TECHNISCHE UNIVERSITÄT WIEN

DIPLOMARBEIT

Local Outlier Detection for Compositional Data

zur Erlangung des akademischen Grades

Diplom-Ingenieur/in

im Rahmen des Studiums

Statistik und Wirtschaftsmathematik

eingereicht von

Lorena Braus

Matrikelnummer 11905848

ausgeführt am Institut für Stochastik und Wirtschaftsmathematik
der Fakultät für Mathematik und Geoinformation der Technischen Universität Wien

Betreuung
Betreuer/in: Univ.Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser

Wien, 3. August 2023

(Unterschrift Verfasser/in) (Unterschrift Betreuer/in)

Abstract

This master thesis explores the application of outlier detection techniques on compositional data, focusing on local outlier detection methods. Compositional data, where relevant information is contained in ratios between the components, require specialized analysis approaches. The thesis begins with an introduction to compositional data and local outlier detection, providing a foundation for the subsequent chapters. The analysis of compositional data involves understanding their geometrical properties and employing the so-called Aitchison geometry on the simplex. Coordinate representations and preprocessing issues are also discussed, highlighting the challenges unique to compositional data analysis. The thesis delves into outlier detection, covering classical and robust statistical analysis techniques and emphasizing their significance in the context of compositional data. The main focus is on local outlier detection methods, specifically exploring two robust methods and the Local Outlier Factor (LOF) technique. The practical application of these methods is demonstrated using spatially dependent geochemical data obtained from the Geological Survey of Finland. The thesis provides a detailed description of the data and explains the necessary data preparation and cleaning steps. The four relevant methods are applied to identify outliers in the data. Furthermore, the identified outliers are analyzed and explained. The thesis concludes with a comprehensive evaluation and comparison of the applied methods, considering their overall effectiveness and performance. Parameters used in the analysis, including the parameter k for the k-nearest neighbors (kNN) method, are discussed to provide insights into their impact on the results.

Kurzfassung

Diese Masterarbeit erforscht die Anwendung von Ausreißererkennungstechniken auf Kompositionsdaten und legt den Fokus dabei auf lokale Ausreißererkennungsmethoden. Kompositionsdaten, bei denen relevante Informationen in den Verhältnissen zwischen den Komponenten enthalten sind, erfordern spezifische Analyseansätze. Die Arbeit beginnt mit einer Einführung in Kompositionsdaten und lokale Ausreißererkennung, um eine Grundlage für die folgenden Kapitel zu schaffen. Die Analyse von Kompositionsdaten beinhaltet das Verständnis ihrer geometrischen Eigenschaften und die Anwendung der sogenannten Aitchison Geometrie am Simplex. Die Darstellung der Koordinaten und Vorverarbeitung werden ebenfalls diskutiert und sie zeigen die speziellen Herausforderungen bei der Analyse von Kompositionsdaten. Die Arbeit vertieft die Ausreißererkennung und behandelt klassische und robuste statistische Analysetechniken, wobei ihre Bedeutung im Kontext von Kompositionsdaten hervorgehoben wird. Der Hauptfokus liegt auf lokalen Ausreißererkennungsmethoden, insbesondere auf der Erforschung zweier robuster Methoden und der Local Outlier Factor (LOF) Methodik. Die praktische Anwendung dieser Methoden wird anhand von geochemischen Daten mit einer räumlichen Komponente demonstriert, die vom Geological Survey of Finland erhalten wurden. Die Arbeit liefert eine detaillierte Beschreibung der Daten und erläutert die erforderlichen Schritte zur Datenbereinigung und -vorbereitung. Die vier relevanten Methoden werden angewendet, um Ausreißer in den Daten zu identifizieren. Darüber hinaus werden die identifizierten Ausreißer analysiert und erklärt. Die Arbeit schließt mit einer umfassenden Bewertung und einem Vergleich der angewendeten Methoden, unter Berücksichtigung ihrer Gesamtwirksamkeit und Leistung. Die verwendeten Parameter in der Analyse, einschließlich des Parameters k für die k-nearest neighbors (kNN) Methode, werden diskutiert, um Einblicke in ihre Auswirkungen auf die Ergebnisse zu geben.

Acknowledgement

I am deeply grateful for all the nice people that surround me with constant help and support. Now is the time to say thank you to some of them.

First and foremost, the biggest thank you goes to the dearest mathematicians in my life, my parents Lidija and Ivo, who have been my life-long source of love, support, and inspiration. Thank you for your belief in me, your constant encouragement and acceptance. Your unwavering faith in my abilities has given me the confidence to pursue my dreams fearlessly.

I would like to extend my heartfelt gratitude to all of my dear friends who have joined me in my academic journey. Your insights, perspectives, and willingness to share knowledge have broadened my understanding and helped shape the person I am today. From celebrating milestones to sharing all the good times, you have been an essential part of my life outside of academia.

I would like to especially thank Prof. Filzmoser for supervising my thesis, for excellent guidance and constant support during the whole research work. Your generosity and expertise have enriched my master thesis immensely. I also want to say many thanks to Patricia Puchhammer for introducing me into the work project and from the very beginning not letting any of my questions go unanswered.

Lorena Braus
Vienna, August 3, 2023

In mathematics, you don't understand things.
You just get used to them.
— John von Neumann

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, 3. August 2023

Lorena Braus

Contents

1 Introduction

Outlier detection belongs to the most important tasks of any statistical analysis. Either by being the most crucial part of the preprocessing and cleaning of the data, or actually being the main goal of our analytical work. The thesis at hand tackles the problem of outlier detection in, more specifically, multivariate and compositional data with the emphasis on local outlier detection, taking into account the data that additionally have a spatial dependency. The following two subsections should give the reader a short introduction into compositional data itself and (local) outlier detection in general.

1.1 What are compositional data?

Compositional data can be simply defined as multivariate observations with positive values that sum up to some constant. They were traditionally seen as constrained data, where the compositional parts sum up to some constant, e.g. to 1 for proportions, or to 100 for percentages. However, what happens if one part of the multivariate composition cannot be measured? Or what happens if we have to round some values so that the sum is not what we initially expected? What if the sum of each observation is different? Because of all the reasons above, it is more useful to see compositional data as data carrying some important relative information between the parts. Compositional data are treated as multivariate observations where relative rather than absolute information is relevant.

Absolute information refers to the data where usual operations in real Euclidean space can be performed. Any rescaling of the original raw data leads to losing information.

Relative information refers to the contributions as a whole. The sum or the total amount is irrelevant. All we care about are the ratios between the components (parts) of a composition.

For compositional data, units like mg/kg, ppm and percentages are expected. One can see that mg/kg or ppm (usually measuring concentrations of chemical elements) impose what the sum of the observations should be, which does not have to be fulfilled in practice.

Another example can be household expenditure or election votes per region. It is clear that in those cases we can expect the total amount (sum) of each observation to be different.

The relevant information is somehow contained in the ratios between the components. One can express the data relative to the sum of each observation. In that case we get a representation in terms of proportions or percentages. Another idea is to express the ratios between the components. One can see that ratios contain much more detailed information than just percentages to the total. They will thus form the representation of relative information that is considered in compositional data analysis. In order to symmetrize the interpretation of ratios, the first choice is to use logarithms of the ratios (log-ratios). Logarithms are easier to handle from a mathematical point of view. In that case the balance is represented by 0.

One can see that log-ratios have problems when it comes to zeros. Zero in the denominator of the ratios lead to infinity, and zero in the numerator will result in minus infinity when we take the logarithm. This is the reason that compositional data are defined with strictly positive values. Any zero components should be specially treated.

The history and interest in compositional data analysis goes back to the 19^{th} century and to the famous statistician Karl Pearson (Pearson, 1897). During the 20^{th} century, the developments were focused on building statistical models to analyze proportional data, or to cope with restrictions resulting from their direct statistical analysis, until in the early 1980s, the Scottish mathematician John Aitchison finally introduced the logratio methodology (Aitchison, 1982).

1.2 What is (local) outlier detection?

The identification of outliers is probably the most important task in any data analysis. Outlier detection (also referred to as anomaly detection) can be understood as the identification of (rare) observations or events that significantly differ from the majority of data and do not "behave normally" or do not fit the wanted structure. Outlier detection finds application in many domains because outliers are widely present in real data sets. Nevertheless, outliers are usually the most interesting observations because they come from some atypical phenomenon that one wants to study and understand.

One can see the "normal behaviour" of the data as following some distribution. The observations not "behaving normally" could then be the observations coming from some other underlying distribution, and thus they shall be recognized.

Compositional data frequently contain outliers because of data inconsistencies, rounding effects, dependencies among the observations, etc. One might want to separate "interesting" outliers from noise or irrelevant outliers.

Many classical statistical methods rely on strict model assumptions, like independence or (multivariate) normal distribution, and violations of the assumptions can lead to biased results. It is preferred to apply robust methods to classical methods in practice. Robust statistics tolerates certain deviations from strict model assumptions. It provides alternatives to classical estimators such as the sample mean or the sample (co-)variance that are highly sensitive to outliers.

The term "local outlier" refers to the setting where observations are grouped in some kind of neighbourhoods. For example, the grouping can be defined according to additional information given by the spatial coordinates of the observations. In that case, local outliers differ from their spatial neighbors. This means that a local outlier does not have to be outlying in a global sense.

Since compositional data are from its nature multivariate, the focus here is on identifying multivariate local outliers.

2 Compositional Data Analysis

The aim of this section is to present and define compositional data and compositional data analysis in a mathematical way. First, we start by introducing special geometrical properties of compositional data, and eventually its own new geometry, called Aitchison geometry. Some specific coordinate representations are defined in Section 2.2. These representations are often simply called transformations, which are applied to the raw data before some standard statistical methodology can be used. In Section 2.3, there is a short reminder of the potential preprocessing issues, specific for compositional data.

2.1 Geometrical properties and Aitchison geometry

For introducing the geometrical properties of compositional data we start by defining the main metrical concepts of the Euclidean space. The Euclidean space is associated with a vector space structure, where we define the scalar product, norm and distance between the two vectors $\mathbf{x} = (x_1, \ldots, x_p)'$ and $\mathbf{y} = (y_1, \ldots, y_p)'$:

$$< \mathbf{x}, \mathbf{y} >= \sqrt{\sum_{i=1}^{p} x_i y_i}$$

$$\|\mathbf{x}\| = \sqrt{< \mathbf{x}, \mathbf{x} >} = \sqrt{\sum_{i=1}^{p} x_i^2}$$

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

Compositional data are carrying relative information between their compositional parts. First let us assume that the compositional parts sum up to some constant κ. Then our sample space is defined as a D-part simplex S^D:

$$S^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)' \in \mathbb{R}^D : x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\}$$

Graphical illustrations of the 2-part and 3-part simplex are shown in Figure 2.1.

We can easily notice that in real world examples, it does not always happen that all the collected observations sum up to the same constant. Moreover, maybe we would want them to all sum up to the same constant (for example, to 1) so we can compare them. In other words, we can show that the sum does not matter because we can always rescale the composition without changing the relative information of the parts. For that reason we define the closure operator:

$$C_\kappa(\mathbf{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right)'$$

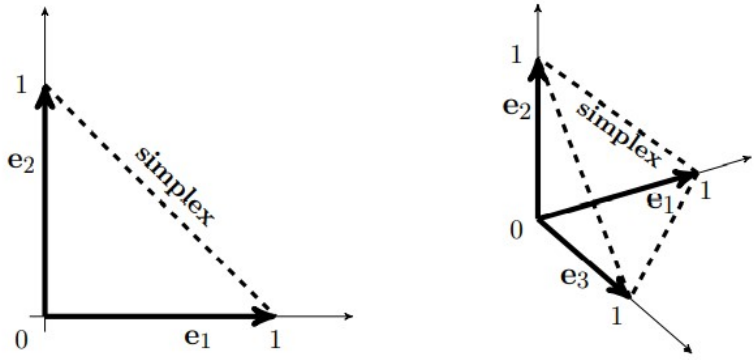

Figure 2.1: 2-part simplex in $\mathbf{R}^2$ (left), and 3-part simplex in $\mathbf{R}^3$ (right), shown by the dashed lines

After applying the closure operator, the compositional parts sum up to κ. Now we say that two compositions $\mathbf{x}$ and $\mathbf{y}$ are **compositionally equal** if $C_\kappa(\mathbf{x}) = C_\kappa(\mathbf{y})$. Figure 2.2 shows visualizations of compositions which are rescaled to sum 1.

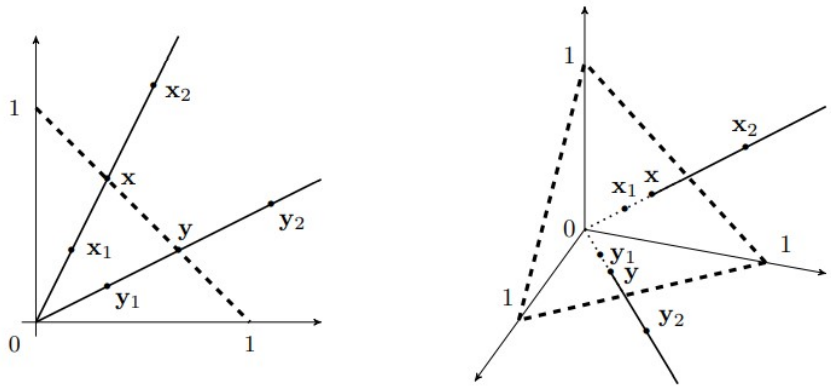

Figure 2.2: In both cases, the compositions $\mathbf{x}, \mathbf{x}_1$ and $\mathbf{x}_2$, as well as the compositions $\mathbf{y}, \mathbf{y}_1$ and $\mathbf{y}_2$ are compositionally equal

Since rescaling the composition does not change the relative information, the new definition of the sample space is given by:

$$\tilde{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)' \in \mathbb{R}_+^D \mid x_i > 0, \forall \kappa > 0 \; \exists! \lambda > 0 : \mathbf{x} = \lambda C_\kappa(\mathbf{x}) \right\}$$

In other words, $\tilde{S}^D$ refers to equivalence classes of compositionally equivalent vectors.

Compositional data do not follow the usual Euclidean geometry. The aim is to define a vector space structure of the previously defined simplex. Consider two compositions $\mathbf{x}$ and $\mathbf{y}$ from the simplex sample space $\tilde{S}^D$. Then the **perturbation** is defined as:

$$\mathbf{x} \oplus \mathbf{y} = (x_1 y_1, \ldots, x_D y_D)'$$

The **powering** is defined as:

$$\alpha \odot \mathbf{x} = (x_1^\alpha, \ldots, x_D^\alpha)'$$

By applying perturbation and powering, it is also possible to define the **perturbation difference** as:

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus [(-1) \odot \mathbf{y}] = (\frac{x_1}{y_1}, \ldots, \frac{x_D}{y_D})'$$

A Euclidean vector space structure can be obtained by defining norm, inner product, and distance in the Aitchison sense:

$$< \mathbf{x}, \mathbf{y} >_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

$$\|\mathbf{x}\|_A = \sqrt{< \mathbf{x}, \mathbf{x} >_A} = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} (\ln \frac{x_i}{x_j})^2}$$

$$d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} (\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j})^2}$$

The definitions are based on logarithms of ratios (logratios) between the compositional parts. Indeed, one can see that the sum of the compositional parts is irrelevant. For example, the compositions $\mathbf{x} = (x_1, \ldots, x_D)'$ and $\mathbf{x}_\lambda = (\lambda x_1, \ldots, \lambda x_D)'$ lead to the same logratios, for any $\lambda > 0$.

2.2 Coordinate representation

Compositional data analysis is associated with applying an appropriate transformation and then using the standard statistical methodology on the transformed data. A transformation can also be viewed as expressing the compositions in a coordinate system with respect to the Aitchison geometry. As one then works just with the transformed data, the interpretation of the results should also be adapted. Here, the emphasis is on the so-called *logratio* coordinates, driven by the Aitchison geometry. It is also possible to get back to the original compositional data – up to a scaling factor.

Additive Logratio (alr) Coordinates

Taking the alr coordinates is a mapping from $\tilde{S}^D$ to $\mathbb{R}^{D-1}$:

$$\mathbf{x}^{(j)} = alr_j(\mathbf{x}) = \left(\ln \frac{x_1}{x_j}, \ldots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \ldots, \ln \frac{x_D}{x_j} \right)'$$

One can notice that prior to taking the alr coordinates, one must choose one of the variables (x_j) to be the ratioing one. This is already one disadvantage of alr coordinates because the choice would depend on the context of the data or suitability for visualizations and data exploration.

However, the alr coordinates move the operations of perturbation and powering to the standard vector addition and multiplication, but this is not fullfilled for the Aitchison inner product, norm and distance. e.g. $< \mathbf{x}, \mathbf{y} >_A \neq < alr_j(\mathbf{x}), alr_j(\mathbf{y}) >$. In other words, the alr coordinates do not build an isometry between Aitchison and Euclidean space!

Moreover, it would be wrong to interpret the alr coordinates in terms of the original parts, e.g., the first coordinate $\ln \frac{x_1}{x_j}$ contains relative information of x_1 only to the j-th part, but not to all the other parts. Thus, the interpretation of a particular coordinate cannot be made in terms of one part.

Nevertheless, alr coordinates are mentioned rather for historical reasons.

Centered Logratio (clr) Coefficients

A composition $\mathbf{x} \in \tilde{S}^D$ is expressed by a vector $\mathbf{y} \in \mathbb{R}^D$, with:

$$\mathbf{y} = (y_1, \ldots, y_D)' = clr(\mathbf{x}) = \left(\ln \frac{x_1}{g_m(x)}, \ldots, \ln \frac{x_D}{g_m(x)} \right)'$$

$$g_m(\mathbf{x}) = \sqrt[D]{\prod_{k=1}^{D} x_k} = \exp\left(\frac{1}{D} \sum_{k=1}^{D} \ln x_k \right)$$

The denominator $g_m(\mathbf{x})$ is called **geometric mean**.

Clr avoids the subjectivity of the choice of the denominator and treats the components symmetrically. Moreover, the clr coefficients represent an isometry: all metric concepts in the simplex are maintained after taking the clr coefficients. This fact is important since then the Euclidean distance after the transformation is equal to the Aitchison distance on the simplex, which would be of great importance if we want to apply any outlier detection method described later in this thesis (see Section 3).

The components sum up to 0, so one ends up with constrained data. Nevertheless, it is easy to notice the geometrical peculiarity that there is one more composition than necessary to form the basis in the Aitchison geometry, being the dimension of $D-1$. Therefore, $\mathbf{y}$ represents coefficients with respect to a generating system (instead of a basis) and we refer to clr *coefficients* (instead of coordinates).

This means that there is not a unique possibility how to form coefficients with respect to the same system of compositions. Thus, it is not possible to consider

just one of the clr coefficients for the analysis without taking also the others into account.

Putting the compositions in a data matrix, after expressing each observation in clr coefficients, the resulting matrix has not full rank in the columns and the corresponding covariance matrix is singular which can be a big problem when one needs to calculate covariance matrix for outlier detection techniques (see Section 3.2).

We can write the coefficient y_1 as:

$$y_1 = \frac{1}{D}\left(\ln\frac{x_1}{x_2} + \ldots + \ln\frac{x_1}{x_D}\right).$$

Thus, the logratios of part x_1 to all other parts are involved and each logratio contributes with the same weight to the first coefficient y_1. That clearly shows that among different coefficients there is "overlap" of information. This problem will be solved by introducing the isometric logratio coordinates (pivot coordinates).

Another interesting fact is that, for example, y_1 can be written as

$$y_1 = \ln x_1 - \frac{1}{D}\sum_{k=1}^{D} \ln x_k,$$

which means that the observations are actually just log transformed and *centered*.

Isometric Logratio (ilr) Coordinates

There are more ways of building an orthonormal basis in the hyperplane formed by clr coefficients. One specific way proposed here are the so-called ilr coordinates (Filzmoser et al., 2018).

One particular proposal of the chosen basis is:

$$ilr(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$$

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}$$

The scaling constants guarantee orthonormality of the coordinate system. By taking the ilr coordinates, one ends up with the full rank in the data matrix. However, the interpretation of the coordinates is not as simple. That is why one part (here x_1) is set to be a pivot and it is contained just in the first coordinate. Thus, the first coordinate has a special meaning and interpretation property. This is the reason why ilr coordinates are also referred to as **pivot coordinates**. The first coordinate expresses the level of dominance of part x_1 with respect to all the other parts and this is not the case for other parts.

So, the first coordiante z_1 contains all relative information about x_1. In other words, this coordinate expresses the level of dominance of part x_1 with respect to the other parts "on average". For positive values of z_1, the first part dominates in the composition and vice versa for negative values. $z_1 = 0$ indicates a balanced state between x_1 and an average behavior of the other parts.

It is important to remember that no other part can be interpreted in such a manner. Note also that the first pivot coordinate z_1 and the first clr coefficient y_1 are proportional up to a scaling factor depending just on a dimension D. Thus, also y_1 can be interpreted like z_1 in terms of the relative dominance of x_1 in the composition. However, $z = (z_1, z_2, \ldots, z_{D-1})'$ are coordinates of an orthonormal basis which is not the case for the clr coefficients. The clr transformation is, thus, leading to singularity of the covariance matrix, which would be a problem for further outlier detection techniques (Filzmoser and Gregorich, 2020).

As mentioned previously, the clr coefficients map a composition $\mathbf{x}$ to a hyperplane: $H : y_1 + \ldots + y_D = 0$, i.e., to a subspace of $\mathbb{R}^D$. Then the ilr coordinates are formed by coefficients expressing $\mathbf{x}$ in an orthonormal basis of this hyperplane. If we denote the $D \times (D-1)$ matrix, containing all the orthonormal basis vectors as columns, as $\mathbf{V}$, one immediately gets the relation between clr coefficients and pivot coordinates as

$$\boldsymbol{y} = \boldsymbol{V}\boldsymbol{z}.$$

Finally, like clr coefficients, also ilr coordinates represent an isometry, which is important for further outlier detection techniques, where one does not want to deform the distance between the samples by taking the coordinates.

