Earth-Science Reviews 277 (2026) 105437

Contents lists available at ScienceDirect

Earth-Science Reviews

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

Review Article

Benchmark-based metric distances as predictors of geochemical processes and data variability: A comprehensive discussion

Caterina Gozzi *, Antonella Buccianti

University of Florence, Dept. of Earth Sciences, Via G. La Pira 4, 50121 Firenze, Italy

ARTICLE INFO

Keywords:
Compositional data analysis
Geochemical variability
System resilience
Tipping points
Mahalanobis distance

ABSTRACT

Understanding compositional changes is the key to investigate how complex natural systems evolve. Traditional geochemical approaches, such as univariate, bivariate, or multivariate analyses, usually focus on variability around a central composition. However, they do not provide insights into evolutionary paths or changes from benchmark states. In this work, we propose a comprehensive conceptual and methodological discussion of a holistic approach that addresses these limitations. We combine compositional data with metrics tailored to the *simplex* structure, specifically the robust Mahalanobis distance (*MD*) and perturbation analysis. Two large-scale datasets were analyzed: the Abyssal Volcanic Glasses Database (AVGD) with primitive compositions (chondrites and MORBs) as benchmarks, and the FOREGS repository of European stream and alluvial sediments, with crustal averages and diluted waters as reference states. By calculating *MD* from benchmark compositions and examining the frequency distribution, skewness, kurtosis, and multimodality, we identify patterns of compositional change, resilience, and potential instability. This work provides a data-driven perspective on how compositional metrics can enhance the understanding of Earth system dynamics. In particular, the proposed framework enables the detection of early-warning signals and tipping points. This offers a powerful tool for multi-scale assessment of geochemical system dynamics and resilience.

Contents

* Corresponding author.
E-mail address: caterina.gozzi@unifi.it (C. Gozzi).

https://doi.org/10.1016/j.earscirev.2026.105437
Received 14 November 2025; Received in revised form 11 February 2026; Accepted 16 February 2026
Available online 6 March 2026

1. Introduction

Understanding the nature of compositional changes allow us to deeply investigate how complex natural systems work. Examining patterns of variation of geochemical data alongside the evolution of natural processes is a fundamental approach when appropriate methods are used. Generally, scientists explore databases using univariate, bivariate and multivariate analysis tools. In all cases, the variance and/or the variance-covariance matrix is crucial for understanding how the data move around their center of gravity. However, this type of analysis provide no information about evolutionary paths. Its main goal is to identify a barycentre and to understand how the variability patterns move around it. In geochemistry, attempts to introduce analyses that provide insights into evolutionary paths are particularly found in the use of binary diagrams. In igneous petrogenesis, bivariate diagrams are a simple way to visualize chemical differentiation resulting from crystal-liquid fractionation processes (e.g., partial melting or fractional crystallization). They can be constructed in terms of major elements, trace elements or a combination of both. Harker diagrams, for example, plot the weight percent of a constituent oxide against wt % SiO_2 on x-axis. They are widely used in scientific literature, notwithstanding the numerical constrain inherent in compositional data (Chayes, 1960; Aitchison, 1982; Cortes, 2009; Egozcue, 2009).

On the other hand, patterns of chemical variation can also involve trace elements, as for example Rare Earth Elements (REEs). In this case, to eliminate the Oddo-Harkins effect, which reflects the higher concentrations elements with even atomic numbers as compared to those with odd atomic numbers, a normalization is required. Generally, this normalization is performed by comparing the concentration of individual REEs in a rock to their abundance in chondritic meteorites, the latter being considered representative of primitive solar system material. Alternatively, Mid-Ocean Ridge Basalts (MORBs) can be used as a potential benchmark of relatively large degrees of partial melting. Following this approach, the concept of spider diagrams has been developed, in which the abundances of a range of incompatible trace elements are normalized to those of the primordial Earth or MORBs. There are several variants of the spider diagrams which order the elements on the x-axis according to decreasing incompatibility or mobility, or from light-REEs to heavy REEs. For example, when less incompatible elements, which are less enriched during partial melting, are located on the right-hand side, a typical curve tilting upward to the left is generated. Conversely, Oceanic Island Basalts (OIBs) show extreme presence of incompatible elements with a peak at $Nb - Ta$, suggesting the enrichment of the source in these elements.

Molar ratio diagrams represent another way to investigate the behavior of chemical elements or species in bio-geochemical and/or mixing processes (Gaillardet et al., 1999). Meanwhile, the analysis of heavy metal patterns in stream and alluvial sediments is often used to link sources to sinks. It is also useful to describe the effects of denudation, sediment transport, deposition and pollution, as well as weathering processes affecting the composition of natural waters (Pandit et al., 2023).

All the previously described methods have some critical limitations related to: (1) the use of relative data affected by a numerical constraint and pertaining to the *simplex* sample space S^D, where the Euclidean geometry is not applicable; and (2) the partial analysis (e.g., binary diagrams with few components) of the relationships between the parts of a composition, when all of them are deeply interdependent (Aitchison, 1986; Gozzi et al., 2021; Gozzi and Buccianti, 2022). A possible solution to the first issue is the use of the log-ratio transformation of data, which allows moving them from the simple S^D to the real space R (Aitchison, 1986; Egozcue and Pawlowsky-Glahn, 2005). Then, the application of multivariate methodologies capable of investigating the variance-covariance structure of the whole log-transformed dataset could help to overcome the second issue. However, this approach does not focus on how the compositional changes manifest their features and properties. In fact, most of multivariate procedures analyze how variability evolves around the data barycenter, without explaining how this state is achieved starting from some pristine (or benchmark) composition.

Consequently, in this work, intended as a conceptual–methodological review with extensive case studies, we propose and discuss a new approach. The method addresses the two previously mentioned issues while maintaining the focus on compositional change. To obtain the target compositional changes measured from benchmark compositions, an appropriate metric was used, considering the nature of compositional data and the need for a holistic (multivariate) approach (Aitchison, 1982; van Rooij et al., 2013). Two different databases were used to demonstrate the benefits of the proposed method:(1) the Abyssal Volcanic Glasses Database (AVGD, Jenner and O'Neill (2012)) with benchmark compositions represented by primitive material (chondrite) and MORBs (N, D and E types), respectively (Wilson, 1989; Rollinson, 2007; Gale et al., 2013) and (2) the repository of the Forum of European Geological Surveys (FOREGS, `http://www.gtk.fi/publ/foregsatlas/`) including European alluvial and stream sediments which were compared to the average composition of the Earth's crust (Faure, 1998) as well as stream waters for which a highly diluted water was chosen as benchmark on a continental scale.

The focus of this work is initially based on the calculation of compositional distances from the selected benchmarks using metrics such as the robust Mahalanobis distance (*MD*) (Pawlowsky-Glahn and Buccianti, 2011). It is computed starting from a log-transformed compositional matrix, thereby avoiding the fallacy associated with the numerical constrain (Aitchison, 1986). Subsequently, the shape of the frequency distribution of *MD* was explored in terms of presence of skewness, kurtosis and multimodality aiming to understand the nature of the variability of the compositional changes. In this perspective, *MD* is considered a probe to achieve this purpose. The results were then compared with those obtained by the investigation of the perturbations matrices calculated from the same benchmark. The perturbation is an operation defined in the *simplex* S^D sample space to measure differences between compositions with D components (variables) (Billheimer et al., 2001).

2. Materials and methods

2.1. Overview and structure of the databases

2.1.1. AVGD databases

The last published version of the AVGD file (Melson et al., 2002) contains 9050 analyses from 664 Atlantic, 28 Caribbean, 89 Indian, and 1304 Pacific localities. Precise trace element analyses were obtained by Laser Ablation (LA)-ICP-MS and Electron Microprobe (EMPA) for REEs, High Field Strength Elements (HFSEs), Large Ion Lithophile Elements (LILEs), First Row Transition Elements (FRTEs) and less commonly analyzed elements such as Li, Be, Cu, Zn, Ga, Ge, As, Se, Mo, Ag, Cd, In, Sn, Sb, W, Tl and Bi. The volcanic glass samples, donated by the Smithsonian Institution, encompasses a broad range of major geographical regions, each characterized by varying spreading rates and local tectonic environments. Only few samples were from the Southwest and Southeast Indian Ridges and the Pacific-Antartic-Ridge. The samples cover a range of major element chemistry from primitive to moderately evolved stages (9.7 to ~ 3.5 wt.% MgO; Jenner and O'Neill (2012)).

The samples originate from a wide range of tectonic settings, classified in the AVGD as: spreading ridges (497), seamounts (37), fracture zones (71), aseismic ridge (4) and backarc basins (7). In some cases, samples classified as spreading centers correspond to ridges associated with hot spot tracks.

To examine compositional variations, specific sub-sets of trace elements, typically employed in spider diagrams, were selected from the AVGD database and grouped as follows: 1) *La*, *Ce*, *Nd*, *Sm*, *Eu*, *Gd*, *Dy*, *Yb*; 2) *Ba*, *Rb*, *Th*, *K*, *Nb*, *Ta*, *La*, *Ce*, *Sr*, *Nd*, *P*, *Sm*, *Zr*, *Hf*, *Ti*, *Tb*, *Y*, *Tm*, *Yb* and 3) *Sr*, *K*, *Rb*, *Ba*, *Th*, *Ta*, *Nb*, *Ce*, *P*, *Zr*, *Hf*, *Sm*, *Ti*, *Y*, *Yb*. The first two groups have been compared with chondrite composition and the third one with the MORBs composition. As reported in Wilson (1989), the order of the elements reflects the progressive decrease in incompatibility and mobility, respectively. For MORBs a more detailed comparison was performed using the average compositions of N-MORB, D-MORB and *E*-MORB as benchmarks, as reported in Gale et al. (2013).

