Advanced Statistical Methods for Geochemical Mineral Exploration

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Univ.-Prof. Dipl.-Ing. Dr. techn. Peter Filzmoser, Institut für Stochastik und Wirtschaftsmathematik (E105)

eingereicht an der Technischen Universität Wien an der Fakultät für Mathematik and Geoinformation

von

Mgr. Dominika Mikšová
Matrikelnummer 11729951

Diese Dissertation haben begutachtet:

doc. RNDr. Karel Hron, Ph.D.	Prof. Pertti Sarala, Ph.D.

Wien, 31. März 2020

Mgr. Dominika Mikšová

Erklärung zur Verfassung der Arbeit

Mgr. Dominika Mikšová
Untere Viaduktgasse 21,1030 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 31. März 2020

Mgr. Dominika Mikšová

Acknowledgements

I would like to express my sincerest thanks to my supervisor Professor Peter Filzmoser for his enduring support, encouragement and advice he has provided me during my PhD studies. I also appreciate the substantial help of the co-authors of my publications, Christopher Rieser and Maarit Middleton. I would like to thank all my present and also past colleagues, their help has made my studies much easier and more fun. I also want to express my gratitude to my family, especially my parents and brother, for their continued support, encouragement and patience.

This thesis come into being, as one of the main outcome, during the UpDeep project (Upscaling deep buried geochemical exploration techniques into European business, 2017-2020), funded by the European Information and Technology Raw Materials, therefore this project was crucial for building up this work. I wish to thank GEUS and BRGM UpDeep researchers (Simon M. Thaarup, Anders Risbjerg, Jeremie Melleton and Bruno Lemiere) for the sample collection and useful information related to data analytics. This activity received funding from the European Institute of Innovation and Technology (EIT), a body of the European Union, under the Horizon 2020, the EU Framework Programme for Research and Innovation.

Next, many thanks of the Lätäseno data go to GTK personel Janne Kivilompolo and Kari Kivilompolo for mountain birch sampling and Jukka Konnunaho for organizing the funding and project management of Mineral potential of northern Finland. Also big thanks of the Tiira data goes to Jukka V̈limaa and Jyrki Korteniemi (Agnico-Eagle Finland Ltd., Kittilä Mine, Finland) for allowing access to drill core geochemical data and plant sampling, and Sami Lepistö and Tero Niiranen (Geological Survey of Finland, GTK) for 3D and lithogeochemical expertise. Päivi Heikkilä, Matti Piekkari, Petro Oravalahti and Markku Virtanen (GTK) performed the plant sampling and sample preparation. The sample collection and analytics was funded by the GTK project Postglacial faults (# 50402-20086, 2016-2019).

Abstract

The detection and identification of mineralization in geochemical exploration contains many tasks that are strongly linked to statistics. A geochemical exploration project starts with sampling planning in the area under investigation in terms of an optimal sampling design. There are of course also several other considerations that need to be taken into account, most importantly the overall costs for sampling, which limits the number of samples to be collected. Once the samples are available, they are analyzed in a laboratory resulting in "geochemical data", which are challenging by their nature. Typically, they are compositional and thus multivariate, spatially dependent, they usually come with detection limit issues, and different kinds of uncertainties are inherent in these data. The last point is particularly addressed with statistical quality control procedures, and this provides the basis for selecting the data that are finally used for the subsequent statistical analyses.

Besides the data quality considerations, data preprocessing is the following important step. Since values below the lower or above the upper detection limit could affect subsequent multivariate data analyses, it is important to first replace these values by appropriately estimated numbers. While methods accounting for the compositional nature of the geochemical data are available to estimate values below the lower detection limit, a novel method dealing with values exceeding the upper detection limit is proposed. Since this regression based procedure acts in a multivariate sense, it has advantages over existing strategies such as replacing right-censored values simply by a constant.

The main statistical task in geochemical exploration is to locate of mineralized zones and to identify the underlying lithogeochemical source. While exploratory data analysis techniques may support this process, they are usually not accounting for the compositional nature of the data. Thus, an unsupervised methodology is introduced which accounts for the log-ratios of all element pairs. Due to the presence of data uncertainties, not the original observations are considered for the log-ratios, but values obtained from smooth fits, derived from Generalized Additive Models (GAMs). A measure incorporating the

overall curvature of a log-ratio pair is introduced to rank the pairs, and to indicate pathfinder elements vectoring towards the mineralization. The procedure is developed for cases where samples located on linear transects, and also extended to cases where samples are taken on a plane. Real geochemical exploration data sets are used to demonstrate the performance of the methods.

Kurzfassung

Die Erkennung von Mineralisation in der geochemischen Exploration ist an viele Aufgaben mit starkem Bezug zur Statistik gekoppelt. Das beginnt bereits bei der Planung der Proben im zu untersuchenden Gebiet, nämlich beim Erstellen eines optimalen Stichprobenplans. Natürlich gilt es dabei auch viele andere Aspekte zu berücksichtigen, zu allererst die gesamten Kosten für die Probennahmen, was die Anzahl der Proben limitiert. Wenn die Proben dann genommen sind, können sie im Labor analysiert werden, und man erhält "geochemische Daten", die aufgrund ihrer Besonderheiten herausfordernd sind. Typischerweise sind das Kompositionsdaten und daher multivariate Daten, sie sind räumlich abhängig, sie haben normalerweise Probleme mit der Nachweisgrenze, und sie sind geprägt von verschiedenen Arten von Unsicherheiten. Speziell der letzte Punkt wird mit den Methoden der statistischen Qualitätskontrolle adressiert, und das liefert die Grundlage für die Selektion der Daten, die letztlich für die weiterführende statistische Analyse herangezogen werden.

Neben Datenqualität spielt auch die Datenaufbereitung eine zentrale Rolle. Nachdem Werte unterhalb oder oberhalb einer Nachweisgrenze nachfolgende multivariate Datenanalysen beeinflussen können, ist es wichtig, solche Werte zuerst durch geeignete Schätzungen zu ersetzen. Während Methoden existieren zur Schätzung von Werten unterhalb einer unteren Nachweisgrenze, auch solche die den kompositionellen Aspekt der Daten berücksichtigen, wird hier eine neuartige Methode vorgestellt, die mit Werten umgehen kann, die eine obere Nachweisgrenze überschreiten. Nachdem diese Methode auf Regression aufbaut und daher multivariat arbeitet, hat sie Vorteile gegenüber herkömmlichen Strategien, wie z.B. das Ersetzen von rechts-zensierten Werten einfach durch eine Konstante.

Die wichtigste statistische Aufgabe bei der geochemischen Exploration ist die Prognose der Lokation von mineralisierten Zonen. Während Techniken der explorativen Datenanalyse diesen Prozess unterstützen, berücksichtigen sie normalerweise nicht den kompositionellen Charakter der Daten. Daher wird eine Methode vorgestellt, welche die Information der log-Verhältnisse aller Paare von Elementen berücksichtigt. Nach-

dem Datenunsicherheiten vorliegen, werden nicht die originalen Beobachtungen für die log-Verhältnisse genommen, sondern Werte, die von glatten Approximationen, hier von Verallgemeinerten Additiven Modellen, kommen. Ein Maß, das die gesamte Krümmung der log-Verhältnisse miteinbezieht, wird zum Ordnen der Paare verwendet, um so Elemente zu identifizieren, die Mineralisation anzeigen können. Diese Prozedur wird entwickelt für den Fall, dass die Proben entlang eines linearen Transektes angeordnet sind, und erweitert auf den Fall der Probennahme in zwei Dimensionen. Echtdaten zur geochemischen Exploration werden verwendet, um die Performanz dieser Methoden in der Praxis zu veranschaulichen.

Contents

Abstract **vii**

Kurzfassung **ix**

Contents **xi**

1 Introduction **1**
- 1.1 Considerations in sampling design . . . 2
- 1.2 Quality assurance and quality control (QAQC) methods . . . 15
- 1.3 Introduction to compositional data analysis . . . 21
- 1.4 Outline of this thesis . . . 24

2 Imputation of values above an upper detection limit in compositional data **27**
- 2.1 Introduction . . . 27
- 2.2 The log-ratio approach for CoDa . . . 29
- 2.3 Method . . . 31
- 2.4 Algorithm . . . 33
- 2.5 Example . . . 34
- 2.6 Numerical experiments . . . 35
- 2.7 Discussion and conclusions . . . 48

3 Identification of mineralization in geochemistry along a transect based on the spatial curvature of log-ratios **51**
- 3.1 Introduction . . . 51
- 3.2 Methodology . . . 54
- 3.3 Algorithm . . . 59
- 3.4 Experimental results . . . 62

3.5 Discussion and conclusions 72

4 A method to identify geochemical mineralization on linear transects 75
4.1 Introduction 75
4.2 Methodology 77
4.3 Results 78
4.4 Summary 84

5 Identification of mineralization in geochemistry for grid sampling using Generalized Additive Models 87
5.1 Introduction 88
5.2 Methodology 90
5.3 Algorithm 93
5.4 Experimental results 95
5.5 Discussion and conclusions 106

6 Summary 109

List of Figures 113

List of Tables 116

Bibliography 117

Curriculum Vitae 125

CHAPTER 1

Introduction

When designing a study for target scale mineral exploration based on data from surface geochemical sampling media, there are several aspects that need to be considered. The success of the study will heavily depend on a selected study area, and in an ideal case there is some pre-knowledge about potential mineralization available. The first step is a carefully designed sampling plan. Obviously, the more samples need to be taken, the more expensive is the study. Thus, there is a natural boundary of the maximum number of samples to be considered, and these need to be placed carefully in order to allow for an accurate prediction of the mineralization. Provided that the sampling has been carried out accordingly, the samples need to be analyzed in a laboratory, yielding concentration data for a number of different chemical elements. Note that usually also different sample media are considered, not only mineral soil samples (from different soil horizon), but also samples from organic materials (leaves, needles, bark, etc.). Once the concentration data are available, a careful quality check of the statistical data needs to be carried out. There are different aspects that are important, such as a high precision of the measurements, of the sampling procedure and sampling material. For subsequent statistical analyses it will also be important that the measurement repeatability is high compared to the spatial variability of the concentration values. Once the elements with reasonable data quality have been selected, the statistical analysis can start. One frequent problem in geochemistry are values below or above a detection limit. There are methods available that allow to replace those values by reasonable estimates. Preprocessing of the data by replacing the censored values is required to repair data structure for multivariate data processing and can lead to a significant improvement of the data quality, with consequences on the resulting multivariate data distribution. After this

step, the processed data are ready for a statistical analysis. Since the interest here is in mineral exploration, appropriate methods need to be considered that allow to predict locations of mineralization.

The first chapter presents general consideration for the sampling design in this context (Section 1.1). Once the samples are available, data quality check have to be carried out, and an overview of such methods is provided in Section 1.2. Geochemical data, like many other data sets as well, are special in a sense that the main interest is in relative rather than in absolute information. This means that these data need to be treated as compositional data. A brief introduction to this subject is provided in Section 1.3. The final Section 1.4 of the chapter gives an overview of the contents of this thesis, and refers to publications that resulted from this work.

1.1 Considerations in sampling design

Sampling design is part of the sampling planning process. In Webster and Lark (2012) and De Geoffroy and Wu (1970) a definition of the distinction between sampling design and sampling planning is found. Sampling design influences data collection and represents determining of sample size and density, location of the sampling stations, analytical strategies in such a way to be able to analyze the data spatially. Using the statistical methods in conjunction with geological and geochemical information we are talking about sampling design. However, the sampling design varies depending on the research question. Therefore, as a first step, before any sampling planning is done, the aim of the planned geochemical research must be clear. Once this is clear one can start with sampling planning. It is a process which should ensure that the initial sample sizes are large enough to yield the required number of assessments. Sampling planning should solve also the question about selection of sampling media and sampling procedure. The next goal is to obtain unbiased information which requires replicated observations according to an appropriate design, in other words, the desire is to avoid investigating nonsignificant sections and conversely, overlooking significant sections. The optimum sampling plan gives a trade-off between those two cases while keeping the cost-effectiveness of geochemical surveys (De Geoffroy and Wu, 1970). While searching for a distinction between both terms, sometimes the difference of these terms may diverge in literature.

Sampling design is the key for a successful target scale surface geochemical project. For the operational geochemical surveying the lack of optimal sampling design procedures is a major bottleneck. Sampling design is a crucial step of any geochemical study but especially in surface geochemical exploration which may produce very local anomalies.

Therefore, a mineralization may be missed if sample density does not reach the minimum spatial variability caused by the potential underlying mineralisation. To detect these types of anomalies, samples have to be placed such that mineralization will not be missed because of too sparse sampling. At the same time, budget is a constraint for projects and oversampling should be avoided not to create unnecessary costs. Background should be well covered spatially because background lithologies may be very variable and also include unknown mineralized lodes. Thus optimal sampling design would be preferred. If the sampling design is poor it may jeopardize the entire project. Although the issue is relevant from the perspective of successful geochemical exploration, it is not discussed very extensively in the literature. Garrett (1983) approaches it from the statistical point of view and Matthews (1996) from the practical point of view. Statistical theories on sampling design are available but often a practical approach (considering budget constraints) is taken which ignores the natural spatial variability of the underlying geochemistry (Gonzalez and Eltinge, 2010).

A limited amount of information is usually available for optimization the sampling design. The sampling design can be aided by:

- Geophysical exploration survey data which are acquired prior to target scale geochemistry. The magnetic, electromagnetic, gravity and radiometric data may, in special cases, be utilized to roughly divide the survey area into lithological units: geophysical exploration target, alteration around the target and lithological units at the background.

- Soil pH and/or self potential measurements (see, e.g., Hamilton, 2007) which may already indicate the location of an underlying mineralization as spatial anomaly patterns in the surface geochemical data. Possibly also electrical conductivity to the depth of 0-30 cm of terrain surface may be helpful to reveal the frequency of the spatial variance.

- Geochemical orientation surveys are conducted to choose the most appropriate surface geochemical sampling media, sample pre-treatment and analytics. With the available orientation survey data it is possible to employ the concepts of geostatistics to optimize sample density. Samples which are acquired with high enough frequency are spatially correlated, and the spatial variance can be estimated with a semi-variogram analysis. The lag distance revealed by a semi-variogram allows estimation of a minimum sampling distance.

Sampling design in practice is becoming increasingly popular especially amongst environmental scientists dealing with agriculture, forestry, environmental monitoring but also amongst geochemists. There is a demand for geochemists to be able to determine the chemical composition at not previously sampled locations, not only for the purpose of mineral exploration, but also for environmental geochemistry.

In these nearby sciences of geochemistry, a lot of literature is available on optimal sampling design. However, these proposed methods usually consider idealized situations, and require certain pre-knowledge. For the purpose of optimal sampling design for mineral exploration there is rarely any literature available. The theoretical background for optimal sampling design was developed during 1930s. One of the key books referring to sampling are Cochran (1977) and De Gruijter et al. (2006). In this literature, however, the methodologies are not devoted particularly to mineral exploration, where prior knowledge such as the length and orientation of the mineralization and the dispersion should be considered.

When designing a surface geochemical survey, a 2D surface projection of the geochemical dispersion patterns should be first conceptualized. Depending on the geological, topographic and climatic factors, the geometry of dispersal patterns may be more complex than imagined. Besides the geochemical deposit halo, also the secondary halo and the alteration halo should be taken into account. The difficulty in conceptualizing the dispersion patterns is caused by lack of geological subsurface knowledge, i.e. stratigraphy, ground water flow, information on bedrock fractures, etc. Any pre-existing data should be recovered, utilized and discussed amongst the exploration crew already at this point.

1.1.1 Geological considerations

Mineralization refers to a bedrock unit where chemical elements with potential economic interest are abnormally abundant in comparison to most common lithologies, forming mineralized bodies or "lodes". An exploration target in detailed scale may be an anomaly in geophysical data or more sparse geochemical data (least commonly an intuition based target), which is considered to be verified by bedrock drilling.

Sampling design is strongly dependent on the target type, i.e. its geochemistry and geometry. Geometrical properties of mineralization types vary considerably and it is difficult to reduce this diversity to simple models. Mineralization is rarely a homogeneous body with sharp contacts with host-rock. In some cases, it corresponds to dissemination at low concentration but in relatively large volume (high tonnage). Dissemination can be relatively homogeneous (e.g. mafic and ultramafic deposits or greisens) or heterogeneous (e.g. stockwork of quartz veins related to magmatic intrusions). In some other cases,

mineralization corresponds to continuous ore bodies that can be represented as layers, beds, strata-bounds or lenses, like banded iron formations, chromite or bauxite deposits or massive sulphide mineralization. The last principal type is related to veins, lodes, pipes or dykes (e.g. gold lode, pegmatites, etc.). Projection at surface of these morphologies is dependent to their extension, strike and dip. It is also important that a mineralizing system often affects the mineralization host-rocks, which results to alteration of lithologies with specific geochemical signatures that can be a leading feature towards an exploration target.

The sampling density and placing of the sampling stations should be adjusted according to the size of the explored mineralization type. A cost-benefit procedure would be to increase the sample density on top of the potential target as well as the surrounding halo, to give a higher resolution of the geochemical anomaly, and to decrease sample density away from the target in the background. This is challenging when no prior information is available. One would need a large amount of samples in order to be sure to have one placed on top of the target. That is often the case for vein type deposits, such as orogenic gold, that may be only a few meters thick. For large deposit such as a porphyry copper deposit, the target may be several kilometers wide (Robb, 2013), meaning the sample grid must be increasingly large to accommodate the geochemical background.

The target population is considered as the entity covering the predefined target area, with all available sample materials (soil types, plant species), and all available elements that can be measured. One of the key questions is the number and placement of the samples. Due to logistic issues, financial costs and efficiency, only a pre-determined portion of the target area can be sampled, that is called the sample population. Afterwards, statistical approaches may be used for appropriate statistical analysis and predictions.

From a practical perspective it is be desirable to have at least two sample points on the underlying mineralization. Because of uncertainties related to sampling (and several other factors), a better option would be to have at least three samples on the anomaly pattern. In order to get sufficient contrast, about two thirds of the samples should be on the background. These rules should be considered for sampling planning. Thus, the existence of pre-knowledge is already very important at this stage, and it would help to optimize costs. Moreover, if something is known about the size of the anomaly, or even on the dispersion pattern, this information can be successfully used to place samples on the target and at the background.

Once exploratory data analysis is provided which is looking at the concentrations after the lab data is received, a threshold for the upper limit of the local background variation can be determined. If this threshold is known, then the goal would be to obtain

e.g. three samples for which the concentration of the corresponding element exceeds this threshold. Consider grid sampling, either regular or samples located randomly in the grid. The expectation that a grid sample intersects with a target mineral resource is proportional to the area of the resource relative to the area of the grid. A practical approach for random sampling (within the grid) might be to inflate the effective area of the grid by a factor of 1.088 (Garrett, 2012).

The orientation of the geological structures affects the survey design such that the survey lines or grids are oriented so that they cross the strike of the expected target at right angles. If the results of an airborne geophysical survey or low, moderate or semi-detailed geochemical survey data are available, then the orientation of the geological structures can be estimated and the lines or grids can be oriented to cross the interesting anomalies at right angles.

1.1.2 Practical survey design considerations

There is a practical issue in the geochemical context. In exploration geochemistry, besides using the commodity elements, element identification of a mineralized body can be facilitated by using so-called pathfinder elements. Pathfinder elements can be trace elements which are often associated with commodity elements. For instance, pathfinder elements of gold deposits are empirically observed to be Cu, Ag, Zn, Cd, As, Bi, Pb, Sb, Hg, W, Mo and Se. Unfortunately, it is not clear yet which sample media and their leaches behold those elements in quantities measurable by commercially available laboratory analytics. The strategy in an orientation study is thus to sample different surface geochemical media, to obtain measurements for various chemical elements, and to investigate if statistically significant differences are observed.

Infrastructure and anthropogenic limitations to sampling area are significant considerations in surface geochemical exploration. Sampling in surface geochemical exploration should be restricted outside of the influence of roads, railways and other significant traffic infrastructure, habited areas, industrial areas and quarries for rock and gravel. The emissions through air and subsurface from the infrastructure vary case specifically. So does also the distance they affect the surface geochemical sampling media, thus buffers of different distances should be constructed in geographic information systems (GIS) to exclude them from the sampling area. All sources of contamination which may overprint the subtle surface geochemical signatures have to be considered.

Besides the infrastructure digital topographic databases can also be used to exclude landscape areas from the suitable survey area for mineral soil and plant samples. Waterbodies, bedrock outcrops, dense boulder fields and peatlands are not suitable sites for

sampling of these materials and can be similarly buffered and removed from the suitable sampling area using GIS. In this instance, logistical limitations should also be taken into account. On demanding terrains, access to sites because of water courses or ability to traverse on peatlands, it may be more cost-efficient to exclude areas from the suitable sampling area. In remote areas, one may also want to restrict the accessible area within a certain distance to roads to avoid excessive hiking in the field. In helicopter surveys the accessibility on foot is not a limitation though. When restricting the sampling area by logistical reason, it has to be ensured that the purpose of the geochemical survey will not be jeopardized. Even the most expensive samples (in terms of time or/and money) should be acquired to fulfill the goals of the sampling campaign.

For distinguishing a target from background it is important to consider some issues. Assume that preliminary experiments have already been carried out in the area of interest. Ideally, an orientation survey is available, which provides some details about element concentrations and their spatial variability. The element concentration of a (target) element under consideration for a mineral deposit will decrease with increasing distance from the target. Therefore, a plot of concentration versus distance from the target will be interesting to determine an average distance from the target, where the signal due to the mineral occurrence is lost in the geochemical noise of the background. This determines a threshold for the element concentration. However, the determination of the threshold is often difficult as even for a single occurrence type it will vary according to bedrock type and physical environment.

A further difficulty is the definition of background, since the background may vary with different type of media, location, size of the survey area and physical environment. According to Garrett (1983), the background represents an area below a particular threshold, resulting from an element concentration which separates the background from mineralization occurrence, more generally from the anomaly, where anomaly refers to a deviation from the norm.

A general assumption is that the main interest in mineral exploration is on unusually 'high' values. However, we should not forget another anomaly patter which occur – negative anomaly, and the so-called rabbit's ear anomaly (Hamilton, 2000). One needs to consider, however, that

- in the background there is always a lot of natural geological variability, also resulting in false high values from the perspective of the mineral exploration;
- thus it is important that the sampling is guided such that the variability of the background will be covered.

1.1.3 Statistical survey design considerations

Simple random sampling is inefficient. The reason is that especially for small areas of mineralization one would need many observations in order to obtain at least 3 samples on top of the mineralization when placing the samples purely in a random manner. Therefore, guidelines for sampling are needed which increase the probability of sampling on the target. We are looking for a trade-off which ensures reasonable costs for sampling and also reaching the target. Of course, this heavily depends on the size of the mineralization type under investigation, but also on its orientation and shape.

Number of sampling locations

Plentiful literature is available on selecting appropriate sample size for a survey, for example in Webster and Lark (2012); Garrett (1983). However, most of the literature refers to clearly controlled experiments, and typically to independent observations meaning two samples taken from the same population but with no relation to each other such that presence of one observation does not have effect on the other one and conversely. In the context of spatially dependent data, meaning that dependency is linked to influence of locations placed in neighborhood, these theoretical approaches have clear limitations. On top of that, budget constraints usually give an upper bound on the number of samples. Existing data can guide the sampling in a way that sampling density is higher on the exploration target and its expected geochemical halo, and in the background it is lower.

In the following it is established known principles how to get an estimation of the number of samples. It brings us to the following question: how many samples should one take? The question is commonly passed to statisticians, therefore let us define N as the desired sample size. Considering a population from a normal distribution with mean μ and standard deviation σ. Then the distribution of the sample mean is normal with mean μ and variance σ^2/N. The variance of the mean, or its square root – the standard error – decreases only slowly if $N \geq 30$. Thus it can be argued that $N = 30$ might be sufficient. This has been standard practice for many years in geochemistry studies in the past. However, Stanley et al. (2010) proposed to consider the coefficient of variation (CV%) to have a more appropriate estimate of the N. Their assumption is that there is an estimate of CV% based on a pre-study conducted in the study area or obtained from a similar geological environment. Then a tolerance level ε (in %) is fixed for the estimation of the mean in the new study. Then the estimated minimum sample size is $N = (\text{CV}\%/\varepsilon)/2$.

The coefficient of variation is based on classical estimators, i.e. the sample mean and the sample variance (standard deviation). These estimators are very sensitive to

outlying observations. Thus, in presence of outliers, which is likely in studies concerning mineral exploration, one should use more robust estimates. A robust counterpart to the arithmetic mean is the median, and a robust counterpart to the classical standard deviation is the MAD (Median Absolute Deviation).

A further reason for outliers can be skewed distributions. Geochemical concentration data are typically right-skewed, and a log-transformation can make the distribution more symmetric.

A completely different approach is based on the concept of compositional data analysis (Filzmoser et al., 2018). Here, the main interest is not in the concentration values themselves, but in the ratios between different elements, taken at the same location. In the simplest case, one would consider only one pair of variables (elements), and use the log-ratio of this pair. If one of these elements is a pathfinder element and the other one is a reference element being more stable in the area, the log-ratio would in the ideal case lead to a peak on top of the mineralization, and more or less constant values on the background. Again, it will be crucial to have sufficiently many observations on top of the mineralization (at least 3), and sufficiently many on the background (1/3 or 1/4). A rule of having at least $N = 30$ is thus again the lower minimum.

Geometric survey configurations When planning a sample survey, there are two basic options: line sampling and grid sampling. When designing a geometric survey configuration of sampling location, the purpose of the survey is the most determining factor. Line surveys are most commonly used in the orientation stage of the exploration project and set up perpendicular to geological structures. Line surveys, i.e. transect sampling, a limited number of transects are used to estimate the properties of the entire region. Thus, the disadvantage of line sampling is that they provide only a limited view on the spatial geochemical patterns. To compensate for this it is possible to use several (parallel) lines, again depending on financial restrictions. Line sampling has the appeal that the results can be easily presented in a two-dimensional plot, with the distance between sampling stations on one axis, and the concentration on the other axis.

An orientation survey which is conducted to choose the appropriate sampling density, sample material and analytical method can be recommended to be done as a line survey. Orientation surveys may be conducted in a geochemical project prior to the collection of the survey samples. The purpose of an orientation survey is to optimize the sampling density and decide what geochemical sampling media to collect if any or which analytical methods to apply on the samples. An orientation survey is most often conducted prior to large projects. In small projects they are not cost-efficient. The problem of conducting

an orientation survey is that it delays the beginning of the survey. In climate zones with very short sampling seasons this might even delay the project by one year. However, conducting an orientation survey may be the only efficient way of determining the optimal sampling density. In addition, orientation surveys may be crucial for finding the correct sampling media/analytical technique.

In the real exploration cases, the samples are mostly planned on grids to visualize the spatial patterns better. The density of samples on a line has to be high enough (over-sampled) to catch the spatial variation in a semi-variogram. In target scale exploration it is often advantageous to understand the spatial anomaly patterns not only along a line but in a 2D map space. The samples are then planned on grids to visualize shape of the geochemical anomalies better. A properly conducted grid design requires usually a much higher number of samples than a line survey. Budget constraints may be the first issue when deciding between line and grid sampling for an exploration crew but it should be always carefully considered if a line survey will really answer the question to be looked for.

In a grid survey, the samples can be placed either on a regular grid, any variation of unaligned grids, clustered designs or random sampling design. Stratified sampling approaches can be used, including unbalanced sampling designs which allow to significantly reduce the number of required samples. The usual strategy for grid sampling is to determine the maximum number of samples (usually based on financial restrictions), to place a regular grid with as many grid cells as the number of samples, and then to randomly select a location in the grid (or select that location which is feasible for being sampled).

Line surveys

Line surveys refer to a sampling arrangement that follows a line, not necessarily a straight line. The results of a line sampling campaign can be conveniently presented in 2D plots to illustrate the variation in the data. Examples can be seen in Figures 1.1 and 1.2 where Co concentrations in crowberry twigs from the Juomasuo Au-Co deposit acquired in the year 2013 are illustrated (Torppa and Middleton, 2017). In Figure 1.1, Co concentrations are illustrated on a map, i.e. x and y coordinates are displayed for each sampled point for a particular location. Colors refer to clusters computed by quantiles. A rainbow color palette was chosen such that red color represents high concentrations and blue low concentrations. In the Figure 1.2 we can see a 2D plot for the same data. In order to extract the 2D plot one needs to compute distances between two points and for this purpose the Euclidean distance calculation in `R` was performed. The distance in meters can be seen on top of the map and on the x-axis the sampled points and their

numbers, such that the subsequent points are always closest to each other. Colored lines at the bottom of the plot refer to known mineralized lodes of the Juomasuo Au-Co ore situated at different depths. The advantage of the 2D plotting is that the fine variation in the data and the spatial patterns the data may contain can be better visualized than in a clustered map presentation (see in Figure 1.1).

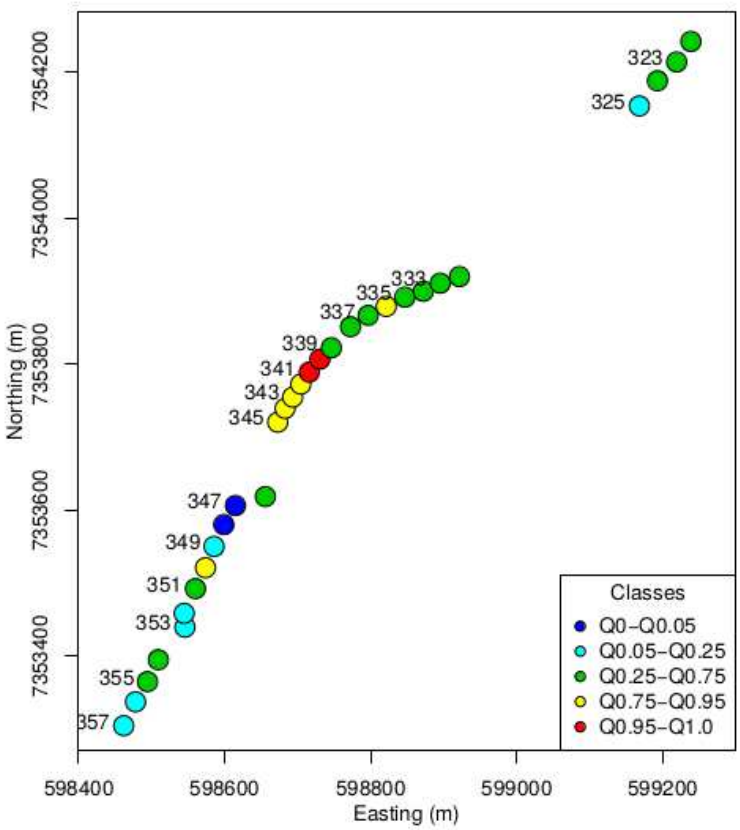


Figure 1.1: Map presentation of cobalt concentrations in Cowberry-twig (modified from Torppa and Middleton, 2017). Concentration values are clustered according to quantiles in the data. Low concentrations are presented with blue and high with red.

Visualizing the variation in the data, which might be significant information caused by geochemical variation or insignificant noise in data, requires a display of the data either in 2D or, in case of grid sampling, in 3D draped surfaces (such as elevation data is often presented), without loosing information. However, grid sampling better brings out geometry of the spatial geochemical anomaly patterns. A compromise is that the parallel lines are combined to compose grid sampling design. A grid is composed of lines which

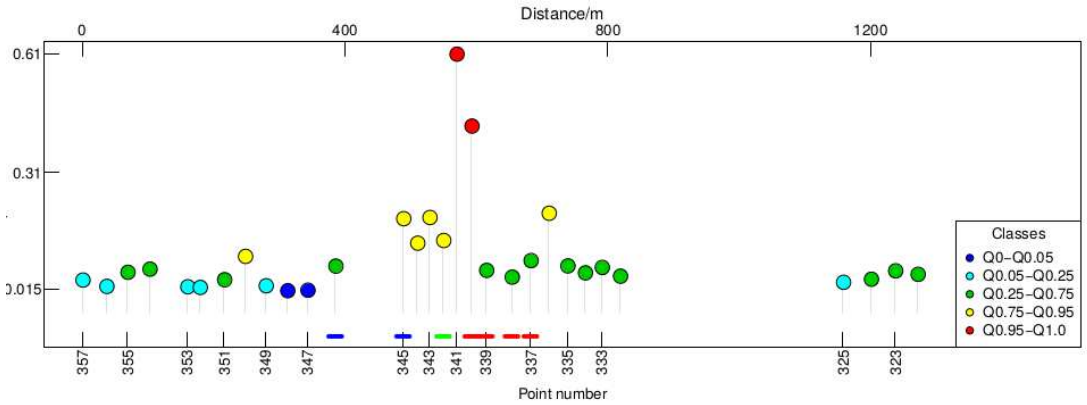


Figure 1.2: Line plot for cobalt, Cowberry-twig/stem (Ultra-LIM project), see Torppa and Middleton (2017).

can be illustrated as single line plots.

As mentioned above, the orientation of the target can be very crucial for sampling design. The orientation can be obtained from geophysical data, geological mapping and geological observations of bedrock outcrops. When line sampling is carried out, the lines should be placed orthogonal to the orientation of the mineralization. Strike and dip of the expected mineralized structures have also an importance for sample location as they should control the shape of the anomaly pattern. This information can be deduced from the local tectonic and structural settings (from geophysical survey or field maping).

Grid sampling

Grid sampling is a favored method for site specific soil because it is unbiased, simple, relatively quick and software exists to facilitate it. It is important that the grid covers the whole are of interest. However, in order to be effective (also concerning costs), a strategy should be used which ideally gives higher priority to regions of potential mineralization as sample locations.

Grid samples can be designed by ellipses, square, rectangular, triangular and hexagonal grids. For better understanding the grid designs are illustrated in Figure 1.3, which show the three types of grid sampling. On the left side the random sampling is examined, in the middle is systematic regular sampling and on the right side of the figure stratified sampling is plotted.

Simple random sampling ensures that every member of the population has an equal chance of being chosen. The disadvantage of this method is that important sample points may not included, and not all of the samples may be representative. Further,

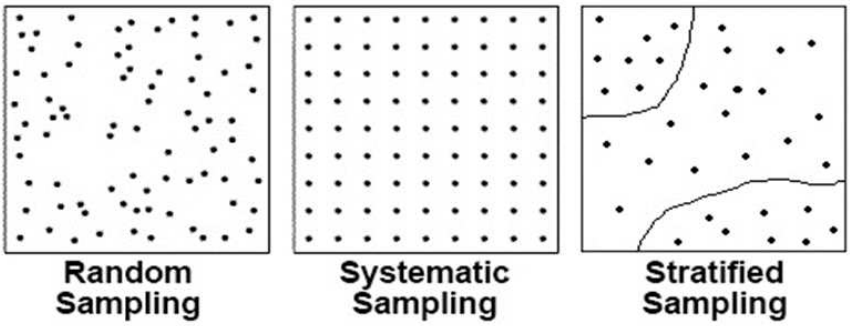


Figure 1.3: Three types of sampling designs (Humboldt State University, 2018).

simple random sampling is inefficient. The reason is that especially for small areas of mineralization one would need many observations in order to obtain at least 3 samples on top of the mineralization when placing the samples purely randomly. Therefore, guidelines for sampling are needed which increase the probability of sampling on the target. We are looking for a trade-off which ensures reasonable costs for sampling and also reaching the target. Of course, this heavily depends on the size of the mineralization, but also on its orientation and shape.

Since simple random sampling following a Poisson distribution brings less efficiency, it is desirable to use other sampling techniques, such as cluster sampling where the sample density at a potential mineralization is increased, leading probably to highly correlated samples. Systematic sampling is considered to be more efficient, but this might require a lot of pre-knowledge, and the resulting estimates might be biased. Finally, stratified random sampling provides a break for sub-regions, and gives unbiased estimates of known variance.

In stratified random sampling the potential area is divided into several small sub-regions of pre-determined size, so called strata. Then a random sample is taken from each stratum. This method requires prior knowledge for defining the strata.

Systematic random sampling is defined as being ideally randomly ordered, providing the most even cover. It preserves the maximum distance between sampling locations and the nearest grid node is the least for a given density of sampling. The disadvantage might be that small mineralization areas are undersampled.

Sampling on a regular grid belongs to a group of systematic sampling, with regularly defined intervals of transects, and within these intervals one sampling point is not randomly placed. This type might be used to achieve target precision of predictions by

kriging. Regular grids are not considered as optimal but still provide spatial coverage. Stratified random grid differs mainly because of randomness of the placed sampling locations within regularly set intervals. Regular sampling and stratified sampling are according to Gallego (2005) more efficient than simple random sampling.

Three grid survey designs have been recommended for mineral exploration purposes: regular grids, stratified random within a grid (Garrett, 2012) and offset grids (Heberlein and Dunn, 2011).

Optimal sampling density

For small mineralization one has to make sure that there are at the very least 3 samples on top of the exploration target. For large scale ones it must be ensured that enough samples will be in the background.

