Journal of Geochemical Exploration 259 (2024) 107425

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

How to choose a proper representation of compositional data for mineral exploration?

Behnam Sadeghi [a,b,*], Hossein Molayemat [c], Vera Pawlowsky-Glahn [d]

[a] CSIRO Mineral Resources, Australian Resources Research Centre, Kensington, WA 6151, Australia
[b] Earth and Sustainability Science Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, NSW 2052, Australia
[c] Department of Mining and Metallurgical Engineering, Yazd University, Yazd, Iran
[d] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

ARTICLE INFO

Keywords:
Compositional data analysis
Logratio
Principal balances
Sweden
Till
Orogenic gold

ABSTRACT

Regional mineral exploration is based on geochemical data of which the nature is compositional and frequently involves a large number of components. Consequently, it mostly needs multivariate dimension reduction methods such as principal component analysis (PCA) and its various robust versions. The application of such methods, defined for real random variables, require the data to be represented in coordinates supported in the real space. However, a common problem in exploration geochemistry is to select the appropriate representation. Using centered (clr) and isometric (ilr) logratio coordinates to discriminate anomalous zones for orogeny gold exploration throughout Sweden revealed that there is, as expected, no difference between the two representation methods. The main difference affects the interpretation of the coordinates used. This is observed for regional scale exploration, while it is also needed to study different ways of representing geochemical data in local scale.

1. Introduction

Geochemical anomaly identification is one of the most significant parts of mineral exploration projects. To this end, a variety of spatial models have been developed for different types of geochemical data. The main purpose of such applications is to define and classify anomalies and the relevant background. This has been attempted by considering the element concentrations through univariate (in global scales) and multivariate (in local scales) points of view (Sadeghi, 2020). Taking into account that the nature of geochemical data is compositional and any approach with raw data might lead to spurious results, it is clear that a compositional approach will provide a better overview of samples' spatial relationships.

A variety of multivariate techniques to assist in clustering samples and separating anomalous observations have been applied in geochemical studies, including canonical variates derived from data dimension reduction methods that reflect geochemical processes (Grunsky et al., 2014; Makvandi et al., 2016; Mueller and Grunsky, 2016; Greenacre, 2010, 2018a, 2018b; McKinley et al., 2018; Grunsky and de Caritat, 2020; Sauro Graziano et al., 2020). These include parametric methods such as the various forms of principal component or factor analysis (PCA and FA) (Chork and Salminen, 1993; Grunsky and Kjarsgaard, 2016) and K-means clustering (Kaski, 1997), that are still commonly used in multivariate geochemical data assessment, and the less common non-parametric ones such as self-organising maps and neural networks (Kohonen, 1995; Foody, 1997; Clare and Cohen, 2001).

Geochemical datasets represent the parts of a whole and are thus, by definition, compositional (Egozcue et al., 2003; Filzmoser et al., 2009; Nazarpour et al., 2015). This means that geochemical analyses provide results that are usually registered as proportions, percentages, moles per liter, or any other units reflecting parts of the whole (Aitchison, 1986; Martín-Fernández et al., 2017). In other words, such values are not absolute values; they provide the elements' relative information considering the whole sample. A composition is actually an equivalence class (Aitchison, 1992; Barceló-Vidal and Martín-Fernández, 2016) implying *scale invariance* of the relationships between different parts. This property justifies the representation of the data in the simplex, a constraint subset of real space. Working in this subset, it is easy to show that standard methods, based on the Euclidean geometry of the real space, might lead to spurious results, a well-known effect since Pearson (1897) coined the term *spurious correlation*. To get this effect resolved, several representations have been proposed, collectively called *logratio transformations* (Aitchison, 1986; Egozcue et al., 2003; Gallo and Buccianti, 2013). Their properties have been analysed in the framework of

* Corresponding author at: CSIRO Mineral Resources, Australian Resources Research Centre, Kensington, WA 6151, Australia.
E-mail addresses: z5218858@zmail.unsw.edu.au, behnam.sadeghi@csiro.au (B. Sadeghi).

https://doi.org/10.1016/j.gexplo.2024.107425
Received 31 August 2023; Received in revised form 28 January 2024; Accepted 6 February 2024
Available online 12 February 2024

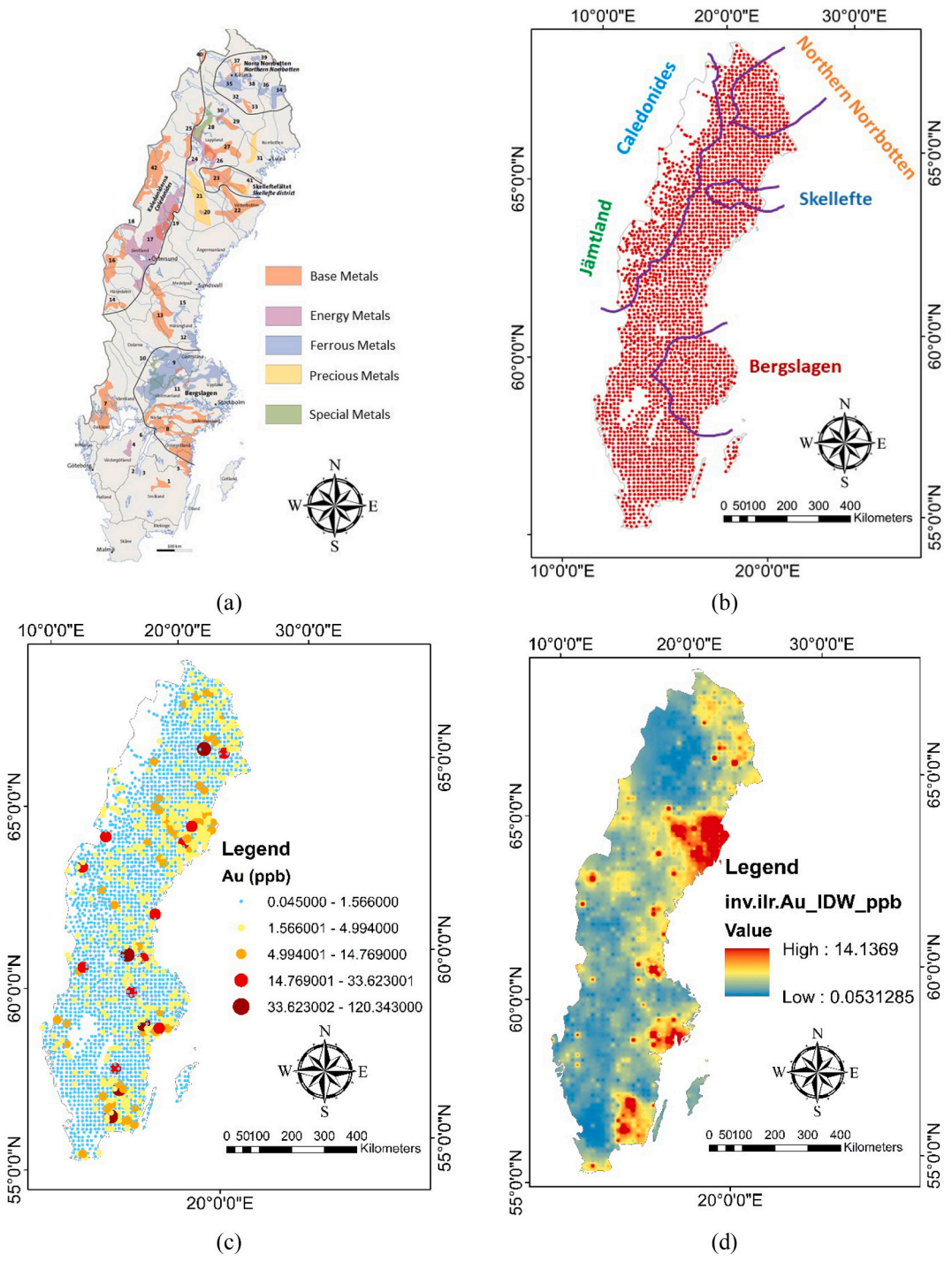

Fig. 1. (a) Geology and main metallogenic areas in Sweden (from Andersson et al., 2014); this map has been originally published as a part of the Geochemical Atlas of Sweden. The numbers refer to some description about different areas throughout the country. The descriptions are available in the Atlas. (b) Till sample locations along with main metallogenic provinces (from Sadeghi, 2020); (c) Au raw data map; and (d) IDW interpolated ilr-Au map.

the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001; Pawlowsky-Glahn et al., 2015).