2.1.2. FOREGS databases

The FOREGS repository was built up to provide high-quality, multipurpose environmental geochemical baseline data for Europe. The need for this type of data was recognized by the first Working Group on Regional Geochemical Mapping immediately after the Chernobyl accident in 1986. The European contribution to the program was carried out by government institutions from 26 countries under the auspices of the former Forum of European Geological Surveys. This activity has now been transferred to EuroGeoSurveys, the Association of the European Geological Surveys in which FOREGS has been merged. The main objectives of this European survey were: (1) to apply standardized methods of sampling, chemical analysis and data management to establish a geochemical baseline across Europe; and (2) to use this reference network to guide national baseline datasets.

As an application example, we consider the composition of stream and alluvial plain sediments for a limited set of heavy metals (i.e., *Cr*, *Ni*, *Cu*, *Zn*, *Pb*, *Co*, *V*,) and investigate compositional changes relative to the average Earth crust composition. The accumulation of metals in these matrices originates from natural sources, primarily through rock weathering but human activities may also contribute to modify their concentrations (Pandit et al., 2023).

Surface waters reflect the interplay between the geosphere, the hydrosphere, and pollution. Stream waters of the FOREGS repository were collected from small, second-order drainage basins (< 100 km^2) at the same site as the active stream sediments. In arid regions, such as Southern Europe, streams remain dry for most of the year. Hence, the sampling, whenever possible, was carried out during the winter and early spring months. In this analysis, only the composition of major anions (i.e., Cl^-, HCO_3^-, SO_4^{2-}) and cations (i.e., Ca^{2+}, K^+, Mg^{2+}, Na^+) was considered and compared with the most diluted water as benchmark.

Information on sampling and sieved procedures, as well as analytical laboratory methods can be found both on the FOREGS website (http://weppi.gtk.fi/publ/foregsatlas/) in the section dedicated to sample preparation and analysis. This includes details on accuracy, precision and detection limits for all sample types.

2.2. CoDA concepts and background

Compositional Data Analysis (CoDA) methods have progressively gained attention across a wide range of scientific disciplines, as the limitations of conventional statistical techniques for analyzing relative data have become increasingly evident. Initially developed within the field of geochemistry, CoDA approaches are now widely applied in numerous research areas, including geology, biometrics, cultural heritage, ecology, economics, medicine, and the social sciences, with an exponential growth in applications observed in recent years. According to a recent bibliometric review (Navarro-Lopez et al., 2022), journals publishing CoDA-related studies are primarily classified within the Geosciences and Multidisciplinary categories (23 papers), followed by Ecology (19) and Geochemistry and Geophysics (16).

Compositional data describe the *D* parts of a whole (the composition) and are commonly represented as vectors of proportions, concentrations, or percentages (Aitchison, 1982). Such data are inherently relative, meaning that their values can only be interpreted in relation to the total composition (Aitchison, 1986). Individual components cannot vary independently, as they are constrained by a constant-sum condition. Consequently, an increase in one part necessarily results in a decrease in the relative contributions of the remaining parts. These fundamental properties, extensively discussed in the literature (e.g, Pawlowsky-Glahn and Buccianti, 2011; Egozcue and Pawlowsky-Glahn, 2005), have important implications for statistical analysis.

Standard Euclidean-based statistical methods are generally inappropriate for compositional data. Such data do not belong to real Euclidean space but to a distinct constrained sample space known as the *simplex*, which is characterized by its own geometry (Aitchison, 1986). Ignoring this constraint and applying conventional methods may lead to misleading or spurious results. To address this issue, compositional data are commonly transformed into dimensionless log-ratios of their components. Transformations such as the centered log-ratio (clr) and the isometric log-ratio (ilr) allow compositional data to be analyzed using classical Euclidean statistical techniques (Egozcue et al., 2003). Alternatively, analyses can be performed directly within the *simplex* by applying operations and distance measures defined under Aitchison geometry, such as perturbation and powering.

A typical example concerns sediment geochemistry, where compositional data commonly describe the relative abundances of major elements, trace metals, or grain-size fractions in a sediment sample. For example, the chemical composition of sediments may be expressed as percentages of SiO_2, Al_2O_3, Fe_2O_3, CaO, and MgO, which always sum to 100 %. Because of this constraint, an increase in one component (e.g. Fe_2O_3) necessarily leads to a decrease in others, even in the absence of any direct geochemical relationship. By focusing on log-ratios between components, CoDA overcomes this constant-sum effect, enabling geochemists to correctly interpret sediment provenance, weathering processes, and contaminant enrichment without generating spurious correlations (Chayes, 1960).

2.3. The distance concept and the compositional question

A metric, or distance function, is a function $d(x, y)$ that defines the distance between elements of a set as a non-negative real number. If the distance is zero, the two elements are considered equivalent under that specific metric. From this perspective, the distance functions provide a way to quantify how close two elements, vectors, matrices or arbitrary objects are. Distance metrics are a key component of many machine learning algorithms, used in both supervised and unsupervised learning, generally to calculate the similarity between data points (Subramanian et al., 2022). From a statistical point of view, the definition of a distance metric is fundamental for defining probability laws. When the *R* domain is assumed as the sample space, the probability laws on the real line are defined through their density functions with respect to the Lebesgue measure, which serves as a reference to define the metric and, consequently, a distance. The Lebesgue measure is frequently used in probability theory, in physics and other research fields and it is of great importance in applications on R^N. It constitutes a generalization of length in R, area in R^2 and volume in R^3, thereby providing the basis for measure theory in the Euclidean space (Burk, 1998).

However, compositional data do not lie in an Euclidean space, as anticipated in Section 2.2, and the question of distance (similarity/dissimilarity) between points, vectors, matrices should to be addressed carefully to avoid misunderstanding. Mathematically, a *D*-part composition is a vector of *D* strictly positive real components,

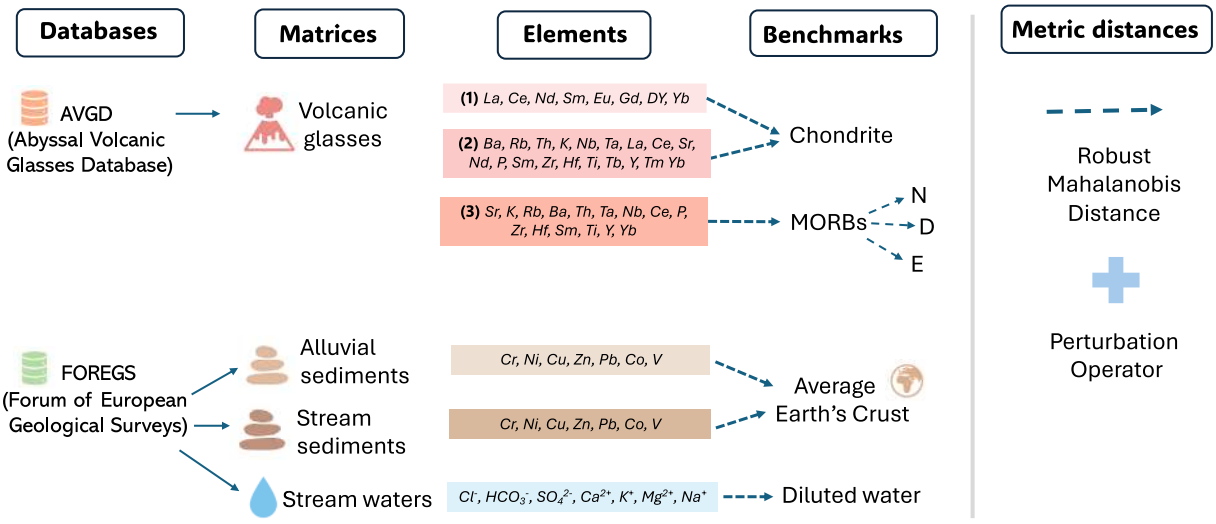

Fig. 1. Workflow of the study describing the databases, the environmental matrices, the adopted benchmarks, and the applied metric distances.

$[x_1 + x_2 + \cdots + x_D]$, such that $x_1 + x_2 + \cdots + x_D = \kappa > 0$. The exact value of κ is not important since the key point is that the parts (variables) of a composition carry only a relative information. Due to this constrained geometry, the appropriate sample space for compositional data is recognized to be the *simplex* S^D (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001), where the rules of the Euclidean geometry cannot be applied without misrepresentations (Egozcue and Pawlowsky-Glahn, 2005; Mateu-Figueras et al., 2011). Consequently, in the *simplex* geometry, tracing compositional changes (i.e., calculating differences or distances between compositions) requires the use of adequate tools to properly capture the behavior of geochemical systems (Gozzi and Buccianti, 2022).