Spatial covariances and the variogram play crucial roles in the description of regionalized variables and in their prediction by kriging. The sampling for kriging should provide an even coverage of a region and can be designed to be efficient once the variogram is known. Kriging, or in other words Gaussian process regression, is a method of interpolation where the interpolated values are modeled by a Gaussian process determined by prior covariances. Following particular assumptions, kriging gives the best linear unbiased prediction of the intermediate values. Sampling for estimation of the variogram requires some degree of spatial nesting. The joint objectives of efficiency for kriging and variogram estimation can be met by a single sampling plan.

The (semi)-variogram which is the basis for kriging provides the information of spatial dependency. The worst situation from perspective of interpolation is a pure nugget effect, which means that even at very small distances, the variance is large – probably as large as for a longer distances. Generally, the prediction will work reasonably well if the variance is small at a small distance, and gradually increases up to a certain higher distance, and at higher distances it remains unchanged. Clearly, the shape of the estimated semi-variogram depends on the sample locations. If the locations are too far from each other, local variability cannot be identified, and the variogram provides an impression of a pure nugget effect. Geochemical knowledge of element distributions shall be studied here, and several elements typically should not show a pure nugget effect. Consequently, the sampling density needs to be increased in order to cover small-scale variability. This strategy can also be used for specific directions, since variograms can be computed along certain directions. Thus, variogram estimation, e.g. based on an orientation survey results, is a very appropriate tool to guide sampling planning and to identify the optimal sample density.

One should also consider that a lot of costs are devoted to reaching a sample location, and the costs for analyzing an additional sample are comparably low. Thus, it is advisable to collect more samples during the field survey, and to analyze them only if the variogram is not sufficiently covering local variability.

Conclusions

Most geochemical surveys are compromises: resources for survey execution are finite. Use of the above described approaches allows various sampling designs to be investigated and evaluated for cost efficiency and geochemical effectiveness prior to implementation.

In all geochemical surveys there are costs in financial and human resources and time, there is the area to be covered, and there is the confidence with which a statement concerning the objective can be made following the survey. Any two of those can be fixed.

There is a lot of literature available on sampling design. However, the theoretical approaches are based on quite strict model assumptions, which are typically not met in exploration geochemistry (spatial dependency, not normally distributed, compositional data, measurement errors, etc.). Moreover, the task is very specific, since one is looking for a design such that a potential target in space (in 3 dimensions) with unknown form, width and orientation would not be missed. The only way to avoid taking hundreds of samples is to incorporate prior knowledge, and it very much depends on the kind of prior knowledge how to proceed. In this section, some ideas and recommendations depending on the type of prior knowledge have been provided.

The ultimately final sample design is the choice of the project geochemist.

1.2 Quality assurance and quality control (QAQC) methods

For carrying out geochemical survey, many steps are involved including sampling, sampling planning, collecting samples, analyzing samples, treating and analyzing data, and finally interpreting the results. The success of the study requires intensive communication between the experts responsible for each individual stage, i.e. mainly between geochemists, analyst in the laboratory, and statisticians/data analysts. Including a good professional practice is necessary to come up with reliable outcome with reasonable quality. One could argue that there is no need to examine and assess the quality of the geochemical data since nowadays most laboratories are nationally accredited, meaning they have the technical competence to perform specific types of measurements following strict

quality control procedures. However, the geochemists and environmental scientist should still be concerned about the quality of their collected data, and also externally monitor the quality of the laboratory results. It is always possible that at some stage in this process mistakes are made, and the source of mistakes can be various, e.g. lack of time or resources to adequately examine data quality, inadequate information given to the laboratory, human fault/error, software mistake, etc. Therefore, any project in applied geochemistry should be subject to the quality check done externally, and not only by the laboratory.

This procedure of checking the data quality, namely Quality Assurance and Quality Control (QAQC), is nowadays a standard practice in all geochemical data handling and should be performed carefully to get an assessment of the quality. A simple definition of quality control can by summarized as efficient, cost-effective and legally defensible. Moreover, the data analysis can only be as good as the quality of the data. The aim of building this protocol is to mainly understand the limitations of the data, quantify the uncertainties, build the confidence on the geochemical data and select elements with reliable quality. The QAQC procedure is the first step of the data analysis followed by the statistical data analysis and interpretation of the results. The technical bottleneck for a complete geochemical consulting business is currently still the incomplete QAQC procedures applicable to surface geochemical techniques. The procedures are routinely applied in large geochemical projects but the concept in medium and small projects needs to be refined and streamlined. Therefore, one of the goals of this work is to develop these analysis techniques. It is crucial that the general overview of the data quality is not just based on statistics, but also on expert knowledge of professionals in geochemistry. Nowadays, a lot of data can be measured, which generally leads to the problem that large amounts of data are available for the statistical analysis. However, any statistical analysis might be misleading if data with low data quality are entered. Therefore, there is an even more urgent need for a thorough quality assurance. This can be done by reporting statistical numbers, but also by displaying/visualizing the data by several appropriate plots.

The QAQC can be an automatic process but the interpretation of the outcome requires the expertise of geochemists. An overview of whole QAQC procedure can be found e.g. in Reimann et al. (2008). We claim that it is really important to view the data in different ways to understand the data quality, structure and content. Measures of uncertainty provide deeper insights of the data, and this is revealed by measuring data accuracy versus precision. Accuracy gives an overview if the results are comparable over time, and precision – how good the repeatability of the results is. Moreover, this analysis

gives an output selection of elements useful for further statistical analysis. Note that each kind of further analysis may require a different choice of elements.

The core part of this procedure is a fast production of the standard QAQC measures and figures in order to devote time also to data interpretation. The guidelines are presented in stepwise order beginning with pre-processing of the data and providing the first overview of the data. A quality control procedure of a project itself should include the following steps:

1. Data overview
2. Process quality
3. Laboratory precision, accuracy and trend
4. Laboratory contamination
5. Field precision

1.2.1 Data overview (QAQC 0)

For each chemical element of a particular sample medium, information about the measurements as well as descriptive statistics are provided. Basic characteristics are, for instance, the measurement unit, the amount of censored data, and the detection limit(s), the number of discretized samples, mean and standard deviation, but also robust counterparts. Depending on the type of measurement, other quantities can be of interest as well.

1.2.2 Process quality (QAQC 1)

The main purpose of this phase is to detect blockiness, i.e. abrupt changes, analytical periodicity, trends, i.e. drifts, outliers by sample mix-ups, censoring and discretization. The plotting of concentrations is done in the same order as samples have been analyzed in the laboratory. Blockiness can be an alarming mistake done by the laboratory, because it would point at time-dependent measurement differences. Outliers can indicate contamination. This phase answers the question how repeatable is a measurement. Routine samples and field duplicates are included in this stage. Requirements for the laboratory are: the samples must be analyzed in a random order (the order should not be changed by the laboratory) and it should also be required that all samples are measured on just one instrument. Figure 1.4 shows an example of a Shewhart chart (right) together with the legend (left) – the concentration of an element in the order of the analysis.

Displayed data are taken as an example from UltraLIM (Torppa and Middleton, 2017) data set, location Saivel is selected, samples of these plant species are acquired in 2013.

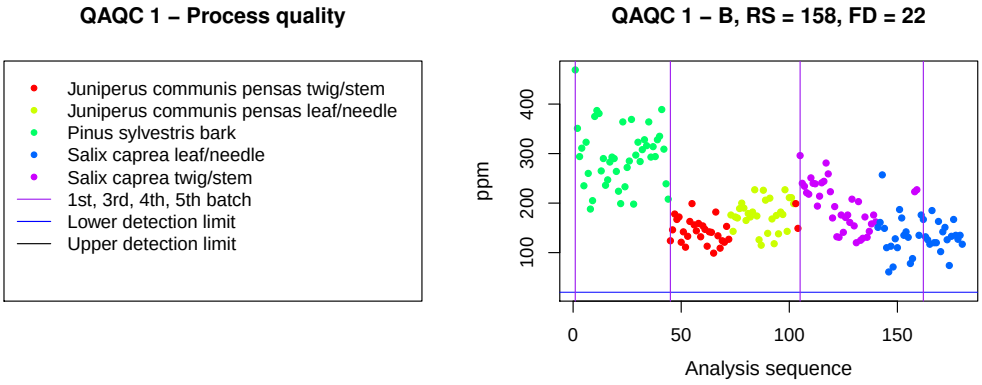


Figure 1.4: Legend for QAQC 1 phase, different colors are used for different plants (left). QAQC 1 plot shows potential blockiness or periodicity (right). Samples were not randomized for the analysis. Thus periodicity cannot be estimated.

1.2.3 Accuracy (QAQC 2)

The purpose of external monitoring of laboratory analytics is monitoring of laboratory precision, laboratory accuracy, i.e. reduction of bias, and analytical trends. When evaluating laboratory precision we compute the relative standard deviation (RSD%) of the reference samples and compare it with the preanalysed project samples. The RSD% is the standard deviation divided by the mean, expressed in percent. A standard rule of thumb is usually used, such as RSD% should be less than 20%. It allows us to compare the variation among all variables. Then, laboratory accuracy can be quantified in term of bias% which compares the mean of the reference samples analyzed along with the project samples to the "real" mean of the reference samples. Note that absolute values are not critical in mineral exploration, therefore the accuracy has a secondary importance. In other words, QAQC 2 monitors how close a result is to a true or accepted value if an appropriate SRMs are available to be inserted into the analysis sequence. Possible sources of error originate from the laboratory analysis. The sample types which can be included in the analysis are CRMs (certified reference materials), SRMs (standard reference materials) and PRMs (project reference materials). Good laboratory precision usually means that most samples fall within the limits of 1st SD calculated of the aliquots taken from a

homogenized SRM bulk. Poor laboratory accuracy can be seen as a systematic error, i.e. bias, however it is not crucial from the perspective of surface geochemical exploration. We are rather interested in the relative differences of concentrations. It is alarming when analytical trends occur; in that case the laboratory should be contacted. The example in Figure 1.5 shows rather good laboratory precision. Selected data are the same as for 1.4, only different element is show, namely strontium (Sr).

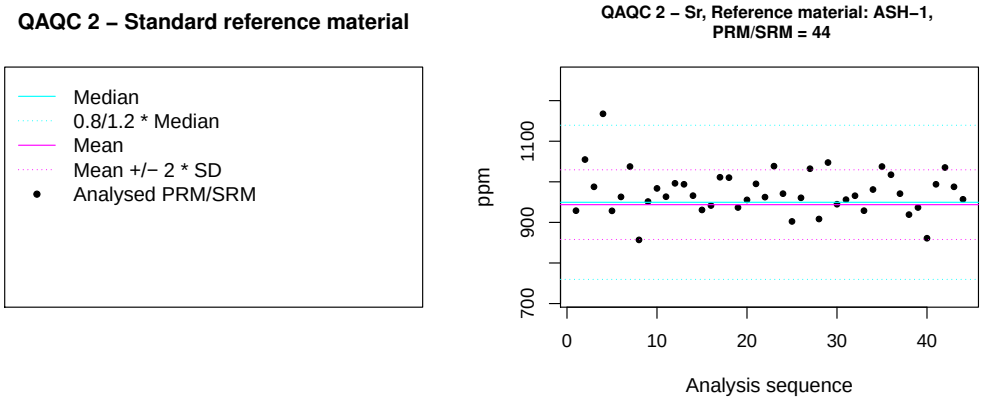


Figure 1.5: Legend for QAQC 2 (left) – Lines indicate measures of standard reference samples (SRM). Plot for QAQC 2 – Analysis sequence versus measured standard reference samples (right).

1.2.4 Laboratory contamination (QAQC 3)

The purpose of QAQC 3 is to estimate the relative impact of the possible element specific laboratory contamination on the analysis results by plotting the analysis results of blank samples in relation to the project sample concentrations. If project sample concentrations are small, even slight deviations of blank samples from zero indicate relatively large contamination. Blank (zero-concentration) samples are inserted by the laboratory and used at this stage of the procedure. Figure 1.6 shows an example of a boxplot comparison for the concentration values of the project samples in comparison to blank samples. Similarly to previous examples, selected subset is chosen for Saivel location, year 2013 and chemical element is mercury.

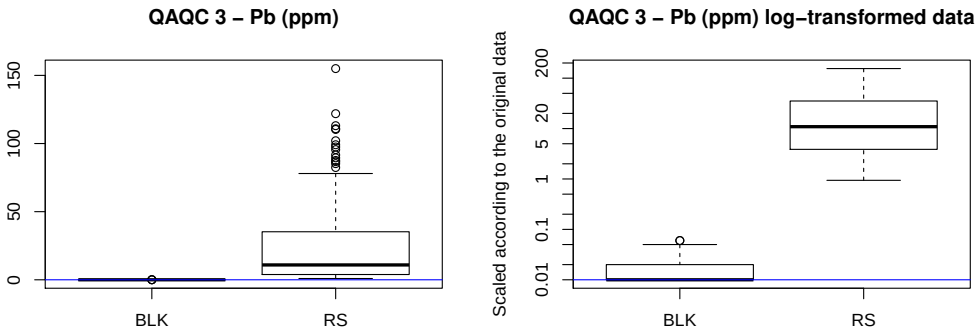


Figure 1.6: QAQC 3 boxplots of the routine sample and blank sample analytical results in a chart for the element Pb for original data (left) and for log-transformed data (right). Blue line indicates lower detection limit.

1.2.5 Field precision (QAQC 4)

The purpose of field precision is to monitor sampling collection quality, and depending how the field duplicate samples were collected, also spatial variance of the sampling media. The intention is to identify how repeatable the sample collection is, and also the geochemical spatial natural variability can be investigated. The goal is then to investigate the relative difference in concentrations of duplicate field samples and quantify the added uncertainty caused in all phases of material handling. Possible sources of error might be caused by sampling, sample handling or laboratory analytics. The example in Figure 1.7 shows the modified Thompson-Howarth plot, where the y-axis shows the absolute difference of the concentration values of a sample pair, divided by the mean of the sample pair. The lines refer to precisions of 10% and 20%. In this example we see good field precision – the S values of sampling material plant are below 20% difference and above the LDL. In this case a subset originating from UltraLIM data refers to Juomasuo location.

1.2.6 Laboratory precision (QAQC 5)

The goal is to monitor laboratory data quality externally. QAQC 5 tells us how repeatable the laboratory analysis is, including milling, ashing, sieving, digestion and analytics. The procedure enables detecting repeatability of analytical routines and another purpose is investigation the relative difference in concentration of duplicate laboratory analyses.

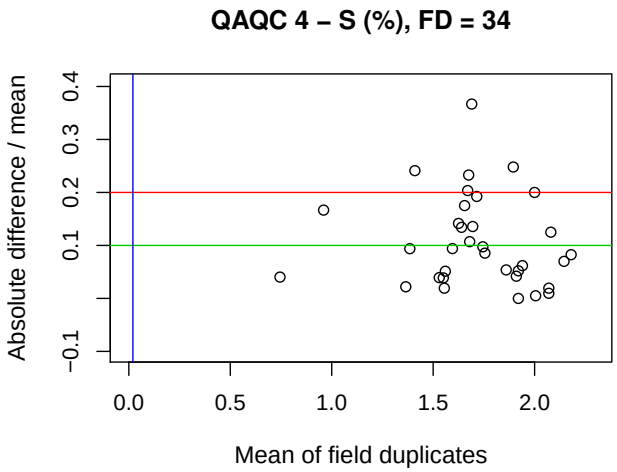


Figure 1.7: QAQC 4 for field precision.

Source of error may appear within a phase in the analytical procedure in the laboratory, e.g. ICP-MS analytics. Note that large enough samples have to be collected to make a split in the lab. The samples have to be labeled in the electronic sample list in order for the lab to make the split on predefined samples. The same type of visualization is used as for QAQC 4, see Figure 1.8. Here, the sample material is not plant but soil, namely for analytical method Bioleach ICP-MS. This example shows the modified Thompson-Howarth plot with good precision.

1.3 Introduction to compositional data analysis

Not only in geochemistry, but also in many other fields, for instance in chemometrics, economics, geology, etc., it is natural to consider the data as compositions. Due to their representation, such data commonly occur in proportions, percentages, ppm (parts per million), ppb (parts per billion), or other units which refer to relative information with respect to some underlying whole or total. Thus, it is natural to deal with these data as compositional data, where only the ratios between the variables, so called compositional parts, contain the relevant information. This leads to a different way how to statistically process the data, namely to *Compositional Data analysis* (CoDa). This type of multivariate data analysis represents relative contributions of parts on a whole. In other words, absolute values do not carry the important information, but it is the

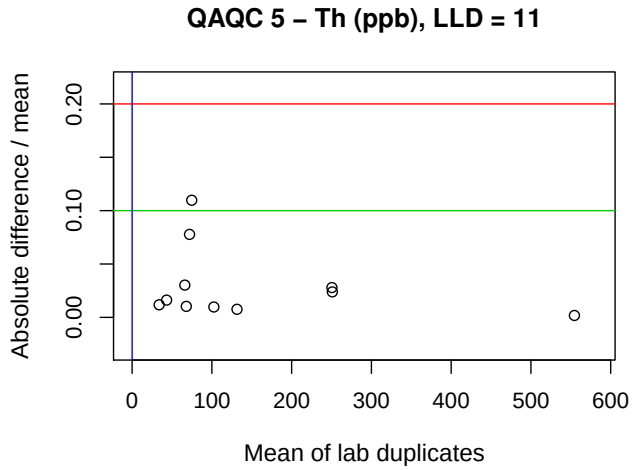


Figure 1.8: QAQC 5 for laboratory precision.

relative information that is of major interest. Using classical statistical methods when the data are compositional by its nature can lead to invalid conclusions.

In fact, the majority of methods for CoDa is based on the so called log-ratio approach, which is mathematically and geometrically concise and easier to handle. The log-ratio methodology was initiated by Aitchison (Aitchison, 1986), and this concept led to a special type of geometry, named as the Aitchison geometry. It was proven that the usual Euclidean geometry is not appropriate for compositional data (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). Note that closure of the data, i.e. the sum of the parts is equal to a constant, is not important. This is because of the property of log-ratios which avoid the phenomena of a constrained sample space. The use of log-ratios makes it necessary that compositional data are positive by definition, although there also exist approaches to deal with zeros in the compositions (see, e.g., Filzmoser et al., 2018).

Consider a composition $\mathbf{x} = (x_1, \ldots, x_D)^t$ with D parts. The sample space of a composition is the simplex $\mathcal{S}^D$, given by

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)^t \in \mathbb{R}^D | x_i > 0, \sum_{j=1}^{D} x_j = \kappa \right\}, \tag{1.1}$$

where κ can be any arbitrary positive real number. The idea of the log-ratio approach is to represent the compositions in the usual Euclidean geometry. The basic information are pairwise log-ratios $\ln(x_k/x_l)$, for $k, l \in \{1, \ldots, D\}$ and $k \neq l$, and thus the sum of parts, κ, does not carry important information. There are several advantages of using log-ratios,

such as the equality of their variances by exchanging numerator and denominator, thus $var(log(x_k/x_l) = var(log(x_l/x_k))$. This property is not possible when only ratios of the parts are applied. Further properties of log-ratios, such as scale invariance, are described for instance in Filzmoser et al. (2018).

The log-ratio approach is the basis for building up different transformations to represent compositions in the Euclidean space. A first transformation introduced in Aitchison (1986) is the additive log-ratio (alr) transformation, defined as

$$\text{alr}(\mathbf{x}) = \left(\ln\frac{x_1}{x_D}, \ln\frac{x_2}{x_D}, \ldots, \ln\frac{x_{D-1}}{x_D} \right)^t. \tag{1.2}$$

For the denominator one can also use a different part, and Aitchison (1986) discussed possible choices from an application point of view. The alr transformation is not isometric, thus distances in the original data space are not preserved, which excludes this transformation for many statistical purposes (Pawlowsky-Glahn et al., 2015).

Another transformation which is frequently applied is the centered log-ratio (clr) transformation, defined as

$$\text{clr}(\mathbf{x}) = \left(x_1^{\text{clr}}, \ldots, x_D^{\text{clr}} \right)^t = \left(\ln\frac{x_1}{g(\mathbf{x})}, \ldots, \ln\frac{x_D}{g(\mathbf{x})} \right)^t, \tag{1.3}$$

where $g(\mathbf{x}) = \sqrt[D]{\prod_{j=1}^{D} x_j}$ represents the geometric mean of the composition. In contrast to alr, the clr transformation preserves distances, and thus it is an isometric mapping. Further, it also preserves the same dimension D as the original data set but leads singularity, because the new D components sum up to zero, $\sum_{j=1}^{D} x_j^{\text{clr}} = 0$. This is undesirable for many statistical methods, because the resulting covariance matrix does not have full rank D. Nonetheless, the new clr variables are useful for the interpretation because they refer to a dominance of the corresponding compositional part on an average behavior (geometric mean) of the values in the composition.

Another conceivable transformation is the so called isometric log-ratio (ilr) transformation, which represents the information from the simplex in the usual Euclidean geometry by constructing orthonormal coordinates (Egozcue et al., 2003). As the name suggests, this transformation is an isometric mapping from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$. In fact, there are infinitely many possibilities to construct such orthonormal coordinates, and one specific choice are so called pivot (log-ratio) coordinates, defined as

$$\text{ilr}(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})^t \tag{1.4}$$

with

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}, \quad j = 1, \ldots, D-1. \tag{1.5}$$

Specifically, the first coordinate z_1 has a clear interpretation for x_1, as it is proportional to the respective clr variable for x_1. Further, x_1 is only contained in z_1, but in none of the other coordinates, which is different from the clr transformation, where each part is contained in each clr variable by the geometric mean. The coordinate z_1 thus extracts all relative information about x_1, and can therefore be interpreted "in terms of" x_1. If another part should be in focus for the interpretation, this part needs to be reordered to the first position. An interpretation of the remaining coordinates is not always straightforward. More discussion on the use of pivot and related coordinates can be found in Filzmoser et al. (2018).

The chapters 3, 4 and 5 build on pairwise log-ratios, which are also the building blocks of the transformations described above. Pairwise log-ratios $\ln(x_k/x_l)$ refer to relative information, and all different (and relevant) $D(D-1)/2$ pairwise log-ratios can be expressed by the $(D-1)$ ilr coordinates without any loss of information. The ilr coordinates, however, aggregate the pairwise log-ratio, because for instance

$$z_1 = \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{\sqrt[D-1]{\prod_{k=2}^{D} x_k}} = \sqrt{\frac{1}{D(D-1)}} \left(\ln \frac{x_1}{x_2} + \ln \frac{x_1}{x_3} + \ldots + \ln \frac{x_1}{x_D} \right).$$

The pairwise log-ratios thus have no geometrical meaning like in terms of coordinates, but they are very easy to interpret. This was also the main motivation for their consideration, because methods for mineral exploration in geochemistry need to result in a clear meaning and interpretation in terms of the involved chemical elements.

1.4 Outline of this thesis

The main part of the thesis is devoted to data preprocessing and prediction of mineralization. Chapter 2 presents a new methodology to estimate values above an upper detection limit. While methods exist to estimate left-censored values, i.e. values below a lower detection limit, this is a novel multivariate approach to estimate right-censored values. Chapters 3–5 propose and test a new methodology to predict mineralization. This method makes use of the information of pairwise log-ratios, i.e. the logarithm of the ratio of the concentration values of element pairs in a specific sample medium. The methodology is able to identify so-called pathfinder elements, which are variables that are

indicative for a particular mineralization, it points at the location of mineralized zones, and it allows to rank the different sample media according to their predictive power to identify mineralization.

Chapter 2: A regression based procedure for imputation of values above an upper detection limit is proposed. This method takes into account the compositional nature of the data. Comparisons with traditional procedures are given, and R code for the new method is provided.
D.M. worked out the theory, did the R implementation and the numerical experiments, and wrote the essential parts of the paper.

D. Mikšová, P. Filzmoser, and M. Middleton. Imputation of values above an upper detection limit in compositional data. *Computers and Geosciences*. To appear. doi.org/10.1016/j.cageo.2019.104383

Chapter 3: A novel methodology for identifying mineralization is proposed. The samples have to be taken on a linear transect. The method uses all variable pairs and approximates the values of the pairwise log-ratios with fits from generalized additive models. Peaks in the resulting curvature indicate mineralized zones.
D.M. contributed to the theoretical development, implemented the methodology, did all the numerical experiments, and wrote the essential parts of the paper.

D. Mikšová, C. Rieser, and P. Filzmoser. Identification of mineralization in geochemistry along a transect based on the spatial curvature of log-ratios. *arXiv, (1912.02867)*, 2019.

Chapter 4: The methodology proposed in Chapter 3 is applied to two data sets from Greenland and France that have been sampled for the purpose of detecting geochemical mineralization.
D.M. contributed to the theoretical development, implemented the methodology, did all the numerical experiments, and wrote the essential parts of the paper.

D. Mikšová, C. Rieser, P. Filzmoser, S. M. Thaarup and J. Melleton. A method to identify geochemical mineralization on linear transects. Accepted in *Austrian Journal of Statistics*. To appear, 2020.

Chapter 5: An extension of Chapter 3 to samples taken from a grid is proposed. The methodology is presented, and numerical challenges in this context are described. The experiments based on real geochemical data show that this approach indeed points at areas with mineralization.

D.M. contributed to the theoretical development, implemented the methodology, did all the numerical experiments, and wrote the essential parts of the paper.

D. Mikšová, C. Rieser, P. Filzmoser, M. Middleton, and R. Sutinen. Identification of mineralization in geochemistry for grid sampling using Generalized Additive Models. Submitted.

Chapter 6 provides a summary of the thesis.

Further publications in the framework of this thesis:

- B. Lemiere, J. Melleton, P. Auger, V. Derycke, E. Gloaguen, L. Bouat, D. Mikšová, P. Filzmoser, and M. Middleton. pXRF measurements on soil samples for the exploration of an antimony deposit: example from the Vendean antimony district (France). Submitted.

CHAPTER 2

Imputation of values above an upper detection limit in compositional data

Geochemical data frequently contain censored values. An imputation method for right-censored compositional data is proposed, based on the Tobit model, in order to get a complete and reliable data set. An algorithm is developed and implemented using regressions in an iterative scheme, where the imputed values are updated step-by-step. Optionally, classical least-squares or robust regressions can be carried out, with or without variable selection. The performance of the algorithm is evaluated using two real geochemical data sets, blackconsidering various different scenarios. Compared to commonly used substitution methods, the proposed method leads to an improved data quality. The procedure is available in the R package robCompositions.

2.1 Introduction

Detection limits typically occur in measurement processes where the measurement instrument is reliable up to a certain minimal or maximal value. In the first case, this refers to the lower detection limit (LDL), while the upper boundary is called the upper detection limit (UDL). Observations with values below the LDL or above the UDL are called (left/right) censored values, and they lead to a truncated distribution of the considered variable (Helsel, 2012; Millard et al., 2012). Naturally, this creates

difficulties for the statistical analysis, for estimating parameters, and in particular for the multivariate analysis if several variables have been measured simultaneously.

In fact, the problem even starts before doing statistics, since laboratories are reporting values below the LDL or above the UDL by non-numeric entries, for example as "< 0.01" for a value which is lower than the LDL value 0.01 for this variable. Reading a data table with such entries in a statistical software package will cause inappropriate data types. For instance, in `R` (R Development Core Team, 2018) one would obtain a factor variable, and each distinct value would create an own factor level, which makes any deeper statistical analysis impossible. As a simple way out, one can substitute these entries by meaningful numbers. For values below the LDL it is common to replace these values by half of the corresponding detection limit. Martín-Fernández et al. (2003) have shown that a better choice is a replacement by 65% of the detection limit, since this minimizes the distortion of the covariance structure (Martín-Fernández et al., 2011). Also for substituting values above the UDL there are different proposals, such as using a factor of 1.7 times the UDL value, or alternatively a factor of 4/3 (Sanford et al., 1993). In the following, such a replacement will be denoted as "simple method", and here the factor 1.2 is considered for comparison.

All these proposals for replacements do not make use of possibly available multivariate data information. This is typically the case for geochemical data, where a plant or soil sample is analyzed for the concentration of several chemical elements. It would then be natural to use non-censored information of other variables to estimate the censored values, if there exists a statistical dependency of the censored variables with the non-censored.

Here the focus is on a specific type of multivariate data, called *compositional data* (CoDa), which consist of strictly positive values, and where the interest is in the relative information between the variables rather than directly in the reported data values (Pawlowsky-Glahn et al., 2015). Data with element concentrations can be considered as CoDa, since already the data unit (mg/kg, pbb, etc.) refers to relative contributions of a certain whole (Egozcue, 2009). Aitchison (1986) introduced the log-ratio approach for the statistical analysis of CoDa. Based on these ideas, an elegant mathematical way for the analysis of CoDa has been established in the last decades (see Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018, for recent books on this subject). As examples of `R` packages dealing with CoDa can be mentioned these `compositions` (van den Boogaart and Tolosana-Delgado, 2008) and `robCompositions` (Templ et al., 2011).

In the literature around CoDa, values below the LDL are called *rounded zeros*. Several proposals for replacing rounded zeros by meaningful numbers are available (e.g. Palarea-Albaladejo and Martín-Fernández, 2015, and references therein). Palarea-Albaladejo

and Martín-Fernández (2008) developed a model-based approach for this purpose, where Tobit regression is used within an iterative algorithm to estimate the rounded zeros. Martín-Fernández et al. (2012) extended this approach, where traditional least-squares regression, but also robust regression is employed and compared based on simulated and real data. In this paper we will modify the method of Martín-Fernández et al. (2012) to estimate values above the UDL. To the best of our knowledge, this is the first proposal for a model-based imputation of values above the UDL for CoDa.

This paper is organized as follows: Section 2.2 provides more detailed information on the concepts of the log-ratio approach for CoDa. Section 2.3 introduces the methodology for model-based imputation of values exceeding the UDL. An algorithm for the imputation procedure is proposed in Section 2.4. A real data example from geochemistry with UDL problems is shown in Section 2.5. Section 2.6 presents numerical experiments based on real geochemical data, and Section 2.7 summarizes and concludes.

2.2 The log-ratio approach for CoDa

Consider a multivariate observation $\mathbf{x} = (x_1, \ldots, x_D)^t$. In the context of CoDa, $\mathbf{x}$ is called a *composition* with D parts, which are all strictly positive. For a composition, the relevant information is included in the ratios between the parts. Historically, a composition is represented in the D-part simplex $\mathcal{S}^D$, which is defined as

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)^t \in \mathbb{R}^D | x_i > 0, \sum_{j=1}^{D} x_j = \kappa \right\}. \tag{2.1}$$

The sum κ can be any positive real value because a composition is an equivalence class (Barceló-Vidal and Martín-Fernández, 2016). For the log-ratio approach, which is based on pairwise log-ratios $\ln(x_k/x_l)$, for $k, l \in \{1, \ldots, D\}$ and $k \neq l$, this sum is irrelevant, because if one would multiply the composition by any positive number c, the resulting log-ratio would be unchanged. Thus, one could also express any composition with sum 1, without any loss of information.

It is possible to define a Euclidean linear vector space structure of the simplex, with all the basic operations that are necessary (Pawlowsky-Glahn and Egozcue, 2001; Egozcue et al., 2003). With this geometrical structure, referred to as the *Aitchison geometry*, the simplex $\mathcal{S}^D$ has dimension $D - 1$.

Using this geometry, a composition is represented by $D - 1$ orthonormal coordinates in the real Euclidean geometry. This may complicate the interpretation of results later on in terms of the original compositions, but it will simplify the use with standard statistical

methods which are based on this Euclidean geometry. One possible approach which became very popular in the last years are so-called isometric log-ratio (ilr) coordinates (Egozcue et al., 2003), and often used for geochemical data because of the compositional nature of the data (Talebi et al., 2019). One particular definition of ilr coordinates are so-called pivot coordinates (Fišerová and Hron, 2011):

$$z_j^{(l)} = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j^{(l)}}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k^{(l)}}}, j = 1, \ldots, D-1 \tag{2.2}$$

Here, $\mathbf{x}^{(l)} = (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})^t = (x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)^t$ stands for the reordered composition, where the l-th part is moved to the first position, for $l \in \{1, \ldots, D\}$. From this construction one can see that the first part $x_1^{(l)}$ (or x_l) appears in the numerator of the first coordinate $z_1^{(l)}$, but in no other coordinate. Thus, this first coordinate expresses all relative information about part x_l in the composition, and it can be interpreted in terms of a dominance of this part with respect to the other parts, which are aggregated by the geometric mean.

Pivot coordinates represent a one-to-one mapping, and thus it is possible to come back to the original compositional parts by computing

$$\begin{aligned}
\tilde{x}_1^{(l)} &= \tilde{\kappa} \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1^{(l)}\right), \\
\tilde{x}_j^{(l)} &= \tilde{\kappa} \exp\left(-\sum_{k=1}^{j-1} \frac{1}{\sqrt{(D-k+1)(D-k)}} z_k^{(l)} + \frac{\sqrt{D-j}}{\sqrt{D-j+1}} z_j^{(l)}\right), \quad j = 2, \ldots, D-1, \\
\tilde{x}_D^{(l)} &= \tilde{\kappa} \exp\left(-\sum_{k=1}^{D-1} \frac{1}{\sqrt{(D-k+1)(D-k)}} z_k^{(l)}\right),
\end{aligned} \tag{2.3}$$

which are normalized by a constant $\tilde{\kappa}$ to sum up to one, i.e. $\sum_{j=1}^{D} \tilde{x}_j^{(l)} = \sum_{j=1}^{D} \tilde{x}_j = 1$, where the $\tilde{x}_j$ refer to the back-permuted version of the $\tilde{x}_j^{(l)}$. Thus, in order to obtain the original sum, the re-expressed compositions are

$$x_j^{(re)} = \tilde{x}_j \cdot \kappa \quad \text{with } \kappa = \sum_{j=1}^{D} x_j \quad \text{for } j = 1, \ldots, D. \tag{2.4}$$

Pivot coordinates are appropriate for a model-based approach to estimate censored values, because the censoring variable can be isolated into one coordinate, for which a model can be established based on the remaining parts in the composition. Note that ilr coordinates are preferable over other representations, in particular if robust methods are applied (Martín-Fernández et al., 2012).

2.3 Method

The Tobit model (Tobin, 1958) described below has been used in Palarea-Albaladejo and Martín-Fernández (2008) and Martín-Fernández et al. (2012) in the context of estimating values below a LDL (left-censored data). In the context of right-censored data, where values above an UDL are not available, the Tobit model introduced in Tobin (1958) describes the relationship between a non-negative censored variable y and p non-censored explanatory variables $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^t$. An unobservable latent variable y^* is assumed, and the observable variable y gives

$$y = \begin{cases} y^* & \text{if } y^* \leq \psi \\ \tau & \text{if } y^* > \psi \end{cases}, \tag{2.5}$$

where ψ represents the value of the UDL, and τ is a so-called truncation point, which is a particular value of the UDL. Denoting the observations of the above variables as y_i, y_i^*, and $\boldsymbol{\xi}_i$, respectively, for $i = 1, \ldots, n$, a regression model $y_i^* = \boldsymbol{\xi}_i^t\boldsymbol{\beta} + \varepsilon_i$ is used for all non-censored observations, with the regression coefficients $\boldsymbol{\beta}$, and the error terms ε_i, assumed to be independent $N(0, \sigma^2)$ distributed. Using likelihood estimation as in the truncated regression model, one obtains (Schnedler, 2005)

$$E[y_i \mid y_i > \psi] = \boldsymbol{\xi}_i^t\boldsymbol{\beta} + \sigma \left[\frac{\phi(\frac{\psi - \boldsymbol{\xi}_i^t\boldsymbol{\beta}}{\sigma})}{1 - \Phi(\frac{\psi - \boldsymbol{\xi}_i^t\boldsymbol{\beta}}{\sigma})} \right], \tag{2.6}$$

where ϕ and Φ are the density and distribution function, respectively, of the standard normal distribution.

For CoDa, values above an UDL can appear in one or more parts of the composition. For now we assume that they only appear in the l-th part x_l. In that case, the pivot coordinates defined in Equation (2.2) are appropriate, because the first coordinate isolates the censored part from the remaining composition. In other words, the first coordinate will be used as the response, and the remaining coordinates as explanatory variables in the previously discussed model.

Assume a CoDa matrix $\mathbf{X}$ with n observations arranged in the rows, and D columns for the compositional parts. Denote $x_{ij}^{(l)}$ as the element (i, j) of the compositional data matrix with the l-th part arranged as the first column. After applying Equation (2.2) one obtains the matrix of coordinates $\mathbf{Z}^{(l)}$, with n rows and $D - 1$ columns. Denote the elements of this matrix by $z_{ij}^{(l)}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, D - 1$, and $\mathbf{z}_{i,-1}^{(l)} = (z_{i2}^{(l)}, \ldots, z_{i,D-1}^{(l)})^t$ the i-th observation with the first entry excluded.

