Philippines: Natural Resources Research, v. 13, p. 173-187.

Cassard, D., M. Billa, A. Lambert, J. Picot, Y. Husson, J. Lasserre, and C. Delor, 2008, Gold predictivity mapping in French Guiana using an expert-guided data-driven approach based on a regional-scale GIS: Ore Geology Reviews, v. 34, p. 471-500.

Castillo-Muñoz, R., and R. J. Howarth, 1976, Application of the empirical discriminant function to regional geochemical data from the United Kingdom: Geological Society of America Bulletin, v. 87, p. 1567-1581.

Cervi, F., M. Berti, L. Borgatti, F. Ronchetti, F. Manenti, and A. Corsini, 2010, Comparing predictive capability of statistical and deterministic methods for landslide susceptibility mapping: a case study in the northern Apennines (Reggio Emilia Province, Italy): Landslides, v. 7, p. 433-444.

Chatterjee, A. K., 1983, Metallogenic map of the Province of Nova Scotia, Department of Mines and Energy, Nova Scotia, Canada, ver. 1, scale 1: 500,000

Chavent, M., V. Kuentz, B. Liquet, and L. Saracco, 2011, Clustofvar: An r package for the clustering of variables: arXiv preprint arXiv:1112.0295.

Chen, M., and E. Vigneau, 2014, Supervised clustering of variables: Advances in Data Analysis and Classification, p. 1-17.

Cheng, Q., 1994, Multifractal modelling and spatial analysis with GIS: gold potential estimation in the Mitchell-Sulphurets Area, northwestern British Columbia, Doctoral Dissertation, School of Graduate Studies and Research, University of Ottawa. 268 pp.

Cheng, Q., 1999, Spatial and scaling modelling for geochemical anomaly separation: Journal of Geochemical exploration, v. 65, p. 175-194.

Cheng, Q., 2007a, Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China: Ore

Geology Reviews, v. 32, p. 314-324.

Cheng, Q., 2007b, Multifractal imaging filtering and decomposition methods in space, Fourier frequency, and eigen domains: Nonlinear Processes in Geophysics, v. 14, p. 293-303.

Cheng, Q., 2008, Non-linear theory and power-law models for information integration and mineral resources quantitative assessments, Progress in Geomathematics, Springer, p. 195-225.

Cheng, Q., 2012a, Ideas and methods for mineral resources integrated prediction in covered areas: Earth Sci-J China Univ Geosci 37:p. 1110–1125 (in Chinese with English abstract)

Cheng, Q., 2012b, Singularity theory and methods for mapping geochemical anomalies caused by buried sources and for predicting undiscovered mineral deposits in covered areas: Journal of Geochemical Exploration, v. 122, p. 55-70.

Cheng, Q., 2014, Vertical distribution of elements in regolith over mineral deposits and implications for mapping geochemical weak anomalies in covered areas: Geochemistry: Exploration, Environment, Analysis, v. 14, p. 277-289

Cheng, Q., 2015, BoostWofE: A New Sequential Weights of Evidence Model Reducing the Effect of Conditional Dependency: Mathematical Geosciences, p. 1-31.

Cheng, Q., G. Bonham-Carter, W. Wang, S. Zhang, W. Li, and X. Qinglin, 2011, A spatially weighted principal component analysis for multi-element geochemical data for mapping locations of felsic intrusions in the Gejiu mineral district of Yunnan, China: Computers & Geosciences, v. 37, p. 662-669.

Cheng, Q., and F. P. Agterberg, 1999, Fuzzy weights of evidence method and its application in mineral potential mapping: Natural Resources Research, v. 8, p. 27-35.

Cheng, Q., Y. Xu, and E. Grunsky, 2000, Integrated spatial and spectrum method for geochemical anomaly separation: Natural Resources Research, v. 9, p. 43-52.

Chin, W. W., 1998, The partial least squares approach to structural equation modeling: Modern methods for business research, v. 295, p. 295-336.

Cho, S., N. C. Poudyal, and R. K. Roberts, 2008, Spatial analysis of the amenity value of green open space: Ecological Economics, v. 66, p. 403-416.

Chun, H., and S. Keleş, 2010, Sparse partial least squares regression for simultaneous dimension reduction and variable selection: Journal of the Royal Statistical Society: Series B (Statistical Methodology), v. 72, p. 3-25.