In this framework, one of the available tools is the perturbation, the group operation of the *simplex* S^D that counterparts translation in real space (classical Euclidean statistics) and rotation on the sphere (circular statistics). It provides information on the relative contributions of the variables in a composition during the change processes. Considering two compositions with D-parts $\mathbf{x}, \mathbf{y} \in S^D$, and denoting by $\oplus$ the perturbation operator, $\mathbf{x} \oplus \mathbf{y}$ it is defined by:

$$\mathbf{x} \oplus \mathbf{y} = \frac{[x_1 y_1, \ldots, x_D y_D]}{(x_1 y_1 + \ldots + x_D y_D)} = C[x_1 y_1, \ldots, x_D y_D] \tag{1}$$

where the closure operator C standardizes the perturbation vector by the sum of its components. Following Aitchison and Ng (2005), the operation $\oplus$ defines an Abelian group on the *simplex*, with identity $e = (1/D)[1, \ldots, 1]$ and inverse $x^{-1} = C[x_1, \ldots, x_D]$. This mathematical property is also fundamental for defining the operation that transforms a D-part composition $\mathbf{x}$ into another D-part composition $\mathbf{y}$, thereby providing a measure of their difference (Pawlowsky-Glahn et al., 2015):

$$\mathbf{p} = \mathbf{y} \ominus \mathbf{x} = C\left[\frac{y_1}{x_1}, \ldots, \frac{y_D}{x_D}\right]. \tag{2}$$

In this work, the perturbation vector for each sample was calculated from the same pristine benchmark composition $\mathbf{x_0}$ given, by the chrondrite, MORBs (N, D, E), average Earth's crust or diluted water composition, depending on the data set. The perturbation, together with the powering operation (analogous to scalar multiplication in real space), both defined in S^D, satisfy the requirements for vector space operations, along with the definition of an inner product, a norm and a distance (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001). After calculating the perturbation matrix, the role of each element or chemical species in the compositional changes can be analyzed, also exploring the connections with insights obtained from other metrics.

Another important tool to investigate compositional changes in the *simplex* sample space is the Aitchison distance, which represents a metric to quantify the difference between compositions. It is defined as:

$$d_a(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{D} \left[log \frac{x_i}{g_m(\mathbf{x})} - log \frac{y_i}{g_m(\mathbf{y})} \right]^2 \right)^{1/2} \tag{3}$$

where $g_m(\cdot)$ is the geometric mean of the components of the composition and $\mathbf{x}$ and $\mathbf{y}$ the two compositions to be compared.

A preferred alternative is the calculation of the robust Mahalanobis distance after having transformed the data in isometric log-ratio (ilr) coordinates (Mateu-Figueras et al., 2011). For a location estimator $\mathbf{t}$ and a covariance estimator $\mathbf{C}$, the squared Mahalanobis distances between the observations expressed in ilr coordinates and the respective location estimator $\mathbf{t}$ are given by (Filzmoser and Hron, 2008):

$$\mathrm{MD}(\mathbf{z}_i)^2 = (\mathbf{z}_i - \mathbf{t})' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{t}), \quad \text{for } i = 1, \ldots, n, \tag{4}$$

In the presence of outliers, the classical arithmetic mean (vector) and the (Pearson) sample covariance matrix are unsuitable as estimators $\mathbf{t}$ and $\mathbf{C}$ in Eq. (4) and should be replaced by robust estimators, like the Minimum Covariance Determinant (MCD) estimators $\mathbf{t}_{\mathrm{MCD}}$ and $\mathbf{C}_{\mathrm{MCD}}$ or the MM-estimator (Modified M-estimator) $\mathbf{t}_{\mathrm{MM}}$ and $\mathbf{C}_{\mathrm{MM}}$ (M-estimators are a broad class of extremum estimators for which the objective function is a sample average; M is for maximum, Maronna et al. (2006)).

The proposed methods, Aitchison or Mahalanobis distances (metrics) and perturbation, are fundamentally different. The metrics provide, for each composition, a single positive number that measure the distance from the benchmark. For geo-referenced data, the distance from a compositional benchmark can also be mapped to discover areas experiencing a higher rate of change or exhibiting a more resilient behavior (Gozzi et al., 2020; Sauro Graziano et al., 2020). In contrast, perturbation data are represented as vectors with D components, if D is the number of the analyzed variables. This allows the relative contribution of each single variable to the compositional changes to be revealed. From this perspective perturbation can be visualized in three forms, as a composition expressed in percentage, as multiplicative (non-closed) factors, and as percentage increase/decrease, which is a traditional way of presenting perturbations.

For clarity, the workflow of the study is outlined in Fig. 1.

The calculations proposed in this work are not complex, and users with some practice in R (R Development Core Team, 2023) or MATLAB (The MathWorks, Inc, 2023) software can easily create appropriate scripts. For less experienced users, it is possible to use the Codapack package, a free, open-source, stand-alone software tool

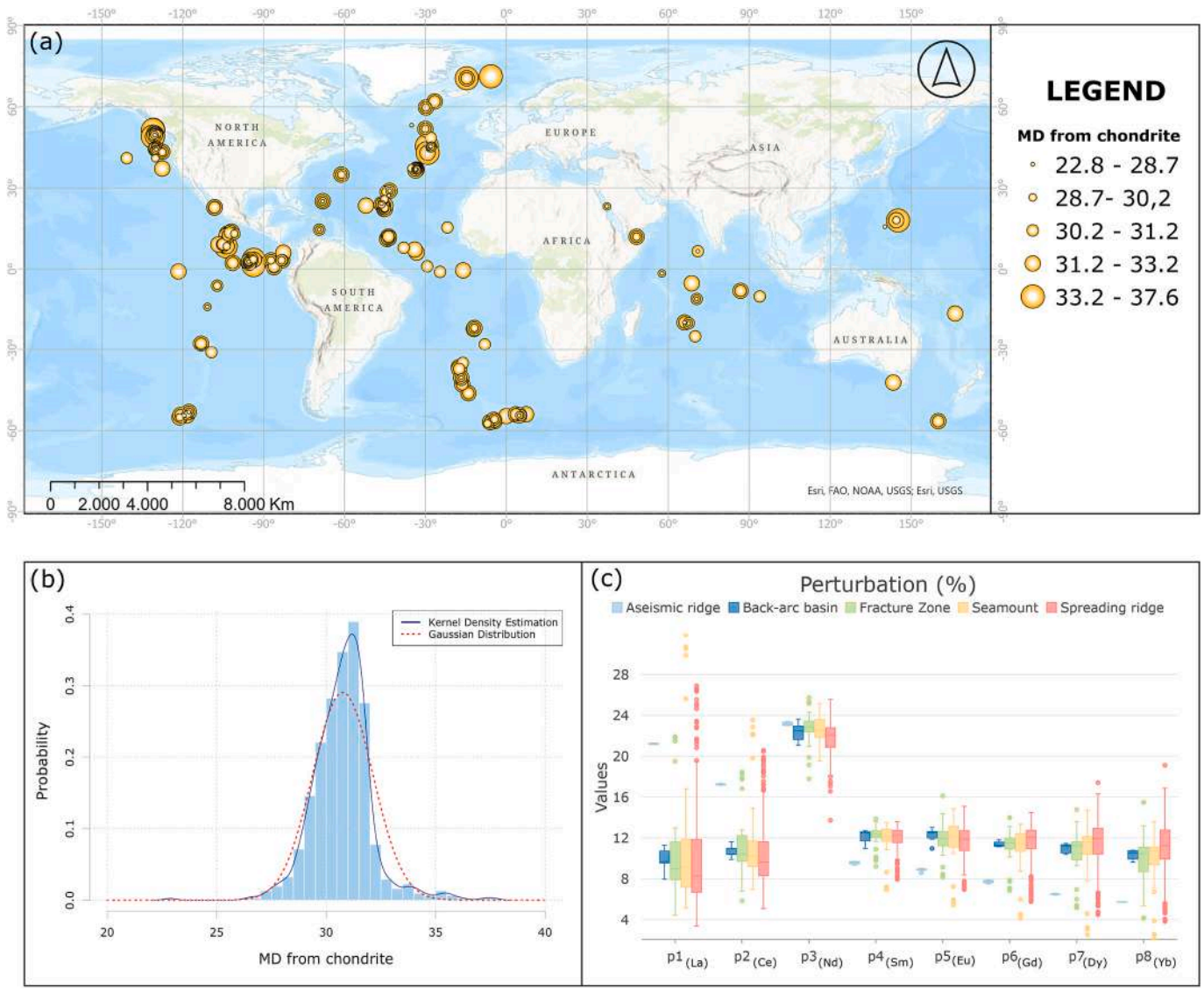

Fig. 2. (a) Map and (b) histogram of the robust *MD* Mahalanobis metric values describing the distance from chondrite composition for a pattern of 8 elements (*La*, *Ce*, *Nd*, *Sm*, *Eu*, *Gd*, *Dy*, *Yb*). (c) Boxplots of the components of the perturbation matrix from the same benchmark ($p_1 = La, p_2 = Ce, p_3 = Nd, p_4 = Sm, p_5 = Eu, p_6 = Gd, p_7 = Dy, p_8 = Yb$) split according to the tectonic setting.

