Journal of Geochemical Exploration 258 (2024) 107392

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

A performance study of local outlier detection methods for mineral exploration with geochemical compositional data

Patricia Puchhammer [a,*], Charmee Kalubowila [b], Lorena Braus [a], Solveig Pospiech [c], Pertti Sarala [b], Peter Filzmoser [a]

[a] TU Wien, Institute of Statistics and Mathematical Methods in Economics, Wiedner Hauptstraße 8-10, Vienna, Austria
[b] University of Oulu, Oulu Mining School, P.O Box 3000, Oulu, Finland
[c] Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Bautzner Landstr. 400, Dresden, Germany

ARTICLE INFO

Keywords:
Robust statistics
Logratio transformation

ABSTRACT

In exploration geochemistry, mineral deposits are typically characterised by an enrichment of the targeted elements, and thus their element composition differs from that of samples in a local neighbourhood. Local outlier detection methods aim at identifying local changes. In contrast to conventional outlier detection procedures, local outlier detection methods are multivariate methods for outlier identification that incorporate the spatial neighbourhood of the samples. It is essential that geochemical data are treated as compositional data, and the requirements for their treatment depend on the specific local outlier detection method. We demonstrate how prominent local outlier detection methods can be used for mineral exploration with geochemical data that vary in scale, in the sampling density, and in data quality. The methods are compared based on known mineralisations, and recommendations for their usefulness are provided.

1. Introduction

Detecting multivariate outliers is one of the most important steps when analysing any kind of data. Such outliers could arise from gross errors during data recording, they could be the result of inappropriate data preprocessing, or they could indicate observations which are indeed very different from the rest and thus point at unusual phenomena (Zimek and Filzmoser, 2018). The problem of outlier detection becomes more difficult when analysing data with additional attributes that need to be considered, such as the locations of observations in a spatial data setting. Here, we are often not interested in the outliers found with standard methods (so-called *global outliers*) but we focus on observations that are anomalous with respect to their spatial surrounding. These observations are called *local outliers*, and they could indicate interesting locations to practitioners, e.g., unknown mineral deposits. On the other hand, methods which use locality (for example geographically weighted methods (e.g. Brunsdon et al., 1998) or geostatistical techniques (see e.g. Cressie, 2015) can also be heavily influenced by local outliers.

While the literature for local outlier detection is not as broad as for global outlier detection, there are still some (multivariate) methods available. We will focus on three methods based on the *pairwise Mahalanobis distance* (see Filzmoser et al., 2013; Puchhammer and Filzmoser, 2023a) defined as

$$\mathrm{MD}_{\Sigma}(x,y) = \left[(x-y)'\Sigma^{-1}(x-y)\right]^{1/2} \quad \text{for } y \in A(x)$$

for two (multivariate) observations x, y with a robust covariance estimate Σ which can depend on the spatial attributes of x and y. The set-valued function $A(x)$ returns observations that are spatially close to an observation x. The three methods differ in their covariance estimation, specifically in the degree of its locality. The fourth method comes from the area of machine learning and is solely distance-based. All of the four methods compare each observation with its k-nearest neighbours ($A(x)$ returns the k-nearest spatial neighbours) and calculate a degree of outlyingness that together with a method-specific cutoff value flag observations as outliers.

The first method introduced by Filzmoser et al. (2013), in the following called *robust local outlier detection method* (ROB), is available in the R-package mvoutlier (Filzmoser and Gschwandtner, 2012) and uses the pairwise Mahalanobis distances together with a global and robustly estimated covariance matrix, ignoring the spatial context of the data. The measure of outlyingness for each observation is based on theoretical

* Corresponding author.
E-mail address: patricia.puchhammer@tuwien.ac.at (P. Puchhammer).

https://doi.org/10.1016/j.gexplo.2024.107392
Received 2 October 2023; Received in revised form 14 December 2023; Accepted 20 December 2023
Available online 9 January 2024

properties connected to χ^2-quantiles. For more details we refer to the respective paper by Filzmoser et al. (2013). In contrast, the method of Ernst and Haesbroeck (2016), here called *regularised spatial detection technique* (REG), estimates local covariance matrices based on the k-nearest spatial neighbours for each observation separately. Thus, for a fixed observation x, the covariance estimation is only based on observations in $A(x)$. The measure of outlyingness (also called next-distance) is just the minimum of all MD, $min_{y\in A(x)}\mathrm{MD}(x,y)$ of each observation x, and the final cutoff value to determine outlyingness is the upper fence of an adjusted boxplot based on all next-distances. Next-distances above the cutoff value indicate local outliers. As a compromise between using only one covariance estimation and using a covariance estimation for the local neighbourhood of each observation individually, the third method of Puchhammer and Filzmoser (2023a) is bridging the gap by partitioning the space into groups (e.g. by country boundaries for socioeconomic data, or via grids or clustering for data without known clear grouping) and estimating a covariance matrix for each group using the so-called *ssMRCD estimator* implemented in the R package ssMRCD (Puchhammer and Filzmoser, 2023b). The concept of next-distances from REG is also applied here to identify outliers. Simulation studies in Puchhammer and Filzmoser (2023a) show that the method ROB tends to have an increased false negative rate since the global covariance matrix seems to not being strict enough in its estimation of the local covariance. The method REG leads to an increased false positive rate, because using only the k-nearest neighbours for the covariance estimation seems to be too strict by not putting the local estimation into the global perspective. Outlier detection based on ssMRCD includes some spatial smoothing among spatially close groups, and thus the broader structure is also taken into account which balances the false positive and false negative rate.

The last considered method for local outlier detection is the *local outlier factor* (LOF) introduced by Breunig et al. (2000) and adapted to the spatial setting according to Schubert et al. (2012). Since the LOF is purely (Euclidean) distance-based and does not use the pairwise Mahalanobis distance, there is no need to estimate a covariance matrix. Instead, a local density based on the Euclidean distance in the feature space is calculated for each observation and compared with the density of its k-nearest spatial neighbours. Formally, the base of the LOF is the so-called reachability distance g_k between two objects x and y which is defined by

$$g_k(x,y) = max\{d_k(x)\,, d(x,y)\,\}$$

where d is the Euclidean distance and $d_k(x)$ the (Euclidean) distance of x to its k-nearest neighbour. The density used, also called the local reachability density, is defined by

$$lrd_k(x) = \left(\frac{\sum_{y\in A_k(x)} g_k(x,y)}{|A_k(x)|}\right)^{-1}$$

with $A_k(x)$ being the spatial k-nearest neighbours. If the density of an observation is considerably lower than the density of its neighbours, measured by a local outlier factor

$$LOF_k(x) = \frac{\sum_{y\in A_k(x)} \frac{lrd_k(y)}{lrd_k(x)}}{|A_k(x)|}$$

bigger than 1, the observation is considered a local outlier. The original LOF method of Breunig et al. (2000) is implemented in the R package DescTools (see Signorell, 2017).

Finding these local outliers is quite important for mineral exploration especially in the context of geochemical data. Though there are a number of methods such as geological mapping, geochemistry, geophysical surveys and remote sensed imagery that are used in mineral exploration to find potential areas for mineral deposits (Marjoribanks, 2010), in this paper, we are focusing on a geochemical approach in connection with local outlier detection. In the areas having transported cover, such as glaciated terrains, mineral deposits are typically found as sub-outcropping under till-cover. In addition, many ore deposits locate buried under the bedrock surface or even hundreds of meters depth in the bedrock without outcrop on the surface. That type of buried deposits are challenging for the mineral exploration due to poor recognition with surface techniques. However, geochemical data of till and bedrock may provide good targeting criteria for identifying both sub-outcropping and buried mineral deposits. Local outliers reveal anomalous data points which highly deviate from the surrounding data variability in geochemical data sets and may be indicators for mineral deposits in geochemical explorations (Filzmoser, 2004). Thus, geochemical anomaly detection in general is crucial for exploring unknown mineral deposits, and applying local outlier detection techniques in particular can be beneficial in achieving this goal. The type of geochemical data (i.e. elements) that should be used to identify outliers and then predict possible deposits may depend on the type of targeted mineral deposit. When detecting Ni - Cu deposits, as an example, outliers can be associate with high Ni, Cu, PGE, Ti, V, S, Cr and Co (Maier, 2015).

In this context, also certain relations of element concentrations are often very insightful. This is connected to the compositional nature of element concentrations which is an essential aspect and needs to be addressed by any method when applied to geochemical data. While the assumption of a normal distribution seems valid for many measurements, the underlying distribution of geochemical data has an inert structure that must not be ignored. Since geochemical measurements (also called analytical results) constitute a composition of elements, the sum of the concentrations or *parts* of each sample is fixed to the same number. Thus, the underlying geometry of the data is not the Euclidean but the Aitchison geometry (Pawlowsky-Glahn et al., 2015) and the relevant information is not in the absolute values but in the pairwise (logarithmic) ratios of the parts. Although this geometry seems complicated, many methods can be applied after appropriately transforming the compositional data to the Euclidean geometry while additionally taking the original structure (i.e. the simplex) or the pairwise (logarithmic) ratios of the parts into account for interpretation.

There are various transformations suited for this task (see, e.g., Filzmoser et al., 2018). We will focus on two of them that are easy to apply and have good theoretical properties. The first transformation leads to the so-called *centered-logratio (clr)* coefficients. For a composition $\boldsymbol{x} = (x_1, \ldots, x_D)$, the clr transformation is defined as

$$clr(\boldsymbol{x}) = \left(ln\frac{x_1}{\sqrt[D]{\prod_{k=1}^{D} x_k}}, \ldots, ln\frac{x_D}{\sqrt[D]{\prod_{k=1}^{D} x_k}}\right), \tag{1}$$

which is essentially the logarithm (per variable) of the observed composition standardised by its geometric mean. The clr-transformation is isometric, meaning that it preserves the distance of the Aitchison geometry when using the Euclidean distance in the transformed space. Also, the interpretation is desirably straightforward and based on relative information with respect to the (geometric) mean. However, a drawback present in many applications is that the transformed data matrix does not have full rank since clr coefficients are based on a generating system and not on a basis of the Aitchison geometry. This can be overcome by using one of infinitely many orthonormal coordinates. The transformation of choice in this paper is based on *isometric logratio* (ilr) coordinates and known under the name *pivot coordinates* (e.g. Filzmoser et al., 2018). The j-th entry of the pivot coordinates of $\boldsymbol{x}$ is defined as

$$ilr(\boldsymbol{x})_j = \sqrt{\frac{D-j}{D-j+1}} ln\left(\frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}\right) \quad (2)$$

for $j = 1, \ldots, D-1$. Since an orthonormal basis is used, we reduce the dimension of the transformed composition by one and resolve the problem of singularity in general present for the clr-transformation. Up to a constant we have equality in the first entry of the ilr and clr transformation, $ilr(\boldsymbol{x})_1 \propto clr(\boldsymbol{x})_1$, and thus, the first entry of ilr coordinates can be interpreted just as easily as the clr-transformation. Note that although there is a close connection between the first coordinates, it should be kept in mind that ilr coordinates represent dominance while clr indicates the average of a composition. However, this close relationship does not apply to the other coordinates of the ilr-transformed composition which is essentially the one major disadvantage of the orthonormal basis. We will use both transformations according to their properties, and choose the transformation based on the questions and requirements arising in our data analysis.