Chung, C. F., and F. P. Agterberg, 1980, Regression models for estimating mineral resources from geological map data: Journal of the International Association for Mathematical Geology, v. 12, p. 473-488.

Comrey, A. L., and H. B. Lee, 2013, A first course in factor analysis, Psychology Press, NewYork. 316 pp.

Crocket, J. H., F. Fueten, P. M. Clifford, and A. Kabir, 1986, Distribution and localization of gold in Meguma Group rocks, Nova Scotia: implications of metal distribution patterns in quartz veins and host rocks on mineralization processes at Harrigan Cove, Halifax County: Atlantic Geology, v. 22. p. 15-33

de Caritat, P., and E. C. Grunsky, 2013, Defining element associations and inferring geological processes from total element concentrations in Australian catchment outlet sediments: multivariate analysis of continental-scale geochemical data: Applied Geochemistry, v. 33, p. 104-126.

Deng, M., 2009, A conditional dependence adjusted weights of evidence model: Natural resources research, v. 18, p. 249-258.

Deng, M., 2010a, A spatially autocorrelated weights of evidence model: Natural resources research, v. 19, p. 33-44.

Deng, M., 2010b, An ordered Weights of Evidence model for ordered discrete variables: Natural resources research, v. 19, p. 83-89.

Dhillon, I. S., E. M. Marcotte, and U. Roshan, 2003, Diametrical clustering for identifying anti-correlated gene clusters: Bioinformatics, v. 19, p. 1612-1619.

Dunn, C. E., W. B. Coker, and P. J. Rogers, 1991, Reconnaissance and detailed geochemical surveys for gold in eastern Nova Scotia using plants, lake sediment, soil and till: Journal of Geochemical Exploration, v. 40, p. 143-163.

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

Fornell, C., 1987, "Corporate Consumer Affairs Departments: Retrospect and Prospect," in The Frontier of Research in the Consumer Interest, E. Scott Maynes (ed.), Columbia, MO: American Council on Consumer Interests, p. 595–619.

Frank, L. E., and J. H. Friedman, 1993, A statistical view of some chemometrics regression tools: Technometrics, v. 35, p. 109-135.

Gao, S., L. Li, W. Li, K. Janowicz, and Y. Zhang, 2014, Constructing gazetteers from volunteered big geo-data based on Hadoop: Computers, Environment and Urban Systems.

Garrett, R. G., V. E. Kane, and R. K. Zeigler, 1980, The management and analysis of regional geochemical data: Journal of Geochemical Exploration, v. 13, p. 115-152.

Geladi, P., and B. R. Kowalski, 1986, Partial least-squares regression: a tutorial: Analytica chimica acta, v. 185, p. 1-17.

Gorney, R. M., D. R. Ferris, A. D. Ward, and L. R. Williams, 2011, Assessing channel-forming characteristics of an impacted headwater stream in Ohio, USA: Ecological Engineering, v. 37, p. 418-430.

Graves, R. H., and P. W. Finck, 1988, The provenance of tills overlying the eastern part of the South Mountain Batholith, Nova Scotia: Atlantic Geology, v. 24. p. 61-70

Grunsky, E. C., U. A. Mueller, and D. Corrigan, 2014, A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: Applications for predictive geological mapping: Journal of Geochemical Exploration, v. 141, p. 15-41.

GSurvey, U. S., 2007, Facing Tomorrow's Challenges—US Geological Survey Science in the Decade 2007–2017: USGS Circular, v. 1309, p. 1-21.

Gustavsson, N., and A. Bjorklund, 1976, Lithological classification of tills by discriminant analysis: J. Geochem. Explor, v. 5, p. 393-395.

Haavelmo, T., 1943, The statistical implications of a system of simultaneous equations: Econometrica, Journal of the Econometric Society, p. 1-12.

Hair Jr, J. F., G. T. M. Hult, C. Ringle, and M. Sarstedt, 2013, A primer on partial least squares structural equation modeling (PLS-SEM), Sage, Thousand Oaks. 328 pp.

Hair, J. F., C. M. Ringle, and M. Sarstedt, 2011, PLS-SEM: Indeed a silver bullet: The Journal

of Marketing Theory and Practice, v. 19, p. 139-152.

Hair, J. F., M. Sarstedt, T. M. Pieper, and C. M. Ringle, 2012, The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications: Long Range Planning, v. 45, p. 320-340.