(https://www.compositionaldata.com/codapack.php), that allows the transformation of data via log-ratio methods (additive log-ratio, centered log-ratio, and isometric log-ratio) and supports operations in the *simplex*, such as centering, perturbation, power transformation, amalgamation, and subcomposition (closure) within the compositional data framework. Alternatively, the same calcolus can be performed in R using packages such as compositions, RobCompositions or compositional (see e.g., https://CRAN.R-project.org/package=Compositional). The computed robust Mahalanobis distances and perturbation values were imported into ArcGIS Pro (Esri, 2024) for spatial visualization and map creation, while R was used to generate histograms with density curves and violin plots for statistical analysis.

3. Results

3.1. The AVGD database

3.1.1. Compositional diversification relative to chondrite

In Fig. 2(a-b), the map of *MDs* computed for each AVGD glass from the chondrite composition is shown together with the histogram of all *MD* values, considering a set of 8 elements (*La*, *Ce*, *Nd*, *Sm*, *Eu*, *Gd*, *Dy* and *Yb* (Wilson, 1989)). In addition, boxplots illustrating the perturbations experienced by each variable relative to the same benchmark are reported in Fig. 2(c). The histogram exhibits a slight asymmetry with a more populated right tail. However, the glasses appear to be compositionally homogeneous around the *MD* median value of 30.78. All of them are more evolved with respect to the benchmark ($MD = 0$), which represents a primitive composition. In this context, the glasses with the highest *MDs* originate from the Pacific Tuzo Wilson seamounts, while the lowest are from the Read Sea spreading ridge. The map allow us to trace the spatial patterns of the changes and identify where the most significant differences are concentrated. When considering the perturbation values, spreading ridges appear to show the higher variability, particularly for *La* and *Ce*, where anomalous values are significant. These two elements, together with *Dy* and *Pb*, appears to generate some diversification among the glasses during the geochemical evolution. Within this framework, *Nd* is the element experiencing the most pronounced relative changes (perturbations) across all geo-structural contexts.

A broader dataset can be selected to calculate the *MDs* relative to the chondrite composition benchmark, as is typically done for a spider di-

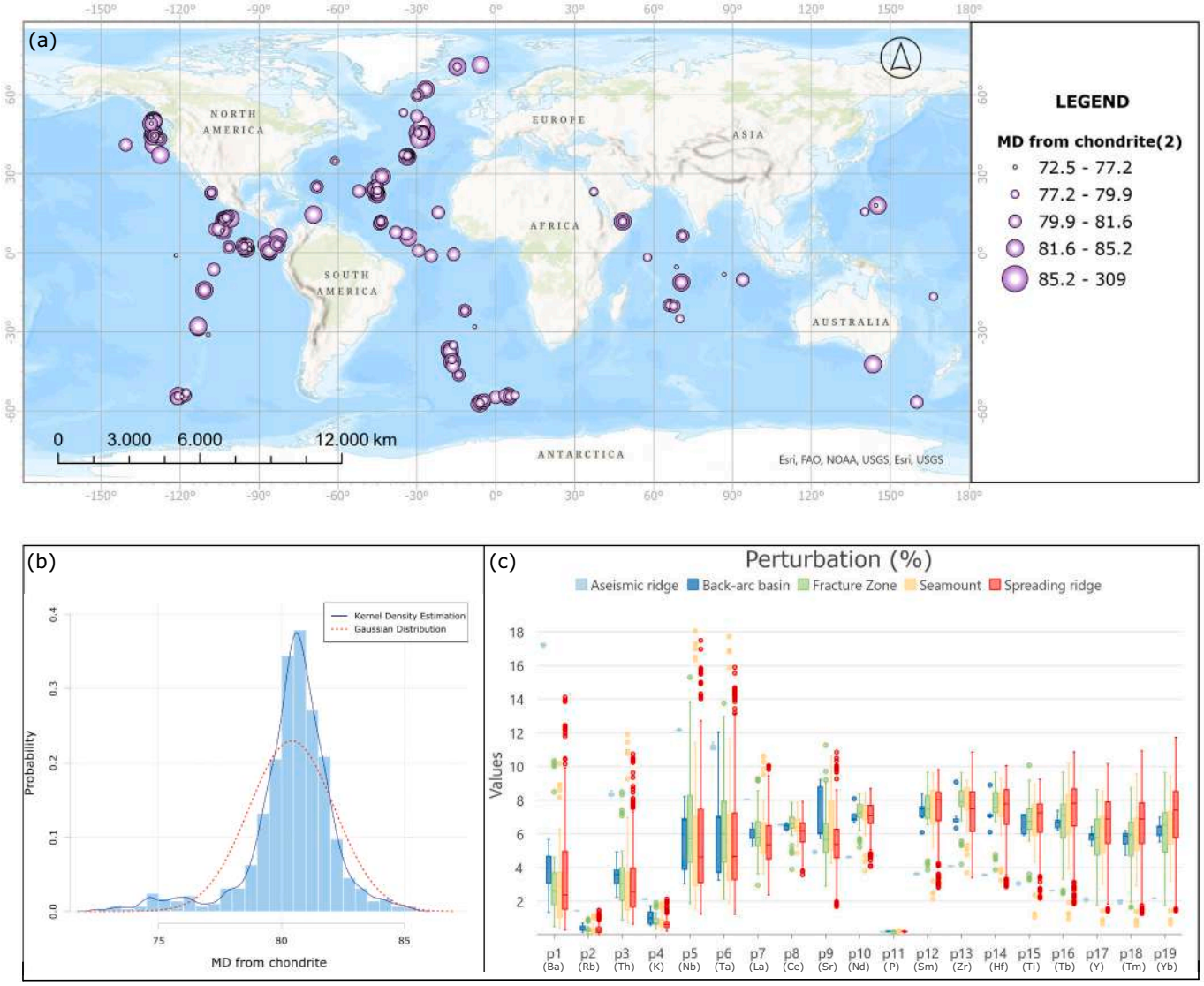

Fig. 3. (a) Map and (b) histogram of the robust *MD* Mahalanobis metric values describing the distance from chondrite composition for a pattern of 19 elements from *Ba* to *Yb* (Wilson, 1989). (c) Boxplots of the components of the perturbation matrix from the same benchmark ($p_1 = Ba, p_2 = Rb, p_3 = Th, p_4 = K, p_5 = Nb, p_6 = Ta, p_7 = La, p - 8 = Ce, p_9 = Sr, p_{10} = Nd, p_{11} = P, p_{12} = Sm, p_{13} = Zr, p_{14} = Hf, p_{15} = Ti, p_{16} = Tb, p_{17} = Y, p_{18} = Tm, p_{19} = Yb$) split according to the tectonic setting.

agram (Wilson, 1989; Rollinson, 2007). A dataset of 19 elements was considered, and the same analytical procedures described above were applied. The elements include *Ba*, *Rb*, *Th*, *K*, *Nb*, *Ta*, *La*, *Ce*, *Sr*, *Nd*, *P*, *Sm*, *Zr*, *Hf*, *Ti*, *Tb*, *Y*, *Tm*, and *Yb*, with their rank order reflecting a progressive decrease in incompatibility. Results are reported in Fig. 3(a-c).

As in the previous case, the *MDs* frequency distribution exhibits good symmetry around a well-defined central value, with a median of 80.6. However, a small group of lower values appears in the left tail of the distribution. Most of these correspond to samples collected from the fracture zone of the Pacific-Antarctic Ridge, the upper northern slope of the ridge in the southern Blanco Fracture Zone and trough, the Tuzo Wilson seamounts, the Siqueiros Fracture Zone, and the Mariana Trough back-arc basin. The spatial distribution of *MD* values, compared to the previous 8-element pattern, reveals several hot-spots where the deviation from the chondrite composition is either lower or higher. The boxplots indicate that perturbation values are generally comparable, except for *Rb*, *K* and *P* showing notably low values. Most elements on the left exhibit high anomalous perturbations, whereas those on the right display low anomalous values. This highlights a contrasting behavior driven by geochemical incompatibility.

3.1.2. *Compositional diversification relative to MORBs*

The AVGD chemical composition can also be compared with that of N-MORB, D-MORB and *E*-MORB (Gale et al., 2013) considering the following set of elements ranked according to decreasing mobility (Wilson, 1989): *Sr*, *K*, *Rb*, *Ba*, *Th*, *Nb*, *Ce*, *P*, *Zr*, *Hf*, *Sm*, *Ti*, *Y* and *Yb*. Results are reported in Fig. 4(a-c). In all cases, the range of *MDs* for N-, D-, and E-MORB is very similar, with medians of 7.56, 7.95, and 9.02, respectively. The histograms exhibit low skewness, with a few higher values noticeable only for D-MORB and E-MORB. The spatial distribution of the data is also very similar, with higher *MD* values observed along the spreading ridge and fracture zone of the Pacific–Antarctic region and at the Tuzo Wilson Seamounts. This similarity in the *MDs* among N-, D- and E-MORB is not maintained in the perturbation values shown in the boxplots, where some differences become evident. Perturbations are more homogeneous across all elements in N-MORB, whereas clear diversification patterns emerge in D-MORB and E-MORB. Overall, variability seems to increase with element mobility. For the N-MORB benchmark, *Rb* and *Ba* exhibit higher variability, with some anomalous cases. This feature is more pronounced in D-MORB, where *Rb* and *Ba* again show the same behavior. In contrast, a different pattern is