In this paper we analyse geochemical data by applying local outlier detection techniques to three data sets differing in scale and data quality. We show the importance of data preprocessing steps and the usage of compositional data analysis methods, describe the problems encountered with data having insufficient quality and debate possible solutions that adequately account for the compositional nature. Moreover, we show how different local outlier detection techniques perform on different scales and analyse in which cases some methods might be less appropriate to find mineral deposits. Some ideas on outlier diagnostics, method evaluation, and filtering of outliers based on common compositional data transformations are also discussed to complete a thorough local outlier analysis in the compositional data setting.

The paper is organised as follows. In Section 2 we describe the three data sets and corresponding preprocessing steps before applying the four local outlier detection methods in Section 3. The final two sections summarise and discuss the findings and provide overall conclusions.

2. Data description and preprocessing

For illustration purposes we choose three data sets differing in spatial scale, sampling scale and data quality to showcase the differences and specifics of the four selected outlier detection methods. The locations of the samples of the different data sets are depicted in Fig. 1.

The first data set is the so-called *GEMAS* data set described in Reimann et al. (2014a, 2014b). The data consists of agricultural soil samples that cover most of Europe in a density of 1 sample per 2500 km^2, see Fig. 1 left. The 2108 samples were analysed by X-ray fluorescence, following tight quality control procedures, resulting in concentration values for 41 chemical elements. Here we use the data set published in the R-package robCompositions (see Templ et al., 2011), named as data set gemas. It contains only elements with less than 3 % of the analytical results below the detection limit, resulting in 18 main elements with good data quality.

The other two data sets are used for till geochemical analysis (regional till geochemistry, targeting till geochemistry and mineral deposits) in Finland. They are provided by the Geological Survey of Finland (1995, 2013, 2016) (GTK) and modified as described below. The regional till data set covers whole Finland and it has been collected during the period of 1983 to 1991. This data set contains the concentrations of 22–26 elements (depending on the map sheet – in the selected area we have 22 elements available), see Table 1. The samples have been collected from the C horizon, which contains unaltered till. The sampling depth is approximately 1.5–2 m. The sampling density is 1 sample per 4 km^2 and the full data set comprises of 82,062 samples. Furthermore, concentrations of 22–26 elements that can be extracted by aqua regia have been analysed for fine fraction of the till material less than 0.06 mm and the data has been published by 1:400,000 map sheets (Salminen and Tarvainen, 1995).

The final data set, the targeting till geochemical data set, contains around 385,000 soil samples collected by GTK along sample lines in certain areas between the years from 1971 to 1983. Most of them are till samples, however samples from sorted mineral soils, weathered bedrock and mixed intermediate forms also exist in the data set. In this paper, only till samples collected using percussion drilling and test pitting methods from the C horizon which contain fine (less than 0.06 mm) fractions are considered. The samples have been collected by 1:100,000 map sheets. The point density of the samples lies between 6 and 12 samples per 1 km^2 where the line interval is 500–2000 m and the distance between two points is 100–400 m. The average depth of the samples is 2 m, and an emission quantometer method has been used to measure the concentration of 17 elements listed in Table 1 (Gustavsson et al., 1979).

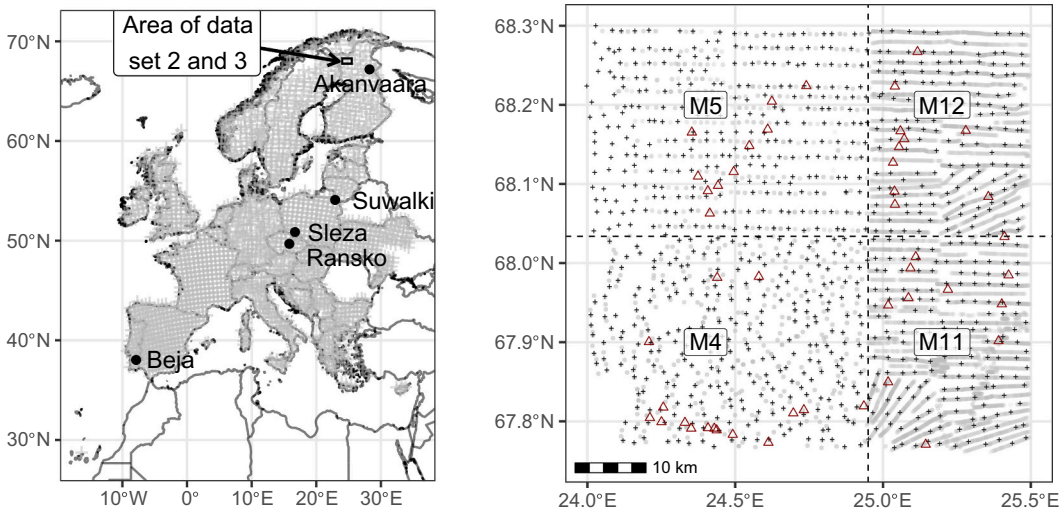

Fig. 1. Map of research areas. *Left:* Grey crosses indicate sample locations of the GEMAS project while the black dots represent the reference sites of the SEMACRET Project (2023). The rectangle in the Northern part of Finland represents the four selected map sheets of the regional and targeting till data set shown on the right. *Right:* Sample locations of regional till (black crosses) and targeting till (grey dots) data, partitioned into four map sheets. Each triangle indicates a known mineral deposits.

Table 1
Percentages of problematic data quality of the targeting till and regional till data set for different elements with respect to corresponding map sheets. The values of the elements per map sheet used after the final data cleaning are underlined.

Element	Targeting till (%)				Regional till (%)			
	M4	M5	M11	M12	M4	M5	M11	M12
Ag	100	100	100	100	–	–	–	–
Al	91.17	93.84	4.64	16.63	0	0	0	0
Ba	–	–	–	–	0	0	0	0
Ca	1.13	2.02	27.09	53.87	0	0	0	0
Co	16.27	18.52	3.04	8.24	0	0	0	0
Cr	86.65	6.82	6.33	4.16	0	0	0	0
Cu	1.27	4.39	0.96	1.93	0	0.36	0	0
Fe	0.03	0.87	0.73	1.80	0	0	0	0
K	1.96	3.44	35.66	9.60	0	8.02	0	0
La	–	–	–	–	0	0	0	0
Li	–	–	–	–	0	0	0	0
Mg	0.03	0.06	0.01	0.03	0	0	0	0
Mn	2.17	1.14	2.71	2.74	0	0	0	0
Na	0.37	0.40	15.90	4.38	–	–	–	–
Ni	0.24	0.33	0.18	0.46	0	1.45	0	0
P	–	–	–	–	0	0	0	0
Pb	99.58	91.95	97.35	97.71	–	–	–	–
Sc	–	–	–	–	0	0	0	0
Si	0	0.47	0.01	0.07	–	–	–	–
Sr	–	–	–	–	0	0	0	0
Th	–	–	–	–	5.88	15.32	9.52	1.92
Ti	0	0.27	0.01	0.01	0	0	0	0
V	0.27	0.94	1.03	3.72	0	0	0	0
Y	–	–	–	–	0	0	0	0
Zn	98	22.04	91.48	60.58	0	0	0	0
Zr	–	–	–	–	11.02	7.66	36.90	44.87

For the data analysis in Section 3 we do not use the complete regional and targeting till data sets, but choose data subsets covering only the area from Central Lapland depicted in Fig. 1 (right), which is partitioned into four smaller areas or *map sheets* by GTK. This area contains many known mineral deposits and provides sufficient data quality in terms of enough reliable measurements, which is not provided in all areas for the targeting till data set. By taking the same sampling area for these two data sets, we are also able to compare their usefulness for mineral exploration with local outlier detection methods.

2.1. Data preprocessing

The element selection based on the detection limit threshold of 3% for the GEMAS data set constitutes a compromise between rejecting too many elements, and keeping too many elements with low data quality. Removing said elements ensures that most of the reliable information of this data set is extracted. Due to the high data quality in general, no further preprocessing is necessary.

The selected subset of the regional till data set generally has good data quality. However, for some elements it contains values below the lower detection limit, and other data quality issues. Therefore, additional data cleaning is required. The right part of Table 1 shows the percentages of values with the data problems mentioned before for the selected 4 map sheets. We decided to exclude Zr and Th for all further analyses. The remaining data quality issues are not connected to detection limit problems, since only Zr and Th where erroneous in this way. A small amount of analytical results have additional markers in the data with unclear encoding. The benefit of the additional information saved trough this procedure outweigh possible negative effects on data analysis and the differently marked analytical results are kept in the data. We refer to Mert et al. (2016) for an analysis of contamination on compositional data transformations. The resulting regional till data set has thus 870 samples and 20 variables.

Compared to the previous data sets, the targeting till geochemical data set has serious data quality issues, typically connected to values related to detection limits, zero and even negative analytical results, and values marked with special symbols. Therefore, extensive data cleaning is required in order to perform further statistical analyses and modelling procedures. The percentages of insufficient quality of samples per element and map sheet areas are calculated and shown in Table 1 left. Eventually, elements which contain more than 30 % of problematic samples over all map sheets (e.g., Ag, Pb and Zn) are removed from all further analyses.

Furthermore, the geochemical analysis of the targeting till data set has been carried out at different times and map sheets. Therefore, it is necessary to analyse a possible mismatch and if the measurements are comparable. Fig. 2 illustrates the spatial concentration of Fe in both till data sets separately. It is evident that there are discontinuities at the map sheet boundaries in the targeting till data set due to inconsistencies during the geochemical analysis done by quantometer method. These discontinuities are not present in the regional till data, where there is a change of geological units from Archean and Proterozoic to only Proterozoic origin visible. Thus, after displaying clearly visible map sheet boundaries and discrepancies between map sheets at least for Fe that are not due to the underlying geology it was decided to analyse the map sheets (1:100,000 scale) separately for the targeting till data set, as the smaller areas also contain enough sample points to carry out the analysis.