Generalization of Pivot Coordinates. It can be of interest to obtain a specific interpretation for some other part within a composition. One can simply permute the compositional parts in a way that the part of interest is placed at the first position:

$$\mathbf{x}^{(l)} = (x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)' =: (x_1^{(l)}, \ldots, x_D^{(l)})'$$

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

Another choice of the pivot coordinate system is just a rotation of the original one.

Symmetric Pivot Coordinates. If one is interested in the relation between two compositional parts, a possible alternative is to construct such coordinates that would treat the dominance of both parts symmetrically.

We would then construct the coordinates denoted as: $z_1^{(1,2)}$ and $z_2^{(1,2)}$ (for the exact definition, see Filzmoser et al., 2018), which can be used for bivariate analysis.

Balances. It can be of interest to interpret the behavior of (non-overlapping) groups of compositional parts. Such coordinates are called **balances** and their construction is called **sequential binary partitioning** (SBP) (Filzmoser et al., 2018).

2.3 Preprocessing issues

While working with compositional data sets, we will not always be able to apply needed transformations and statistical methods immediately and there will be a need for special preprocessing.

Since we are working with logratios, the presence of missing values in some compositional parts, but also zero values, should be handled beforehand. Zeros are basically excluded from the definition of compositional data. According to the nature of zero values, rounded, count and structural zeros need to be considered.

Rounded zeros occur when either small values are rounded to zeros, or a measuring device/tool has a detection limit that automatically sets values below this limit to zero. Count zeros result similarly from insufficient sample size. For the imputation of missing values, rounded and count zeros, model-based algorithms have been developed. On the other hand, structural zeros are a result of a structural process, and thus imputing them to obtain a full data set is not meaningful.

Values below the detection limits are often set to zero, to half of the value of the detection limit, or to the negative value of the detection limit. The meaning, however, remains the same: the value cannot be measured and thus we can consider it as a zero. It is also possible that some parts can be measured with higher precision than others. If the percentage of data with values under the detection limit is not too high (less than 10%), the simplest solution is to impute zeros with $\frac{2}{3}$ of the detection limit (DL). This minimizes the distortion of the covariance structure but the univariate character of this imputation ignores the multivariate complexity of the compositional data and multivariate imputation methods outperforms it (Filzmoser et al., 2018).

Structural zeros are the most challenging type of zeros for compositional data preprocessing, since that means that a component is truly zero. Their replacement might lead to artificial outliers which becomes a serious problem for the (outlier detection) analysis. One way to deal with them is to simply consider two populations, one with zero values and one with some strictly positive values for a specific component (Aitchison and Kay, 2003). Another way is to use amalgamation (Aitchison, 1982), that aims to aggregate the values of parts containing zero values to such part(s) that never have this effect, but are thematically related. For more insights and tool propositions see also Templ et al. (2017).

Due to the relative nature of the compositional data, there are some peculiarities that are also worth to be considered. Particularly, observations are rarely measured in terms of ratios and almost always the absolute values are produced. We would maybe want to add a certain positive value to avoid negative and/or zero values or they are resulting from calibration of the measurement device. Clearly, this adjustment could completely destroy the source information. The influence can be particularly severe for small (absolute) concentrations.

The topic of preprocessing data for compositional data analysis is discussed in more detail in Filzmoser et al. (2018), with proposed algorithms that deal with zeros and missing values.

3 Outlier Detection

This section is aimed at explaining common tools and techniques of outlier detection. At the start of any statistical data analysis it is usually needed to make some assumptions of the underlying distribution of the data at hand and to estimate the parameters. Section 3.1 approaches the difference between robust and classical statistical estimates. In Section 3.2, the simplest outlier detection methods are introduced for univariate and multivariate data, but also the importance of robustness in outlier detection is emphasized.

3.1 Classical and robust statistical analysis

Compositional data, like any other data collected from practical applications, can contain outliers, inconsistencies, measurements errors, etc. By applying statistical methods we would not like those violations to give us biased results. That is why robust statistics was developed. The basic idea behind robust statistical methods is to fit a model to the majority of data points, and not to satisfy every single data point. The point is that a robust estimator does not change arbitrarily in presence of contamination.

There are different measures of robustness. One of them is the *influence function*. For example, the arithmetic mean can go towards infinity if just one observation is moved arbitrarily far away from the data set, while the median remains more stable.

Another measure is the *breakdown point*. Intuitively, one can say that it measures the minimal fraction of arbitrary contamination that drives the estimator beyond all bounds or, in other words, causes the estimator to yield an arbitrarily large result. Seheult et al. (1989) offers a more precise definition: Let T be an estimator and $X = (x_1, \ldots, x_n)$ any sample of n data points. Now, let us consider all possible "corrupted" samples X' that are obtained by replacing any m of the original data points by arbitrary values. Let us denote by $bias(m; T, X)$ the maximum bias that can be caused by such a contamination:

$$bias(m; T, X) = sup_{X'}||T(X') - T(X)||,$$

where the supremum is over all possible X'. If $bias(m; T, X)$ is infinite, this means that m outliers can have an arbitrarily large effect on T, which may be expressed by saying that the estimator "breaks down." Therefore, the (finite-sample) breakdown point of the estimator T at the sample X is defined as:

$$\epsilon_n^*(T, X) = min\left\{\frac{m}{n} : bias(m; T, X) = \infty\right\}.$$

For the arithmetic mean the breakdown point equals $\frac{1}{n}$, which tends to zero for increasing sample size n, so we say that the asymptotic breakdown point is 0. On the other hand, the breakdown point of the median is 0.5, which is the highest possible value.

Univariate Location and Scale

Let us take a look at a simple example, the classical estimator of the location parameter of a normal distribution, i.e. at the **arithmetic mean**. A robust alternative is the **median**, which is given as the innermost value of the sorted data. Although, as already seen, it has very good robustness properties, the median has very low efficiency. Increasing the efficiency in general, requires using more of the available data information. There is also the **trimmed mean** as a compromise between good robustness and efficiency, with a tuning parameter $\alpha \in (0, 0.5)$ that regulates the amount of trimming.

We are still considering a normal distribution $N(\mu, \sigma^2)$. The parameter σ refers to the scale of the distribution. The classical scale estimator for given data $x_1, \ldots, x_n$ is the **empirical standard deviation**

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the arithmetic mean.

In the case of the scale estimator, a robust alternative is the **median absolute deviation**, defined as

$$s_{MAD} = 1.4826 \cdot median_i|x_i - \tilde{x}|,$$

where $\tilde{x}$ denotes the median of the sample, and the factor 1.4826 makes the estimator consistent under normality.

Another possible robust scale estimator is the **interquartile range**,

$$s_{IQR} = 0.7413 \cdot (q_{0.75} - q_{0.25}),$$

where q_k denotes the k-quantile.

Multivariate Location and Covariance

Let us now consider multivariate (non-)compositional observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$, which form the rows of the $n \times p$ data matrix $\boldsymbol{X}$. In the following, these observations will typically be the compositions, expressed in one of the introduced coordinates.

The classical estimators for location and covariance are the arithmetic mean vector $\bar{\boldsymbol{x}}$ and the sample covariance matrix $\boldsymbol{S}_{\boldsymbol{x}}$, defined as

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$$

$$\boldsymbol{S}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'.$$

Both estimators' breakdown point is zero. Thus, they are both highly sensitive to outliers and we would like to rather use robust alternatives. Nevertheless, the classical location and covariance estimators are *affine equivariant*, which means that they transform properly under affine transformations. An affine transformation of the data matrix $\boldsymbol{X}$ is given by a non-singular $p \times p$ matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b}$ of length p as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A} + \mathbf{1}_n \boldsymbol{b}'$$

So, $\boldsymbol{Y}$ as outcome is any shifted, rotated and rescaled version of $\boldsymbol{X}$. Thus, estimators of location $\boldsymbol{t}$ and covariance $\boldsymbol{C}$ are called **affine equivariant** if they satisfy

$$\begin{aligned} \boldsymbol{t}(\boldsymbol{Y}) &= \boldsymbol{t}(\boldsymbol{X})\boldsymbol{A} + \boldsymbol{b}, \\ \boldsymbol{C}(\boldsymbol{Y}) &= \boldsymbol{A}' C(\boldsymbol{X})\boldsymbol{A}. \end{aligned}$$

The most popular robust estimator of location and covariance is the **minimum covariance determinant (MCD)** estimator. It is defined by that subset $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_h}$ of h observations whose sample covariance matrix has the smallest determinant among all possible subsets of size h. Denote this index set of the resulting h observations by I_h. The MCD location estimator $\boldsymbol{t}_{MCD}$ is given by the arithmetic mean of those h observations, and the MCD covariance estimator $\boldsymbol{C}_{MCD}$ by their sample covariance, multiplied by a factor c_{MCD} for consistency,

$$\begin{aligned} \boldsymbol{t}_{MCD} &= \frac{1}{h} \sum_{i \in I_h} \boldsymbol{x}_i \\ \boldsymbol{C}_{MCD} &= c_{MCD} \cdot \frac{1}{h-1} \sum_{i \in I_h} (\boldsymbol{x}_i - \boldsymbol{t}_{MCD})'(\boldsymbol{x}_i - \boldsymbol{t}_{MCD}). \end{aligned}$$

The number h refers to the *data majority*. It can be taken as an integer in the interval $[(n+p+1)/2, n]$. The highest breakdown point is achieved for the smallest value of h, but this also leads to low efficiency. In practice, the compromise is to take $h \approx 0.75 \cdot n$.

The MCD estimators $\boldsymbol{t}_{MCD}$ and $\boldsymbol{C}_{MCD}$ are also affine equivariant which will be important for outlier detection methods with compositional data, because we do not want to depend on the chosen basis of (ilr) coordinates and the potential outliers will remain identical, no matter what transformation of the data was used (Filzmoser and Gregorich, 2020).

From the computational side, it is important to mention the routine tool for computing MCD estimates, **FAST-MCD algorithm**, which efficiently computes the exact MCD for smaller data sets and gives satisfying accurate estimates for larger data sets (Rousseeuw and Driessen, 1999).

The limitation is that the MCD estimator does not work for data sets with more variables than observations, because the determinant of the covariance matrix of any

subset would always be zero. To ensure the positive definiteness of the estimated covariance matrix, the **regularized MCD estimator** can be used. The basic idea is to replace the subset-based covariance by a regularized covariance estimate as seen in Fritsch et al. (2011) and Ernst and Haesbroeck (2017). Here, a ridge regularization is considered and one wants to keep the good breakdown point properties of the MCD estimator.

The regularized MCD estimator is obtained by the maximization of the penalized negative log-likelihood function restricted to a subset of $\frac{n}{2} \leq h \leq n$ observations, given by the index set I_h:

$$(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) = \arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \left(\log|\boldsymbol{\Sigma}| + \frac{1}{h} \sum_{i \in I_h} (\boldsymbol{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) + \lambda \; tr(\boldsymbol{\Sigma}^{-1}) \right)$$

Two parameters are to be considered and tuned: the *coverage* h and the *regularization* parameter λ.

1. The coverage is chosen according to the breakdown point of the robust estimator one wants to achieve by taking $h = \lceil n \times (1 - \alpha) \rceil$ where $0 < \alpha < \frac{1}{2}$ is the chosen breakdown value.

2. The regularization parameter equal $\frac{tr(\boldsymbol{\Sigma})}{np}$ would yield an unbiased estimation of the trace of the covariance matrix.

As explained in Fritsch et al. (2011), the FAST-MCD algorithm from Rousseeuw and Driessen (1999) can be adopted to compute the regularized version of the MCD estimator as well.

3.2 Univariate and multivariate outliers

In the last subsection, we saw how to use robust counterparts instead of classical estimators so that the always-present data outliers do not spoil our statistical analysis. In general, we do not want to remove data outliers from our dataset. We want to find them and analyze them. Even more, they are usually the most interesting observations, because some other atypical phenomenon is responsible for it.

Univariate outliers

If we are assuming, for example, that some univariate data are following a normal distribution with certain parameters, where are the possible outliers coming from? They could come from a different distribution completely, or from a normal distribution with different mean and/or variance.

Thus, under the assumption of normality, the regular observations are generated from some normal distribution $N(\mu, \sigma^2)$, where mean μ and variance σ^2 are unknown. So, we expect that the "inner" 95% of the data are in the interval

$$[\mu - 1.96 \cdot \sigma, \mu + 1.96 \cdot \sigma].$$

The left boundary corresponds to quantile $Q_{0.025}$ and the right one to $Q_{0.975}$ of the distribution. In other words, 5% of the data would fall outside of the interval and we could see them as "unusual" observations, or simply as outliers. Nevertheless, in practice it is unclear if those outliers are generated from a completely different distribution, or they are just located on the extremes of the current distribution.

Going back to the point of the last subsection, we already assume that the outliers are present in the sample and they would potentially have an effect on the classical estimators, and consequently on the interval boundaries.

We can prevent that by using robust alternatives from the last subsection. So, the robust interval looks like this,

$$[\tilde{x} - 1.96 \cdot s_{MAD}, \tilde{x} + 1.96 \cdot s_{MAD}],$$

where $\tilde{x}$ denotes the median and s_{MAD} the median absolute deviation.

Another possible robust alternative is

$$[\tilde{x} - 1.96 \cdot s_{IQR}, \tilde{x} + 1.96 \cdot s_{IQR}],$$

where s_{IQR} denotes the interquantile range.

It is important to notice that for a normally distributed sample with very large n, the resulting intervals will be identical to the one with classical estimators, $[\bar{x} - 1.96 \cdot s, \bar{x} + 1.96 \cdot s]$.

A fourth possibility for identifying outliers is the **Tukey boxplot**. Figure 3.1 shows three examples visually. In this case, outliers are the observations outside of the interval

$$[Q_{0.25} - 1.5 \cdot IQR, Q_{0.75} + 1.5 \cdot IQR],$$

with the *interquantile range* $IQR = Q_{0.75} - Q_{0.25}$.

One can adjust the factor 1.5 of the boxplot fences to the needs of the distribution. For example, if the distribution is rather skewed, the **adjusted boxplot** can be used (Hubert and Vandervieren, 2008).

Multivariate outliers

Multivariate outliers are harder to find than univariate ones, since we can no more rely on a visual inspection of the data. Moreover, they are not any more extremes along one coordinate, but could be anywhere in the whole multivariate space. Statistical theory of multivariate outlier detection is based either on univariate projection of the multivariate data or on the estimation of the empirical covariance structure to assign to each multidimensional data point a distance to the centre of the bulk of the data (Filzmoser and Gregorich, 2020).

In this section, it is assumed that the multivariate data are already compositions. It is common to first express the data in the compositional space with some representation in logratios coordinates, and then apply the multivariate outlier detection methods.

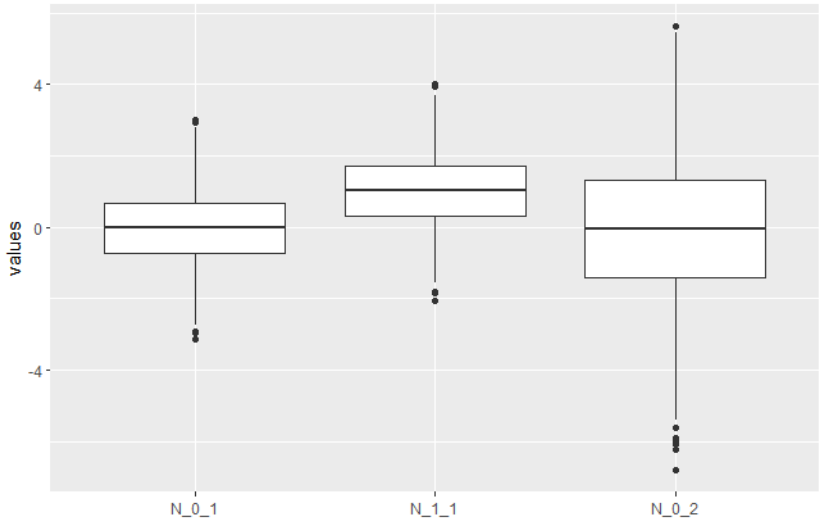

Figure 3.1: The boxplots of samples of 1000 generated values from normal distributios $N(0,1), N(1,1)$ and $N(0,4)$.

However, there are a few points to think about while detecting outliers in compositional data. First, it is important to realize what are the sources of outlyingness in the compositional data, for example which parts are responsible for having deviating logratios. We also want to have the methods that are able to discover outliers independently of the choice of the coordinate system (coordinate representation). Another point is that the parts with low concentrations will tend to produce more outliers. This is closely connected to approaching the detection limits of the measurement devices, where the precision of measurements is lower.

Suppose we have a compositional data set $\mathbf{X}$, with n observations and D compositional parts. Representing the data in ilr coordinates results in an $n \times (D-1)$ matrix $\mathbf{Z}$ with observations $\mathbf{z}_1, \ldots, \mathbf{z}_n$.

Following the standard multivariate outlier detection procedure, we have to assume that the observations of $\mathbf{Z}$ are generated by a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

For the location estimator $\mathbf{t}$ and the covariance estimator $\mathbf{C}$, the squared **Mahalanobis distance** is defined as

$$MD(\mathbf{z}_i)^2 = (\mathbf{z}_i - \mathbf{t})' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{t}), \text{ for } i = 1, \ldots, n.$$

For robustigying the Mahalanobis distance, the robust MCD estimates can be used for $\mathbf{t}$ and $\mathbf{C}$. The key strength of the robust MCD estimator is the property of affine equivariance. As a result of this property, the robust Mahalanobis distance also remains unaffected under affine transformations (Filzmoser and Gregorich, 2020).

The squared Mahalanobis distances are approximately following a χ^2 distribution with $D-1$ degrees of freedom, χ^2_{D-1}. Thus, similar as in the univariate case, it is possible to use the 0.975 quantile, $\chi^2_{D-1;0.975}$ as a cut-off value for identifying outliers. The observations $\mathbf{z}_i$ with

$$MD(\mathbf{z}_i)^2 > \chi^2_{D-1;0.975}$$

are marked as multivariate outliers.

As already mentioned, the reason for outlyingness of observations is usually unclear. There are, however, some tools (plots) proposed in the literature that help with the interpretation of multivariate outliers (Filzmoser et al., 2012).

4 Local Outlier Detection Methods

In the next sections, three different local outlier detection techniques are described.

In the previous section, the tools for the outlier detection (for compositional data) were introduced. In general, we were talking about *global* outlier detection. Now, let us assume that our (compositional) data has some spatial component. Meaning, that each observation has a pair of spatial coordinates. Thus, identifying global outliers means identifying outliers overall in the whole data set without considering the spatial dependency. Nevertheless, finding *local* outliers could be of interest in many fields where the data have a spatial component, for example environmental data. That means that those outliers are outliers in some defined neighborhood.

As a result, we should keep in mind that we are, on one side, dealing with the spatial space (spatial coordinates) and, on the other side, with the variable space (compositional space).

A local outlier can, but does not have to be a global outlier. Every global outlier usually is a local outlier, but it does not have to be. It could be also of interest to compare outliers' local and global nature.

4.1 Robust local outlier detection technique

At the end of the last section, the most common measure of outlyingness was introduced, the Mahalanobis distance. This distance measure in the variable space assigns to each observation the distance from the "centre" of the multivariate data set. As already emphasized several times throughout this work, it is crucial how one estimates the location and the covariance matrix. Thus, robust estimates have to be used.

If we define some cut-off value for the Mahalanobis distance (for example, the 0.975 quantile, $\chi^2_{D-1;0.975}$), then all the observations $\mathbf{z}_i$ with $MD(\mathbf{z}_i)^2 > \chi^2_{D-1;0.975}$ are marked as multivariate outliers. Those are considered global outliers, since no spatial dependency was taken into account.

This method relys on comparing the so-called *pairwise* Mahalanobis distances in the spatially defined neighborhoods (with respect to pairwise Euclidean distances in the spatial/geographical space).

For the coordinates (geographical locations) $(\boldsymbol{s_i}, \boldsymbol{s_j})$, let us consider the **Euclidean distance**:

$$ED(\boldsymbol{s_i}, \boldsymbol{s_j}) = \sqrt{(\boldsymbol{s_i} - \boldsymbol{s_j})'(\boldsymbol{s_i} - \boldsymbol{s_j})}.$$

In the following, the geographical location (spatial dependency) will always be a pair of coordinates, meaning $\boldsymbol{s_i}$ is a vector of length 2.

For the two multivariate observations $\mathbf{z}_i$ and $\mathbf{z}_j$, the **pairwise Mahalanobis distance** is defined as

$$MD(\mathbf{z}_i, \mathbf{z}_j) = \sqrt{(\mathbf{z}_i - \mathbf{z}_j)'\mathbf{C}^{-1}(\mathbf{z}_i - \mathbf{z}_j)},$$

where $\mathbf{C}$ is the robust MCD estimator of the covariance matrix of the data set.

Let us consider a sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of i.i.d. random vectors of dimensions D following a Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu} \in \mathbf{R}^D$ and covariance $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a $D \times D$ symmetric positive definite matrix.

It is known that the $MD(\mathbf{z}_i)^2$, $i = 1, \ldots, n$, are i.i.d. and follow a χ^2_D distribution. Moreover, the conditional distribution of the pairwise squared Mahalanobis distances $MD(\mathbf{z}_i, \mathbf{z}_j)^2$, $j = 1, \ldots, n$, given $\mathbf{z}_i$, is a non-central χ^2 distribution with D degrees of freedom and non-centrality parameter $MD(\mathbf{z}_i)^2$ (Filzmoser et al. (2014)).