Further, denote the value of the UDL for the l-th compositional part as $\tau^{(l)}$, the set $U^{(l)}$ as the index set containing the indexes of observations of part x_l which exceed $\tau^{(l)}$,

and $O^{(l)} = \{1, \ldots, n\} \setminus U^{(l)}$ the set with the remaining indexes. The UDL needs to be expressed in the same coordinate system using Equation (2.2), and one obtains

$$\psi_i^{(l)} = \sqrt{\frac{D-1}{D}} \ln \frac{\tau^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_{ij}^{(l)}}}. \tag{2.7}$$

Note that the re-expressed UDL is an individual value for each observation, thus depending on the index i.

The regression problem for the non-censored observations is

$$z_{i1}^{(l)} = \boldsymbol{z}_{i,-1}^{(l)t} \cdot \boldsymbol{\beta}^{(l)} + \varepsilon_i \quad \text{for } i \in O^{(l)}. \tag{2.8}$$

After employing an appropriate regression method, one obtains the estimated regression coefficients $\hat{\boldsymbol{\beta}}^{(l)}$, as well as the estimated standard deviation of the residuals $\hat{\sigma}^{(l)}$. Following Equation (2.6), these estimates are used to predict the values above the UDL,

$$\hat{z}_{i1}^{(l)} = \boldsymbol{z}_{i,-1}^{(l)t} \cdot \hat{\boldsymbol{\beta}}^{(l)} + \hat{\sigma}^{(l)} \left[\frac{\phi\left(\frac{\psi_i^{(l)} - \boldsymbol{z}_{i,-1}^{(l)t} \cdot \hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}} \right)}{1 - \Phi\left(\frac{\psi_i^{(l)} - \boldsymbol{z}_{i,-1}^{(l)t} \cdot \hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}} \right)} \right] \quad \text{for } i \in U^{(l)}. \tag{2.9}$$

Finally, the estimated values from Equation (2.9) need to be represented in terms of the original composition, which is done by using Equation (2.3) for all observations with index in $U^{(l)}$. An important issue is the choice of the appropriate factor κ according to Equation (2.4), which in fact is now an individual value κ_i, with $i \in U^{(l)}$. Since the total sum for these observations is not available because of the values exceeding the UDL, one can first apply Equation (2.3) for all $i \in U^{(l)}$ to obtain values $\tilde{x}_{ij}^{(l)}$, for $j = 1, \ldots, D$. Then the factors are obtained by comparing the total sum of the original and the back-transformed compositions, the l-th part excluded,

$$\kappa_i^{(l)} = \frac{\sum_{j=2}^{D} x_{ij}^{(l)}}{\sum_{j=2}^{D} \tilde{x}_{ij}^{(l)}}. \tag{2.10}$$

Finally, the re-expressed estimated right-censored values of the l-th part are obtained by

$$x_{il}^{(re)} = \kappa_i^{(l)} \tilde{x}_{il} \quad \text{for } i \in U^{(l)}, \tag{2.11}$$

see Equation (2.4).

2.4 Algorithm

Section 2.3 has outlined a methodology to estimate right-censored values, i.e. values that exceed an UDL in one compositional part. In practice it might happen that right-censoring appears in several compositional parts. For example, when analyzing concentrations of chemical elements in plants using ashed samples, right-censoring typically happens in elements such as K, Mn, P, or Rb (Beinrohr et al., 1991). Therefore, the following algorithm based on iterative updating is proposed:

Step 1 Assume that right-censoring is present in r compositional parts, with $1 \leq r \leq D$. To simplify notation, assume that the compositional parts are arranged in a way that the part with the highest amount of right-censored data is at the first position, the part with the second highest amount at the second position, etc. Parts without right-censored values are arranged at the last positions in their original order, although this is not relevant for the method. Initialize all right-censored values by 1.2 times the UDL of the corresponding part. The following steps intend to improve the initially imputed values.

Start with part $l = 1$.

Step 2 Consider the corresponding value $\tau^{(l)}$ of the UDL. Represent the composition with the l-th part reordered to the first position in coordinates using Equation (2.2), and use Equation (2.7) for a coordinate representation of $\tau^{(l)}$.

Step 3 Estimate the regression coefficients in the model Equation (2.8) and the censored observations in coordinates by Equation (2.9). Express these estimated values in the simplex using Equation (2.3) to obtain the values $\tilde{x}_{ij}^{(l)}$, for $i \in U^{(l)}$ and $j = 1, \ldots, D$.

Step 4 Compute the constants $\kappa_i^{(l)}$ not as in Equation (2.10), but as

$$\kappa_i^{(l)} = \frac{\sum_{j \in A_i} x_{ij}}{\sum_{j \in A_i} \tilde{x}_{ij}} \quad \text{for } i \in U^{(l)}, \tag{2.12}$$

where A_i denotes the indexes of parts without right-censored values for the i-th observation. Use Equation (2.11) to re-express the right-censored values in this part with the appropriate total sum.

Step 5 Apply Steps 2-4 in turn for $l = 2, \ldots, r$. Denote the resulting values of the re-expressed composition by $x_{ij}^{[1]}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, D$.

Step 6 Iterate Steps 2-5 until the relative squared error

$$\sum_{i=1}^{n} \sum_{j=1}^{D} \left(\frac{x_{ij}^{[m]} - x_{ij}^{[m+1]}}{x_{ij}^{[m+1]}} \right)^2 \tag{2.13}$$

is smaller than a given threshold, or until a maximum number of iterations has been reached. Here, $x_{ij}^{[m]}$ are the values of the re-expressed composition after the m-th iteration. Note that this criterion would not appropriately reflect a relative distance in CoDa sense, but it turned out to be useful as a convergence criterion.

Variable selection:

Templ et al. (2016) have introduced an imputation method for rounded zeros for high-dimensional data, with an option for variable selection. We are not particularly dealing with high-dimensional data, but for data with smaller numbers of samples this might still give an advantage. Consider the regression model in Equation (2.8) where for the prediction of right-censored values in the l-th part all remaining parts in the composition are used. One could select fewer predictors as well in order to get a more stable regression model. This can be done by computing the variation matrix elements between x_l and all remaining parts, $\text{var}(\ln \frac{x_l}{x_j})$, for $j \in \{1, \ldots, D\}$, $j \neq l$ (see e.g. Pawlowsky-Glahn et al., 2015). Small values indicate stronger association with x_l, larger values point at weaker association. The predictor variables are sorted according to these values in ascending order, and regression models according to Equation (2.8) with only the first k predictors are considered, for $k \in \{2, \ldots, D-1\}$. Using a cross-validation scheme, that model will be selected which leads to the smallest cross-validated prediction error.

This algorithm has been implemented in the software environment `R` and is available as the function `imputeUDLs()` in the package **robCompositions** (Templ et al., 2011).

2.5 Example

Within the project Mineral potential of northern Finland, carried out by the Geological Survey of Finland (GTK), mountain birch twig samples on an iron oxide-copper-gold mineral deposit Lätäseno norhernmost Finnish Lapland, were collected, ashed in 475°C and elemental concentrations were measured with inductively coupled plasma mass spectrometry (ICP-MS) and optical emission spectrometry (ICP-OES) at the Acme Analytical Laboratories (Vancouver, Canada) from a dissolution of 0.25 g aliquot of ash digested in 1:1:1 HCl:HNO3:H_2. This resulted in data of 99 observations and 64 elements, which after quality check was reduced to 27 variables (element concentrations). Two of the variables, phosphorus (P) and zinc (Zn) have values above the UDL, which typically happens when the samples are ashed. For P 20% of the values were above the UDL, and for Zn 26%. Usually, the laboratory would only report the UDL. In this case, however, the laboratory also unofficially provided the ICP calibrated values >UDL.

Usually, the laboratory would not provide access to these data because the values could be unreliable. In the following we will estimate the values which exceed the UDL, but use the "unofficially" reported values for comparison, with a caution that the "true" values might not be very accurate.

We will compare the following procedures:

- *Simple method*: Values above the UDL are substituted by 1.2 times the UDL.
- *Classical method*: The regression coefficients in the model in Equation (2.8) are estimated by the classical least-squares estimator.
- *Robust method*: The regression coefficients in the model in Equation (2.8) are estimated by the robust MM-estimator (Yohai, 1987), as it is implemented in the function `lmrob` of the R package `robustbase` (Rousseeuw et al., 2009). The MM-estimator is highly robust against outliers and highly efficient at the same time.

Figure 2.1 shows the results of the three methods, and they are compared to the values reported by the laboratory. The dashed lines in the plots represent the value of the UDL. The simple method seems inappropriate, especially if one would have to use these estimated values in a subsequent (multivariate) statistical analysis. There is not too much difference between the results of the classical and the robust method, perhaps because there were no severe outliers present, and both correspond to the trend of the reported values (Filzmoser et al., 2009, 2012). For the classical as well as for the robust method, the algorithm converged after 3 iterations, in a fraction of a second.

Any of the above methods should be taken with care if the values above the UDL are originating from a different process. In that case, a structural break in the element distribution will be visible, typically in a QQ-plot. Figure 2.2 shows the QQ-plots for the reported values of P (left) and Zn (right), with a comparison against the quantiles of a standard normal distribution (horizontal axes). The dashed lines indicate the UDL. There is no break visible, and thus a regression-based replacement seems useful.

2.6 Numerical experiments

We will demonstrate the performance of the imputation procedure by using simulations based on a real data set. This data set originates from a project of the Geological Survey of Norway (NGU) in a 100 km transect in Gjøvik, Norway (Reimann et al., 2018). In total, 41 sample sites have been investigated in an area where four zones of

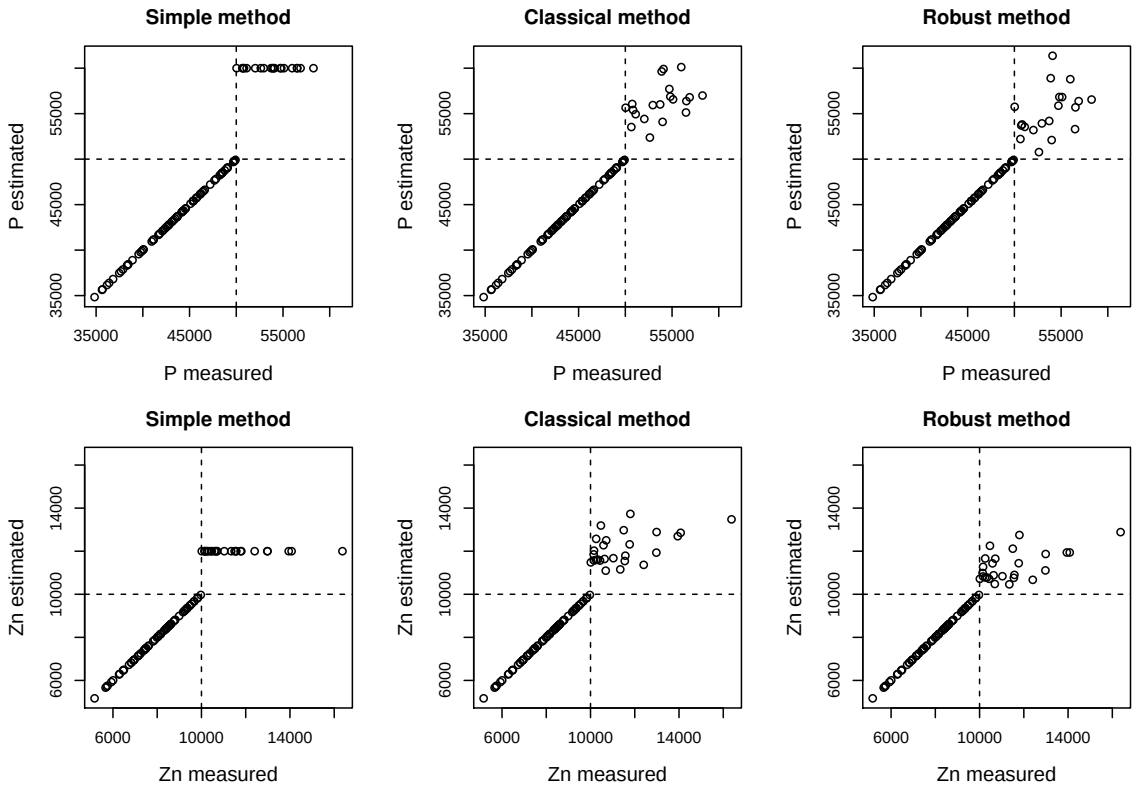


Figure 2.1: Plots of measured versus estimated values for P and Zn for the Lätäseno data set.

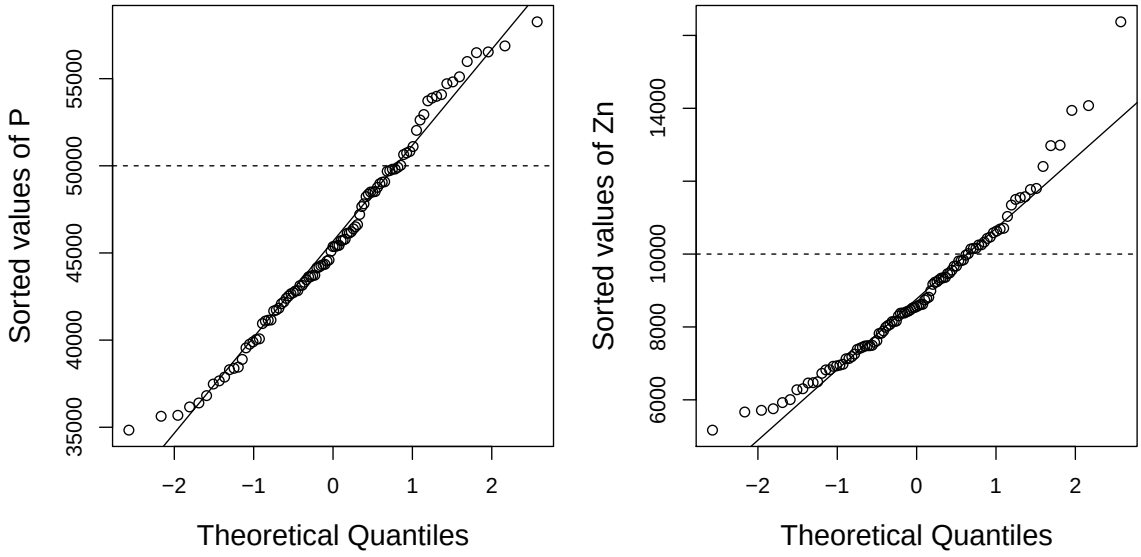


Figure 2.2: QQ-plots of the reported values for P and Zn for the Lätäseno data set.

mineralization are crossing. At each site, 15 different sample materials (birch, spruce, cowberry, mushroom, O- and C-horizon for soil, etc.) have been collected and analyzed for the concentration of 53 chemical elements. The data are made available in the R package `robCompositions` as data set `gjovik`. In this data set, no right-censored values occurred, and therefore we will simulate such scenarios and compare with the available information. In order to avoid any data problems, we use 13 selected elements with sufficiently good data quality in all the sampled media (reasonably small proportions of left-censored or rounded values). From a statistical point of view it is now of interest if the quality of the imputation differs among the sample media, if it depends on the size of the data set, etc.

For the comparison of the simple, the classical and the robust imputation method, we will consider two evaluation measures that focus on two aspects, on the covariance structure and on the distances between the observations (see also Martín-Fernández et al., 2012; Buccianti, 2013):

- *Relative difference in covariance matrix* (RDCM): Denote $\mathbf{X}^{(re)}$ the imputed (re-expressed) data set, as returned from the algorithm described in Section 2.4. Then the original data set $\mathbf{X}$ and the re-expressed data are represented in coordinates, say $\mathbf{Z}$ and $\mathbf{Z}^{(re)}$, respectively, using the same ilr representation. The sample covariance matrix $\mathbf{S}$ of $\mathbf{Z}$ with elements s_{jk}, and the sample covariance matrix $\mathbf{S}^{(re)}$ of $\mathbf{Z}^{(re)}$ with elements $s_{jk}^{(re)}$ are computed, for $j,k \in \{1,\dots,D-1\}$. The measure RDCM is defined as

$$\frac{\|\mathbf{S}-\mathbf{S}^{(re)}\|_F}{\|\mathbf{S}\|_F} = \frac{\sqrt{\sum_{j,k=1}^{D-1}\left(s_{jk}-s_{jk}^{(re)}\right)^2}}{\sqrt{\sum_{j,k=1}^{D-1} s_{jk}^2}}, \tag{2.14}$$

  where $\|\cdot\|_F$ denotes the Frobenius norm.

- *Compositional error deviation* (CED): Denote the observations of $\mathbf{Z}$ by $\mathbf{z}_i$, and those of $\mathbf{Z}^{(re)}$ by $\mathbf{z}_i^{(re)}$, for $i=1,\dots,n$. Further, call C the index set referring to observations where at least one part has been imputed, and the number of those observations by n_C. The measure CED is defined as

$$\frac{\frac{1}{n_C}\sum_{i\in C} d(\mathbf{z}_i,\mathbf{z}_i^{(re)})}{\max\limits_{\{\mathbf{z}_r,\mathbf{z}_s\in\mathbf{Z}\}}\{d(\mathbf{z}_r,\mathbf{z}_s)\}}, \tag{2.15}$$

  where $d(\cdot,\cdot)$ stands for the Euclidean distance. The denominator is the maximum distance of any two observations in the original data set.

2.6.1 Example with R code

In order to be more specific, the 41 observations from the plant species "SPR" (spruce) from the Gjøvik data set are used for the selected 13 variables. Initially, a detection limit problem is introduced in the variable "Fe": the value of the UDL is set to the quantile 0.8 of this variable, and thus the upper 20% of the values of Fe are right-censored. All remaining variables will not have any values above an UDL, and thus the UDL is set to the maximum for these variables. The R code looks as follows.

```
R> library(robCompositions)                # load package
R> data(gjovik)                            # load data set
R> sv <- c("Al","Ba","Cd","Ce","Co","Cs","Cu","Fe","Mn","Mo",
         "Na","Ni","Zn")
R> dat <- gjovik[gjovik$MAT=="SPR",sv]     # select species SPR
                                           # and variables
R> UDL <- apply(dat,2,max)                 # UDL value for each
                                           # variable
R> names(UDL) <- names(dat)
R> UDL["Fe"] <- quantile(dat[,"Fe"], probs = 0.8)
R> whichudl <- dat[,"Fe"] > UDL["Fe"]    # which cells are > UDL
```

As mentioned above, the simple imputation method is replacing UDL values in Fe by 1.2 times the corresponding UDL:

```
R> imp.simple <- dat
R> imp.simple[whichudl,"Fe"] <- UDL["Fe"]*1.2
```

For the imputation using the function `imputeUDLs`, values above the UDL need to be set to infinity (`Inf`):

```
R> imp.lm <- dat
R> imp.lm[whichudl,"Fe"] <- Inf
R> res.lm <- imputeUDLs(imp.lm,dl=UDL,method="lm",
             variation=TRUE)
R> imp.lm <- res.lm$x
```

In the above code, the classical method using least-squares regression is applied, and variable selection is carried out with `variation=TRUE`. In the resulting object

`res.lm$x`, the whole data matrix is saved, but only the cells which have been set to `Inf` are modified.

Imputation with the robust regression method, using again variable selection, can be done as follows:

```
R> imp.lmrob <- dat
R> imp.lmrob[whichudl,"Fe"] <- Inf
R> res.lmrob <- imputeUDLs(imp.lmrob,dl=UDL,method="lmrob",
                 variation=TRUE)
R> imp.lmrob <- res.lmrob$x
```

Figure 2.3 shows a visual comparison of the measured values of Fe versus the estimated (imputed) values for the three different methods, together with the "artificial" UDL of 100 mg/kg as dashed lines. The simple imputation seems to destroy the data structure much more than the regression-based imputations. In this case, there is almost no difference visible between the classical and the robust regression method.

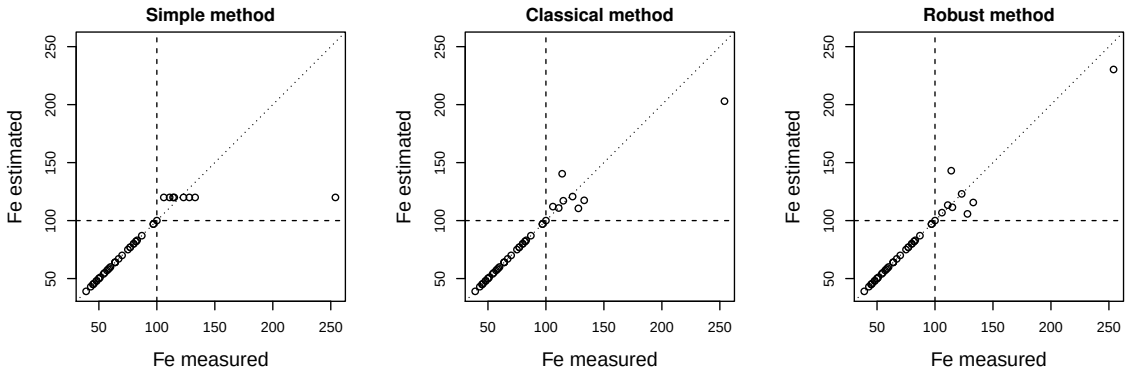


Figure 2.3: Plot of measured versus estimated values for Fe when the UDL is equal to the quantile 0.8 (subset spruce from the Gjøvik data).

For each imputation method, both error measurements can be computed. For instance, for the classical least-squared imputation, the measures RDCM and CED are computed as follows:

```
R> rdcm(dat,imp.lm)
R> ced(dat[whichudl,], imp.lm[whichudl,], sum(whichudl)) /
       max(dist(pivotCoord(dat)))
```

The results for all three methods are shown in Table 2.1. While the regression-based methods are indeed quite comparable, the simple method performs worse, and the imputed data show a larger error in the covariance structure and for the distances.

Table 2.1: Resulting error measurements for the three imputation methods after imputing the upper 20% of the values of Fe.

Error measurement	Simple method	Classical method	Robust method
RDCM	0.031	0.0158	0.0060
CED	0.024	0.0166	0.0168

Figure 2.4 shows for the same data set another imputation example where the upper 20% of the values of the variables Ce and Cs have been imputed by the simple and the classical regression method. The resulting data structure is represented by log-ratio biplots (Aitchison and Greenacre, 2002) of the first two principal components for the original data (left), for the simple method (middle) and for the classical method (right). The explained variance in these presentations is around 60%, and for better readability the original variable names are plotted. Although there are not big changes, one can see that the data structure is somehow distorted with the simple method, compared to the classical one.

2.6.2 Convergence of the algorithm

The algorithm as presented in Section 2.4 stops after the relative squared error in Equation (2.13) is smaller than a given threshold, or after a maximum number of iterations has been reached. The following experiment should provide more detailed insights into the convergence behavior, and for this purpose the threshold is set to 0.01, and the maximum number of iterations to 100. As in the previous section, the spruce data from the Gjøvik data set is considered with the 13 selected variables. To mimic a right-censored situation, variables are randomly selected, and the UDL of these variables is set to the quantile 0.8. This is done simultaneously for an increasing number of variables, starting from 1 until 10. Higher numbers can lead to instabilities of the algorithm in terms of much higher numbers of iterations – depending on the position of the values above UDL. Figure 2.5 shows the resulting numbers of iterations of the algorithm for 100 replications of this simulation, using the classical and the robust method, respectively, for imputation.

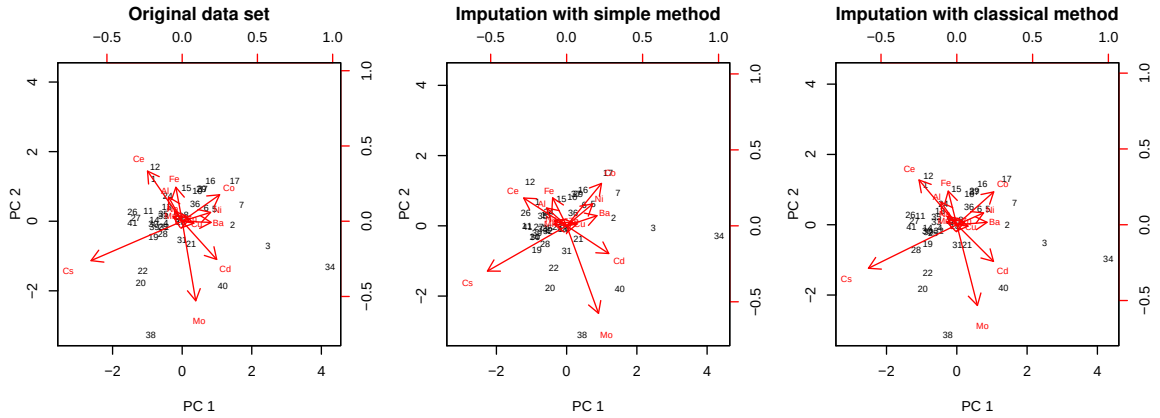


Figure 2.4: Log-ratio biplots for the original data, and for data based on imputation with the simple and the classical method in the variables Ce and Cs, where the UDL is equal to the quantile 0.8 (subset spruce from the Gjøvik data).

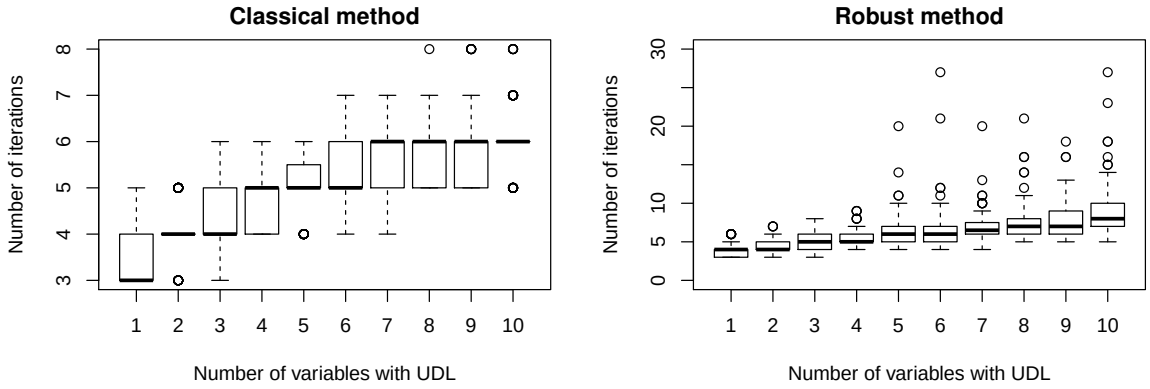


Figure 2.5: Classical and robust imputation for the spruce data of Gjøvik with 13 variables. The UDL is set to the quantile 0.8 for each of the 1 to 10 randomly selected variables. The plots show the numbers of iterations of the algorithm for the imputation, for 100 simulation replications.

Figure 2.5 reveals a clear difference between the classical and the robust method: in the latter case, the number of iterations is clearly higher. This can be explained by the fact that imputed values in one iteration could become outliers in the next one, or the other way around. Thus, the observations which are downweighted (outliers) would also vary, leading to a certain instability in the algorithm. Note that the vertical axis of the plot for the robust method was cut at 30. In rare cases, the number of iterations reaches almost the maximum 100. It can also be seen from the plots, that the number of iterations increases if imputation has to be carried out in more and more variables. However, this increase is still in very limited bounds, in particular for the classical method.

2.6.3 Effect of sample size

The effect of varying sample size is evaluated in Figure 2.6; on the horizontal axes the considered specific sample sizes are shown. Accordingly, data sets with these sample sizes are randomly (without replacement) selected from the pool of all observations in the Gjøvik data set, irrespective of the material (plant species). Then, a UDL is set according to the quantile 0.8 in one randomly selected variable, and classical and linear regression (without variable selection) is applied. The boxplots in Figure 2.6 represent the resulting error measures for 100 simulations. It can be seen that with increasing sample size, the errors decrease but also stabilize (smaller variability). The error reduction e.g. from 30 to 80 observations is remarkable. Despite both methods may be affected by mixing observations from different sub-populations together, the robust method clearly shows poorer performance compared to the classical one. Robust regression will fit only the data majority corresponding to the common data structure, and treat the remainder as outliers, which are then not appropriately modeled.

2.6.4 Increasing the number of variables for imputation

In the following simulation experiment, again the subset for plant spruce of the Gjøvik data set is used with the same 13 selected variables as before. As in the previous section, the UDL is set to the quantile 0.8 of a specific variable, and the number of such variables is increased from 1 to 8. Classical and robust regression is compared, both with (`TRUE`) and without (`FALSE`) variables selection. The evaluation of the results is based on the RDCM and CED measures, and their 10% trimmed means over 100 simulations are presented in Figure 2.7; trimming is used to suppress the effect of outliers in the evaluation. Clearly, the more variables have to be imputed the worse the error measures are, since both the covariance structure and the distances are more and more distorted. However, also the

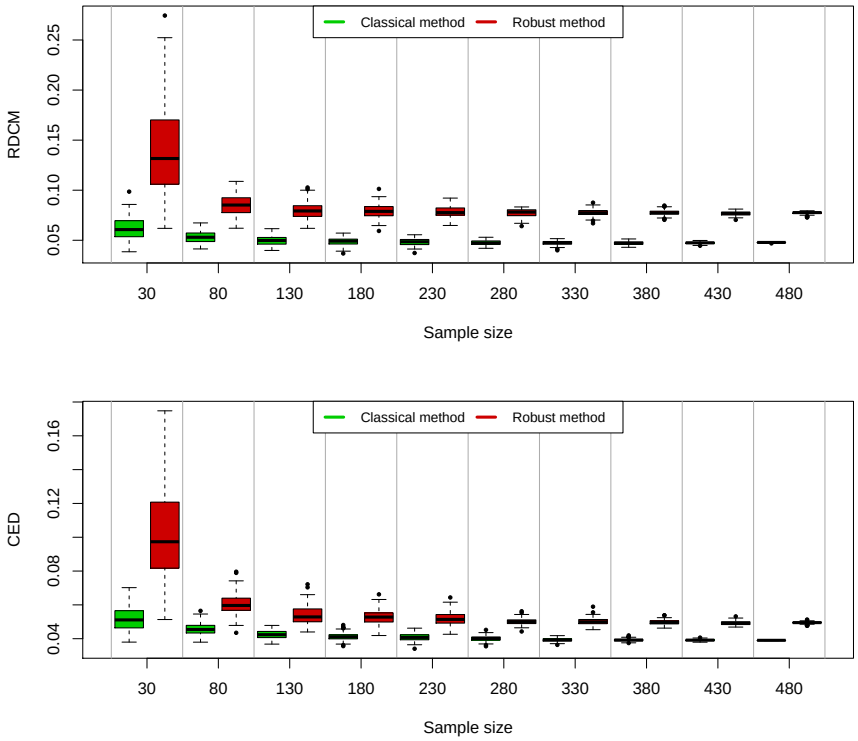


Figure 2.6: Error measurements depending on sample size: Data sets with the indicated sample size are randomly drawn from the Gjøvik data set, and imputation is done in one randomly selected variable, where the UDL is set to the quantile 0.8.

difference in the performance for the methods gets more pronounced with increasing numbers of variables for imputation. If imputation has to be done in more than half of the 13 variables, the methods using variable selection show some instability. The classical method without variable selection shows the overall best performance. Depending on the measure and the situation, variable selection sometimes leads to an advantage. This is clearly seen for the robust method, for up to 6 imputed variables.

2.6.5 Increasing the proportion of values above the UDL

For the next experiment, again the spruce data subset from the Gjøvik data set with the 13 selected variables is considered. UDL values are generated in one randomly selected variable, and the quantile (UDL value) is modified between 0.5 and 0.95. The 10% trimmed average of the RDCM and CED measure over 100 simulations are visualized in Figure 2.8 for the different methods, also considering variable selection (`TRUE`/`FALSE`).

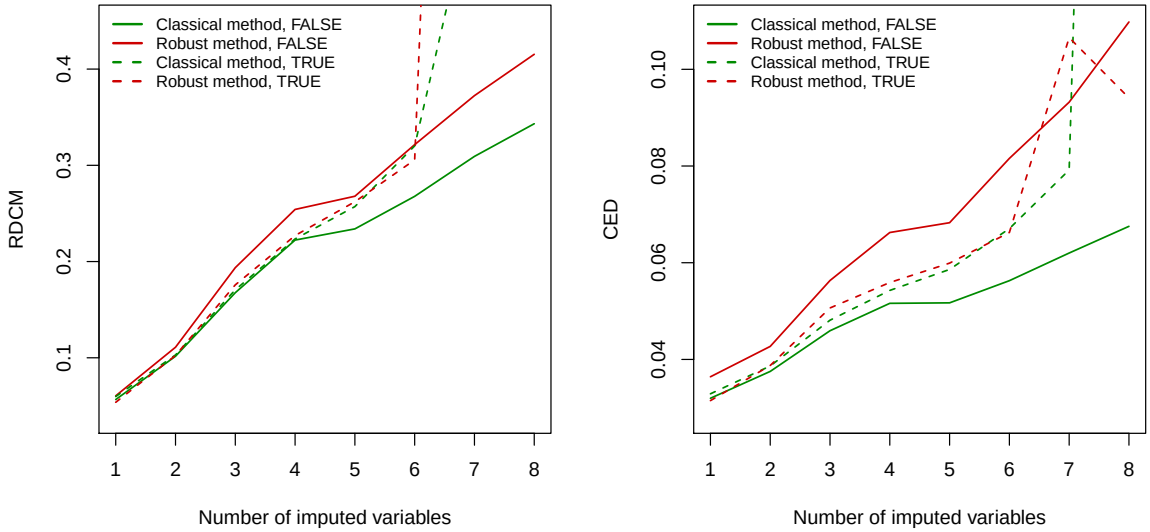


Figure 2.7: Comparison of classical and robust regression imputation with (TRUE) and without (FALSE) variable selection for the spruce data subset. The number of variables to be imputed is increased, and imputation needs to be done for the upper 20% of the values.

In several situations, variable selection leads to better results, and the robust method is in most cases superior to the classical one. Especially for data with low sample size (we have 41 spruce samples), the prediction from regression with a larger number of variables (here 13) leads to an overfit and can become quite unstable, see also Figure 2.7. Variable selection avoids the effect of overfitting. The simple method leads to a rather poor performance. It is also interesting to note that for an increasing censoring level especially the RDCM measure suffers (the covariance structure is destroyed), but the CED is still quite stable (distances are still reasonable).

We carried out this experiment also for the complete Gjøvik data set with 604 observations (instead of 41 observations for the spruce data), where variable selection did no longer reveal any advantage. This is also to be expected, since with high numbers of observations compared to the number of parts, the regression models become more stable within the iterative scheme.

2.6.6 Comparison for different data subsets

In the following experiment, the imputation is done for each sub data set corresponding to the different sample materials. Their data structure is potentially very different, which

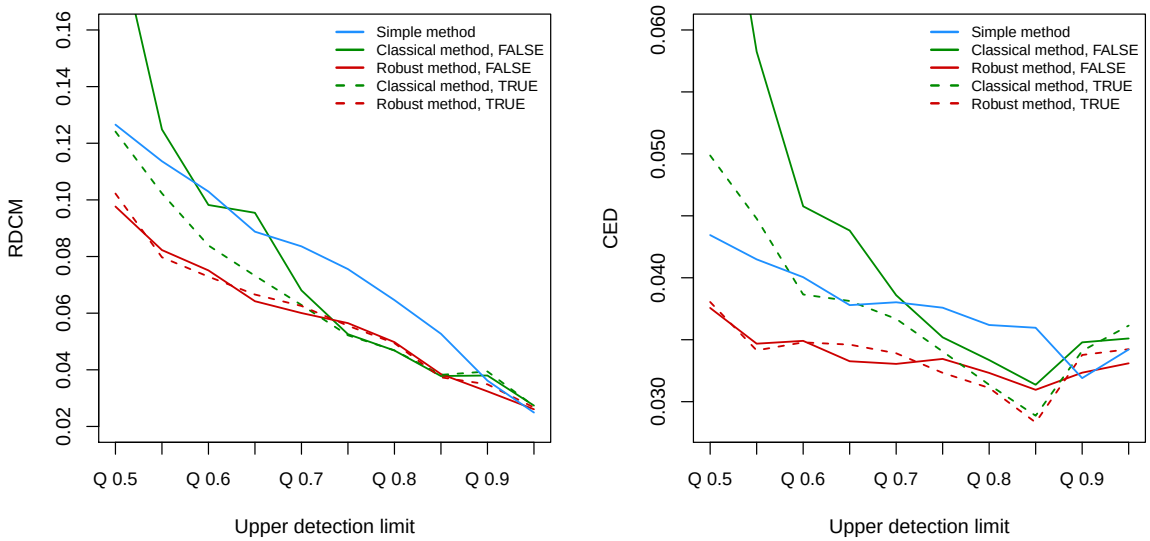


Figure 2.8: Comparison of the simple method, and of classical and robust regression imputation with (TRUE) and without (FALSE) variable selection for the spruce data from Gjøvik. Imputation is done in only one randomly selected variable, by modifying the UDL value from the quantile 0.5 to 0.95.

could also lead to a difference in the performance of the imputation. Out of the considered 13 variables, one variable is randomly picked, and the UDL value is modified between the quantile 0.5 and 0.95. Then the average of the resulting error measures is computed for the simple and the regression-based methods (without variable selection). Figure 2.9 shows the results for 100 simulations. The robust method shows some instability, visible as very right-skewed distributions. Here, the sample size of each plant material was around 41 (sometimes there were observations with missings which had to be eliminated). According to Figure 2.6, this small sample size indeed leads to higher instability of the regression-based methods, and this becomes even more severe if the proportion of values to be imputed increases, see also Figure 2.8. In this case, the overall recommended procedure could be the classical method. It is remarkable that the only two soil sample materials, C- and O-horizon, are ranked among the worst error results.