To improve data cleaning further, also Q-Q plots are used to examine the distribution of concentrations between different map sheets in the regional as well as in the targeting till data set where only elements with less than 30% of quality issues are included. For the Q-Q plots, we focus on the elements Co, Cr, Cu, Fe, Ni, V, and Ti, which are important ore metals in ultramafic rocks, and thus of special interest for mineral exploration. As example, the concentration values of Fe for all four map sheets separately are shown by Q-Q plots for the targeting till data set in Fig. 3(a) and for the regional till data set in Fig. 3(c) as well as the corresponding clr transformed values in Fig. 3(b) and in Fig. 3(d), respectively. The Q-Q plots for Fe vary between map sheets (M4, M5, M11, M12) but especially between the two data sets. With respect to the average concentration level per map sheet we even see adverse ordering in original as well as clr transformed values for regional and targeting till, which is congruent with Fig. 2. Note that Q-Q plots alone are not sufficient to diagnose map sheet levelling problems but the adverse

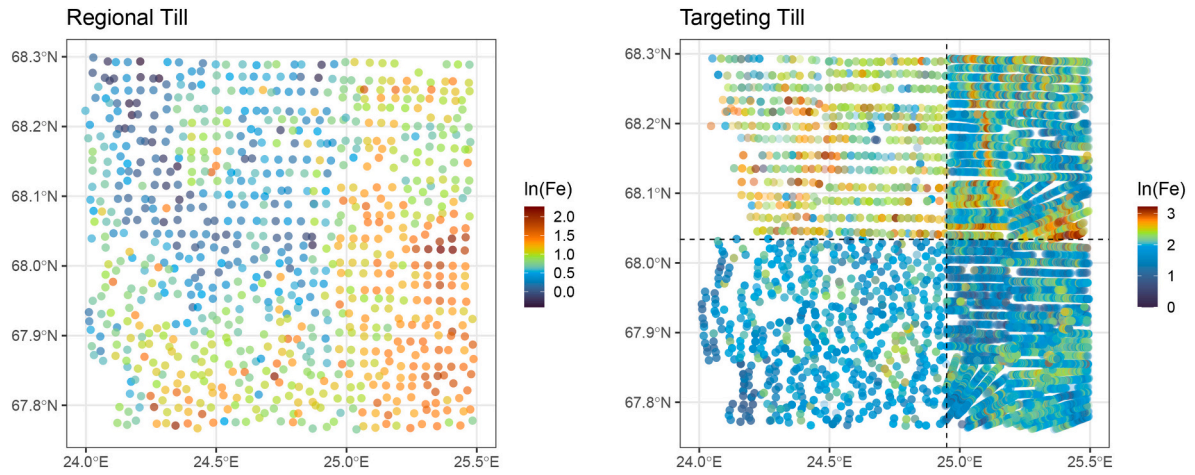

Fig. 2. Illustration of discontinuities of Fe (%) between map sheets in the targeting till (right) compared to regional till data set (left). Clear boundaries are visible between the map sheets in the targeting till data set.

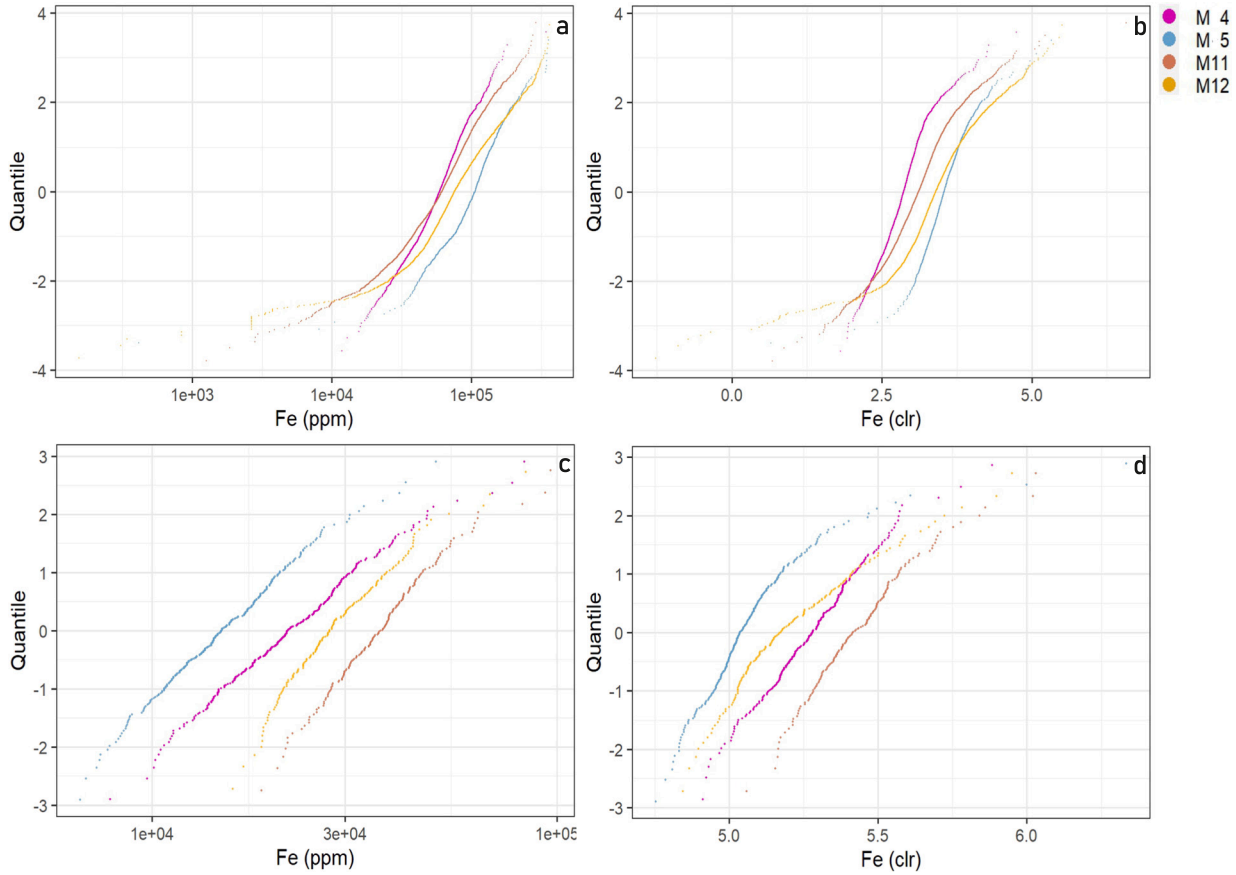

Fig. 3. Q-Q plots of Fe: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till, n (targeting till) = 16,460, n (regional till) = 870.

ordering could still be a strong indicator of them. However, other quantitative differences between the two data sets might be mainly due to different analytical techniques. Interestingly, the clr transformation based in the regional till data set reorders the average relative level of Fe between map sheet M4 and M12 indicating that using the appropriate compositional data structure adds important information which would be ignored otherwise. Regarding differences between the map sheets per data set for other elements (Co, Cr, Cu, Ni, V, and Ti) shown in Appendix A it is less clear whether they originate from map sheet problems or from spatially changing lithology. Finally, the distributions of elements for all map sheets, elements and data sets seem to be plausible. Apart from some lower detection limit problems in the targeting till data set, which will be taken care of in the next step, we do not need to account for any extensive rounding, grouping or other distributional issues that might occur.

After the extensive map sheet analysis of the targeting till, the final data cleaning is necessarily done per map sheet. In order to use as many elements as possible, we start by removing samples that have at least one

zero value of element concentration. Also observations with more than 30% of problematic values over all elements provide a restricted amount of reliable information and are removed. Due to the high sample density, we still keep enough observations to make sensible analysis when applying the rather strict row cleaning (M4: 2417 samples, M5: 1399 samples, M11: 5821 samples, M12: 4557 samples). Note, that for data sets of lower sample density, the decision between having less samples or less elements available after data cleaning is less clear than in this case. Finally, only the elements that have less than 5% problematic values per map sheet are used, which are underlined in Table 1. This is again rather strict, but we hope to reduce the number of local outliers connected to poor data quality or detection limit problems. Imposing the even stricter limit of 3% similar to the GEMAS data set is not applicable for this data set, since many more elements would be lost for the analysis. However, note that the effects of some data quality issues on compositional transformations will be present but limited (Mert et al., 2016).

After the final data cleaning of both data sets, targeting and regional till, the last preprocessing step is to address the compositional nature of the geochemical data (for targeting till again per map sheet). Although the clr-transformation is isometric regarding the Aitchison geometry and easy to interpret, the linear dependency introduced is problematic. In the case of covariance estimation we would get a singular matrix which is not invertible. However, this is a necessity for the pairwise Mahalanobis distance and the three methods based on it. Thus, it is sensible to choose ilr coordinates for the regularised spatial outlier detection technique, the robust local outlier detection method and the ssMRCD-based outlier detection technique to avoid this problem. Since LOF does not need a covariance estimation and is strictly (Euclidean) distance-based, any transformation for compositional data which is isometric can be applied. Thus, both ilr and clr can be used, and due to isometry both lead to the same local outlier factor and thus, to the same local outliers.

3. Data analysis

After finishing the preprocessing and data cleaning steps and the compositional data transformations, the four outlier detection methods can be applied to the transformed data. Regarding the parameters, we generally adhere to default settings wherever sensible. For all four methods we compare each observation with the same amount of k nearest neighbours. Setting k influences the locality of the local outlier detection in all methods since we are looking for an anomaly compared to samples from a larger area. For a more detailed analysis of the effects of different k values we refer to Braus (2023). Since the considered data sets differ in sample size and density, k is adjusted to the data sets. The parameters for the ssMRCD-based method are a smoothing parameter $\lambda = 0.5$, representing a compromise between local and global covariance estimation. Neighbourhoods are defined data specific and the weighting matrix for smoothing between neighbourhoods is defined as the pairwise inverse distance of the neighbourhood spatial centres, which is the most natural choice for data without inherent spatial structural breaks. For LOF a value above 1.5 is flagged as outlying, and for the regularised spatial detection technique we want to include all observations ($\beta_{REG} = 1$) independent of local heterogeneity. As regularised covariance estimator, the Minimum Regularised Covariance Determinant estimator (Boudt et al., 2020) with a trimming percentage $\alpha = 75\%$ is chosen, meaning that 75% of the k-nearest neighbours are used for the local covariance estimation. Regarding ROB, all other parameters including the amount of neighbours allowed to be similar as well as the cutoff value are adjusted to the spatial scale of the data.