Using this fact, it is possible to define the outlyingness in a local sense. Since local outliers are supposed to be different from their neighbors, one could define a quantile of the non-central chi-square distribution that can be used as a cut-off value for the pairwise Mahalanobis distances. However, this approach has to be taken with caution, since we deal with the additional problem of violation from the i.i.d. assumption. The data observations are, in general, not independent, since we are dealing with spatially dependent data.

Let $\mathbf{z}_j$ be the next neighbor of $\mathbf{z}_i$, i.e. the Euclidean distance between $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$ is the smallest among all the neighbors of observation $\mathbf{z}_i$. Thus, the pairwise squared Mahalanobis distance $MD(\mathbf{z}_i, \mathbf{z}_j)^2$ is equal to a certain $\alpha(j)$-quantile $\chi^2_{D;\alpha(j)}(MD(\mathbf{z}_i)^2)$ of the non-central chi-square distribution. The value $\alpha(j)$ will be called the **degree of isolation** from the next neighbor.

Note that just by chance the next neighbor could be close but a third neighbor far away and local outlyingness of an observation implies that the observation is very different from most of its neighbors. Therefore, β will denote a fraction, and $\lceil n(i) \cdot \beta \rceil$ is the number of neighbors out of $n(i)$ of them in the neighborhood of $\mathbf{z}_i$ that can be similar to $\mathbf{z}_i$ but the remaining neighbors have to be reasonably different (here,$\lceil x \rceil$ means rounding to an integer not smaller than x). Note that $0 \leq \beta < 0.5$ aims at looking for local outliers, but for $0.5 \leq \beta \leq 1$ it is possible to search for homogeneous regions.

Definition of local neighborhoods

Local outlier detection requires the definition of local neighborhoods. For this purpose, two common approaches are proposed:

1. Fix the distance d_{max} and define the neighbors of an observation $\mathbf{z}_i$ as all points $\mathbf{z}_j, (j = 1, \ldots, n; j \neq i)$ where the distance $ED(\boldsymbol{s_i}, \boldsymbol{s_j})$ is not larger than d_{max}.

2. Define the neighborhood of an observation $\mathbf{z}_i$ with the k nearest neighbors (kNN) method. For an observation $\mathbf{z}_i$ we have to consider the sorted distances $ED(\boldsymbol{s_i}, \boldsymbol{s_{(1)}}) \leq ED(\boldsymbol{s_i}, \boldsymbol{s_{(2)}}) \leq \ldots \leq ED(\boldsymbol{s_i}, \boldsymbol{s_{(k)}}) \leq \ldots \leq ED(\boldsymbol{s_i}, \boldsymbol{s_{(n)}})$ to all other observations and the kNN to $\mathbf{z}_i$ are all observations where $ED(\boldsymbol{s_i}, \boldsymbol{s_j}) \leq ED(\boldsymbol{s_i}, \boldsymbol{s_{(k)}})$ for $j = 1, \ldots, n; j \neq i$

Clearly, if kNN is used, the number of neighbors of any observation $i \in \{1, \ldots, n\}$ is always fixed, $n(i) = k$.

Let $MD(\mathbf{z}_i, \mathbf{z}_{(j)})^2$ be the sorted squared pairwise Mahalanobis distances of observation $\mathbf{z}_i$ to all of its neighbors $\mathbf{z}_{(j)}$, with $j \in N_i = \{i_1, \ldots, i_{n(i)}\}$. As already discussed, the degree of isolation of an observation $\mathbf{z}_i$ from a fraction $(1 - \beta)$ of its neighbors can be characterized by the $\alpha(i)$-quantile

$$\chi^2_{D;\alpha(j)}(MD(\mathbf{z}_i)^2) = MD(\mathbf{z}_i, \mathbf{z}_{\lceil n(i)\cdot\beta \rceil})^2$$

If a large amount of neighbors is taken into account and all the assumptions are fulfilled (independence and normal distribution of the observations), then $\alpha(i)$ should be approximately equal β.

However, if $\alpha(i)$ is substantially larger (e.g. two times) than β, observation $\mathbf{z}_i$ is considered a local outlier.

In the following we demonstrate the proposed techniques at an example, which is a subset of monthly weather data of Austria GeoSphere Austria (2021). It represents 183 stations in Austria where 10 different variables were noted (including spatial coordinates, longitude and latitude). The focus here is just on two measurements, average daily sum of precipitation in mm (rsum) and altitude of the weather station in meters (alt).

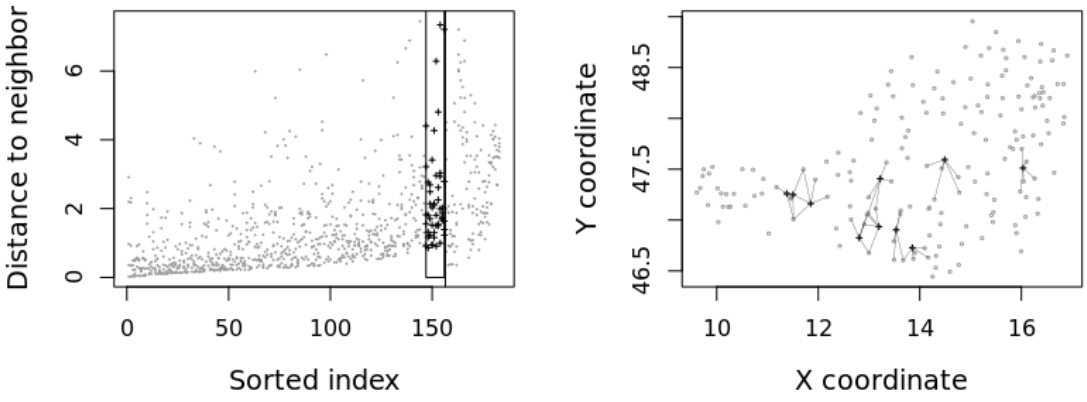

Figure 4.1: Local outlier detection of two variables of the GeoSphere Austria (2021). Left: observations sorted by their degree of isolation with distances to their neighbors; right: the selected 10 most extreme local outliers with their 5 nearest neighbors in the geographical sense.

The neighborhood size for the kNN is fixed to $k = 5$, and the parameter $\beta = 0.1$. For each observation the degree of isolation is computed and then the observations are sorted by the degree of isolation on the x-axis of Figure 4.1 (left). Vertically there are neighbors to each observation marked by their distance on the y-axis. The plot is also split into local (left) and global (right) outliers. The most extreme local outliers were selected and they are marked on the right plot of Figure 4.1, in the spatial coordinate space with their 5 nearest neighbors connected to them.

The plots were made using the function `locoutSort` that can be found in the `R` package `mvoutlier` (Filzmoser and Gschwandtner (2021)).

With the method described in this subsection, all together 12 local outliers were identified. As one can see in Figure 4.2 (top), they are not necessarily extreme in the data cloud. There are other observations that clearly deviate from the majority in the variable space, but those would refer to global multivariate outliers. However, the identified local outliers can possibly indicate regions with stronger local variability, such as the stations which are on higher altitude (in the mountains), but all their 5 neighbors are lower in the valleys. There are of course all the other measured variables that one can take into account, or that can influence rsum or alt, so this just shows a simple example in the two-dimensional variable space, since this way we can still plot the observations.

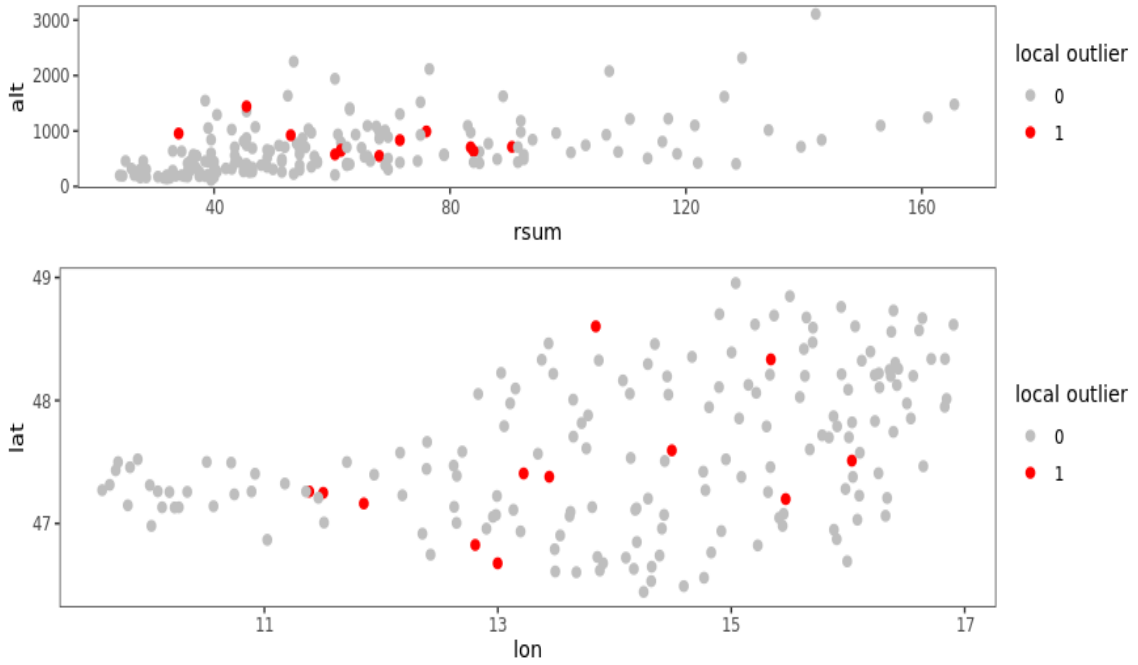

Figure 4.2: The marked 12 most extreme identified local outliers. Top: scatter plot in the space of variables rsum and alt; bottom: scatter plot in the geographical space.

4.2 Regularized spatial outlier detection technique

In the last section, local outlier detection for spatially dependent data was introduced. The method in this section extends the previous method with two proposed improvements. This method was introduced in Ernst and Haesbroeck (2017).

We start with the idea that the observations can be categorized in four groups: the local outliers, the global outliers, the local and global outliers and the regular observations. It is of interest here to tackle the problem of detecting local outliers. Thus, it would be important to try to improve the method's local nature.

The way of defining the neighborhoods should be done independently before applying the detection technique. Therefore, it is still proposed to either fix the maximal allowed distance d_{max} or to fix the parameter k for the kNN distance.

In Filzmoser et al. (2014), the technique is based on calculating the (squared) pairwise Mahalanobis distances

$$MD(\mathbf{z}_i, \mathbf{z}_j)^2 = (\mathbf{z}_i - \mathbf{z}_j)' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{z}_j)$$

of all the observations $\mathbf{z}_j$ in the neighborhood of the observation $\mathbf{z}_i$, relying on the robust estimation $\mathbf{C}$ of the *global* covariance matrix. This leads to the computation of the isolation degree as this degree is a quantile of a non-central χ^2 distribution.

Therefore, to increase the local nature of this procedure, it is suggested to take locally estimated covariance matrices into the definition of the pairwise squared distance, yielding so-called **local squared Mahalanobis distances**

$$MD(\mathbf{z}_i, \mathbf{z}_j)^2 = (\mathbf{z}_i - \mathbf{z}_j)' \boldsymbol{C_i}^{-1} (\mathbf{z}_i - \mathbf{z}_j).$$

Here, $\boldsymbol{C_i}$ is the robust MCD estimator of the covariance matrix $\boldsymbol{\Sigma}_i$ of the observations belonging to the neighborhood of $\mathbf{z}_i$, and $\mathbf{z}_i$ itself.

Nevertheless, some care is needed while computing the local estimator $\boldsymbol{C_i}$, because the number of neighbors $n(i)$ can be very small, sometimes even smaller than the dimension D. We want to ensure the positive-definiteness of the estimated covariance matrix. Therefore, the regularized version of the MCD estimator already introduced in Section 3.1 should be used.

One should consider more detailed analysis of choosing the tuning parameters for the regularized MCD estimator of each neighborhood:

1. The coverage is chosen according to the breakdown point of the robust estimator one wants to achieve by taking $h = \lceil n \times (1 - \alpha) \rceil$, where $0 < \alpha < \frac{1}{2}$ is the chosen breakdown value. It can be shown that the breakdown point of the regularized MCD is given by $\frac{\min(h, n-h+1)}{n}$. It is not expected to have more than 25% of outliers in each neighborhood, so it is decided to set the coverage rate to the proportion 0.75 for all the local estimations (i.e. $h_i = (n(i)+1) \times 0.75$).

2. The regularization parameter equal to $\frac{tr(\Sigma)}{np}$ would yield an unbiased estimation of the trace of the covariance matrix. That means that λ_i can locally be set to $\frac{\widehat{tr(\Sigma_i)}}{h_i p}$, where $\widehat{tr(\Sigma_i)}$ should be robustly estimated.

All the proposed parameters are taken from Ernst and Haesbroeck (2017).

A second improvement can be made to the methodology in Filzmoser et al. (2014) in order to take into account the possible heterogeneity of the local neighborhoods.

One can measure the *concentration* in the space of the non-spatial attributes (variable space) of the observations in a neighborhood as the volume of the ellipsoid centered at the robust estimation of the local mean and shaped according to the locally and robustly estimated covariance matrix.

More precisely, let the $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$ be the regularized MCD estimations for the neighborhood of the observation $\mathbf{z}_i$. The estimated covariance $\widehat{\boldsymbol{\Sigma}}_i$ is characterized by its *size*, i.e. its determinant and a *shape* defined by $\widehat{\boldsymbol{V}}_i = \widehat{\boldsymbol{\Sigma}}_i / \sqrt[p]{|\widehat{\boldsymbol{\Sigma}}_i|}$. Notice that the determinants of all shape matrices are equal to 1 and therefore, using them to construct the ellipsoids yield the ellipsoids of comparable volumes.

Moreover, a measure of concentration in the i-th neighborhood, c_i, can be defined as follows,

$$c_i = \frac{1}{h_i} \sum_{j:z_j \in H_i} MD_{\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{V}}_i}(\mathbf{z}_i)^2,$$

where H_i is the optimal subset corresponding to the regularized MCD estimations. An alternative option for computing a measure c_i of concentration would be to replace the mean by the median.

Finally, the resulting "volumes" c_i are ranked from the smallest (most concentrated ellipsoid) to the largest. Only the observations having neighborhoods among the smallest $\lceil \beta \times n \rceil$ ellipsoid volumes are considered further in the local outlier detection technique.

Taking the parameter β too large means to allow more heterogeneous zones to enter our analysis which can increase the false positive rates. On the other hand, taking β too small could be too restrictive as some neighborhoods can contain more severe local outliers. It is advised to take a grid of proportions for β, for example $0.1, 0.25, 0.5, 0.75, 0.9$, and for each of those continue the analysis.

The last change in the method from the Section 4.1 is the parametric approach with the χ^2 distribution. This parametric approach has been replaced by a non-parametric one. The *next neighbor* of the observation is considered, being the closest neighbor in the non-spatial (variable) space. The (local squared Mahalanobis) distance between the observation and its closest neighbor is called *next distance* and when this distance is large, the observation is considered to be a local outlier.

Now, for each of the β from the chosen grid, the next distances are plotted by means of a boxplot. Nevertheless, asymmetry of the distribution of the distances should be taken into account, so the **adjusted boxplot** is preferred. Full details on the adjusted boxplot can be found in Hubert and Vandervieren (2008).

The fences of the boxplot used here are

$$[Q_{0.25} - 1.5e^{-4MC} \cdot IQR, Q_{0.75} + 1.5e^{3MC} \cdot IQR],$$

where MC denotes the medcouple, which measures the skewness of the sample and is defined by

$$MC = median_{z_i \leq Q_{0.5} \leq z_j} \frac{(z_j - Q_{0.5})(Q_{0.5} - z_i)}{z_j - z_i}.$$

The observations having the biggest next distances may be considered local outliers. Thus, the natural cut-off value is the upper whisker of the adjusted boxplot.

To demonstrate the usage of the proposed method, let us get back to the data set of the GeoSphere Austria (2021) monthly weather data (from the Section 4.1). We are again observing just two variables rsum and alt. After applying all the necessary steps, one can study the adjusted boxplot of the "next distances" of the observations (Figure 4.3). For demonstration purposes, the parameter β is set to 1, meaning all the neighborhoods are considered for local outlier detection. The parameter k for kNN is still set to be 5.

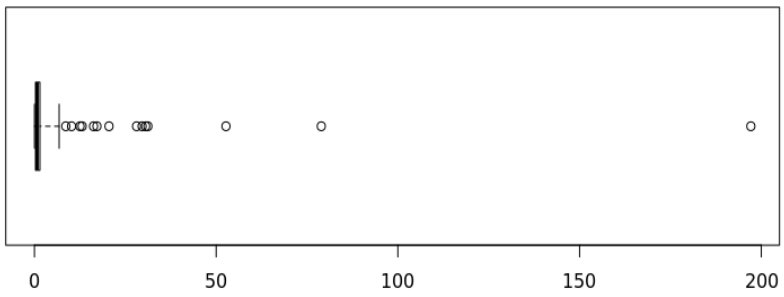

Figure 4.3: Adjusted boxplot of the "next distances" values.

Fourteen observations had their "next distance" value over the upper whisker and are thus considered as the most extreme local outliers. They are marked in the variable space of rsum and alt, as well as in the spatial coordinates in the scatter plots of Figure 4.4. Again, the local outliers do not have to be the most extremes in the whole data cloud.

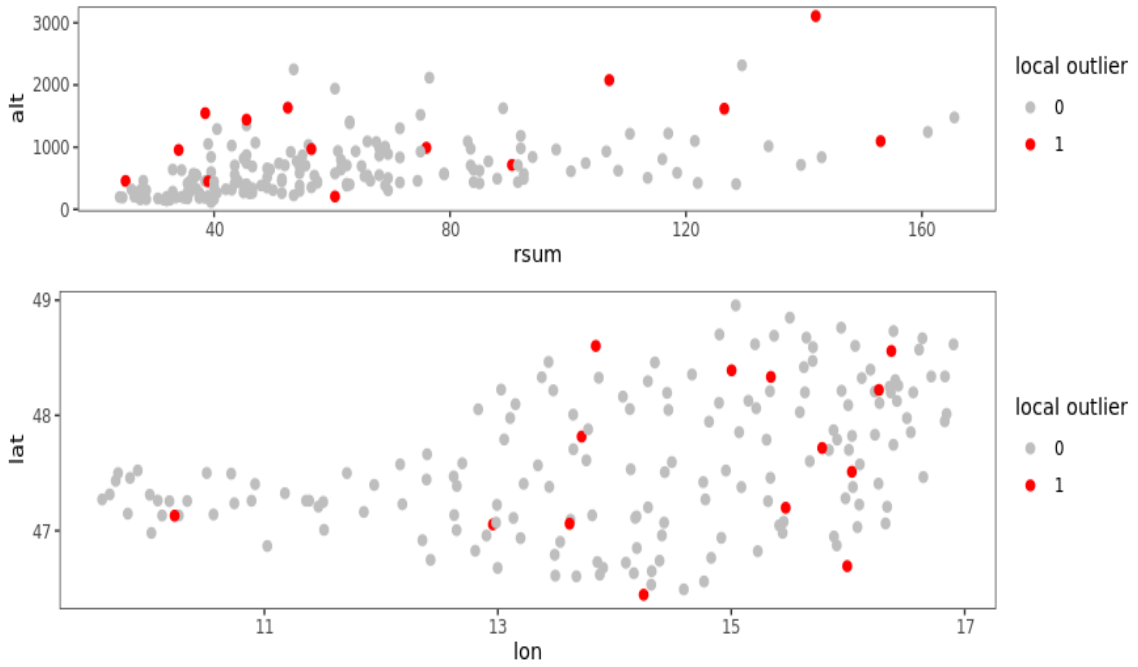

Figure 4.4: The marked 14 most extreme found local outliers. Top: scatter plot in the space of variables rsum and alt; bottom: scatter plot in the geographical space.

4.3 LOF

The following method was introduced in Breunig et al. (2000). This method, as the previous two, defines some "measure" of being outlier. For the start we would assume that the property of being an outlier is a binary property. That is, either an object in the data set is an outlier or not. For many cases, the situation is more complex, so as before, some *degree* of being an outlier will be assigned to the observations.

The outliers are local in the sense that only some restricted spatial neighborhood of each observation is taken into account. This happens in the space of spatial coordinates. After the local neighborhoods are defined, in the following we will be focusing just on the variable space.

Contrary to the previous two approaches, the LOF method does not use the MCD estimates. It proposes a completely new measure of outliyngness, and outliers are not defined based on the probability distribution of the data, that is in general unknown.

We begin by defining the *k-distance* of object p: Let k be a positive integer. The k-distance of object p, denoted as k-distance(p), is the distance $d(p, o)$ between p and an object $o \in O$ such that:

- for at least k objects $o' \in O \setminus \{p\}$ it holds that $d(p, o') \leq d(p, o)$, and
- for at most $k - 1$ objects $o' \in O \setminus \{p\}$ it holds that $d(p, o') < d(p, o)$.

In other words, it is the distance of the object to the k-th nearest neighbor. Note that the set of the k nearest neighbors includes all objects at this distance, which in some cases can be more than k objects. We then define the *k-distance neighborhood* of the object p as

$$N_k(p) = \{q \in O \setminus \{p\} | d(p,q) < k\text{-}distance(p)\}.$$

The chosen distance d can be any distance in the variable space. For simplicity, let us assume it is the Euclidean distance in the multivariate variable space. The object p is then some vector (observation) from that space. However, we will continue to use the notions distance and objects to stay on the abstract level for this subsection.