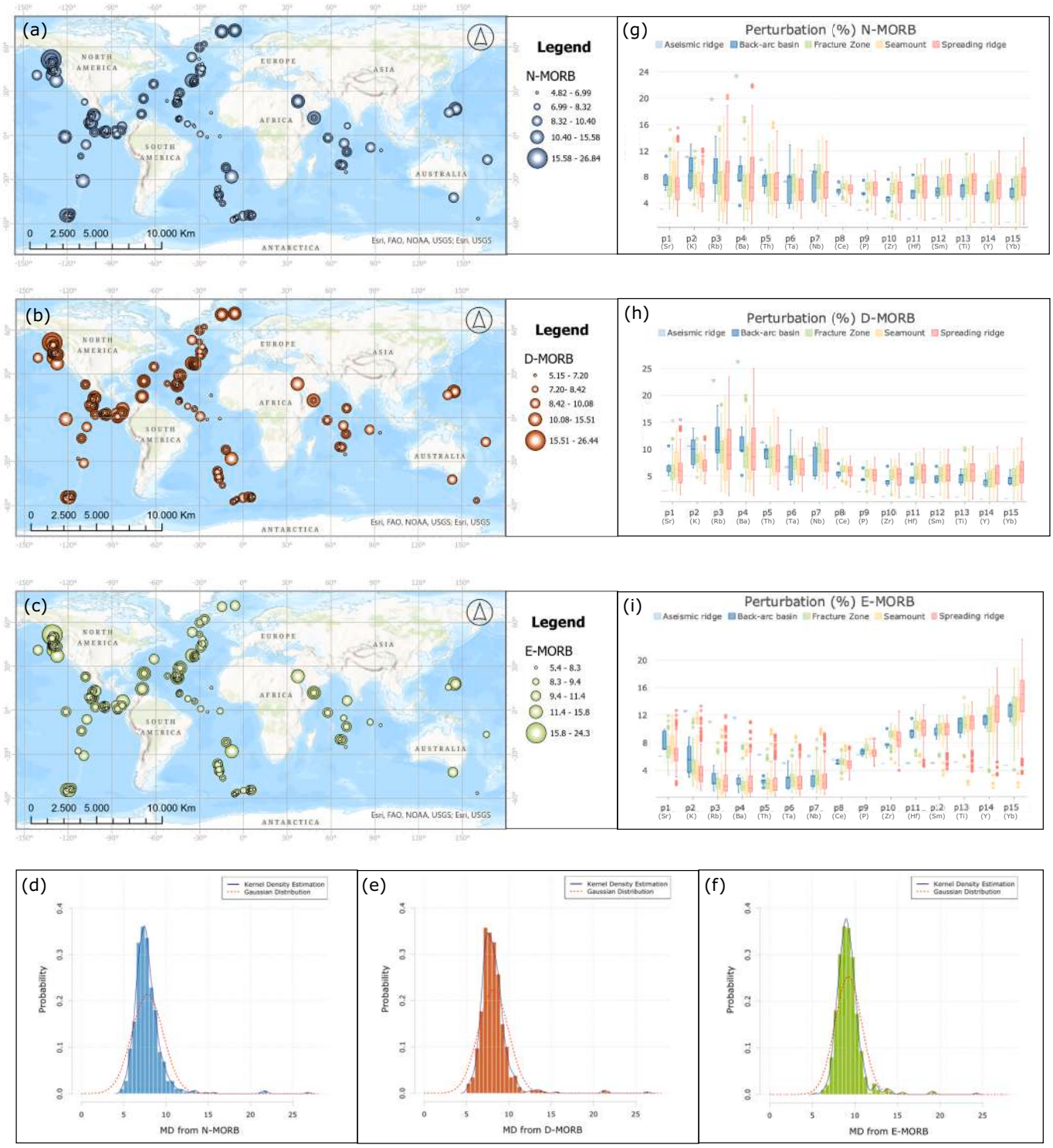

Fig. 4. (a-c) Maps and (d-f) histograms of the Mahalanobis distance *MD* values for benchmarks given by N-MORB, D-MORB and E-MORB for 15 elements, *Sr*, *K*, *Rb*, *Ba*, *Th*, *Ta*, *Nb*, *Ce*, *P*, *Zr*, *Hf*, *Sm*, *Ti*, *Y*, *Yb* (treated line = Gaussian distribution, continuous line = Kernel density estimation). (g-i) Boxplots of the components of the perturbation matrix from the MORBs composition (N-MORB, D-MORB, E-MORB) chosen as benchmark with 15 elements ($p_1 = Sr$, $p_2 = K$, $p_3 = Rb$, $p_4 = Ba$, $p_5 = Th$, $p_6 = Ta$, $p_7 = Nb$, $p_8 = Ce$, $p_9 = P$, $p_{10} = Zr$, $p_{11} = Hf$, $p_{12} = Sm$, $p_{13} = Ti$, $p_{14} = Y$, $p_{15} = Yb$) split according to the tectonic setting.

observed for the E-MORB benchmark, where the occurrence of high or low anomalous perturbations is clearly related to geochemical mobility.

3.2. The FOREGS repository

3.2.1. Compositional diversification across alluvial and stream sediments

For stream and floodplain sediments the patterns of *Cr*, *Ni*, *Cu*, *Zn*, *Pb*, *Co* and *V* were analyzed in comparison with the composition of the average continental crust. Results are shown in Fig. 5 for stream

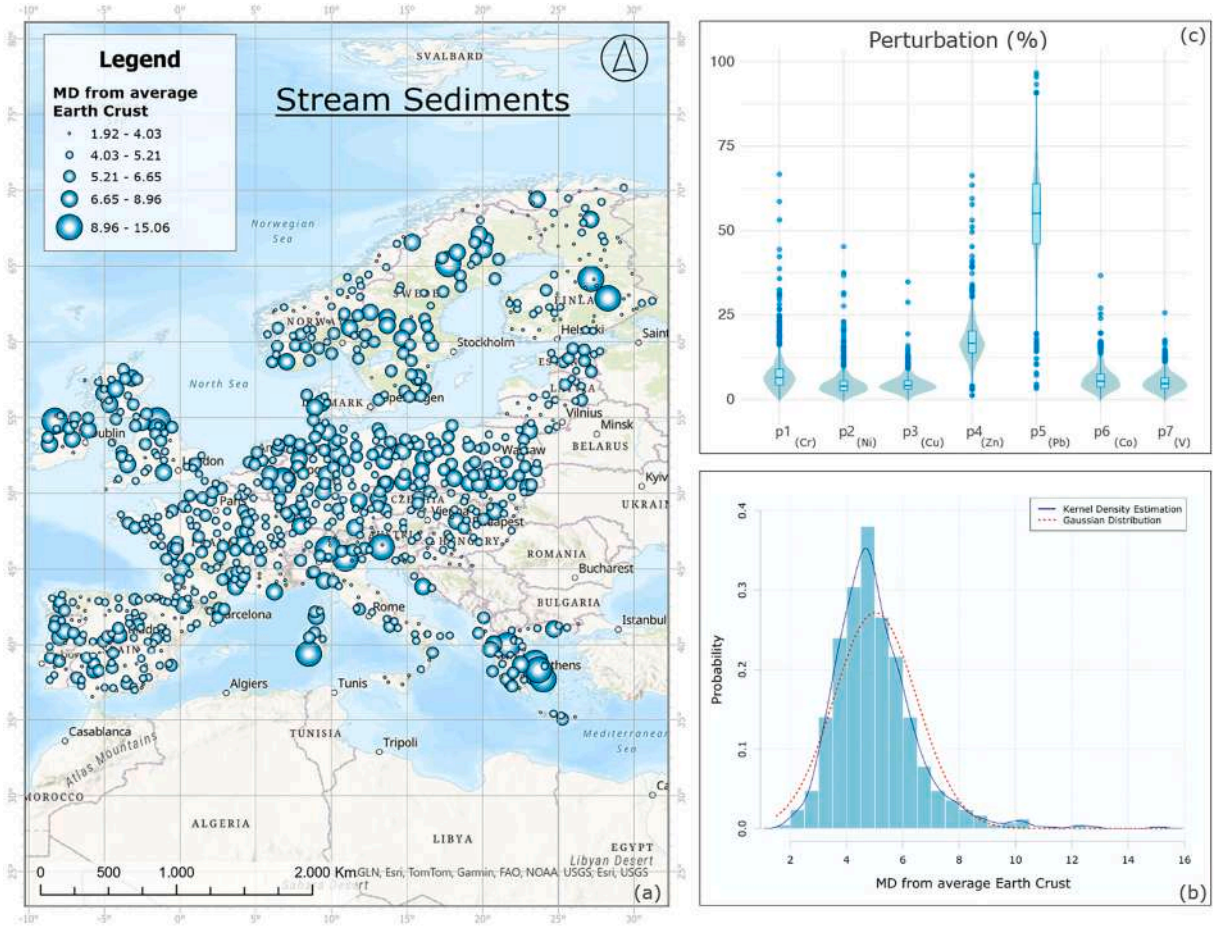

Fig. 5. Metric distance of stream sediment composition from the average Earth crust chosen as benchmark for 7 elements. (a) Map and (b) histogram of the Mahalanobis distance values and (c) boxplots of the components of the perturbation matrix with $p_1 = Cr$, $p_2 = Ni$, $p_3 = Cu$, $p_4 = Zn$, $p_5 = Pb$, $p_6 = Co$, $p_7 = V$.

sediments and in Fig. 6 for floodplain sediments.