3.1. GEMAS data

For the GEMAS data set we use the following parameter setting for the different local outlier detection methods: We choose $k = 10$ for all methods, a standard settings, and $N = 50$ neighbourhoods for the ssMRCD estimation which reflects an appropriate level of locality given the sample density and ensures sufficient observations per neighbourhood. The neighbourhoods are selected by k-means clustering based only on the spatial attributes of the data and checked for reasonably sized spatial clusters. For the robust local outlier detection method (ROB) we allow 10% of the neighbours to be similar ($\beta_{ROB} = 0.1$) due to the large area covered and the sparse sampling of the GEMAS data set, and choose an isolation degree bigger than 0.2 as cutoff value. These parameter choices are also supported by the work of Braus (2023). As some indicator for performance we also include the reference sites of the SEMACRET Project (2023) in Fig. 1 (left). Finding these reference sites can be interpreted as analysis goals. However, on the one hand this approach is unbalanced since a high number of found outliers already leads to an improved performance, and on the other hand unknown mineral deposits are not taken into account. Nevertheless, we get more insight into possible drawbacks of different methods.

The flagged outliers per method are shown in Fig. 4. Starting with the most global method, the robust local outlier detection method (ROB), problems connected to the global covariance estimation are evident. For the robust covariance estimation, the MCD estimator (Rousseeuw, 1985) is used which selects a subsample of the data with the lowest determinant of the sample covariance based on this subsample. On the GEMAS data set, the subsample contains mainly observations from Middle to Northern European countries, thus leading to a covariance not representing Southern Europe and to an unbalanced and somewhat biased spatial distribution of the outliers. For the regularised local outlier detection technique (REG), the outliers are more or less evenly distributed. However, the problem of a high number (almost 15% of the observations) of outliers arises which is likely connected to an increased false positive rate. The high amount of outliers makes it difficult to get more valuable insight into the data rendering the method essentially useless without further processing. Both, the ssMRCD-based and the LOF-based outlier detection method seem promising, however it seems that the ssMRCD finds most reference sites, including a strong signal very close to the ultra-mafic intrusion body in Beja (see Fig. 5b). Note that the mineral deposit in Suwalki in North Poland is assumed to be multiple hundreds of metres deep under the surface, so it is unlikely that it affects the soil sufficiently. Moreover, LOF does not flag the soil sample from the Canary Islands as outlying which would be sensible given that the ten nearest neighbours are located far away somewhere in South Spain.

Although the element selection in the GEMAS data set might not be oriented towards mineral exploration, the data quality is very good and it is well suited to discuss further processing steps. After applying the four methods we end up with many potential locations for mineralisation or other anomalous observations. Thus, a closer look at the identified outliers is quite important since finding mineralised areas can be very expensive, and additional analysis can improve the identification of important locations. We are interested in utilising potential mineralisations by mining, hence high values of elements in clr coordinates (meaning high concentrations relative to other elements) but also in total concentrations are desirable. Thus, we employ a filtering procedure on all flagged outliers keeping only those observations which have clr and measured concentration levels simultaneously above the global 95% quantile for at least one element. In Table 2 the total numbers of outliers, filtered and unfiltered, are shown for all three data sets. Depending on geological knowledge and the type of mineral deposits that should be found, the selection of elements for the filtering procedure can be further specified in concrete applications. For finding potential Ni-Cu mineralisation, the elements in the filtering procedure can be tailored specifically to high Ni and Cu and other connected elements (see also Section 3.3).

The number of outliers is reduced by the filtering procedure, but for single observations we can still improve on the analysis to increase the chance of finding valuable mineral deposits. A possible diagnostic tool is

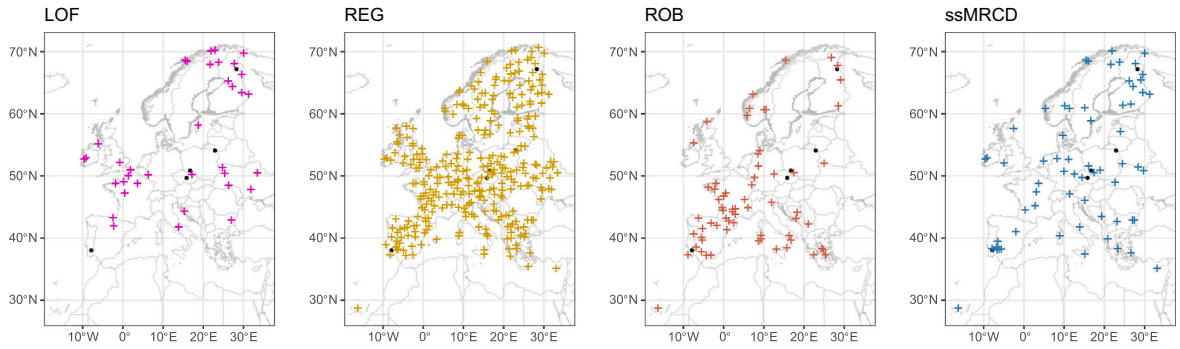

Fig. 4. Spatial locations of flagged outliers (marked as cross) by each method separately on the GEMAS data set. The black dots represent areas of interest in the SEMACRET Project (2023) where mineral deposits are anticipated.

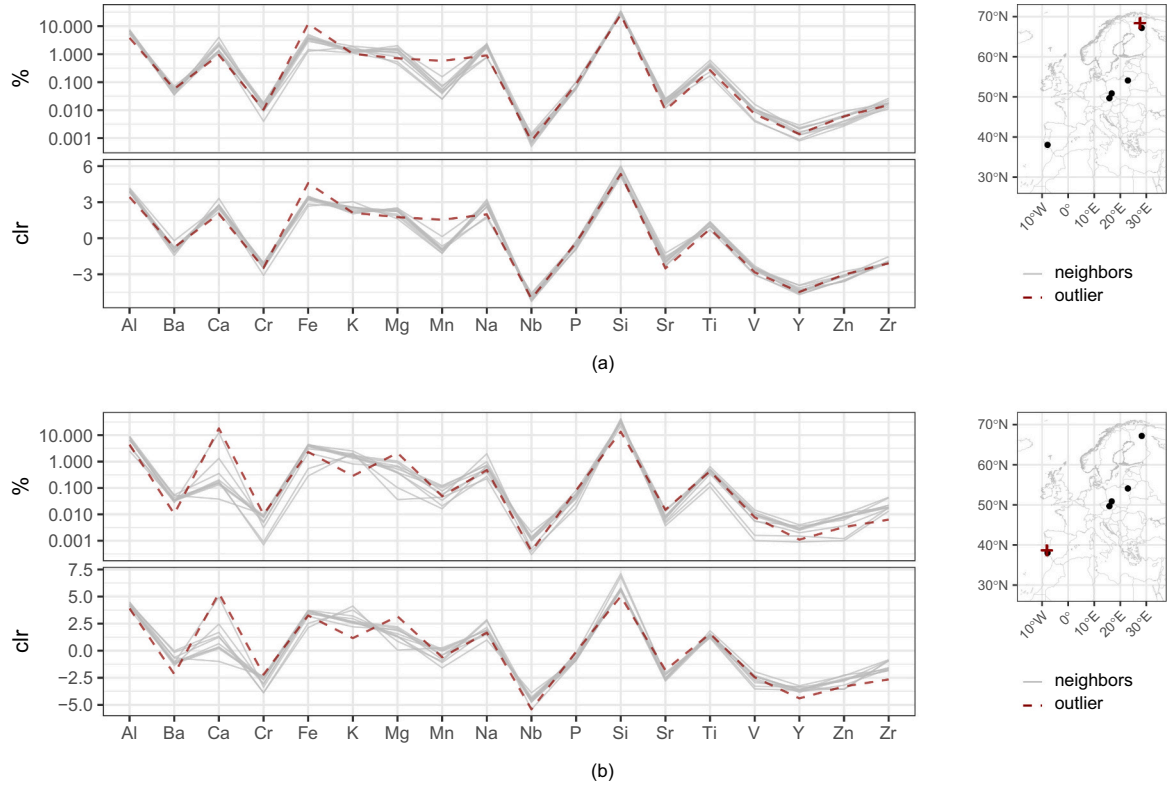

Fig. 5. Outlier diagnostics for (a) observation 530 which is closest to the Akanvaara area, and for (b) observation 189, closest to the ultra-mafic intrusion body in Beja, each colourised in red. The two parallel coordinate plots show the multivariate structure of the observations and corresponding 10 nearest neighbours in grey, once in percent (upper part) and once in clr-transformed values (lower part). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Number of flagged outliers for each method and data set, unfiltered and filtered by high values in clr values and non-transformed measurements in at least one element.

Method	Unfiltered			Filtered		
	GEMAS	Regional	Targeting	GEMAS	Regional	Targeting
LOF	36	17	420	24	15	359
REG	311	115	943	182	59	640
ROB	66	13	595	28	5	167
ssMRCD	64	26	431	48	21	379
# samples	2108	870	14,194	–	–	–

based on parallel coordinate plots which can give insight into the multivariate structure. Each observation is represented by a line, and the values of each variable on the horizontal axis are connected. We focus on the comparison with the k-nearest neighbours. Together with insight into the underlying bedrock, the corresponding observation can be interpreted as interesting new target for further exploration or discarded as uninteresting anomaly. In Fig. 5a and b, two of the flagged outliers which are closest to the Akanvaara area and Beja are analysed in

comparison to their 10 nearest neighbours (coloured in grey) using the parallel coordinate plot.

Regarding the outlier next to Beja in Portugal, which was flagged only by the ssMRCD-based method, we see high values in Ca and Mg and particularly low values in K. This fits well to the known geology in this region. While the neighbours are mostly located on sand (3 samples) and on the South-Portuguese Flysch zones (4–5 samples) which are composed of higher Al, Si, Fe, K as well as hardly any Ca and Mg (Jorge et al., 2013), respectively, the flagged outlier lies on the layered Gabbroic Sequence at Beja which is consistent with the elemental composition of the outlier as it contains olivine bearing gabbroic rocks which are bordered by heterogeneous diorites (Jesus et al., 2014). Gabbro usually contains minerals which associate with Ca and Mg such as pyroxene, plagioclase, and olivine of which weathering release Ca and Mg. The depicted high values in Ca and Mg are thus indicators for the Caliche type of weathering, which is typical in that type of climate for (ultra-)mafic lithologies. Also, low Si and slightly higher Cr with respect to neighbours indicate weathering of gabbroic rocks.

For the outlier indicated by the methods LOF, REG and ssMRCD near the Akanvaara deposit and the so-called Koitelainen deposit north-western of Akanvaara, higher values can be observed for Fe and Mn with respect to the nearest neighbours (Fig. 5a). The Akanvaara deposit is located in Northern Finland (eastern part of the Central Lapland greenstone belt) and it is considered as a layered mafic intrusion which hosts vanadium mineralisation in layers of magnetite gabbro and also in chromitite layers within gabbro. These two layers have been mineralised by massive, semi-massive and disseminated magnetite, pyrite, chalcopyrite and chromite (Lutynski, 2019). Koitelainen also an ultramafic deposit which is enriched by commodities such as Cr_2O_3, V, Fe and PGE. The flagged outlier is closer to the Koitelainen deposit than the Akanvaara deposit where the distances from the outlier to the deposits are approximately 17 km and 72 km respectively. Thus, when considering the flagged outlier for these deposits, elevated amounts of elements such as Cr, V, Cu are also expected other than Fe in order to identify it as an indicator for Akanvaara and Koitelainen. However, the GEMAS data are for grassland areas and Akanvaara locates inside largely forested area without close vicinity to grasslands. Furthemore, since this flagged outlier associates with only high Fe and Mn, it cannot be 100 % certain that it indicates the Akanvaara or Koitelainen deposits, but it is certain that it indicates a mafic environment where there is a possibility for a mineral deposit.