We continue by defining the *reachability distance of an object p from o*. Let k be a natural number. The *reachability distance* of object p from o is defined as

$$reach\text{-}dist_k(p,o) = \max\{k\text{-}distance(o), d(p,o)\}.$$

Intuitively, if object p is far away from o, then the reachability distance between the two is simply their actual distance. However, if they are "sufficiently" close, the actual distance is replaced by the k-distance of o (see Figure 4.5). Note that this is not a distance in the mathematical definition, since it is not symmetric.

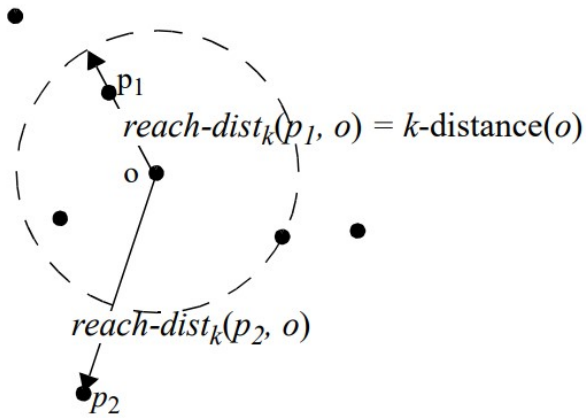

Figure 4.5: Illustration of reachability distance with $k = 4$.

By now, we are able to control the parameter k, specifying the number of objects in a k-distance neighborhood. In order to talk about the densities of the neighborhoods, we define *local reachability density* of an object p as:

$$lrd_k(p) = 1/\left(\frac{\sum_{o \in N_k(p)} reach\text{-}dist_k(p,o)}{|N_k(p)|}\right)$$

Note that it is not the average reachability of the neighbors from p, but the distance at which p can be "reached" from its neighbors. With duplicate points, this

value can become infinite but we will assume that there are no duplicate observations in our data set.

Finally, the *(local) outlier factor* of an object p is defined as:

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}.$$

The outlier factor of object p captures the degree to which we call p an outlier. A value of approximately 1 indicates that the object is comparable to its neighbors (and thus not an outlier). A value below 1 indicates a denser region (which would be an inlier), while values significantly larger than 1 indicate outliers. For the binary choice if an object is an outlier or not, 1.5 can be a rule of thumb for the cutoff value for LOF.

The LOF method has many good properties and is also applicable in clustering the observations. For example, one can show that most objects inside of a cluster have the LOF value approximately equal to 1. One can also give a lower and upper bound on the LOF and analyze the tightness of bounds (Breunig et al., 2000).

An interesting question comes to mind concerning the value k. Does LOF increase or decrease monotonically by monotonically increasing the value k? Unfortunately, the LOF does not increase or decrease by any rule. Given an increasing sequence of k values, one can however see that LOF eventually stabilizes to some value.

One can determine the values k_L and k_U as the "lower bound" and the "upper bound" of the range for k. The k_L can be regarded as the minimum number of objects a "cluster" has to contain, so that an object can be local outlier relative to this cluster. This can depend on the context of the application and can be as small as 2. The value k_U is then regarded as the maximum cardinality of a cluster, if this cluster is seen as "close by" objects. Then k_U can be set to the maximum number of objects (observations) in the data set (or already some spatially defined local neighborhood). Having determined k_L and k_U, we can compute for each object its LOF value within this range. It is proposed to rank all objects with respect to the maximum LOF value within the specified range. That is, the ranking of an object p is based on: $\max\{LOF_k(p) | k_L \leq k \leq k_U\}$. It is proposed to take the maximum to highlight the instance at which the object is the most outlying, but one can take other aggregates, such as the minimum or the mean.

In order to illustrate the LOF method, we will once again make use of the GeoSphere Austria (2021) monthly weather data with the same variables, rsum and alt.

The spatial neighborhoods are again defined using kNN (with $k = 5$). The LOF values of all the observations, considering other observations in their neighborhoods, are being calculated (in other words, the observations from their spatial neighborhoods are being taken into account for their k-distance neighborhoods in the variable space). The ones with the LOF values more than 1.7 are considered to be local outliers. Figure 4.6 shows the identified 16 local outliers, and they are marked in the variable space (top) and in the geographical space (bottom). As already seen, that

means that they are locally close to their neighbors, but "further away" from them in the variable space (or outside of the "cluster").

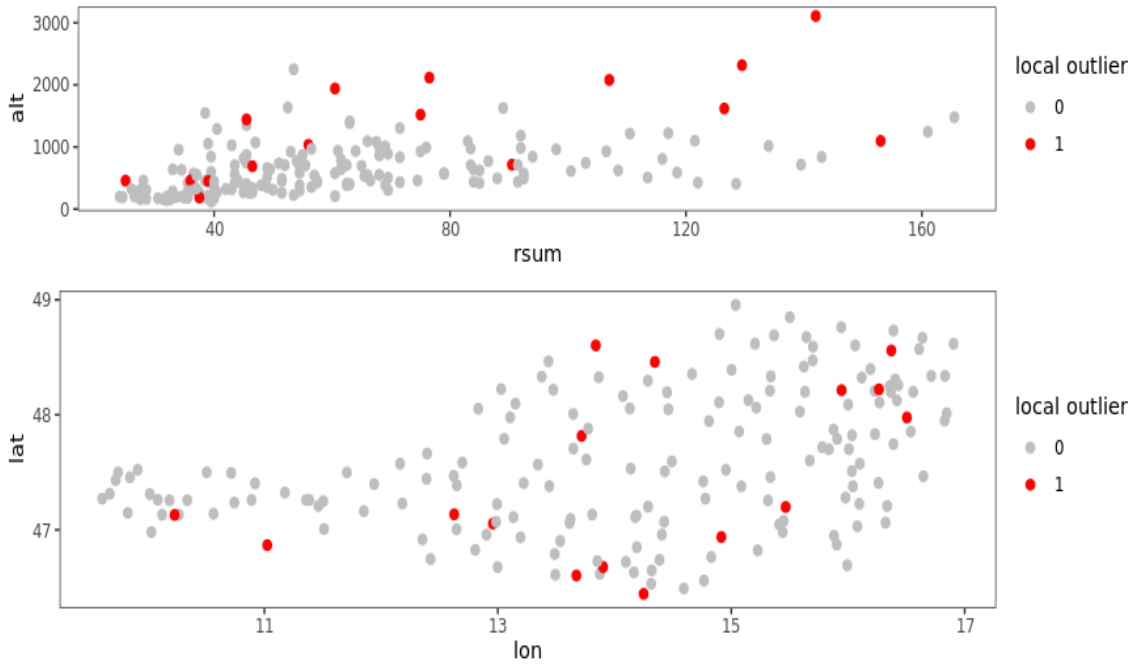

Figure 4.6: The marked 16 most extreme found local outliers. Top: scatter plot in the space of variables rsum and alt; bottom: scatter plot in the geographical space.

In Figure 4.7, observations with the highest LOF score (around 3.5) are marked, together with their 5 nearest neighbors. In the spatial space (top), they make a small "cluster", but in the variable space, the neighbors are making a "cluster" and the marked outlier is far outside of it.

On the other hand, another observation with a LOF value being around 0.73 is not marked as local outlier. Moreover, it should point to some denser region in the variable space, where the observation is inside of the "cluster". This observation (with its 5 nearest neighbors) is plotted in Figure 4.8, again in the sense of the geographical space and variable space.

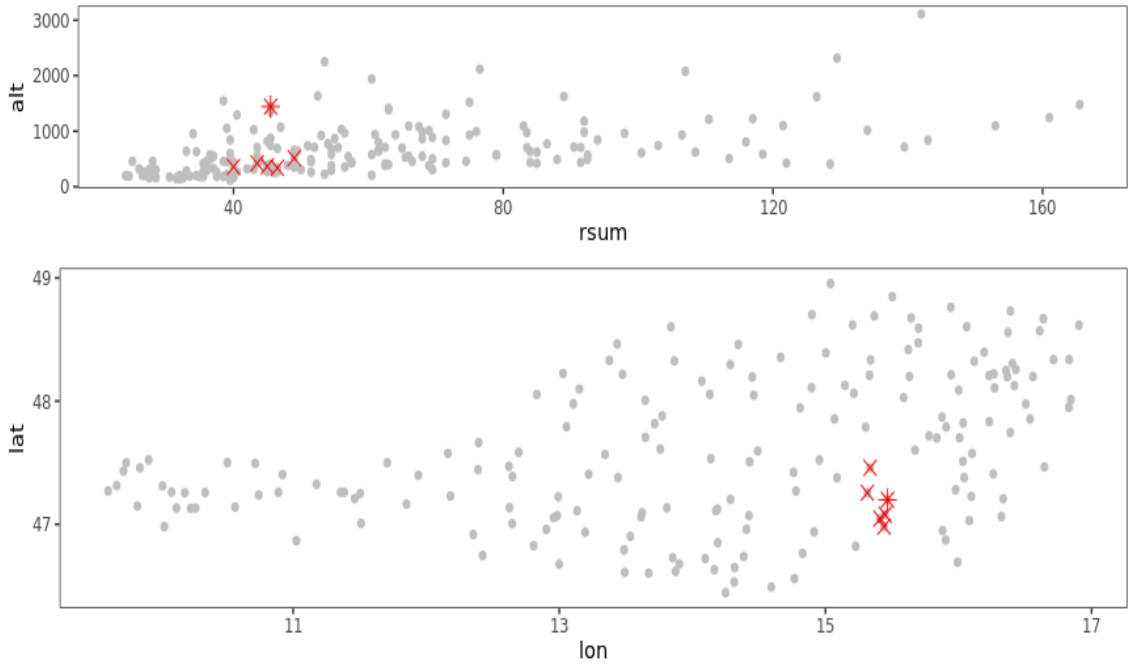

Figure 4.7: The most extreme local outlier (star) with its 5 nearest neighbors (crosses). Top: scatter plot in the space of variables rsum and alt; bottom: scatter plot in the geographical space.

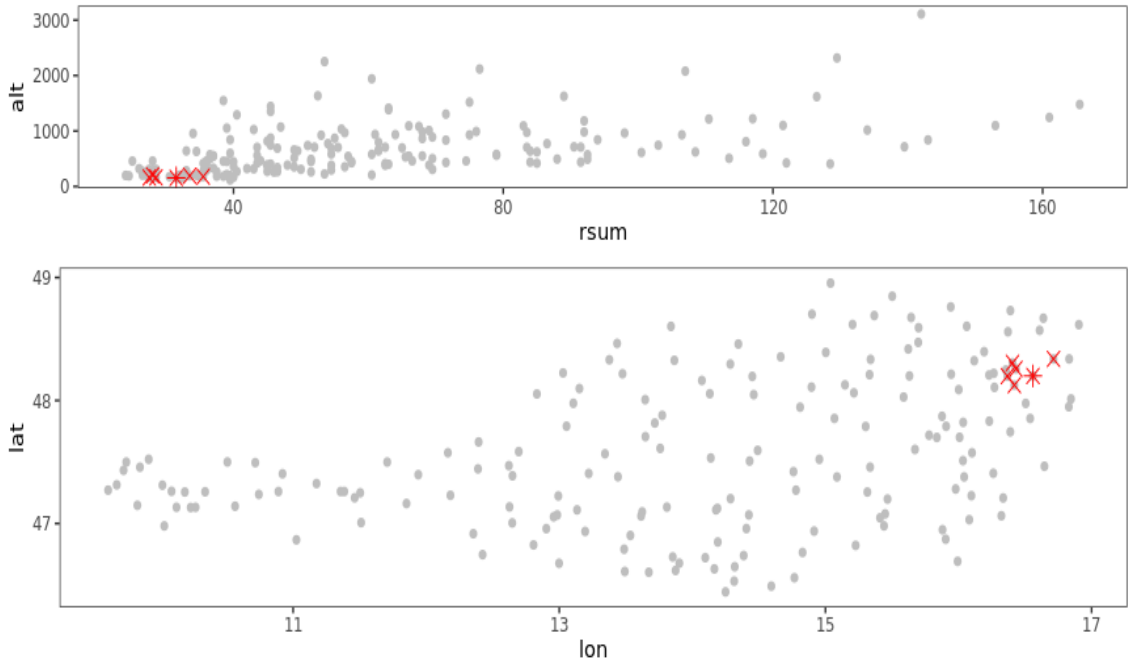

Figure 4.8: A local inlier (star) with its 5 nearest neighbors (crosses). Top: scatter plot in the space of variables rsum and alt; bottom: scatter plot in the geographical space.

5 Application to Geochemical Data

This section is aimed at describing the data that are analyzed in this thesis and applying the relevant local outlier detection methods. As the process of data cleaning and preparation (preprocessing) is a crucial step in every data analysis, it will be presented in this section in detail. The rest of the section aims at describing the set up and application of local outlier detection methods on the prepared data set in order to find and evaluate local outliers. The three methods from the Section 4 will be applied as well as one more method with further adapted MCD estimator. We also aim to explain and discuss the identified outliers. Finally, the methods should be evaluated and compared and some parameters should be discussed and tested.

5.1 Data description

Following the last sections in this thesis, the goal of the practical part is to apply the presented concepts and methods of finding local outliers for real world compositional data, where those findings can be of use.

As already discussed in Section 2 on compositional data, it is the type of multivariate data where some relative information of the variables is more relevant than absolute one. One of those are geochemical data, often seen in the geological sciences, where samples were collected in different layers of soil and the concentration of different elements was measured. The consequence of finding spatial local outliers in such type of data is detecting where in some smaller local regions some elements could be found with unusual concentration. Most interest here will be in critical raw materials. By detecting possible new deposits, new mines could be potentially put in place which would lead to more sustainable (local) further exploration and development.

The data for the practical part of this thesis include targeting till data from Geological Survey of Finland (2013). The spatial space in Finland, where the samples were taken, was divided in 170 mapsheets in the scale $1 : 1000000$ and the samples were analyzed per mapsheet. The samples were taken by a portable percussion drill. All samples have been analyzed using the same analytical equipment, but because of ten years period during which the sampling took place, some variation can be seen, especially on the borders of the mapsheets.

For the purpose of this thesis, the data set used was from chemically unchanged C-horizon till and one chosen mapsheet of interest at a time. At one point of spatial coordinates, there were samples taken at different depths of the soil, but they will be seen as different samples having the same spatial coordinates.

The mapsheets usually consist of six sub-mapsheets, and the taken samples plotted in the spatial space look like in Figure 5.1. One can see that sampling was not uniformly distributed on the mapsheet but rather along some more or less horizontal lines.

The working data set, apart from the spatial coordinates (NORTHING_YKJ, EASTING_YKJ), has 17 different elements as variables (Si, Al, Fe, Mg, Ca, Na, K, Ti, V, Cr, Mn, Co, Ni, Cu, Zn, Pb and Ag). Some of the concentrations of

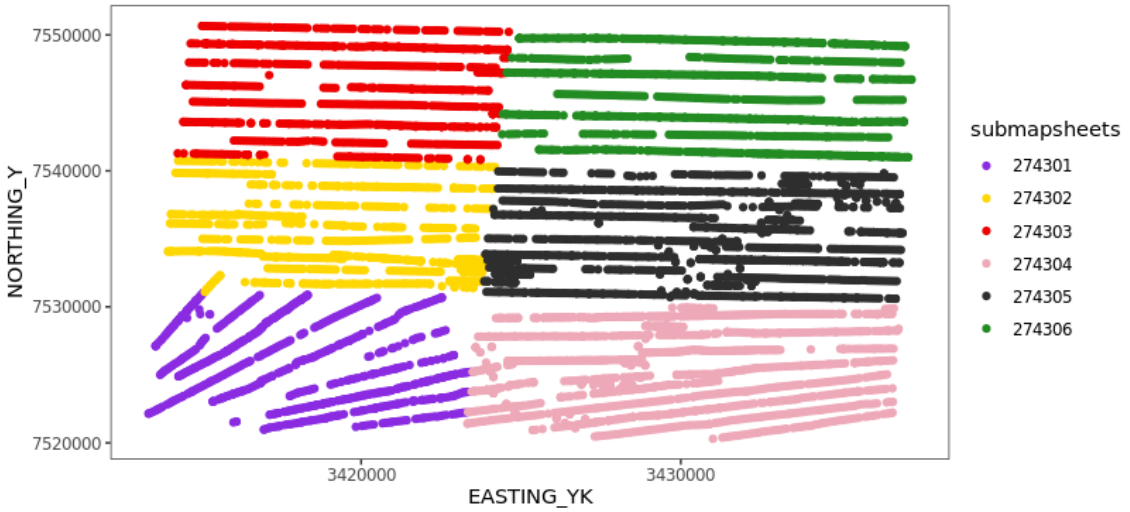

Figure 5.1: Mapsheet 2743 divided in its 6 sub-mapsheets.

the elements are expressed in percentages and some in ppm (parts per million). However, as already emphasized in the previous sections, the choice of the units for the parts of compositional data is not important, because the absolute sum is not important.

For all the (compositional) data where some measures are taken with appropriate tools, it is common that the tool used can detect the concentration down to some limit, called the detection limit. It means that the concentrations under that limit will not be detected correctly, but rather rounded to zero, rounded to half of the value of the detection limit or left as missing value. One should specially mark and take care of those measurements.

In our case, each of the elements measured has its own detection limit (see Figure 5.2) and they are known. One can notice that if already one element in a sample had a strange (or by the definition of the compositional data not allowed) value, it can influence the whole sample, or in the words of local outlier detection, the sample can be marked as local outlier just because already one element of the sample (in its local neighborhood) had an inappropriate value (zero, missing value, or simply really small concentration).

However, in the practical part of this thesis, one is rather interested in finding local outliers that have one or more elements with extremely high concentration compared to other samples in its local neighborhood. One is not interested in finding local outliers because of having elements with concentrations under their detection limits. Even if those are for some reason found and marked as outliers, one should think of filtering them for the ones that we are interested in. Thus, some univariate analysis of the elements is needed before as well as after applying the local outlier detection techniques.

After applying the local outlier detection techniques, one is interested in evaluating the identified local outliers and evaluating the performance of the method. Apart from univariate analysis of each element, some reference is needed for already

Table 1. The determined elements, the used wave lengths of spectrum, limit of quantitation (LOQ) in aqua regia matrix diluted 1:100 and median values in soil according to Rose et al. (1979).

Alkuaine	Aallonpituus (λ) nm	Määritysraja (LOQ) ppm	Maaperän mediaanipitoisuus ppm
Ag	328,068	5	
Al	308,215	70	
As	193,215	40	7,5
B	499,356	20	29
Ba	493,409	8	300
Ca	317,933	40	
Cd	457,604	8	0,3
Co	228,616	2	10
Cr	267,716	4	43
Cu	324,754	8	15
Fe	259,940	50	21 000
K	766,940	1300	11 000
La	379,478	3	
Li	670,784	7	22
Mg	279,079	20	
Mn	257,610	1	320
Mo	202,030	3	2,5
Na	589,592	120	
Ni	463,208	10	17
P	429,828	60	300
Pb	220,353	30	17
Sb	413,666	30	2
Sc	361,384	1	
Si	251,611	100	
Sr	421,552	1	67
Th	401,913	20	13
Ti	334,941	4	
U	367,007	1300	1
V	292,402	7	57
W	207,911	40	1
Y	371,030	2	
Yb	425,348	12	
Zn	213,856	2	36
Zr	339,198	3	270

Figure 5.2: Example table of detection limits (column LOQ) and other information for each element. (Salminen, 1995)

known local regions (known mineral deposits) so one can evaluate if local outliers that were output of applied methods are actually found near them. In other words, if those would not already be known mineral deposits, would they have been found by local outlier detection methods. For that purpose, it was taken care that there are already known mineral deposits in the mapsheets we want to work with. So apart from the working data set, the data of mineral deposits (Geological Survey of Finland, 2016) was also used for the purpose of this thesis. Information about each mineral deposit is already available, such as mine type, discovery year, commodity measures etc. However, in this thesis, the exact elements whose concentrations make

some sample a local outlier are not discussed and the only information taken from the mineral deposits data set was their spatial coordinates (locations) that fall into a mapsheet of interest.

In the first place, we want to focus on the mapsheet 2743 (see Figure 5.3) in the central Lapland area, where there are 12 known mineral deposits.

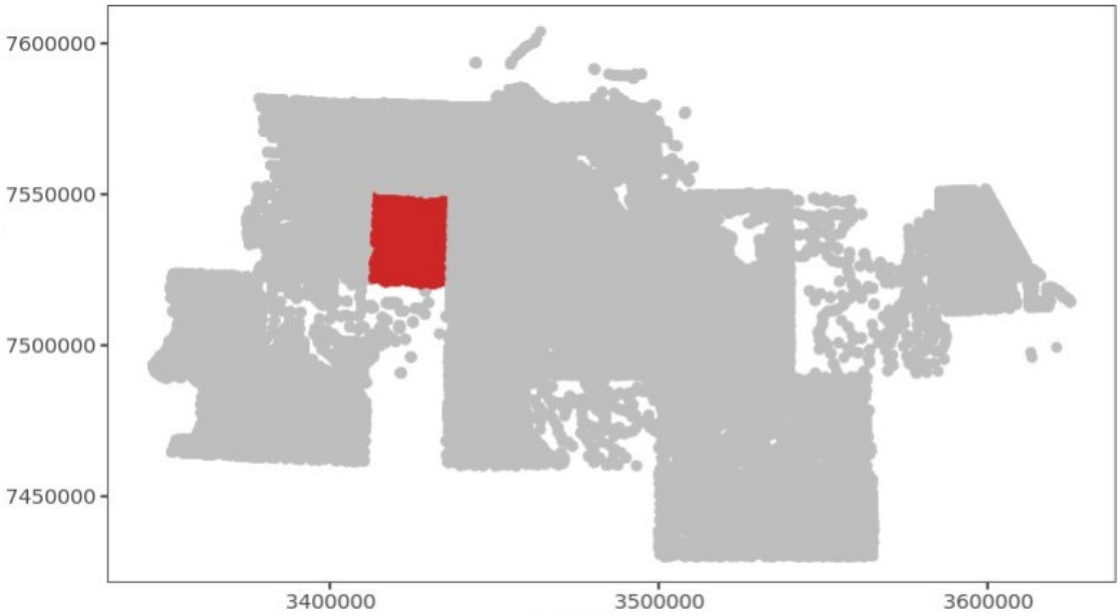

Figure 5.3: Sampling in the central Lapland area with marked mapsheet 2743.

