Journal of Geochemical Exploration 265 (2024) 107545

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Mapping geochemical domains using stream sediment geochemistry: An approach based on compositional indicators in the Volturno River basin (South Italy)

Maurizio Ambrosino [a], Javier Palarea-Albaladejo [b], Stefano Albanese [c,*], Domenico Cicchella [a]

[a] Department of Science and Technology, University of Sannio, 82100 Benevento, Italy
[b] Department of Computer Sciences, Applied Mathematics and Statistics, University of Girona, 17003 Girona, Spain
[c] Department of Earth, Environmental and Resources Sciences, University of Naples Federico II, 80126 Naples, Italy

ARTICLE INFO

Keywords:
Compositional indicators
Geochemical domains
Log-ratio analysis
Principal balances
Stream sediments geochemistry

ABSTRACT

When dealing with environmental problems, it is of fundamental importance to establish reference values (geochemical baselines) against which to determine the presence or absence of active contamination processes.

In the effort to develop a method to assess the geochemical baselines for territories featuring complex geological settings and a well-established anthropic environmental pressure, we combined compositional data analysis (CoDA) with geolithological information to reduce the degree of uncertainty possibly affecting the results. The proposed approach comprises (1) a knowledge-driven step to select a number of sample subsets from a geochemical dataset each with a high probability of having its composition strongly influenced by only one of the lithologies outcropping in the study area; (2) a data-driven step to compute compositional principal balances and define geochemical indicators to be used to assign each of the observations in the dataset to one of the geochemical domains associated to a mayor lithologies outcropping in the study area; (3) the determination for each geochemical domain of baseline values based on the samples assigned to them by the data-driven step.

The method was tested using the geochemical data referring to 887 stream sediment samples collected across the Volturno River catchment basin (Southern Italy), featuring a relevant lithological heterogeneity.

The results obtained were easily interpretable as they fitted well with the geomorphological, geochemical, and geodynamic processes characterizing the study area.

Despite the use of stream sediments for the specific case study presented, the application principles of the method hold for any environmental media and for any territory for which there is a need to define baseline values. However, for a successful application of the method, it is crucial to have a fair knowledge of the geological settings of the study area.

1. Introduction

In recent years, we have witnessed a critical research effort to quantify the impact of human-induced activities on the environment and their associated risks. However, determining the geochemical baseline values is a critical step towards establishing this. A geochemical baseline for an element refers to its natural variations in concentration in a given environmental compartment (Salminen and Tarvainen, 1997), and it strongly depends on geological characteristics such as mineral composition, grain size distribution, and organic matter content (Dung et al., 2013). The geochemical baseline can be estimated using both empirical and statistical methods. In a nutshell, we could say that the empirical approach usually involves the elaboration of data related to samples proceeding from areas not affected by anthropogenic activities (which are also referred to as preindustrial samples), and the statistical methods are based on the identification and elimination of outliers (Nawrot et al., 2021). However, recognizing samples unaffected by human pollution and quantifying the amount of a pollutant introduced into the environment by human activities is a fairly difficult task. In many cases, the degree of contamination of an environmental media is determined by assuming a unique reference value (i.e geochemical baselines) throughout the whole study area (Boente et al., 2022; Chen et al., 2019; Sundaray et al., 2011); this makes sense in studies conducted at a local scale, but it can be too stringent in studies conducted at a regional scale

[*] Corresponding author.
E-mail address: stefano.albanese@unina.it (S. Albanese).

https://doi.org/10.1016/j.gexplo.2024.107545
Received 3 January 2024; Received in revised form 19 May 2024; Accepted 18 June 2024
Available online 3 July 2024

due to the presence of lithological variations characterizing the geological context and different degree of weathering affecting rocks.

The extent of environmental contamination is typically evaluated by computing indices that utilize geochemical baselines as reference values. For instance, amongst others, this applies to measures such as the enrichment factor, the contamination factor or the geo-accumulation index. Although there are studies in the literature where the pollution indices are based on local geochemical baselines (Cicchella et al., 2023; Nawrot et al., 2021), the reference level can also be based on the average content of elements in the earth's crust. Recent research (Cicchella et al., 2022; Wang et al., 2021) has indicates that while the identification of the spatial distribution of geological units serves as a solid foundation for assessing geochemical baselines, their form cannot be directly used to delineate areas with a singular reference geochemical composition. This is because geological units do not encompass any geochemical data, but solely contain chronological and geodynamic information. This is particularly true for territories with a complex geological setting (e.g., Campania region in Italy) where rocks with very different chemical compositions (such as limestones, clays, and pyroclastic levels) can coexist within a geological unit (Piana et al., 2017; Vitale et al., 2011; Vitale and Ciarcia, 2018). Therefore, to define geochemical baselines, geological information should be integrated, to the best of possibilities, with geochemical data to determine "geochemical domains" where the geochemical variability found in environmental media can be associated with a specific pool of processes and lithologies. Determining geochemical domains should represent a starting point for environmental, agricultural, and crop production studies. Therefore, this study represents a first step towards an adequate evaluation of geochemical baselines and the environmental state of soils, proposing a valuable approach to identifying geochemical domains.

The development of well-principled statistical methods for compositional data analysis (CoDA) based on log-ratio coordinates (Aitchison, 1982; Pawlowsky-Glahn et al., 2015) has dramatically contributed to overcoming the limitations and pitfalls of traditional statistical methods when directly applied to raw geochemical data (Chayes, 1962). Thus, analyses based on log-ratio data representations, such as additive log-ratio, centered log-ratio, or isometric log-ratio coordinates, are now common in the literature and have notably enhanced the significance of statistical analysis in the geological and environmental sciences. Some illustrative examples include the evaluation of soil baselines (Cicchella et al., 2022), soil pollution (Cicchella et al., 2020; Aruta et al., 2022), soil mineralogy (Butler et al., 2020), water quality (Glendell et al., 2019), water dynamics (Graziano et al., 2020), air pollution (Jarauta-Bragulat et al., 2016), environmental factors leading to diseases (McKinley et al., 2020; McKinley et al., 2021); and health risk assessment (Tepanosyan et al., 2020). In the evaluation of soil pollution, the CoDA approach has led to the definition of new pollution indices (or compositional indicators) for studies at both local (Boente et al., 2022) and regional scales (Petrik et al., 2018).

Based on the above considerations, we propose a novel approach to recognize and differentiate geographical areas in terms of their geochemical baselines, aiming to minimize the uncertainty around it. The method combines inputs from the geological knowledge of the study area with evidence drawn from geochemical data to generate indicators that allow distinguishing geochemical domains. These are understood as areas with low compositional variation, which exhibit a characteristic geochemical signature in response to natural phenomena such as bedrock type and weathering degree. Stream sediment data from the catchment basin of the Volturno River (Southern Italy) are used here as a case study. In order to recognize only the natural phenomena that affect soil geochemistry, emphasis has been put on chemical elements whose concentration in the environment is typically not affected by anthropic activities and varies considerably according to lithologies. For these elements, the geochemical baselines are estimated for each geochemical domain. The geochemical baselines of elements that may present an anthropogenic contribution will be elaborated in future research, focusing on the individual geochemical domains recognized in this study. In our view, focusing on the individual geochemical domains facilitates the identification and removal of the anthropic contribution, since there is no overlapping of multiple natural sources.

2. Study area

The Volturno River basin is located along the main axis of the southern Italian Apennines. The Matese massif, Meta mountains, Daunia mountains, Picenti mountains, Partenio mountains, and the Roccamonfina volcano represent its main water and sediment alimentation sources (Fig. 1). The watershed size is 5500 km^2 and the more anthropized areas are located in the south-western sector, between the cities of Avellino, Caserta and Benevento where industrial, agricultural and livestock activities are present. Although in the central-western sector of the study area there has been a considerable increase in urbanization in recent decades (Ruberti and Vigliotti, 2017), a large part of its extension is characterized by the presence of small urban centres, which are undergoing a drastic demographic decline (Forte et al., 2020). The study area is crossed by the Volturno River, which originates from the Matese massif, and its tributaries (Fig. 2). The most important tributary is certainly the Calore River, which has a hydrographic basin of 3050 km^2 and originates from the Picentini mountains, collecting the waters of the south-eastern sector of the basin. The Calore River flows into the Volturno River between Benevento and Caserta, 108 km from its spring. The complex hydrographic network of the Volturno River basin crosses rocks that present extreme compositional variability (Cicchella et al., 2023; Vitale and Ciarcia, 2018). These rocks are linked with the orogeny of the Apennines and can be grouped into three macro categories (Vitale and Ciarcia, 2013, 2018):

Pre-orogenic units include the carbonate platform, known as the Apennine platform, and the Lagonegrese Molise basin. The Apennine platform domain has a very simple lithology and is mainly made up of Mesozoic limestones and dolomites at the base and Chalcyclastic successions at the top. These rocks characterize most of the mountainous areas of the Volturno River Basin, such as the Matese, Picentini, Partenio, and Taburno mountains (Fig. 1). The geochemistry of Apennine platform rocks, and therefore also of soils and stream sediments deriving from their alteration, is very simple and show enrichment of Ca, Mg, and Sr compared to other rocks occurring in the central-southern Apennines (Ambrosino et al., 2022; Cicchella et al., 2022, 2023) The Lagonegrese Molise basin domain is defined by a late Mesozoic basinal sedimentary succession, rich in siliciclastic lithologies, including clays, marls, fine-grained sandstones, siliciferous clays and a minor calcareous component consisting of calcilutites and calcarenites. Most of these siliciclastic deposits crop out in the southeastern sector of the Volturno River basin (Fig. 1). According to their lithology, The geochemistry of Lagonegrese Molise basin rocks is very variable. However, recent studies (Cicchella et al., 2022, 2023) have shown that its geochemical signal consists of an enrichment in Ni, Co, Cr, and Mn associated with the presence of the clay-rich portion and an enrichment of Si, Ba, Al, Ga associated with siliciferous clays and marls. Their carbonate-rich portion shows the same geochemical association as the Apennine platform.