3.2. Regional till data

For the regional till data set some parameter settings are adjusted. We again compare single observations with their $k = 10$ nearest neighbours. For the ssMRCD-based method, each of the 4 map sheets is chosen as an own neighbourhood. This choice is due to the very dense sampling grid, but simulations in Puchhammer and Filzmoser (2023a) also suggest that the method is rather insensitive to the number of neighbours, as long as some smoothing by the parameter λ is performed. For the robust local outlier detection method (ROB), we increase the percentage of neighbours allowed to be similar to 30% ($\beta_{ROB} = 0.3$) due to the smaller scale of the sampling area and choose an appropriate cutoff value for the isolation degree of 0.4. We refer to Braus (2023) for sensitivity analyses with respect to the choice of these parameters.

Interestingly, due to the smaller scale of the data we have the advantage of known mineral deposits (Geological Survey of Finland, 2016). There are 48 known mineral deposits of various types in the research area depicted as red rectangles in the right part of Fig. 1. Ideally, our methods find these locations. However, since generally there are no samples directly on the deposits, we define a deposit to be found if an outlier is located 4 km or closer to the deposit. This might seem like quite far, but for an average density of one sample per 4 km^2 and historical glacial movement this distance is quite reasonable. Note, that this is not a guarantee that the outlying sample detecting the deposit has a typical element composition connected to the specific deposit type. Hence, it might be possible, that the sample is outlying due to other processes. Moreover, it would be preferable if the methods find the deposits as the most extreme outliers. Thus, we rank the outliers according to their outlyingness value, and analyse how many deposits are found until which outlier rank.

The left part of Fig. 6 shows how many deposits are found by outliers up to the rank depicted on the horizontal axis for the regional till data set, with and without the filtering procedure described in the prior subsection. We see that filtering outliers reduces the number of outliers overall. However, for the regularised spatial outlier detection technique, the ssMRCD-based method and also the LOF there is an improvement in accuracy, meaning more deposits are found earlier. The degree of the improvement differs among methods, from strong for REG to negligible for LOF. Nevertheless, the filtering tool proves to be valuable if a subselection of outliers is necessary.

3.3. Targeting till data

Finally, the targeting till data set is used for the analysis. As discussed in Section 2 it is most sensible to analyse the four map sheets separately. The data also provides a structure of smaller sub-mapsheets, 12 for M4 and M5 and 6 for M11 and M12, respectively, that are used as neighbourhood structure for the ssMRCD-based method. The only other parameter setting that is changed compared to the regional till data analysis is k, the number of neighbours to be compared with each observation. Due to the high sampling density, we increase k to 30 to find appropriate local outliers. Since we have 16 times more observations than in the regional till data set for the same area, the necessary distance to a known mineral deposit for it to be defined as found is reduced to 1 km in order to compare the performance of the data sets fairly.

Due to the separation of the map sheets in the analysis and the fact that we have a different set of elements per map sheet, we cannot compare the degree of outlyingness without adjustments. Thus, for each map sheet the outlyingness is standardised with its cutoff value to reduce the effects of separate analysis, and then the observations is ranked jointly by the standardised outlyingness.

The results can be seen in the right panel of Fig. 6, again with unfiltered and filtered outliers. The methods flag many samples as outlying and for ROB and REG filtering significantly reduces the number of outliers while increasing accuracy. Ideally, the curves would jump at the very beginning up towards the number of known deposits. We can see that the ssMRCD-based method is closest to the ideal, both with and without filtering of outliers.

As mentioned before it is also possible to use a specific set of elements for filtering that match a deposit type of interest. As illustration we now try to find a (known) Ni-Cu deposit by filtering according to Ni, Cu, Ti, V, Co and Cr (see Section 1). Three of the four methods (LOF, REG, ssMRCD) flag the sample analysed in Fig. 7 as outlier, which is less than 1 km away from the Saattopora-Cu deposit hosting Cu together with Au, Ni, Co and Ag. High values in Ni, Ti and Co of the flagged sample imply that the Ni-Cu deposit is connected to its outlyingness.

Comparing the results of the two data sets shown in Fig. 6, we can clearly see that significantly more mineral deposits are potentially found by less flagged outliers using the regional till data set. In the case of mineral exploration this is definitely desirable since each outlier would need to be analysed more closely. By providing that valuable outliers have high ranks in outlyingness, the effort and time spent on additional analysis is reduced. Note that outlier detection with the regional till data set might be more accurate than with the targeting till data set just because of the availability of more elements. This seems to be an important factor in finding certain types of ore deposits compared to a higher sampling density.

Another interesting approach is to analyse if the (potentially) found mineral deposits are the same or if the data sets lead to different results.

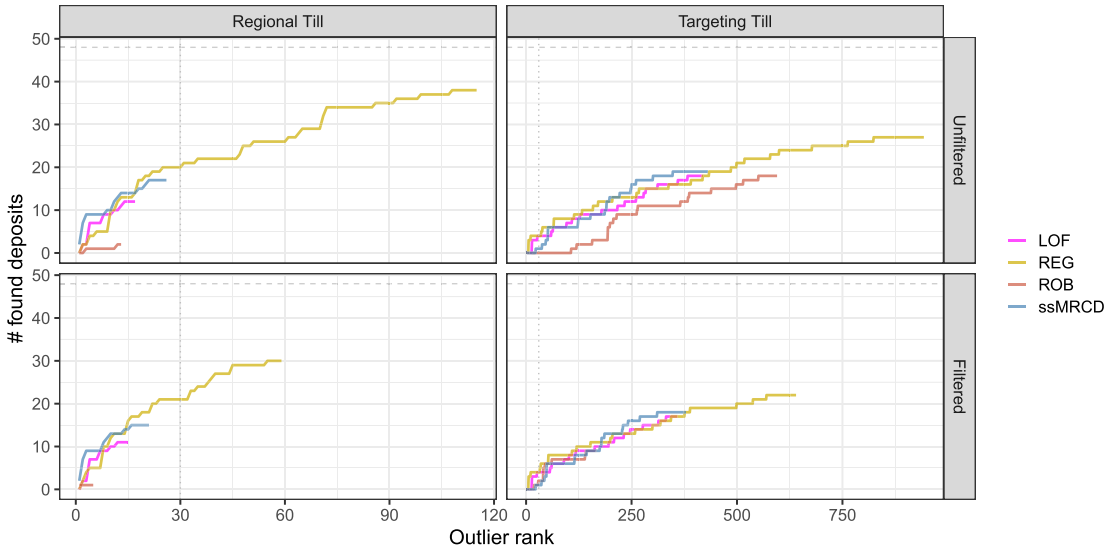

Fig. 6. Performance of local outlier detection methods on regional (left) and targeting (right) till data for filtered and unfiltered flagged outliers. The dashed line represents the number of known mineral deposits. The methods applied to the regional till data set have better performance than for targeting till, as can be seen for the first 30 outliers (dotted line).

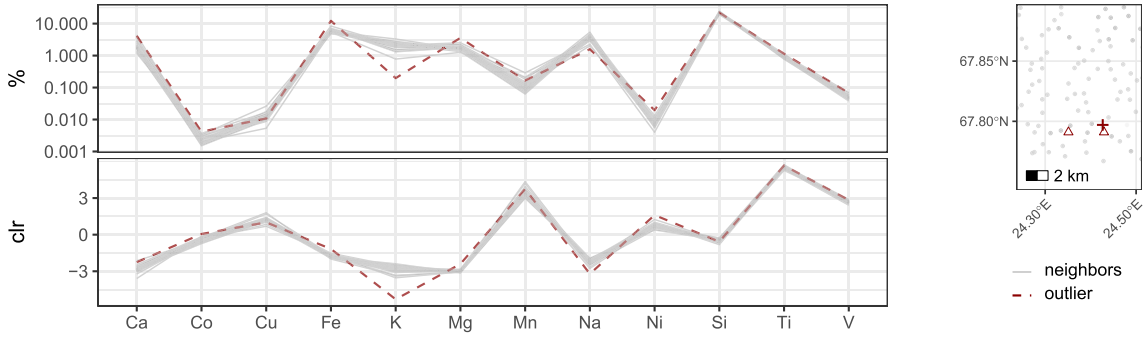

Fig. 7. Outlier diagnostics for an outlier (red cross) close to the Saattopora-Cu deposit (right triangle). The two parallel coordinate plots show the multivariate structure of the observations and its corresponding 30 nearest neighbours in grey, once in percent (upper part) and once in clr-transformed values (lower part). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

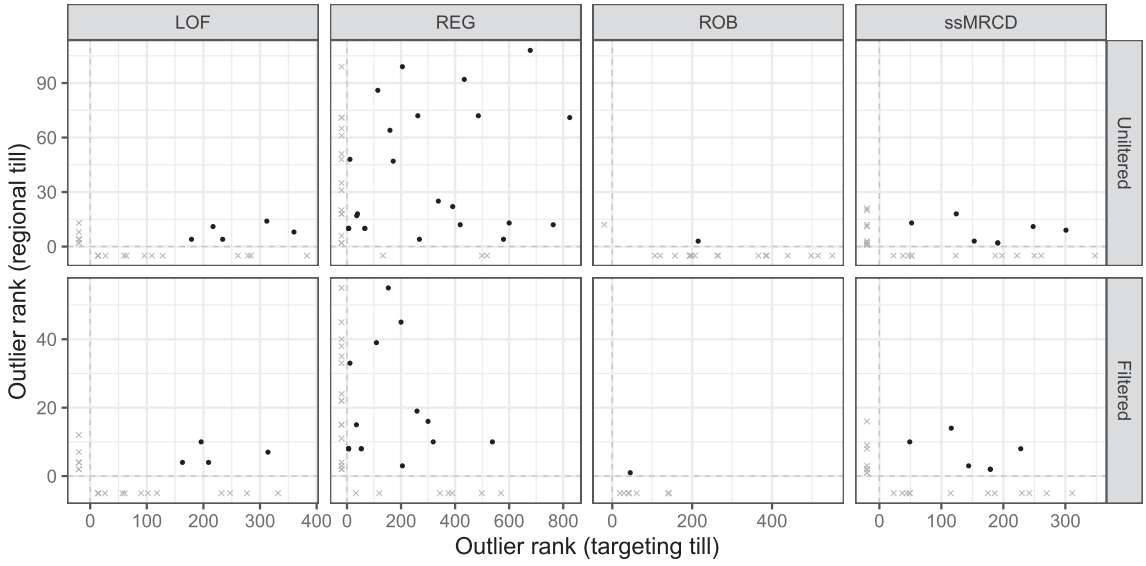

Fig. 8. Found mineral deposits per outlier rank for unfiltered outliers (a) and filtered outliers (b) in either the regional and/or the targeting till data set. Jointly found deposits are marked by dots, deposits found only by one method are marked as grey crosses.