2.6.7 Comparison in terms of different variables

In this last experiment, all observations from the Gjøvik data set are used, and 30 selected variables are considered for imputation. Those 30 variables still have reasonable data quality, because there are reasonably small proportions of left-censored or rounded values.

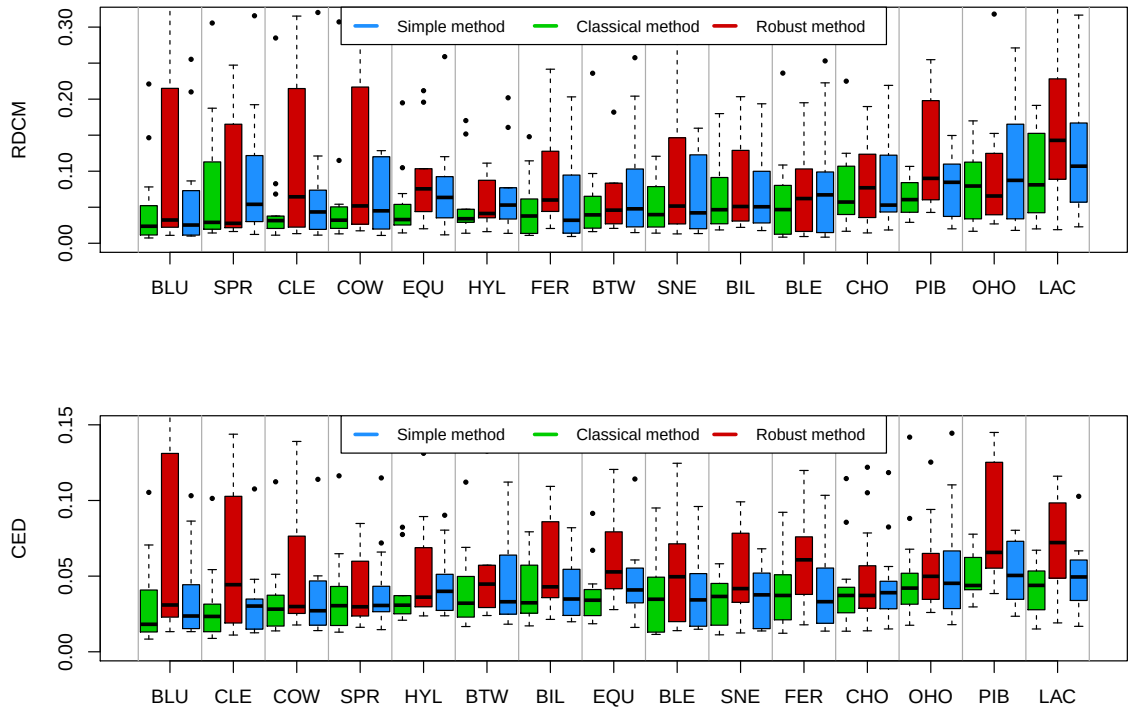


Figure 2.9: Imputation for each individual sample material (listed on the horizontal axis). The upper detection limit is modified from the quantile 0.5 to 0.95 of the values for a randomly selected variable, and the average of the error measures is computed. The boxplots show the outcomes for 100 simulations. The order of the sample materials is according to the median performance of the classical method.

Imputation is done only in one variable at a time, and the UDL is modified from the quantile 0.5 to 0.95. The average error among this range is computed for each individual variable, and the boxplots in Figure 2.10 show the results for 100 simulations. The boxplots are sorted according to the median for the classical method. There are quite large differences among the variables visible. The simple method sometimes leads to very poor results, e.g. for Pb, Mo, La, Ce or Co. The performance of the classical and robust method is quite comparable. The reason for the different performance of the variables has to be based on the strength of the relationship to the other variables, but it could also be affected by the fact that the different sample materials result in subpopulations in the data set.

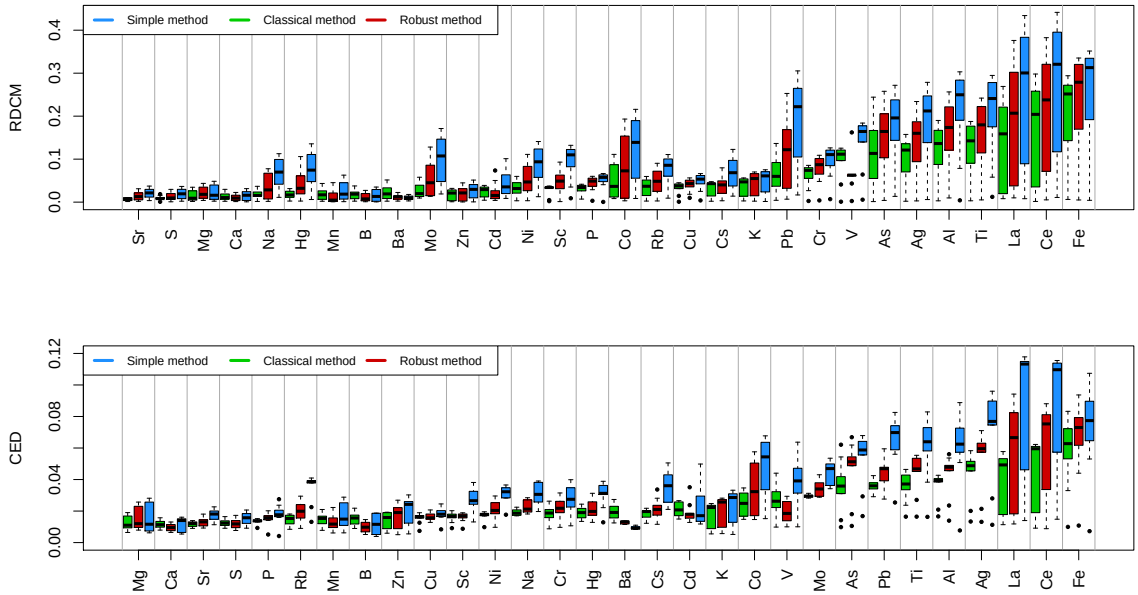


Figure 2.10: Imputation for the complete data set. The upper detection limit is modified from the quantile 0.5 to 0.95 of the values for a particular variable, and the average of the error measures is computed for each variable separately. The boxplots show the outcomes for 100 simulations. The order of the variables is according to the median performance of the classical method.

2.6.8 Summary

The main conclusions from the numerical experiments can be summarized as follows:

- The algorithm is stable and converges in few iterations, but convergence is slower for the robust method (in few cases there is very slow convergence). The algorithm still works well for a high fraction of variables to be imputed, and for a high proportion of censoring.

- Regression-based imputation usually outperforms simple imputation, which is to be expected since the initialization is done by the simple imputation method. It depends very much on the data structure and size if robust regression leads to better results than classical least-squares regression. Often, the results are quite comparable. In particular, we found that the robust method is preferable when the number of censored values is high (more than 30%). In most other experiments there is no clear advantage over the classical method.

- For a low number of observations (also compared to the number of variables), variable selection within the regression method leads to better results than regression with all the available variables, especially for a high proportion of censoring. The more observations, the better the imputation results of the classical and robust regression methods.

- The performance of the imputation can differ a lot for different variables in the same data set – depending on the data structure.

2.7 Discussion and conclusions

In this paper we proposed a regression-based method for the imputation of right-censored values in compositional data, making use of the Tobit model. An algorithm was developed based on the ideas of an algorithm for left-censored data (Palarea-Albaladejo and Martín-Fernández, 2008; Martín-Fernández et al., 2012), with the option for least-squares or robust regression with or without variable selection. The algorithm is available in the software environment `R` as the function `imputeUDLs()` in the package `robCompositions` (Templ et al., 2011).

The performance of the algorithm was investigated in detail and compared to a simple imputation method, imputing a value of 1.2 times the value of the upper detection limit. All experiments were based on two geochemical data sets, one with several data subgroups which were used in the numerical experiments for comparison (Reimann et al., 2018). For reproducibility of the presented results in Section 6, the second data set has been made available as data `gjovik` in the package `robCompositions`. This data set did not contain any censored values, and thus right-censoring was artificially introduced, and the imputed values could be compared to the reported values.

In real data sets, the proportion of samples above the upper detection limit could be high for some variables, probably even higher than 80%. In that case, the proposed algorithm would have to use less than 20% of the observations to estimate the parameters in the regression model in Equation (2.8), which could become very unstable, in particular if the data set has many variables. In such cases it might be more advisable to omit such variables from the analysis.

A further problem which appears with real data is that for some variables one could have both values above an UDL and values below a LDL. A simple strategy to deal with this problem could be as follows:

- Determine the proportion of UDL and LDL values separately for each variable.

- Omit variables with proportions higher than e.g. 30% to obtain higher stability in the iterative algorithm.

- – If the proportion of the remaining variables is "generally" higher for UDL than for LDL, do UDL model-based imputation by setting the LDL values to 65% of the corresponding LDL (Martín-Fernández et al., 2003). Afterwards, impute the LDL values using a regression method.
  – Otherwise, set UDL values to 1.2 times the corresponding UDL, and do LDL model-based imputation. Another multiplier such as 4/3 or 1.7 (Sanford et al., 1993) might work as well. Afterwards, impute the UDL values using the algorithm of Section 4.

One could iterate the last two steps until some error measure is smaller than a given threshold, see Section 2.4.

An alternative algorithm could consider the regression problem in Equation (2.8) only for the non-censored observations, and use the estimated regression coefficients to estimate the values above the UDL and below the LDL. As outlined in Section 2.4, one would iterate through all variables for the imputation, and then repeat the procedure until the results stabilize. The performance of both options will heavily depend on the pattern of the LDL and UDL values, but also on the number of observations/variables in the data set. Particularly difficult will be situations where the UDL and LDL values are present in more or less the same variables.

The above recommendations always depend on the purpose of the analysis. For instance, in geochemical exploration it would not be desirable to omit a variable which is considered as pathfinder element for mineralization, just because of a very high proportion of values above the UDL. In this case it might not even be necessary to estimate the values above the UDL, because the information of "high" values at specific locations might be sufficient for the purpose.

CHAPTER 3

Identification of mineralization in geochemistry along a transect based on the spatial curvature of log-ratios

Detecting subcropping mineralization but also deeply buried mineralization is one important goal in geochemical exploration. The identification of useful indicators for mineralization is a difficult task as mineralization might be influenced by many factors, such as location, investigated media, depth, etc. We propose a statistical method which indicates chemical elements related to mineralization along a transect. Moreover, the method determines along a transect the potential area of the deposit. The identification is based on General Additive Models (GAMs) for the element concentrations across the spatial coordinate(s). The log-ratios of the GAM fits are taken to compute the curvature, where high and narrow curvature is supposed to indicate the mineralization area. By defining a measure for the quantification of high curvature, the log-ratios can be ranked, and elements can be identified that are indicative of the anomaly patterns.

3.1 Introduction

Identifying geochemical processes as mineralization is defined as the presence of higher concentrations of particular chemical elements compared to the background concentration.

However, it is challenging to define the background concentration, since the threshold between background and mineralization will in general not be characterized by a single number (see Reimann and Garrett, 2005). Nevertheless, it would be expected that a biogeochemical anomaly in mineralization exploration is indicated by a rapid spatial change in the concentration on top of the mineralization, depending on the type and extent of the mineralization. Nowadays, identifying geochemical features related to geochemical signature, and the separation of background and target zones of future local mineral exploration are becoming popular challenges in geochemistry.

Geochemical data in the form of chemical element concentrations are naturally compositional data, which are strictly non-negative values, forming parts of a whole. In this context we talk about compositional data analysis, and the log-ratio methodology introduced by Aitchison (1986) is the most common approach in this context. The important information to be analyzed is reflected in the log-ratios between the variables rather than in the absolute values. This "relative" information is employed for a proper understanding of the data.

There are two main difficulties with such an approach for practical geochemical data sets: (a) Nowadays it is possible to measure the concentration of dozens of chemical elements, and this leads to hundreds (or more) possible pairwise log-ratios. Filtering out the elements which may indicate mineralization is thus challenging. (b) Especially for mineral exploration there might not be many observations available, because often they are the result of a pre-study of the area. This creates further difficulties for the prediction of the location of a potential mineralization.

Since the identification of mineralization is a very relevant topic in practice, there are numerous publications available in the literature. This problem is also known under geochemical anomaly mapping, referring to a map presentation of geochemical uni-element or multi-element soil or plant data. Related to the log-ratio methodology, the works of Buccianti et al. (2015), Carranza (2017) and Tolosana-Delgado and van den Boogaart (2014) aim to predict an anomalous presence of a mineral commodity. In these papers, a log-ratio transformation is applied, and then different mapping techniques are used to reveal the mineralization. The first mentioned paper uses centered log-ratio (clr) coefficients and isometric log-ratio (ilr) coordinates, since these representations preserve metric properties. Anomalous compositions then originate from the robust barycentre. Using robust methods, the variables are split into two groups, then the variation of log-ratios gives ratios being in geochemical relationships, however the interpretation can be relatively weak and not specific enough. The work Carranza (2017) uses enrichment factors and log-ratios, where log-ratios are in terms of ilr coordinates. Based on kriging,

a spatial correlation analysis was performed, where ilr values have much stronger positive spatial correlation with the known gold deposits. The paper also concludes that for mapping of significant anomalies, it is better to use ilr-transformed soil geochemical data than enrichment factors. A limitation is that this procedure is a supervised method, meaning that the deposits need to be known for the input. Another example, proposed in Tolosana-Delgado and van den Boogaart (2014), shows that compositional data analysis is useful as a first step to identify geochemical features linked to natural phenomena. Logistic regression-like techniques are used to obtain a combination of variables that favor the presence of mineralization. Geostatistics is used to interpolate the composition to unsampled locations. However, the two proposed methods – the Fisher method and Poisson processes – can lead to incomparable results. The methods rather rely on information about high log-ratios, not in consideration of any spatial changes. However, this still gives informative results combining potential areas of interest and also relevant favorability in sense of ratios.

The idea behind the presented method is that pairwise log-ratios would rapidly change towards a mineralized area. A rapid change would imply that the values of the log-ratio show strong curvature. However, based on the observation data, "curvature" can only be computed numerically, and this is infeasible if the rapid change is expressed only by very few observations. For this reason we will approximate the underlying element concentration data by smooth values, producing a continuous signal which allows to extract as many data points as necessary to compute a curvature later on. Smoothness is important at this stage, because otherwise one could obtain arbitrary jumps in the log-ratios, leading to artifacts in the curvature. One could also argue that the smoothed concentration values allow to suppress the effect of measurement uncertainties.

We decided to use Generalized Additive Models (GAMs) for the estimation of a smooth signal (Wood, 2017) and (Yee, 2015). The smoothness can be regulated by a tuning parameter, which is selected by cross-validation. A GAM fit is based on natural cubic splines with knots at every data point. Depending on the tuning parameter, one can obtain the whole spectrum from the linear fit to the very non-smooth exact fit. Once the GAM fits for both input variables of the log-ratio are available, the curvature of the log-ratio of the smoothed concentrations can be computed, and an unsupervised learning method is employed, leading to a hitlist of log-ratios most suitable for finding mineralization. The proposed method has been tested on two real data sets, where the mineralized zones are known, and the results seem to be reliable and carrying out promising prospects.

The paper is organized as follows: Section 2 introduces the methodology and provides

a closer description of GAMs, additional information on the concepts of the curvature and its measurement. A detailed algorithm for the whole procedure of ranking of log-ratios is proposed in Section 3. Section 4 demonstrates numerical experiments based on two real geochemical data sets. The last Section 5 concludes and provides possible extensions of the proposed method.

3.2 Methodology

3.2.1 Motivating artificial example

The principal idea is to investigate the curvature of log-ratios. The higher an index involving the curvature, the more likely a point of mineralization has been identified.

For illustration purposes, let us consider the function $x \mapsto (1 + (\frac{x}{\sigma})^2)^{-1}$, for fixed $\sigma > 0$. Figure 3.1 shows on the top row the function itself for different values of σ, and the bottom row presents the corresponding curvatures. An appropriate measure of curvature will be defined later in Section 2.3. As it can be seen on the top row, the lower σ becomes, the quicker the spatial change of the function is. Looking at the bottom row, this translates into a growing value of the curvature at zero. The case of $\sigma \to \infty$, not displayed here, corresponds to the function being constant equal to one, thus having zero curvature. The methodology developed in the following involves the curvature as a measure of how quick a signal undergoes spatial change. Of course, this measure needs to be normalized appropriately in case of several peaks with possibly different curvature.

3.2.2 GAMs

Since in many studies only few observations are available, the original concentration data are approximated by smooth curves originating from Generalized Additive Models (GAMs), before log-ratios for computing curvature are considered. This is preferable to computing a curvature measure directly from the log-ratios of the observations, since with smooth fits one can in principle generate arbitrarily many observations, and derive a more stable value for the curvature. GAMs have the advantage that the degree of smoothness of the fit to the data can be tuned.

Denote $(x_1, y_1), \ldots, (x_n, y_n)$ the n observed data of the measured concentration y at position x of a certain element, where we assume that $x_1 \leq \cdots \leq x_n$ are measured along a linear transect. At the heart of GAMs is a weighted penalized log-likelihood problem

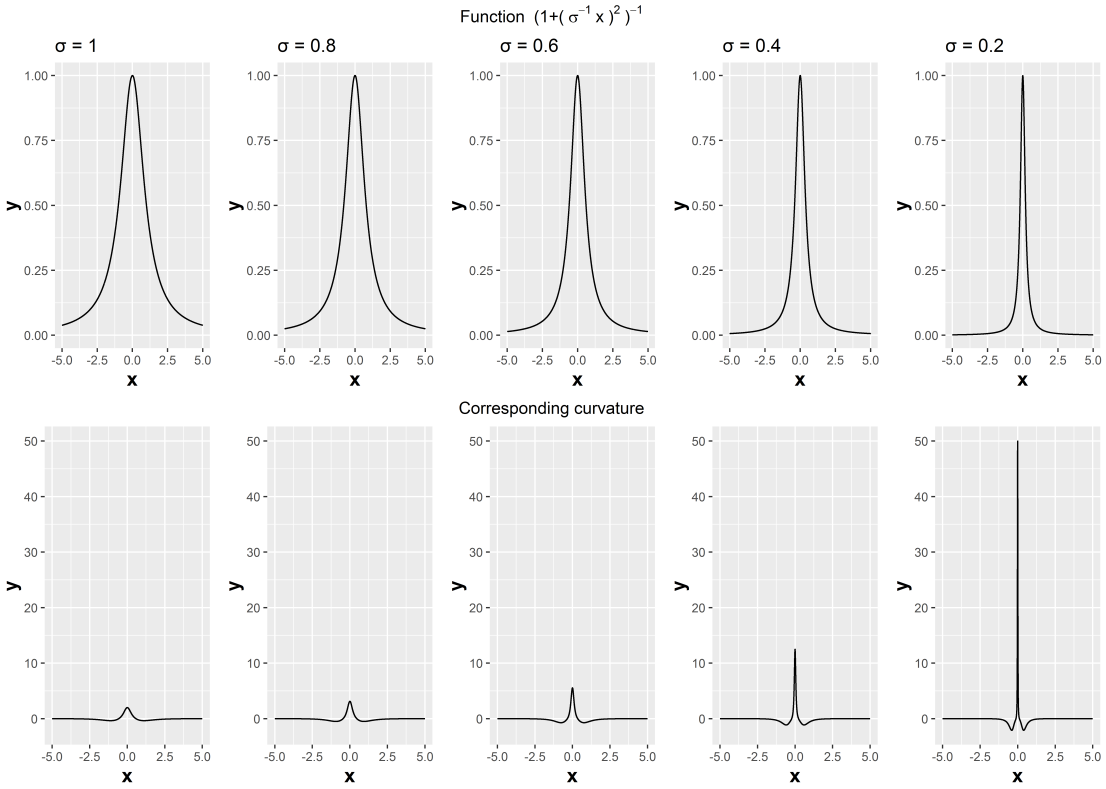


Figure 3.1: Top row: function $x \mapsto (1+(\frac{x}{\sigma})^2)^{-1}$ for different σ. Bottom row: corresponding curvature (to be defined in Section 2.3).

over a suitable function space $\mathcal{H}$, estimating the presumably smooth linear predictor η

$$\hat{\eta} = \arg\max_{\eta \in \mathcal{H}} \sum_{i=1}^{n} \omega_i l(y_i|x_i;\eta) - \lambda \int (\eta''(x))^2 dx, \tag{3.1}$$

where λ is the so called smoothing parameter, l is the log-likelihood function, and ω_i are predefined weights. To explain this further, we note that in the GAM framework one necessary assumption is that the response y belongs to the exponential family, thus its density is of the form

$$f(y|x) = \exp\left(\frac{\theta y|x - b(\theta)}{a(\psi)} + c(y|x,\psi)\right),$$

with parameters θ, ψ and given functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. It can be shown that with this assumption we are able to rewrite the log-likelihood in dependence of the conditional mean $\mathbb{E}(y|x)$. Given now a so called link function $h(\cdot)$ – a smooth monotonic function which the user must normally choose, except in some special cases for which we get a canonical

link – we model the composition of the expectation of y with $h(\cdot)$ as $h(\mathbb{E}(y|x)) = \eta(x)$ and are therefore able to implicitly write the log-likelihood in dependence of the linear predictor η. Modeling the mean in such a way, we are able, once η is estimated, to make predictions of $y|x$ through $h^{-1}(\eta(x))$. The choice of $h(\cdot)$ is in many cases not crucial as long as its domain matches with the range of possible values of $\mathbb{E}(y|x)$.

The smoothing parameter λ controls the trade-off between smoothness and the fit to the data; the bigger λ becomes, the smoother the function will be, as in the case $\lambda \to \infty$ we get $\eta'' \equiv 0$ and therefore η is a linear function. Typically, the smoothing parameter λ is chosen in a data dependent way by using either Generalized Cross Validation (GCV) or Restricted Maximum Likelihood (REML).

For fixed λ, the function η solving problem (3.1) can be written in terms of a cubic B-spline basis, see Friedman et al. (2001), for example. Therefore, $\eta(x) = \sum_{j=1}^{n} h_j(x)\beta_j$, where h_j are the cubic B-spline basis functions, and β_j the coefficients.

For a more thorough introduction to GAMs, GCV, REML and how problem (3.1) is solved algorithmically, we refer to Wood (2017).

In the applications of our method to the data of Section 4, we decided to model $y_i|x_i$ belonging to the family of Tweedie distributions with a log-link function $h \equiv \log$. This means that we model

$$h(\mathbb{E}(y|x)) = \eta(x) \qquad \mathbb{V}(y|x) = \mathbb{E}(y|x)^p \frac{\psi}{\omega(x)} \tag{3.2}$$

for some $p \in (1, 2)$. Furthermore, to capture the effect of outliers in the response we use predetermined weights ω_i; for example we took $\bar{\omega}_i := \max\left(\frac{|y_i - \hat{\mu}|}{\hat{\sigma}}, 1\right)$, where $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and the sample standard deviation of y, and then put $\omega_i = \frac{\bar{\omega}_i}{\sum \bar{\omega}_i}$.
In these particular examples, the choice of this family of distributions and this specific link function is motivated by the fact that it provides a very flexible range of modelling the mean-variance relationship – as it comprises many distributions through the additional parameter p. Furthermore, the predetermined weights have been chosen in a way such that outliers get upweighted, i.e. if a point y_i is bigger than $\hat{\mu} + \hat{\sigma}$ then it will get upweighted proportionally. This seems necessary as the variance will likely be higher on top of mineralization and thus over- or under-dispersion in our model might otherwise appear. All in all, inspecting the linear predictor vs. the residuals as well as the fitted values vs. the response plots show consistency with our choice of link-function and thus also with model (3.2). Some of these plots are shown in Section 4.

3.2.3 Curvature of log-ratios

Since we are interested in the log-ratios of two chemical elements, we denote for an element el_1 and an element el_2 their respective GAM fits on the response scale $\hat{f}_{el_1}(x) := h^{-1}(\hat{\eta}_{el_1}(x))$ and $\hat{f}_{el_2}(x) := h^{-1}(\hat{\eta}_{el_2}(x))$. The log-ratio of the fits

$$g(x) := \log\left(\frac{\hat{f}_{el_1}(x)}{\hat{f}_{el_2}(x)}\right) = \log\left(\hat{f}_{el_1}(x)\right) - \log\left(\hat{f}_{el_2}(x)\right)$$

is then shifted and scaled to $c(g(x) - \min g(x))$, with the scaling constant $c := |\max_{x\in[x_1,x_n]} g(x) - \min_{x\in[x_1,x_n]} g(x)|^{-1}$, whenever g is not constant. This is done to make our method comparable across different log-ratios. In the special case of g being constant we can set c to one, as such functions will be ranked lowest by the measure described below.

As a next step we will define a measure for identifying important log-ratios based on the curvature. The curvature κ of the shifted and scaled function is defined as

$$\kappa(x) := \frac{|cg''(x)|}{(1 + (cg'(x))^2)^{\frac{3}{2}}}, \tag{3.3}$$

see Kline (1998), and thus, as

$$g'(x) = \frac{\hat{f}'_{el_1}}{\hat{f}_{el_1}}(x) - \frac{\hat{f}'_{el_2}}{\hat{f}_{el_2}}(x)$$
$$g''(x) = \frac{\hat{f}''_{el_1}}{\hat{f}_{el_1}}(x) - \left(\frac{\hat{f}'_{el_1}}{\hat{f}_{el_1}}(x)\right)^2 - \frac{\hat{f}''_{el_2}}{\hat{f}_{el_2}}(x) + \left(\frac{\hat{f}'_{el_2}}{\hat{f}_{el_2}}(x)\right)^2$$

holds, this amounts to calculating the first and second derivatives of the GAM fits on the response scale of the individual elements.

3.2.4 Curvature exceeding a threshold

For each such a combination of elements we calculate the mean and the variance of the curvature κ, namely

$$\mu = \int_{x_1}^{x_n} \kappa(x)dx \tag{3.4}$$
$$\sigma^2 = \int_{x_1}^{x_n} (\kappa(x) - \mu)^2 dx \tag{3.5}$$

and define the set of crossings with the threshold $\mathcal{T} := \mu + \sigma$ by

$$\mathcal{C} := \{x \in [x_1, x_n] | \kappa(x) = \mathcal{T}\} \bigcup \{x \in \{x_1, x_n\} | \kappa(x) \geq \mathcal{T}\}, \tag{3.6}$$

where the second set contains x_1 and/or x_n depending if they are above the threshold or not. The purpose of defining this set is to detect the points where the curvature κ crosses the threshold and subsequently exceeds it, so that we are left with only a few high local maxima – see for example Figure 3.2. Of course these maxima depend on the definition of the threshold, and it seems reasonable to take the mean plus the standard deviation over the whole range, because, as it is implied by Chebyshev's inequality, the further we get from this threshold the less likely an observation is.

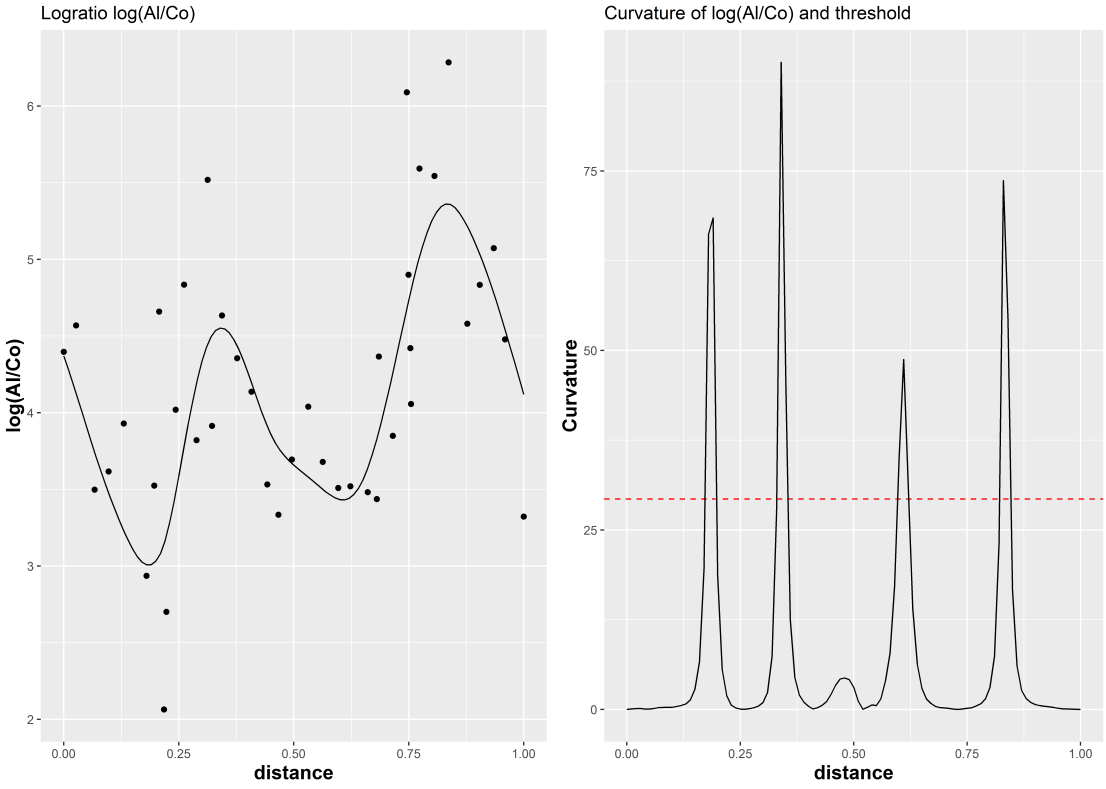


Figure 3.2: Example of a log-ratio plot of the elements Al and Co and the corresponding curvature plot with the threshold (dashed line). One can see the correspondence between local maxima above the threshold in the right plot with the peaks in the left plot.

Remark. *If the set $\mathcal{C}$ is finite it follows that its cardinality is even by definition. For example, if the second set in (3.6) is empty and the first is non-empty, then the set must have even cardinality as otherwise before or after the last crossing it must always remain above the threshold, thus $\kappa(x_1) > \mathcal{T}$ or $\kappa(x_n) > \mathcal{T}$.*

Instead of analytically solving the equation $\kappa(x) = \mathcal{T}$ we decided to uniformly

sample a high number of points N in the interval (x_1, x_n) - where we always add $\{x \in \{x_1, x_n\} | \kappa(x) \geq \mathcal{T}\}$ to the set - and then check for each of these points x_l if $\kappa(x_l) - \mathcal{T}$ has a change of sign. If so, we add this point to the set and we denote $\hat{\mathcal{C}}$ the set constructed in this way. We do the latter so that we can continue to work with a finite set which is of even cardinality, and one could see the measure constructed in the following as an approximation to the analytical case.

3.2.5 Measure for comparing curvature

We define the measure to compare the curvature values of the log-ratio of two elements el_1 and el_1 as

$$c(el_1, el_2) := \frac{2}{L} \sum_{l=1}^{\frac{L}{2}} \max_{x \in [x_{j_{2l-1}}, x_{j_{2l}}]} (\kappa(x) - \mathcal{T})_+^2, \tag{3.7}$$

where $(\cdot)_+$ denotes $\max(\cdot, 0)$, and where $x_{j_1} < ... < x_{j_l}$, $l = 1, ..., L$ are the ordered points of the finite set $\hat{\mathcal{C}}$. We will call this measure c-value in the following.

The more the curvature κ exceeds the threshold in $[x_{j_{2l-1}}, x_{j_{2l}}]$, the bigger the maximum and thus also the measure will be. Therefore, a relatively fast change in the original signal will contribute a lot to this measure. By including the factor $\frac{1}{2L}$, the measure $c(el_1, el_2)$ becomes the mean, and thus $c(el_1, el_2)$ is high if the peaks above the threshold $\mathcal{T}$ are high on average.

Since this measure is normalized, it can be used to compare all different pairs of log-ratios, and it can even be used to compare different data sets taken at the same locations, such as measurements from different sample materials or soil layers. The log-ratio pairs can be ordered according to the value of the measure, and the pairs corresponding to the highest ranking will be most promising for the identification of mineralization.

3.3 Algorithm

In the following we will describe the algorithm using the methodology above, which takes as an input the element concentrations for n observations, denoted as the vectors $\mathbf{y}_{el_1}, \ldots, \mathbf{y}_{el_m}$ of length n, where $(y_{el_k})_i$ is the i-th observation of the measured concentration of the k-th element, and the location vectors $\mathbf{x}$ of length n, where x_i is the i-th observed location. The output is a matrix C with entries $c(el_r, el_s)$ for different elements el_r and el_s.

Step 1: Before fitting the GAM model we scale the entries of $\mathbf{x}$ to the range $[0, 1]$, and then we calculate the weights $(\omega_{el_k})_i$ for the element concentrations (here for the k-th element), see definition below model (3.2).

Step 2: As a next step we fit a GAM model to each element, meaning that for the measurements $(x_i, (y_{el_k})_i)$ we solve

$$\max_{\eta_{el_k}} \sum_{i=1}^{n} (\omega_{el_k})_i \; l((y_{el_k})_i | x_i; \eta_{el_k}) - \lambda \int (\eta''_{el_k}(x))^2 dx.$$

For the applications presented in Chapter 4 we have decided to use the Tweedie family and the log-link function. The fitting is done with the help of the R package `mgcv` (Wood, 2017), and the smoothing parameter λ is tuned automatically by using the implemented REML criterion.

Step 3: Once all the elements have been fitted, thus once we have computed all $\hat{\eta}_{el}$, we compute for a high number of points $x \in [x_1, x_n]$, typically we used N = 3000, all the possible shifted and scaled log-ratios at these points. Therefore, denoting $\mathcal{X}$ the ordered set of these points x, we calculate for each possible pair of elements el_1 and el_2, $c := |\max_{x\in\mathcal{X}} g(x) - \min_{x\in\mathcal{X}} g(x)|^{-1}$, where $g(x) = \log(\hat{f}_{el_1}(x)) - \log(\hat{f}_{el_2}(x))$; $\hat{f}_{el_1}(x) := h^{-1}(\hat{\eta}_{el_1}(x))$ and $\hat{f}_{el_2}(x) := h^{-1}(\hat{\eta}_{el_2}(x))$. If g is constant we set $c = 1$.

Step 4: As a next step we calculate the curvature of these log-ratios. Thus, for each such pair of elements el_r and el_s and all $x \in \mathcal{X}$ we need to determine $g'(x)$ and $g''(x)$ first. This is done numerically. For a small ϵ, say 10^{-3}, we compute the approximate derivatives for all elements

$$\hat{f}'_{el}(x) \approx \frac{1}{2\epsilon}(\hat{f}_{el}(x+\epsilon) - \hat{f}_{el}(x-\epsilon))$$
$$\hat{f}''_{el}(x) \approx \frac{1}{\epsilon^2}(\hat{f}_{el}(x+\epsilon) - 2\hat{f}_{el}(x) + \hat{f}_{el}(x-\epsilon))$$

and as

$$g'(x) = \frac{\hat{f}'_{el_1}}{\hat{f}_{el_1}}(x) - \frac{\hat{f}'_{el_2}}{\hat{f}_{el_2}}(x)$$
$$g''(x) = \frac{\hat{f}''_{el_1}}{\hat{f}_{el_1}}(x) - \left(\frac{\hat{f}'_{el_1}}{\hat{f}_{el_1}}(x)\right)^2 - \frac{\hat{f}''_{el_2}}{\hat{f}_{el_2}}(x) + \left(\frac{\hat{f}'_{el_2}}{\hat{f}_{el_2}}(x)\right)^2$$

holds, it is easy to compute $\kappa(x) := \frac{|cg''(x)|}{(1+(cg'(x))^2)^{\frac{3}{2}}}$ for each pair of elements and $x \in \mathcal{X}$.

Step 5: After this we compute an approximation to the treshold τ by approximating (3.4) and (3.5). Thus we define

$$\tau := \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\kappa(x) + \sqrt{\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\left(\kappa(x) - \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\kappa(x)\right)^2}.$$

Step 6: Next we compute an approximation to the set $\mathcal{C}$. We define the set $\hat{\mathcal{C}}$ as all the x for which we have $\kappa(x) = \tau$ or for which $\kappa(x)$ is smaller than τ and then the next element in $\mathcal{X}$, say $\bar{x}$, we have $\kappa(\bar{x}) > \mathcal{T}$. Also we add x_1 and/or x_n if κ is bigger or equal than τ there. This seems like a reasonable approximation as long as the cardinality of $\mathcal{X}$ is high enough.