Among the dimension reduction methods, PCA is one of the most robust models. However, the loadings are mostly non-zero, and the PCs involve all the original parts; that is why the PCA interpretation can be extremely demanding. Several approaches have been developed in the real space framework as a solution for such complexities which are mostly based on either the rotation of the PCs or variable selection techniques (Jolliffe, 2002; Jolliffe et al., 2003). Note that in the framework of compositional data no component can be suppressed without redoing the whole calculation, as the sum of the coefficients has to be necessarily zero. Therefore, in the framework of compositional data we find, to this end, the approach known as *principal balances* (Martín-Fernández et al., 2017; Pawlowsky-Glahn et al., 2011a, 2011b).

To classify the geochemical anomalies detected based on element values or their relevant PCs, various models have been developed such as fractal/multifractal models (Daya Sagar et al., 2023; Sadeghi, 2024). Such methods include a variety of models in 2*D* and 3*D* and for different types of data samples (Zuo and Wang, 2016; Cheng and Agterberg, 2021). Some of the well-established fractal/multifractal models are

Table 1
Descriptive compositional statistics of the 6 elements in the present study.

Element	Center	Min.	1st quartile	Median	3rd quartile	Max.
Ag (ppb)	44.45	4.62	32.51	44.93	59.31	517.42
As (ppm)	2.431	0.16	1.06	2.31	5.41	253.28
Au (ppb)	0.5546	0.05	0.27	0.67	1.29	120.34
Sb (ppm)	0.1420	0.02	0.08	0.13	0.25	2.33
Te (ppm)	17.77	1.63	11.61	17.34	25.88	328.19
W (ppm)	0.3552	0.01	0.21	0.37	0.60	8.85

Table 2
Scales of till geochemical samples.

Scale	Sampling density (sites per km^2)	Dispersal length trains detected	Objectives, target
Reconnaissance	<0.01 (0.15 in Sweden)	500 to 1000 km	Geochemical provinces, continental glacial dynamics
Regional	0.01 – 0.1	10 to 100 km	Mineral belt, kimberlite cluster
Local	1 – 10	1 to 5 km	Mineralised ground, tails of dispersal trains from orebodies
Detailed	>10	100 m to 1 km	Individual orebodies

(After McMartin and McClenaghan (2001).)

Table 3
Loadings and vector of the variances of PCs obtained from clr-data.

	PC1	PC2	PC3	PC4	PC5
clr_{Au}	0.88	− 0.15	0.14	− 0.047	− 0.90
clr_{Ag}	− 0.024	0.32	− 0.22	− 0.16	0.81
clr_{As}	− 0.30	− 0.59	0.33	0.50	0.20
clr_{Sb}	− 0.31	− 0.24	0.10	− 0.78	− 0.24
clr_{Te}	− 0.094	− 0.014	− 0.79	0.27	− 0.37
clr_{W}	− 0.16	0.68	0.44	0.22	− 0.31
Var.	1.41	0.72	0.51	0.26	0.16
% cum. var.	46.19	69.61	86.29	94.87	100.00

Table 4
Loadings and vector of the variances of RPCA applied to ilr-data.

	RPC1	RPC2	RPC3	RPC4	RPC5
ilr_1	0.45	− 0.36	− 0.21	− 0.28	0.74
ilr_2	0.61	0.59	0.23	0.45	0.16
ilr_3	0.63	− 0.40	− 0.19	− 0.01	− 0.64
ilr_4	− 0.039	− 0.57	0.71	0.40	0.10
ilr_5	− 0.17	− 0.20	− 0.61	0.74	0.12
Var.	1.42	0.712	0.501	0.262	0.140
% cum. var.	46.75	70.23	86.73	95.38	100.00

number-size (Mandelbrot, 1983; Sadeghi et al., 2012), concentration-area (C-A: Cheng et al., 1994, see Sadeghi, 2021c); spectrum-area (Cheng et al., 1999, see Sadeghi, 2021c); singularity (Cheng, 2007, see Sadeghi and Agterberg, 2021); concentration-concentration (Sadeghi, 2021a), concentration-distance from centroids (Sadeghi and Cohen, 2021b), simulated-fractal models (Sadeghi et al., 2015; Sadeghi, 2021b); category based (Sadeghi and Cohen, 2021a), and their 3D equivalents. The critical point is all about selecting the most appropriate, consistent, and efficient model, not only in classification but also in multivariate inference (Egozcue et al., 2015).

In this research, an orogenic gold commodity (OGC), represented by a vector of chemical elements including Au, Ag, As, Sb, Te, and W, is selected among the whole set of elements available, obtained from till samples collected and analysed by the Geological Survey of Sweden (SGU: Andersson et al., 2014) from throughout the whole country. One objective here is to demonstrate what transformation works more efficiently. To evaluate the OGC from a multivariate point of view, principal component (PCA) and robust principal component (RPCA) analysis are applied respectively to isometric logratio (ilr) and centred logratio (clr) transformed data, and are compared with principal balances (PB), which are useful to obtain sparse balances which are as close as possible to the principal components gained by clr and are useful in the interpretation of the PCs when they are close. In the end, the continuous interpolated maps of the first principal components, PC1 and RPC1, are generated and classified using a number-size (N-S) fractal model, and the OGC anomalies in Sweden are identified.

2. Geology and mineralization of Sweden

Sweden is geologically dominated by the Fennoscandian Shield (FS) metamorphic and crystalline rock in the north, Caledonian orogeny in NW and northern Sweden and, shortly, Phanerozoic sedimentary units in the south (Beckholmen and Tirén, 2009). Around 22% of the country area is covered by 42 metallogenic sub-provinces (Hallberg et al., 2012; Andersson et al., 2014), including 16 base-metal deposits (Cu, Pb, Zn, Ni and Co), 13 ferrous metal deposits (Fe, Mn, Ti, V and Cr), and four precious metal deposits (Au, Ag and PGE). All these sub-provinces are classified as four main mining districts (Allen et al., 1996; Eilu, 2012; Andersson et al., 2014; Sadeghi and Cohen, 2021b, 2021a): (1) the Northern Norrbotten, (2) Skellefte districts, both in Northern Sweden, dominated by sulfidic mineral deposits, (3) the Caledonides in NW Sweden, and (4) the Bergslagen area in South-Central Sweden, dominated by sulfidic mineral deposits (Fig. 1a). The detailed geology of Sweden is out of the scope of this paper. It can be found in Andersson et al. (2014); Sadeghi (2020); Sadeghi and Cohen (2021b) or in Sadeghi and Cohen (2021a). A number of 2, 578 till samples, collected by the Geological Survey of Sweden (SGU), are used in the present research to illustrate how compositional data methods can focus on the Au mineralization (through a multivariate study on Au, Ag, As, Sb, Te, and W) throughout Sweden (Andersson et al., 2014, Fig. 1b). Table 1 shows descriptive statistics of the 6 elements.

The sample density and sample spacing correspond to the reconnaissance studies defined by McMartin and McClenaghan (2001) (Table 2).

The Au raw data map is presented in (Fig. 1c). An inverse distance weighted (IDW) interpolated map was generated using the above-mentioned data to give an initial overview of the Au throughout Sweden (Fig. 1d). This map is generated by interpolation of the whole ilr vector and backtransformation of the interpolated values to the simplex. The map demonstrates that the high concentrations of Au are mainly located in Southern Sweden and Bergslagen (Eastern Sweden), followed by Skellefte, Southern Caledonides (i.e., Jämtland) and Norrbotten (NE Sweden).