In stream sediments, the deviation from the average continental crust is associated with some skewness in the *MD* distribution. Spatially, lower values are mainly observed in the northwestern part of Europe, where the most extensive areas of exposed Precambrian rocks occur. The median *MD* is 4.78, and higher values appear to be associated with ophiolite belts, such as in Greece and Albania, or with mineralizations, such as in southern Sardinia and Lavrion, Greece. From the boxplots, *Pb* appears to be the most variable element in both floodplain and stream sediments, whereas in other elements, only the presence of anomalous cases leads to higher perturbation values. For floodplain sediments, the median *MD* is 4.85, similar to stream sediments, but with a clearly skewed frequency distribution. From a spatial perspective, high *MD* values tend to correspond to areas with vein-type mineralization historically mined, massive sulfide (Pb-Zn) outcrops hosted in Cambro-Ordovician schists and limestones, and Silurian black shales (eastern Pyrenees, southern Norway, Massif Central in France, Iberian Pyrite Belt, Roman Alkaline Province, etc.), as well as regions affected by atmospheric deposition (e.g., fly ash from power plants in northern Czech Republic) and industrial emissions.

3.2.2. *Compositional diversification in stream waters*

Another application of the proposed method concerns natural waters. Compositional changes in European stream waters were evaluated using as a benchmark the sample with the lowest TDS (Total Dissolved Solids, or low conductivity), from Norway. The chemical composition of the stream waters is represented by HCO_3^-, Cl^-, SO_4^{2-}, K^+, Na^+, Mg^{2+}, Ca^{2+} in mg/L, with the benchmark sample having concentrations of 0.72, 0.45, 0.52, 0.0050, 0.3, 0.048 and 0.24 mg/L, respectively. Increases in *MD* values are expected to reflect the intensity of weathering processes and changes in lithology and environmental conditions where contributions from different sources are present. Results are shown in Fig. 7. The histogram of *MD* values (median = 3.03) is fairly symmetrical and is characterized by a low variability, with only a small group of data in the right tail of the distribution showing higher values. From a spatial perspective, lower values are associated with the more acidic environments across southern and central Fennoscandia, while higher values occur in Mediterranean regions, in the calcareous environments of northern France and southeastern England, and in areas affected by intensive agriculture and wide use of Ca-Mg fertilizers (e.g., Germany, Poland, and the Baltic states). The observation of the boxplots indicates that Ca^{2+} is the chemical species most affected by variability and perturbation, followed by Na^+, Cl^-, and Mg^{2+}. Anomalous values are observed in elements with a lower interquartile range, such as HCO_3^-, SO_4^{2-}, K^+ and Mg^{2+}.

4. Discussion

Allègre and Lewin (1995) described geochemical systems as governed by "operators" and "distributions". Geochemical distributions are defined as a set of values of concentration in a specific environmental medium, which can be analyzed using histograms or spatially represented through maps. Geochemical operators, such as volcanism, sedimentation, weathering processes, climate changes and human impact, transform one geochemical distribution into another. Due to the complexity of natural environments, these operators (also referred to as

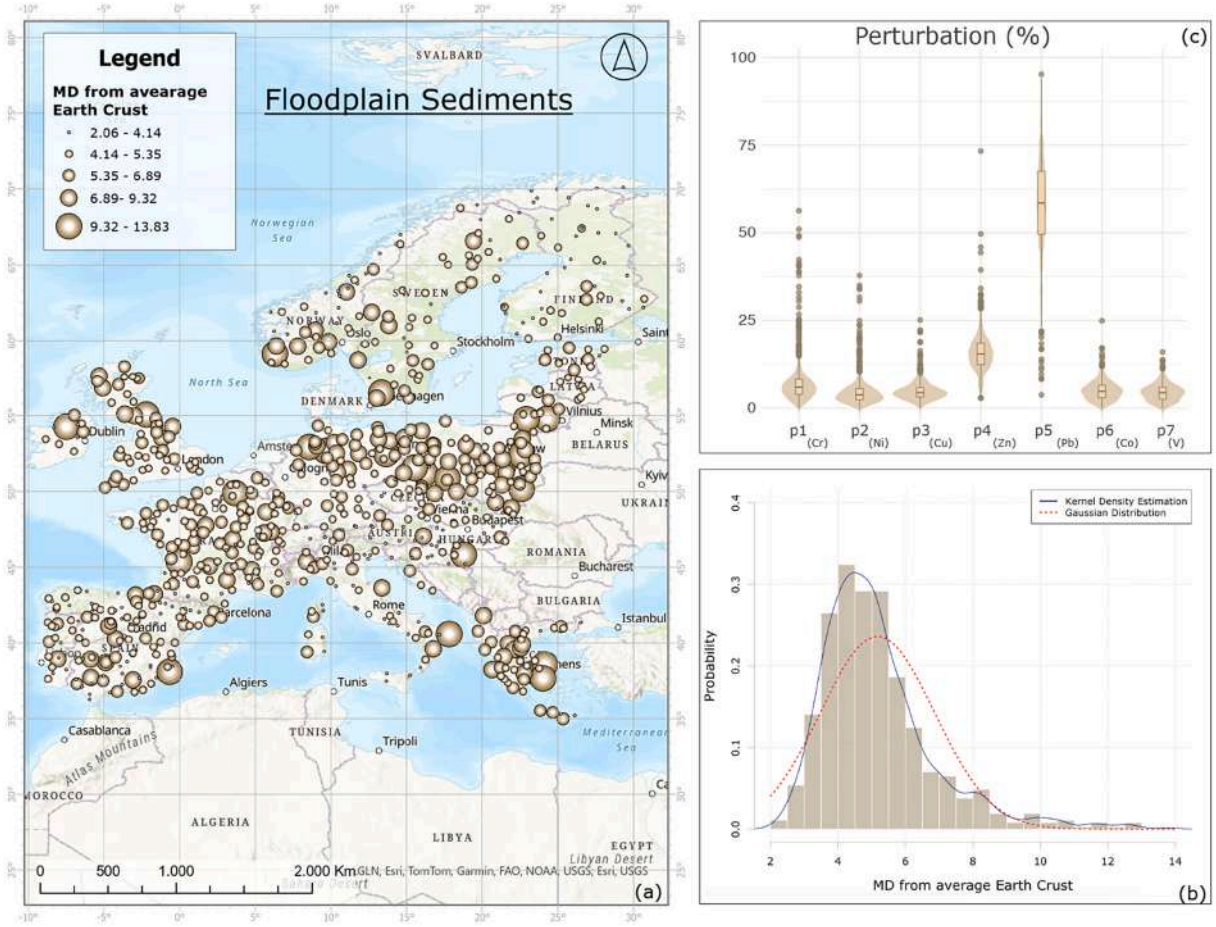

Fig. 6. Metric distance of floodplain sediment composition from the average Earth crust chosen as benchmark for 7 elements. (a) Map and (b) histogram of the Mahalanobis distance values and (c) boxplots of the components of the perturbation matrix with $p_1 = Cr$, $p_2 = Ni$, $p_3 = Cu$, $p_4 = Zn$, $p_5 = Pb$, $p_6 = Co$, $p_7 = V$.

drivers), may combine in an additive or multiplicative way, thereby affecting the shape of frequency distribution of concentrations. In this context, the repeated action of mixing operators tend to produce symmetrical (normal) distributions, whereas incomplete mixing is often associated with bimodal, or even multimodal distributions (Allègre and Lewin, 1995). Following Scheffer et al. (2012), multimodality can indicate a system's tendency to undergo flickering, a typical early-warning signal in highly stochastic systems. The presence of skewed (lognormal) or heavy-tailed distributions, instead, is more commonly related to complex dynamics driven by differentiation, multiplicative or cascade processes, as well as feedback mechanisms (van Rooij et al., 2013).

The analysis of the AVGD and FOREGS datasets reveal that changes can be detected when differences among cases are properly measured from a benchmark composition, using different yet complementary compositional approaches. A metric such as the robust Mahalanobis distance (*MD*), calculate on log-transformed data, provide a first synthesis of how differences evolve over time or space. It improves the detection of time periods or spatial regions with larger differences, indicating more efficient geochemical processes, contribution of multiple sources and, consequently, increased geochemical complexity. It could also help the identification of alternative states when a clear bimodality is present, thereby revealing flickering processes underlying major environmental changes (Scheffer et al., 2012), as observed, for instance, in the Amazon forest system (Flores et al., 2024). When an environment loses stationary conditions due to long-term feedbacks among its components, new interaction structures emerge, altering system resilience and increasing the risk of reaching a tipping point and triggering a critical transition. From this perspective, compositional benchmark-based MD, represents an interesting early-warning signal of resilience loss and increasing instability in geochemical systems. In Fig. 8, the kernel density distribution of the *MD* values is shown on an inverted scale.