In Fig. 8 the outlier rank of the found deposits for both data sets are shown, for filtered and unfiltered outliers, and summarised in Table 3. In most cases, the number of found deposits is hardly affected by the filtering procedure. This indicates that the filtering process designed for subselecting outliers really leads to more accuracy in finding mineral deposits. Again, we can see that analysing outliers from the regional till data set is effectively detecting ore deposits since many of them are found with a much lower outlier rank. Interestingly, the ore deposits found differ between the data sets used. This reflects also the size and type of ore deposits which would mean that with sparse sampling grids bigger outcropping or sub-outcropping mineralisations are possibly found but with increased sampling density the detection of smaller sub-outcropping and buried deposits is improved.

4. Summary and discussion

In this paper we demonstrated the suitability of local outlier detection methods for the purpose of mineral exploration in geochemistry. Generally, local outlier detection incorporates the spatial neighbourhood of the samples in order to identify local anomalies in the multivariate element composition. The analysed data sets are of different scale, sample density and data quality, and they also vary in the number of available element concentrations. However, the geochemical data sets have in common that they are of compositional nature, which made it necessary to process them with tools from compositional data analysis.

The different methods for multivariate local outlier detection mainly vary in the way how they estimate the covariance matrix to compute pairwise Mahalanobis distances. The simplest approach is to use a joint global covariance matrix. The other extreme is to use separate covariance estimates for each local neighbourhood. A third, recently proposed methods tries to find a compromise between those two extremes, with the idea that the robust covariance estimation should change smoothly across the neighbourhoods. These methods are compared to a procedure called LOF (Local Outlier Factor), which incorporates Euclidean distances between the mutlivariate observations, and thus is based on a very different concept.

While all methods find mineralisations, we have shown that they also have their limitations, ranging from biased covariance estimation to an extensive flagging of outliers and not finding reasonable spatial outliers. With known mineral deposits it is possible to evaluate the methodologies on real data and analyse their performance in more detail. However, the considered mineral deposits are of very different type, and one might have to go into much more detail to see if the compositions of the identified outliers really reflect the type of mineralisation, or if the elements used in the analysis are even appropriate for this purpose. Moreover, it can also happen that some of the identified outliers point at new yet unknown mineralisations, which makes the evaluation used in this paper biased.

Thus, next to appropriate outlier detection methods, it is also important to use diagnostic tools to verify if the indicated outliers indeed point at mineralisations. We introduced exploratory procedures that combine relative and absolute information, as outliers are supposed to be atypical in the multivariate compositional data space, but at the same time they are supposed to have high concentration values for particular elements.

Next to a data subset from the GEMAS project we evaluated the procedures for two data sets from the same area in Finland, measured in different years, with a very different sampling density, and yielding different sets of elements with different data quality. The main question was if higher sampling density would also lead to higher accuracy for mineral identification. However, the crucial point for mineral identification seems that not only the commodity elements need to be available, but also complementing elements that allow to understand and characterise the geological situation.

Table 3
Number of found deposits for each method and data set, unfiltered and filtered by high values in clr values and non-transformed measurements in at least one element. Maximum number of deposits possible to find is 48.

Method	Unfiltered			Filtered		
	Regional	Targeting	Both	Regional	Targeting	Both
LOF	12	18	5	11	17	4
REG	39	27	24	31	22	15
ROB	2	18	1	1	9	1
ssMRCD	17	19	7	15	18	6

5. Conclusions

A general but possibly obvious conclusion is that also for local outlier detection, data quality is more important than quantity. However, it is not just quality which matters, it is also the set of elements which needs to be big enough in order to cover the complexity of the geochemistry that experts would expect to find at mineralised zones. Here, rare elements such as gold could be very valuable, provided that they are measured with sufficient quality. Elements measured with low quality, as for example with a high proportion of values below the detection limit, will negatively affect the log-ratio transformations used in compositional data analysis. In more detail, an observation where just one element has a value below the detection limit could end up in a multivariate observation of the compositionally transformed data set with all entries being distorted. This could lead to a very high proportion of outliers, where local outlier detection methods could fail to work correctly.

For the tested local outlier detection methods it is known that some are very sensitive and may lead to a too strict rule for indicating outliers. Also the way how the methods work internally is very different, and therefore these methods are flagging different sets of outliers. From a theoretical point of view, the ssMRCD method will be preferable over the methods REG and ROB in case where the investigated area shows geochemical differences, e.g. as a result of different underlying processes (pollution sources, soil formation, environmental conditions, etc.). The LOF method tends to identify data points that are isolated in the multivariate space. Thus, if the sampling is dense and the observations continuously change towards the mineralisation, this method may fail to see samples on top of mineralised zones as outliers. Nevertheless, a strategy could be to use multiple local outlier detection methods to balance their advantages and limitations.

For sampling strategies it follows that a lower density with more analysed elements is desirable to high density sampling with low data quality. When interesting locations are found with sparse data, the density can then still be increased in further studies adjusted to the specific ore type and deposit size to also find smaller targets (for example, vein type or small sub-outcropping deposits). Nevertheless, statistical analysis alone is limited and always needs cooperation with experts providing interpretation of outliers and classifying them as potential mineral deposits worth to be analysed further.

CRediT authorship contribution statement

Patricia Puchhammer: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Charmee Kalubowila:** Data curation, Visualization, Writing – review & editing. **Lorena Braus:** Conceptualization, Formal analysis, Methodology. **Solveig Pospiech:** Supervision, Writing – review & editing. **Pertti Sarala:** Supervision, Writing – review & editing. **Peter Filzmoser:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

We would like to thank our colleague Ana Jesus for sharing her geological knowledge with us without any hesitation contributing to more interesting insights. Co-funded by the European Union (SEMACRET, Grant Agreement no. 101057741) and UKRI (UK Research and Innovation). The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