3. Methodologies and data processing

3.1. Characterisation and basic assumptions

Geological evaluation and models are highly influenced by a variety of errors and sources of uncertainties such as (Sadeghi, 2020; Sadeghi et al., 2021): lack of samples, unclear understanding of the geology, relevant errors in geochemical lab analysis, data analysis, interpolation, misunderstanding of geological and geochemical processes, measurement and analysing instrument errors, etc. Moreover, further potential biases may be caused by not taking into account the above-mentioned compositional nature of the data. Such biases are a consequence of an inappropriate sample space and/or structure and can be avoided using an adequate framework, consistent with the nature of the data. Such a framework is given for compositional data by considering the simplex—endowed with the Aitchison geometry—as sample space (Pawlowsky-Glahn and Egozcue, 2001, 2006; Filzmoser et al., 2009; Carranza, 2011; Pawlowsky-Glahn and Egozcue, 2011; Gallo and

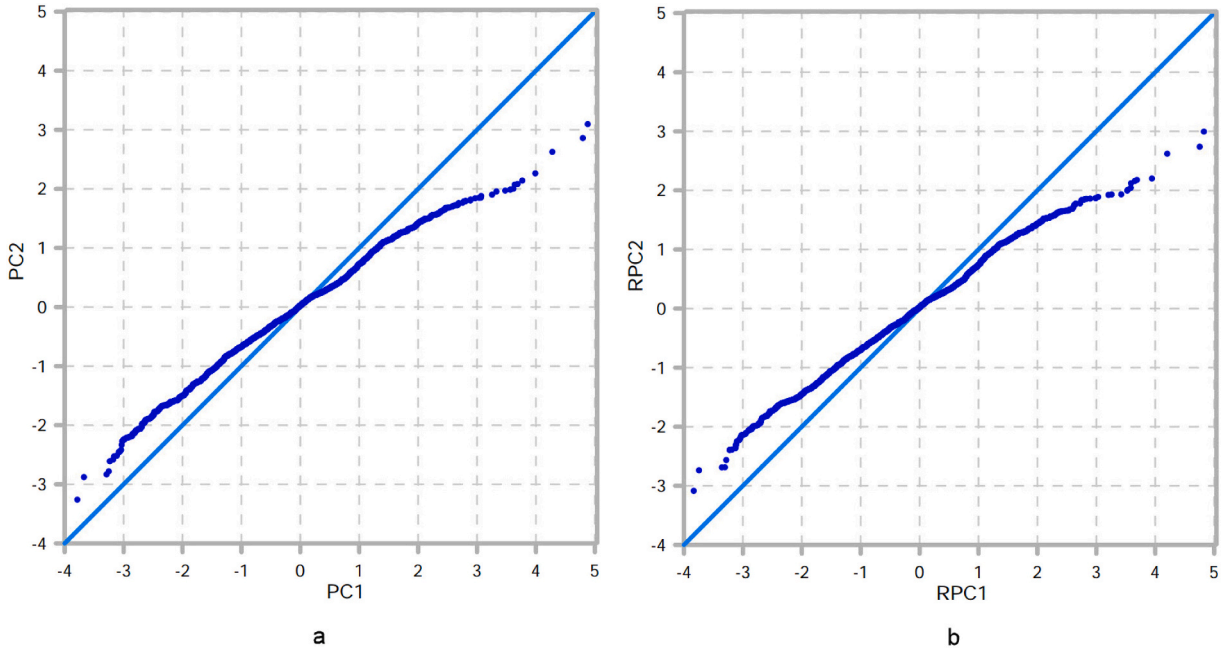

Fig. 2. Q-Q plots of the a) PC1 vs PC2 clr-scores, and b) RPC1 vs RPC2 ilr-scores in Sweden.

Buccianti, 2013; Buccianti and Grunsky, 2014; Nazarpour et al., 2015; Daya Sagar et al., 2018; Zuzolo et al., 2018; Pospiech et al., 2020; Sadeghi et al., 2021). In Pawlowsky-Glahn and Egozcue (2001) it was shown that for $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ a D-part composition, and κ a constant representing the sum of the components, the simplex, defined as

$$\mathscr{S}_c^D = \left\{ x = (x_1, \ldots, x_D) \,|x_i > 0, i = 1, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\}, \quad (1)$$

with perturbation and powering as alternatives of addition and multiplication by a scalar in real space, can be endowed with a Euclidean space structure by adding an inner product and its associated norm and distance (Pawlowsky-Glahn and Egozcue, 2001). It would be rational to work with compositions in the simplex, as it defines an interpretable and meaningful structure.

Note that compositions can be projected onto particular directions by the inner product and then compositional vectors can be studied for the angle between vectors and, in particular, for the orthogonality between them. With the norm, the length of a composition, or distance from the neutral element, where all components or parts are identical, can be calculated. For the computation of distances in the simplex, the Aitchison distance is available. It is a measure of the difference between two compositions. In practice, the Aitchison distance can be calculated as the Euclidean distance of two clr transformed data. The Aitchison distance is known to be scale invariant, permutation invariant, perturbation invariant, and sub-compositionally dominant. Therefore, it is a sensible criterion for calculating the dissimilarity - namely, the distance between two compositions.

All above-mentioned operations define a real Euclidean space structure in the simplex, which was termed in Pawlowsky-Glahn and Egozcue (2001) for the first time *Aitchison geometry* to avoid confusion with the usual Euclidean geometry in real space. This property is crucial, as real Euclidean spaces have the appropriate property that working in coordinates using standard methods devised for real random variables is equivalent to working in the simplex with the operations mentioned above (Queysanne, 1973). It is important to be aware that all standard methods assume explicitly or implicitly the real space endowed with the usual Euclidean geometry as the sample space.

3.2. Exploratory data analysis

An exploratory analysis of compositional data starts with the calculation of the center, variation matrix, and total variance. Initially, it is mandatory to check for outliers, censored data, and especially, rounded zeros and, if necessary, to apply methods that take the outliers into account. A series of recent developments can be found in Martín-Fernández et al. (2012) or Palarea-Albaladejo et al. (2022).

Note that standard descriptive statistics, such as arithmetic mean and variance, are not suitable representatives of central tendency and data dispersion. The reason is that these statistics do not fit with the Aitchison geometry. Alternatively, the concept of center, variation matrix, and total variance were introduced by Aitchison (1986). The center is simply defined as the geometric mean of the desired parts in their raw, closed form. A variation matrix can be defined as the matrix of variances of logratios of pairs of parts. In the following, the total variance is the sum of all values of the calculated variation matrix divided by $2D$ where D is the number of parts. It is notable that the two latter parameters do not depend on the closure constant, as the constant cancels out when taking ratios.

Another step for compositional data exploratory analysis is to calculate and interpret biplots. A biplot can be built by considering any two principal components (PCs) of clr transformed data; e.g. the first two PCs, which explain the greatest variance. Biplots can be defined as graphical tools to simultaneously represent the rows and columns of a matrix by means of a rank-2 approximation (Aitchison, 1983; Aitchison and Greenacre, 2002; Pawlowsky-Glahn et al., 2015). The usefulness of clr biplots as a tool to identify the mutual behaviour of parts in a proper geometry is discussed in numerous studies. This property helps the analyst to properly perform the grouping of parts and subsequently, define the intended isometric balances. Examples of this procedure can be found in Olea and Luppens (2015), Molayemat et al. (2018) and Molayemat et al. (2022).

More complicated for exploratory analysis of compositional data is to interpret the PCs computed on clr transformed data. In addition to the availability of dimension reduction, this technique can be used as an appropriate modelling tool when the presence of a trend in the data is suspected without knowing about any possible external controlling variable.