Synorogenic units (wedge-top or piggy-back basins) consist of sedimentary successions that can be grouped according to their age into Miocene sedimentary deposits and Pliocene sedimentary deposits. Miocene sedimentary deposits and Pliocene sedimentary deposits lie above the accretionary prism (made up of the pre-orogenic units). Therefore, they were formed due to the Apennine platform and Lagonegrese Molise basin disgregation. In the Volturno River basin, the synorogenic deposits outcrop mainly in the eastern and northern sector (Fig. 1). The Miocene sedimentary deposits are a basin environment succession consisting of turbidite sequences including calcilutides, marls, sandstones, clay (coming from Lagonegrese Molise basin disgregation) and carbonate (Apennine platform disgregation) olistostromes. Pliocene sedimentary deposits were formed in a continental-

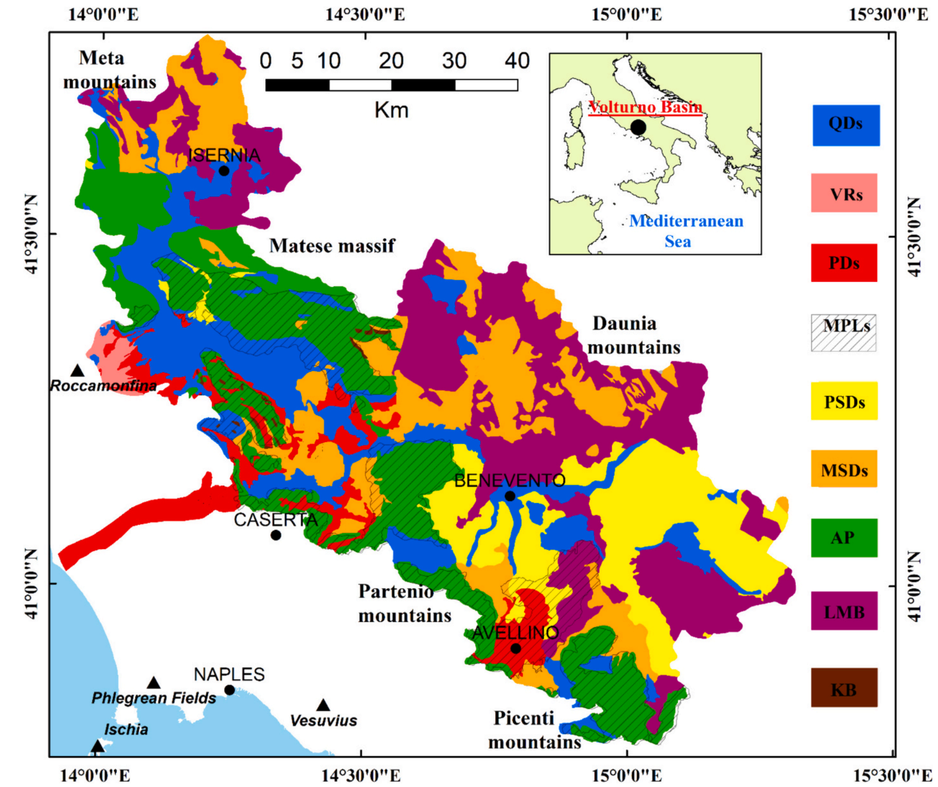

Fig. 1. Geological map (with underlying shaded relief) of the Volturno River basin (modified after Vitale and Ciarcia, 2022). QDs: Quaternary deposits, VRs: volcanic rocks, PDs: pyroclastic deposits, MPLs: minor pyroclastic layers, PSDs: Pliocene sedimentary deposits, MSDs: Miocene sedimentary deposits, AP: Apennine platform, LMB: Lagonegrese Molise basin, KB: karst bauxite.

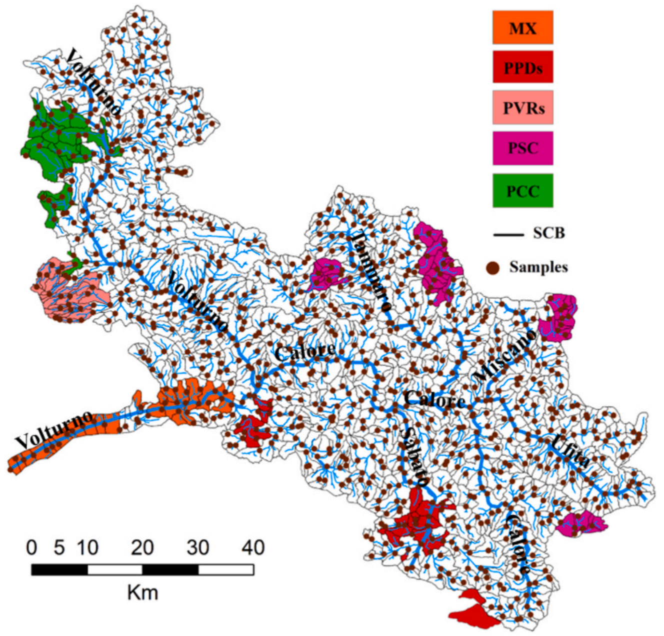

Fig. 2. Location of stream sediment samples and their sample catchment basin (SCB) outlines. Samples within the coloured areas were labelled as follows: mixed sediments (MX), predominantly pyroclastic deposits (PPDs), predominantly volcanic rocks (PVRs), predominantly siliciclastic component (PSC), and predominantly carbonate component (PCC).

transitional environment, consisting of polygenic conglomerates, shallow-water sands, calcarenites, silts, and clays. Both Miocene sedimentary deposits and Pliocene sedimentary deposits produce geochemical signals attributable to the Apennine platform or Lagonegrese Molise basin based on the paleogeographic domain that mostly fed the basin (Cicchella et al., 2022).

Post-orogenic units include Quaternary deposits, volcanic rocks connected with the activity of Roccaminfina Volcano (630–50 ka), and pyroclastic deposits related to the most recent activity of Ischia (150 ka – 1302 CE), Campi Flegrei-Procida (80 ka – 1538 CE) and Somma-Vesuvio (39 ka - 1944 CE) volcanoes. The pyroclastic products relating to the various eruptive steps affected large areas of the region (Bisson et al., 2007; Giaccio et al., 2008), but thick layers of tephra outcrop only in the south-western sector of the study area (Fig. 1). Minor pyroclastic layers are also present in areas far from volcanic buildings (di Gennaro et al., 2002). Their geochemical signal in the soils is evident up to >100 km from the sources (Ambrosino et al., 2022). All volcanic products are enriched in a suite of chemical elements, including K, Na, Ba, Be, U, Th, and Zr, compared to other rocks outcropping in the southern Apennines (Albanese et al., 2007; Ambrosino et al., 2022; Zuzolo et al., 2020). However, the same authors have also shown that volcanic rocks exhibit different chemical compositions in soils and stream sediments based on their age and, therefore, on their degree of weathering. The early volcanic products are enriched in low-mobility elements (e.g., Th, Zr, Ce, La, Y), while the late volcanic products are enriched in high-mobility elements (e.g., K, Na).

Other rocks in the Volturno River basin, unrelated to the Apennines' formation, are karst bauxite deposits of uncertain origin (Mondillo et al., 2011). These deposits are located in small outcrops scattered in the central-northern sector of the study area (Fig. 1).

3. Material and methods

3.1. Dataset

The dataset used includes geochemical data relative to 887 steam sediment samples collected along the whole hydrological system of the Volturno River basin with an average sampling density of about 1 sample every 6 km^2 (Fig. 2). To ensure that the samples used in this study are representative of the local geochemistry, artificial banks, channels, and levees were avoided during the sampling procedure. Each sample was made of composite material taken from five points over a stream stretch of 20–100 m. Finer grain size material (< 2 mm) was collected from the center of the stream beds avoiding, where possible, the collection of organic matter. The sampling procedure and the sample preparation followed the protocol described in detail by Albanese et al., 2007. Chemical analyses were carried out at ACME Analytical Lab. Ltd. (Vancouver, Canada). A total of 35 elements were determined for each sample, following a modified aqua regia digestion, combining inductively coupled plasma mass spectrometry (ICP-MS) and inductively coupled plasma emission spectrometry (ICP-ES).

To study only natural processes, we selected those chemical elements whose concentrations, according to previous studies (Ambrosino et al., 2022; Cicchella et al., 2022, 2023), were strongly related to lithology. Therefore, the dataset used in this study includes the concentration of 16 chemical elements (Al, Ba, Ca, Co, Fe, Ga, K, La, Mg, Mn, Na, Ni, P, Sr, Th, Ti) for which details regarding analytical quality and raw data distribution structure are reported in Table 1.