Step 7: Finally, we compute for each pair of elements the measure $c(el_r, el_s) := \frac{2}{L}\sum_{l=1}^{\frac{L}{2}}\max_{x\in I_l}(\kappa(x) - \mathcal{T})_+^2$, where $I_l := [z_{j_{2l-1}}, z_{j_{2l}}]$, z_j is a point of the set $\hat{\mathcal{C}}$ and where L is the cardinality of $\hat{\mathcal{C}}$.

In short, the steps above can be subsumed into the following algorithm:

Algorithm 1 Log-ratio measures

for k = 1,...,m **do**
  Calculate weigths, i.e. calculate $\hat{\mu}$ and $\hat{\sigma}$ of $\mathbf{y}_{el_k}$ and set $(\omega_{el_k})_i := \max(\frac{(y_{el_k})_i - \hat{\mu}}{\hat{\sigma}}, 1)$
  Solve $\max_{\eta_{el_k}} \sum_{i=1}^{n}(\omega_{el_k})_i\ l((y_{el_k})_i|x_i;\eta_{el_k}) - \lambda\int(\eta''_{el_k}(x))^2dx$
end for
for (r,s) in {1,...,m} **do**
  Calculate curvature κ for the shifted and scaled log $\left(\frac{f_{el_r}}{f_{el_s}}\right)$
  Compute μ and σ^2 as described in (3.4) and (3.5) and set $\mathcal{T} := \mu + \sigma$
  Draw uniformly N points from (x_1, x_n) and add crossing points of κ with $\mathcal{T}$ to $\hat{\mathcal{C}} = \{x \in \{x_1, x_n\}|\kappa(x) \geq \mathcal{T}\}$
  Set $c(el_r, el_s) := \frac{2}{L}\sum_{l=1}^{\frac{L}{2}}\max_{x\in I_l}(\kappa(x) - \mathcal{T})_+^2$
end for
Define matrix C with entries $c(el_r, el_s)$
return Matrix C

Once these values $c(el_r, el_s)$ are obtained for all element combinations, we can compute a ranked list for the log-ratios from highest to lowest, or a heatmap based on the matrix C – where we scale the entries by the maximum entry first. As the measure $c(el_r, el_s)$ is comparable across materials, it is also possible to use these matrices to explore accumulated heatmaps for a comparison of different plant materials, see next section.

The whole algorithm as described above has been implemented in the software environment `R` (R Development Core Team, 2018) using mainly the `gam()` function implemented in the package **mgcv** (Wood, 2012). This software is available from the authors upon request.

3.4 Experimental results

The proposed method has been tested on two real geochemical data sets with known mineralization. Both data sets are sampled along a (more or less linear) transect. The known locations of mineralization can be used to evaluate the proposed procedure.

3.4.1 Juomasuo data

The Juomasuo data set is described in detail in Middleton et al. (2018), and it originates from the UltraLIM project, where biogeochemical samples have been taken in the years 2013 and 2014 in a subarctic region in northern Finland. Juomasuo, among all available sites, is the largest of the known Au deposits in the area. We take the data from 2013, where three different sample plant materials have been collected (Crowberry, Bilberry and Labrador tea). The investigated tissue of the plant species is either twig/stem or leaf/needle, and they have been analyzed for the concentration of more than 40 chemical elements. Depending on the plant material, 27 to 30 samples are available, and we focus on the concentration values of 27 chemical elements with reasonable data quality. Moreover, plants showed strong positive apical anomaly patterns for cobalt (Co), iron (Fe), thorium (Th), uranium (U) and rare earth elements (REE), such as cerium (Ce), lanthanum (La) and neodymium (Nd) in more than one species.

The sample locations are approximately along a line, and as a first step the distances between the samples are computed. The distance between the most extreme sites is 1271 meters, an on average one sample has been taken per 45 meters. For reasons of comparability, the distances are normalized from 0 to 1 as an input for the GAM models (see horizontal axis of Figure 3.3). The known mineralization is at the following (normalized) distances: 0.3, 0.38, 0.43, 0.48, 0.51, 0.53 and 0.55.

Figure 3.3 shows the GAM fits applied on eight selected variables measured in Crowberry twigs, together with the original concentration values (dots). The GAM fits result in a very smooth signal, even around normalized distance 0.75, where a gap in the sampling procedure occurred due to a peat bog. As mentioned above, the known mineralization is at distances between 0.3 and 0.55, and this is visible also in Figure 3.3, where one can clearly see anomaly patterns around these distances. Due to the choice of

the weights for the GAM fit, the outliers have a stronger impact on the fits, which is a desirable effect for the purpose of anomaly detection.

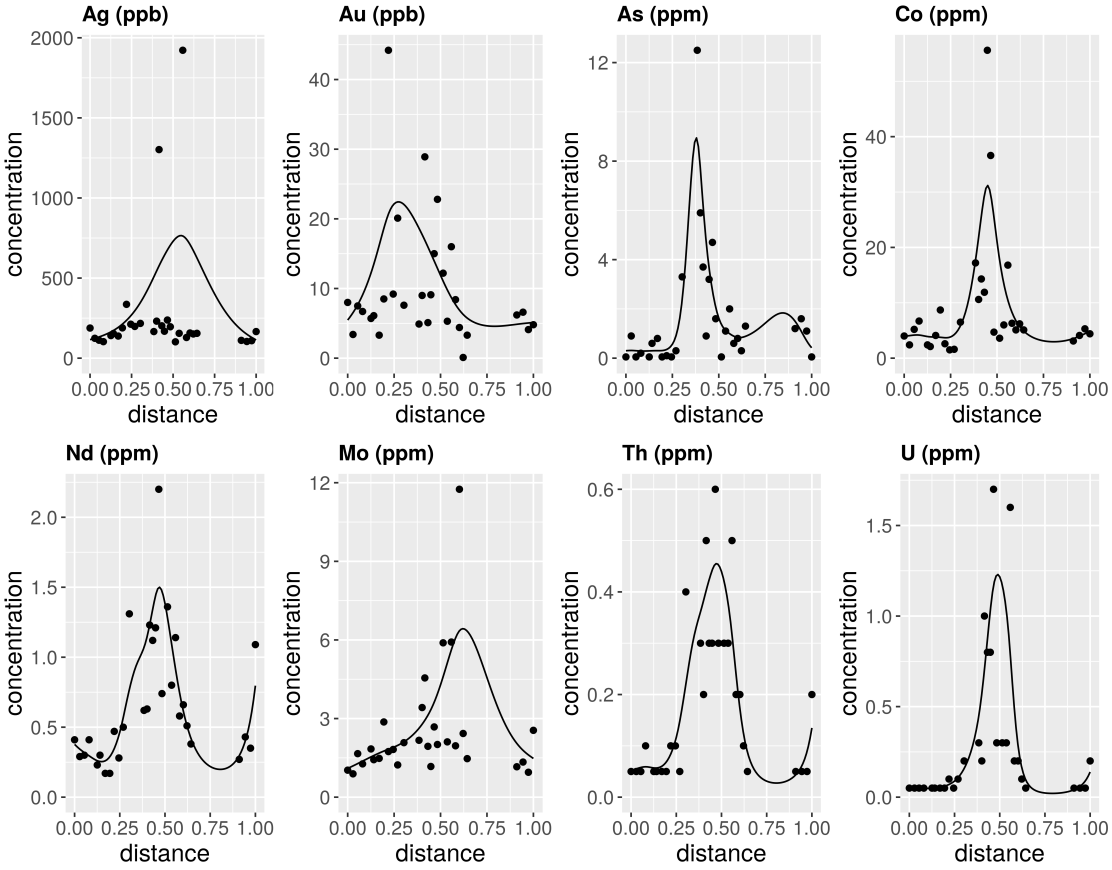


Figure 3.3: GAM fits (lines) for eight selected elements measured in Crowberry twigs from the Juomasuo data set are displayed together with their original concentrations (dots).

Once the GAM fits are available for all elements, their log-ratios for all different element pairs can be computed, together with the curvature measure. Figure 3.4 shows four examples of such log-ratios, their curvature, and the corresponding thresholds (dashed lines). These examples indeed reveal some of the known mineralization, shown by large spatial variability which is reflected by high curvature. Note that due to the use of log-ratios, we are not necessarily interested in high peaks but also in low ones.

Because of their relatively high curvature values and very narrow peaks, the pairwise log-ratios shown in Figure 3.4 have high values for our c-value measure defined in Equation (3.7). In fact, these c-values are assigned to the first top ranked 6 log-ratios among 351 log-ratios available in total for the particular sample material. Table 3.1

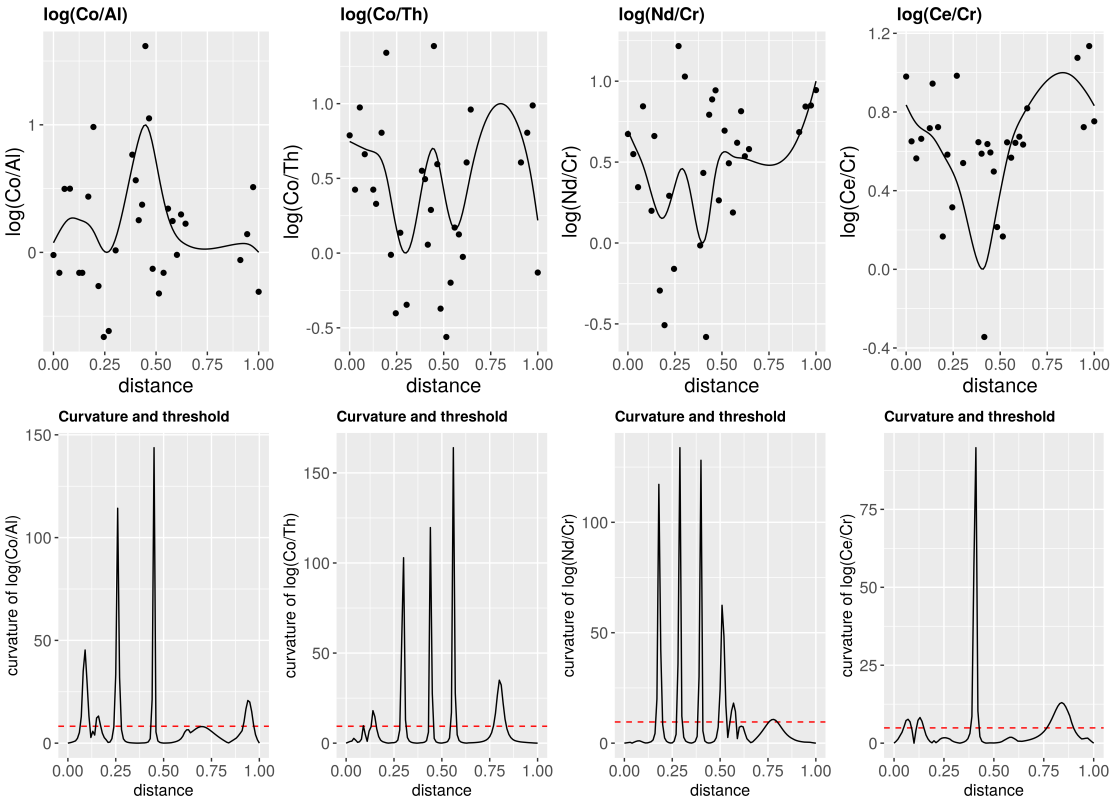


Figure 3.4: Upper part: four different log-ratios of GAM fits. Lower part: corresponding curvature together with the threshold (dashed red line).

presents the pairwise log-ratios for the top ten curvature measures for each sample material. Here, the measures have been scaled to the interval $[0, 1]$ first (for each sample material individually) for reasons of comparability. For example, for Crowberry-twig one can see that cobalt is involved often in the top ranked log-ratios, and thus this element seems to be a "pathfinder" for mineralization. Indeed, cobalt plays an important role concerning gold deposits. Depending on their element uptake, different plant materials can involve different elements in the top log-ratio pairs.

Similarly, the interpretation for individual combinations of plants can be provided. This information is stored in Table 3.2, where the column "element" provides the first top 10 elements which appear most often in the best ranked log-ratios for the particular plant media. For example, for plant species Crowberry and plant tissue twig, the elements Co, As, Cr, Bi, Nd (in this order) most frequently appear among the best ranked log-ratios. These elements can thus be considered as pathfinder elements.

The information contained in the ranked lists can also be visually summarized in

Table 3.1: Top 10 ranked log-ratios and its scaled c-values for the plant materials Crowberry (CRO), Bilberry (BIL), and Labrador tea (LBT).

	CRO-twig		CRO-leaf		BIL-twig		BIL-leaf		LBT-twig		LBT-leaf	
	pair	c	pair	c	pair	c	pair	c	pair	c	pair	c
1	Co/Al	1	Au/As	1	Se/Na	1	La/Th	1	As/Th	1	Se/Ag	1
2	Co/Fe	0.95	As/Sc	0.99	Ba/Se	0.92	U/Fe	0.86	Co/As	0.93	Ag/Al	0.82
3	Co/Ce	0.84	Bi/Pb	0.99	Fe/Se	0.9	U/S	0.81	As/Mo	0.88	Ce/Ag	0.81
4	Co/Th	0.78	As/Ti	0.95	Se/Th	0.83	U/Al	0.79	As/Sc	0.78	La/Ag	0.78
5	Ce/Cr	0.73	Nd/As	0.93	S/Se	0.8	U/Ni	0.77	As/Fe	0.74	Ag/Sc	0.77
6	Nd/Cr	0.73	As/Cr	0.9	Pb/Se	0.76	Cu/U	0.73	U/As	0.74	Ag/Y	0.73
7	Co/Cr	0.72	As/Ce	0.9	Se/Al	0.76	U/V	0.7	Co/Th	0.73	Fe/Ag	0.73
8	La/Cr	0.67	As/Al	0.88	Se/Ti	0.76	U/Na	0.63	As/Ce	0.7	Nd/Ag	0.72
9	Co/V	0.67	As/Na	0.88	Au/Se	0.7	U/La	0.5	As/La	0.67	Cu/Ag	0.7
10	Co/La	0.64	Bi/Ni	0.88	Ce/Se	0.7	U/Th	0.5	As/Y	0.66	Ag/Ti	0.67

Table 3.2: Top 10 ranked log-ratios and its elements for each material.

	CRO-twig	CRO-leaf	BIL-twig	BIL-leaf	LBT-twig	LBT-leaf
	element	element	element	element	element	element
1	Co	As	Se	U	As	Ag
2	As	Bi	U	Bi	Co	Y
3	Cr	Co	W	La	U	Nd
4	Bi	Sc	Tl	Th	Fe	Ce
5	Nd	Cr	Ag	Cr	W	Au
6	Ce	Fe	Co	Fe	Th	Al
7	Fe	Al	Fe	S	Se	Co
8	La	Y	Na	Ce	Al	Fe
9	Al	La	Bi	Y	Mo	La
10	Y	Ti	Ba	Cu	Ag	U

heatmaps. The scaled c-value measures need to be mapped to colors, where in the following representation 0 was mapped to white, and 1 to dark blue, with a continuous spectrum between these extremes. Figure 3.5 shows the resulting heatmaps for the sample media twigs of the different plant species, as well as a heatmap for the accumulated values of all sample materials (upper left). The heatmaps represent the different elements in the rows and columns, with symmetry around the diagonal. Each cell in the heatmap refers to the scaled c-values of the corresponding pairwise log-ratio. For instance, the upper left plot for accumulated panel shows that silver (Ag) is involved in many log-ratios with high values of the c-values. Also arsenic (As), bismuth (Bi), cobalt (Co), selenium (Se) and uranium (U) are present in many important log-ratios. The interpretation of those mentioned elements can be partially seen in Middleton et al. (2018). Ag, Bi and Se are elements verified by lithogeochemistry and also elements exhibiting anomalous spatial patterns

over the mineralization. Cobalt creates together with gold the underlying hydrothermal deposit, where Uranium is one of the elements showing spatial multi-elemental anomaly patterns for Au-Co deposits. Arsenic is one of the interesting pathfinder elements from the perspective of geochemical exploration. The heatmaps for the individual plant materials provide different information, because it depends very much on the plant materials which elements are enriched by a potential mineralization. The heatmap of the sample medium Crowberry-twig (upper right) shows a couple of highly ranked log-ratios indicated by dark blue color, where cobalt seems to be involved in several log-ratios, for instance log(Co/Al), log(Co/Fe), log(Co/Ce), log(Co/Th), etc. One of the conclusions in Middleton et al. (2018) is that Crowberry twigs were the most efficient plant tissues for revealing the location of the mineralized lodes determined as background for Au-Co deposits. Another investigated plant was Bilberry twig (lower left heatmap) clearly showing two elements with high c-values, i.e. Selenium and Uranium. In fact, it turned out that Bilberry twig was rather a poor quality indicator for many elements, forming spatial anomaly patterns for only a few elements. The last heatmap (lower right) for Labrador tea and its twig clearly shows Arsenic as the mostly involved in highly ranked log-ratios. Arsenic belongs to the group of so called pathfinder elements; another important element seems to be cobalt and Uranium forming the group of ore elements.

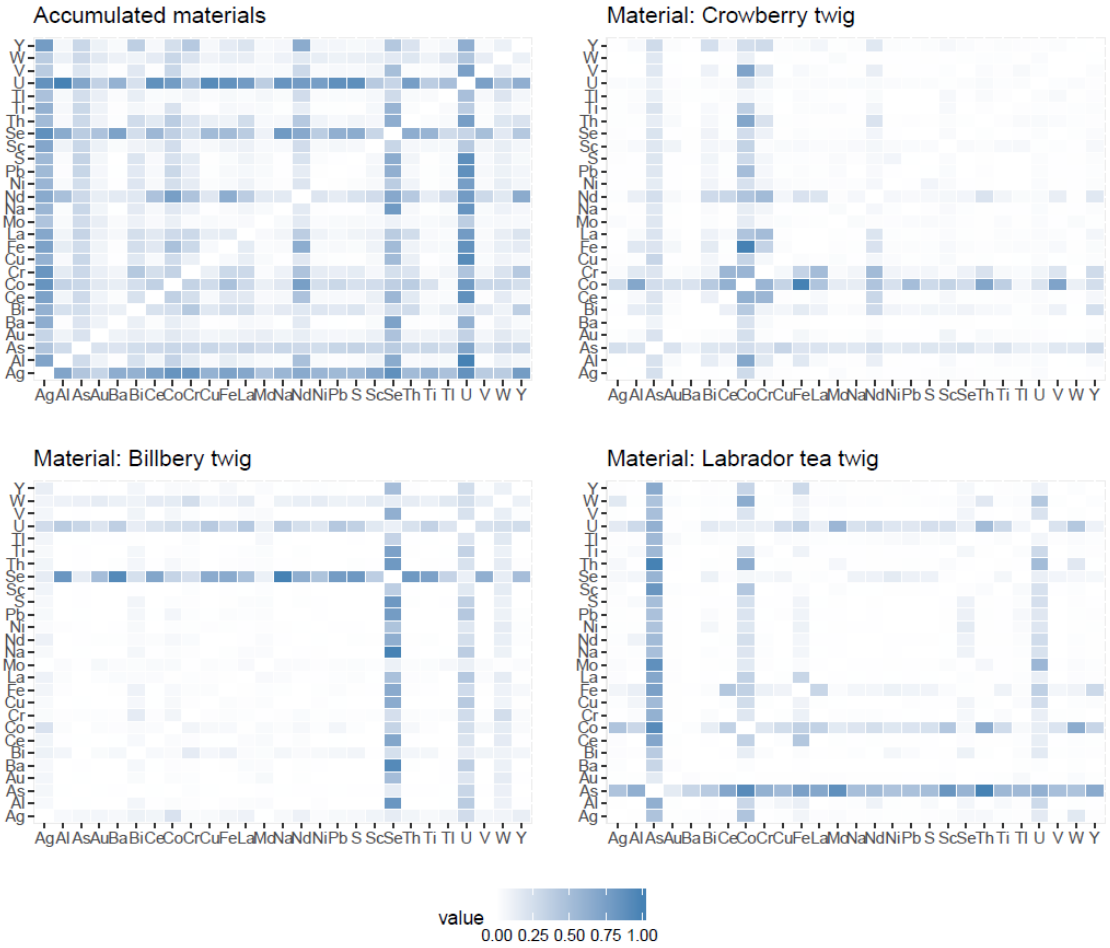


Figure 3.5: Heatmaps of the c-values per element for all possible log-ratios of the tissue twig for all plant species, and the accumulated values of all materials (upper left).

3.4.2 Gjøvik Data

As a second application we use the Gjøvik data set, which originates from a project of the Geological Survey of Norway (NGU) in a 100 km transect in Gjøvik, Norway (Reimann et al., 2018). In total, 15 different sample materials have been investigated, soil as well as plants, and approximately 40 samples are available for each subdata set. They have been sampled more or less on a linear transect, and for our method we first derived the distances between the samples by projection onto a line. We selected 30 chemical elements with reasonable data quality. The GAM fits have been computed for each element and each sample material, followed by computing the log-ratios and the curvature measure.

Figure 3.6 shows the resulting heatmaps for four selected sample materials, Birch leaves (BIL), Blueberry leaves (BLE), Cowberry leaves (CLE), and Spruce needles (SNE). Three of these plots show that lead (Pb) seems to be a pathfinder element, but also Tl (thallium), Mo (molybdenum), Sn (tin), and Ti (titanium) result in log-ratios with high c-values. In fact, the Gjøvik data set has been investigated because there is known mineralization of lead (Pb) and molybdenum (Mo).

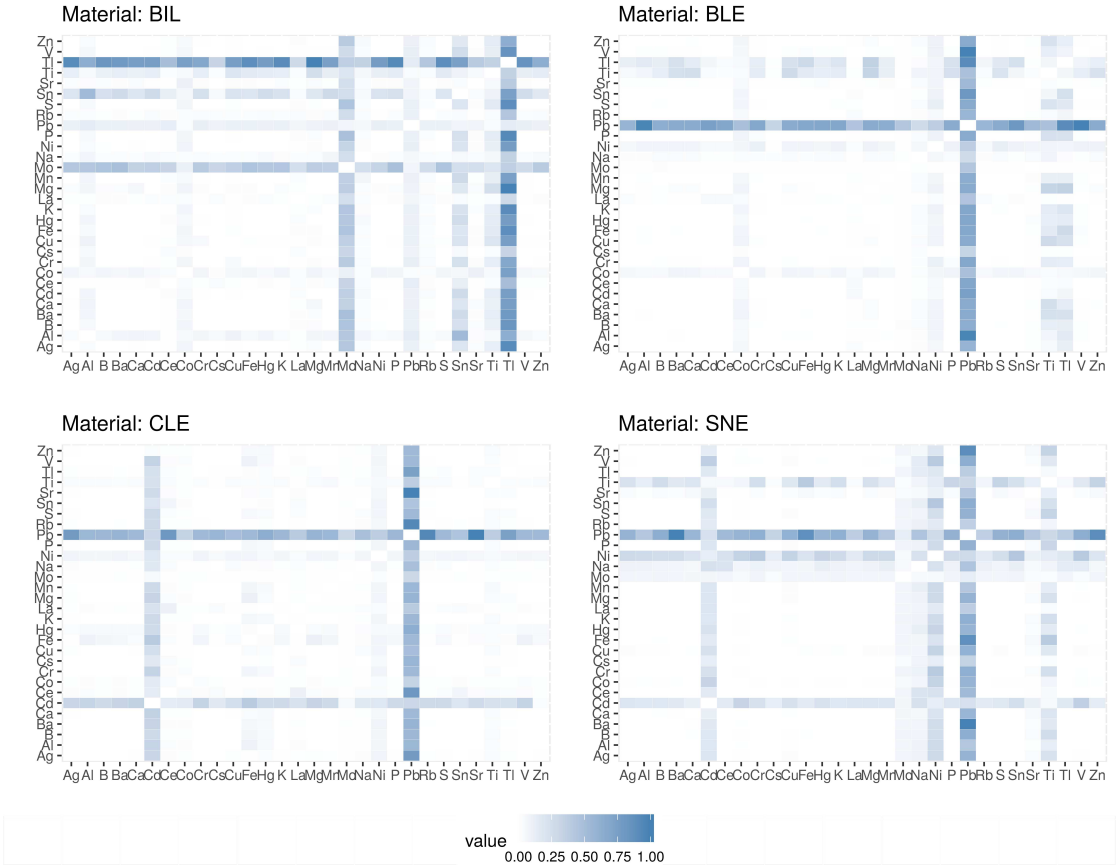


Figure 3.6: Heatmaps of the c-values for all possible log-ratios of four media – BIL, BLE, CLE, SNE.

Figure 3.7 focuses on the two elements Mo and Tl for the sample material Birch leaves. Both elements may be relevant for identifying mineralization. The solid line in the plot is the log-ratio of the GAM fits for these two elements, and the dashed line corresponds to the threshold used inside the algorithm to compute the curvature measure. The blue points indicate the predicted areas of mineralization, while the red points indicate the known mineralization for lead and molybdenum. There is a strong overlap of the known and predicted areas, and in addition to that there might be new

predicted areas worthwhile to be explored.

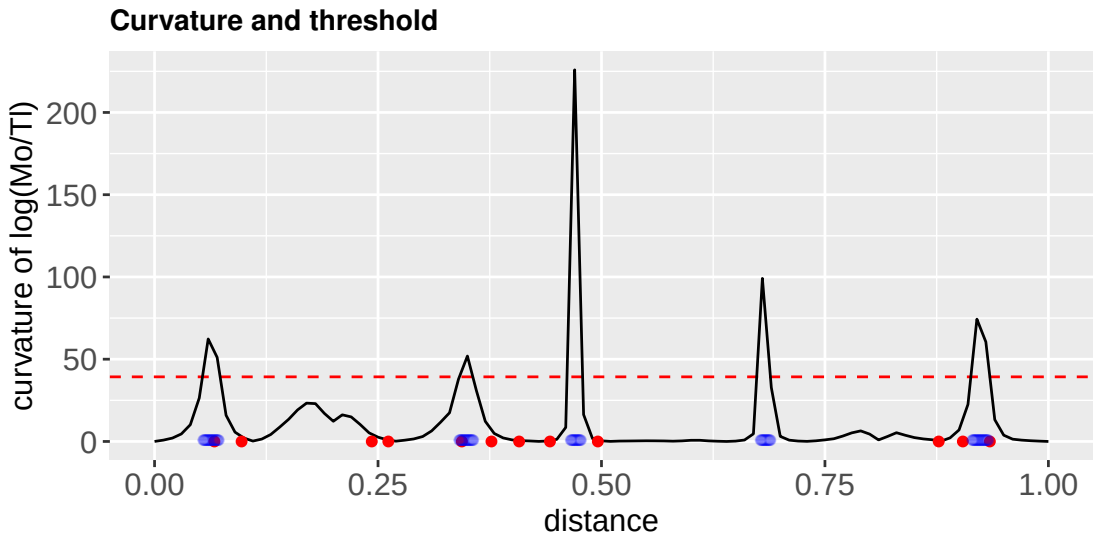


Figure 3.7: Curvature of log(Mo/Tl) in sample material Birch leaves, where known mineralization area (red points) and mineralized points identified by the method (blue points) are displayed.

Another example of an interesting log-ratio is displayed in Figure 3.8. This log-ratio of lead (Pb) versus aluminium (Al) is ranked as the second most important log-ratio (according to our c-value) in sample material BLE. In this figure we also show the original measured concentrations for Pb and Al, their GAM fits, and the locations of the known lead anomalies (red points). The log-ratio of the GAM fits clearly indicates the area around the known mineralization. The second known Pb mineralization around distance 0.9 is not indicated. This is because there is no increased measured Pb value around this distance, or the sampling survey has missed an appropriate measurement.

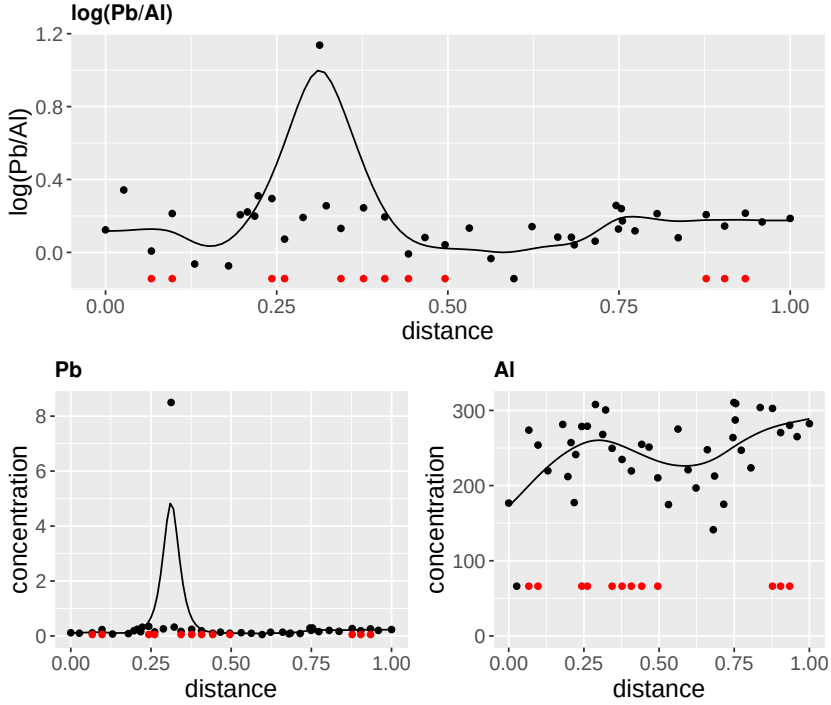


Figure 3.8: Upper part: Log-ratio of lead and aluminium constructed by using GAM fits of its individual elements – displayed on lower part of plot. Sample material is BLE. The red points indicate areas of known mineralization.

In order to stress the importance of the individual sample materials, Figure 3.9 displays the 70 top ranked c-values (unscaled) of these materials. Every line in the plot corresponds to a particular material, and the top-ranked c-values are connected by the line. The highest c-values are obtained for the O-horizon (OHO) samples. A quick decline of the curve means that the c-values of lower-ranked log-ratios are clearly smaller than for the top-ranked ones. One still needs to be careful with the interpretation, because high c-values can be obtained by few wide peaks in the log-ratios of the GAM fits, and not necessarily by several sharp peaks. In practice it will be definitely worthwhile to inspect the top-ranked log-ratios for several sample media, based on the information of Figure 3.9.

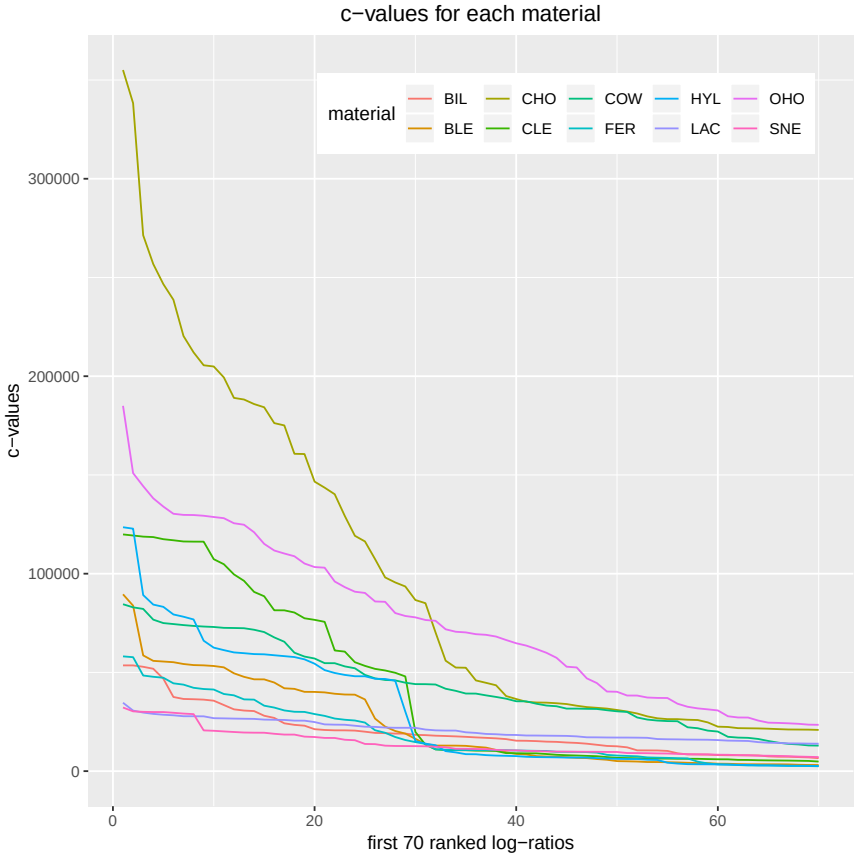


Figure 3.9: Top-ranked 70 (unscaled) c-values for each sample material. The horizontal axis represents the rank.

3.5 Discussion and conclusions

The identification of mineralization is usually based on a pre-study in the prospective area, where only few sample locations are considered. The sample locations are supposed to cross the potential mineralized zones, and thus the samples are frequently arranged on a linear transect. This is the setting which we considered in this paper.

Due to the compositional nature of geochemical data, log-ratios of the element concentrations are considered as informative. A further important property is the scale-invariance of log-ratios, which is very important when comparing log-ratios of different elements. However, analyzing log-ratios of the measurements of only few sample sample points may lead to a lot of uncertainty and instability. For this reason, the concentration data have been smoothed first using Generalized Additive Models (GAMs). The advantage of GAMs is that the smoothness can be tuned with a parameter, and the tuning parameter is selected using the underlying data (with cross-validation). Thus, the smoothing is adapted to the data, and once the smoothed signal is available, an arbitrary number of "concentration" values can be generated. Taking log-ratios of such generated values allows to compute the curvature, which involves the first and the second derivative, and these can be numerically obtained. Finally, a measure of "overall" curvature, which we called c-value, can be obtained. The c-value is not depending on the measurement units, and it can thus be compared for different element combinations, and even across different sample media. Moreover, due to the symmetry of log-ratios, an exchange of nominator and denominator would result in exactly the same c-value, which reduces the number of potential element combinations for the algorithm significantly.

In the experimental part we have demonstrated using two geochemical mineral exploration studies, that this methodology is indeed promising to identify pathfinder elements for mineralization. Those elements that appear in the top-ranked log-ratios (ranking according to the c-value) are considered as most informative. In addition, the inspection of the top-ranked log-ratios gives an indication of the location of the mineralization, and even of the extent of the mineralized areas. Furthermore, it is informative to inspect the magnitude of the top-ranked c-values for the different sample media in order to get an idea about their importance for the task of mineral exploration.

It is important to mention that the algorithm is unsupervised. This means that prior knowledge on the mineralized locations is not necessary. In our studies we only used this prior knowledge for the verification of the results.

In our future work we will extend this approach to the two-dimensional setting, i.e. where the samples are not taken along a linear transect but at locations in a two-

dimensional (irregular) grid. Although computationally more challenging, GAM fits can be extended to the two-dimensional case, but curvature also needs to be extended to this setting, together with an appropriate measure of overall curvature.

CHAPTER 4

A method to identify geochemical mineralization on linear transects

Mineral exploration in biogeochemistry is related to the detection of anomalies in soil, which is driven by many factors and thus a complex problem. Mikšová et al. (2019b) have introduced a method for the identification of spatial patterns with increased element concentrations in samples along a linear sampling transect. This procedure is based on fitting Generalized Additive Models (GAMs) to the concentration data, and computing a curvature measure from the pairwise log-ratios of these fits. The higher the curvature, the more likely one or both elements of the pair indicate local mineralization. This method is applied on two geochemical data sets which have been collected specifically for the purpose of mineral exploration. The aim is to test the technique for its ability to identify pathfinder elements to detect mineralized zones, and to verify whether the method can indicate which sampling material is best suited for this purpose.

4.1 Introduction

The identification of mineralized zones belongs to the important challenges in applied geochemistry. The difficulty is that the targeted mineralization could be of any arbitrary size, and in any depth, depending on the type of mineralization. The common procedure to discover mineralized zones is based on sampling, using strategic sampling designs in order to be as economic as possible. Samples can be taken from different soil layers, but also from different trees and plants around the presumed target. Since samples and their analysis of element concentrations is cost intensive, the sampling is often done on

linear transects, crossing the presumed mineralized zones. If there is more evidence, drilling is also used in order to obtain a depth profile of the element concentrations. More information on different sampling strategies can be found in Mikšová et al. (2019a).