Hair, J. F., Babin, B., Money, A. H., and Samouel, P., 2003, Essentials of business research methods. New York: John Wiley & Sons. 576 pp.

Hair, J. F., R. L. Tatham, R. E. Anderson, and W. Black, 2006, Multivariate data analysis, v. 6, Pearson Prentice Hall Upper Saddle River, NJ. 899 pp.

Hanesch, M., R. Scholger, and M. J. Dekkers, 2001, The application of fuzzy c-means cluster analysis and non-linear mapping to a soil data set for the detection of polluted sites: Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy, v. 26, p. 885-891.

Harman, H. H., 1976, Modern factor analysis, 3rd ed, Chicago, University of Chicago Press. 469 pp.

Hart, J. K., and K. Martinez, 2006, Environmental Sensor Networks: A revolution in the earth system science, Earth-Science Reviews, v. 78, p. 177-191.

Helland, I. S., 1990, Partial least squares regression and statistical models: Scandinavian

Journal of Statistics, p. 97-114.

Hendry, D. F., 1976, The structure of simultaneous equations estimators: Journal of Econometrics, v. 4, p. 51-88.

Höskuldsson, A., 1988, PLS regression methods: Journal of chemometrics, p. 211-228.

Howarth, R. J., 1973, The pattern recognition problem in applied geochemistry. In Geochemical Exploration 1972 (ed. M. J. JONES), Institution of Mining and Metallurgy, London. p. 259-273

Hoyle, R. H., 1995, “The structural equation modeling approach: Basic concepts and fundamental issues”, in R.H. Hoyle (ed.), Structural Equation Modeling, Concepts, Issues, and Applications, Sage Publications,, p. 1–15.

Hoyle, R. H., and A. T. Panter, 1995, "Writing About Structural Equation Models", in R. H. Hoyle, ed., Structural equation modeling: Concepts, issues, and applications, Sage Publications, p. 158-176.

Iacobucci, D., 1994, Classic factor analysis: Principles of marketing research, p. 279-316.

Ihaka, R., and R. Gentleman, 1996, R: a language for data analysis and graphics: Journal of computational and graphical statistics, v. 5, p. 299-314.

Jain, A. K., and R. C. Dubes, 1988, Algorithms for clustering data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ, 334 pp.

James, L. R., and B. K. Singh, 1978, An introduction to the logic, assumptions, and basic analytic procedures of two-stage least squares: Psychological Bulletin, v. 85, p. 1104-1122.

Jarvis, C. B., S. B. MacKenzie, and P. M. Podsakoff, 2003, A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research: Journal of Consumer Research, v. 30, p. 199-218.

Jensen, J. R., 2009, Remote Sensing of the Environment: An Earth Resource Perspective, 2nd ed; Pearson Prentice-Hall, Pearson Education, Inc: Upper Saddle River, NJ, USA. 608 pp.

Ji, H., D. Zeng, Y. Shi, Y. Wu, and X. Wu, 2007, Semi-hierarchical correspondence cluster analysis and regional geochemical pattern recognition: Journal of Geochemical Exploration, v. 93, p. 109-119.

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin, 2003, A modified principal component technique based on the LASSO: Journal of Computational and Graphical Statistics, v. 12, p. 531-547.

Jöreskog, K. G., 1970, A general method for analysis of covariance structures: Biometrika, v.

57, p. 239-251.

Jöreskog, K. G., 1978, Structural analysis of covariance and correlation matrices: Psychometrika, v. 43, p. 443-477.

Jöreskog, K. G., and D. Sörbom, 1996, LISREL 8 user's reference guide, Chicago: Scientific Software International. 378 pp.

Journel, A. G., 2002, Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses: Mathematical geology, v. 34, p. 573-596.

Kaplan, D., 2008, Structural equation modeling: Foundations and extensions, v. 10, Sage Publications. 267 pp.

Kaufman, L., and P. J. Rousseeuw, 2009, Finding groups in data: an introduction to cluster analysis, v. 344, John Wiley & Sons. 368 pp.

Kemp, L. D., G. F. Bonham-Carter, and G. L. Raines, 1999, Arc-WofE: Arcview extension for weights of evidence mapping: Geological Survey of Canada, United States of Geological Survey. User Guide, v. 76.