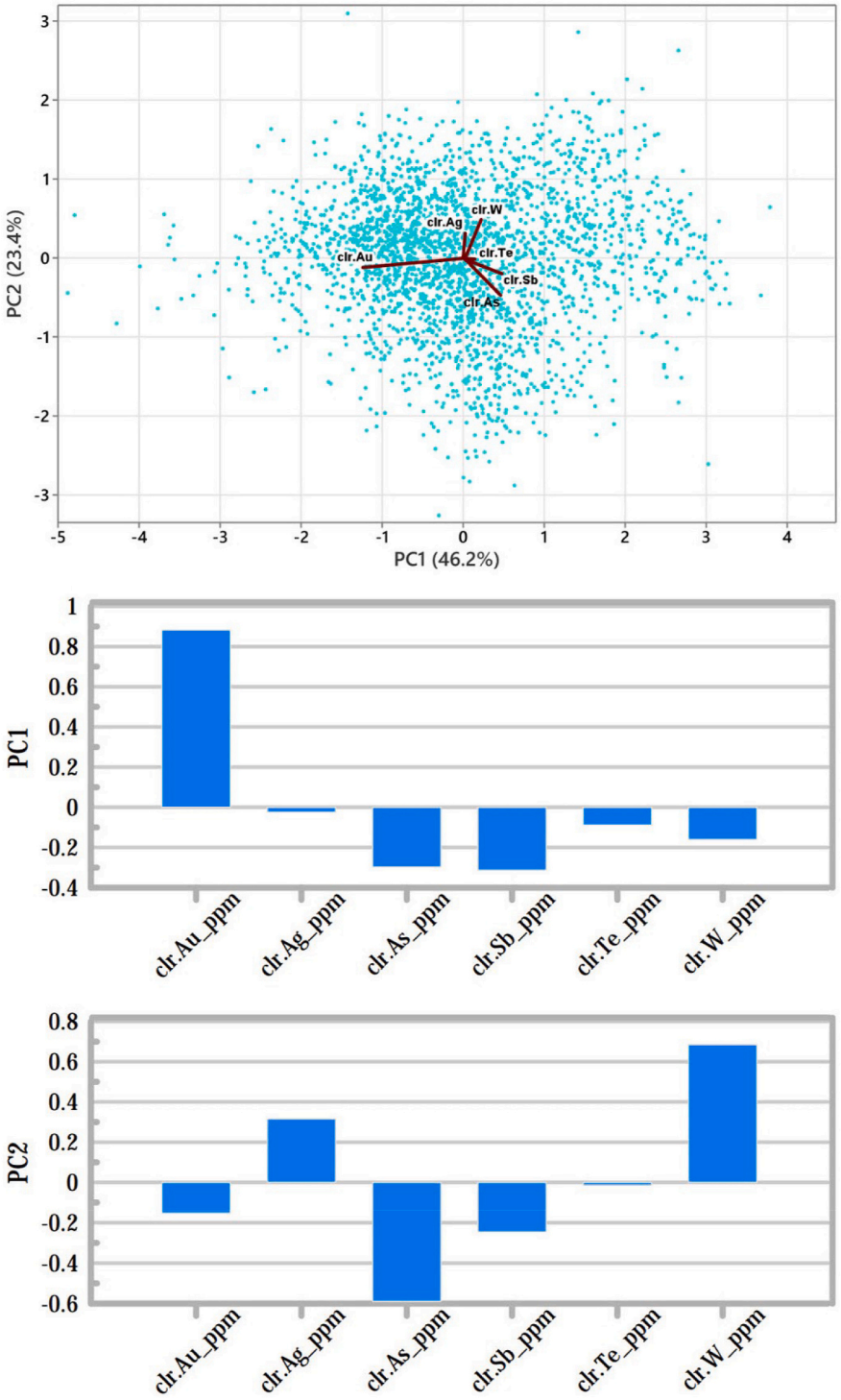

Fig. 3. Covariance biplot and eigenvector plots of PCA applied to clr-data (69.6 % explained variance).

Another debatable aspect of compositional data exploratory analysis is the computation and interpretation of correlation coefficients, which are by definition used to measure pairwise association. The problem is that the known correlation coefficients are subcompositionally incoherent. It is crucial to be aware that logratio measures such as the correlation of clr components, pivot coordinates (Stefelova et al., 2021), and symmetric balances (Kynčlová et al., 2017) are also subcompositionally incoherent (Egozcue et al., 2019; Egozcue and Pawlowsky-Glahn, 2023). They appear when closure is involved in the coefficient functions and may drastically affect geological and mineral exploration studies, as they use the correlations to choose pathfinder elements (e.g., for gold exploration), to delineate a target area for detailed exploration, and to reveal geochemical patterns in general. A known treatment is to define a sequential binary partition (SBP) matrix, based on the exploratory project's requirements, and then compute the desired correlation coefficients between the calculated orthonormal balances. Examples can be found in Pawlowsky-Glahn et al. (2015) and Egozcue et al. (2019). A newer way of quantifying the relation of compositional parts is introduced in Egozcue and Pawlowsky-Glahn (2023). There, a Proportionality Index of Parts (PIP) is proposed. The PIP ranges from 0 to 1 and

Table 5
Loadings and vector of the variances of the PBs generated by O, C, and W hierarchical clustering (defined in Section 3.3). The PB order is based on the variances.

	O1	O4	O2	O5	O3
Au	0.87	0.00	− 0.22	0.00	− 0.18
Ag	0.00	0.00	0.00	0.00	0.91
As	− 0.29	0.41	− 0.22	0.71	− 0.18
Sb	− 0.29	0.41	− 0.22	0.71	− 0.18
Te	0.00	0.00	0.89	0.00	− 0.18
W	− 0.29	− 0.82	− 0.223	0.00	− 0.18
Var.	1.39	0.65	0.48	0.29	0.25
% cum. var.	45.48	66.68	82.53	91.85	100.00

	C1	C4	C3	C2	C5
Au	− 0.87	0.00	− 0.18	− 0.22	0.00
Ag	0.00	0.00	− 0.18	0.89	0.00
As	0.29	− 0.41	− 0.18	− 0.22	− 0.71
Sb	0.29	− 0.41	− 0.18	− 0.22	0.71
Te	0.00	0.00	0.91	0.00	0.00
W	0.29	0.82	− 0.18	− 0.22	0.00
Var.	1.39	0.65	0.44	0.29	0.29
% cum. var.	45.48	66.68	81.07	90.67	100.00

	W1	W2	W3	W4	W5
Au	0.91	0.00	0.00	0.00	0.00
Ag	− 0.18	− 0.37	− 0.41	0.00	0.71
As	− 0.18	0.55	0.00	0.71	0.00
Sb	− 0.18	0.55	0.00	− 0.71	0.00
Te	− 0.18	− 0.37	− 0.41	0.00	− 0.71
W	− 0.18	− 0.37	0.82	0.00	0.00
Var.	1.36	0.68	0.48	0.29	0.26
% cum. var.	44.35	66.50	82.31	91.64	100.00

coherently is defined based on the properties linking the space of parts and observations. The PIP can be computed using the Aitchison distance in the space of parts and is invariant under subcomposition.

3.3. Logratio transformations

Aitchison (1986) demonstrated a transformation, which is compatible with the equivalence class character of compositional data. This transformation is a scale-invariant logratio, which is in fact a log-contrast:

$$\sum_{i=1}^{D} a_i ln x_i = ln\left(\prod_{i=1}^{D} x_i^{a_i}\right), \qquad \sum_{i=1}^{D} a_i = 0, \tag{2}$$

where the data set is an $(n \times D)$ matrix $\mathbf{X}$, $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ is a row of $\mathbf{X}$, and a_i, $i = 1, 2, \ldots, d$ are real numbers subject to a zero sum constraint.

In the framework of the Aitchison geometry, CoDa can be represented in several different systems of log-contrasts, which are more or less convenient depending on the research question to be answered. As a general rule, orthonormal logratio coordinates, denoted as olr-coordinates or ilr-coordinates, as they were previously known (Egozcue and Pawlowsky-Glahn, 2005, 2019a, 2019b; Pawlowsky-Glahn and Egozcue, 2011; Martín-Fernández et al., 2017) can be used without further limitations. One particular type of ilr coordinates is *balances*, which can be defined based on Sequential Binary Partitions (SBP).

Frequently, for computational convenience, the clr representation is applied as:

$$\mathrm{clr}_k\left(\mathbf{x}\right) = ln\frac{x_k}{\left[\left(\prod_{i=1}^{D} x_i\right)^{1/D}\right]}, \quad k = 1, 2, \ldots, D. \tag{3}$$

The computational simplicity of clr and its isometric properties make it comparable to ilr. However, it is essential to note that clr is not a representation in orthonormal coordinates. It is, mathematically, representing vectors that are subject to a zero-sum constraint. These problems could be handled e.g., with generalised inverses. In other words, the major problem is not computational, but about the interpretation. As pointed out by Aitchison (2002), there is a potential risk in associating the components of $\mathrm{clr}_k(\mathbf{x})$ solely with the numerator, overlooking the fact that all parts present in the composition contribute to each term. This poses a significant issue as it may lead to misinterpretations (see Egozcue and Pawlowsky-Glahn (2023) for further details).

There is no standard or canonical ilr representation. For instance, PCs obtained using the clr result in an ilr which in each component involves all parts, i.e. it is an ilr not made up of balances.

Balances (Egozcue and Pawlowsky-Glahn, 2005) are a number of logratios with the geometric mean of r_k parts in the numerator and of s_k parts in the denominator and a normalising coefficient as stated in 4,

$$b_k = \sqrt{\frac{r_k \cdot s_k}{r_k + s_k}} ln \frac{\left(x_{n_1} \cdots x_{n_{r_k}}\right)^{1/r_k}}{\left(x_{d_1} \cdots x_{d_{s_k}}\right)^{1/s_k}}, \quad k = 1, 2, \ldots, D-1, \tag{4}$$

where both sets of parts are disjoint. To guarantee that the obtained logratios define coordinates on an orthonormal basis it is sufficient that they follow a sequential binary partition (SBP) (Egozcue and Pawlowsky-Glahn, 2006; Pawlowsky-Glahn et al., 2011a).

Once an SBP has been chosen, it is interesting to visualise it in a CoDa-dendrogram (Pawlowsky-Glahn et al., 2015), particularly when there are groups in the data. In this case, the CoDa-dendrogram can be considered as a visual ANOVA, as it is easy to detect which balances are very different and which behave similarly in the groups.