In this work we assume that the sampling has been carried out on a linear transect, or that the available samples can be aggregated to such linear transects. This means that the spatial locations can be considered along a line, and thus it is simple to graphically investigate the spatial variability of the measured elements by simply plotting the element concentrations against the locations (Torppa and Middleton, 2017). However, in modern geochemistry, the number of elements that can be reliably measured is in the range of 30-60, and if the samples have been obtained from several different sampling media, it is a challenging task to study all resulting plots for abrupt changes in the concentration values. Such changes could indicate mineralized zones, since their signals could lead to sudden increases of element concentrations. There is, however, the problem that due to the (economic) sampling procedure, only very few samples might have been taken on top of the mineralization, and together with measurement and analysis uncertainties, the resulting concentration changes might not be clearly expressed. The second problem is that there is an interplay of the concentration values among the elements, because geochemical data are compositional by their nature (Aitchison, 1986; Filzmoser et al., 2018).

Consider a composition $x_1^m, \ldots, x_{D_m}^m$, consisting of D_m chemical elements, measured in $m = 1, \ldots, M$ different sample materials. For the analysis of compositional data it has become popular to use the so-called log-ratio methodology, introduced by Aitchison (1986). This refers to the use of logarithms of ratios, and the basic information are log-ratio pairs $\ln(x_j^m/x_l^m)$, for $j, l \in \{1, \ldots, D_m\}$. The use of log-ratios eventually leads to a sound geometrical concept, referred to as the Aitchison geometry (Pawlowsky-Glahn et al., 2015). There are, however, also practical reasons why log-ratios are useful, such as symmetry around zero, and equal variance if numerator and denominator are exchanged.

A further argument for considering (log-)ratios is the assumption that there could be elements which are stable and thus not affected by a mineralization, and others are very indicative of mineralized zones. The log-ratio of such elements could even better express the local change around a mineralization, because measurement and analysis uncertainties could cancel each other out (Mikšová et al., 2019b).

On the other hand, D_m different elements would lead to $D_m(D_m - 1)/2$ different (and relevant) pairwise log-ratios, which makes a visual inspection practically impossible. For this reason, Mikšová et al. (2019b) have introduced a procedure to rank the list of log-ratio pairs according to their ability to indicate mineralization. This is done by first

approximating the individual element concentration by a smooth fit, taking log-ratios of the smooth fits, and computing a measure of curvature. The higher the curvature, the more likely (at least) one of the log-ratio pair elements shows sudden changes. In addition, the visualization of the smooth fits and their log-ratio allows to localize the presumed mineralized zones.

In this paper we briefly review the method of Mikšová et al. (2019b). Then we apply this procedure to two geochemical data sets, originating from surveys carried out in Greenland and France, respectively, in the frame of the ongoing project "UpDeep" (UpDeep, 2017–2020), which aims at developing and implementing a methodology to identify mineralization.

4.2 Methodology

As already indicated in the introduction, the main idea of the methodology developed in Mikšová et al. (2019b) is that at the beginning and end of a transect crossing a potential mineralization, important log-ratios of an element pair display a very quick spatial change which can be captured by a measure based on the curvature of the latter.

The first step of the methodology consists in fitting a so called GAM model, see Wood (2017), to each element, with concentration values y_i at locations x_i, for $i = 1, \ldots, n$. After considering the nature of our data and after inspection of the corresponding residual plots we decided to model the data belonging to the Tweedie family with a log-link and additional weights. Modelling the element concentrations in such a way means that for each element the following optimization problem based on the log-likelihood function l is solved to obtain a linear predictor η

$$\hat{\eta} = \arg\max_{\eta \in \mathcal{H}} \sum_{i=1}^{n} \omega_i l(y_i | x_i; \eta) - \lambda \int (\eta''(x))^2 dx,$$

with predefined weights ω_i, upweighting certain points, a suitable function space $\mathcal{H}$ and a smoothing parameter λ.

This results in GAM fits $\hat{f}_{el_1}$ and $\hat{f}_{el_2}$ for each pair of elements el_1 and el_2, and the log-ratio of the fits for any location x along the transect can be obtained subsequently as

$$\begin{aligned} g(x) :&= \log\left(\frac{\hat{f}_{el_1}(x)}{\hat{f}_{el_2}(x)}\right) \\ &= \log\left(\hat{f}_{el_1}(x)\right) - \log\left(\hat{f}_{el_2}(x)\right) \\ &= \log\left(h^{-1}(\hat{\eta}_{el_1}(x))\right) - \log\left(h^{-1}(\hat{\eta}_{el_2}(x))\right), \end{aligned}$$

where $h(\cdot)$ stands for link function.

Its curvature is then computed by

$$\kappa(x) := \frac{|kg''(x)|}{(1 + (kg'(x))^2)^{\frac{3}{2}}},$$

where k is a scaling factor allowing the curvatures to be comparable amongst different pairs.

Finally, for each pair of log-ratios, the following quantity is introduced to measure quantitatively important spatial changes potentially indicating the beginning and the end of a mineralization, namely:

$$c(el_1, el_2) := \frac{2}{L} \sum_{l=1}^{\frac{L}{2}} \max_{x \in [x_{j_{2l-1}}, x_{j_{2l}}]} (\kappa(x) - \mathcal{T})_+^2. \tag{4.1}$$

This measure is denoted as the c-value in the following. Here, $(\cdot)_+$ denotes $\max(\cdot, 0)$, and $\mathcal{T}$ is a threshold, L is the number of times that the curvature κ crosses the threshold, and $[x_{j_{2l-1}}, x_{j_{2l}}]$ are the corresponding points where this happens. It is easy to see that only points x for which the curvature is above the threshold are influencing this measure $c(el_1, el_2)$. This avoids any influence of small values of $\kappa(x)$, meaning that only very high signal changes of the log-ratio are taken into account. Summing up over all maximum leads to a quantity measuring the mean number of high signal changes.

For a more detailed description of the weights ω_i, the smoothing parameter λ, the scaling factor k, the threshold $\mathcal{T}$, and the numerical computation of the derivative, as well as the measure $c(el_1, el_2)$ we refer to Mikšová et al. (2019b).

Since we are dealing with compositional data, one could argue that not the absolute element concentrations should be used for the GAM fits, but rather the log-ratios of all pairs of elements. Although this could be a reasonable approach, there are several arguments against this idea: (a) The GAM fits may require some manual adjustment and tuning, which is not feasible for all pairwise log-ratios. (b) Typically, the number of observations is rather low, and there could be some data quality issues as well. GAM fits on the raw data could, to some extent, "repair" this effect, particularly if there is uncertainty in small concentrations, and ideally the data quality after the log-ratios of the GAM fits increases.

4.3 Results

One important part of the UpDeep project has been to take samples in two countries, namely in Greenland and France. This sampling procedure was successfully accomplished

by the sample providers GEUS (Geological Survey of Denmark and Greenland) and BRGM (The French Geological Survey), respectively. The sampling was performed by executing geochemical sampling surveys according to the established protocols in geologically well-known mineralized areas. The samples in both countries were taken in the years 2017 and 2018, however in this context we focus on one specific year 2018.

4.3.1 GEUS data

The sampling areas were chosen due to known mineralization and exploration in the area. The interest is in the area Isortoq, which is situated in the very south of Greenland, see Figure 4.1 for a detailed map, where the map background is obtained using Google maps (Kahle and Wickham, 2013). In total, three traverses were sampled which are 300 meters apart. The samples from the different traverses are shown in different color in the map. Green color refers to the locations of known mineralization. In this case the deposit is an iron (Fe), vanadium (V), titanium (Ti) deposit. A possible proxy for V could be scandium (Sc) (since V tends to be analyzed poorly). In our analyses we merge the samples from the three traverses into one linear transect, which means that all samples have been taken, but their locations are set to a linear transect in the center of the three traverses. The individual samples are now at a distance between 50 to 400 meters. The total length of the transect is about 12 km.

Two different plant species and soil samples have been investigated, namely *Salix Glauca* and *Empetrum Nigrum* with 49 samples, and soil comprises 47 observations containing only so called routine samples.

Following the procedure of Section 2, Figure 4.2 shows the log-ratio pair of the GAM fits of the elements Ti and Ca (calcium) measured in soil, and Figure 4.3 displays the resulting log-ratio for Fe and P (phosphorus) in soil. Both log-ratios yield top-ranked c-values, see Equation (4.1). In these plots, the mineralized zones are shown by red points. The blue points are the predicted mineralized zones, when the curvature exceeds the threshold $\mathcal{T}$, which is indicated by the horizontal dashed line. The predictions confirm the presumed mineralized zones very well, and they do not indicate new mineralized areas.

A useful tool to display the overall information about meaningful log-ratios is the heatmap. The input for the heatmap is a matrix of c-values, computed from the GAM fits of all pairwise log-ratios of a specific sampling material (plants or soil). Figure 4.4 shows the resulting heatmap for *Salix Glauca* (left), *Empetrum Nigrum* (right), and soil (bottom). Obviously, the heatmaps are symmetric due to the symmetry of the log-ratios. The darker the blue color, the higher is the c-value obtained from the corresponding

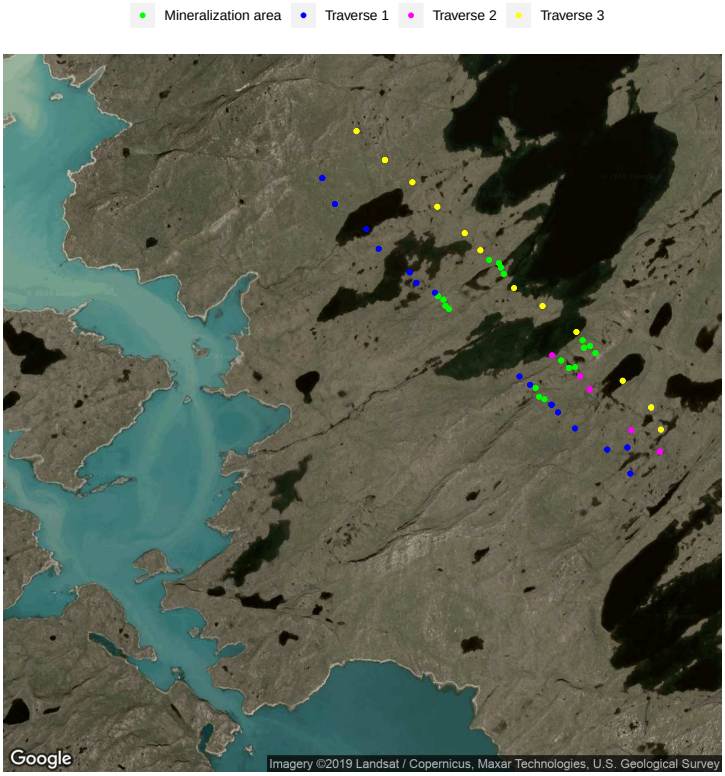


Figure 4.1: Map of the locations of the samples taken by GEUS in the Isortoq South Area.

log-ratio. A dark blue row or column in the heatmap indicates so-called pathfinder elements, which potentially refer to mineralization. From a geochemical point of view, most of the elements with higher c-values are related to the deposit (Fe-V-Ti).

These heatmaps can also be used to identify potentially interesting pathfinder elements that could indicate new mineralized zones. For example, the heatmap for soil (bottom plot in Figure 4.4) shows a high c-value for the pair Na (sodium) and Pb (lead). Figure 4.5 presents the corresponding curvature plot of this log-ratio. Indeed, there are several locations where abrupt signal changes are visible. One would have to further explore these locations.

4.3.2 BRGM data

The second data set originates from the Vendée area in middle-west France, which has been sampled in 2018. The area was investigated because of some historical knowledge of the occurrence of rare elements. Moreover, an easy access allowed for a valuable

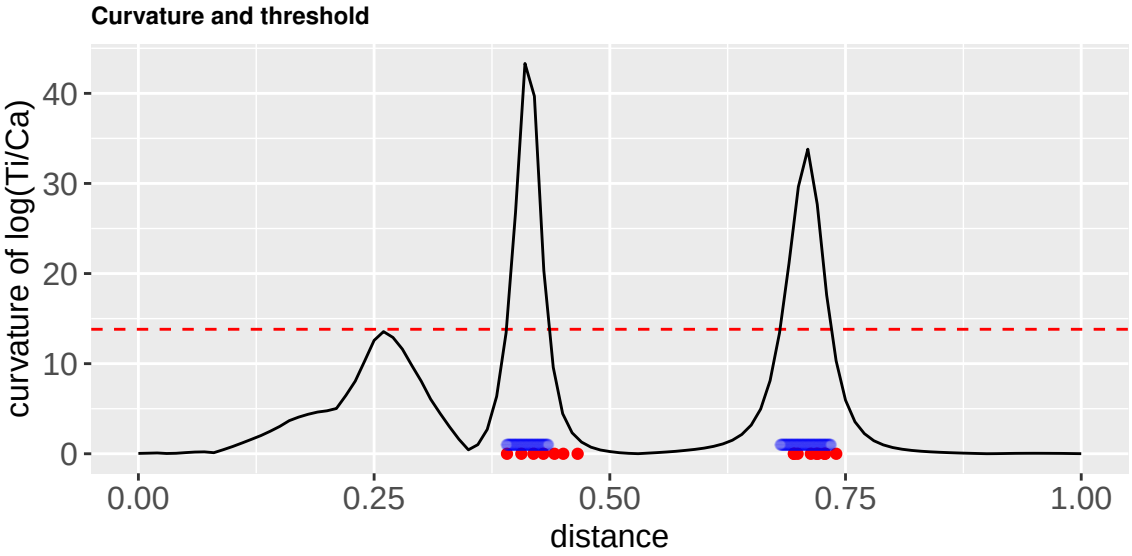


Figure 4.2: Curvature of the log-ratio of the GAM fits of Ti and Ca in soil.

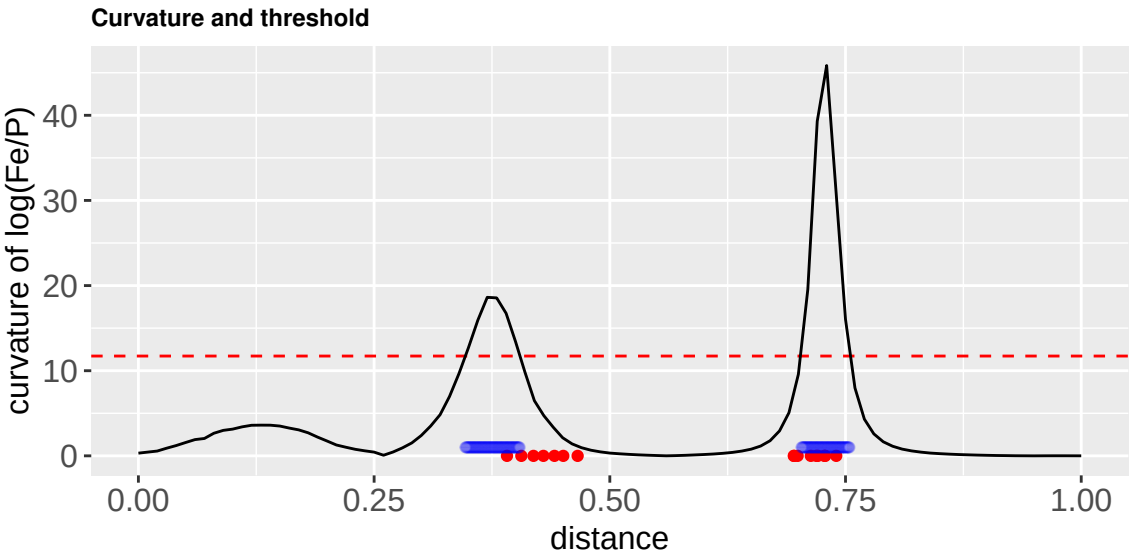
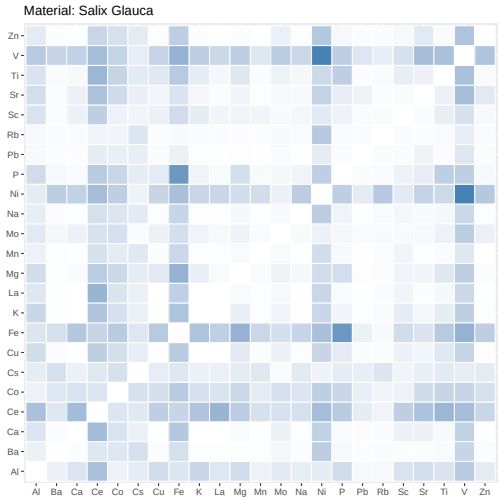
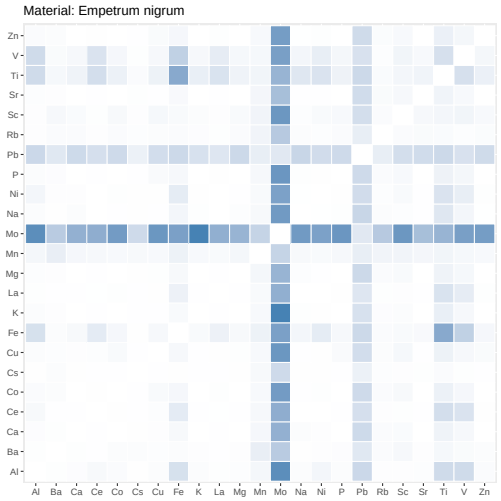


Figure 4.3: Curvature of the log-ratio of the GAM fits of Fe (iron) and P (phosphorus) in soil.

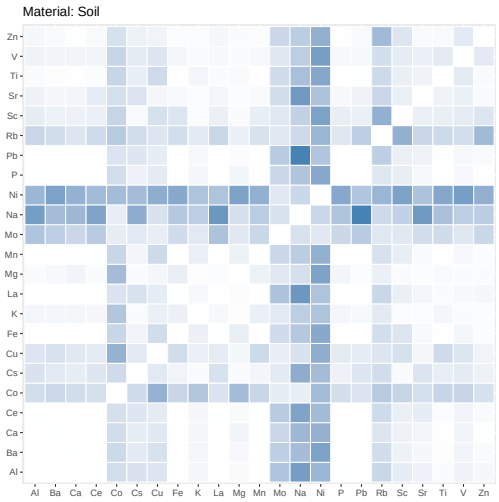
recognition of the area prior to sampling. Figure 4.6 shows a satellite map of the area where the samples have been taken from three different sites. Each of these subareas



(a) *Salix Glauca*



(b) *Empetrum Nigrum*



(c) Soil

Figure 4.4: Heatmaps of the c-values for the plant materials and soil.

contains two traverses which are again merged to one transect in our procedure in order to increase the number of observations per site. The first site in the south-west of Figure 4.6 holds approximately 30 samples, the second (middle) site about 40, and the third (north-eastern) site only 18 samples. The presumed mineralization type on all sites is antimony (Sb) and gold (Au). Due to pre-studies it turned out that the second site has the highest concentrations of Sb. The element Au is in any case difficult to measure.

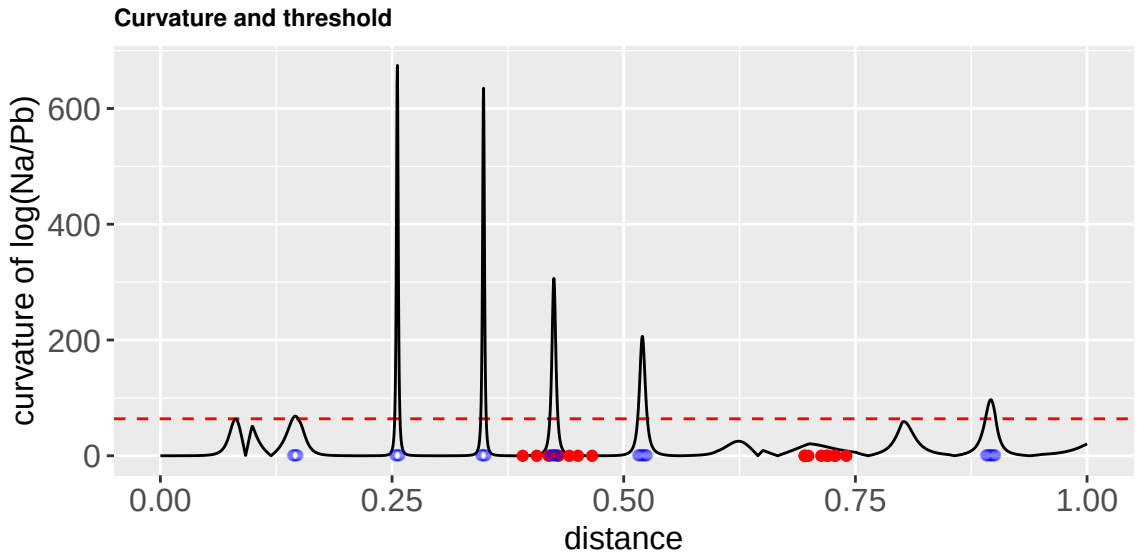


Figure 4.5: Curvature of the log-ratio of the GAM fits of Na (sodium) and Pb (lead) in soil.

This data set provides in total 6 different sample materials, namely Ah horizon with Aqua Regia leach (AhAQ), Ah with deionized water leach (AhL1), Ah with sodium pyrophosphate leach (AhL3), Bramble branch (BB), Bramble leaves (BL), and Oak bark (OB). Rather than investigating again the curvature plots, we focus now on the task to identify the most promising sample material indicating mineralization. An answer would be highly relevant, because sampling of the different materials is very time- and cost-intensive.

Figure 4.7 presents for each sample site the top-ranked 70 c-values from all pairwise log-ratios of the GAM fits, separated by sample material. Since the log-ratios of the fitted values are scaled to the interval $[0, 1]$, the c-values are comparable, regardless of sample site and sample material. We obtain the highest c-values for Site 2, which is the most reliable sample site due to the higher number of observations. The plot for Site 2 reveals a clear difference in the top ranked c-values for the mineralization, while the soils seem to be highly informative. All sites show that sample material OB performs worst in terms of the c-values, and thus this is the least interesting sample material.

The heatmaps in Figure 4.8 confirm our findings. The left plot for the soil material AhAQ identifies Sb (and to a lesser extent Zn) as important pathfinder element of mineralization. The right plot for plant BB uses the same color scheme, but represents

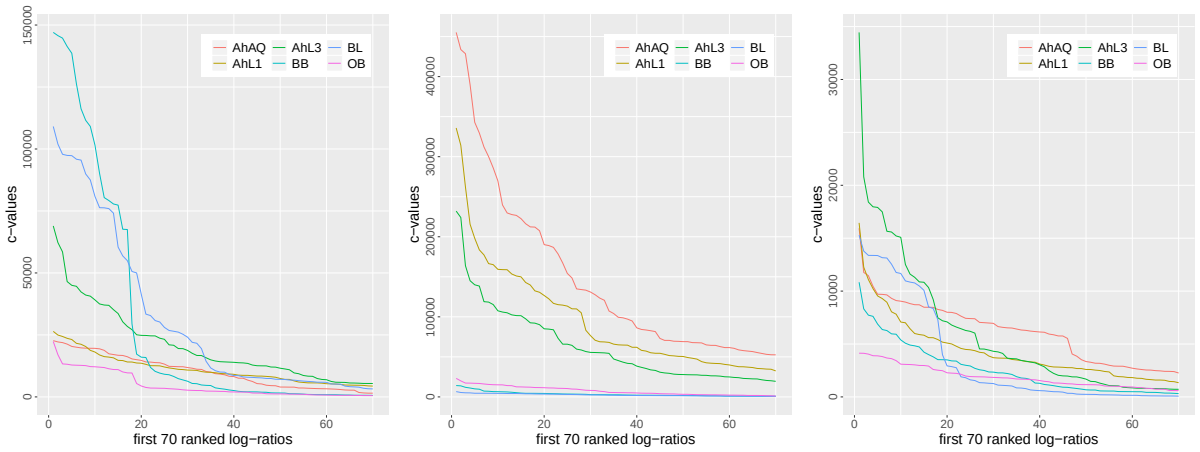
Figure 4.6: Map with the sample locations taken by BRGM in the Vendée area in 2018

much lower c-values (see Figure 4.7, middle). This heatmap shows a rather inhomogeneous structure and thus no clear pathfinder elements.

4.4 Summary

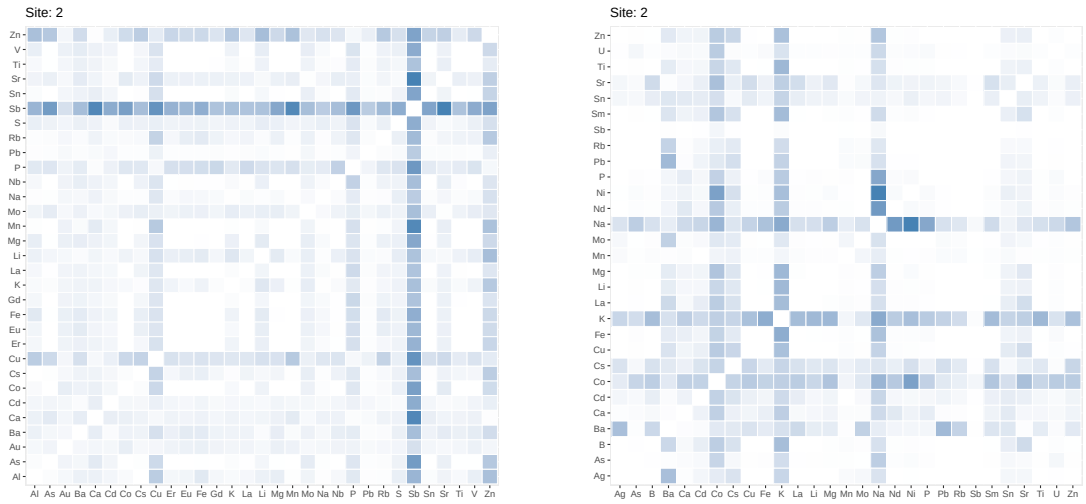
Due to the technological developments, mineral exploration nowadays belongs to the most important tasks in geochemistry. Although many chemical elements can be investigated for their concentration in different sample materials, sampling is still time- and cost-intensive, and this is the reason why usually only 20-60 samples are available at a potentially mineralized zone. The common strategy is to position the samples on (a) linear transect(s), crossing the mineralized zones, and mineralization would then appear in terms of increased element concentrations.

Rather than investigating single element concentrations, Mikšová et al. (2019b) have developed a method based on considering log-ratios of all pairs of elements. Since the number of possible pairs increases quickly with the number of investigated chemical



(a) Site 1 (b) Site 2 (c) Site 3

Figure 4.7: Top-ranked 70 c-values, computed for 6 different sample materials and the three different sample sites.



(a) Ah horizon with Aqua Regia leach (b) Bramble branch

Figure 4.8: Heatmaps of the c-values for the BRGM data - soil and plant material.

elements, a strategy has been proposed to rank the element pairs according to their relevance for mineral exploration. This strategy uses a measure of curvature for log-ratios of smooth fits of the concentration values. The resulting c-values are normalized and can be compared across different pairs, and even across different sample materials and sites.

In this paper we have demonstrated the usefulness of this procedure based on two

data sets that have been collected specifically for the purpose of mineral exploration. For the first data set originating from Greenland it has been shown that the c-values indeed identify important pathfinder elements to confirm presumed mineralized zones, but they seem also promising to point at new locations with potential mineralization. The second data set from France was employed to investigate which sample material is most promising to detect mineralization. It turned out that the soil samples are much more informative than the plant samples, but this may again depend on the type of mineralization, and probably even on further factors.

In our future work we will extend the methodology of Mikšová et al. (2019b) to the case where the samples are not necessarily taken along a linear transect, but on a sample grid with different x- and y-coordinates. This means that the smooth fits as well as the curvature measure need to be extended to the two-dimensional case.

CHAPTER 5

Identification of mineralization in geochemistry for grid sampling using Generalized Additive Models

The important goals of mineral exploration geochemistry are detection and identification of underlying mineralization. This paper deals with element concentration data analyzed of surface geochemical samples acquired from soil horizons or plants. A new unsupervised procedure is proposed for this purpose when the samples have been taken on a regular or irregular grid in the area under investigation. The methodology is based on Generalized Additive Model (GAM) fits on absolute concentration data. Then new data points are taken of the surface of the smooth fits across the entire sampling area as a regular grid. Pairwise log-ratios of elements are then calculated of these points, and curvature of the log-ratio pairs is computed. High curvature indicates abrupt spatial changes, which could point at locations of mineralized zones. A measure called c-value evaluates the overall curvature and thus serves as an importance measure of the log-ratio pair. The methodology is tested on two real surface geochemical data sets collected in areas with known underlying mineralization, and the results confirm existing pre-knowledge.

5.1 Introduction

Mineral exploration with surface geochemical data, including soil partial extraction and biogeochemical datasets, is challenging because an underlying significant mineralization may be observed in absolute concentration data as a multitude of different spatial anomaly patterns with a variety of characteristics (e.g., Dunn, 2007; Torppa and Middleton, 2017; Taivalkoski et al., 2019). Traditionally, occurrence and type of mineralization is determined as the presence of an anomaly in a single element or multi-variable element data. Spatial anomaly patterns appear in a variety of shapes such as symmetric or skewed, narrow single point or narrow or wide multi point positive or negative apical anomalies. They may also be so-called 'rabbit ear' anomalies located at the flanks of the mineralization, or alteration indicating halos around the mineralization itself. In addition, data may be noisy and significant anomalies may be weakly contrasting. In many cases, anomalies are captured in the data collection phase only from a few sampling stations because studies are conducted as cost-effectively as possible. One then deals just with very few samples, typically tens to hundreds. From a practical point of view, it would be desirable to have at least two sampling stations on top of an underlying mineralization forming an anomaly pattern. However, this requirement is not always easy to meet because pre-knowledge of an underlying deposit type can usually be based on knowledge from other locations or geophysical measurements. Furthermore, anomaly formation is influenced by many factors including target characteristics such as type and depth of mineralization. Anomalies in these dataset are not purely related to underlying exogenic signals from the mineralization but may also be resulting from lithogeochemical variation, bedrock fracturing, land use, anthropogenic contamination, environmental factors, characteristics of the sample media, quality of measured chemical concentrations in the laboratory etc. These factors makes a study for detecting and identifying potentially mineralized zones from surface geochemical data complex.

Significant anomaly patterns in these datasets however, are mostly local with low spatial dispersion. Additionally, they may retain information of presence of deep lying mineralizations which could also be buried under sediment cover (Cheng, 2012). The confidence in detecting and identifying an underlying mineralization is high when the anomaly patterns are detected with many coinciding or zonal elemental patterns including commodity, pathfinder and alteration elements and the anomaly to background contrast is high. Often pathfinder elements, which may not commonly be associated with the mineralization type, form more significant anomalies than the commodity elements. For example, plants may control the uptake of essential elements, thus unexpected pathfinder

elements form higher contrasting anomalies to background than commodity elements do. Thus, it is essential to analyze a wide range of elements and pay attention to spatial patterns to all of them.

In practice, two strategies for the sampling design are commonly used: sampling along a linear transect, and sampling on a grid (Webster and Lark, 2012). The focus in this paper is on mineral exploration when the samples have been taken on a grid on a plane to be plotted as a map. There is a lot of literature available tackling this problem. Fractal and multi-fractal models are considered (Dahooei et al., 2016), and machine learning methods are frequently used. For a recent review of different approaches, see Zuo (2017). Some of the methods are limited to identifying specific types of mineralization, and some try to incorporate additional information such as geophysics data as prospectivity models (Darabi golestan et al., 2013). There are also supervised techniques which need prior information about mineralized areas in order to find other mineralized zones (Roshani et al., 2013).

A different approach was used in Mikšová et al. (2019b) regarding the compositional nature of these data. The method was specifically developed for geochemical samples taken along a linear transect and for cases when no knowledge of an existing mineralization is available. As a first step, a set of chemical elements containing information of a potential mineralization are determined. This is based on expert knowledge but also on quality assurance and quality control of the data in hand. Since the number of samples is usually low, and variability of the concentration values caused by different kinds of uncertainties exists, these concentration values are firstly smoothed along the spatial dimension. For this purpose, Generalized Additive Models (GAMs) are used. Once all concentration values are smoothed, a much larger number of "artificial" samples can be generated by using the fitted values on a regular grid. Now pairwise log-ratios of the smooth fits are computed for all element pairs, and a curvature measure is derived. High curvature corresponds to strong changes in the signal, which is hypothesized to indicate presence of an underlying mineralization. Possibly, high curvature may also indicate other phenomena, not only the mineralization. This method allows for a ranking of the element pairs, and also a ranking of the sample media for their potential to indicate mineralization.

This work extends the method of Mikšová et al. (2019b) to grid sampling. For many studies, the grid sampling strategy might be more useful than line sampling, since the data may better capture the two-dimensional signal including shape, orientation and size of an underlying mineralized load in a spatial coordinate space. Again, all concentration values of the different elements are smoothed, but the smooth fit needs to be bivariate,

over the sample locations in the two-dimensional grid. This leads to several computational and conceptual challenges. Also a modified measure of curvature has to be developed which takes the log-ratios of the GAM smoothed fits as inputs. Pairwise log-ratios are derived on a regular grid with higher resolution than the original sampling density. As a result, the curvature measure should indicate promising element pairs which point towards the spatial locations of mineralized zones. Moreover, a comparison of different sample media for their effectiveness for mineral identification is obtained.

The paper is structured as follows. In Sect. 2, the concept of the methodology is introduced. A detailed algorithm for the whole procedure of ranking log-ratio pairs is proposed in Sect. 3. This procedure is applied in Sect. 4 to two geochemical data sets, originating from plant and soil sampling surveys carried out in Finland and Greenland. The final Sect. 5 summarizes and concludes.

5.2 Methodology

The main idea in Mikšová et al. (2019b) was to take the curvature of log-ratios as a basis for constructing a measure which helps to identify different kinds of mineralization. The interesting pairs of log-ratios involved in mineralization should display a sharp spatial change on top of or in immediate surroundings of an underlying mineralization rather than exhibiting a flat behavior throughout the area. However, in two dimensions there are multiple definitions of curvature. The most prominent ones are mean and Gaussian curvature (Goldman, 2005).

At first, as in the one-dimensional case, before looking at the curvature of log-ratios, GAMs are fitted to the concentration data (z_i^{el}, x_i, y_i), where z_i^{el} is the concentration of a fixed element el at location (x_i, y_i) and $i = 1, ..., n$, for all elements. The motivations for using GAMs are manifold. Firstly, the goal is to get rid of any noise and excess curvature in the concentration signal whilst only retaining high curvature to get a good fit. Secondly, data is typically scarce and non-uniformly sampled over an area. That excludes other smoothing methods such as LOESS (Cleveland et al., 1992). Thirdly, concentrations are a positive quantity, and thus the smoothed signal should also have positive values only. At last, the method presented in this paper requires partial derivatives up to the second order, which again excludes many other smoothing techniques. Considering all these requirements, GAMs seem to be a very natural choice.

In one dimension, smoothing splines are widely used for modeling the linear predictor in GAMs. They are motivated by finding the solution of minimal curvature. However, instead of curvature usually only the second derivative is considered as an approximation,

because this makes the optimization problem easier. In the two-dimensional case there is a multitude of smoothing spline methods, and the generalization is not so clear. It is known from differential geometry, especially minimal surface theory that, under certain circumstances, for a function $\eta : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3$ satisfying

$$\frac{\partial}{\partial x}\left(\frac{\eta_x}{\sqrt{1+\eta_x^2+\eta_y^2}}\right)+\frac{\partial}{\partial y}\left(\frac{\eta_y}{\sqrt{1+\eta_x^2+\eta_y^2}}\right)=0, \tag{5.1}$$

where η_x and η_y denote the partial derivatives, its solution has mean curvature of zero, see Fomenko and Tuzhilin (2005). Equivalently, this means that Eq. (5.1) describes a surface locally having a minimal area. Such a property seems to be adequate to exclude fits with any excess curvature as it was presented in the one-dimensional case (Mikšová et al., 2019b). The latter ideas lead to so-called soap film smoothers, derived in Wood et al. (2008) and also allowing complex areas Ω – which might be the case for geographical areas represented by irregular grids.
Thus, as a first step, a soap film smoother is fit to the concentrations of each element el, meaning that the following problem is solved,

$$\hat{\eta}_{el} = \arg\max_{\eta\in\mathcal{H}} \sum_{i=1}^{n} w_i^{el} l(z_i^{el}|x_i, y_i; \eta) - \lambda \int_\Omega \left(\frac{\partial^2}{\partial x^2}\eta(x,y)+\frac{\partial^2}{\partial y^2}\eta(x,y)\right)^2 dxdy, \tag{5.2}$$

where $\mathcal{H}$ is the function space of sufficiently smooth functions such that the penalty of η exists, $\lambda > 0$ is the so-called smoothing parameter, l is an appropriately chosen log-likelihood function, w_i^{el} are predefined weights, and Ω is a user chosen area for the fit. Additionally, one also chooses a link function h such that the fitted concentration for a fixed element is obtained by $\hat{f}_{el}(x,y) = h(\hat{\eta}_{el}(x,y))$. An appropriate choice of w_i^{el}, l and h for modeling concentrations in the framework of mineralization detection for line sampling is discussed in Mikšová et al. (2019b), but it is also case dependent. For a more thorough introduction to soap film smoothers, see Wood et al. (2008) and Wood (2017).