Kerswill, J. A., ed., 1988, Lithogeochemical indicators of gold potential in the eastern Meguma Terrain of Nova Scotia: second progress report: Mines and Mineral Branch, Report of Activities 1988, v. 88-3, p. 215-217.

Kim, J., and C. W. Mueller, 1978a, Factor analysis: Statistical methods and practical issues, Newbury Park: Sage Publications, Beverly Hills, CA. 88 pp.

Kim, J., and C. W. Mueller, 1978b, Introduction to factor analysis: What it is and how to do it, Newbury Park: Sage Publications, Beverly Hills, CA. 80 pp.

Kontak, D. J., P. K. Smith, R. Kerrich, and P. F. Williams, 1990, Integrated model for Meguma Group lode gold deposits, Nova Scotia, Canada: Geology, v. 18, p. 238-242.

Kontak, D. J., R. J. Horne, H. Sandeman, D. Archibald, and J. K. Lee, 1998, 40Ar/39Ar dating of ribbon-textured veins and wall-rock material from Meguma lode gold deposits, Nova Scotia: implications for timing and duration of vein formation in slate-belt hosted vein gold deposits: Canadian Journal of Earth Sciences, v. 35, p. 746-761.

Kontak, D. J., and R. Kerrich, 1997, An isotopic (C, O, Sr) study of vein gold deposits in the Meguma Terrane, Nova Scotia; implication for source reservoirs: Economic Geology, v. 92, p. 161-180.

Kramar, U., 1995, Application of limited fuzzy clusters to anomaly recognition in complex geological environments: Journal of Geochemical Exploration, v. 55, p. 81-92.

Krishnan, S., A. Boucher, and A. G. Journel, 2004, Evaluating information redundancy through the tau model, Geostatistics Banff 2004, Springer, p. 1037-1046.

Lawley, D. N., and A. E. Maxwell, 1967, Factor analysis as a statistical method: Journal of the Royal Statistical Society, Series D (The Statistician), 12(3), p. 209 - 229.

Lê Cao, K., D. Rossouw, C. Robert-Granié, and P. Besse, 2008, A sparse PLS for variable selection when integrating omics data: Statistical applications in genetics and molecular biology, v. 7(1).

Lee, R. H., B. Kim, I. Choi, H. Kim, H. Choi, K. Suh, Y. C. Bae, and J. S. Jung, 2004, Characterization and expression analysis of mesenchymal stem cells from human bone marrow and adipose tissue: Cellular Physiology and Biochemistry, v. 14, p. 311-324.

Lee, S., and J. Choi, 2004, Landslide susceptibility mapping using GIS and the weight-of-evidence model: International Journal of Geographical Information Science, v. 18, p. 789-814.

Lee, S. Y., and J. Q. Shi, 2001, Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data: Biometrics, v. 57, p. 787-94.

Lewis, B. W., M. B. Lewis, and S. RUnit, 2014, Package ‘rredis’, Available from: ftp://apache.cs.uu.nl/mirror/CRAN/web/packages/rredis/rredis.pdf, 101 pp.

Lohmöller, J., 1989, Latent variable path modeling with partial least squares, Heidelberg: Physica. 286 pp.

MacKenzie, S. B., P. M. Podsakoff, and C. B. Jarvis, 2005, The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions: Journal of Applied Psychology, v. 90, p. 710-730.

Madden, R. A., and P. R. Julian, 1994, Observations of the 40-50-day tropical oscillation-A review: Monthly Weather Review, v. 122, p. 814-837.

Maechler, M., P. Rousseeuw, A. Struyf, and M. Hubert, 2005, cluster: Cluster Analysis Basics and Extensions. R package version 1.15. Available from: http://cran.rproject.org/web/packages/cluster/index.html.

Martens, H., and T. Naes, 1992, Multivariate calibration, John Wiley & Sons. 438 pp.

Mateos-Aparicio, G., 2011, Partial least squares (PLS) methods: Origins, evolution, and application to social sciences: Communications in Statistics-Theory and Methods, v. 40, p. 2305-2317.

Mawer, C. K., 1986, The bedding-concordant gold-quartz veins of the Meguma Group, Nova Scotia: Turbidite-Hosted Gold Deposits: Geological Association of Canada, Special Paper, v. 32, p. 135-148.

McArdle, J. J., and K. M. Kadlec, 2013, Structural equation models: The Oxford Handbook of Quantitative Methods, Vol. 2: Statistical Analysis, 295 pp.