5.2 Data preparation and cleaning

As already stated in the previous section, the preprocessing and data cleaning step can be crucial for this type of analysis. One has to take special care of the rounded zeros, negative values and values under the detection limits. We are focusing for the moment just on the mapsheet 2743 including 6662 samples.

Zeros and negative values

Zero values and negative values should be taken care of first. For the data set of the mapsheet 2743 there are around 3% of the rows (observations) that have for at least one element zeros or negative values. For such a small percentage it is acceptable to simply omit those observations from the analysis. One is then still left with 6442 observations.

Another option would be to impute those values. However, imputing has to be done very carefully and in the context of the domain. For instance, for geochemical data, the detection limit of measurement devices needs to be taken into account. Imputing can mess later with the local outlier detection, since there is a danger of finding observations with imputed values as outliers, which would obviously not be

the goal of this analysis. 3% or lower is in any case a percentage small enough to simply kick out those observations and continue with the rest of unchanged data.

Filtering variables

Every cell of the data set, where the value is below or above its detection limit, is already noted as negative or zero value, or if it is known that the value is some mistyped or wrongly measured value it is marked with signs $<, >, ?, *$, respectively. In technical terms, for each column (element) with concentration values there is an adjacent column marking if it was such a value in the corresponding cell.

After the first step, our data set has no zero or negative values, which was needed in the first step since we want to handle the data as compositions. On the other hand, with the univariate analysis of the elements, one should check how many values in each column were actually marked with the mentioned signs and omit some of the elements from the further analysis (for the record, the element Ag had even 100% of the marked values, which means that there is no relevant information from that element whatsoever). The choice was to omit the elements (columns) that have more than 5%. The reason is that we do not want our detection techniques to find outliers based on having "wrong" values since the variable would spread that information after applying transformations for compositional data. In this case it is a better option to omit some elements from the analysis completely.

Applying this step for the mapsheet 2743, one is left with 10 elements, Si, Al, Fe, Mg, Ti, V, Mn, Co, Ni and Cu.

Filtering samples

On the other hand, there were more than 6400 samples in the chosen mapsheet and sampled spatially in a quite dense way since some observations were under each other (in different depths of the soil). Inspired by the previous preprocessing step, one can afford to similarly omit samples (rows) that have more than 30% of the elements with marked values (signs). Thus, one omits the observations that have more chance of being marked as local outliers for having strange and unreliable values.

After applying this step for the mapsheet 2743, one is, however, still left with 6436 samples.

Univariate analysis

The previous preprocessing steps of filtering possible lower quality variables and observations is a crucial step before possibly applying local outlier detection techniques. However, since the spatial position plays a big role in finding local outliers, the chosen spatial space (mapsheet) should have comparable sub-mapsheets.

In other words, one does some univariate anaylsis element by element but with respect to sub-mapsheet in order to compare them for each element. There are more ways of plotting the empirical data so one can judge the underlying distribution (Reimann et al., 2011). We will use the tool called QQ-plot. One can plot sorted raw values against the quantiles of a lognormal distribution (which is know to be expected from this type of data). Since lognormal distribution is simply the logarithm of

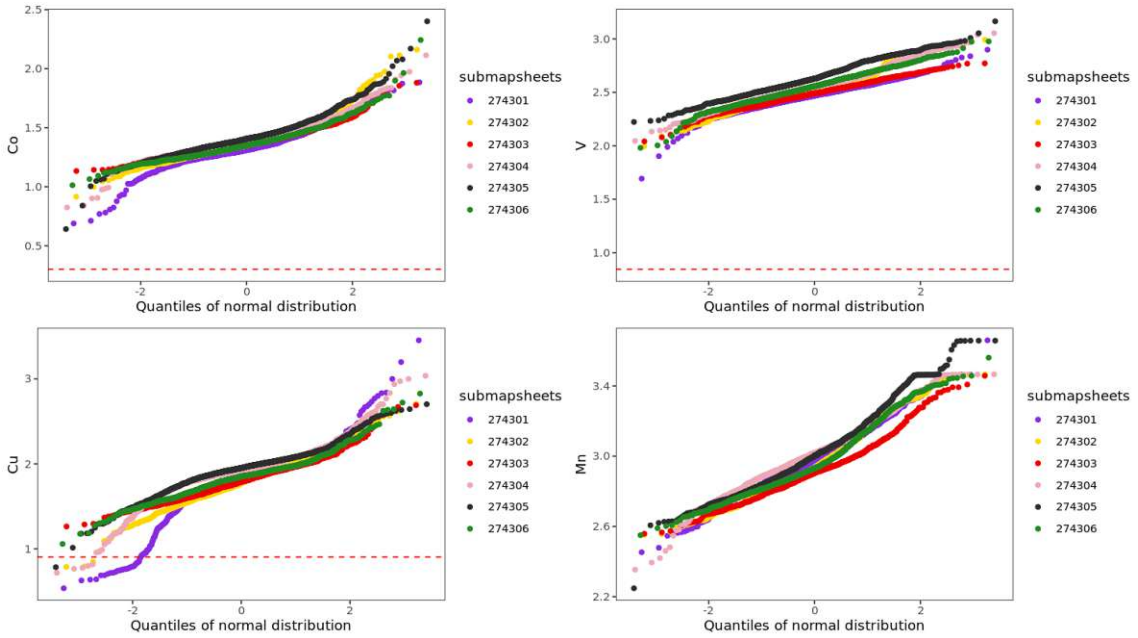

Figure 5.4: QQ-plots of the elements Co, V, Cu and Mn per sub-mapsheet.

the normal distribution, one can also plot the log-transformed values against the quantiles of a normal distribution. It is easy to detect changes from an expected distribution if the points fail to follow a straight line. However, we do not expect the perfect straight line, since we expect the data to contain outliers (local but also global ones).

As an example, Figure 5.4 shows such QQ-plots for the elements Co, V, Cu and Mn. One can see that the sub-mapsheets are in principle comparable. For the element V, the lines are a bit further from each other, but over all the quantiles are still parallel. The red dashed lines are (log-transformed) detection limits of the elements. The elements Co, V and Mn have all their values above the detection limit. It means that those elements had quite stable measurements. Cu is more unstable. We expect a bit of dispersion on the both ends of the lines, but Cu shows bigger dispersion with two sub-mapsheets' values going under the detection limits, which means that those values were more or less rounded, but not rounded to zero (since those values were already omitted). The element Mn has its values way over the detection limit but one can notice a rather rarer problem. The values are rounded on the upper end of the lines (sub-mapsheets 274304 and 274305). It means this element had a problem with the upper detection limit.

The univariate analysis is important to inspect the values in relationship with their detection limits and their distribution. Although some elements show a certain instability we do not want to impute or omit values, since imputing or omitting the ends of the shown lines would be omitting a part of the distribution's end. And although the data do not show a perfect quality, one can do a univariate analysis after the outlier detection procedure to conclude how the method has handled these instabilities.

Preparation of compositional data

In the practical part of this thesis, the aim is to work with compositional data and apply and compare the methods for local outlier detection, since data are spatially dependent. All the previous steps of pre-processing were needed, so one can transform the prepared data set into the chosen coordinates. The question is what coordinates (transformation) shall be chosen?

The alr transformation does not build an isometry and is thus not suitable for outlier detection, since taking the coordinates can change the distance between the samples in the variable space. This problem is solved with the clr coefficients. However, the clr are the coefficients of a generating system so the resulting data matrix has not full rank in the columns. Thus, the covariance matrix would end up being singular which is completely unsuitable for the local outlier detection methods which use some global or local covariance structure (4.1, 4.2).

As discussed in Section 2.2, the ilr coordinates aim at building an orthonormal basis and avoid the singularity issue while computing covariance matrices. They also built an isometry between Aitchison and Euclidean space. These special properties are needed for the further statistical methods that we want to apply, thus, the resulting data set will be transformed into pivot coordinates. Finally, one ends up with a prepared data set with 6436 observations (rows) and 9 variables (after taking the ilr coordinates one ends up with one less variable than prepared raw data with 10 chemical elements).

As seen in this subsection, the preprocessing and pre-analysis is very important and sometimes quite time-consuming before one can indulge in the whole outlier detection analysis. The appropriate analytical skills are needed as well as domain knowledge, since after possibly imputing values and filtering the observations, one should still be left with the data set that mirrors real-life data and captures the most real information as possible. And having compositional data at hand means from the very definition that one should kick out, impute or filter all the values and observations having zero or negative values. Although having multivariate data to deal with, one should still not forget to do some univariate statistical analysis. Outlier detection aims at finding outliers, meaning observations that in some way deviate from the majority, but one does not want to locate imputed, wrong values or extremely small values that fall under the detection limit. However, after the applied preprocessing steps and analysis (for the mapsheet 2743), we can finally proceed to applying the local outlier detection methods described in the previous sections of this thesis.

5.3 Application of robust local outlier detection technique

The goal in this and in the next sections is to show the set-up and the results of local outlier detection methods from the theoretical part of this thesis (see Section 4).

Let us start by applying the Robust local outlier detection technique from Section 4.1 on the prepared data set of the mapsheet 2743. For this and the next proposed outlier detection technique, the choice of the definition of local neighborhoods

is to be made beforehand and independently from the outlier detection method. We will use kNN, so the parameter k, that defines how many nearest neighbors of the observations should be taken into account for the local neighborhood, should be set. Seeing the sampling of the working data set, the method of setting the radius for the local neighborhood d_{max} should not give much different results.

Another possible parameter to set is the parameter β, defining a fraction of the data in the local neighborhood, that are allowed to be similar to the observation at hand. Often seen in practice is to set β to 0.1 and for the observation to be a local outlier one wants to have the $\alpha(i)$-quantile of the i-th observation more than two times larger than β. That means that the local outlier found would have $\alpha(i)$-quantile of the chi-quare distribution of their Mahalanobis distance larger than 0.2.

The `R` package `mvoutlier` (Filzmoser and Gschwandtner, 2021) gives a tool to explore the stability of finding global and local outliers for different values of k. Figure 5.5 shows the output of the function `locoutNeighbor` from this package. The parameter β is set to 0.1, which means that we allow 10% of the data in the neighborhood to be similar to the observation and analyse the remaining 90% of its neighbors. The plot is done separately for globally outlying observations (not taking into account the spatial component) and all the other observations ("regular observations"). Local outliers can be found in each of those two groups, but some global outlier can, however, be regular in its local neighborhood. The y-axis measures the degree of isolation, and the number of neighbors (k) changes on the x-axis. Each line represents one observation.

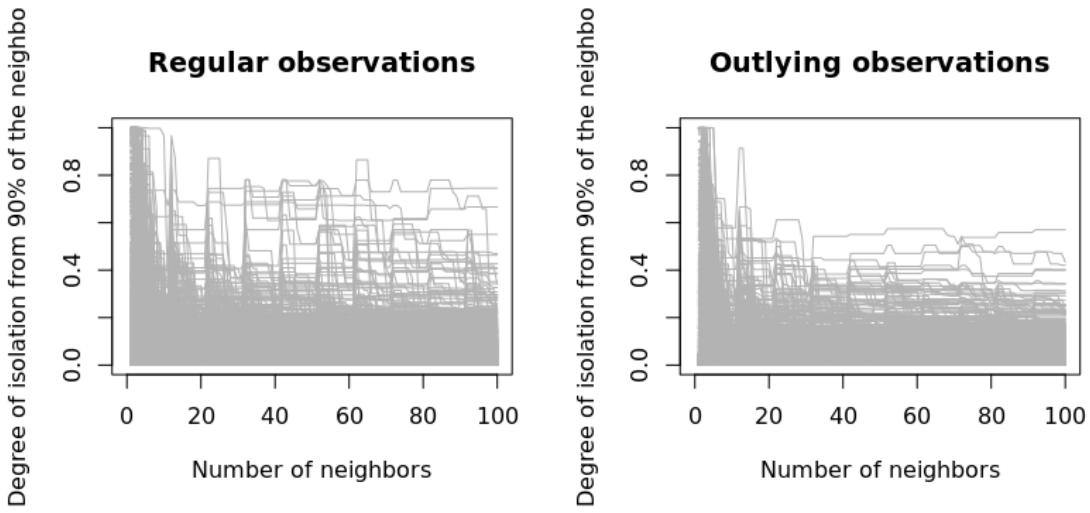

Figure 5.5: Degree of isolation (vertical axis) of each observation (lines) from 90% of the neighbors. The size of the neighborhood is changed on the horizontal axis from 0 to maximal 100. Separate plots are drawn for regular (left) and globally outlying observations (right).

For very small neighborhoods (under 20) one can observe more instability. The reason is that just by chance two observations could be close in the spatial sense but very different in the variable space. In our case, some observations have the

spatial (Euclidean) distance between them even 0 (the observations on the different depths in the ground). There are up to 10 samples on top of each other, but this maximum is reached for just a couple of observations. However, for this reason we want to set our k to be more than 10. For a larger neighborhood the local outlier measure should become more reliable but it does not stabilize completely reaching 100, which is to be expected since the size of our data set is over 6000. Nevertheless, for now the k value will be set to 30, since for that neighborhood size one seems to catch local outliers among regular observations and the outlying ones and for the bigger values, the outcome does not seem to change that much. Different k values will be explored in the last subsection.

There is another tool from the `mvoutlier` package that does a similar analysis. The function `locoutPercent` analyzes, for a fixed value of k, different β values. Figure 5.6 shows the output of this function.

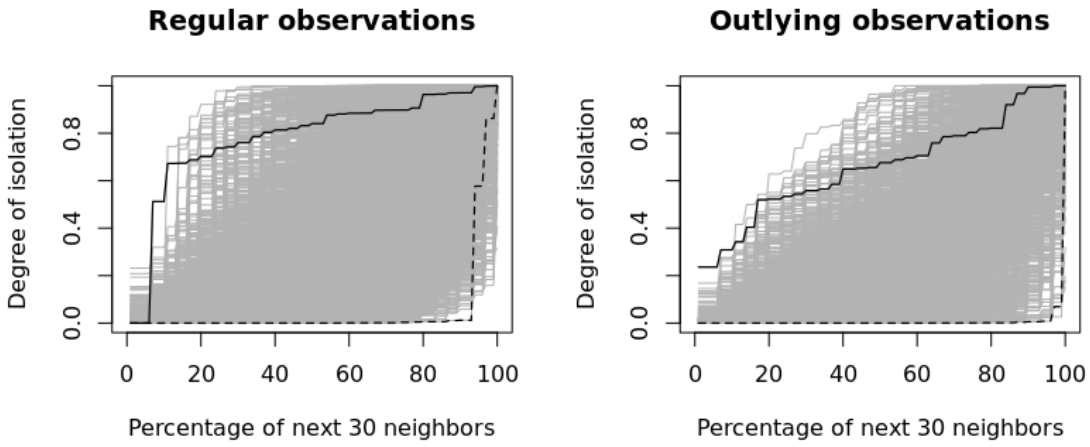

Figure 5.6: Degree of isolation of each observation (lines) from its nearest neighbors. The percentage of the neighbors taken is changed on the horizontal axis. Separate plots are drawn for regular (left) and globally outlying observations (right).

One can see that for $\beta = 0.1$, and with a cut-off value for the isolation degree set to 0.2, one does not seem to catch as many local outliers (the lines are first rising up after 0.2). If one would take $\beta = 0.3$ and the cut-off value a bit less than twice as large as β, for example 0.4, one would end up "in the denser" regions of both plots, and seem to capture the local outliers. Let us set β and the cut-off value like this for the moment being.

Once the parameters for the method are discussed and set, one can simply apply the method and plot the identified local outliers in the spatial space together with the found mineral deposits in the mapsheet 2743.

Figure 5.7 shows by marked red symbols the 255 found local outliers. The symbols + refer to the known mineral deposits in that mapsheet. By observing the figure, one cannot claim that all mineral deposits would be found by this local outlier detection technique. It is of course a topic for itself, how does one define

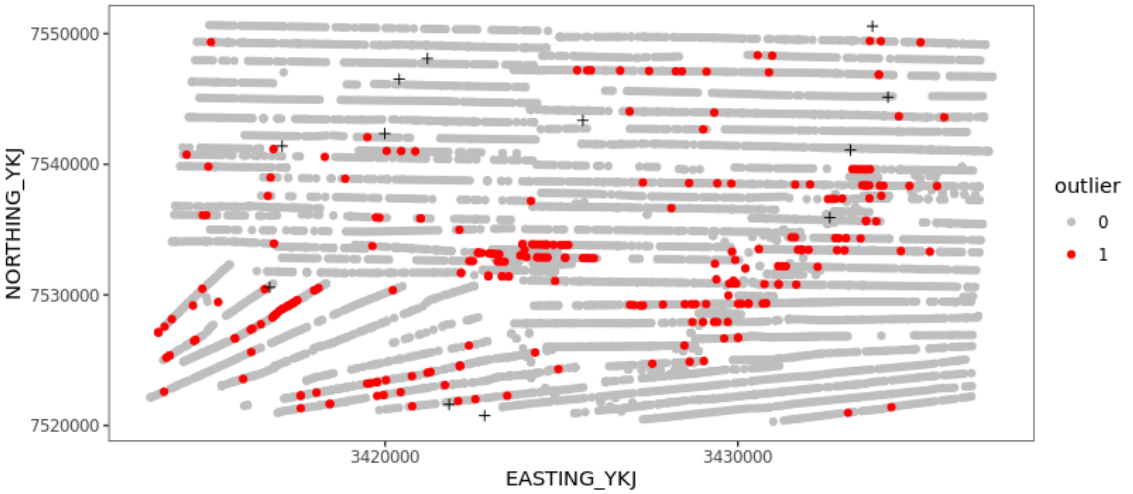

Figure 5.7: Marked (red) found local outliers with the Robust local outlier detection technique in the mapsheet 2743, and known mineral deposits (black symbols +).

if the mineral deposit is found or not. Moreover, we cannot tell if a found local outlier not connected to some known mineral deposit would be a false positive or not. In Figure 5.7 one can notice a "cluster" of outliers in the middle and a more dense line on the right-hand side, that are worth checking for. The main idea is to find local outliers (possible local regions) that have new and unknown potential for some or more elements to have higher concentrations than its local neighbors (potential new mineral deposits). Nevertheless, one should take into account how many of the known mineral deposits are "found" by each method. The evaluation and comparison of the methods is, however, discussed in the Section 5.9.

The point of the method from the Section 4.1 was to emphasize the difference of finding global outliers and local outliers and to introduce the concept of the spatial component and spatial local outliers. Already in Figure 5.5 (right), one can see that many observations (lines) were marked as globally outlying observations from the very beginning. It would be interesting to analyze how many exactly are globally outlying and where they are in the geographical space. As a reminder, globally outlying observations are simply the ones that have their squared Mahalanobis distance larger than the cut-off value, which is the 0.95 quantile of the squared chi-squared distribution. They are marked red in Figure 5.8, again with the known mineral deposits (+).

There are even 1582 marked global outliers. That can point to the fact that the observations in the data set are quite varying (which is true since the data set is "big" and covers a large surface), so many of them have a (global) Mahalanobis distance quite larger than the 0.95 quantile of the chi-squared distribution, which is around 17.5. Nevertheless, this fact adds to the reasons why one should use a **local** outlier detection approach for this type of problem. Same as for the output of Robust local outlier detection technique, there are two dense regions of found outliers, the cluster in the middle and the diagonal line to the right.

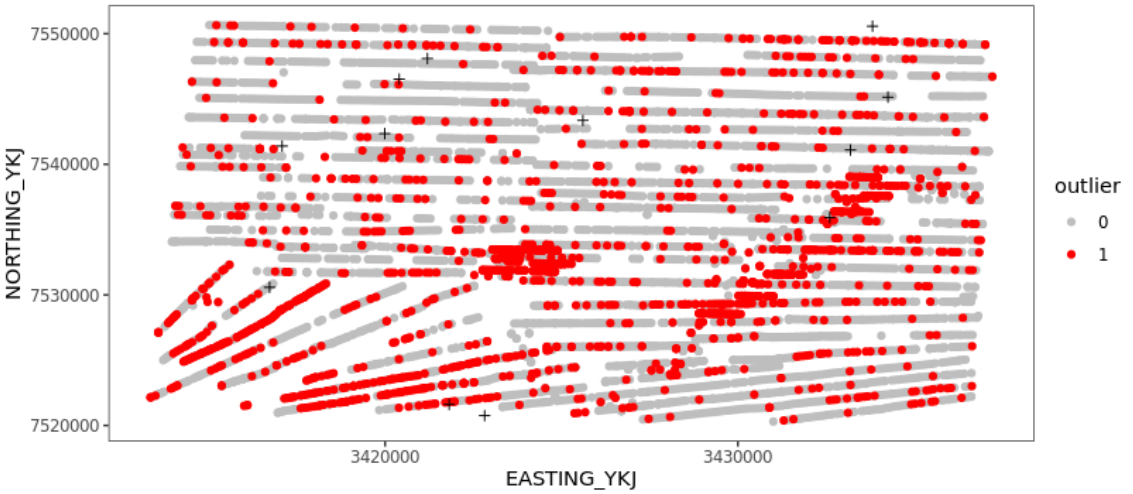

Figure 5.8: Marked (red) found global outliers in the mapsheet 2743 and known mineral deposits (symbol +)

5.4 Application of the regularized spatial outlier detection technique

The Regularized spatial outlier detection technique from Section 4.2 provides some changes for the Robust local outlier detection technique. It increases the local nature of the method by taking local covariance matrices for the calculation of the Mahalanobis distances. It also switches to the regularized MCD estimator instead of the traditional MCD and takes into account the possible heterogeneity of the local neighborhoods, where one can set the fraction of the smallest ellipsoid volumes neighborhoods that are considered further in the local outlier detection technique. With that, the parametric approach considering the chi-squared distribution is also omitted and one deals with the so-called *next distances* and the cut-off value from the upper whisker of the adjusted boxplot.