We can now interpret the vertical axis (the inverse of frequency or probability) as the potential of the system, with the valleys representing the shape of different attraction basins where the *MDs* are most frequently found (Scheffer et al., 2012). The deeper the valley, the lower the variability, implying greater stability and, consequently, a stronger concentration of data around a defined and representative *MD* value. As shown in Fig. 8 and the previous figures, most systems appear resilient at the analyzed broad scale, with major differences restricted to localized hot-spot conditions. When the surficial environment is involved, variability, and thus instability, increase due to strong interactions with multiple reservoirs, each with a complex geochemistry. When information from *MD* values on system stability is combined with that from perturbation values relative to the same compositional benchmark, valuable insights can be gained into the behavior of individual elements (or groups of elements) during geochemical processes. Our analysis reveals which elements in the compositions are more susceptible to perturbation and/or more likely to generate extreme perturbations, enhancing bimodality, flickering, and instability that are closely linked to source incompatibility or geochemical mobility. In this framework, the behavior of *Zn* and *Pb* in stream and floodplain sediments is particularly relevant from an environmental perspective, as these elements exhibit the highest perturbation values compared with the average continental crust. Both are chalcophile metallic elements that

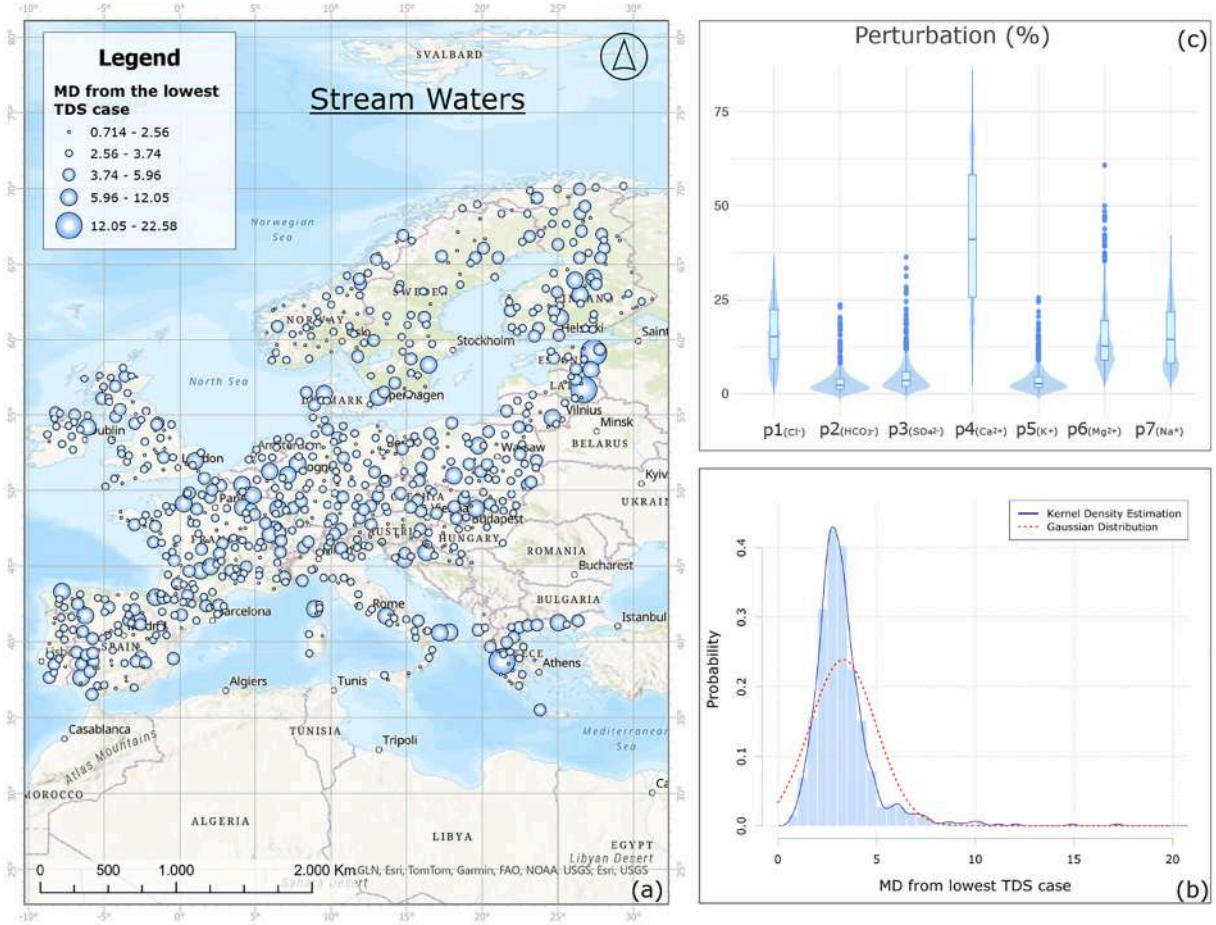

Fig. 7. Metric distance of stream water composition from the lower TDS case used as benchmark for 7 chemical species, Cl^-, HCO_3^-, SO_4^{2-}, Ca^{2+}, K^+, Mg^{2+} and Na^+ in mg/L. (a) Map and (b) histogram of the Mahalanobis distance values and (c) boxplots of the components of the perturbation matrix with $p_1 = Cl^-$, $p_2 = HCO_3^-$, $p_3 = SO_4^{2-}$, $p_4 = Ca^{2+}$, $p_5 = K^+$, $p_6 = Mg^{2+}$, $p_7 = Na^+$.

form several minerals, but they are also widely dispersed as trace elements in pyroxene, amphibole, mica, garnet, and magnetite (*Zn*), or in *K*–feldspar, plagioclase, mica, zircon, and magnetite (*Pb*). High perturbation values likely reflect differences in sources, affinities with organic matter or presence of contamination (e.g., mineralization, human activity). This leads to greater variability, as seen in stream and floodplain sediments that act as sinks in geochemical transfer processes. In general, these elements, compared to the average continental crust, exhibit more fragmented concentrations at the European scale. Another noteworthy aspect is the variability of the perturbation values of Ca^{2+} in stream waters at the analyzed scale, where this chemical component appears particularly sensitive. This behavior is primarily driven by changes in lithology, from acidic igneous and metamorphic (schist) rocks to calcareous formations, as well as by potential contributions from *Ca–Mg* fertilizers. Understanding system stability and how it can be lost, represents a crucial area of investigation in the context of climate change and growing pressure on natural environments. Our method, applicable across different spatial and temporal scales, can be highly useful for probing the system state and the behavior of individual variables within the overall composition, without ignoring their interdependence. In many cases, the loss of stability occurs through tipping points, where abrupt responses arise from small changes in external drivers. The ability to identify such tipping points is fundamental for improving our understanding of complex geochemical systems.

5. Conclusions

Aiming to provide a conceptual–methodological review with extensive case studies, this work revisits the foundations of compositional distance, and consequently the nature of geochemical processes. We demonstrate how robust Mahalanobis distances (*MD*) and the perturbation approach can serve as quantitative indicators of compositional change, enhancing the interpretation of the stability and variability of geochemical systems. Most of the analyzed systems appear resilient at broad spatial scales, with major differences generally limited to localized hot-spots. Surficial environments, however, show greater variability and instability due to strong interactions among multiple reservoirs. By combining *MD* values with perturbation metrics, we can identify elements most susceptible to change: *Zn* and *Pb* exhibit high perturbation values reflecting multiple sources, geochemical mobility, and potential contamination, whereas Ca^{2+} in stream waters is particularly sensitive to lithological changes and anthropogenic inputs. Stream and floodplain sediments act as sinks, capturing variability from different processes, highlighting the heterogeneity of geochemical systems relative to average continental crust benchmarks. The proposed distance-based metric can reveal compositional trajectories and evolutionary patterns more effectively than traditional geochemical approaches. These methods allows the detection of early signs of instability and potential tipping points, revealing abrupt responses to small external perturbations. Overall, the combined use of *MD* and perturbation metrics offers a powerful approach to investigate individual elements within the full compositional context, preserving component