McDonald, J. B., 1977, The k-class estimators as least variance difference estimators: Econometrica: Journal of the Econometric Society, p. 759-763.

McIntosh, A. R., F. L. Bookstein, J. V. Haxby, and C. L. Grady, 1996, Spatial pattern analysis of functional brain images using partial least squares: Neuroimage, v. 3, p. 143-157.

Mehta, J. S., and P. A. Swamy, 1978, The existence of moments of some simple Bayes estimators of coefficients in a simultaneous equation model: Journal of Econometrics, v. 7, p. 1-13.

Minasny, B., A. B. McBratney, and S. Salvador-Blanes, 2008, Quantitative models for pedogenesis—A review: Geoderma, v. 144, p. 140-157.

Mouillot, F., M. G. Schultz, C. Yue, P. Cadule, K. Tansey, P. Ciais, and E. Chuvieco, 2014, Ten years of global burned area products from spaceborne remote sensing—A review: Analysis of user needs and recommendations for future developments: International Journal of Applied Earth Observation and Geoinformation, v. 26, p. 64-79.

Mulaik, S. A., 2009, Foundations of factor analysis, NY: McGraw-Hill. 548 pp.

Muthen, B., 1984, A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators: Psychometrika, v. 49, p. 115 - 132.

Neuhäuser, B., and B. Terhorst, 2007, Landslide susceptibility assessment using "weights-of-evidence" applied to a study area at the Jurassic escarpment (SW-Germany): Geomorphology, v. 86, p. 12-24.

Nykiforuk, C. I., and L. M. Flaman, 2009, Geographic information systems (GIS) for health promotion and public health: a review: Health promotion practice. Health Promotion Practice, 12(1), 63–73.

Oztekin, A., Z. J. Kong, and D. Delen, 2011, Development of a structural equation modeling-based decision tree methodology for the analysis of lung transplantations: Decision Support Systems, v. 51, p. 155-166.

Pawlowsky-Glahn, V., and A. Buccianti, 2011, Compositional data analysis: Theory and applications, John Wiley & Sons, Ltd., London, 400 pp.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado, 2015, Modeling and Analysis of Compositional Data, John Wiley & Sons, 272 pp.

Pearl, J., 2000, Causality: models, reasoning and inference, v. 29, Cambridge Univ Press. 400 pp.

Polyakova, E. I., and A. G. Journel, 2007, The Nu expression for probabilistic data integration: Mathematical Geology, v. 39, p. 715-733.

Pearson, K., 1896, Mathematical Contributions to the Theory of Evolution.--On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs: Proceedings of the Royal Society of London, v. 60, p. 489-498.

Porwal, A., I. González-Álvarez, V. Markwitz, T. C. McCuaig, and A. Mamuse, 2010, Weights-of-evidence and logistic regression modeling of magmatic nickel sulfide prospectivity in the Yilgarn Craton, Western Australia: Ore Geology Reviews, v. 38, p. 184-196.

Punniyamoorthy, M., P. Mathiyalagan, and P. Parthiban, 2011, A strategic model using structural equation modeling and fuzzy logic in supplier selection: Expert Systems with Applications, v. 38, p. 458-474.

Qannari, E. M., E. Vigneau, P. Luscan, A. C. Lefebvre, and F. Vey, 1997, Clustering of variables, application in consumer and sensory studies: Food quality and preference, v. 8, p. 423-428.

Qannari, E. M., E. Vigneau, and P. Courcous, 1998, Une nouvelle distance entre variables; application en classification.: Revue de Statistique, v. XLVI(2), p. 21–32.

Ramsey, J. B., 1978, Nonlinear estimation and asymptotic approximations: Econometrica: Journal of the Econometric Society, p. 901-929.

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.

Rao, C., A. Sahuquillo, and J. L. Sanchez, 2008, A review of the different methods applied in environmental geochemistry for single and sequential extraction of trace elements in soils and related materials: Water, Air, and Soil Pollution, v. 189, p. 291-333.

Regmi, N. R., J. R. Giardino, and J. D. Vitek, 2010, Modeling susceptibility to landslides using the weight of evidence approach: Western Colorado, USA: Geomorphology, v. 115, p. 172-187.

Reimann, C., P. Filzmoser, and R. G. Garrett, 2002, Factor analysis applied to regional geochemical data: problems and possibilities: Applied Geochemistry, v. 17, p. 185-206.