Since in the end, the goal is also to compare the outcomes and performance of each method, we will stick to the defined local neighborhoods as kNN. Moreover, the value k will still be set to 30, which means that for each observation, one is taking into account its 30 nearest neighbors (in the spatial sense).

Once the neighborhoods are defined, one can measure the volumes of the ellipsoids of the neighborhoods in the variable space and take into account just the fraction β of the smallest ones. As proposed in Ernst and Haesbroeck (2017), one runs the algorithm for a grid of proportions for β, for example $0.1, 0.25, 0.5, 0.75, 0.9, 1$, and for each of these, analyze the "next distances" of each spatial unit by means of an adjusted boxplot. For a start, let us consider all the neighborhoods, by taking $\beta = 1$. Different values of β will be explored in the last section.

The coverage h of the regularized MCD estimator and the regularization parameter λ are set as in Ernst and Haesbroeck (2017) and described in Section 4.2 of this thesis.

After all the parameters have been discussed and set, one can apply the method on the prepared data. For a start, the adjusted boxplot of the next distances of all observations is plotted.

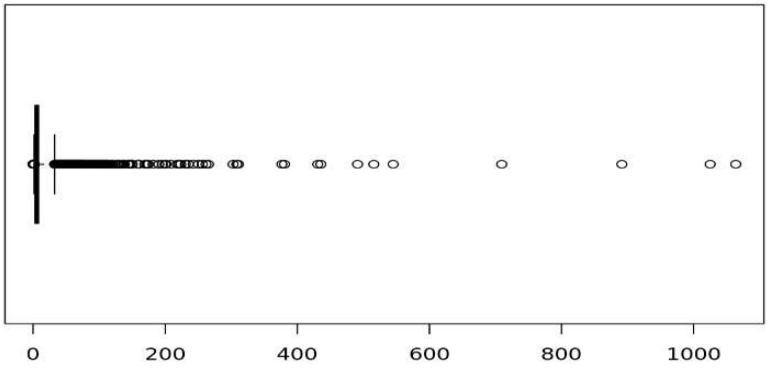

Figure 5.9: Adjusted boxplot of the next distances for the observations of the mapsheet 2743.

The adjusted boxplot after running the algorithm is shown in Figure 5.9. It is hard to interpret the boxplot, since the distribution is still heavily skewed. The Figure 4.3 from the two-dimensional example in Section 4.2 shows the boxplot a bit more obviously. However, there is an upper whisker, and all the observations with their next distances above the upper whisker are marked as local outliers.

The marked local outliers are then plotted in the spatial coordinates, again with the known mineral deposits in Figure 5.10.

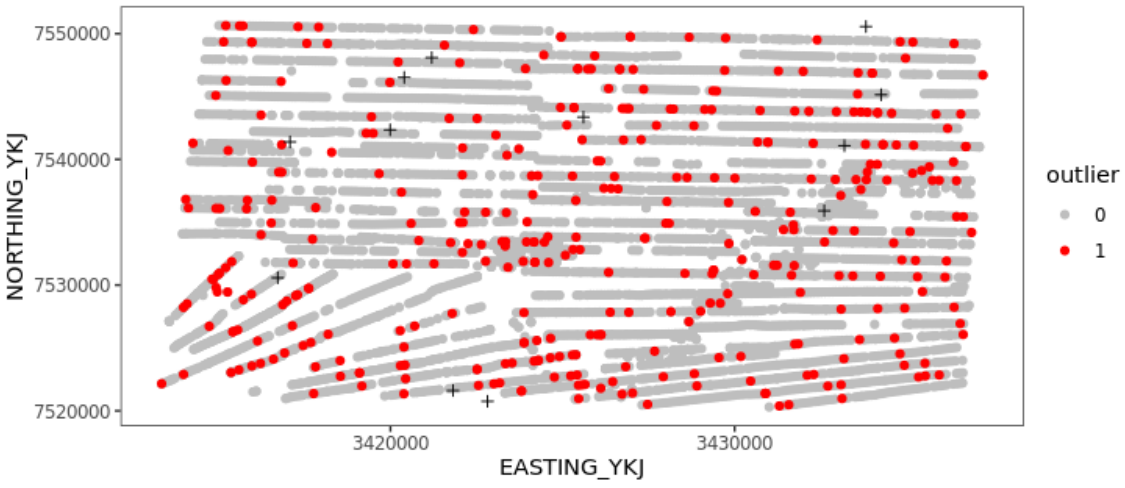

Figure 5.10: Marked (red) found local outliers with the Regularized spatial outlier detection technique in mapsheet 2743, and known mineral deposits (symbol +).

There are 361 found local outliers, which is quite a lot when compared to the number of found outliers by the Robust local outlier detection technique. However,

as already mentioned in Section 3.2, multivariate outliers are harder to find and to interpret because they are not anymore extremes along one coordinate, but could be anywhere in the whole multivariate space. One idea is to still do some analysis of the found outliers in the univariate sense. Since our goal is to find local outliers, having higher concentrations for one or more elements, one can filter the found outliers that have for at least one element a higher concentration (provided that there is some discussed definition, depending also on the context of the data, what "higher" stands for).

Nevertheless, Sections 5.9 and 5.8 provide more analyses of the identified outliers, evaluations of the methods and their comparisons.

5.5 Application of LOF

Finally, the LOF method as described in Section 4.3 is applied to the data set. In contrast to the previous two methods, LOF does not use the MCD estimates. It proposes a completely different measure of outliyngness. For each observation, the LOF (local outlier factor) is being calculated.

Again, the value k for the kNN, defining the neighborhoods should be discussed and set. Following the previous two methods, let us consider again $k = 30$. The spatially set neighborhoods will then define the $k + 1$ observations for the "cluster" in the variable space, where the LOF method is applied.

Another parameter to discuss would be the cut-off value for the LOF. All the observations having a LOF value higher than the set cut-off value will be considered local outliers. As explained in Section 4.3, a LOF value significantly larger than 1 should indicate outliers (or an observation outside of "its cluster"). For a start, we will use the cut-off value of 1.5 as a rule of thumb.

Let us once again mark the identified local outliers (Figure 5.11) in the spatial coordinates together with the known mineral deposits.

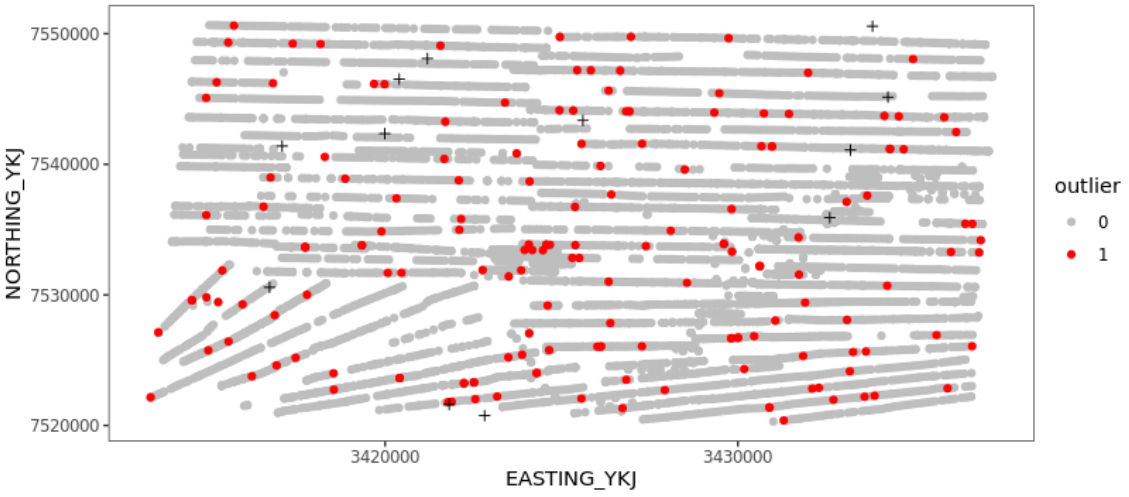

Figure 5.11: Marked (red) found local outliers with the LOF method in mapsheet 2743 and known mineral deposits (symbol +).

There were 166 found local outliers. The LOF method finds the smallest number of outliers. It is still needed, in the further analysis, to inspect the found outliers and possibly filter the ones out that seem irrelevant.

5.6 Application of ssMRCD

There is one more method (or more precisely, an extension of the methods using MCD estimators) that could be applied to our prepared data set. The idea is to not just use the regularized version of the MCD estimator (as done in the Regularized spatial outlier detection technique), but also to spatially smooth it. The underlying estimator thus constitutes a compromise between a global covariance structure and a local one for individual neighborhoods (Puchhammer and Filzmoser, 2023a). The `R` package `ssMRCD` (Puchhammer and Filzmoser, 2023b) provides the needed functions and outlier detection method. The function `local_outliers_ssMRCD` returns the indices of the found local outliers. They are again marked and plotted together with the known mineral deposits in Figure 5.12. There are 197 found local outliers.

Nevertheless, one needs to discuss, evaluate and compare all the used different local outlier detection methods. The evaluation will be based on how many known mineral deposits are found and the comparision of the methods needs to be defined as well. This will be done in Section 5.9.

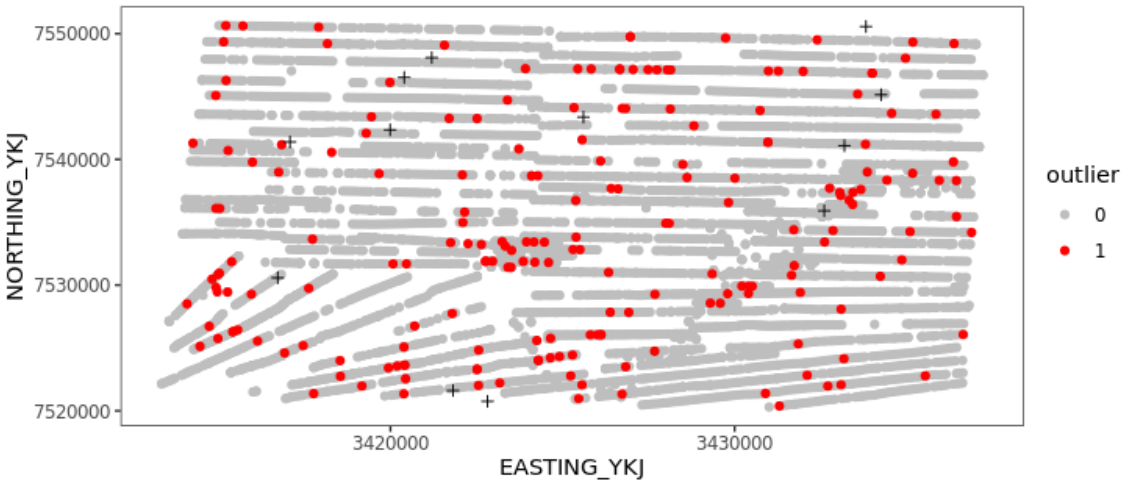

Figure 5.12: Marked (red) found local outliers with the ssMRCD method in mapsheet 2743 and known mineral deposits (symbol +).

5.7 Another example

As already described, there are a lot of mapsheets in the data set, on which we could apply (for this thesis) relevant local outlier detection methods. However, not every mapsheet has as many found mineral deposits, which we could use to evaluate the application of our methods. To show once more the outputs of the proposed methods, we will take another mapsheet from the central Lapland area, just north

of the mapsheet 2743, namely the mapsheet 2744. The sampling divided in its 6 sub-mapsheets is shown in Figure 5.13.

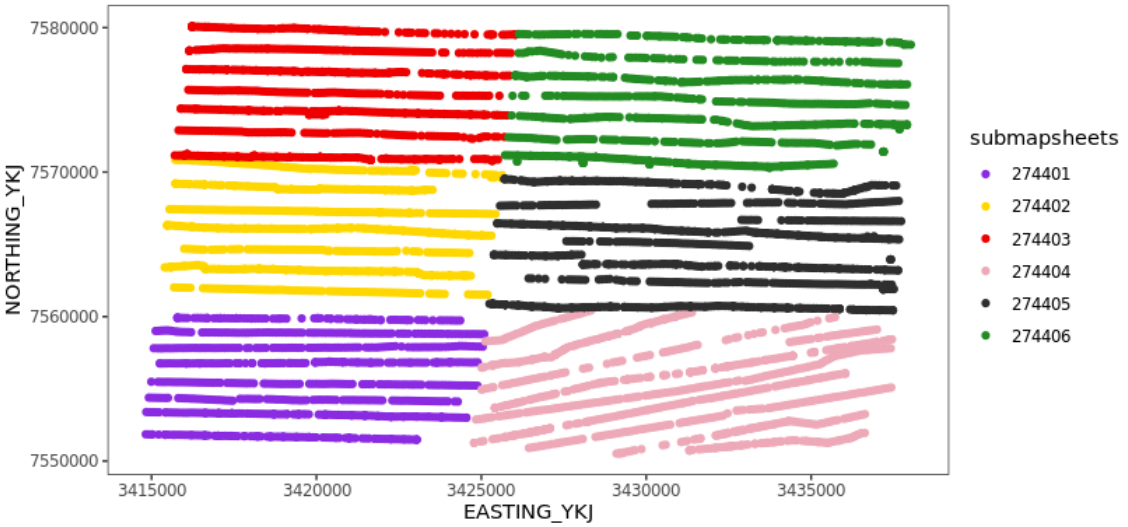

Figure 5.13: Mapsheet 2744 divided in its 6 sub-mapsheets.

The pre-processing of this data set is as described before. First, we omit completely the samples having zeros or negative values (8%). After that, we move to the filtering the strange, wrong and under detection limit values (the values marked with signs). We omit from the further analysis those elements having more than 5% of such values in its column. We are left with more elements than for the mapsheet 2743. There are 11 elements left: Si, Al, Fe, Mg, Na, Ti, V, Cr, Mn, Ni and Cu. Finally, one wants to omit the whole samples that have more than 3 elements with the marked values.

After applying those steps for the mapsheet 2744, one is still left with 4557 samples. Finally, we transform the data set into pivot coordinates before applying the outlier detection methods.

All the discussed and set parameters from applying the methods for the mapsheet 2743 stay the same, in order to have exactly the same procedure for both mapsheets. Thus, the value k for kNN in the geographical space equals 30.

There are 192 found local outliers after applying the Robust local outlier detection. The highest number of outliers is again found from the Regularized spatial outlier detection (323). The LOF method finds 150 and the ssMRCD method 145 local outliers. Figure 5.14 shows all the found local outliers marked in red, together with the known mineral deposits in that mapsheet for each method. Since the methods have a different way of defining what means to be a local outlier, it is expected that they are found on some different locations. However, none of the methods seems to catch the known mineral deposits closer together on the left lower side. The Robust local outlier detection technique finds most of the outliers in the upper right corner (sub-mapsheet 274401). This method uses the global covariance structure while finding local outliers, so the reason might be that the whole sub-mapsheet had a different sampling scheme than the other ones.

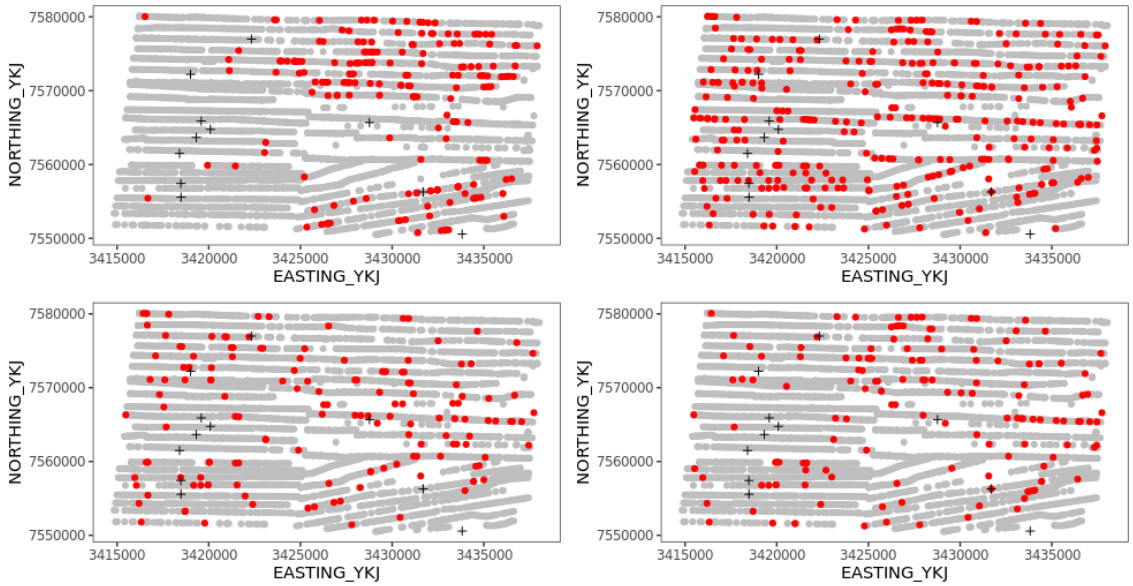

Figure 5.14: Marked (red) found local outliers in the mapsheet 2744 and known mineral deposits (symbol +) with the Robust, Regularized, LOF and ssMRCD method, respectively, top left, top right, bottom left, and bottom right

5.8 Explanation of the identified outliers

After applying the local outlier detection techniques on the prepared data set of mapsheets 2743 and 2744, one can see that different methods yielded different results. The number of found local outliers differs as well as where the local outliers are to be found. Thus, it would be of interest to try to explain the found outliers from the output of each method.

The goal of finding local outliers for the purpose of this thesis and in the context of the data, is to find new possible locations where one or more elements could be found with higher concentrations than for the other geographical points in its neighborhood. The first idea to explain the identified outliers would be to check if those local outliers have some element (at least one) with higher concentration than some decided threshold. For a start, higher concentration would mean higher value in the global sense (taking all the data points into account ignoring any spatial dependency). That way, one can filter the identified local outliers. For example, one can require that an outlier has to have at least one element value above the 0.95 quantile of all the values for this element to be considered a local outlier. We are talking about "raw" values in the units ppm and percent. This step would refine the outputs of the local outlier detection methods so that outcome and found outliers are more suitable for our goal. Moreover, the Regularized spatial outlier detection method gives by far more outliers than the other methods. Thus, it would be of interest to filter the found outliers that are (more) relevant and then evaluate and compare the methods.

To get more insight where the outlier points are for each element, one can plot univariate scatter plots and mark the values for which the outlier was found.

Figure 5.15 shows univariate scatter plots for 4 elements with marked found outliers for the mapsheet 2743 from the Robust local outlier detection technique. Moreover, they are marked from yellow to red by their outlier value, meaning how strong of an outlier they are, which for the Robust outlier detection is the *isolation degree*.

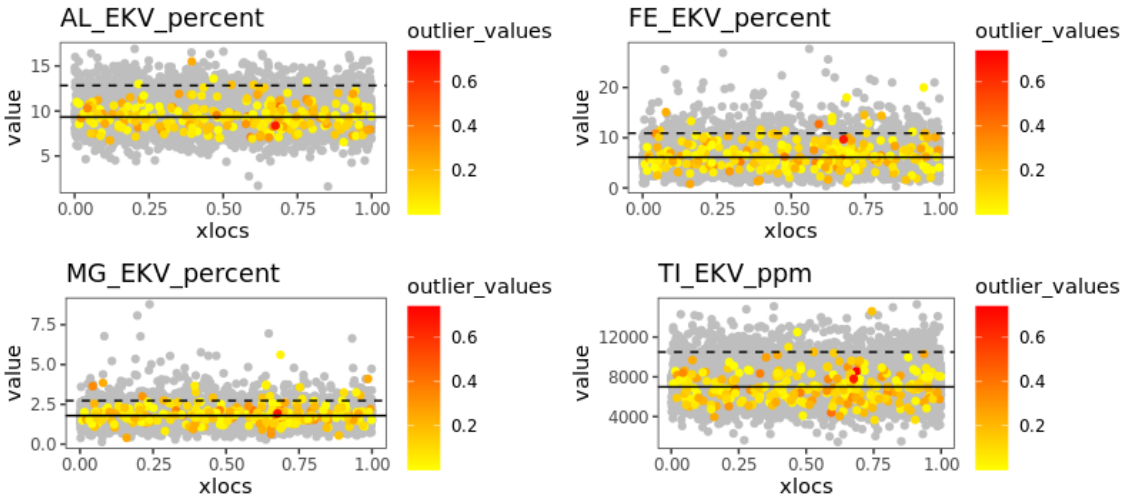

Figure 5.15: Univariate scatter plots for 4 elements with marked found outliers for the mapsheet 2743 from the Robust local outlier detection technique. The solid line represents the median value and the dashed line the 0.95 quantile value.

Globally, one can see that some more intense local outliers are plotting above the 0.95 quantile values of those 4 elements. The element that in general makes one observation a local outlier is not obvious per se. Moreover, the values of the different elements are also dependent of one another and we are dealing with compositional data. But the univariate scatter plots can still show us if there are elements with more local outliers over its 0.95 quantile value and where the outliers with different outlier values are regarding each element. As already explained, one would also like to filter the found outliers and prefer the ones that have at least one element value over 0.95 quantile.