A particular type of balances is principal balances. They have the same structure, but they are built in a different way (Martín-Fernández et al., 2017). They attempt to emulate the PCs. They are built as sparse log-contrasts, the first one maximising the explained part of the total variance in the data, and the second one maximising the explained part of the remaining variance, and this is continued until the last principal balance is created. They can be written as log-linear functions:

$$pb_k = \sum_{i=1}^{D} a_{ki} ln x_i, \quad k = 1, 2, \ldots, D-1, \tag{5}$$

in which x_i is the random part for $i = 1, 2, \ldots, D$, $X = (X_1, X_2, \ldots, X_D)$ is a D-part random composition, and $a_k = (a_{k1}, a_{k2}, \ldots, a_{kD})$ is a vector of constants that can take up only three values, namely zero for those parts that do not participate in the balance, those that are positive have an expression $\sqrt{s/[(r+s)r]}$ and those that are negative are $-\sqrt{r/[(r+s)s]}$, with r the number of the positive loadings and s the number of negative loadings. Note that the coefficients in balances always sum to zero, as they are log-contrasts. The PBs maximize the explained variance. They include three main algorithms, namely: (1) optimal PBs (O), (2) constrained PCs (C), and (3) Ward hierarchical clustering (W) (Table 5). For a reduced number of parts, optimal PBs can be computed with affordable time and computer memory (Martín-Fernández et al., 2017).

Historically, one of the first transformations that was introduced was the additive logratio (alr) (Aitchison, 1986). For a D-part composition of $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, alr is defined as:

$$\mathrm{alr}\left(\mathbf{x}\right) = \left(ln\frac{x_1}{x_D}, ln\frac{x_2}{x_D}, \ldots, ln\frac{x_{D-1}}{x_D}\right), \tag{6}$$

where the denominator x_D can be any part of the composition. In the framework of the Aitchison geometry, this transformation represents the data in oblique coordinates, a fact that introduces a new difficulty to any analysis.

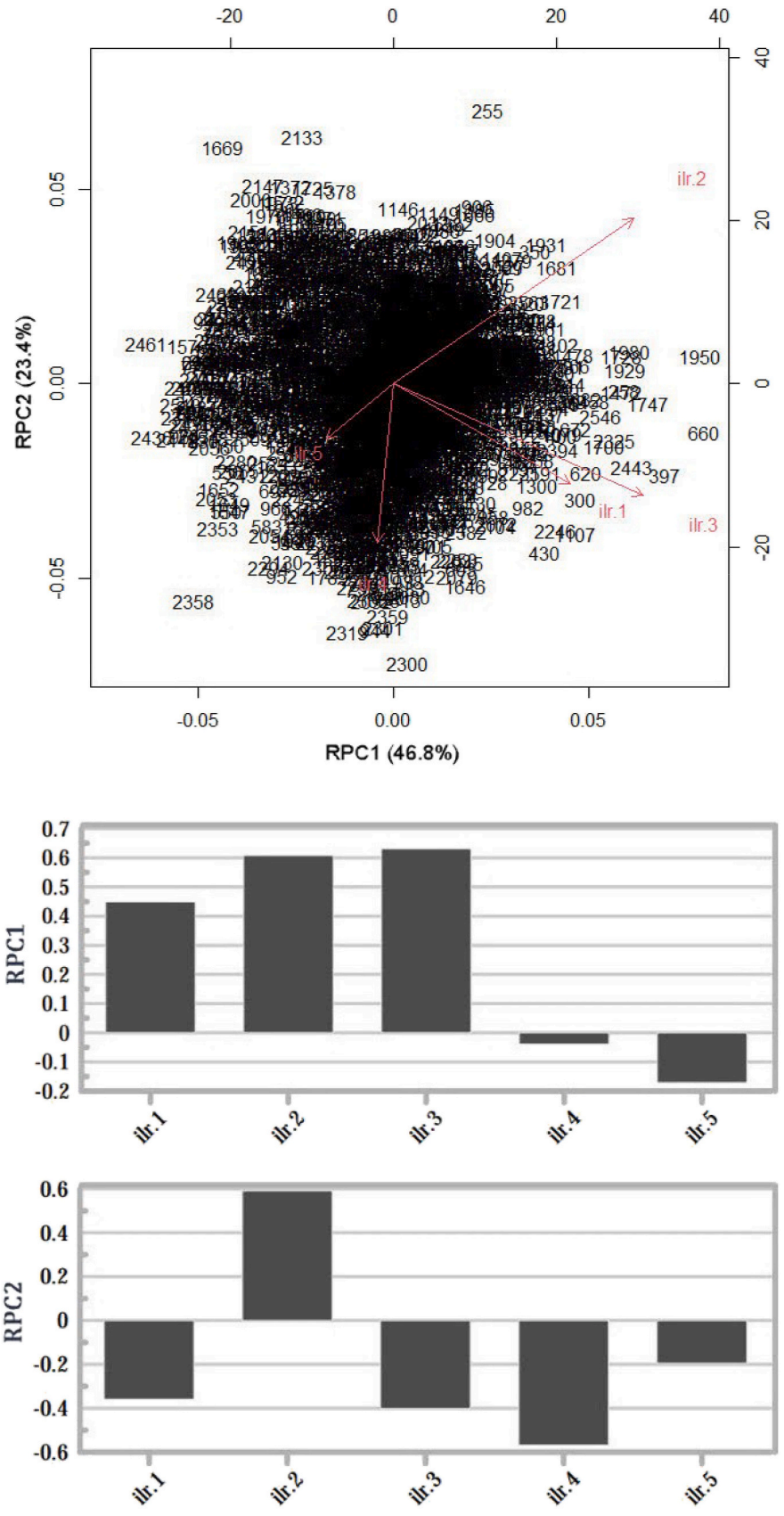

Fig. 4. Covariance biplots and eigenvector plots of RPCA applied to ilr-data (70.2 % explained variance).

Table 6
Thresholds obtained by number-size fractal model applied to PC1 and RPC1.

	PC1	RPC1
Background	<0.09	<0.17
Weakly Anomalous	0.09 – 0.59	0.17 – 0.66
Moderately Anomalous	0.59 – 1.0	0.66 – 1.70
Strongly Anomalous	1.0 – 2.19	1.70 – 4.83
Very Strongly Anomalous	2.19 – 4.88	–

3.4. Principal component analysis (PCA) and its robust version (RPCA)

One of the most significant and commonly used dimension reduction models in multivariate analysis is PCA. Such dimension reduction models provide a simple and easy overview of the whole data set, i.e., if a big number of elements are available to review, a limited number of representative factors are provided by such models (Reimann, 2005). This results in a straightforward statistical analysis. PCA is based on variance-covariance matrices, and it is extremely sensitive to the available data and particularly to extreme values such as outliers. Therefore, their results could be highly biased if the raw data with retained outliers is evaluated (Zuo et al., 2013; Nazarpour et al., 2015; Sadeghi, 2020). To diminish the outliers' side effects on the PCA results, RPCA was developed (Filzmoser et al., 2009). It needs to be applied to ilr values (Egozcue et al., 2003) as the covariance matrix is not necessarily singular. RPCA works based on the minimum covariance determinant estimator (MCD) rather than the classic covariance matrix.

In the following, a number-size (N-S) multifractal model is considered to study the geochemical patterns via modelling PCA and RPCA of clr and ilr transformed data. Data was processed using CoDaPack and ioGAS v.7.4 in addition to an R code developed by Martín-Fernández et al. (2017). which is available in the R-package zCompositions.