Reimann, C., and P. Filzmoser, 2000, Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data: Environmental geology, v. 39, p. 1001-1014.

Ringle, C. M., S. Wende, and A. Will, 2005, SmartPLS 2.0 M3, Hamburg, Available at

http://www.smartpls.de.

Rogers, P. J., A. K. Chatterjee, and J. W. Aucott, 1990, Metallogenic domains and their reflection in regional lake sediment surveys from the Meguma Zone, southern Nova

Scotia, Canada: Journal of Geochemical Exploration, v. 39, p. 153-174.

Rogers, P. J., M. A. MacDonald, D. W. Rigby, and N. Scotia, 1985, Regional lake sediment survey of the Meguma Zone, southern Nova Scotia: new analytical data, N.S. Dep. Mines and Energy, Open File Rep. 605, 65 pp., 2 maps.

Rogers, P. J., R. F. Mills, and P. A. Lombard, 1987, Regional geochemical study in Nona Scotia: Mines and mineral branch, report of activities, v. 87, p. 147-154.

Rogers, P. J., and R. G. Garrett, 1987, Lithophile elements and exploration using centre-lake bottom sediments from the East Kemptville area, Southern Nova Scotia, Canada: Journal of Geochemical Exploration, v. 28, p. 467-478.

Rollinson, H. R., 2014, Using geochemical data: evaluation, presentation, interpretation, Routledge. Longman Group UK Ltd., New York (1993), 352 pp.

Romero-Calcerrada, R., F. Barrio-Parra, J. Millington, and C. J. Novillo, 2010, Spatial modelling of socioeconomic data to understand patterns of human-caused wildfire ignition risk in the SW of Madrid (central Spain): Ecological Modelling, v. 221, p. 34-45.

Romero-Calcerrada, R., and S. Luque, 2006, Habitat quality assessment using Weights-of-Evidence based GIS modelling: The case of< i> Picoides tridactylus as species indicator of the biodiversity value of the Finnish forest: Ecological Modelling, v. 196, p. 62-76.

Ryan, R. J., and W. Ramsay, 1997, Preliminary comparison of gold field in the Meguma Terrain, Nova Scotia, and Victoria, Australia: Mines and Mineral Branch. Report of Activities 1996, 97-1, p. 157–162.

Sammel, M. D., and L. M. Ryan, 1996, Latent variable models with fixed effects: Biometrics, v. 52, p. 650-663.

Sánchez, B. N., E. Budtz-Jørgensen, L. M. Ryan, and H. Hu, 2005, Structural equation models: a review with applications to environmental epidemiology: Journal of the American Statistical Association, v. 100, p. 1443-1455.

Sangster, A. L., 1990, Metallogeny of the Meguma Terrane, Nova Scotia: Mineral deposit studies in Nova Scotia, v. 1, p. 90-8.

Sargan, J. D., 1978, On the existence of the moments of 3SLS estimators: Econometrica: Journal of the Econometric Society, v. 46, p. 1329-1350.

Särndal, C. E., B. Swensson, and J. Wretman, 1992, Model assisted survey sampling. Springer-Verlag, New York, 694 pp.

Savinykh, V. P., and V. Y. Tsvetkov, 2014, Geodata as a systemic information resource: Herald of the Russian Academy of Sciences, v. 84, p. 365-368.

Schaeben, H., 2012, Comparison of mathematical methods of potential modeling:

Mathematical Geosciences, v. 44, p. 101-129.

Schaeben, H., 2014, A mathematical view of weights-of-evidence, conditional independence, and logistic regression in terms of markov random fields: Mathematical Geosciences, v. 46, p. 691-709.

Selva, D., B. G. Cameron, and E. F. Crawley, 2014, Rule-Based System Architecting of Earth Observing Systems: Earth Science Decadal Survey: Journal of Spacecraft and Rockets, v. 51, p. 1505-1521.

Simon, H. A., 1977, Causal ordering and identifiability: Models of Discovery, Springer, v. 54, p. 53-80.

Sinnwell, J. P., D. J. Schaid, and Z. Yu, 2007, haplo.stats: Statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous: URL http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm.

Steinberg, A. N., 2009, Context-sensitive data fusion using structural equation modeling: Information Fusion, 2009. FUSION'09. 12th International Conference on, p. 725-731.