3.2. Data preparation

A vector map of sample catchment basins was generated using the ArcGIS 10.8 spatial analyst tool (Fig. 2) by using a regional digital elevation model and the stream sediment sample locations as inputs. The digital elevation model allowed us to determine flow direction and accumulation paths corresponding to the river basin's active segments and primary runoff directions. Sample locations were used along the accumulation paths as virtual outlets of catchments limited upstream by the next sample. Each sample catchment basin corresponds to the upstream area which mainly influences the composition of the related sample at its outlet (Carranza, 2009; Dominech et al., 2022).

Subsequently, to recognize the geochemical fingerprints on sediments of the main lithological associations occurring in the study area, we selected from the dataset those samples having their sample catchment basin meeting the following requirements: i) to belong to a geological unit consisting of a dominant lithology (pre-orogenic units or volcanic products), ii) to cover a single geological unit, iii) to be placed in the upper or middle course of the Volturno River basin to limit the effect of sediment mixing, iv) to be featured by a chemical composition characteristic of a given lithological domain. Through this labelling criterion, samples were selected and grouped in four domains (end-members): i) predominantly pyroclastic deposits (PPDs), referring to sample catchment basins lying on pyroclastic deposits; ii) predominantly volcanic rocks (PVRs) referring to those lying on volcanic rocks; iii) predominantly siliciclastic component (PSC) referring to those lying on Lagonegrese Molise basin and iv) predominantly carbonate component (PCC) referring to those lying on Apennine platform. Additionally, acknowledging that the physical displacement of sediments can result in highly mixed products, we labelled the samples of mixed sediments to identify their geochemical signature. It is known that extensive sediment mixing is prevalent in downstream regions, particularly at the junction of two rivers (Lane et al., 2008; Umar et al., 2018). This level of mixing significantly influences the geochemical signature of stream sediments (Lipp et al., 2021; Caracciolo, 2020). Thus, samples falling along the highest-order streams in the downstream area, namely after the junction between the Calore and Volturno rivers, were manually labelled to form an additional reference group of mixed sediments (featuring the "MX" label). To have a similar number of samples in each group (which is convenient for the classification algorithm applied below), we did not select all the samples that satisfied the above criteria, especially in the case of those geological units widely present in the study area (e.g., Lagonegrese Molise basin). Furthermore, samples lying on Pliocene

Table 1

Summary statistics of chemical components (in mg/kg) in the studied area.

	Al	Ba	Ca	Co	Fe	Ga	K	La	Mg	Mn	Na	Ni	P	Sr	Th	Ti
[*]**Accuracy (%)**	0	0.3	3.9	0	0.7	3.2	6.3	3.5	0	0.5	3.6	0.6	0	5.3	5.1	0
Precision (%RPD)	1.8	1.5	2.2	2.7	1.3	2.2	5.3	3.4	1.5	1.9	2.9	1.7	3.6	2.4	3.6	5.7
IDL	100	0.5	100	0.1	100	0.1	100	0.5	100	1	10	0.1	10	0.5	0.1	10
Min	2900	27	3100	1.9	3300	1	400	1.5	500	115	75	2.5	140	29	0.8	10
Q25	10,925	117	45,100	8.6	15,825	3	2000	12.1	3600	669	200	15.1	540	124	4.1	100
Mean	19,104	197	75,392	12.1	21,818	5	3800	24.1	7387	1031	657	24.2	845	183	8.1	639
Median	15,300	163	67,150	11.2	19,900	4	2800	17.7	5000	888	310	22.1	680	172	5.9	290
Geometric mean	16,134	167	62,020	11.1	20,031	5	2998	19.3	5476	908	387	21.6	741	163	6.4	297
Q75	23,800	241	98,850	14.4	26,100	6	4100	30.3	7200	1189	620	30.8	950	219	9.3	970
Q95	44,715	430	163,345	22.4	39,030	12	10,600	64.2	18,715	2113	2209	45.9	1820	342	22.1	2189
Max	78,400	1429	271,900	59.3	69,200	15	38,200	120.6	88,100	5599	6700	84.8	6700	823	71	4520
MAD	9086	89	34,056	3.9	6914	2	2065	13	4572	402	556	9.3	349	65.4	4.6	595
Skewness	1.4	2.3	1.1	1.9	1.2	1.2	3.7	1.9	4.9	2.7	7.6	1.1	3.5	2.1	2.9	1.7
rCV (%)	59.4	54.6	50.7	34.8	34.7	45.5	73.8	73.4	91.4	45.3	179.4	42.1	51.3	38.0	78.0	205.2

[*] Accuracy, precision and instrumental detection limits (IDL) of the applied analytical method (RPD: relative percent difference).

sedimentary deposits, Miocene sedimentary deposits, and Quaternary deposits were not considered due to their lithological inhomogeneity. Finally, a selection of samples was manually assigned to the groups defined above (Fig. 2): 21 to the mixed sediments (featuring the "MX" label), 22 to the predominantly carbonate component (featuring the "PCC" label), 21 to the predominantly siliciclastic component (featuring the "PSC" label), 24 to the predominantly volcanic rocks (featuring the "PVRs" label), and 16 to the predominantly pyroclastic deposits (featuring the "PPDs" label).

3.3. *Obtaining compositional indicators for different geochemical signatures*

Compositional Data Analysis relies on log-ratio coordinate representations of the original geochemical compositions. Unlike with raw data, using log-ratios generally leads to results that do not depend on the scale of measurement of the data or the size of the composition. Thus, orthonormal log-ratio coordinates (olr; a.k.a. isometric log-ratio coordinates) allow the consistent projection of the information contained in a composition into the ordinary real space and facilitate statistical analysis. The orthonormality property guarantees that distances and other measures of differences between compositions are preserved by their log-ratio counterparts. Given a composition of D parts $\mathbf{x} = (x_1, \ldots x_D)$, we will define compositional indicators (CIs) from olr-coordinates in the form of *balances* (Egozcue and Pawlowsky-Glahn, 2005), which have general expression given by

$$\sqrt{\frac{r \bullet s}{r+s}} ln \left[\frac{\left(\prod_{h=1}^{r} x_h^{+}\right)^{\frac{1}{r}}}{\left(\prod_{l=1}^{s} x_l^{-}\right)^{\frac{1}{s}}} \right] \tag{1}$$

and represent contrasts between the geometric means of mutually exclusive subsets of r and s parts placed, respectively, into the numerator and denominator of the log-ratio term (+ and – superscripts). Note that this is equivalent to the normalized aggregation of all possible pairwise log-ratios between parts from both groups: $\sqrt{\frac{1}{rs(r+s)}} \sum_{h=1}^{r} \sum_{l=1}^{s} ln \frac{x_h^{+}}{x_l^{-}}$. Hence, depending on the parts going into the + and − groups and the relative weight of one group against the other, these balances will be meaningful as compositional indicators of phenomena such as pollution, mineralized veins, or lithologies. For instance, let us consider a balance that confronts elements typically found in high concentration in volcanic soils and low concentration in carbonate soils (group +) against elements showing the opposite pattern (group −). This could, therefore, be a relevant compositional indicator of soil parental material, where volcanic and carbonate rocks occur, with high values indicating the volcanic parental material and low values doing so for carbonatic parental material.

3.3.1. *Computation of compositional indicators through principal balances*

A common procedure to compute balances (Eq. 1) is through a sequential binary partition (SBP) of the original composition: from the original composition, a hierarchical structure is defined through successive nested splits into two mutually exclusive subsets of parts until no more splits are possible. This results in a system of D-1 balances, one per split, fully representing the information in the original D-part composition (Egozcue and Pawlowsky-Glahn, 2005). Respecting the SBP rules, the splits can be predefined based on expert knowledge so that the resulting balances (at least one of them) are meaningful as compositional indicators (see, e.g. Boente et al., 2022 or Liu et al., 2018).

The following table illustrates an example of SBP used to generate balances that distinguish between volcanic (rich in Na, Ti, La, and Th) and non-volcanic (rich in Ca, Mg and Ni) soils in the study area. In this example, the first balance could be used as a compositional indicator of the bedrock type, while the second one would relate to the degree of weathering. The additional balances derived via SBP, such as those involving La, Th, and Ti, lack a geochemical significance that is indicative of a specific phenomenon; consequently, they are not designed to serve as compositional indicators.

Alternatively, a structure of balances can be inferred from available data using some criterion (data-driven approach). Amongst the latter, the method of principal balances constructs an SBP system so that the corresponding collection of D-1 (principal) balances successively maximizes the fraction of the original data variance explained (Martín-Fernández et al., 2018). Principal balances approximate the properties of ordinary principal components analysis, but they generally improve interpretability, since they do not necessarily involve all original parts. As the number of parts of the studied composition increases, the exhaustive search for optimal principal balances can be computationally intensive. Thus, some algorithms have been proposed to ease this task at the expense of losing some explanatory power. We will use here the constrained method described in Martín-Fernández et al. (2018).