4. Discussion and results

Here, to avoid the closure effect of the data, PCA was applied to the clr transformed data, and RPCA was applied to ilr transformed data of Au, Ag, As, Sb, Te and W, as the specific elements in orogenic gold mineralization (Tables 3 and 4; Figs. 2 and 3), with the aim of demonstrating the higher efficiency of the latter one in case a sub-composition (a group of elements, i.e., a commodity among the whole available elements) is studied (Figs. 1 and 3). Fig. 2 shows that the Q-Q plot of PC1-PC2 and RPC1-RPC2 pairs are almost identical. This is a sign of similar distribution of the two pairs. At the first sight, it is expected to see

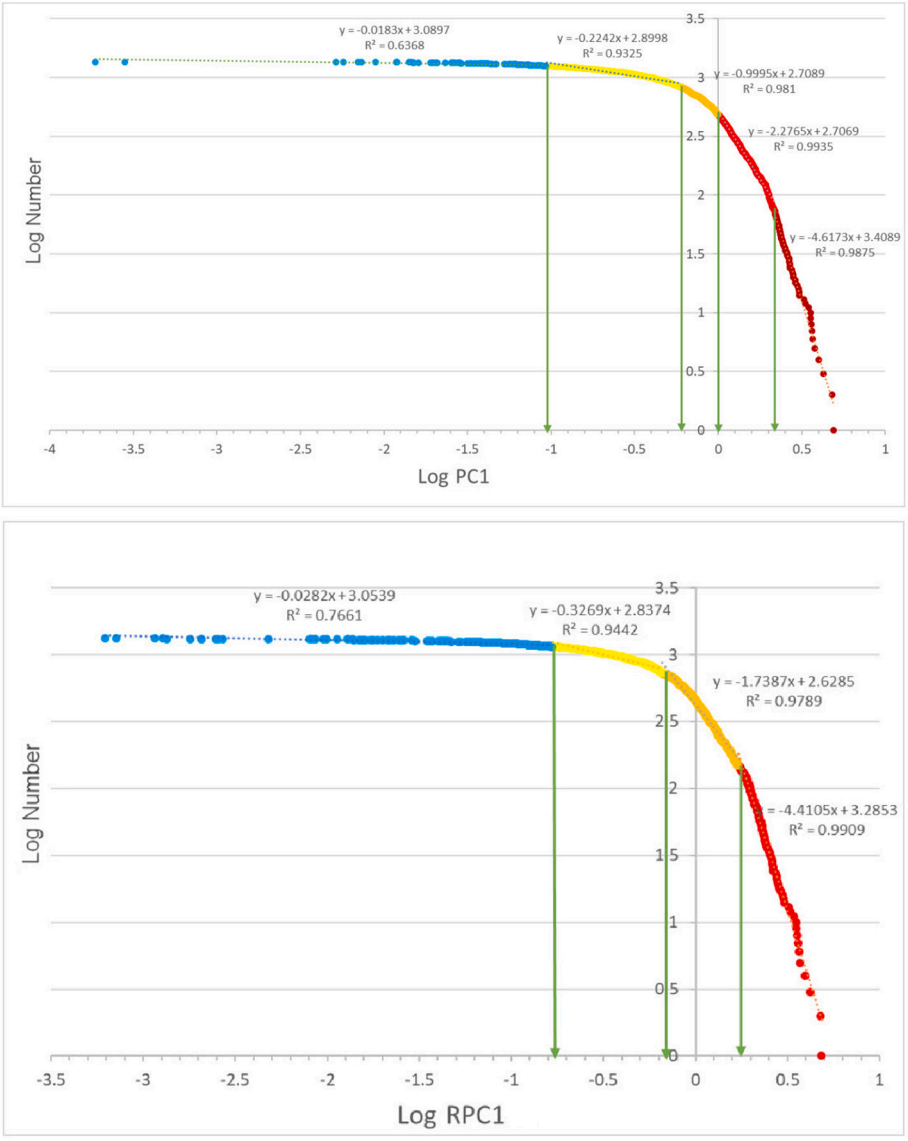

Fig. 5. Number-size fractal modelling applied to a) PC1 and b) RPC1.

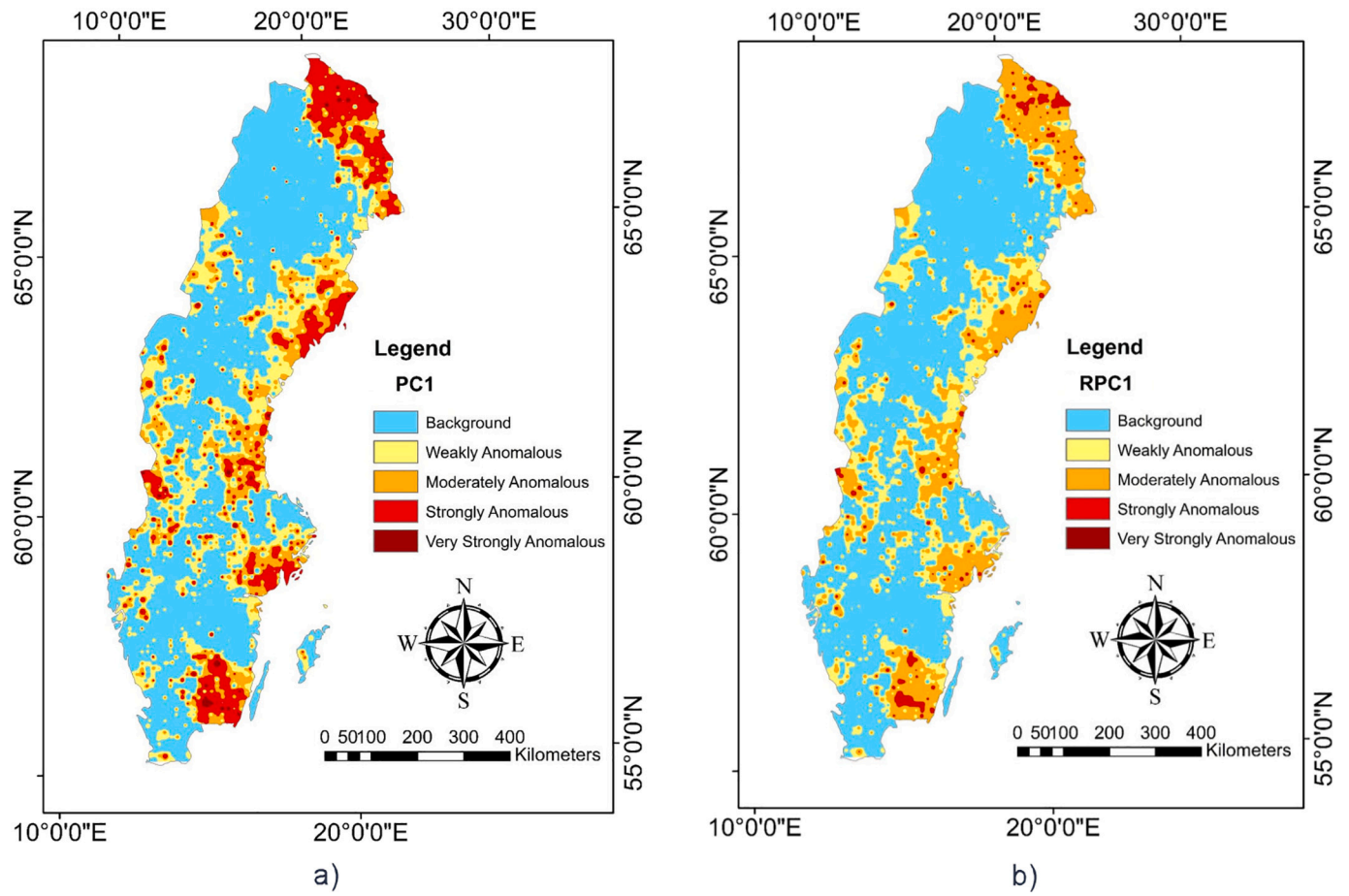

Fig. 6. Number-size fractal model classified maps based on a) PC1 and b) RPC1. The samples are categorized in five and four classes based on the model applied to PC1 and RPC1, respectively. It is notable that RPC1 map shows no very strong anomalous zones while the PC1 map indicates very strong anomalies. Moreover, ilr suggests relatively smaller strong anomalous areas, especially in the north part of the map.

similar maps and outputs derived from PCA and RPCA variables.

The biplots of the first couple of clr PCs (i.e., PC1 with explained variance 46.19 % and PC2 with cumulative explained variance 69.61 %; Table 3) demonstrates positive loading of clr_{Au}, but negative loadings of clr_{Ag}, clr_{As}, clr_{Te}, clr_{Sb} and clr_{W} for PC1, in addition to the positive loadings of clr_{Ag} and clr_{W} and negative loadings of clr_{Au}, clr_{As}, clr_{Sb}, and clr_{Te} for PC2 (Fig. 3).

The biplots of the couple of ilr RPCs (i.e., PC1 with 46.75 % explained variance and PC2, explaining jointly 70.23 % of the variance; Table 4) show that in PC1, ilr1, ilr2 and ilr3 have positive loadings, and ilr4 and ilr5 have negative loadings; also in PC2, only ilr2 is positive and the rest of the ilrs are negative.