After producing the GAM fits to the absolute concentration data of all elements and calculating the log-ratio pairs, curvature should be calculated at each point (x_0, y_0) – just as it was done for the one-dimensional case in Mikšová et al. (2019b). However, for surfaces the notion of curvature is different from that of a curve, because there are multiple directions to go at each point of the surface. Basically, in the case of curves, one can define curvature as the length of the vector obtained by differentiating the unit normal to the curve with respect to the arc length. For surfaces, the same can be done with respect to any element in the tangent space. More precisely, let $\mathcal{S}$ denote the surface induced by the function $g : \Omega \mapsto \mathbb{R}^3$, thus $g(\Omega)$, and $\mathbf{n} : \mathcal{S} \mapsto \mathbb{R}^{\mathbf{3}}$ its unit

normal field, meaning that $\mathbf{n}$ is of unit length and is orthogonal to any tangent vector to the surface, at each point $s \in \mathcal{S}$, as well as sufficiently smooth. For any smooth curve $\gamma : (-\epsilon, \epsilon) \to \Omega \subset \mathbb{R}^2$, with $\gamma(0) = (x_0, y_0) \in \Omega$, the derivative of $g(\gamma(t))$ in $t = 0$ defines an element in the tangent space at the point $g(x_0, y_0) \in \mathcal{S}$. Thus, if $\alpha(t) := g(\gamma(t))$ is parameterized by arc length, meaning that the norm of its derivative is one at each point, its curvature in t is obtained as $\|\ddot{\alpha}(t)\|$, where one dot over α denotes the derivative in t. As shown in Do Carmo (2016), this leads to

$$-\left(\frac{d}{dt}\mathbf{n}(\alpha(\mathbf{t}))\right)' \cdot \dot{\alpha}(\mathbf{t}) = (\mathbf{n}(\alpha(\mathbf{t})))' \cdot \left(\ddot{\alpha}(\mathbf{t})\right). \tag{5.3}$$

As the norm of $\ddot{\alpha}(t)$ is the curvature of the curve α, this is valid because $\mathbf{n}(\alpha(\mathbf{t}))$ is in the same – or exactly opposite – direction as $\ddot{\alpha}(t)$. The left side of Eq. (5.3), called the second fundamental form, can be seen to contain information of how the surface is curved at a point $g(x_0, y_0)$ for any taken curve γ and therefore $g(\gamma)$. It turns out that Eq. (5.3) can be expressed as a bilinear form in $\dot{\gamma}(0)$, see Do Carmo (2016), namely as

$$(\dot{\gamma}(0))'\left(\frac{1}{\sqrt{1 + (g_x)^2 + (g_y)^2}} \begin{bmatrix} g_{xx} & g_{xy} \\ g_{xy} & g_{yy} \end{bmatrix}\right)\dot{\gamma}(0). \tag{5.4}$$

The mean curvature is then defined as the mean of the two eigenvalues of the matrix in Eq. (5.4). However, here the following term

$$\kappa(x, y) := \frac{1}{2}(|k_1(x, y)| + |k_2(x, y)|) \tag{5.5}$$

is used to construct a measure, where k_1 and k_2 denote the eigenvalues at (x, y). The reason for this is that the mean curvature can become zero when the two eigenvalues cancel each other out. Since the goal is to detect any kind of spatial change, the absolute values of the latter are taken to avoid this issue.

Once that κ is obtained for the surface defined by the fitted log-ratio $g := \log(\frac{\hat{f}_{el_1}}{\hat{f}_{el_2}})$ for each fixed pair of elements el_1 and el_2, we define

$$\mu = \frac{1}{|J|}\sum_{j \in J} \kappa(x_j, y_j) \tag{5.6}$$

$$\sigma^2 = \frac{1}{|J| - 1}\sum_{j \in J} (\kappa(x_j, y_j) - \mu)^2, \tag{5.7}$$

denoting J the index set of a fine enough mesh of Ω. These two measures are approximations to $\int_\Omega \kappa(x, y)dxdy$ and $\int_\Omega (\kappa(x, y) - \mu)^2 dxdy$. Similar to Mikšová et al. (2019b), the

total curvature measure, in the following called c-value, for a pair of elements is defined as

$$c(el_1, el_2) := \frac{1}{|J|} \sum_{j \in J} (\kappa(x_j, y_j) - (\mu + \sigma))_+, \tag{5.8}$$

where $(\cdot)_+$ denotes $\max(0, \cdot)$. Similarly to the univariate case, this measure is high when κ is high. This means that points at which the signal of log-ratios changes quickly influence this measure a lot, and small values of κ have no influence at all. It thus serves to identify pairs of log-ratios which on average have a lot of quick spatial changes over Ω, therefore identifying interesting pairs.

5.3 Algorithm

This section provides the full algorithm to obtain all the c-values for each pair of elements el_1 and el_2. As an input, the algorithm takes the boundary $\partial\Omega$ of a set Ω, and the measured concentrations z_i^{el} at (x_i, y_i) for each element. Also a mesh χ of Ω is provided. Such a mesh can be obtained by dividing the range of x_i and of y_i, thus $[\min(x_i), \max(x_i)]$ and $[\min(y_i), \max(y_i)]$, for $i = 1, \ldots, n$, into M equidistant parts and only keeping the grid points contained in Ω. The corresponding index set is denoted by J.

Step 1: As first step, weights are calculated. The following weights for each element el are used for the specific examples below,

$$w_i^{el} := \begin{cases} 2\frac{z_i - \mu^{el}}{\sigma^{el}} & \text{if } \frac{z_i - \mu^{el}}{\sigma^{el}} \geq \frac{1}{2} \\ 1 & \text{else} \end{cases}$$

where μ^{el} and σ^{el} are the mean and the standard deviation of the measured concentrations z_i^{el}.

Step 2: Soap film smoothers are fit to each element, i.e. the problem stated in Eq. (5.2) is solved. We decided to model the data using the Tweedie-family as it contains a big range of probability distributions for positive responses. For this purpose, the R package `mgcv` (Wood, 2012) is used. The smoothing parameter is chosen according to the REML criterion.

Step 3: Following Step 2, all the log-ratios are now given by $g := \log\left(\frac{\hat{f}_{el_1}}{\hat{f}_{el_2}}\right)$. To make things comparable for each element pair, g is scaled by a constant $\Gamma := |\max_{(x,y)\in\Omega} g(x, y) -$

$\min_{(x,y)\in\Omega} g(x,y)|^{-1}$ if g is not constant, and one otherwise. In the following, g is replaced by $\Gamma^{-1}g$.

Step 4: By choosing $\epsilon = 10^{-2}$, for example, an approximation to all the partial derivatives in Eq. (5.4) and implicitly also in Eq. (5.5) is computed at the points given by the mesh χ of Ω, thus (x_j, y_j) with $j \in J$. As each log-ratio is the difference of individual logarithms of an element $\Gamma^{-1}\log(\hat{f}_{el})$, only the latter needs to be considered to obtain all the partial derivatives involved in a pair. Thus, for example, the first and second partial derivatives in x and y are given by

$$\begin{aligned}
\left(\Gamma^{-1}\log(\hat{f}_{el})\right)_x &= \Gamma^{-1}\frac{(\hat{f}_{el})_x}{\hat{f}_{el}} \\
\left(\Gamma^{-1}\log(\hat{f}_{el})\right)_{xx} &= \Gamma^{-1}\left(\frac{(\hat{f}_{el})_{xx}}{\hat{f}_{el}} - \left(\frac{\hat{f}_{el_x}}{\hat{f}_{el}}\right)^2\right) \\
\left(\Gamma^{-1}\log(\hat{f}_{el})\right)_{xy} &= \Gamma^{-1}\left(\frac{(\hat{f}_{el})_{xy}}{\hat{f}_{el}} - \frac{\hat{f}_{el_x}\hat{f}_{el_y}}{\hat{f}_{el}^2}\right),
\end{aligned}$$

where the two covariates can be interchanged to obtain the remaining partial derivatives. Each partial derivative of $\hat{f}_{el}$ is approximated in a finite difference way, meaning that at a point (x_j, y_j) for $j \in J$

$$\begin{aligned}
(\hat{f}_{el})_x &\approx \frac{\hat{f}_{el}(x_j+\epsilon, y_j) - \hat{f}_{el}(x_j-\epsilon, y_j)}{\epsilon} \\
(\hat{f}_{el})_{xx} &\approx \frac{\hat{f}_{el}(x_j+\epsilon, y_j) - 2\hat{f}_{el}(x_j, y_j) + \hat{f}_{el}(x_j-\epsilon, y_j)}{\epsilon^2} \\
(\hat{f}_{el})_{xy} &\approx \frac{\hat{f}_{el}(x_j+\epsilon, y_j+\epsilon) - \hat{f}_{el}(x_j+\epsilon, y_j-\epsilon) - \hat{f}_{el}(x_j-\epsilon, y_j+\epsilon) + \hat{f}_{el}(x_j-\epsilon, y_j-\epsilon)}{4\epsilon^2}
\end{aligned}$$

and equivalently for interchanged x and y.

Step 6: From Step 4, an approximation to κ is obtained for each pair of elements at the mesh points (x_j, y_j). Furthermore, for all the mesh points, mean and standard deviation are calculated, see Eq. (5.6) and Eq. (5.7). The final c-value for each pair of elements, namely $c(el_1, el_2)$ is then obtained by using Eq. (5.8).

Step 7: Finally, a ranked list of important log-ratios can be obtained from the c-values by sorting them in a descending order. Also, one can plot heatmaps, similar as in Mikšová et al. (2019b), to obtain a good overview of important elements.

5.4 Experimental results

This section presents applications of the proposed methodology on two real data sets from surface geochemical exploration. The first data set is sampled on an orogenic Au mineralization in northern Finland. Sampling at this Tiira site was designed as a stratified random grid with denser sampling on the known lodes. The other dataset was acquired at Isortoq Fe-Ti-V showings in southern Greenland. The two mineralized lodes were covered with three parallel lines with rather constant distance between the sampling stations. In both examples, 3D information about mineralized lodes is available, but it is only used in a final stage to evaluate the unsupervised methodology.

5.4.1 Tiira data set

The Tiira orogenic gold deposit is located in the Central Lapland Greenstone Belt, 12 km north of the largest gold producer in Europe, the Agnico-Eagle Finland Ltd.'s Kittilä Mine (25° 26' 2.73" E, 68° 1' 34.63" N), see Härkönen et al. (2000); Molnár et al. (2018); Geological Survey of Finland. Estimates based on historical resources are 121,000 t at 2.7 g/t Au for the Tiira ore bodies (Härkönen et al., 2000), not including the most recent 2018 exploration work (pers. comm. Jukka Välimaa, June 20, 2018). The Tiira mineralization is situated in the intersection of the NE-trending Kiistala shear zone and other NE-SW and NW-SE fractures (see Molnár et al., 2018) Gold appears native and refractory within arsenopyrite and pyrite (Härkönen et al., 2000). Eighty-four drill cores and their lithogeochemistry and a block deposit model (Agnico Eagle, 2015; Geological Survey of Finland) were visualized in 3D. The ore zone (app. 100 m x 400 m) consists of several 10-100 m wide lenses of auriferous quartz–carbonate– sulphide veins and breccias that strike NE-SW and dip slightly to the east. The mineralization is characterized with elevated Au, As, Bi and S, and sporadically with Cu, Zn and Pb. They are surrounded by mafic volcanic host rocks showing intensely altered very narrow zones of carbonatization, albitization and pyritization around the veins that can be observed as elevated K and Na. On the basis of the 3D visualization, the plant sampling stations were classified into the following categories: 1) sample on a subcropping mineralization, 2) sample on deep penetrating mineralization, depth < 50 m, 3) 50-100 m, 4) 100-150 m, 5) 150-200 m, 6) 200-250 m and 7) sample is not directly above a known mineralization. Figure 5.1 presents the sampling area, together with these different categories of sample locations.

Norway spruce (NS), common juniper (CJ), lingonberry (LB) twig and needle/leaf, and Norway spruce bark samples were collected on 60 sampling stations across the Tiira deposit. The stratified grid sampling design was also extended on the background towards

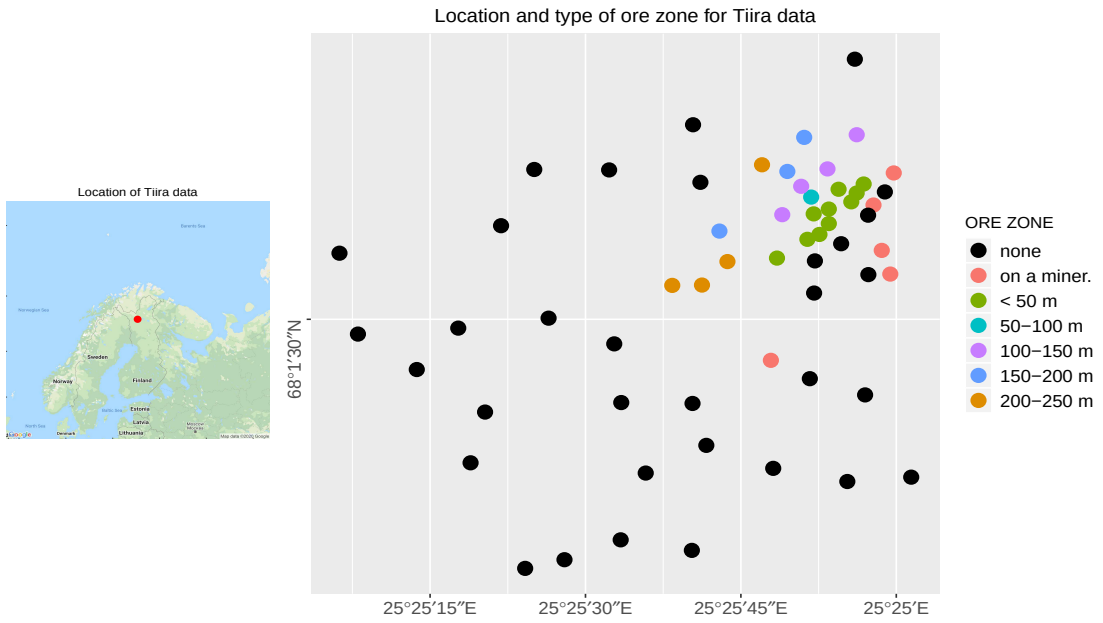


Figure 5.1: Left plot shows location (red point) of Tiira data using Google map (Kahle and Wickham, 2013). Plant sampling stations were placed on top of the known gold mineralization and covering the surrounding background based on year 2015 3D modeling data at the Tiira study site, northern Finland (right plot).

the SW to cover only areas where mature (> 100 years) Norway spruce trees were available for sampling. Lingonberry was sampled within a 10 m radius and juniper 20 m radius from the Norway spruce trees. Samples were dried (40°C for 48 h) and needles/leaves were separated from the twig. Samples were milled (< 1 mm), ashed (475°C for 24 h), 0.25 g aliquot was digested in hot aqua regia (1:1:1 HCl:HNO3:H2O mixture at 95°C for 1 h with a sample-to-acid ratio of 1:6) and analyzed for 64 elements with inductively coupled plasma mass spectrometry and optical emission spectrometry in Bureau Veritas Commodities Canada Ltd., AcmeLabs (Vancouver, Canada). Samples were randomized for analysis, field duplicates were collected (8.7 %) and standard reference materials (9.4%, CDV1-Ash, eucalyptus foliage from Western Australia, Colin Dunn Consulting, North Saanich, Canada) were inserted in the analysis sequence.

Before applying the proposed procedure, relevant elements were selected from a data quality point of view. Elements with a high proportion (more than 40%) of values below a lower or above an upper detection limit, or discretization (rounded values close to detection limit) were discarded, which resulted in a total of 25 elements for the analysis. For the purpose of comparability, the same 25 elements have been selected in each sample

material.

GAM fits of original data values

As described in the algorithm in Sect. 5.3, the first two steps perform GAM fits of the concentration values using predefined weights which emphasize high concentrations. Some examples of those GAM fits are shown in the following. Figure 5.2 (left) shows a 3D presentation of the GAM fit for the concentration values of gold (Au) from the Norway spruce twig samples. The original data are shown in orange, and the values of the GAM fit surface are according to the color scheme in the legend to the very right. Two clear peaks are visible in the GAM fit, and they follow the two extreme Au values in a continuous manner. Figure 5.2 (right) presents this information as contour plot in two dimensions. The original data values are shown as points by using the same color scheme as for the surface. This visualization is easier to compare with the map presentation from Fig. 5.1. It can be seen that the highest Au values can be expected just in the south-west of the densely sampled area, and in the center of the investigated region.

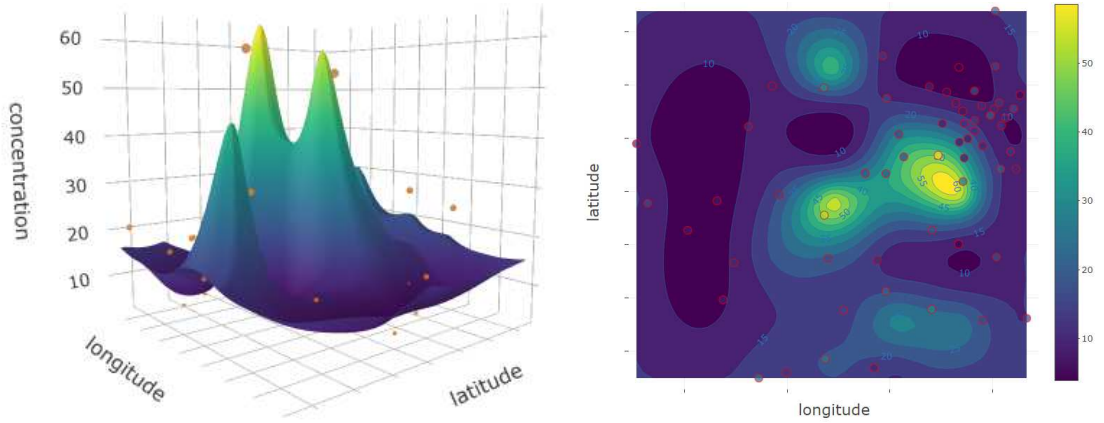


Figure 5.2: GAM fit for Au with original concentration values (points) of Norway spruce twig samples from the Tiira data: left plot for 3D fit, right plot for projection into 2D, using a contour presentation with isolines.

Further GAM fits are shown in Fig. 5.3 as contour plots: arsenic (As) and copper (Cu) of Norway spruce twig samples (top), and magnesium (Mg) and sulfur (S) of Norway spruce bark samples (bottom). The plot for S shows that there can be a lot of local variability of the concentration values, but the GAM fits aim at smoothening those values. Values with higher concentration receive higher weight for the GAM fit (see Step 1 of the algorithm), with the aim that upper concentration should also result in clear signals in the fit.

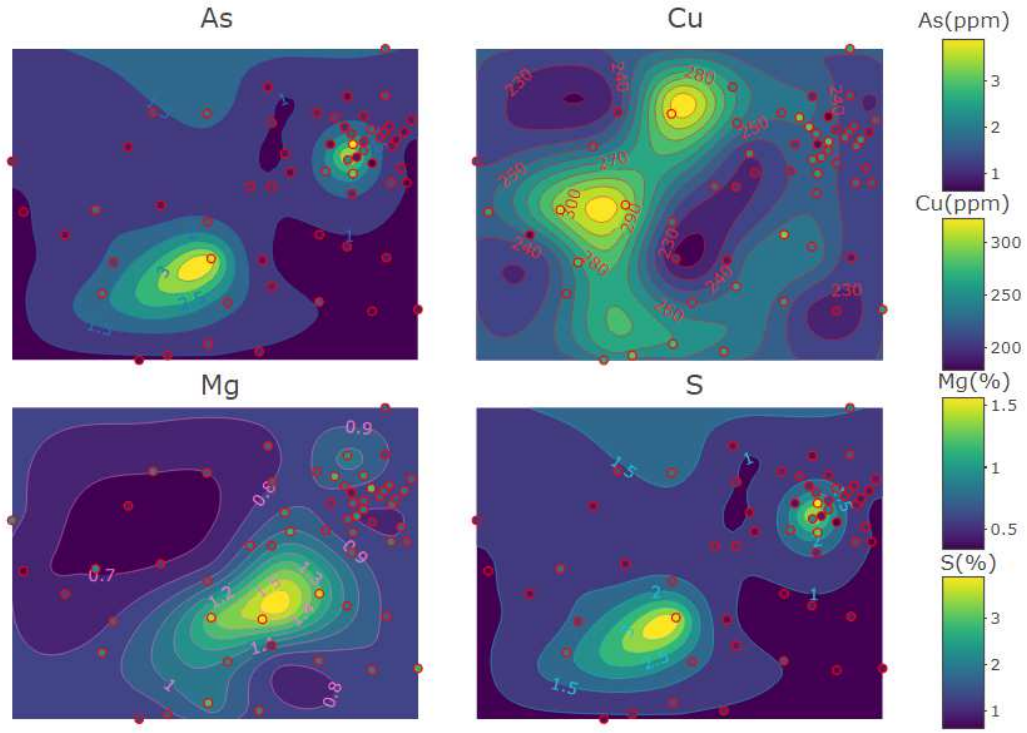


Figure 5.3: GAM fits for the elements As and Cu (Norway spruce twig) and Mg and S (Norway spruce bark), together with the original concentration values (colored dots).

Log-ratios of GAM fits and curvature

Once the GAM fits are available for all elements and all plant materials, the log-ratios of the fits for all possible element pairs per material can be computed. Note that the log-ratios are formed not from the originally measured concentration values, but only from the GAM fits, which are constructed on a regular grid with 100 horizontal and 100 vertical grid points, thus 10.000 values. This number of "artificial" data points turned out to be practical for this purpose, also considering computational issues. Since the GAM fits are smooth, also the log-ratios will appear as smooth surfaces. Examples of two log-ratio pairs can be seen in Fig. 5.4 (left column). This information is used in the following to identify mineralized zones, which should be indicated by rapid spatial changes. A log-ratio shows such local changes if either both elements have local changes, or one element is stable, reflecting background, and the second element has strong local variability. In either case, since the logarithm is used, it does not matter which element

is used in the numerator, and which in the denominator, as long as a subsequent measure of local change does not consider the sign. This is indeed the case with the curvature measure proposed in Eq. (5.5). Figure 5.4 visualizes the resulting curvature values of the GAM fits for the log-ratios log(Au/Cu) and log(Mg/S), see Fig. 5.3 for the GAM fits. Note that the curvature values are typically not big at the locations of the log-ratio peaks, but they are big at locations of abrupt changes of the log-ratios, for instance around the peaks. Both curvature plots show several local changes, and the next step is to evaluate the overall local changes in curvature using the c-value of Eq. (5.8).

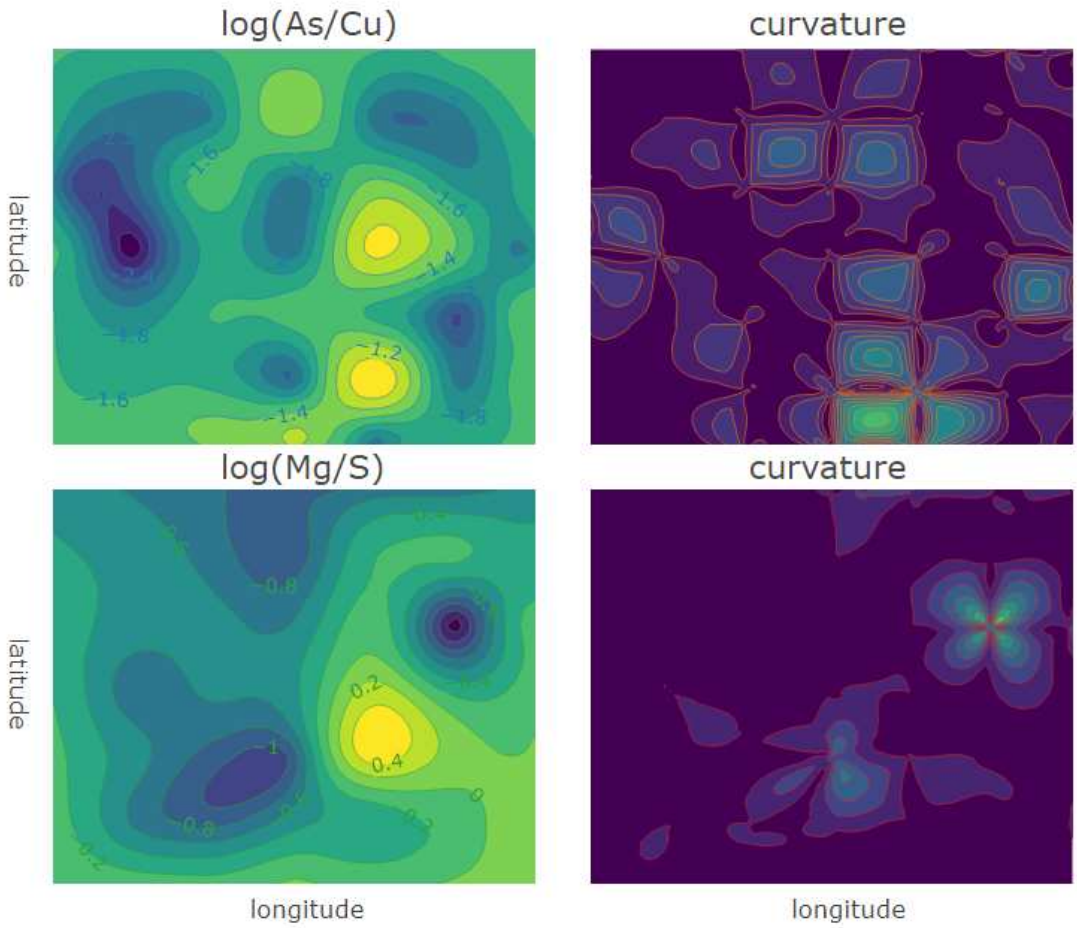


Figure 5.4: Log-ratios of the GAM fits shown in Fig. 5.3 (left column), and corresponding curvatures (right column).

Comparison of log-ratios and different sample media

Since there are in total 25 elements of interest per sample medium, this results in 300 different pairwise log-ratios, for each of the six sample materials. The c-value of Eq. (5.8) allows to perform a ranking of the log-ratios, because it neither depends on the scale of the elements, nor on the scale of the log-ratio, and thus is comparable across all curvature values of the log-ratios, and even across those from different sample materials.

Table 5.1 shows the top-10 ranked log-ratios, together with the resulting c-values, for the sample materials twig, needle and bark of Norway spruce (the values of the other materials are not shown due to shortage of space). It can be seen that there are elements which are involved in many of the log-ratios, such as As in NS twig, see also Fig. 5.3 (top left plot). The log-ratio log(As/Cu) from NS twig, which is shown in Fig. 5.4 (top left plot), thus has a c-value of 16.8×10^{-5}, and is listed with rank 3. It should therefore be highly informative because of its strong curvature in the area, and As is also referring to the anomaly. The anomaly is also characterized by S, and the log-ratio log(Mg/S) in NS bark (see Fig. 5.4 bottom left) also has a high c-value (the value is 9.2×10^{-5}). Since S is in the denominator of this log-ratio, low values point at the relevant locations.

Table 5.1: Top-10 ranked log-ratios for Norway spruce twig, needle and bark samples of the Tiira data.

Media	NS twig		NS needle		NS bark	
Ranking	log-ratio	c-value ($\times10^{-5}$)	log-ratio	c-value ($\times10^{-5}$)	log-ratio	c-value ($\times10^{-5}$)
1	As/Mg	19.8	Al/Sm	9.0	Pb/Zn	19.5
2	As/Zn	17.1	Nd/Y	9.0	Mg/Pb	17.2
3	As/Cu	16.8	Rb/Sm	8.5	Na/Pb	15.4
4	Nd/Sm	15.4	S/Sm	8.0	Ni/Pb	14.8
5	As/Sb	14.1	Co/Sm	7.8	Cu/Pb	14.7
6	As/S	14.1	Al/Y	7.6	Pb/Rb	13.6
7	As/Ni	14.1	Ni/Y	7.4	Co/Pb	12.1
8	As/Pb	13.5	Ni/Sm	7.2	Mo/Pb	10.9
9	As/La	13.2	Pr/Y	7.2	La/Pb	10.9
10	Ni/Sb	12.9	Cu/Sm	7.0	Sb/Sc	10.0

Figure 5.5 shows the c-values of the top-100 ranked log-ratios as lines, where one line is used for each sample material. For example, the highest c-value was achieved by a log-ratio from material NS twig, with a value of about 0.0002, see also Tab. 5.1. The top-25 log-ratios of this material have higher c-values than any log-ratio pair of material CJ twig, because the curve for this material generally shows very low values. One would conclude that NS twig is the most promising sample material to identify mineralization,

and CJ twig the least interesting one, provided that this unsupervised procedure indeed points at the relevant locations.

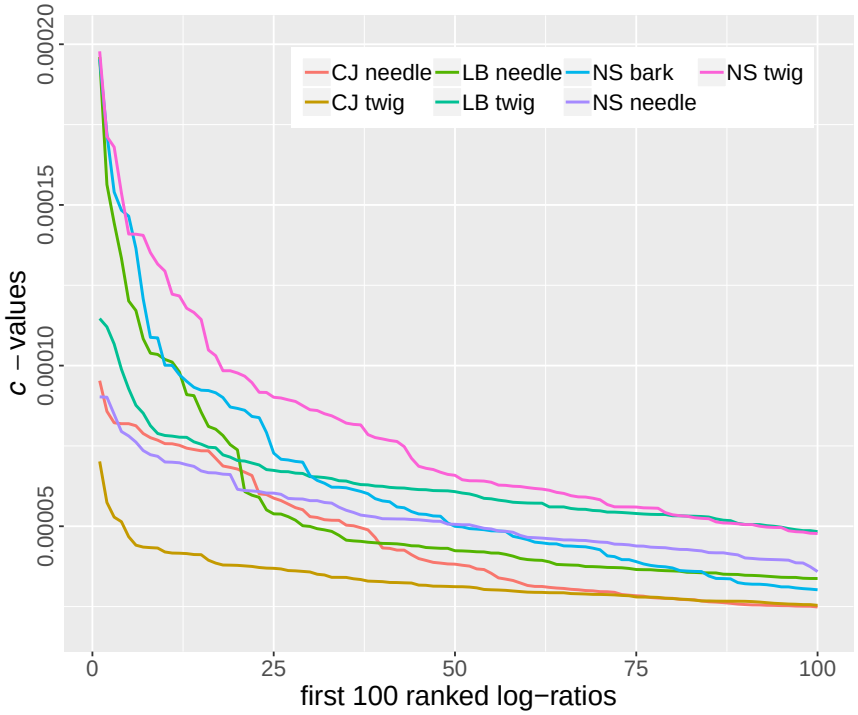


Figure 5.5: Comparison of the top-100 ranked c-values for all plant materials.

As a final step, it may be interesting to identify so-called pathfinder elements which indicate mineralization. For this purpose, one could count how often each element is involved in all 300 investigated log-ratio pairs, and show the ranked list. This is done in Tab. 5.2, which provides the top-10 elements per material being involved in all pairwise log-ratios. This list can be compared with the list of elements that – from a geochemical point of view – characterize the mineralization, namely Au, As, Bi and S, and also Cu, Zn and Pb (see beginning of this section). Many of these elements can be found in the hitlists of the different sample materials. Since NS twig was considered the most relevant material, As measured in NS twig should be highly indicative as a pathfinder element, see Fig. 5.3, upper left plot. Lead (Pb), neodym (Nd), yttrium (Y) and praseodymium (Pr) are in the top-10 list in several plant materials, but also gold (Au) appears as a top pathfinder element.

Table 5.2: Top-10 mostly involved elements in all pairwise log-ratios for each plant material.

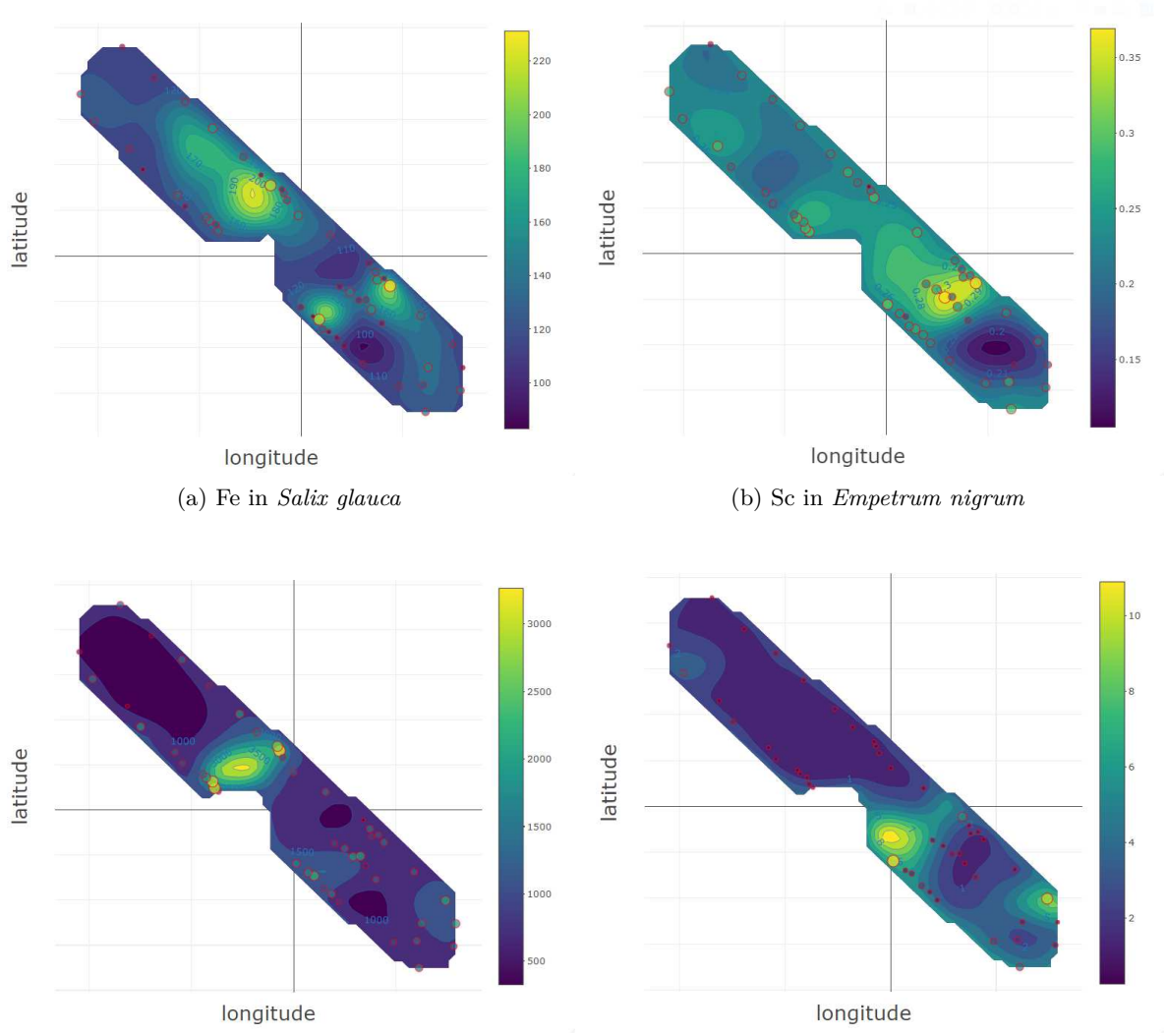
ranking	NS twig	NS needle	NS bark	LB twig	LB needle	CJ twig	CJ needle
1	As	Sm	Pb	Co	Pb	Rb	Pb
2	Sb	Y	Au	Pb	Pr	Au	Rb
3	Sm	Pr	Sb	Pr	Y	Sc	Cu
4	Y	Nd	S	Nd	Nd	Mo	Ni
5	Nd	Zn	Na	Au	Co	Cu	Co
6	Pr	Sb	Y	Sm	Mo	Pr	Mo
7	Ce	As	As	Y	Au	Pb	Nd
8	Co	Ce	Co	Sc	Ce	Mg	Sb
9	Pb	Mg	Mg	Nb	Nb	Co	Ti
10	Au	Co	Ni	Ce	Rb	Ti	Nb

5.4.2 Greenland data set

This data set originates from the area Isortoq, situated in the southwest of Greenland, and it was provided by GEUS, the Geological Survey of Denmark and Greenland. Extensive exploration by mining companies have confirmed the existence of distinct dykes consisting of mineralized troctolite with high abundance of titano-magnetite Fe-Ti-V minerals. Furthermore, the Isortoq area is affected by the Gulf Stream currents which ensure good conditions for vegetation and thus for sampling. Linear transects have been selected for sampling in two areas. The first area consisted of three traverses, where each line was 4 km in length and had 34, 34 and 29 sample stations, respectively. The second area had two traverses, at 2 km each. Soil samples and two plant species *Salix glauca* and *Empetrum nigrum* were chosen as sample media. After the geochemical analysis, in total 21 elements have been selected in each sample material, following quality control considerations. More details on the data are available in Mikšová et al. (2020), where the analysis for mineral identification for samples along a line has been applied. This analysis can now be compared to the two-dimensional extension of the methodology.