3.3.2. *Handling different geochemical signatures with common elements*

By construction, the SBP procedure described above is limited when intending to define multiple compositional indicators that share common elements. For example, we might be interested in having two compositional indicators to distinguish, respectively, different parental materials and the weathering degree of soils. The first compositional indicator should contrast the elements enriched in the volcanic parental material (e.g. Al, Na, La, Th) to the carbonate parental material (e.g. Ca, Mg). The second compositional indicator should contrast the low mobility elements (e.g. Th, Al, La), enriched in soils and exhibiting a high degree of weathering, to the high mobility elements (e.g. Ca, Na), enriched in soils and exhibiting a low degree of weathering. However, these two compositional indicators cannot be obtained from a single SBP, such as the one shown in Table 2, as the ratio of Al and Th to Ca cannot appear more than once. Thus, once the compositional indicator of the parental material has been defined (see the first balance of Table 2), it is not possible to obtain that defining the weathering degree since the second balance of the SBP in Table 2 can contrast only elements that are present in the numerator or denominator of the first balance. This issue can be overcome by conveniently picking out compositional indicators from different SBP systems and adequately combining them into the data analysis. We thus apply a hybrid approach to build compositional indicators of natural sources, comprising: (1) selection of a meaningful composition (after removing primary pollutants) and labelling reference samples based on expert knowledge (Section 3.1; knowledge-driven step); and (2) using such curated data to objectively inform the definition of compositional indicators through principal balances which stress the features of interest (data-driven step). For this, the data were first split into three subsets of samples distinguishing pairs of geochemical signatures:

i) All labelled samples (ALS), including the predominantly pyroclastic deposits, the predominantly volcanic rocks, the predominantly carbonate component, and the predominantly siliciclastic component;

ii) Volcanic labelled samples (VLS), including the predominantly pyroclastic deposits and the predominantly volcanic rocks;

iii) Sedimentary labelled samples (SLS), including the predominantly carbonate component and predominantly siliciclastic component.

Table 2
Exemplary balances obtained through a knowledge-driven sequential binary partition. The symbols (+ and −) indicate elements placed in the numerator and denominator of the balances, respectively.

Balance	Na	Ti	La	Th	Ca	Mg	Ni
Volcanic vs. non-volcanic	+	+	+	+	−	−	−
Low mobility vs. High mobility	−	+	+	+			
La, Th vs. Ti		−	+	+			
La vs. Th			−	+			
Ca, Mg vs. Ni					+	+	−
Ca vs. Mg					+	−	

Secondly, three compositional indicators were separately computed from each subset and denoted by CI_{ALS}, CI_{VLS} and CI_{SLS} labels, respectively, for reference. Technically, they corresponded with the first balances of the respective three SBP systems obtained by the method of principal balances (i.e., the compositional indicators corresponded with the balances between elements explaining the highest fraction of variation in each subset).

3.3.3. Joint compositional biplot display

Compositional biplots-based principal component analysis (PCA) of centred log-ratio transformed data are ordinarily used to jointly display samples and compositional parts (i.e. the geochemical elements) in low dimensions to facilitate interpretation (Aitchison and Greenacre, 2002). The biplot display is defined from a matrix of *loadings* (determining the length and orientation of rays representing the compositional parts) and a matrix of *scores* (determining the coordinates of the samples in the low-dimensional, typically 2D, space). To devise such a graphical representation including all the collected samples along with the three compositional indicators (CI_{ALS}, CI_{VLS} and CI_{SLS} associated with three different principal-balance-based SBP systems), we adapted the strategy depicted in Štefelová et al. (2023) to combine balances from different but orthogonal coordinate systems into a single biplot display. In brief, this involved firstly applying each of the three SBPs (learned from the ALS, VLS and SLS subsets above) to the entire data set, including labelled and unlabelled samples. Note that thanks to the orthogonality between SBP coordinate systems, all of them led to the same matrix of scores (which coincides with that of the standard compositional PCA biplot). Then, the vectors of loadings associated with each one of the compositional indicators (first principal balances) were extracted and combined into an ensemble loadings matrix. Lately, such ensemble loadings matrix and the common matrix of scores were used as input to build a biplot display.

3.4. Samples classification and assessment of geochemical baselines

As will be shown below in the “Results and discussion” section, the reference samples in the five groups of labelled sediments showed notable differences in compositional indicator values. It made sense to use them to build a classification model to allocate the remaining unlabeled samples. Thus, labelled samples served as input to train a random forest classification algorithm (Biau and Scornet, 2016), using compositional indicators as predictor variables. In brief, a random forest model iteratively fits many decision tree classifiers to random subsamples and subsets of variables, using averaging to reduce over-fitting and improve predictive accuracy when assigning new samples to groups. Model training, including tuning of model parameters, was conducted through a 5-time repeated 5-fold cross-validation (CV) pipeline. Thus, the input data were randomly partitioned into five folds, with four of them used to train the model and one used as a validation set, sequentially. This fold randomization was repeated five times, contributing to a fairer assessment of the predictive ability of the model with independent unlabeled samples (99 % CV accuracy reached). This predictive analysis was implemented using the “Caret – v.6.0-94” R package (Kuhn, 2008).

Following the random forest classification process, each of the unlabelled samples in the dataset was assigned to one of the five groups including the four geolithological domains (i.e., PPDs, PVRs, PSC, PCC) and the mixed sediments (MX), respectively.

Finally, the upper baseline limits (UBLs) were determined for the considered chemical elements and the distinct groups of stream sediments using the US EPA's (United States Environmental Protection Agency) ProUCL 5.1.0 software package (Singh and Maichle, 2015). This approach was applied to the raw data following recent applications on soils (Meloni et al., 2023) and stream sediments (Cicchella et al., 2022). Specifically, following a Rosner test to eliminate outliers from the single group dataset, the upper baseline limits were determined for each element by using the upper tolerance limit at the 95 % confidence limit of the 95th percentile of the data distribution (UTL95–95 value) as a reference (Singh and Maichle, 2015).

4. Results and discussion

4.1. Exploratory analysis

Fig. 3 shows the compositional PCA biplot of the entire data set. Coloured points are used to distinguish the original labelled samples selected as representative of the lithological domains and thin rays represent the geochemical elements. The first two axes (PC1 and PC2) explain 68.8 % of the total variability of the data, which is consistent with similar analyses in other studies (Reimann et al., 2012; Buccianti et al., 2014; Cicchella et al., 2023). The compositional biplot suggests the existence of four geochemical associations (groups of elements showing moderate to high proportionality between them), broadly arranged according to the four biplot quadrants. This is further corroborated by analyzing the co-dependence structure between elements based on their pairwise log-ratio variances (see Supplementary Material: Fig. SM1). Interestingly, the centroids of the labelled samples are also distributed according to the biplot quadrants, thus suggesting a predominant influence of the selected lithological units on the geochemical composition of the stream sediments.

The PC1 axis separates the predominantly pyroclastic deposits and volcanic rocks domains (linked to the enrichment of Ti, Na, Th, La, K, Al, Ga, Ba) (with positive loadings), from the predominantly siliciclastic and carbonate component domains (linked to the enrichment of Ca, Mg, Ni, Co, Mn, Fe, Sr, P) (with negative loadings). The PC2 axis separates the predominantly volcanic rocks and siliciclastic component domains (linked to the enrichment of Mn, Th, Ni, Co, La, Fe, Ba, Al, Ga) (with positive loadings) from the predominant carbonate component and pyroclastic deposits domains (linked to Ca, Mg, K, P, Sr, Na, Ti) (with negative loadings). These results suggest that PC1 separates the samples according to their parental material (volcanic or sedimentary rocks), whereas PC2 separates the samples according to their weathering degree. This latter observation is more significant when considering the different mobility of the elements separated by PC2. In fact, except Ti, the elements with positive loadings for PC2 have low mobility under environmental conditions, while those with negative values have high mobility (Reimann et al., 2014).

It is observed that some of the labelled samples are located notably far from the centroid of their group; particularly, some samples of the predominant carbonate component domain appear fairly close to the centroid of the predominant pyroclastic deposits domain; this occurrence is most probably attributable to a labelling error as the pyroclastic covers are dispersed throughout the study area and lie mainly on the lithologies belonging to the Apennine platform (i.e., Ca-rich lithologies). Therefore, even if the areas where the pyroclastic covers mainly occur have been defined, their presence cannot be excluded in areas proximal to these limits (Bisson et al., 2007; Giaccio et al., 2008). Furthermore, the compositional biplot shows a few samples of the predominantly siliciclastic component and volcanic rocks domains falling far away from their cluster centroids, and located around the predominantly pyroclastic deposits domain; as a consequence, two considerations arise: i) the pyroclastic deposits represent the primary chemical natural contaminant of the lithological domains present in the study area, and ii) using only geological and geomorphological evidence to select samples with a similar chemical composition can lead to errors. Although the labelling performed on a geological basis allowed us to interpret differences in the data, not all labelled samples can be considered compositional end-members. To avoid the propagation of this error in calculating the compositional indicators, these presumably wrongly labelled samples were deemed unlabeled in Fig. 3.