Subedi, S., A. Punzo, S. Ingrassia, and P. D. McNicholas, 2013, Clustering and classification via cluster-weighted factor analyzers: Advances in Data Analysis and Classification, v. 7, p. 5-40.

Team, R. C., 2012, R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; Available from: http://www.R-project.org/.

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

Tillé, Y., and A. Matei, 2009, Sampling: survey sampling. R package version 2.2. Available from: http://cran.r-project.org/web/packages/sampling/index.html

Thurstone, L. L., 1947, Multiple factor analysis. Chicago, IL: University of Chicago Press.

Ullman, J. B., and P. M. Bentler, 2003, Structural equation modeling, In J. A. Schinka & W. F. Velicer (Eds.), Handbook of psychology: Vol. 2. Research methods in psychology (p. 607–634). Hoboken, NJ: Wiley.

Vigneau, E., and E. M. Qannari, 2003, Clustering of variables around latent components: Communications in Statistics-Simulation and Computation, v. 32, p. 1131-1150.

Vistelius, A. B., 1960, The skew frequency distributions and the fundamental law of the geochemical processes: The journal of geology, p. 1-22.

Vriend, S. P., P. Van Gaans, J. Middelburg, and A. De Nijs, 1988, The application of fuzzy c-means cluster analysis and non-linear mapping to geochemical datasets: examples from

Portugal: Applied Geochemistry, v. 3, p. 213-224.

Wang, W., and Q. Cheng, 2008, Mapping mineral potential by combining multi-scale and multi-source geo-information: Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International, p. II-1321-II-1324.

Wathne, K., J. Roos, and G. von Krogh, 1996, Towards a theory of knowledge transfer in a cooperative context, Sage Publications: London.

Weng, Q., 2014, Global Urban Monitoring and Assessment through Earth Observation, Crc Press, 420 pp.

Wielicki, B. A., B. R. Barkstrom, E. F. Harrison, R. B. Lee III, G. Louis Smith, and J. E. Cooper, 1996, Clouds and the Earth's Radiant Energy System (CERES): An earth observing system experiment: Bulletin of the American Meteorological Society, v. 77, p. 853-868.

Wold, H., 1966, Estimation of principal components and related models by iterative least squares: Multivariate analysis, v. 1, p. 391-420.

Wold, H., 1982, Systems under indirect observation using PLS: A second generation of multivariate analysis, v. 1, p. 325-347.

Wold, H., 1985, Systems analysis by partial least squares: Measuring the unmeasurable, p.

221-251.

Wright, S., 1921, Correlation and causation: Journal of agricultural research, v. 20, p. 557-585.

Xie, X., D. Liu, Y. Xiang, G. Yan, and C. Lian, 2004, Geochemical blocks for predicting large ore deposits—concept and methodology: Journal of Geochemical Exploration, v. 84, p. 77-91.

Xu, Y., and Q. Cheng, 2001, A fractal filtering technique for processing regional geochemical maps for mineral exploration, v. 1, p. 147-156.

Yuan, K. H., and P. M. Bentler, 2000, Three likelihood‐based methods for mean and covariance structure analysis with nonnormal missing data: Sociological methodology, v. 30, p. 165-200.

Yuan, K., and P. M. Bentler, 1997, Mean and Covariance Structure Analysis: Theoretical and Practical Improvements: Journal of the American Statistical Association, v. 92, p. 767-774.

Zellner, A., 1978, Estimation of functions of population means and regression coefficients including structural coefficients: A minimum expected loss (MELO) approach: Journal of Econometrics, v. 8, p. 127-158.

Zentilli, M., M. C. Graves, T. Mulja, I. MacInnis, and J. R. Matheson, 1985, Geochemical characterization of the Goldenville-Halifax Transition of the Meguma Group of Nova Scotia; preliminary report: Geological Survey of Canada Paper 86-1A, pp. 423–428.

Zhang, S., Q. Cheng, S. Zhang, and Q. Xia, 2009, Weighted weights of evidence and stepwise weights of evidence and their applications in Sn-Cu mineral potential mapping in Gejiu, Yunnan Province, China: Earth Sci J China Univ Geosci, v. 34, p. 281-286.

Zou, H., T. Hastie, and R. Tibshirani, 2006, Sparse principal component analysis: Journal of computational and graphical statistics, v. 15, p. 265-286.