All the showed plots were made for raw values of elements (either percentage or ppm). One can make similar plots with the clr values of each element. The clr univariate scatter plots for the same 4 elements are plotted in Figure 5.16. One might prefer the clr coordinates over raw values for this type of plots because taking clr values means log-transforming and centering the real variables which can make the visualization more insightful. By comparing Figure 5.15 and Figure 5.16, one can see that the "clouds" of clr values are more "centered" in the shown plots and the dashed line for the 0.95 quantile is closer to the full line representing the median. This means that one would also expect more marked outliers over the 0.95 quantile value. Similar to filtering outliers with the ones having for at least one element its real value over 0.95 quantile, one can filter the ones having for at least one element the clr value over 0.95 quantile.

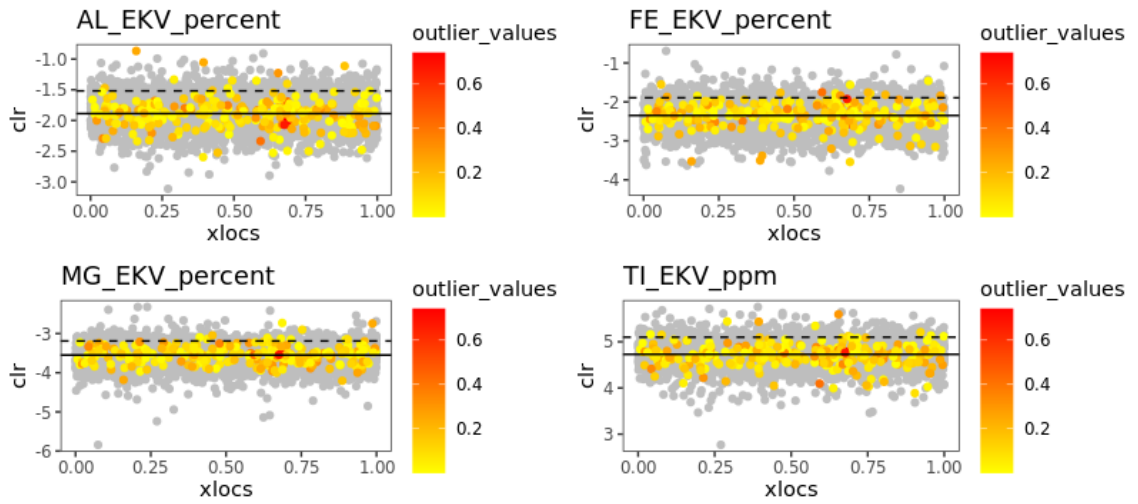

Figure 5.16: Univariate scatter plots of clr coordinates for 4 elements with marked found outliers for the mapsheet 2743 from the Robust local outlier detection technique. Full line represents the median value and the dashed line the 0.95 quantile value.

Filtering the identified local outliers

After filtering the identified local outliers with the two mentioned ways of filtering, in the Table 5.1 are the numbers of remaining local outliers from all the 4 methods and for both mapsheets 2743 and 2744.

	Mapsheet 2743			*Mapsheet 2744*		
Method	unfiltered	raw filter	clr filter	unfiltered	raw filter	clr filter
Robust	255	75	86	192	74	104
Regularized	361	233	392	323	241	280
LOF	166	124	165	150	131	147
ssMRCD	197	158	195	145	130	141

Table 5.1: Number of identified outliers for each method and for both mapsheets, unfiltered, after applying the raw values filter and after applying the clr values filter

The numbers of local outliers are always reduced, but there are more found outliers left while filtering with the 0.95 quantile values of clr coordinates than with those of raw values. It is interesting to notice that for the LOF and ssMRCD method, the clr values filter almost does not reduce the number of identified outliers. However, the evaluation of found outliers after filtering will be described later in this section.

Found global and local outliers in the spatial coordinate space

In the previous sections, the plots show where the local outliers were found together with the known mineral deposits. One can notice that some local outliers

are found close to one another. Some spatial "regions" have more outliers than the others. That is especially to be seen in Figures 5.7 and 5.8, where the local outliers from the Robust local outlier detection technique were marked as well as all the global outliers, respectively.

In both figures there is a noticeable group of outliers close together somewhere in the middle of the investigated area and another thicker diagonal "line" on the right. The sampling was already denser on those two places and different from all the other more or less horizontal lines. The reason behind finding those denser regions can be because of some different sampling on those places. As already said in Section 5.1, the sampling was made using the same analytical equipment, but some variations can be noticed between different years when the sampling took place or on the borders of the mapsheets. However, the denser outlier regions do not seem to follow especially the borders of some mapsheets.

In Figure 5.17, the samples from mapsheet 2743 are plotted with marked year of their sampling. One can clearly see where the dense groups of found outliers from Figures 5.7 and 5.8 come from. There was another sampling in the years 1975 and 1977 exactly on those spots. Since global covariance structure is taken into account while finding global outliers in the Robust local outlier detection method, it seems that those years have been sampled somehow differently than in the other two years. Locally they are still found near other samples from the years 1973 and 1974, thus they are easily marked as local outliers because for the Robust local outlier detection technique the global covariance structure was taken into account.

One solution to this problem could be simply omitting the samples sampled in 75 and 77. But other methods like the Regularized spatial outlier detection technique and the LOF method do not use global covariance structure, so from Figures 5.10 and 5.11 one can notice that those dense groups of outliers are not present anymore.

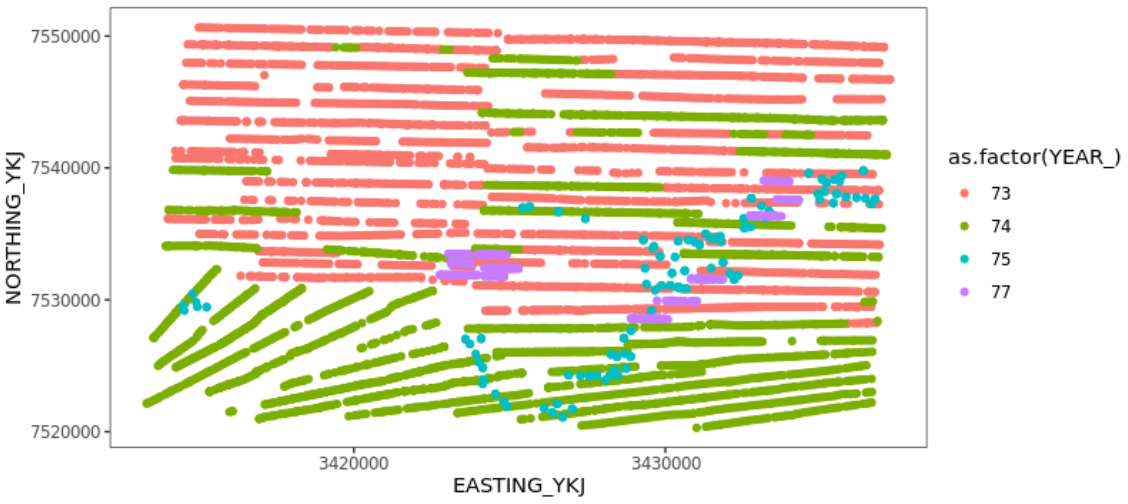

Figure 5.17: Samples from the mapsheet 2743 per year when the sample was collected.

5.9 Overall evaluation and comparison

The main reason for trying out different local outlier detection methods is to evaluate them and compare their results. After the application of the methods, each of them gives specific results that are finally to be analyzed and compared.

The evaluation of the performance of any method is usually a problem for itself. Without some further information it is hard to conclude if our methods are actually finding real local outliers as we would like to.

For this specific problem we have the known mineral deposits in the mapsheets taken into account. They are places where the sources of one or more elements are already found and used. The goal is then that our local outlier detection techniques "find" those mineral deposits. For this we need to define what does it mean for a mineral deposit to be "found".

One could define some radius around a mineral deposit that one wants at least one local outlier to fall into for this mineral deposit to be "found". Here, 1000 units in the investigated area (coordinates EASTING_YKL, NORTHING_YKL) equals 1 km. Since the horizontal lines of the sampling are $1-2$ km apart from each other, let us set the wanted radius to 2 km. This means if in the 2000 units of radius around a mineral deposit there is at least one local outlier to be found, this mineral deposit is considered to be found. Thus one can evaluate a method by how many mineral deposits were found by its (filtered) local outliers.

However, different methods found different numbers of outliers. Naturally, the more outliers the method identified, the more mineral deposits would be found with those local outliers. It would be more meaningful to compare how "fast" is a certain number of mineral deposits found by each method.

Each method has some outlier value: isolation degree for the Robust method, next distance for the Regularized method and LOF value for the LOF method. Thus, outliers can be sorted by how strong of an outlier they are, from the ones with the biggest outlier value until the ones with the smallest. Then, taking the local outliers that are already filtered as mentioned before, one can calculate how many mineral deposits are found with respect to how many of the found local outliers one takes into account by the order with the outlier value. Therefore, for each method one takes the strongest local outlier and looks how many mineral deposits are found by this one outlier (usually still none of them are found). Then one takes the second strongest to see how many mineral deposits are found by those two, and so on. If one plots the number of mineral deposits found with respect to the number of outliers taken into account, those methods with graphs having the biggest area under the curve are identifying the mineral deposits the fastest.

The methods are then easily compared for the same amount of outliers taken into account and also by the total amount of mineral deposits found.

In mapsheet 2743 there are in total 12 mineral deposits. Figure 5.18 shows the mentioned curves for all four applied outlier detection techniques those identified outliers which are left after filtering by their raw element value as mentioned before.

The curves show clearly how well each method is finding the mineral deposits. For example, let us consider the ssMRCD (green curve) and the Regularized method (red curve). The red curve reaches the longest to the right since the Regularized

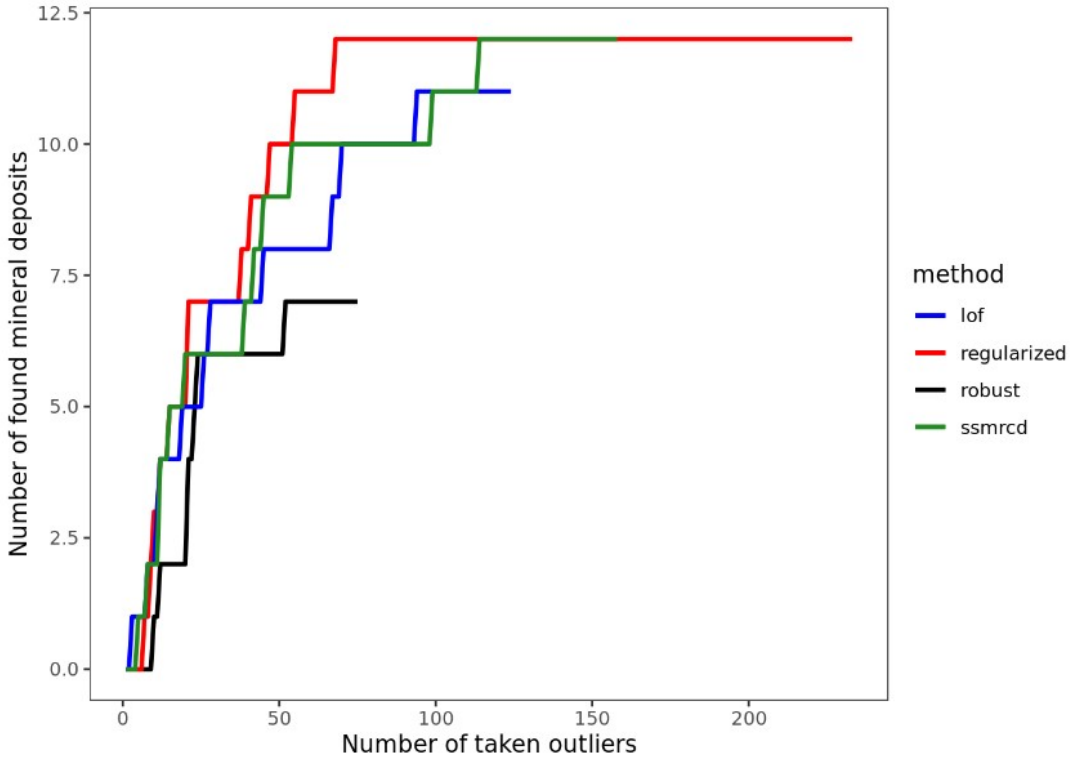

Figure 5.18: The graphs for the 4 outlier detection methods with the number of identified mineral deposits with respect to the number of raw value filtered outliers taken into account for the mapsheet 2743.

method is finding the highest number of outliers, but they both reach the highest number of found mineral deposits (12). However, if it is enough to have just 10 deposits found, then the red line (and the Regulirezed method) finds 10 mineral deposits "the fastest" (it reaches 10 found mineral deposits with the lowest number of outliers taken into account).

The methods can be evaluated and compared by the same plot taking into account one by one outlier after they were filtered with respect to the clr values and 0.95 quantile as described before. Figure 5.19 shows the 4 curves for the 4 applied methods taking into account the clr values filtered outliers.

The results change just slightly in Figure 5.19 compared to Figure 5.18. If one filters the outliers by the elements' clr values, both methods, the Regularized spatial outlier detection method and the ssMRCD find again all the mineral deposits. They do not find them "at the same time"; the Regulirezed method seems to perform "faster". However, the ssMRCD, after taking into account outliers after applying the raw value filter, found all the mineral deposits already with around 120 outliers taken. The Robust local outlier detection method finds more mineral deposits than the output outliers after the raw value filter, namely 8 instead of just 7. For the LOF method's performance one can draw similar conclusions from both ways of filtering found outliers and both plots. However, after filtering the identified outliers by their clr values, the curves are more similar to one another and they seem to have a

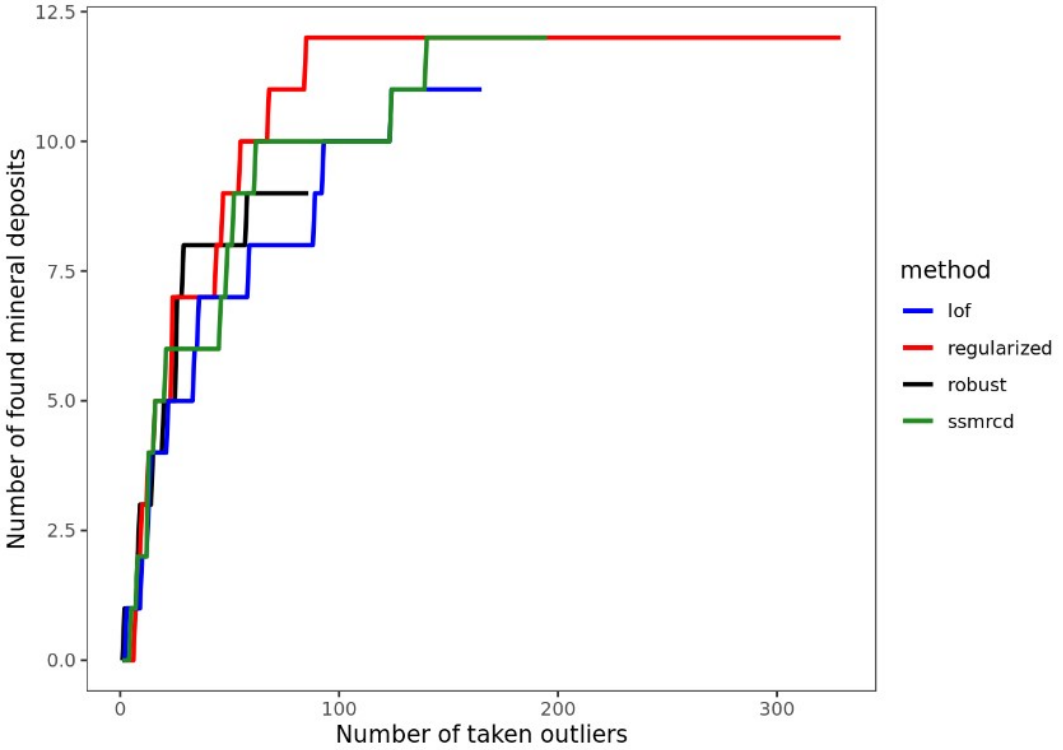

Figure 5.19: The graphs for the 4 outlier detection methods with the number of found mineral deposits with respect to the number of clr value filtered outliers taken into account for the mapsheet 2743.

bigger area under the curve. Nevertheless, the choice of filtering outliers as well as the choice of the method used are strongly dependent on what one wants to achieve. That is why the curves are made by how many outliers are taken into account while finding mineral deposits. That way, one does not compare just how many mineral deposits are found but also "how fast".

As another example, the data from the mapsheet 2744 were also prepared and the 4 local outlier detection methods were applied to the data explained in the Section 5.7. As for mapsheet 2743, one can make the plots to evaluate and compare the 4 methods for the data set of mapsheet 2744. Figure 5.20 shows the 4 curves for the outliers taken into account after the raw value filter (left) and the clr value filter (right).

There are 11 known mineral deposits in mapsheet 2744. For this example, just the Regularized method seems to find all the mineral deposits. The LOF and the Regularized methods are outperforming the ssMRCD if one considers the total amount of mineral deposits found. The LOF method also finds 9 mineral deposits the fastest, which is the point where it outperforms the Regularized method but does not find more mineral deposits than that. The curves for both choices of filtering outliers are quite similar, but as before, the number of outliers taken into account after the clr value filter is bigger for all the 4 methods.

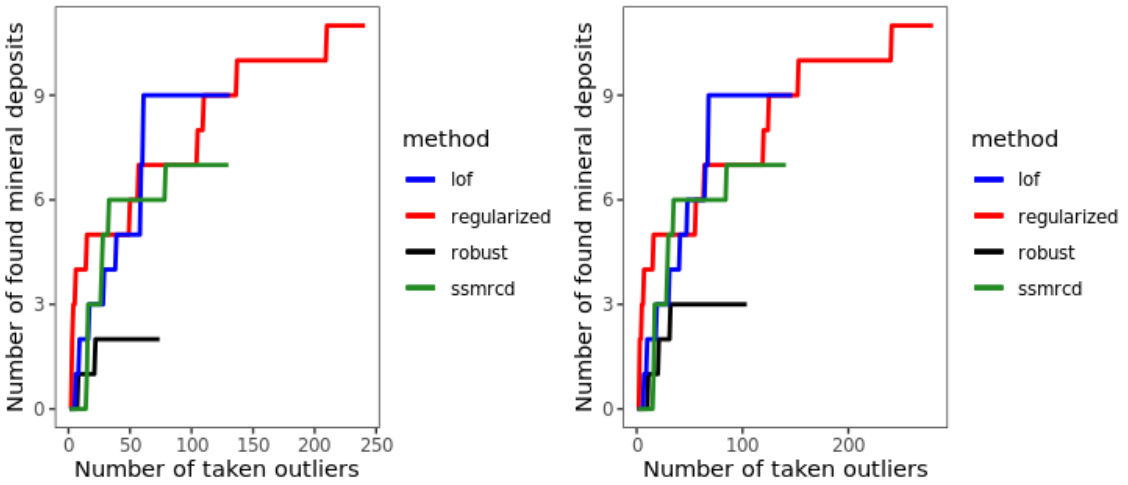

Figure 5.20: The graphs for the 4 outlier detection methods with the number of found mineral deposits with respect to the number of raw value filtered (left) and clr value filtered (right) outliers taken into account for the mapsheet 2744.

5.10 Parameters

Finally, one would want to get back and discuss some parameters that were set in the application of the outlier detection techniques. All the parameters were set without much discussions and usually by some propositions seen in practice. In this last section, we would like to try out some different parameter settings and compare the results for mapsheet 2743 per each method.

The parameters of the Robust local outlier detection technique

The first proposed local outlier detection method seen in this thesis was the Robust local outlier detection technique. Not taking into account the parameter k for the size of neighborhoods, there are two parameters to be set for this method. The parameter β, defining a fraction of the data in the local neighborhood, that are allowed to be similar to the observation at hand, and the cut-off value for the $\alpha(i)$-quantile of the chi-square distribution of the Mahalanobis distances. The cut-off value is usually set to be two times larger than β. Already seen in practice is to take β equal 0.1 and then the cut-off value 0.2, but analyzing the plots with the tools from the package `mvoutlier` (see Section 5.3), one notices that the parameters should be set higher (namely 0.3 for β and 0.4 for the cut-off value). Analyzing the plots from Figure 5.6 one notices that for β set to be 0.1 one would probably not catch all the local outliers. We would expect a small number of found local outliers.

After the parameter setting as in Section 5.3 one gets 255 identified local outliers. However, with the parameter $\beta = 0.1$ and the cut-off value 0.2 one identifies only 62 local outliers (the parameter k stays 30 for the whole time being). They are marked in Figure 5.21 again with the mineral deposits from mapsheet 2743.

Obviously, the parameter setting with $\beta = 0.1$ would find a far too small number of local outliers and performs worse. When compared to the other local outlier

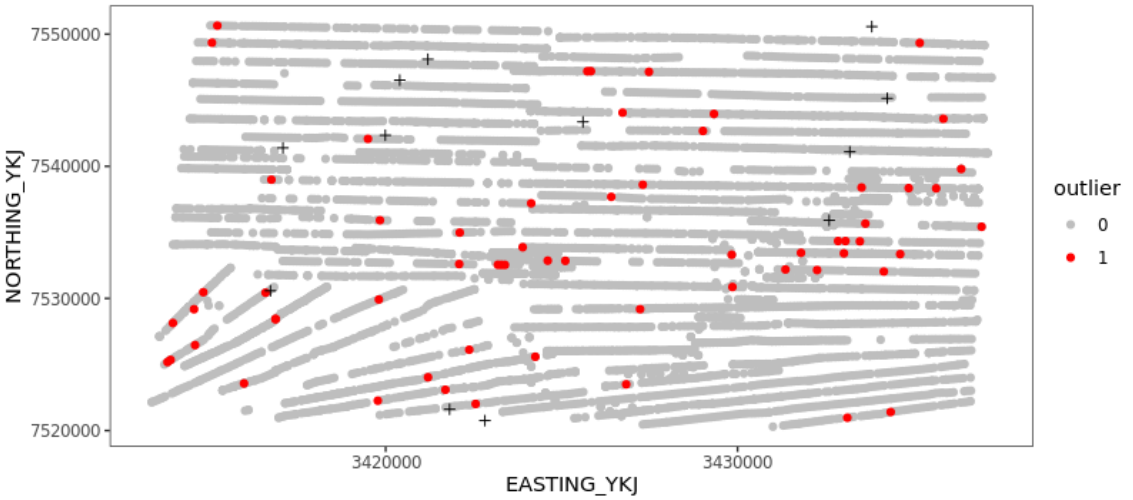

Figure 5.21: Marked (red) found local outliers with the Robust local outlier detection technique with the parameter $\beta = 0.1$ in mapsheet 2743 and known mineral deposits (symbol +).

detection methods and with the data set of over 6000 samples, 62 is a too small number to identify all the wanted local outliers.