4.2. Compositional indicators and classification of unlabeled samples

Once the compositional end-members were selected, compositional

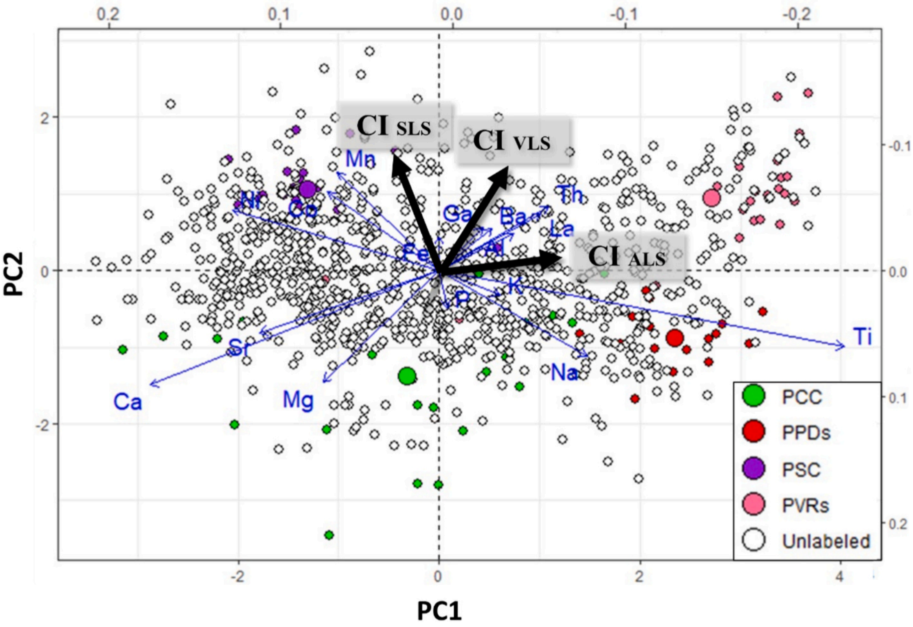

Fig. 3. Compositional PCA biplot display including projection of compositional indicators (CIs) computed according to subsets of labelled samples (ALS, VLS, and SLS, see text for details). Labelled samples showed in color: PCC, predominant carbonatic component; PPDs, predominant pyroclastic deposits; PSC, predominant siliciclastic component; PVRs, predominant volcanic rocks.

indicators were calculated. The resulting CI_{ALS} (Eq. 2), based on all labelled samples, contrasts elements that are more abundant in volcanic parental materials (predominantly pyroclastic deposits and volcanic rocks) (in the numerator) against elements that are more abundant in sedimentary parental materials (predominantly carbonate and siliciclastic components) (in the denominator):

$$CI_{ALS} = \sqrt{\frac{12}{7}} ln\left(\frac{(\mathrm{Na} \bullet \mathrm{Ti} \bullet \mathrm{La} \bullet \mathrm{Th})^{\frac{1}{4}}}{(\mathrm{Ca} \bullet \mathrm{Mg} \bullet \mathrm{Ni})^{\frac{1}{3}}}\right) \tag{2}$$

Therefore, CI_{ALS} can be understood as a compositional indicator of the parental material, where high values correspond to volcanic parental material and low values correspond to sedimentary parental material. It is obvious that this indicator cannot be applied worldwide (considering that Ni can be abundant in ultramafic rocks), but it can be applied at least to all areas of the Roman Comagmatic Province in Italy, which includes most of the Latium and Campania regional territories.

The CI_{VLS} (Eq. 3), obtained using only samples with a predominant volcanic component (predominantly pyroclastic deposits and volcanic rocks), contrasts low-mobility elements (in the numerator) against high-mobility elements (in the denominator):

$$CI_{VLS} = \sqrt{\frac{20}{9}} ln\left(\frac{(\mathrm{Ti} \bullet \mathrm{Ga} \bullet \mathrm{La} \bullet \mathrm{Mn} \bullet \mathrm{Th})^{\frac{1}{5}}}{(\mathrm{Ca} \bullet \mathrm{Mg} \bullet \mathrm{K} \bullet \mathrm{Na}\ )^{\frac{1}{4}}}\right) \tag{3}$$

Hence, CI_{VLS} can be interpreted as an indicator of the degree of weathering. Considering that the selected end-members come from the same magmatic province, their most significant differences can be associated with their alteration degrees. Higher values of CI_{VLS} indicate samples with an advanced weathering degree, whereas lower values indicate less alteration. In this respect, it is essential to note that Ti, together with other low mobility elements (i.e., Ga, La, Mn, Th), is associated with a higher weathering degree, unlike what PC2 in the compositional biplot suggests (Fig. 3). However, note that the direction of the Ti axis in the biplot is influenced by other data populations, such as that related to the predominantly carbonate component domain, which is enriched in high mobility elements (i.e., Ca and Mg) but also has high Ti concentrations.

The third indicator, CI_{SLS} (Eq. 4), was obtained using only samples with a predominantly sedimentary component (carbonate and siliciclastic components).

This indicator contrasts elements more abundant in siliciclastic-rich sediments (in the numerator) against elements more abundant in carbonate-rich sediments (in the denominator):

$$CI_{SLS} = \sqrt{\frac{12}{7}} ln\left(\frac{(\mathrm{Fe} \bullet \mathrm{Co} \bullet \mathrm{Mn} \bullet \mathrm{Ni})^{\frac{1}{4}}}{(\mathrm{Ca} \bullet \mathrm{Mg} \bullet \mathrm{Ti})^{\frac{1}{3}}}\right) \tag{4}$$

The involvement of Ti in CI_{SLS} confirms its higher concentration in the predominantly carbonate component compared to the siliciclastic one. However, this might only be a feature of the study area, where carbonate rocks are often in spatial association with pyroclastic covers. Titanium allows us to better distinguish between predominantly carbonate and siliciclastic component domains for this research, confirming the versatility of the method. Therefore, CI_{SLS} can be considered an indicator of the siliciclastic component (fine-grained sandstones, marls, and clays of the Lagonegrese Molise basin domain), with higher values corresponding to an enrichment of siliciclastic material and lower values corresponding to an enrichment of carbonate-rich material.

A feature of the proposed method is that it allows the recognition of elements (e.g., Al, Ba, P, Sr) showing slightly variable concentrations between them that are unimportant in separating end-members. Amongst these elements, the case of Al is worth noting since it is often used to assess the degree of alteration, for instance using the chemical index of alteration.

In general, our results agree with Cicchella et al. (2023), where it was shown that volcanic soils always exhibited a chemical index of alteration >60 in the study area, whereas soils developed on sedimentary bedrock rarely exceeded these values. This evidence suggests that a parental material with higher concentrations of Al can generate a higher chemical index of alteration, regardless of the degree of alteration. Furthermore, the elements present in the numerator of CI_{VLS} (Ti, Ga, La, Mn, Th) coincide with those indicated in Ambrosino et al. (2022) as the main proxies for the degree of alteration. Therefore, the absence of Al in CI_{VLS} confirms that this element should not be used as a proxy for the degree of alteration in the study area.

Rays indicating the direction of increasing values of the three compositional indicators were projected onto the compositional biplot shown in Fig. 3, thus providing further insight and facilitating joint interpretation. This biplot display allows us to evaluate the relative position of the samples according to their compositional indicators and their predominant parental material and degree of alteration. Furthermore, Fig. 4A and B show the samples in scatterplots according to the three compositional indicators (CI_{ALS}, $\text{CI}_{\text{VLS,}}$ and CI_{SLS}) and the labels of the reference samples. These samples were used to train a random forest model to assign unlabeled samples (indicated in light grey in the graphs) to one of the lithological groups used (MX, PCC, PPDs, PSC, and PVRs). The resulting classification of all samples based on the compositional indicators as predictors is represented in Fig. 4C and D. For each compositional indicator, regions including borderline samples, where some overlapping of different geochemical domains can occur, are highlighted by vertical and horizontal segments, including the corresponding boundary compositional indicator values. For CI_{ALS}, two segments identify the transition between sedimentary and mixed domains (from -4.5 to -4.4) and between mixed and volcanic domains (from -3.7 to -3.6). For CI_{SLS}, a horizontal segment (from 3.6 to 3.4) represents the transition from predominantly siliciclastic to carbonate component domain. Regarding CI_{VLS}, the segment shown (from 5.6 to 5.4) separates predominant pyroclastic deposits from volcanic rock domains. Samples farther apart from these regions are classified across the five recognized geochemical domains with greater certainty. Thus, the boundary values could be used as a reference for the characterization of future samples collected in the study area and their allocation into one of the five geochemical domains. It is worth highlighting that calculating the values of the three compositional indicators (involving just nine chemical elements) makes it feasible to classify new samples of stream sediments and get insight into their geochemical baselines (Table 3).

According to the above partition, the predominant volcanic rocks domain is associated with $\text{CI}_{\text{ALS}} > -3.6$ e $\text{CI}_{\text{VLS}} > -5.4$. The low number of samples it includes reflects the spatial distribution of the volcanic products of the Roccamonfina volcano, in agreement with the geological map (Fig. 1). The predominant siliciclastic component domain corresponds to $\text{CI}_{\text{ALS}} < -4.5$ and $\text{CI}_{\text{SLS}} > -3.4$. The high number of samples allocated to this geochemical domain can be justified by the abundance of clayey material in the Miocene sedimentary deposits and Lagonegrese Molise basin. The predominant carbonate component domain is associated with $\text{CI}_{\text{ALS}} < -4.5$ and $\text{CI}_{\text{SLS}} < -3.6$. It reflects the broad spatial distribution of the Apennine platform and the Pliocene deposits, the geological units most enriched in the carbonate component. The predominant pyroclastic deposits domain corresponds to $\text{CI}_{\text{ALS}} > -3.6$ and $\text{CI}_{\text{VLS}} < -5.6$. The relative frequency of samples allocated to it is somehow more extensive than expected for the reduced spatial distribution of the volcanic products reported in the geological map (Fig. 1). However, this result can be explained by the fact that the geological map shows pyroclastic covers only where they outcrop in high thicknesses. The MX domain is associated with $-4.4 < \text{CI}_{\text{ALS}} < -3.7$, and the non-negligible number of samples included in it reflects the complex geological setting of the study area.