Appendix A. Q-Q plots

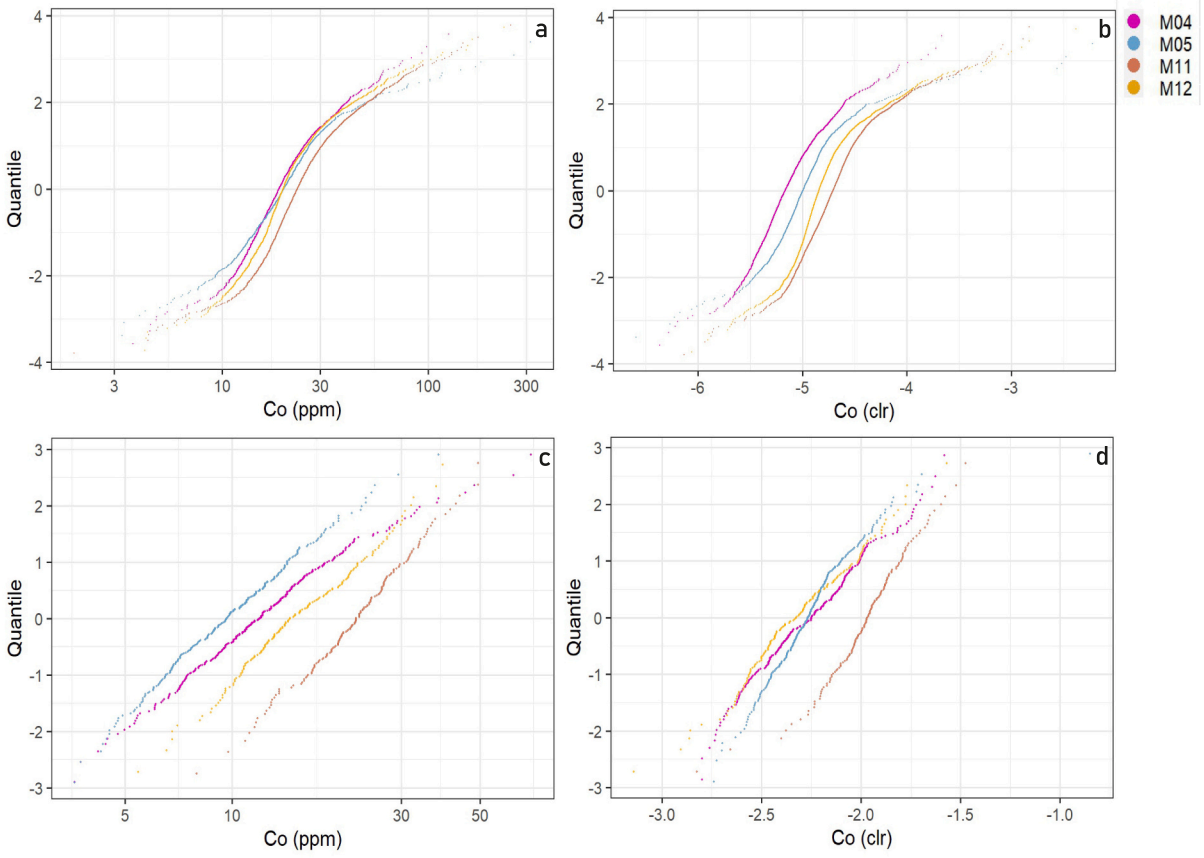

Fig. A.1. Q-Q plots of Co: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

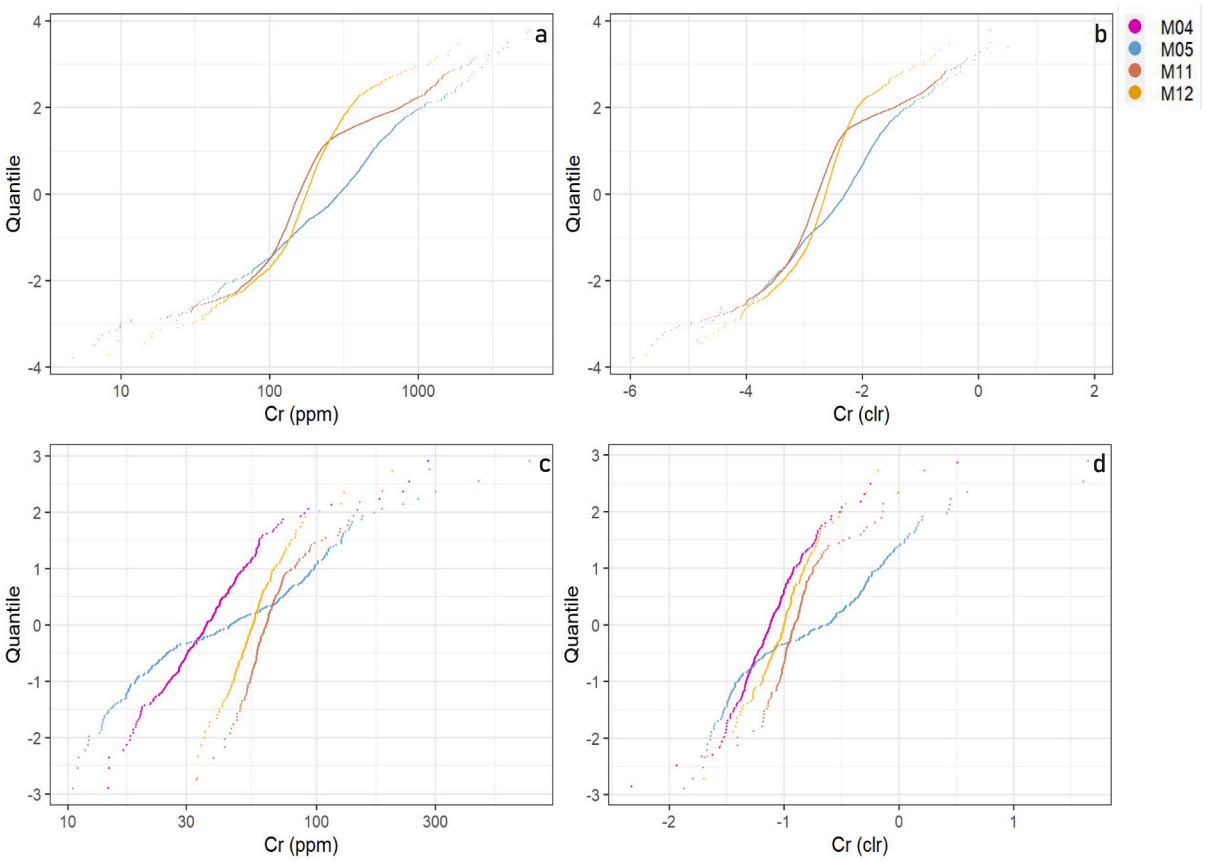

Fig. A.2. Q-Q plots of Cr: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

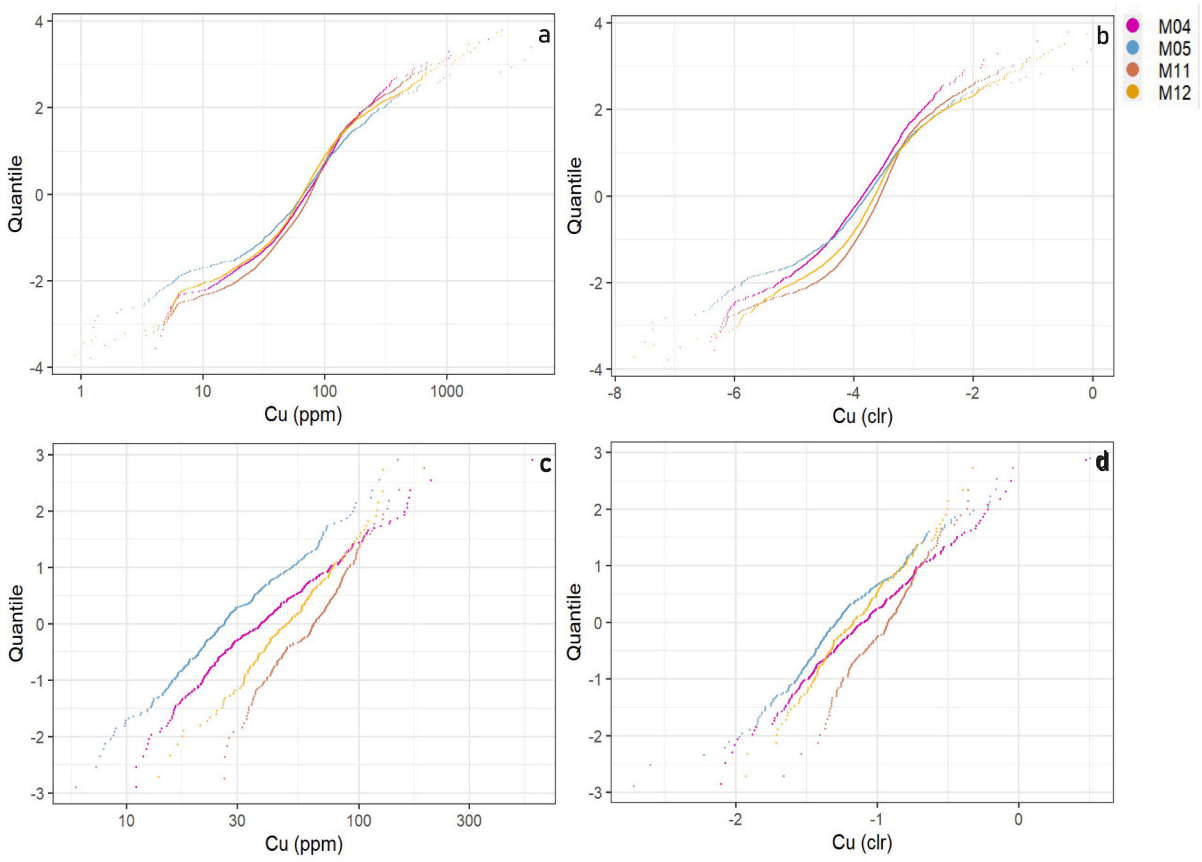

Fig. A.3. Q-Q plots of Cu: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

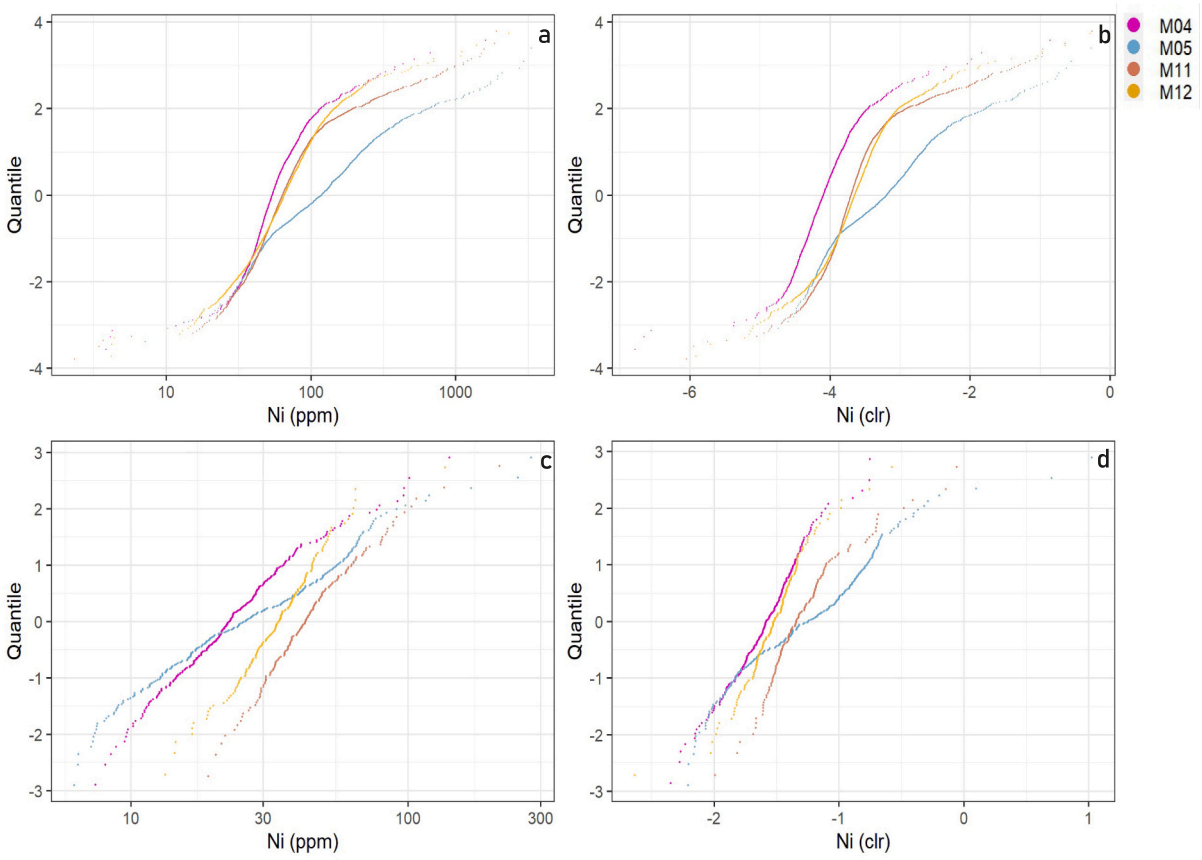

Fig. A.4. Q-Q plots of Ni: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

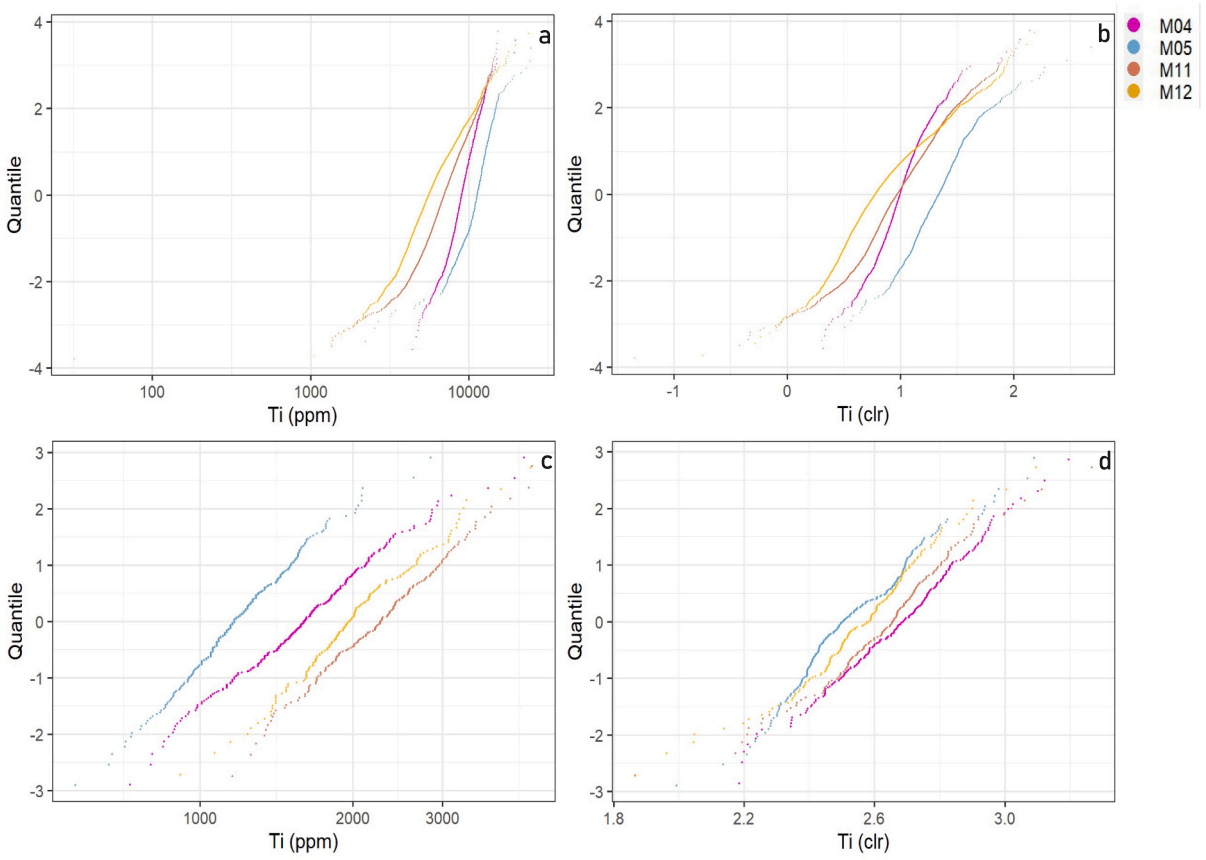

Fig. A.5. Q-Q plots of Ti: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

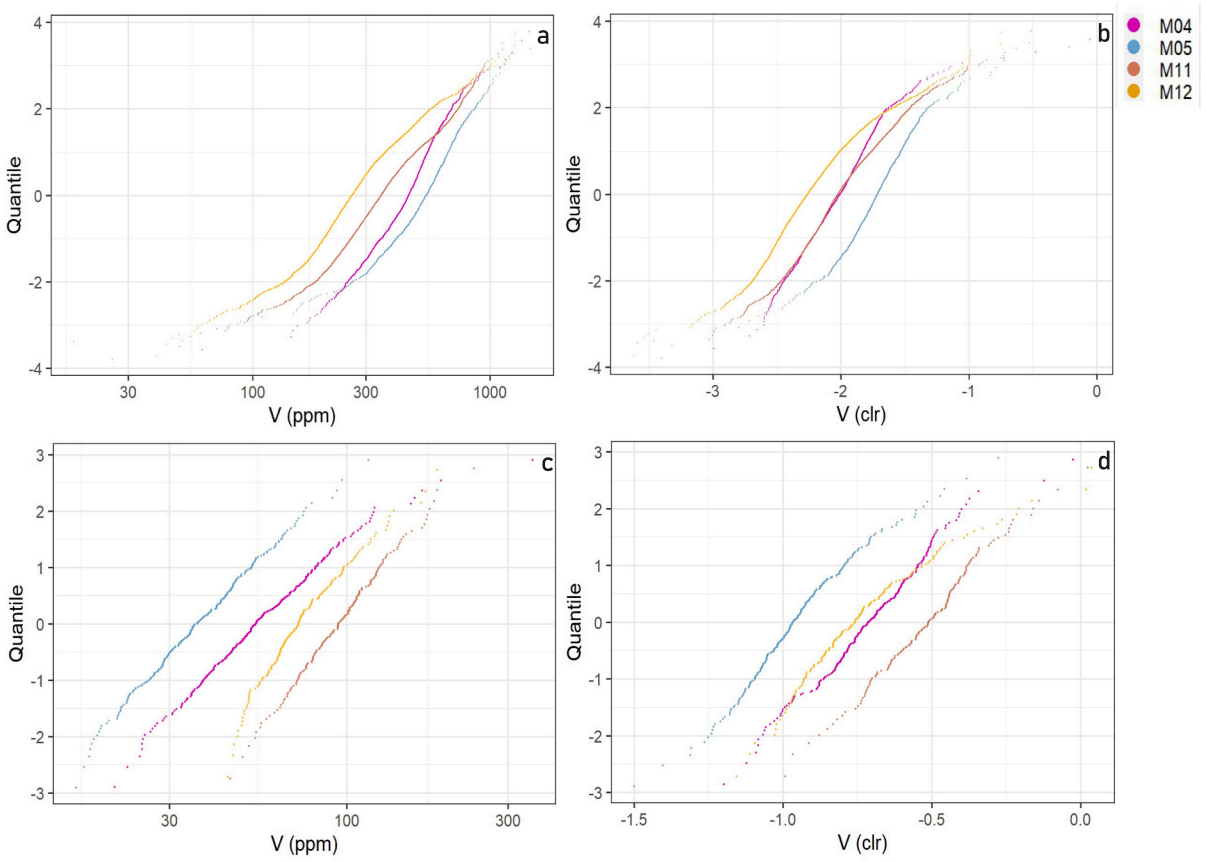

Fig. A.6. Q-Q plots of V: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

References

Boudt, K., Rousseeuw, P.J., Vanduffel, S., Verdonck, T., 2020. The minimum regularized covariance determinant estimator. Stat. Comput. 30, 113–128. https://doi.org/10.1007/s11222-019-09869-x.

Braus, L., 2023. Local Outlier Detection for Compositional Data (Diploma thesis). Technische Universität Wien. https://doi.org/10.34726/hss.2023.105504.

Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93–104. https://doi.org/10.1145/342009.335388.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. J. R. Stat. Soc. Ser. D Stat. 47, 431–443.

Cressie, N., 2015. Statistics for Spatial Data. John Wiley & Sons.

Ernst, M., Haesbroeck, G., 2016. Comparison of local outlier detection techniques in spatial multivariate data. Data Min. Knowl. Disc. 31, 371–399. https://doi.org/10.1007/s10618-016-0471-0.

Filzmoser, P., 2004. A Multivariate Outlier Detection Method. Citeseer.

Filzmoser, P., Gschwandtner, M., 2012. mvoutlier: multivariate outlier detection based on robust methods (R package version 2.1.1) URL: https://cran.r-project.org/package=mvoutlier.

Filzmoser, P., Ruiz-Gazen, A., Thomas-Agnan, C., 2013. Identification of local multivariate outliers. Stat. Pap. 55, 29–47. https://doi.org/10.1007/s00362-013-0524-z.

Filzmoser, P., Hron, K., Templ, M., 2018. Applied Compositional Data Analysis. Springer, Cham.

Geological Survey of Finland, 1995. Regional till geochemistry. URL: https://hakku.gtk.fi/en/locations/search.

Geological Survey of Finland, 2013. Targeting till geochemistry. URL: https://hakku.gtk.fi/en/locations/search.

Geological Survey of Finland, 2016. Mineral deposits. URL: https://hakku.gtk.fi/en/locations/search.

Gustavsson, N., Noras, P., Tanskanen, H., 1979. Summary: Report on Geochemical Mapping Methods, Technical Report. Geological Survey of Finland. URL: https://tupa.gtk.fi/julkaisu/tutkimusraportti/tr_039.pdf.

Jesus, A.P., Mateus, A., Munhá, J., 2014. Internal architecture and Fe-Ti-V oxide ore genesis in a Variscan synorogenic layered mafic intrusion, the Beja Layered Gabbroic Sequence (Portugal). Lithos 190-191, 111–136. https://doi.org/10.1016/j.lithos.2013.12.001.