Based on Table 5, O1 indicates that the maximum variability in the data corresponds to the relative proportion of Au compared to As, Sb, and W, followed by O4 reflecting the relative proportion of W compared to As and Sb, and similar for the others. Note that O1 and C1 are the same, only the sign is changed, which does not affect the interpretation; same for O4 and C4. O3 and C3 are similar, but the role of Ag and Te is exchanged, same as in C2 and O2. O5 and C5 are again identical except for the sign (Fig. 4).

The clr_{Au} ray separation from the other rays clearly demonstrates the most widespread features which are the Au primary hosts in Sweden. To help with better imagination of the relevant anomalies, the spatial distribution maps were generated based on PC1 and RPC1, and then classified using N-S multifractal model (Tables 5 and 6 and Figs. 5 and 6).

Fig. 6 demonstrates that PCA applied to clr looks similar to RPCA applied to ilr in the definition of the orogenic gold anomalies in Northern Norrbotten and Bergslagen. This is also the same in Skellefte and Jämtland. Moreover, it is evident that no very strong anomalous zone is introduced by the RPCA, while the PCA demonstrated sparse very strong anomalous zones in the south, center, and north of the map. Another visible result is that the RPCA map delivers proportionally smaller anomalous zones in general.

5. Conclusions

Mineral exploration, particularly on a regional scale, involves handling compositional data with a multitude of components. Effectively representing geochemical data necessitates employing dimension reduction techniques, such as PCA, while also translating them into coordinates applicable in real space. In this research, we employed PCA on clr transformed data derived from till geochemical samples throughout Sweden. A challenge arises from the inherent singularity of the covariance matrix in clr-transformed data, rendering robust versions of PCA inapplicable. The reason for the clr not constituting an orthonormal coordinate system is that the sample space of CoDa is D-1 dimensional and there are D clr components; so, there is one more than needed. To address this, an ilr transformation is employed, supplemented by SBP to define balances. The SBP is tailored to incorporate pathfinder elements indicative of orogeny gold. To discriminate the anomalous areas and background, the N-S multifractal model was applied. The results exhibit that while RPCA identifies anomalous zones more accurately, PCA and RPCA converge in highlighting almost similar regions. It can be inferred that for regional geochemical data representation, both PCA and RPCA yield analogous maps. However, their applicability might vary on a local scale or under different sampling

densities. This study underscores the need for additional research to achieve optimal representation of compositional data in scenarios involving local scale analysis or sparse sampling.

CRediT authorship contribution statement

Behnam Sadeghi: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Hossein Molayemat:** Conceptualization, Data curation, Visualization, Writing – review & editing. **Vera Pawlowsky-Glahn:** 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 data that has been used is confidential.

Acknowledgements

The authors express their gratitude to the Geological Survey of Sweden, particularly Dr. Martiya Sadeghi, for providing the valuable data. They also extend their appreciation to J. A. Martn-Fernández for the support offered regarding the PBs. VPG acknowledges the financial support received from the Spanish Ministry of Science and Innovation under the projects "CODA-GENERA" (Ref. PID2021-123833OB-I00) and "CONBACO" (Ref. PID2021-125380OB-I00).

References

Aitchison, J., 1983. Principal component analysis of compositional data. Biometrika 70 (1), 57–65.
Aitchison, J., 1986. The statistical analysis of compositional data. In: Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK) (Reprinted in 2003 with additional material by The Blackburn Press, 416 p.).
Aitchison, J., 1992. On criteria for measures of compositional difference. Math. Geol. 24, 365–379.
Aitchison, J. (2002). A Concise Guide for the Algebraic-Geometric Structure of the Simplex, the Sample Space for Compositional Data Analysis. (unpublished).
Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. J. R. Stat. Soc. Ser. C Appl. Stat. 51, 375–392.
Allen, et al., 1996. Setting of Zn-Cu-Au-Ag massive sulfide deposits in the evolution and facies architecture of a 19 Ga marine volcanic arc, Skellefte district, Sweden. Econ. Geol. 91, 1022–1053.
Andersson, M., Carlsson, M., Ladenberger, A., Morris, G., Sadeghi, M., Uhlback, J., 2014. Geochemical Atlas of Sweden. Geological Survey of Sweden (SGU), Uppsala (SV). https://www.sgu.se/en/mineral-resources/geokemisk-kartlaggning/geochemical-atlas-of-Sweden/.
Barceló-Vidal, C., Martín-Fernández, J.-A., 2016. The mathematics of compositional analysis. Austrian Journal of Statistics 45, 57–71.
Beckholmen, M., Tirén, S., 2009. The geological history of the Baltic Sea: a review of the literature and investigation tools. In: Swed. Rad. Saf Auth. Rpt. 40.
Buccianti, A., Grunsky, E., 2014. Compositional data analysis in geochemistry: are we sure to see what really occurs during natural processes? J. Geochem. Explor. 141, 1–5.
Carranza, E., 2011. Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. J. Geochem. Explor. 110, 167–185.
Cheng, Q., 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol. Rev. 32, 314–324.
Cheng, Q., Agterberg, F., 2021. Fractal geometry in geosciences. In: Daya Sagar, B., Cheng, Q., McKinley, J., Agterberg, F. (Eds.), Encyclopedia of Mathematical Geosciences (2021).
Cheng, et al., 1994. The separation of geochemical anomalies from background by fractal methods. J. Geochem. Explor. 51, 109–130.
Cheng, Q., Xu, Y., Grunsky, E., 1999. Integrated spatial and spectral analysis for geochemical anomaly separation. In: Proceedings of IAMG'1999, Trondheim, Norway.
Chork, C., Salminen, R., 1993. Interpreting exploration geochemical data from Outokumpu, Finland: a MVE-robust factor analysis. J. Geochem. Explor. 48, 1–20.
Clare, A., Cohen, D., 2001. An unsupervised neural network approach to the analysis of multi-element stream sediment data, NE NSW, Australia. Geochemistry: Exploration. Environment, Analysis 1, 119–134.
Daya Sagar, B., Cheng, Q., Agterberg, F., 2018. Handbook of Mathematical Geosciences. Springer (914 p.).
Encyclopedia of mathematical geosciences. In: Daya Sagar, B., Cheng, Q., McKinley, J., Agterberg, F. (Eds.), 2023. Encyclopedia of Earth Sciences Series. Springer, Cham. https://doi.org/10.1007/978-3-030-26050-7-334-1.
Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37 (7), 795–828.
Egozcue, J., Pawlowsky-Glahn, V., 2006. Exploring compositional data with the CoDa-Dendrogram. In: Pirard, E., Dassargues, A., Havenith, H.B. (Eds.), Proceedings of IAMG'2006, Liège (B). University of Liège, Belgium, CD-ROM.
Egozcue, Pawlowsky-Glahn, 2019a. Compositional data: the sample space and its structure. TEST 28 (3), 599–638.
Egozcue, Pawlowsky-Glahn, 2019b. Rejoinder on: compositional data: the sample space and its structure. TEST 28 (3), 658–663.
Egozcue, J., Pawlowsky-Glahn, V., 2023. Subcompositional coherence and a novel proportionality index of parts. SORT 47 (2), 229–244 (DOI: 10.57645/20.8080.02.7).
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.
Egozcue, J., Pawlowsky-Glahn, V., Cheng, Q., 2015. Compositional approach to mineral singularity assessment. In: Proceedings of IAMG'2015.
Egozcue, J.J., Pawlowsky-Glahn, V., Molayemat, H., 2019. Correlations, subcompositional coherence, and misinterpretations. In: Proceedings of IAMG'2019, The 20th Annual Conference of the International Association for Mathematical Geosciences.
Eilu, P., 2012. Mineral deposits and metallogeny of Fennoscandia. In: Geological Survey Finland Special Paper 53.
Filzmoser, P., Hron, K., Reimann, C., Garrett, R., 2009. Robust factor analysis for compositional data. Comput. Geosci. 35, 1854–1861.
Foody, G., 1997. Fully fuzzy supervised classification of land cover from remotely sensed imagery with an artificial neural network. Neural Comput. & Applic. 5, 238–247.
Gallo, M., Buccianti, A., 2013. Weighted principal component analysis for compositional data: application example for the water chemistry of the Arno river (Tuscany, Central Italy). Environmetrics 24, 269–277.
Greenacre, M., 2010. Logratio analysis is a limiting case of correspondence analysis. Math. Geosci. 42, 129–134.
Greenacre, M., 2018a. Compositional Data Analysis in Practice. Chapman and Hall/CRC (136 p.).
Greenacre, M., 2018b. Variable selection in compositional data analysis, using pairwise log-ratios. Math. Geosci. 51, 649–682.
Grunsky, E., de Caritat, P., 2020. State-of-the-art analysis of geochemical data for mineral exploration. Geochem.: Explor., Environ., Anal. 20 (2), 217–232.
Grunsky, E., Kjarsgaard, B., 2016. Recognizing and validating structural processes in geochemical data. In: Martn-Fernández, J.A., Thió-Henestrosa, S. (Eds.), Compositional Data Analysis, Springer Proceedings in Mathematics and Statistics, vol. 187, pp. 85–116.
Grunsky, E., Mueller, U., Corrigan, D., 2014. A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: applications for predictive geological mapping. J. Geochem. Explor. 141, 15–41.
Hallberg, A., Bergman, T., Gonzales, J., Larsson, D., Morris, G., Perdahl, J., Ripa, M., Niiranen, T., Eilu, P., 2012. Metallogenic areas in Sweden. In: Geological Survey Finland Special Paper 53, pp. 139–206.
Jolliffe, I., 2002. Principal component analysis. In: Springer Series in Statistics, 2nd ed. Springer, New York, NY.
Jolliffe, I., Trendafilov, N., Uddin, M., 2003. A modified principal component technique based on the lasso. J. Comput. Graph. Stat. 12, 531–547.
Kaski, S., 1997. Data exploration using self-organizing maps. Acta Polytech. Scand. Mathematics, Computing and Management in Engineering 82, 57.
Kohonen, T., 1995. Self-organizing Maps. Springer, Berlin, Heidelberg.
Kynčlová, P., Hron, K., Filzmoser, P., 2017. Correlation between compositional parts based on symmetric balances. Math. Geosci. 49, 777–796.
Makvandi, S., Ghasemzadeh-Barvarz, M., Beaudoina, G., Grunsky, E., McClenaghan, M., Duchesne, C., 2016. Principal component analysis of magnetite composition from volcanogenic massive sulfide deposits: case studies from the Izok Lake (Nunavut, Canada) and Halfmile Lake (New Brunswick, Canada) deposits. Ore Geol. Rev. 75, 60–85.
Mandelbrot, B., 1983. The Fractal Geometry of Nature, 2nd edn. W. H. Freeman and Company, San Francisco.
Martín-Fernández, J., Pawlowsky-Glahn, V., Egozcue, J., Tolosona-Delgado, R., 2017. Advances in principal balances for compositional data. Math. Geosci. 50, 273–298.
Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J., 2012. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. Computational Statistics and Data Analysis 56, 2688–2704.
McKinley, J., Grunsky, E., Mueller, U., 2018. Environmental monitoring and peat assessment using multivariate analysis of regional-scale geochemical data. Math. Geosci. 50, 235–246.
McMartin, I., McClenaghan, M., 2001. Till geochemistry and sampling techniques in glaciated shield terrain: a review. Geol. Soc. Lond. Spec. Publ. 185, 19–43.