Why setting the cut-off value lower than two times β when deciding for $\beta = 0.3$? Simply because from Figure 5.6, one can see that for 0.3 on the x-axis and 0.6 for the y-axis one again jumps out of the "dense" region and identifies too few outliers. For $\beta = 0.3$ and the cut-off value 0.6, one would find 92 local outliers which is simply again not performing well enough when compared with other methods.

The parameters of the Regularized spatial outlier detection technique

For the parameters of the Regularized spatial outlier detection technique, the coverage of the regularized MCD estimator h and the regularization parameter λ will not be furthered discussed but set as in Ernst and Haesbroeck (2017) and described in Section 4.2.

With the exception of those two parameters, one can set and analyze the parameter β, for the fraction of the smallest volumes neighborhoods taken into account for the local outlier detection. One runs the algorithm for a grid of proportions for β, for example $0.1, 0.25, 0.5, 0.75, 0.9, 1$, and for each of these, analyze the "next distances" of each spatial unit by means of an adjusted boxplot. Until now, in Section 5.4 we considered all the neighborhoods by taking $\beta = 1$. As a reminder, the Regularized method finds by far the largest number of local outliers. Taking β too big might tend to increase the false detection rate. Let us consider kicking out some of the samples whose neighbors have a heterogeneous pattern from the outlier detection by setting, for example, $\beta = 0.5$. The parameter k for the size of the neighborhoods is set to 30 as for all the applications until now.

The Regularized method with $\beta = 1$ identified 361 local outliers in mapsheet 2743, but by setting $\beta = 0.5$, one identifies just 198 local outliers. They are marked in Figure 5.22 again with the mineral deposits from mapsheet 2743.

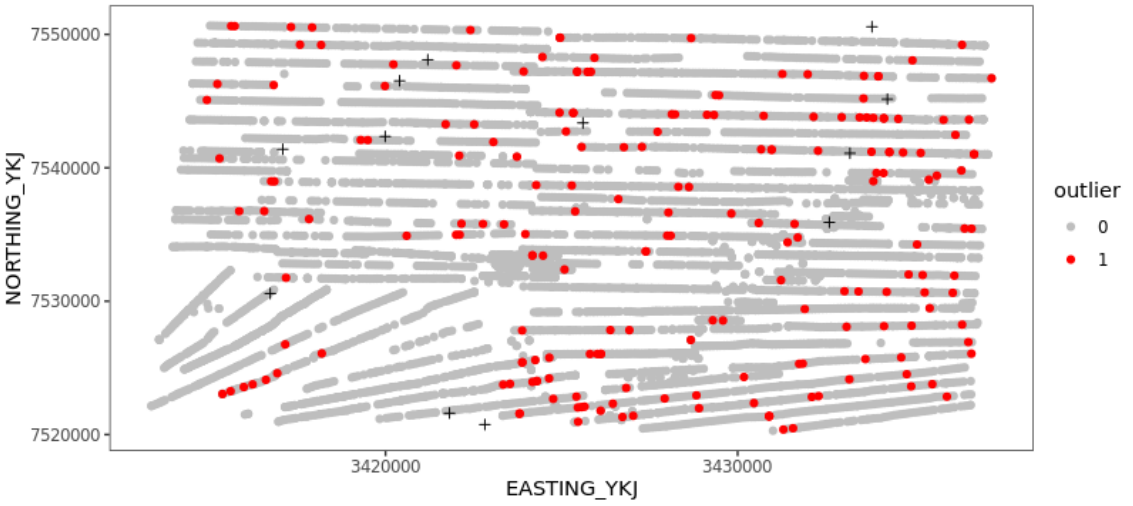

Figure 5.22: Marked (red) found local outliers with the Regularized spatial outlier detection technique, with the parameter $\beta = 0.5$ in mapsheet 2743 and the known mineral deposits (symbol +).

However, with fewer outliers found, it is obvious one does not seem to catch more mineral deposits. Setting the β parameter lower would lower the number of found local outliers but that does not mean that the rest of the outliers were filtered meaningfully for our purpose. Nevertheless, it depends strongly on the context and the goal if one wants to even think about omitting the more heterogeneous neighborhoods, and the question is why would somebody even want to do that. Even for a more heterogeneous neighborhood one still wants to say if the observation at hand is an outlier in its neighborhood and most probably not omit the whole observation from further analyses.

The parameters of the LOF method

The LOF method proposes a completely new measure of outliyngness, called the Local Outlier Factor. For this method there are no parameters to be set but the cut-off for the LOF value when the observation is to be considered an outlier. The LOF value significantly larger than 1 should indicate outliers or an observation outside of "its cluster". The rule of thumb would be 1.5 as used in Section5.5.

Instead of just setting the cut-off value, in this subsection one wants to analyze more closely the calculated LOF values of the observations in mapsheet 2743. After applying the LOF method, each observation has a LOF value calculated for it. Figure 5.23 shows the density of the LOF values of all the observations in the data set. This way one can get an impression about the distribution of LOF values in this data set.

The vertical line marks the value 1.5 which is our cut-off value. One can notice that the right tail is more unstable and deviates from normality and the curve that one would find as normal. Also the value 1 is right in the middle of the density plot where the peak of the distribution happens. Setting the cut-off value simply just as 1 would identify far too many observations as outlying. The value 1.5 seems to be

a good rule of thumb since it is further away in the tail of the unstable part of the distribution.

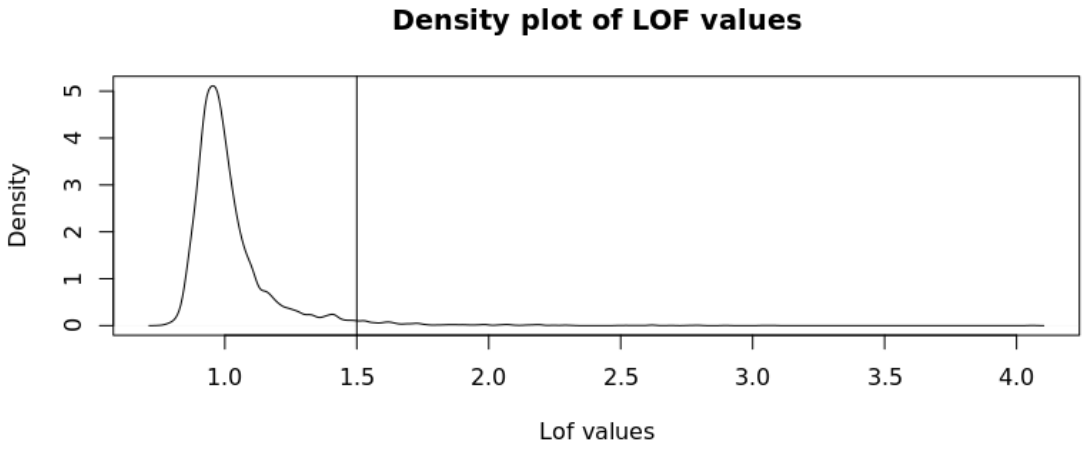

Figure 5.23: The density plot of the LOF values from the observations in mapsheet 2743.

5.11 Parameter k for kNN

In order to do the local outlier detection one needs to define local neighborhoods. As explained in Section 4.1, there are two possible ways of defining local neighborhoods: by the radius d_{max} or by setting k for the kNN distance. For the sampling as we have in mapsheet 2743 it does not make much difference what type of definition of local neighborhoods one uses. Thus, we used kNN for all the applications in this thesis. In other words, one is taking into account 30 nearest neighbors (in the spatial sense) to define the local neighborhood of an observation at hand.

The parameter k was analyzed more closely for the application of the Robust method in Section 5.3 together with the parameter β. The value $k = 30$ fits the purpose, since for values higher than 30 the plots in Figure 5.5 do not stabilize any more and the outcome does not seem to change that much.

For a comparison of the methods, it made sense to fix $k = 30$ for all the methods applied. However, in this subsection one wants to at least check the results one would get by setting some other (more extreme values for k). Let us compare the results of the Regularized spatial outlier detection technique and the LOF method when setting the parameter k to $30, 10$ and 100, while leaving all the other parameters set as in the application sections for those two methods.

The application of the Regularized spatial outlier detection technique

In principle, one would expect less identified outliers from the Regularized method while setting the parameter k bigger. For the larger local neighborhoods one expects a more homogeneous neighborhood taken into account and less chances for an observation to be identified as significantly different than the rest of its neighbors.

In the following we give an overview of the number of found outliers for each of the parameters k:

- $k = 30$: 361 found outliers
- $k = 10$: 933 found outliers
- $k = 100$: 207 found outliers

To compare the performance of the application of the Regularized method for the three different k parameters, one can make similar curves as in Section 5.9. We are once again going to filter the identified outliers by the raw and clr values of the elements. Figure 5.24 shows the 3 curves for the outliers taken into account after the raw value filter (left) and the clr value filter (right) for these 3 parameters k.

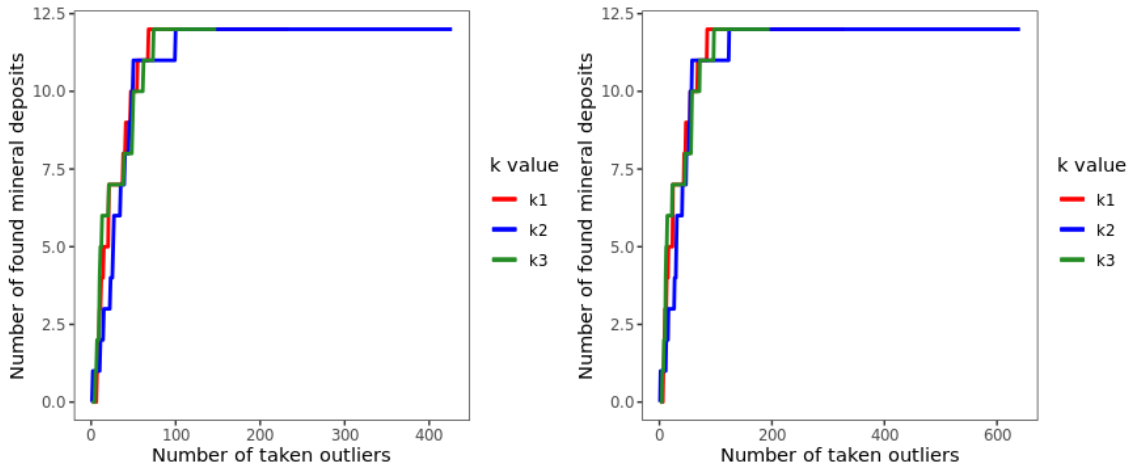

Figure 5.24: The graphs for the 3 values of the parameter k ($k1 = 30$, $k2 = 10$, $k3 = 100$) in the application of the Regularized method; with the number of found mineral deposits with respect to the number of raw value filtered (left) and clr value filtered (right) outliers taken into account for mapsheet 2743.

It is interesting to see that the performance of finding mineral deposits with the identified local outliers is pretty much the same for differetn values of k. The reason behind it could be that the first outliers with the largest outlier value (next distance for the Regularized method) are always identified, not depending of the size of their local neighborhood taken into account. As a reminder, the Regularized method was boosting the local nature of the local outlier detection by calculating the robust MCD estimator for each neighborhood.

The application the LOF method

For the LOF method, the parameter k defining the spatial local neighborhoods will define the same neighborhood of an observation in the variable space where the LOF value will be calculated. How the "cluster" looks like in the variable space is then crucial for the outcome of the LOF values. One cannot predict if some lower value k will yield less or more identified outliers.

However, in the following we give an overview of the number of found outliers for each of the parameters k:

- $k = 30$: 166 found outliers
- $k = 10$: 178 found outliers
- $k = 100$: 181 found outliers

For both $k = 10$ and $k = 100$ one gets more outliers in the output than for $k = 30$. There is no special rule to determine or filter how many outliers one wants to find by setting the parameter k.

In Figure 5.25 one again compares the curves of the numbers of taken outliers versus the number of found mineral deposits. One again filters the outliers by the raw and clr values as in Section 5.9.

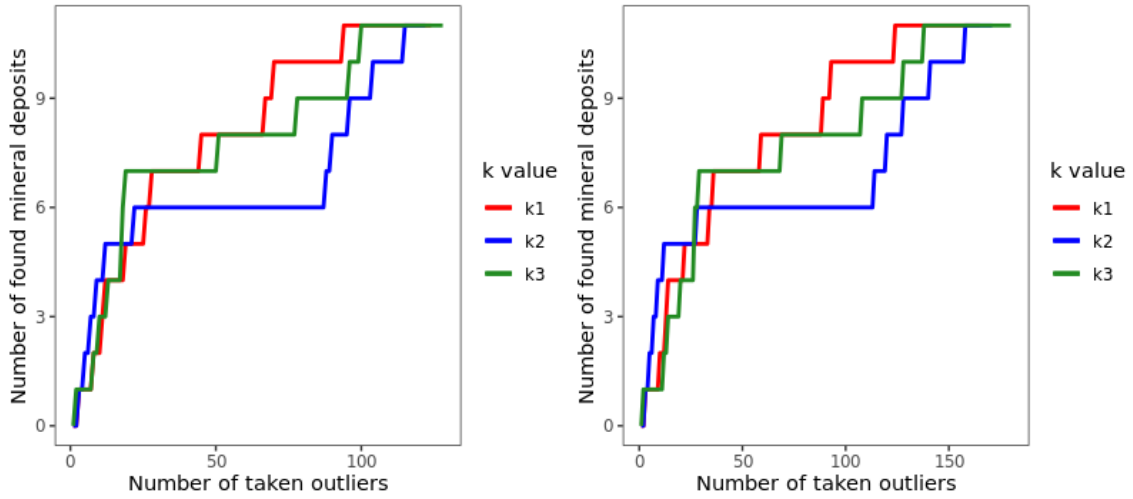

Figure 5.25: The graphs for the 3 values of the parameter k ($k1 = 30$, $k2 = 10$, $k3 = 100$) in the application of the LOF method; with the number of found mineral deposits with respect to the number of raw value filtered (left) and clr value filtered (right) outliers taken into account for mapsheet 2743.

It seems to not matter if one filters the outliers by the raw or clr values of the elements. The LOF method finds the same amount of mineral deposits by all the three parameters k. However, it seems that it found the mineral deposits "slower" for $k2 = 10$ than for the values 30 or 100. Meaning, if one for some reason wants to take just the first 70 outliers with the largest outlier values (LOF value), the parameter setting with $k = 30$ or $k = 100$ would find more mineral deposits than the setting with $k = 10$. In that case, the decision of the value k would be context and goal dependent. Nevertheless, the choice of the value $k = 30$ fits our purpose in the application.

6 Discussion and Conclusions

The point of this thesis was to introduce and apply local outlier detection techniques to compositional data. After the application of the methods, each of them gives different results that were finally to be analyzed for more insight and understanding. The spatial coordinate plots give different and interesting locations of identified outliers together with the known mineral deposits.

6.1 Data

The first noticeable thing about the spatial components of the data of both used mapsheets is the sampling. The sampling was made along some more or less horizontal lines with randomly added sampling from some other years (see Figure 5.1). The application of the spatial outlier detection and evaluation methods would be easier if the sampling was done more uniformly in the investigated area. It would cover more intensively the whole space (mapsheet) and identifying the mineral deposits would not be dependent on the position of the mineral deposits in or out of the horizontal line. That is why the evaluation and the definition of what it means to "find" a known mineral deposit needs to be considered carefully.

In the data description one finds out that some variations can be seen although all samples have been analyzed using the same analytical equipment. This sets a bigger challenge for the (local) outlier detection methods because one wants to detect outliers whose outlyingness lies in the real element values and not in some external factors. The external factors are also detection limits of the analytical tools. Because of that, the data preparation and cleaning was done carefully and the data were prepared in the sense of their compositional nature. All the zero and negative values are by definition out of question and one needs to apply the appropriate preprocessing that does not distort the real-world data.

Moreover, the local outlier analysis depends strongly on the part of the data set taken for the analysis (mapsheet(s) taken into account). The data from different (sub-)mapsheets need to be comparable for all the elements that the analysis is done for (see Figure 5.4). Thus it is not possible to take randomly selected mapsheets at the same time while applying the outlier detection methods.

However, the spatially dependent geochemical data are a perfect example data set for the thesis at hand. It is a great example of compositional data, where the relative information between different components (elements) form the relevant information about the data. The spatial component is needed for the local part of the outlier detection, where the outlyingness happens in the smaller local neighbourhoods and not (just) in the global sense of the whole data set.

6.2 Application of the methods

The evaluation of the applied methods on the mapsheets at hand are a subject on its own. Since there are known mineral deposits spatially positioned in the mapsheets, the strategy used was to define what it means for a mineral deposit to be found by the identified local outliers.

It is not the best strategy since every mineral deposit has its peculiarities, and in this thesis, we do not go into details of which element(s) make an observation a (local) outlier. Also, the mineral deposits and the samples around could yield global outliers but not outliers in the local sense, or in our defined local neighbourhood, thus they might not be identified as local outliers by the methods applied in this thesis.

From Figure 5.18, where the methods were compared by how well are their returned outliers identifying the known mineral deposits, one could conclude that the ssMRCD and the Regularized method perform the best. ssMRCD is the method that extends the Robust and Regulirezed method by spatially smoothing the MCD estimator (Puchhammer and Filzmoser, 2023a). This method is identifying the mineral deposits "fast", and with a bit over 100 strongest outliers it finds all the known mineral deposits. The LOF method is, however, not far behind, although it is the most different technique in this thesis with a completely different logic behind calculating the outlier values (Breunig et al., 2000).

The local outliers were, however, tried to be analysed and explained. In the plots as in Figures 5.15 and 5.16 one wants to determine where the found outliers are located in the univariate sense per element. However, from the plots one can conclude that for each element, the outlier values still are concentrated around the median line. With that conclusion, it is hard to determine what makes local outliers real outliers.

The ssMRCD is the newest method of all the described and applied methods for the purpose of this thesis and does perform really well, even with the challenges of the data mentioned before. It shows that some previously introduced local outlier detection methods indeed have potential for improvement, changes and adaptation to better fit the real-world data and problems.

Nevertheless, compositional data as well as local outlier detection both stay in the peak of the interests and ongoing research for mathematicians, statisticians as well as computer and data scientists; as long as new challenging data and problems arise.

References

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

Aitchison, J. and Kay, J. W. (2003). Possible solution of some essential zero problems in compositional data analysis. In Thió-Henestrosa, S. and Martín-Fernández, J., editors, *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings*. Universitat de Girona, Spain.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 93–104.

Ernst, M. and Haesbroeck, G. (2017). Comparison of local outlier detection techniques in spatial multivariate data. *Data Mining and Knowledge Discovery*, 31:371–399.

Filzmoser, P. and Gregorich, M. (2020). Multivariate outlier detection in applied data analysis: global, local, compositional and cellwise outliers. *Mathematical Geosciences*, 52(8):1049–1066.

Filzmoser, P. and Gschwandtner, M. (2021). *mvoutlier: Multivariate Outlier Detection Based on Robust Methods*. https://CRAN.R-project.org/package=mvoutlier.

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

Filzmoser, P., Hron, K., and Templ, M. (2018). Applied compositional data analysis. with worked examples in R. *Cham: Springer*.

Filzmoser, P., Ruiz-Gazen, A., and Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers*, 55:29–47.

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.-B., and Thirion, B. (2011). Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part III 14*, pages 264–271. Springer.

Geological Survey of Finland (2013). The local geochemical mapping of till in 1971-1983. GTK, Espoo, Finland.

Geological Survey of Finland (2016). Mineral deposits. GTK, Espoo, Finland.

GeoSphere Austria (2021). Monthly weather data. Data retrieved from https://data.hub.zamg.ac.at/dataset/klima-v1-1m.

Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, 52(12):5186–5201.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60(359-367):489–498.

Puchhammer, P. and Filzmoser, P. (2023a). Spatially smoothed robust covariance estimation for local outlier detection. arXiv 2305.05371.

Puchhammer, P. and Filzmoser, P. (2023b). *ssMRCD: Spatially Smoothed MRCD Estimator*. https://CRAN.R- project.org/package=ssMRCD.

Reimann, C., Filzmoser, P., Garrett, R., and Dutter, R. (2011). *Statistical data analysis explained: Applied environmental statistics with R*. John Wiley & Sons.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Salminen, R. (1995). *Alueellinen geokemiallinen kartoitus Suomessa vuosina 1982-1994*. Tutkimusraportti. Geologian Survey of Finland.

Seheult, A., Green, P., Rousseeuw, P., and Leroy, A. (1989). Robust regression and outlier detection. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 152:133.

Templ, M., Hron, K., and Filzmoser, P. (2017). Exploratory tools for outlier detection in compositional data with structural zeros. *Journal of Applied Statistics*, 44(4):734–752.