Jorge, R., Fernandes, P., Rodrigues, B., Pereira, Z., Oliveira, J., 2013. Geochemistry and provenance of the Carboniferous Baixo Alentejo Flysch Group, South Portuguese Zone. Sediment. Geol. 284-285, 133–148. URL: https://www.sciencedirect.com/science/article/pii/S0037073812003302 https://doi.org/10.1016/j.sedgeo.2012.12.005.

Lutynski, P., 2019. Akanvaara project, Finland, Technical Report, Strategic Resources Inc, Canada. URL: https://strategic-res.com/site/assets/files/3618/akanvaara_43-101_10-06-2019.pdf.

Maier, W., 2015. Geology and petrogenesis of magmatic Ni-Cu-PGE-Cr-V deposits: an introduction and overview. In: Mineral Deposits of Finland. Elsevier, pp. 73–92.

Marjoribanks, R., 2010. Geological Methods in Mineral Exploration and Mining. Springer Science & Business Media.

Mert, M.C., Filzmoser, P., Hron, K., 2016. Error propagation in isometric log-ratio coordinates for compositional data: theoretical and practical considerations. Math. Geosci. 48, 941–961.

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

Puchhammer, P., Filzmoser, P., 2023a. Spatially smoothed robust covariance estimation for local outlier detection. J. Comput. Graph. Stat. 0, 1–30. https://doi.org/10.1080/10618600.2023.2277875.

Puchhammer, P., Filzmoser, P., 2023b. ssMRCD: spatially smoothed MRCD estimator. https://cran.r-project.org/package=ssMRCD (URL: R package version 0.1.0).

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., 2014a. Chemistry of Europe's Agricultural Soils - Part A: Methodology and Interpretation of the GEMAS Data Set. Schweizerbart Science Publishers, Stuttgart, Germany. URL: http://www.schweizerbart.de//publications/detail/isbn/9783510968466/Geologisches_Jahrbuch_Reihe_B_Heft_B102.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., 2014b. Chemistry of Europe's agricultural soils - Part B: general background information and further analysis of the GEMAS data set, Schweizerbart Science Publishers, Stuttgart, Germany. URL: http://www.schweizerbart.de//publications/detail/isbn/9783510968473/Geologisches_Jahrbuch_Reihe_B_Heft_B103.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. Math. Stat. Appl. 8, 37.

Salminen, R., Tarvainen, T., 1995. Geochemical mapping and databases in Finland. J. Geochem. Explor. 55, 321–327.

Schubert, E., Zimek, A., Kriegel, H.-P., 2012. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier

detection. Data Min. Knowl. Disc. 28, 190–237. https://doi.org/10.1007/s10618-012-0300-z.
SEMACRET Project, 2023. URL. https://semacret.eu.
Signorell, A., et al., 2017. DescTools: tools for descriptive statistics (R package version 0.99.23) URL: https://cran.r-project.org/package=DescTools.
Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: An R-package for Robust Statistical Analysis of Compositional Data. John Wiley and Sons. https://doi.org/10.1002/9781119976462.ch25.
Zimek, A., Filzmoser, P., 2018. There and back again: outlier detection between statistical reasoning and data mining algorithms. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8, e1280.