The relative frequency distribution of samples allocated to each domain can be arranged in decreasing order as PSC (32.4 %) > PCC (28.5 %) > PPDs (17.9 %) > MX (14.4 %) > PVRs (6.8 %).

4.3. *Spatial distribution of clusters and their geochemical baselines*

The spatial distribution of the geochemical domains (Fig. 5A) was obtained by combining the classification results above with the sample catchment basin information. Thus, the catchment area of each sample was labelled according to the classification results. The map obtained shows substantial differences from the geological map in Fig. 1. The predominant carbonate component domain covers areas linked to the Apennine platform, Quaternary deposits and Pliocene sedimentary deposits. Although these latter originate from different environments and exhibit considerable diachronism, their chemical composition is similar. In the Apennine platform and Pliocene sedimentary deposits, the enrichment in carbonate minerals is explained by their lithology, which is rich in carbonate minerals. This finding is consistent with Ciarcia and Torre (1996), which shows that the Pliocene sedimentary deposits consist of a suite of sedimentary rocks with a carbonate matrix and an abundance of calcareous clasts up to 87 %. Concerning the Quaternary deposits, enrichment in the carbonate component should be due to the precipitation of calcium carbonate, which occurs in the active depositional zones even for a low state of water saturation (Manzo et al., 2012).

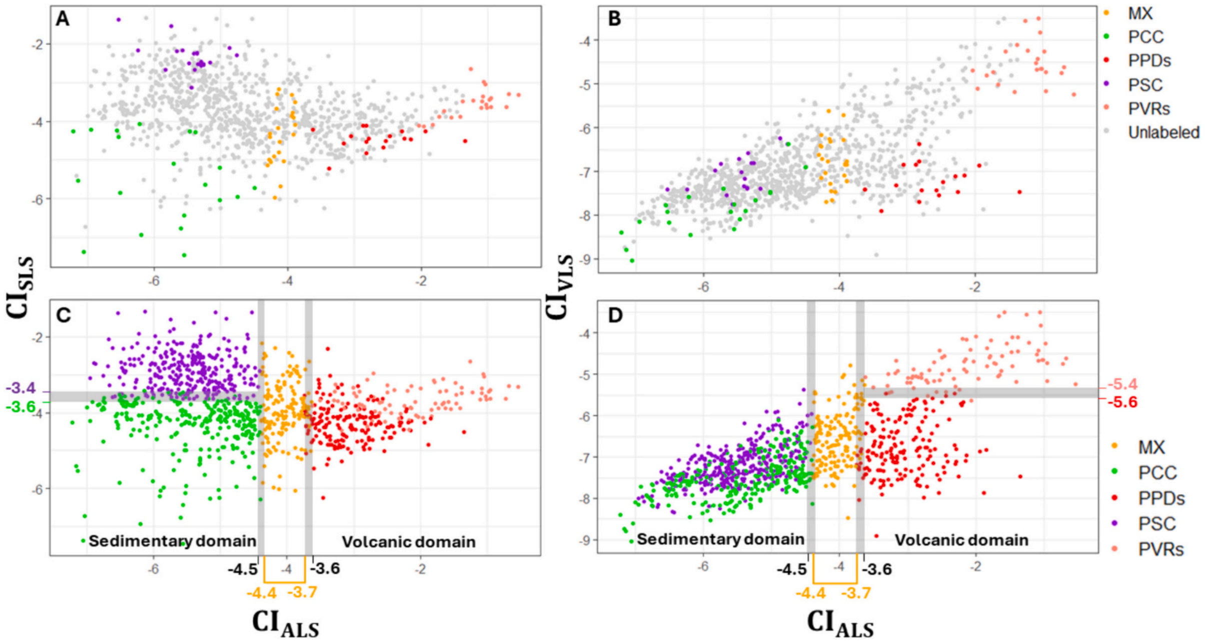

Fig. 4. Reference samples used to train the random forest model (A and B) based on the three compositional indicators (CI_{ALS}, $\text{CI}_{\text{VLS,}}$ and CI_{SLS}) and predicted allocations for all samples (C and D). Vertical and horizontal segments indicate approximate geochemical domain boundaries and associated thresholds. MX: mixed; PCC: predominant carbonate component; PPDs: predominant pyroclastic deposits; PSC: predominant siliciclastic component; PVRs: predominant volcanic rocks.

Table 3
Upper background levels calculated using ProUCL for each defined geochemical domain. MX: mixed sediments, PCC: predominant carbonate component, PVRs: predominant volcanic rocks, PPDs: predominant pyroclastic deposits, PSC: predominant siliciclastic component.

Geochemical domain	Al	Ca	Fe	K	Mg	Na	P	Ti	Ba	Co	La	Mn	Ni	Sr	Th
MX	39,351	126,389	36,030	7498	17,520	1415	1740	1597	382	22	56	2172	41	256	15
PCC	21,854	195,739	23,704	3874	38,945	550	1424	654	215	12	27	1190	31	390	7
PVRs	60,490	51,419	55,296	5267	7613	1720	2095	3461	525	19	126	2917	36	286	50
PPDs	58,739	91,094	46,981	24,142	14,391	6847	2889	2917	730	17	61	1545	25	340	20
PSC	22,095	160,474	32,843	4468	9718	544	1034	198	267	37	28	2494	58	398	9

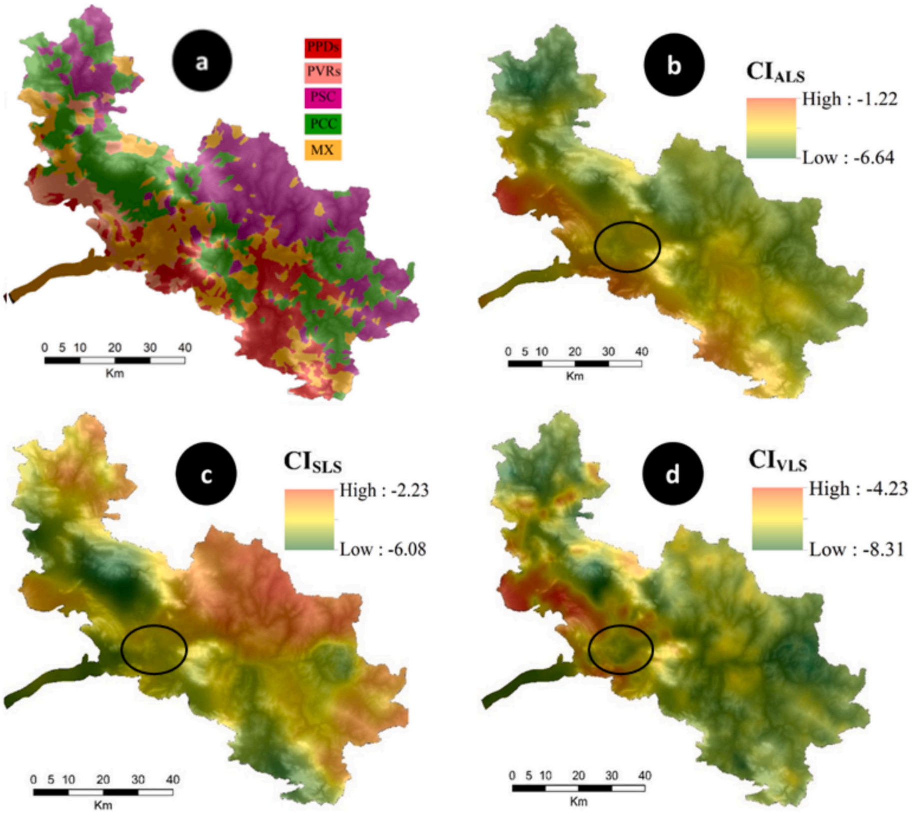

Fig. 5. Spatial distribution of the defined geochemical domains (A) and compositional indicators (B, C, D). Ovals indicate areas in which all compositional indicators exhibit intermediate values, which in the downstream areas can be attributed to a high sediment mixing.

The predominantly volcanic rocks domain covers a relatively small area near the Roccamonfina volcano and falls within the Apennine platform, Quaternary deposits, and volcanic rocks geological units.