Molayemat, H., Torab, F.M., Pawlowsky-Glahn, V., Hossein Morshedy, A., Egozcue, J.J., 2018. The impact of the compositional nature of data on coal reserve evaluation, a case study in Parvadeh IV coal deposit, Central Iran. Int. J. Coal Geol. 188, 94–111.
Molayemat, H., Torab, F.M., Pawlowsky-Glahn, V., Egozcue, J.J., Hossein-Morshedy, A., 2022. A compositional approach to in-situ evaluation of polymetallic deposits. A case study at Sungun Cu-Mo deposit, NW Iran. J. Geochem. Explor. 237, 106981.
Mueller, U., Grunsky, E., 2016. Multivariate spatial analysis of lake sediment geochemical data; Melville Peninsula, Nunavut, Canada. Appl. Geochem. 75, 247–262.
Nazarpour, A., Rashidnejad Omran, N., Rostami Paydar, G., Sadeghi, B., Matroud, F., Mehrabi Nejad, A., 2015. Application of classical statistics, logratio transformation and multifractal approaches to delineate geochemical anomalies in the Zarshuran gold district. NW Iran. Chemie der Erde/Geochemistry 75 (1), 117–132.
Olea, R.A., Luppens, J.A., 2015. Mapping of coal quality using stochastic simulation and isometric logratio transformation with an application to a Texas lignite. Int. J. Coal Geol. 152, 80–93.
Palarea-Albaladejo, J., Antoni Martn-Fernández, J., Ruiz-Gazen, A., Thomas-Agnan, C., 2022. lrsvd: an efficient imputation algorithm for incomplete high-throughput compositional data. J. Chemom. 36 (12), 34–59.
Pawlowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15 (5), 384–398.
Pawlowsky-Glahn, Egozcue, 2006. Spatial analysis of compositional data: a historical review. J. Geochem. Explor. 164, 28–32.
Pawlowsky-Glahn, V., Egozcue, J., 2011. Exploring compositional data with the Coda-Dendrogram. Austrian Journal of Statistics 40 (1 & 2), 103–113.
Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2011a. Principal balances. In: Proceedings of the 4th International Workshop on Compositional Data Analysis.
Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2011b. Principal balances to analyse the geochemistry of sediments. In: Proceedings of IAMG'2011, pp. 878–887.
Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and analysis of compositional data. In: Statistics in Practice. John Wiley & Sons, Chichester UK (272 pp.).
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. Proc. R. Soc. Lond. LX, 489–502.
Pospiech, et al., 2020. Discriminant analysis for compositional data incorporating cell-wise uncertainties. Math. Geosci. 53, 1–20.
Queysanne, M., 1973. Álgebra Básica. Editorial Vicens Vives, Barcelona (E) (669 p.).
Reimann, C., 2005. Geochemical mapping: technique or art? Geochemistry: Exploration. Environment, Analysis 5, 359–370.
Sadeghi, B., 2020. Quantification of Uncertainty in Geochemical Anomalies in Mineral Exploration. University of New South Wales (Ph. D. thesis).
Sadeghi, B., 2021a. Concentration-concentration fractal modelling: a novel insight for correlation between variables in response to changes in the underlying controlling geological-geochemical processes. Ore Geol. Rev. 128.
Sadeghi, B., 2021b. Simulated-multifractal models: a futuristic review of multifractal modeling in geochemical anomaly classification. Ore Geol. Rev. 128.
Sadeghi, B., 2024. Fractals and Multifractals in the Geosciences. Elsevier, 250 P.
Sadeghi, B., 2021c. Spectrum-area method. In: See Encyclopedia. https://doi.org/10.1007/978-3-030-26050-7-334-1.
Sadeghi, B., Agterberg, F., 2021. Singularity analysis. In: See Encyclopedia. https://doi.org/10.1007/978-3-030-26050-7-334-1.
Sadeghi, B., Cohen, D., 2021a. Category-based fractal modelling: a novel model to integrate the geology into the data for more effective processing and interpretation. J. Geochem. Explor. 226.
Sadeghi, B., Cohen, D., 2021b. Concentration-distance from centroids (c-dc) multifractal modeling: a novel approach to characterizing geochemical patterns based on sample distance from mineralization. Ore Geol. Rev. 137.
Sadeghi, B., Moarefvand, P., Afzal, P., Yasrebi, A., Daneshvar Saein, L., 2012. Application of fractal models to outline mineralized zones in the Zaghia iron ore deposit, Central Iran. J. Geochem. Explor. 122, 9–19.
Sadeghi, B., Madani, N., Carranza, E., 2015. Combination of geostatistical simulation and fractal modelling for mineral resource classification. J. Geochem. Explor. 149, 59–73.
Sadeghi, B., Grunsky, E., Pawlowsky-Glahn, V., 2021. Uncertainty quantification. In: See Encyclopedia. https://doi.org/10.1007/978-3-030-26050-7-334-1.
Sauro Graziano, R., Gozzi, C., Buccianti, A., 2020. Is compositional data analysis (coda) a theory able to discover complex dynamics in aqueous geochemical systems? J. Geochem. Explor. 211, 106465.
Stefelova, N., Palarea-Albaladejo, J., Hron, K., 2021. Weighted pivot coordinates for partial least squares-based marker discovery in high-throughput compositional data. Statistical Analysis and Data Mining: The ASA Data Science Journal 14 (4), 315–330.
Zuo, R., Wang, J., 2016. Fractal/multifractal modeling of geochemical data: a review. J. Geochem. Explor. 164, 33–41.
Zuo, R., Xia, Q., Zhang, D., 2013. A comparison study of the C–A and S–A models with singularity analysis to identify geochemical anomalies in covered areas. Appl. Geochem. 33, 165–172.
Zuzolo, et al., 2018. Exploring uni-element geochemical data under a compositional perspective. Appl. Geochem. 91, 174–184.