Figure 5.6 shows GAM fits of the selected elements Fe (iron), Sc (scandium), Ti (titanium), and Mo (molybdenum) in different sample materials, together with the original data values. The peaks of the GAM fits in the southern part of the area correspond to the places of known mineralization, and indeed the known deposits are Fe, Ti and V (vanadium) mineralization. Since the measurements of V are not very reliable, Sc might serve as a pathfinder.

Similar to the previous example, the longitude and latitude of the investigated area is divided into a grid of 100×100, and the GAM fits of the grid points are used to compute



(a) Fe in *Salix glauca*

(b) Sc in *Empetrum nigrum*

(c) Ti in *Salix glauca*

(d) Mo in soil

Figure 5.6: GAM fits for the elements iron (Fe), scandium (Sc), titanium (Ti), and molybdenum (Mo) of the Greenland data.

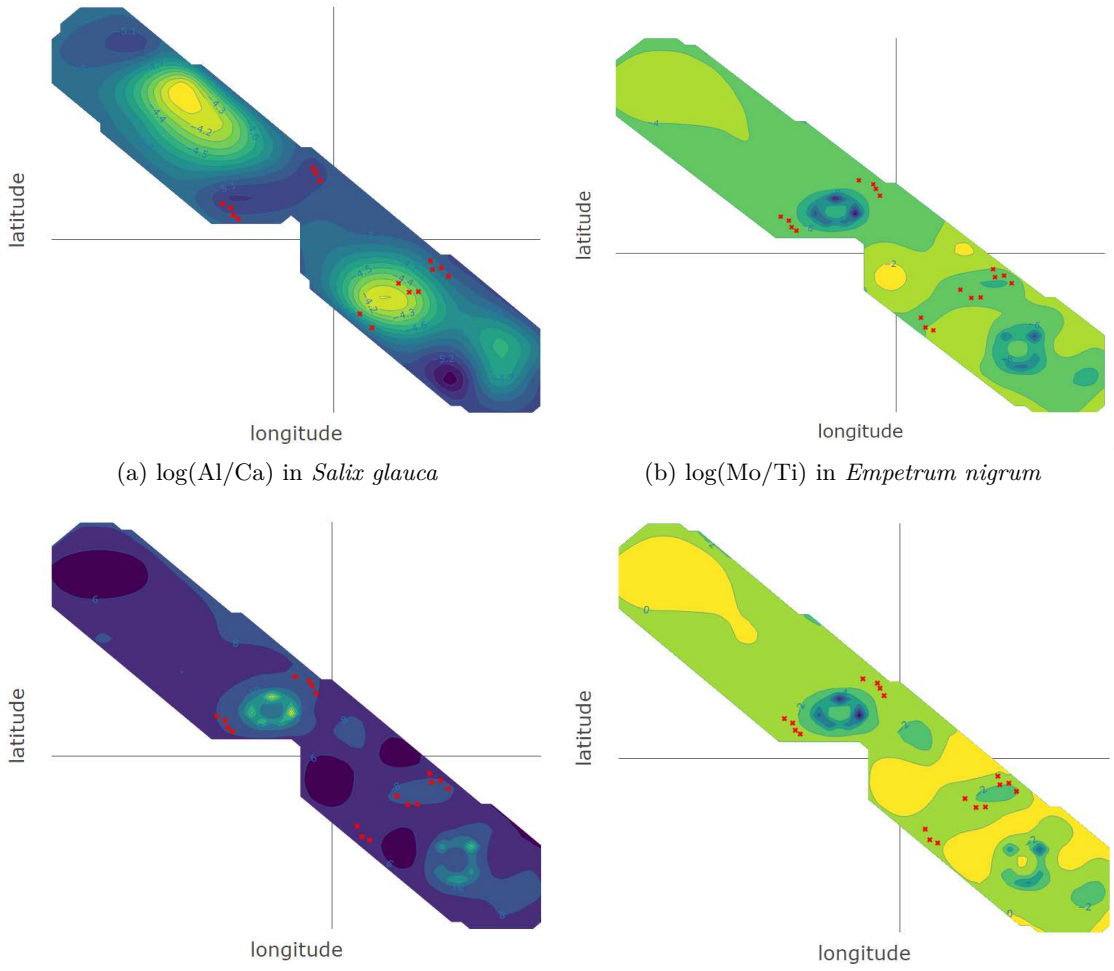
the pairwise log-ratios per sample material. Only those grid points have been taken which are falling in the colored areas shown in Fig. 5.6. Finally, the c-value is computed, and Tab. 5.3 shows the top-15 ranked log-ratios for each sample material, together with the resulting c-values. The element Mo is dominant in many of those log-ratio pairs in *Empetrum nigrum* and soil samples, which corresponds to the findings in the univariate consideration of the problem, see Mikšová et al. (2020).

Table 5.3: Top-15 ranked log-ratios for each sample material for the Greenland data.

Media	*Salix glauca*		*Empetrum nigrum*		Soil	
Ranking	log-ratio	c-value	log-ratio	c-value	log-ratio	c-value
1	Ba/Zn	20.7	Cs/Mo	38.4	Al/Mo	40.6
2	Al/Sc	19.8	Mo/Pb	36.3	Fe/Mo	40.5
3	La/Ti	13.2	Mo/Ti	35.9	Mo/Ni	38.6
4	Al/Ca	12.6	Mo/Rb	35.8	Mo/Ti	38.6
5	Ba/Ce	12.6	La/Mo	35.5	Mo/Pb	38.6
6	Pb/Rb	12.4	Ba/Mo	35.1	Mo/V	38.6
7	La/Sr	12.3	Mn/Mo	34.9	Mo/Sc	38.5
8	Cs/Pb	12.2	Fe/Mo	34.8	Ce/Mo	38.5
9	La/Sc	11.1	Al/Mo	34.6	La/Mo	38.5
10	Cs/Rb	10.9	Ce/Mo	34.6	Mo/Zn	38.5
11	Ca/K	10.8	Mo/Zn	34.3	Mn/Mo	38.5
12	Cs/Ni	10.8	Mg/Mo	34.1	Cs/Mo	38.5
13	La/P	10.7	Ca/Mo	34.0	Mo/P	38.4
14	K/La	10.6	Mo/Sr	33.9	Mo/Sr	38.4
15	K/Sc	10.3	Co/Mo	33.9	Co/Mo	38.4

Figure 5.7 presents some selected log-ratios: log(Al/Ca) in *Salix glauca*, shown as upper left plot, is the second-ranked log-ratio, see Tab. 5.3 in this sample medium. The red dots are the locations of the known mineralization, and a peak is visible at the location of the southern mineralized zone. In addition, there is a peak at a new location, which could be a potentially mineralized area. The remaining three plots in Fig. 5.7 show essentially a quite similar structure. All these log-ratio pairs are in the top ranks according to their c-value, see Tab. 5.3, and they have a strong peak at the mineralization higher to the north. The elements involved in these log-ratios are clearly related to the mineralization (Fe, Ti, Sc). However, there is also a second peak in the southern part of the area visible, and this could again be a promising location for further investigation.

Finally, the sample materials are compared according to the c-values. Figure 5.8 shows the ranked c-values for all 231 log-ratio pairs of the different sample materials.



(a) log(Al/Ca) in *Salix glauca*

(b) log(Mo/Ti) in *Empetrum nigrum*

(c) log(Fe/Mo) in soil

(d) log(Mo/Sc) in soil

Figure 5.7: Plots of some of the log-ratios with high c-values, see Tab. 5.3; red crosses represent known locations of subcropping mineralization.

There is a strong decline of the values for *Empetrum nigrum* and soil after about 23 ranks. The few very high values are caused by single elements leading to important log-ratios, such as Mo, see Tab. 5.3. *Salix glauca* in general has much lower *c*-values and thus might not be considered to have the same importance as the other sampling media.

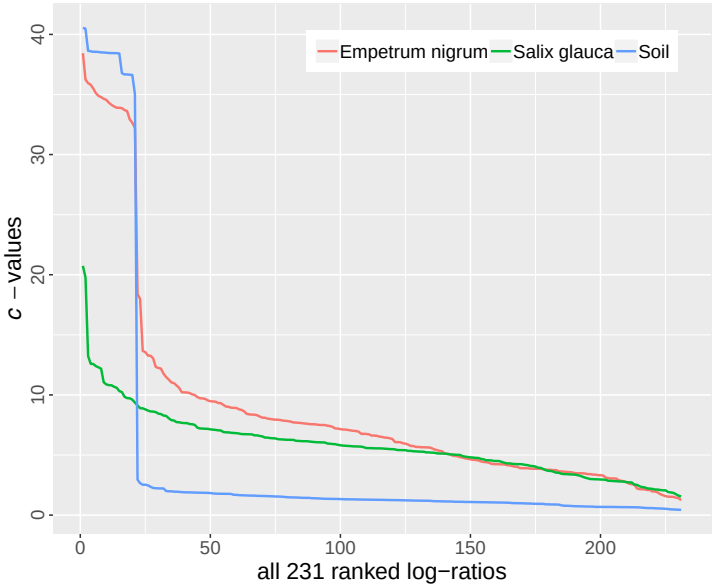


Figure 5.8: Comparison of three sample media of the Greenland data by the ranked *c*-values.

5.5 Discussion and conclusions

A completely unsupervised mathematical procedure has been proposed to identify an underlying mineralization and to point out its location. The methodology is developed for surface geochemical samples from soil horizons and plants which have been sampled following a 2D grid design which geometry may be regular or irregular. In a first step, the absolute element concentration data are approximated by a smooth surface fitted by a Generalized Additive Model (GAM) in the sampling area. Then, an arbitrary number of data points is taken from the surface of the smooth fits, representing new artificial concentration data for an element. Pairwise log-ratios of elements are then computed on this artificial smoothed data, curvature of the log-ratios is calculated, and finally a measure called *c*-value is calculated, reflecting the overall curvature of the log-ratio pair. The higher the *c*-value is, the more pronounced spatial changes are present in the specific

log-ratio pair. Regions of abrupt spatial changes can be indicative of an underlying mineralized zone. The extension of the procedure (Mikšová et al., 2019b) to the 2D case presented in this paper is not trivial, because there are several technical and numerical issues which had to be considered.

It is demonstrated with the two real data applications that the procedure is able to spot at potentially interesting locations and highlight elements which form these abrupt changes in the pairwise log-ratios. Based on these results of these experiments, the method can be considered as a first pass data exploration technique which guides the geochemist towards potentially interesting exploration targets and elements. In each sample medium where the concentrations for p chemical elements are available, $p \cdot (p-1)/2$ pairwise log-ratios are considered for the methodology, and they are finally ranked according to their c-values. This does not necessarily mean that only the top-ranked pairs are interesting or relevant, and others contain no information. Anomalies in these datasets are also caused by other factors besides exogenic geochemical signals traveling from underneath the sampling locations. The challenge is to separate these false positive signals from the true positive drill targets which may lead to mineral discovery. Rather, the ranking is intended to draw attention and guide the practitioner to possibly interesting elements and their log-ratio pairs, because manually inspecting $p \cdot (p-1)/2$ map plots of log-ratios is tedious. Practically, one would start looking at the corresponding maps of the log-ratios from top-rank upwards, and also consider geochemical knowledge for the validity or relevance of the involved elements in the target at hand. Especially in this process it is desirable that geochemists do not blindly follow a mathematically optimized strategy because it may lead to cost-intensive bedrock drilling of false positive targets. An alternative way applying the algorithm would be a pre-selection of the input elements not only based on quality standards, such as done in this paper, but by narrowing them down based on conceptual geological and geochemical models presented by an exploration geologist and geochemist familiar with the geological environment.

CHAPTER 6

Summary

The identification of geochemical features related to deposit signature, and the separation of background and target zones for future local mineral explorations are becoming popular challenges in geochemistry, since it might lead to ore discoveries which is of high interest.

The main objective of this thesis was to develop statistical methods for mineral exploration in geochemistry. For this task, the whole process is important, starting from sampling planning, the collection of the samples, quality control, data processing, statistical analysis, and interpretation of the results. The first steps, the sampling design and quality control, were treated in the introduction of the thesis. At least for the sampling design it turned out that statistical considerations alone would not be useful in this context; there are many practical aspects that need to be taken into account as well. Nevertheless, optimal sampling design procedures are the key for a successful geochemical exploration study. The next step when dealing with geochemical data is to conduct a thorough quality control procedure to gain confidence in the investigated geochemical data. For this purpose it is necessary to evaluate the laboratory accuracy and the field precision. These procedures were improved and completed such that they are applicable to surface geochemical techniques. The quality control methods are usually based on the concentration data as they are reported from the laboratory. Subsequent statistical analyses, however, are (or should be) based on the composition, thus they need to incorporate the multivariate information. For this reason, a brief introduction to compositional data analysis is provided in Chapter 1.

Even if the data quality as a result from quality control analyses is appropriate, there will in general be measurements with censoring, because the concentrations of some elements are either very low or very high, exceeding a lower or an upper threshold. This

threshold, called the detection limit, depends on the instrument, and values reported below/above the detection limit would be very unreliable. Since statistical methods for compositional data are based on the multivariate information, it is important that values below/above the detection limit are taken into account appropriately. While several statistical procedures are available to estimate values below a lower detection limit, also from a compositional point of view, no multivariate procedures have been proposed so far to deal with the estimation of values above an upper detection limit. Chapter 2 proposes a regression based method for the imputation of right-censored values in compositional data. This methods makes use of the non-censored information of the remaining elements, and it follows the aspects of compositional data analysis. The replacement of values above the upper detection limit by this method turns out to be more reasonable than the widely used rule of multiplying the reported upper detection limit value by a factor.

The main focus of the thesis was on developing methods for identifying mineralized zones for surface geochemical exploration. The main difficulty is that usually only few measurements, say 30-40, are available, and only very few of them will be taken on top of a mineralization such that certain element concentrations clearly deviate from those which are taken from the background. This makes it difficult to distinguish "true signals" from values that are somehow deviating due to different kinds of uncertainties. The main idea is thus to eliminate uncertainties by estimating the "signal", which is done by fitting the element concentrations with a smooth curve or surface, here with GAMs (Generalized Additive Models). Once the GAM fits are available, it is possible to generate a desired number of grid points in the area under investigation and to compute the fitted values on the smooth curve/surface. While Chapter 3 and 4 treat the special case of measurements given on a linear transect, Chapter 5 considers the more general case of measurements taken on a grid in the plane. Thus, the GAM fits in the former case are smooth curves, while in the latter case they are surfaces, being computationally more challenging. In either case, once the GAM fits are available, log-ratios for element pairs are considered, and a measure of overall curvature is computed. High curvature indicates abrupt signal changes, which could point at locations of mineralized zones. The measure of overall curvature allows for a ranking of the log-ratio pairs, which supports the geochemist in identifying pathfinder elements for mineralization. Several geochemical data sets for mineral exploration have been used in these chapters as demonstration examples, and according to experts, the results are reasonable and helpful for the purpose.

An interesting possible extension of the methodology proposed in Chapters 3–5 would be the incorporation of further dimensions. In addition to longitude and latitude, also the depth of the samples, the depth of drilling cores, the time when the samples

have been taken, or other significant information could be incorporated. This could lead to a more accurate prediction of possibly mineralized zones.

List of Figures

1.1 Map presentation of cobalt concentrations in Cowberry-twig (modified from Torppa and Middleton, 2017). Concentration values are clustered according to quantiles in the data. Low concentrations are presented with blue and high with red. . . . 11
1.2 Line plot for cobalt, Cowberry-twig/stem (Ultra-LIM project), see Torppa and Middleton (2017). . . . 12
1.3 Three types of sampling designs (Humboldt State University, 2018). . . . 13
1.4 Legend for QAQC 1 phase, different colors are used for different plants (left). QAQC 1 plot shows potential blockiness or periodicity (right). Samples were not randomized for the analysis. Thus periodicity cannot be estimated. . . . 18
1.5 Legend for QAQC 2 (left) – Lines indicate measures of standard reference samples (SRM). Plot for QAQC 2 – Analysis sequence versus measured standard reference samples (right). . . . 19
1.6 QAQC 3 boxplots of the routine sample and blank sample analytical results in a chart for the element Pb for original data (left) and for log-transformed data (right). Blue line indicates lower detection limit. . . . 20
1.7 QAQC 4 for field precision. . . . 21
1.8 QAQC 5 for laboratory precision. . . . 22

2.1 Plots of measured versus estimated values for P and Zn for the Lätäseno data set. . . . 36
2.2 QQ-plots of the reported values for P and Zn for the Lätäseno data set. . . . 36
2.3 Plot of measured versus estimated values for Fe when the UDL is equal to the quantile 0.8 (subset spruce from the Gjøvik data). . . . 39
2.4 Log-ratio biplots for the original data, and for data based on imputation with the simple and the classical method in the variables Ce and Cs, where the UDL is equal to the quantile 0.8 (subset spruce from the Gjøvik data). . . . 41

2.5 Classical and robust imputation for the spruce data of Gjøvik with 13 variables. The UDL is set to the quantile 0.8 for each of the 1 to 10 randomly selected variables. The plots show the numbers of iterations of the algorithm for the imputation, for 100 simulation replications. 41

2.6 Error measurements depending on sample size: Data sets with the indicated sample size are randomly drawn from the Gjøvik data set, and imputation is done in one randomly selected variable, where the UDL is set to the quantile 0.8. 43

2.7 Comparison of classical and robust regression imputation with (TRUE) and without (FALSE) variable selection for the spruce data subset. The number of variables to be imputed is increased, and imputation needs to be done for the upper 20% of the values. 44

2.8 Comparison of the simple method, and of classical and robust regression imputation with (TRUE) and without (FALSE) variable selection for the spruce data from Gjøvik. Imputation is done in only one randomly selected variable, by modifying the UDL value from the quantile 0.5 to 0.95. 45

2.9 Imputation for each individual sample material (listed on the horizontal axis). The upper detection limit is modified from the quantile 0.5 to 0.95 of the values for a randomly selected variable, and the average of the error measures is computed. The boxplots show the outcomes for 100 simulations. The order of the sample materials is according to the median performance of the classical method. 46

2.10 Imputation for the complete data set. The upper detection limit is modified from the quantile 0.5 to 0.95 of the values for a particular variable, and the average of the error measures is computed for each variable separately. The boxplots show the outcomes for 100 simulations. The order of the variables is according to the median performance of the classical method. 47

3.1 Top row: function $x \mapsto (1+(\frac{x}{\sigma})^2)^{-1}$ for different σ. Bottom row: corresponding curvature (to be defined in Section 2.3). 55

3.2 Example of a log-ratio plot of the elements Al and Co and the corresponding curvature plot with the threshold (dashed line). One can see the correspondence between local maxima above the threshold in the right plot with the peaks in the left plot. 58

3.3 GAM fits (lines) for eight selected elements measured in Crowberry twigs from the Juomasuo data set are displayed together with their original concentrations (dots). . . . 63
3.4 Upper part: four different log-ratios of GAM fits. Lower part: corresponding curvature together with the threshold (dashed red line). . . . 64
3.5 Heatmaps of the c-values per element for all possible log-ratios of the tissue twig for all plant species, and the accumulated values of all materials (upper left). . . . 67
3.6 Heatmaps of the c-values for all possible log-ratios of four media – BIL, BLE, CLE, SNE. . . . 68
3.7 Curvature of log(Mo/Tl) in sample material Birch leaves, where known mineralization area (red points) and mineralized points identified by the method (blue points) are displayed. . . . 69
3.8 Upper part: Log-ratio of lead and aluminium constructed by using GAM fits of its individual elements – displayed on lower part of plot. Sample material is BLE. The red points indicate areas of known mineralization. . . . 70
3.9 Top-ranked 70 (unscaled) c-values for each sample material. The horizontal axis represents the rank. . . . 71

4.1 Map of the locations of the samples taken by GEUS in the Isortoq South Area. 80
4.2 Curvature of the log-ratio of the GAM fits of Ti and Ca in soil. . . . 81
4.3 Curvature of the log-ratio of the GAM fits of Fe (iron) and P (phosphorus) in soil. . . . 81
4.4 Heatmaps of the c-values for the plant materials and soil. . . . 82
4.5 Curvature of the log-ratio of the GAM fits of Na (sodium) and Pb (lead) in soil. 83
4.6 Map with the sample locations taken by BRGM in the Vendée area in 2018 . 84
4.7 Top-ranked 70 c-values, computed for 6 different sample materials and the three different sample sites. . . . 85
4.8 Heatmaps of the c-values for the BRGM data - soil and plant material. . . . 85

5.1 Left plot shows location (red point) of Tiira data using Google map (Kahle and Wickham, 2013). Plant sampling stations were placed on top of the known gold mineralization and covering the surrounding background based on year 2015 3D modeling data at the Tiira study site, northern Finland (right plot). . . . 96

5.2 GAM fit for Au with original concentration values (points) of Norway spruce twig samples from the Tiira data: left plot for 3D fit, right plot for projection into 2D, using a contour presentation with isolines. 97
5.3 GAM fits for the elements As and Cu (Norway spruce twig) and Mg and S (Norway spruce bark), together with the original concentration values (colored dots). 98
5.4 Log-ratios of the GAM fits shown in Fig. 5.3 (left column), and corresponding curvatures (right column). 99
5.5 Comparison of the top-100 ranked c-values for all plant materials. 101
5.6 GAM fits for the elements iron (Fe), scandium (Sc), titanium (Ti), and molybdenum (Mo) of the Greenland data. 103
5.7 Plots of some of the log-ratios with high c-values, see Tab. 5.3; red crosses represent known locations of subcropping mineralization. 105
5.8 Comparison of three sample media of the Greenland data by the ranked c-values. 106

List of Tables

2.1 Resulting error measurements for the three imputation methods after imputing the upper 20% of the values of Fe. 40

3.1 Top 10 ranked log-ratios and its scaled c-values for the plant materials Crowberry (CRO), Bilberry (BIL), and Labrador tea (LBT). 65
3.2 Top 10 ranked log-ratios and its elements for each material. 65

5.1 Top-10 ranked log-ratios for Norway spruce twig, needle and bark samples of the Tiira data. 100
5.2 Top-10 mostly involved elements in all pairwise log-ratios for each plant material. 102
5.3 Top-15 ranked log-ratios for each sample material for the Greenland data. . . 104

Bibliography

Agnico Eagle. Annual report, 2015. http://s1.q4cdn.com/150142668/files/doc_financials/2015/2015-Annual-Report.pdf.

J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, London. (Reprinted in 2003 with additional material by The Blackburn Press), 1986.

J. Aitchison and M. Greenacre. Biplots for compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 51(4):375–392, 2002.

C. Barceló-Vidal and J. Martín-Fernández. The mathematics of compositional analysis. *Austrian Journal of Statistics*, 45(4):57–71, 2016.

E. Beinrohr, B. Sileš, J. Štefanek, and V. Rattay. Determination of traces of sodium and potassium in gallium arsenide by graphite furnace atomic absorption spectrometry and flame atomic emission spectrometry. *Chemical Papers*, 45(1):61–68, 1991.

A. Buccianti. Is compositional data analysis a way to see beyond the illusion? *Computers & Geosciences*, 50:165–173, 2013.

A. Buccianti, A. Lima, S. Albanese, C. Cannatelli, R. Esposito, and B. De Vivo. Exploring topsoil geochemistry from the CoDa (Compositional Data Analysis) perspective: The multi-element data archive of the Campania Region (Southern Italy). *Journal of Geochemical Exploration*, 159:302–316, 2015.

E. J. M. Carranza. Geochemical mineral exploration: should we use enrichment factors or log-ratios? *Natural Resources Research*, 26(4):411–428, 2017.

Q. Cheng. 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*, 122:55–70, 2012. doi: 10.1016/j.gexplo.2012.07.007.

W. Cleveland, E. Grosse, and W. Shyu. Local regression models. In J. Chambers and T. Hastie, editors, *Statistical Models in S*, pages 309–376. Wadsworth & Brooks/Cole, Pacific Grove, CA, 1992.

W. Cochran. *Sampling Techniques*. Wiley, New York, USA, 1977.

H. Dahooei, P. Afzal, M. Lotfi, and A. Jafarirad. Identification of mineralized zones in the Zardu area, Kushk SEDEX deposit (Central Iran), based on geological and multifractal modeling. *Open Geosciences*, 8:143–153, 2016.

F. Darabi golestan, R. Ghavami, R. Khalokakaie, H. Asadi, and M. seyedrahimi Niaraq. Interpretation of lithogeochemical and geophysical data to identify the buried mineralized area in Cu-Au porphyry of Dalli-Northern Hill. *Arabian Journal of Geosciences*, 6:4499–4509, 2013.

J. De Geoffroy and S. Wu. Design of a sampling plan for regional geochemical surveys. *Economic Geology*, 65(3):340–347, 1970.

J. De Gruijter, D. Brus, and M. Bierkens. *Sampling for Natural Resource Monitoring*. Springer-Verlag, Berlin, 2006.

M. P. Do Carmo. *Differential geometry of Curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016.

C. Dunn. *Biogeochemistry in Mineral Exploration, Handbook of Exploration and Environmental Geochemistry*, volume 9. Elsevier, Amsterdam, 2007. ISBN 978-0-444-53074-5.

J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3): 279–300, 2003.

J. J. Egozcue. Reply to "On the Harker variation diagrams; ..." by J. A. Cortés. *Mathematical Geosciences*, 41(7):829–834, 2009.

P. Filzmoser, K. Hron, C. Reimann, and R. Garrett. Robust factor analysis for compositional data. *Computers & Geosciences*, 35(9):1854–1861, 2009.

P. Filzmoser, K. Hron, and C. Reimann. Interpretation of multivariate outliers for compositional data. *Computers & Geosciences*, 39:77–85, 2012.

P. Filzmoser, K. Hron, and M. Templ. *Applied Compositional Data Analysis. With Worked Examples in R*. Springer Series in Statistics, Springer, Cham, Switzerland, 2018.

E. Fišerová and K. Hron. On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43(4):455–468, 2011.

A. T. Fomenko and A. A. Tuzhilin. *Elements of the Geometry and Topology of Minimal Surfaces in Three-Dimensional Space*, volume 93. American Mathematical Society, 2005.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York, 2001.

F. Gallego. Stratified sampling of satellite images with a systematic grid of points. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(6):369–376, 2005.

R. Garrett. Sampling methodology. In R. Howarth and G. Govett, editors, *Handbook of Exploration Geochemistry*, volume 2, pages 83–110. Elsevier, Amsterdam, 1983.

R. Garrett. *Sampling Strategies. A Short Course and Tutorial*. Geological Survey of Canada, Ottawa, April 16-20, 2012.

Geological Survey of Finland. Mineral deposit report, 2020. electronic resource available at. http://tupa.gtk.fi/karttasovellus/mdae/raportti/387_Kuotko.pdf.

R. Goldman. Curvature formulas for implicit curves and surfaces. *Computer Aided Geometric Design*, 22(7):632–658, 2005.

J. Gonzalez and J. Eltinge. Optimal survey design: A review. *Section on Survey Research Methods–JSM*, 2010.

S. Hamilton. Spontaneous potentials and electrochemical cells. In *Handbook of Exploration Geochemistry*, volume 7, pages 81–119. Elsevier, 2000.

S. Hamilton. *A prospector's guide to the use of selective leach and other deep penetrating geochemical techniques in mineral exploration*. Ontario Geological Survey, 2007. Open File Report 6209.

I. Härkönen, H. Pankka, and S. Rossi. Summary report: The Iso-Kuotko gold prospects, northern Finland. Technical Report Report No. C/M06/2744/00/1/10, Geological Survey of Finland (GTK), 2000.

D. Heberlein and C. Dunn. The application of surface organic materials as sample media over deeply buried mineralization at the Kwanika Central Zone, north-central British Columbia (NTS 93N). *Geoscience BC, Report*, 3:1–74, 2011.

D. Helsel. *Statistics for Censored Environmental Data using Minitab and R*. Wiley, Hoboken, 2nd edition, 2012.

Humboldt State University. Conducting accuracy assessment, 2018. URL http://gis.humboldt.edu/OLM/Courses/GSP_216_Online/lesson6-2/accuracy.html.

D. Kahle and H. Wickham. ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013. URL https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

M. Kline. *Calculus: An Intuitive and Physical Approach*. Courier Corporation, 1998.

J. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278, 2003.

J. Martín-Fernández, J. Palarea-Albaladejo, and R. Olea. Dealing with zeros. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis: Theory and Applications*, pages 43–58. Wiley, Chichester, 2011.

J. A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics and Data Analysis*, 56:2688–2704, 2012.

M. Matthews. Importance of sampling design and density in target recognition. In D. Schumacher and M. Abrams, editors, *Hydrocarbon migration and its near-surface expressions*, volume 66, pages 243–253. AAPG Special Volumes, Tulsa, 1996.

M. Middleton, J. Torppa, P. R. Wäli, and R. Sutinen. Biogeochemical anomaly response of circumboreal shrubs and juniper to the Juomasuo hydrothermal Au-Co deposit in northern Finland. *Applied Geochemistry*, 98:141–151, 2018.

D. Mikšová, P. Filzmoser, C. Rieser, J. Melleton, S. Thaarup M., and M. Middleton. Application of the R codes on the existing and new geochemical data to produce interpretation results. 2019a.

D. Mikšová, C. Rieser, and P. Filzmoser. Identification of mineralization in geochemistry along a transect based on the spatial curvature of log-ratios. *arXiv*, (1912.02867), 2019b.

D. Mikšová, C. Rieser, P. Filzmoser, S. Thaarup, and J. Melleton. A method to identify geochemical mineralization on linear transect. *arXiv*, (2003.10268), 2020.

S. Millard, N. Neerchal, and P. Dixon. *Environmental Statistics with R.* CRC Press, Boca Raton, USA, 2nd edition, 2012.

F. Molnár, A. Middleton, H. Stein, H. O'Brien, Y. Lahaye, H. Huhma, L. Pakkanen, and B. Johanson. Repeated syn- and post-orogenic gold mineralization events between 1.92 and 1.76 Ga along the Kiistala Shear Zone in the Central Lapland Greenstone Belt, northern Finland. *Ore Geology Reviews*, 101:936–959, 2018.

J. Palarea-Albaladejo and J. Martín-Fernández. zcompositions – R package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, 2015.

J. Palarea-Albaladejo and J. A. Martín-Fernández. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computer & Geosciences*, 34(8): 902–917, 2008.

V. Pawlowsky-Glahn and J. J. Egozcue. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 15 (5):384–398, 2001.

V. Pawlowsky-Glahn, J. Egozcue, and R. Tolosana-Delgado. *Modeling and Analysis of Compositional Data.* Wiley, Chichester, 2015.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `http://www.R-project.org/`.

C. Reimann and R. G. Garrett. Geochemical background–concept and reality. *Science of the Total Environment*, 350(1-3):12–27, 2005.

C. Reimann, P. Filzmoser, R. Garrett, and R. Dutter. *Statistical Data Analysis Explained: Applied Environmental Statistics with R.* Wiley, Chichester, 2008.

C. Reimann, P. Englmaier, B. Flem, O. Eggen, T. Finne, M. Anderson, and P. Filzmoser. The response of 12 different plant materials and one mushroom to Mo and Pb mineralization along a 100-km transect in southern central Norway. *Geochemistry: Exploration, Environment, Analysis*, 18(3):204–215, 2018.

L. Robb. *Introduction to Ore-forming Processes*. Blackwell, Oxford, 2013.

P. Roshani, A. Mokhtari, and S. Tabatabaei. Objective based geochemical anomaly detection – Application of discriminant function analysis in anomaly delineation in the Kuh Panj porphyry Cu mineralization (Iran). *Journal of Geochemical Exploration*, 130:65–73, 2013.

P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, and M. Maechler. ***robustbase**: Basic Robust Statistics*, 2009. URL `http://CRAN.R-project.org/package=robustbase`. R package version 0.4-5.

R. Sanford, C. Pierson, and R. Crovelli. An objective replacement method for censored geochemical data. *Mathematical Geology*, 25(1):59–80, 1993.

W. Schnedler. Likelihood estimation for censored random vectors. *Econometric Reviews*, 24(2):195–217, 2005.

C. Stanley, N. O'Driscoll, and P. Ranjan. Determining the magnitude of true analytical error in geochemical analysis. *Geochemistry: Exploration, Environment, Analysis*, 10 (4):355–364, 2010.

A. Taivalkoski, J. Torppa, D. Mikšová, M. Middleton, and P. Sarala. Ultra low-impact exploration methods in the subarctic (UltraLIM) project: The analytical data of the soil samples processed with R. 2019. URL `http://tupa.gtk.fi/raportti/arkisto/90_2019.pdf`.

H. Talebi, U. Mueller, and R. Tolosana-Delgado. Joint simulation of compositional and categorical data via direct sampling technique – Application to improve mineral resource confidence. *Computers & Geosciences*, 122:87–102, 2019.

M. Templ, K. Hron, and P. Filzmoser. *robCompositions: Robust Estimation for Compositional Data.*, 2011. URL `http://CRAN.R-project.org/package=robCompositions`. R package version 1.5.0.

M. Templ, K. Hron, P. Filzmoser, and A. Gardlo. Imputation of rounded zeros for high-dimensional compositional data. *Chemometrics and Intelligent Laboratory Systems*, 155:183–190, 2016.

J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26 (1):24–36, 1958.

R. Tolosana-Delgado and K. G. van den Boogaart. Towards compositional geochemical potential mapping. *Journal of Geochemical Exploration*, 141:42–51, 2014.

J. Torppa and M. Middleton. Biogeochemical data analysis methods and R implementation in the UltraLIM project. Geological Survey of Finland, GTK archive report 8/2017. 2017. URL `http://tupa.gtk.fi/raportti/arkisto/8_2017.pdf`.

UpDeep. KAVA Reference: 16329, UpDeep project, Upscaling deep buried geochemical exploration techniques into European business, 2017–2020.

K. G. van den Boogaart and R. Tolosana-Delgado. "compositions": A unified R package to analyze compositional data. *Computers & Geosciences*, 34(4):320–338, 2008.

R. Webster and M. Lark. *Field Sampling for Environmental Science and Management*. Routledge, London, 2012.

S. N. Wood. *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.*, 2012. URL `https://CRAN.R-project.org/package=mgcv`. R package version 1.8.0.

S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton, USA, 2017.

S. N. Wood, M. V. Bravington, and S. L. Hedley. Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):931–955, 2008.

T. W. Yee. *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer, New York, 2015.

V. Yohai. High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics*, 15:642–665, 1987.

R. Zuo. Machine learning of mineralization-related geochemical anomalies: A review of potential methods. *Natural Resources Research*, 26(4):457–464, 2017.

Curriculum Vitae

Dominika Mikšová

Contact address

TU Wien
Institute of Statistics and Mathematical Methods in Economics
Research Group Computational Statistics (CSTAT)

Wiedner Hauptstraße 8-10
A-1040 Vienna, Austria
Email: dominika.miksova@tuwien.ac.at, miksovadominika1@gmail.com
Phone: +420 724113069

Education

since 07/2017	**TU Wien, Vienna (Austria)**
	Ph.D. candidate in Statistics
2014 – 2017	**Palacký University Olomouc (Czech Republic)**
	Masters in Applications of Mathematics in Economy
	Thesis: *Regression analysis using the partial least squares method*
2011 – 2014	**Palacký University Olomouc (Czech Republic)**
	Bachelor in Mathematics - Economics with focus on banking
	Thesis: *Hedging of the Interest and Currency Risk in Czech Firms and Fourier Analysis*
02/2015 – 07/2015	**Catholic University of Leuven (Belgium)**
	Erasmus program

Working experience

since 07/2017	**TU Wien, Vienna (Austria)**
	Project assistant (UpDeep project)
	Developing methods for exploration mineralization in geochemistry
2016–2018	**Transport Research Centre (Czech Republic)**
	Researcher and Data Scientist
	Sector: Road safety evaluation and strategies

Publications

D. Mikšová, P. Filzmoser, and M. Middleton. Imputation of values above an upper detection limit in compositional data. *Computers and Geosciences*. To appear. doi.org/10.1016/j.cageo.2019.104383

D. Mikšová, C. Rieser, and P. Filzmoser. Identification of mineralization in geochemistry along a transect based on the spatial curvature of log-ratios. *arXiv, (1912.02867)*, 2019.

D. Mikšová, C. Rieser, P. Filzmoser, S. M. Thaarup and J. Melleton. A method to identify geochemical mineralization on linear transects. Accepted in *Austrian Journal of Statistics*. To appear.

D. Mikšová, C. Rieser, P. Filzmoser, M. Middleton, and R. Sutinen. Identification of mineralization in geochemistry for grid sampling using Generalized Additive Models. Submitted.

B. Lemiere, J. Melleton, P. Auger, V. Derycke, E. Gloaguen, L. Bouat, D. Mikšová, P. Filzmoser, and M. Middleton. pXRF measurements on soil samples for the exploration of an antimony deposit: example from the Vendean antimony district (France). Submitted.