In the area surrounding the volcano, the enrichment of Ti, Ga, La, Mn, and Th in the Quaternary deposits and Apennine platform geological units could be attributable to old pyroclastic covers. In the distal areas, the few samples classified as predominantly volcanic rocks (located around the Matese Massif) are mainly located in the zones where bauxitic deposits are present. Here, the presence of the predominantly volcanic rocks domain could be associated with residual soils and old pyroclastic covers. The predominantly pyroclastic deposits domain appears throughout the study area but is primarily present in the central-southern sector, where the thicknesses of the volcanic fall deposits are quite relevant (Bisson et al., 2007; Giaccio et al., 2008). The predominantly pyroclastic deposits pattern allows us to recognize the areas where the pyroclastic covers are currently present, improving the information provided by the geological map and the volcanological models. The geological map indicates only the areas where the pyroclastic covers are present in considerable thicknesses, while the volcanological models reconstruct the thicknesses of the volcanic fall deposits immediately after the eruption, thus ignoring the subsequent erosion and transport phenomena. The predominantly siliciclastic component domain is present in the eastern and northern sectors of the Volturno River basin and is located in the Lagonegrese Molise basin and Miocene sedimentary deposits geological units. This result suggests that the Miocene sedimentary deposits were mainly fed by siliciclastic material from the Lagonegrese Molise basin, while the Pliocene sedimentary deposits were made by carbonate rocks from the Apennine platform unit. The mixed domain includes sediments with a geochemical signature between volcanic and sedimentary products. In the downstream areas (central-western sector), the mixed domain is the result of mixing between distal sediments from the upstream areas (mainly of a sedimentary nature) and proximal sediments of volcanic nature. In the upstream areas, sediment mixing does not reflect the transport of distal sediments but reveals the presence of weathered pyroclastic deposits mixed with the sedimentary bedrock. Far from the volcanoes (eastern sector), the mixed domain is associated with a low degree of weathering, which allows for the preservation of the chemical signal of the thin pyroclastic levels that are deposited. In the areas proximal to the volcanoes (southern sector), the mixed domain is associated with the high slopes of the topographic peaks (Picentini mountains), where the morphological structure and the high weathering did not allow a high accumulation of pyroclastic material, which partially mixed with the bedrock (Celico and Guadagno, 1998).

Interpolated distribution maps of the three compositional indicators (CI_{ALS}, CI_{VLS}, and CI_{SLS}) were generated in the ArcGIS package using ordinary kriging interpolation (Cressie, 1992; Isaaks and Srivastava, 2010) (Fig. 5B-D) to highlight the geochemical fingerprint of stream sediments. Due to the irregular distribution of the samples, the lag size (7.4 km) was chosen according to the built-in average nearest neighbour method.

The pattern generated by CI_{ALS} shows that volcanic products (high CI_{ALS} values) are present throughout the western and central sectors of the Volturno River basin. The CI_{ALS} values are higher in the western margin, where the sediments result exclusively from volcanic material, and relatively lower in the central sector, where there is a greater mixing with sedimentary rocks. The lowest values of CI_{ALS} occur in the northern sector and at the highest altitudes of the Matese Massif. In these areas, the pyroclastic covers could be completely eroded and their geochemical signal erased. The CI_{SLS} pattern shows that the geochemical signal produced by carbonate rocks (low CI_{SLS} values) is strong only in the Matese Massif, the Picentini mountains, and the Volturno River plain. In the Volturno River plain, the high CI_{SLS} values are instead attributable to chemical precipitation phenomena that favour the deposition of carbonate minerals. The pattern generated by CI_{VLS} allows for highlighting the areas with the highest degree of weathering associated with an accumulation of elements with low mobility. The highest values are found in the central-western sector, where the lavas of the Roccamonfina volcano, the bauxitic deposits, and, probably, the oldest volcanic covers outcrops. The CI_{VLS} also allows the recognition of the different degrees of alteration occurring within the same geochemical domain. For example, CI_{VLS} reveals that the predominantly siliciclastic component domain has a low degree of weathering where Pliocene sedimentary deposits occur (south-eastern sector) and relatively higher degree where Miocene sedimentary deposits occur (central-eastern sector), which is in agreement with Cicchella et al. (2023). The sediments with the lowest degree of weathering are scattered throughout the southern and northern sectors of the study area. Combining the information from the compositional indicator maps with the geochemical domain map it is possible to improve the characterization of the obtained geochemical domains. In fact, it is possible to recognize that the predominantly pyroclastic deposits domain is enriched in the carbonate component (low CI_{SLS} values) in the southern sector, probably due to slight contamination with the calcareous bedrock. The mixed domain shows an enrichment of the carbonate component in the Volturno river plain (western sector), attributable to a greater precipitation of carbonate minerals. The predominantly carbonate component domain presents a greater siliciclastic component (higher CI_{SLS} values) in the northern sector, due to the more diffuse presence of marl compared to the Matese massif which is dominated by limestone and dolomite. Moreover, the predominantly siliciclastic component domain presents the above-discussed difference in the degree of weathering, highlighted by CI_{VLS}.

The upper baseline limits (Table 3) calculated for the five defined geochemical domains show notable differences for all the elements considered. Specifically, elements featuring high variability in values are Ca, Mg, K, Na, Ti, Co, and La. For example, upper background limits of Co and Na in predominantly carbonate and siliciclastic components domains (Table 3) are even lower than their respective mean values in the entire data population (Table 1), thus highlighting the usefulness of determining geochemical domains before assessing any reference value. The effectiveness of the proposed method is underscored by the lower skewness and robust coefficient of variation (rCV), exhibited by the individual variables within each geochemical domain when compared with the entire dataset. For more details, please refer to the supplementary material (Table SM1).

Furthermore, it is also noteworthy that the sediments of the predominantly pyroclastic deposits domain exhibit high values of the upper baseline limits for K, Na, and P, which are indispensable macronutrients for the development of many plant species (Havlin, 2020; Osman, 2013). Finally, the upper baseline limits of Co and Ni in the predominantly siliciclastic component domain and of La, Ti, and Th in the predominantly volcanic rocks domain, respectively, are above the concentrations of the 90th percentile of European stream sediments data (Salminen et al., 2005), indicating that in the study area there are geological and environmental conditions able to generate a natural enrichment of these elements.

4.4. *Limitations and scope of application*

The proposed approach allowed us to recognize five geochemical domains and characterize their geochemical signature. However, the compositional indicators obtained do not enable a clear-cut distinction of the transition from one domain to another. This transition therefore generates borderline regions which are the result of the complex geological setting of the study area, and the different degrees of weathering and sediment mixing. A key part of the approach presented here is the selection of end-members, which is fruitful only if extensive knowledge of the study area exists. Furthermore, incorrect labelling of some end-members could mislead the compositional indicators deducted from them and, therefore, the classification of the samples. However, mislabeled samples can be removed before computing compositional indicators as reported above. Another critical aspect concerns the applicability of the proposed compositional indicators to other study areas. The ones obtained in this study cannot be blindly used for any other regions; they are very much confined to the current study area or, at most, to central-southern Italy. However, the procedure described is general and can be similarly applied and adapted to other areas to obtain specific compositional indicators or to focus on the geochemical signals generated by specific anthropogenic phenomena. For example, focusing on the pollution of the study area and using samples from known polluted sites would make it possible to create compositional indicators for a concrete type of pollution in the different geochemical domains. Once the anthropic contribution is known, removing it and estimating the reference values of other PTEs would be possible.

Furthermore, using samples from the most important eruptions that occurred in the study area, it would be possible to create compositional indicators capable of distinguishing them and mapping them within the predominantly pyroclastic deposits domain. In our view, the proposed methodology could also help map mineralized areas, generating compositional indicators between the mineralization and the host rock. Therefore, although the proposed method has some limitations, we consider that it can significantly improve geochemical surveys and the evaluation of geochemical baselines in areas with prior knowledge of the geochemical framework.

5. Conclusions

A novel approach was presented to quantitatively define geochemical baselines for geochemically homogeneous areas (geochemical domains), combining geological and geochemical knowledge with compositional data analysis and supervised statistical learning.

The method was applied using chemical data relating to stream sediments proceeding from an area characterized by a relevant lithological heterogeneity. The map of geochemical domains generated shows substantial differences compared to the geological map, allowing a better understanding of the data variability concerning geology and some of the natural processes affecting the environmental media (e.g., weathering).

A key strength of the proposed method is that it allows for building compositional indicators that reveal the main natural phenomena (parental material, weathering degree and mixing) affecting stream sediment geochemistry. The obtained compositional indicators effectively group samples with a similar geochemical signature before assigning them reference values to be used as the upper limits of their geochemical background.

In the case of the present study, it has been sufficient to produce three compositional indicators, referring to the presence of volcanic material (CI_{ALS}), siliciclastic material (CI_{SLS}), and the degree of weathering (CI_{VLS}), to recognize five geochemical domains. These refer to the predominance of early volcanic products, late volcanic products, clayey, carbonate and mixed sediments. The results obtained are easily interpretable as they respond precisely to the geomorphological, geochemical, and geodynamic processes of the study area. Compositional indicators represent an easy and intuitive tool for identifying geochemical domains and assessing their spatial variation and geochemical baselines. In addition, the maps of individual compositional indicators can help to recognize how the intensity of the identified phenomena varies across the space and how overlaps occur. Therefore, by comparing the maps of the compositional indicators with the map of the geochemical domains, it is possible to obtain even more detailed information regarding the processes within the individual geochemical domains.

Although the method presented was applied to stream sediment data, its application principles are valid for any environmental media for which there is a need to define reference values for a specific territory (even of high complexity) and purpose (e.g., environmental pollution, mineral exploration or groundwater provenance studies). However, constructing compositional indices implies that a prior knowledge of the geological evolution and settings of the study area is available as the experience of the practitioner is a fundamental ingredient for the success of the method.

Funding sources

Javier Palarea-Albaladejo was partly supported by the Spanish Ministry of Science and Innovation (MCIN/AEI/10:13039/501100011033) and ERDF A way of making Europe [grant number PID2021-123833OB-I00] and the Department of Research and Universities of the Generalitat de Catalunya [grant number 2021SGR01197].