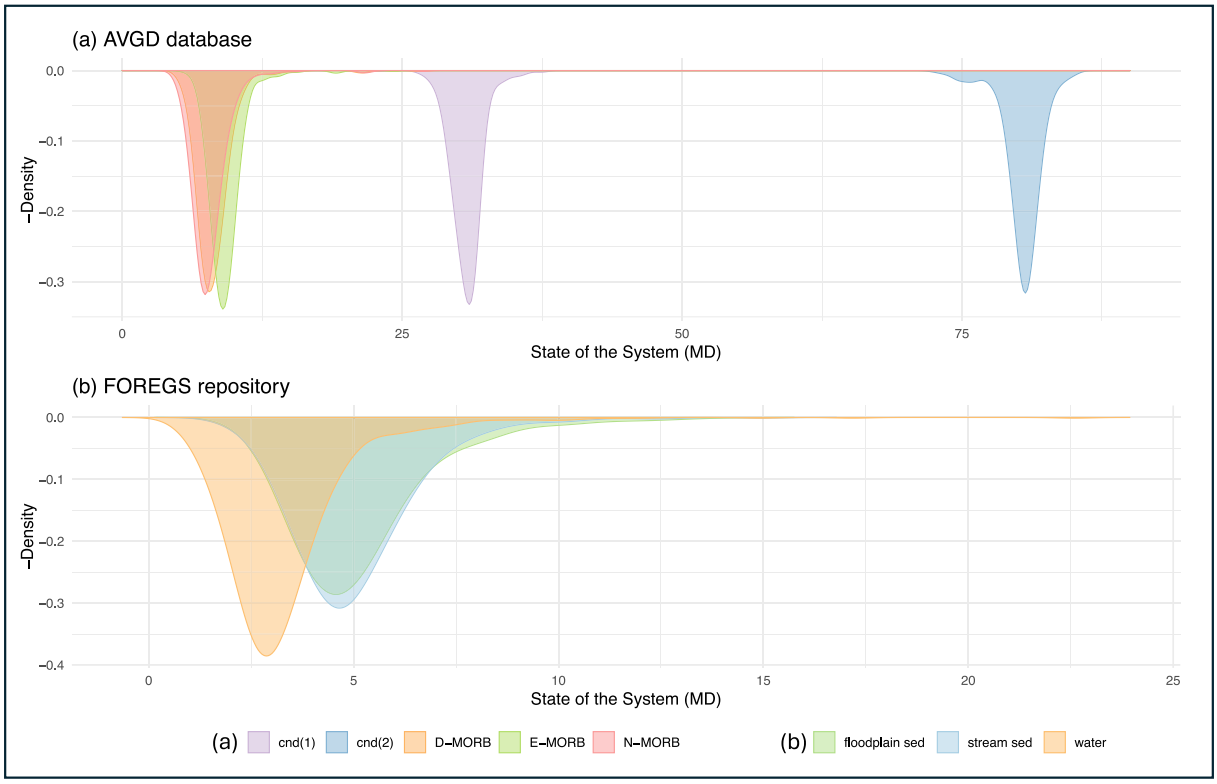

Fig. 8. The inverted kernel density distribution of the *MD* values is shown for (a) the AVGD database and (b) the FOREGS repository.

interdependence and enabling multi-scale geochemical assessments.

It is important to emphasize that the explicit use of benchmark compositions (e.g., chondrite, MORBs, average continental crust, diluted waters) allows compositional variability to be interpreted as a deviation from physically and geochemically meaningful reference states, rather than as purely statistical dispersion around the barycenter. Moreover, the proposed framework complements existing CoDA approaches by providing a robust, benchmark-oriented metric structure that enhance the integration of compositional constraints into multivariate statistics and machine/deep learning workflows, ultimately enabling more reliable and geochemically consistent data-driven analyses.

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.

Acknowledgments

The University of Florence is acknowledged for the financial assistance to the present research through University funds (A.B, C.G.). This work was funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree no. 3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Award Numbers: Project codes CN00000033 and CN00000013, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP: B83C22002910001 and B83C22002830001, Project titles “National Biodiversity Future Center - NBFC” (A.B and C.G) and “National Centre for HPC, Data and Quantum Computing” (A.B). We thank two anonymous reviewers for their helpful comments and suggestions.

Data availability

Data will be made available on request.

References

Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). J. R. Stat. Soc. Ser. B 44 (2), 139–177.
Aitchison, J., 1986. The statistical analysis of compositional data. In: Monographs on Statistics and Applied Probability. Chapman & Hall Ltd, London.
Aitchison, J., Ng, K., 2005. The role of perturbation in compositional data analysis. Stat. Model. 5, 173–185.
Allègre, C.J., Lewin, E., 1995. Scaling laws and geochemical distributions. Earth Planet. Sci. Lett. 132, 1–13. https://doi.org/10.1016/0012-821X(95)00049-I.
Billheimer, D., Guttorp, P., Fagan, W., 2001. Statistical interpretation of species composition. J. Am. Stat. Assoc. 96, 1205–1214. https://doi.org/10.1198/016214501753381850.
Burk, F., 1998. Lebesgue Measure and Integration. An Introduction. Pure and Applied Mathematics. John Wiley & Sons Inc., Canada, p. 292.
Chayes, F., 1960. On correlation between variables of constant sum. J. Geophys. Res. 65 (12), 4185–4193.
Cortes, J., 2009. On the harker variation diagrams; a comment on “the statistical analysis of compositional data. Where are we and where should we be heading?” by Aitchison and Egozcue (2005). Math. Geosci. 41, 817–828.
Egozcue, J., 2009. Reply to “on the harker variation diagrams; ...” by J.A. Cortés. Math. Geosci. 41, 829–834.
Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geosci. 37, 795–828. https://doi.org/10.1007/s11004-005-7381-9.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logration transfomations for compositional data analysis. Math. Geol. 35 (3), 270–300.
Esri, 2024. ArcGIS Pro: Version 3.2. Redlands, California. URL. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview.software.
Faure, G., 1998. Principles and Applications of Geochemistry: A Comprehensive Textbook for Geology Students. Hewlett Packard Professional Books, Prentice Hall.
Filzmoser, P., Hron, K., 2008. Outlier detection for compositional data using robust methods. Math. Geosci. 40, 233–248. https://doi.org/10.1007/s11004-007-9141-5.
Flores, B., Montoya, E., Sakschewski, B., Nascimento, N., Staal, A., Betss, R., Levis, C., Lapola, D., Esquivel-Muelbert, A., Jakovac, C., Nobre, C., Oliveira, R., Borma, L., Nian, D., Boers, N., Hecht, S., ter Steege, H., Arieira, J., Lucas, I., Berenguer, E., Maengo, J., Gatti, L., Mattos, C., Hirota, M., 2024. Critical transitions in the amazon forest system. Nature 626, 555–564.

Gaillardet, J., Dupré, B., Louvat, P., Allègre, C., 1999. Global silicate weathering and CO_2 consumption rates deduced from the chemistry of large rivers. Chem. Geol. 159, 3–30.
Gale, A., Dalton, C., Lanmuir, C., Su, Y., Schillin, J., 2013. The mean composition of ocean ridge basalts. Geochem. Geophys. Geosyst. 14 (3), 489–518.
Gozzi, C., Buccianti, A., 2022. Assessing indices tracking changes in river geochemistry and implications for monitoring. Nat. Resour. Res. 31 (2). https://doi.org/10.1007/s11053-022-10014-1.
Gozzi, C., Sauro Graziano, R., Buccianti, A., 2020. Part–whole relations: new insights about the dynamics of complex geochemical riverine systems. Minerals 10. https://doi.org/10.3390/min10060501.
Gozzi, C., Dakos, V., Buccianti, A., Vaselli, O., 2021. Are geochemical regime shifts identifiable in river waters? Exploring the compositional dynamics of the Tiber River (Italy). Sci. Total Environ. 785, 147268. https://doi.org/10.1016/j.scitotenv.2021.147268.
Jenner, F., O'Neill, H., 2012. Analysis of 60 elements in 616 ocean floor basaltic glasses. Geochem. Geophys. Geosyst. 13, 1. https://doi.org/10.1029/2011GC004009.
Maronna, R., Martin, R., Yohai, V., 2006. Robust Statistics: Theory and Methods. John Wiley & Sons, New York. ISBN 978–0–470-01092-1.
Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J., 2011. The principle of working in coordinates. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd, Chichester, pp. 34–42. https://doi.org/10.1002/9781119976462.ch3.
Melson, W., O'Heaarn, T., Jarosewich, E., 2002. The mean composition of ocean ridge basalts. Geochem. Geophys. Geosyst. 3 (4), 1023. https://doi.org/10.1029/2001GC000249.
Navarro-Lopez, C., Linares-Mustaros, S., Mulet-Forteza, C., 2022. "The Statistical Analysis of Compositional Data" by John Aitchison (1986): a Bibliometric Overview. SAGE Open 12. https://doi.org/10.1177/21582440221093366.
Pandit, P., Saini, A., Sahu, N., Mehra, R., 2023. Geochemical evaluation and environmental risk assessment of heavy metals: a case study from Ireland using tellus stream sediment data (2011–2017). Gorundwater Sustain. Develop. 23, 1–16.
Pawlowsky-Glahn, V., Buccianti, A., 2011. Compositional Data Analysis: Theory and Applications. John Wiley & Sons Ltd., p. 400
Pawlowsky-Glahn, V., Egozcue, J., 2001. Geometric approach to statistical analysis on the simplex. Stoch. Env. Res. Risk A. 15, 384–398.
Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. Statistics in Practice. John Wiley & Sons Ltd., p. 272
R Development Core Team, 2023. R: A Language and Environment for Statistical Computing. Version 4.2.0.. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.
Rollinson, H., 2007. Early Earth Systems. A Geochemical Approach. Blakwell Publishing, UK, p. 285.
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. https://doi.org/10.1016/j.gexplo.2020.106465.
Scheffer, M., Carpenter, S., Lenton, T., Bascompte, J., Brock, W., Dakos, V., van de Koppel, J., van de Leemput, I., Levin, S., Van Nes, E., Pascual, M., Vandermeer, J., 2012. Anticipating critical transitions. Science 338 (6105), 344–348. https://doi.org/10.1126/science.1225244.
Subramanian, B., Paul, A., Kim, J., Chee, K., 2022. Metrics space and norm: taxonomy to distance metrics. Sci. Program. https://doi.org/10.1155/2022/1911345.
The MathWorks, Inc, 2023. MATLAB: Version 9.15 (R2023b). Natick, Massachusetts. URL. https://www.mathworks.com/products/matlab.html.software.
van Rooij, M., Nash, B., Rajaraman, S., Holden, J., 2013. A fractal approach to dynamic inference and distribution analysis. Front. Physiol. 4 (1), 1–11.
Wilson, M., 1989. Igneous Petrogenesis. Unwin Hyman, London, p. 466.