CRediT authorship contribution statement

Maurizio Ambrosino: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Javier Palarea-Albaladejo:** Writing – review & editing, Visualization, Supervision, Software, Methodology, Funding acquisition, Formal analysis, Data curation. **Stefano Albanese:** Writing – review & editing, Visualization, Validation, Supervision, Methodology. **Domenico Cicchella:** Writing – review & editing, Writing – original draft, Resources, Project administration, Funding acquisition.

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

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gexplo.2024.107545.

References

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. B 44 (2), 139–160.

Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. J. R. Stat. Soc. C Appl. Stat. 51 (4), 375–392.

Albanese, S., De Vivo, B., Lima, A., Cicchella, D., 2007. Geochemical background and baseline values of toxic elements in stream sediments of Campania region (Italy). J. Geochem. Explor. 93 (1), 21–34.

Ambrosino, M., Albanese, S., De Vivo, B., Guagliardi, I., Guarino, A., Lima, A., Cicchella, D., 2022. Identification of Rare Earth Elements (REEs) distribution patterns in the soils of Campania region (Italy) using compositional and multivariate data analysis. J. Geochem. Explor. 243, 107112.

Aruta, A., Albanese, S., Daniele, L., Cannatelli, C., Buscher, J.T., De Vivo, B., Petrik, A., Cicchella, D., Lima, A., 2022. A new approach to assess the degree of contamination and determine sources and risks related to PTEs in an urban environment: the case study of Santiago (Chile). Environ. Geochem. Health 45, 275–297.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227.

Bisson, M., Pareschi, M.T., Zanchetta, G., Sulpizio, R., Santacroce, R., 2007. Volcaniclastic debris-flow occurrences in the Campania region (Southern Italy) and their relation to Holocene - Late Pleistocene pyroclastic fall deposits: Implications for large-scale hazard mapping. Bull. Volcanol. 70 (2), 157–167.

Boente, C., Albuquerque, M.T.D., Gallego, J.R., Pawlowsky-Glahn, V., Egozcue, J.J., 2022. Composi- tional baseline assessments to address soil pollution: an application in Langreo, Spain. Sci. Total Environ. 812, 152383.

Buccianti, A., Nisi, B., Martín-Fernández, J.A., Palarea-Albaladejo, J., 2014. Methods to investigate the geochemistry of groundwaters with values for nitrogen compounds below the detection limit. J. Geochem. Explor. 141, 78–88.

Butler, B.M., Palarea-Albaladejo, J., Shepherd, K.D., Nyambura, K.M., Towett, E.K., Sila, A.M., Hillier, S., 2020. Mineral–nutrient relationships in African soils assessed using cluster analysis of X-ray powder diffraction patterns and compositional methods. Geoderma 375, 114474.

Caracciolo, L., 2020. Sediment generation and sediment routing systems from a quantitative provenance analysis perspective: review, application and future development. Earth Sci. Rev. 209, 103226.

Carranza, E.J.M., 2009. Chapter 5: catchment basin analysis of stream sediment anomalies. In: Handbook of Exploration and Environmental Geochemistry, vol. 11. Elsevier, pp. 115–144. https://doi.org/10.1016/S1874-2734(09)70004-X.

Celico, P., Guadagno, F.M., 1998. L'instabilità delle coltri piroclastiche delle dorsali carbonatiche in Campania: attuali conoscenze. Quaderni di geologia applicata 5 (1), 129–188.

Chayes, F., 1962. Numerical correlation and petrographic variation. J. Geol. 70 (4), 440–452.

Chen, R., Chen, H., Song, L., Yao, Z., Meng, F., Teng, Y., 2019. Characterization and source apportionment of heavy metals in the sediments of Lake Tai (China) and its surrounding soils. Sci. Total Environ. 694, 133819.

Ciarcia, S., Torre, M., 1996. I ciottoli dei conglomerati medio-pliocenici dell'Appennino campano: Provenienza, elaborazione, ambiente di deposizione. Soc. Geol. It. 115, 569–581.

Cicchella, D., Zuzolo, D., Albanese, S., Fedele, L., Di Tota, I., Guagliardi, I., Thiombane, M., De Vivo, B., Lima, A., 2020. Urban soil contamination in Salerno (Italy): concentrations and patterns of major, minor, trace and ultra-trace elements in soils. J. Geochem. Explor. 213, 106519.

Cicchella, D., Ambrosino, M., Gramazio, A., Coraggio, F., Assunta, M., Caputi, A., Avagliano, D., Albanese, S., 2022. Using multivariate compositional data analysis (CoDA) and clustering to establish geochemical backgrounds in stream sediments of an onshore oil deposits area. The Agri River basin (Italy) case study. J. Geochem. Explor. 238, 107012.

Cicchella, D., Ambrosino, M., Albanese, S., Guarino, A., Lima, A., De Vivo, B., Guagliardi, I., 2023. Major elements concentration in soils. A case study from Campania Region (Italy). J. Geochem. Explor. 247, 107179.

Cressie, N., 1992. Statistics for spatial data. Terra Nova 4 (5).

di Gennaro, A., Aronne, G., De Mascellis, R., Vingiani, S., Sarnataro, M., Abalsamo, P., Cona, F., Vitelli, L., Arpaia, G., 2002. I sistemi di terre della Campania. Monografia e carta 1, 250.000, con legenda.

Dominech, S., Yang, S., Aruta, A., Gramazio, A., Albanese, S., 2022. Multivariate analysis of dilution-corrected residuals to improve the interpretation of geochemical anomalies and determine their potential sources: the Mingardo River case study (Southern Italy). J. Geochem. Explor. 232, 106890.

Dung, T.T.T., Cappuyns, V., Swennen, R., Phung, N.K., 2013. From geochemical background determination to pollution assessment of heavy metals in sediments and soils. Rev. Environ. Sci. Bio/Technology 12, 335–353.

Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37 (7), 795–828.

Forte, F., Maffei, L., De Paola, P., 2020. Which future for small towns? Interaction of socio-economic factors and real estate market in irpinia. Val. e Valut. 25, 2020.

Giaccio, B., Isaia, R., Fedele, F.G., Di Canzio, E., Hoffecker, J., Ronchitelli, A., Sinitsyn, A. A., Anikovich, M., Lisitsyn, S.N., Popov, V.V., 2008. The Campanian Ignimbrite and Codola tephra layers: two temporal/stratigraphic markers for the Early Upper Palaeolithic in southern Italy and eastern Europe. J. Volcanol. Geotherm. Res. 177 (1), 208–226.

Glendell, M., Palarea-Albaladejo, J., Pohle, I., Marrero, S., McCreadie, B., Cameron, G., Stutter, M., 2019. Modeling the ecological impact of phosphorus in catchments with multiple environmental stressors. J. Environ. Qual. 48 (5), 1336–1346.

Graziano, R.S., 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.

Havlin, J.L., 2020. Soil: Fertility and Nutrient Management, Landscape and Land Capacity, 2020. CRC Press, pp. 251–265.

Isaaks, E.H., Srivastava, R.M., 2010. An introduction to applied geostatistics. Geogr. Anal. 6 (3).

Jarauta-Bragulat, E., Hervada-Sala, C., Egozcue, J.J., 2016. Air quality index revisited from a compositional point of view. Math. Geosci. 48 (5), 581–593.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.

Lane, S.N., Parsons, D.R., Best, J.L., Orfeo, O., Kostaschuk, R.A., Hardy, R.J., 2008. Causes of rapid mixing at a junction of two large rivers: Río Paraná and Río Paraguay, Argentina. J. Geophys. Res. Earth 113 (F2).

Lipp, A.G., Roberts, G.G., Whittaker, A.C., Gowing, C.J., Fernandes, V.M., 2021. Source region geochemistry from unmixing downstream sedimentary elemental compositions. Geochem. Geophys. Geosyst. 22 (10), e2021GC009838.

Liu, Y., Zhou, K., Carranza, E.J.M., 2018. Compositional balance analysis for geochemical pattern recognition and anomaly mapping in the western Junggar region, China. Geochem. Explor. Environ. Anal. 18 (3), 263–276.

Manzo, E., Perri, E., Tucker, M.E., 2012. Carbonate deposition in a fluvial tufa system: processes and products (Corvino Valley - southern Italy). Sedimentology 59 (2), 553–577.

Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosona-Delgado, R., 2018. Advances in principal balances for compositional data. Math. Geosci. 50, 273–298.

McKinley, J.M., Mueller, U., Atkinson, P.M., Ofterdinger, U., Jackson, C., Cox, S.F., Pawlowsky-Glahn, V., 2020. Investigating the influence of environmental factors on the incidence of renal disease with compositional data analysis using balances. Applied Computing and Geosciences 6, 100024.

McKinley, J.M., Mueller, U., Atkinson, P.M., Ofterdinger, U., Cox, S.F., Doherty, R., Pawlowsky-Glahn, V., 2021. Chronic kidney disease of unknown origin is associated with environmental urbanisation in Belfast, UK. Environ. Geochem. Health 43, 2597–2614.

Meloni, F., Nisi, B., Gozzi, C., Rimondi, V., Cabassi, J., Montegrossi, G., Rappuoli, D., Vaselli, O., 2023. Background and geochemical baseline values of chalcophile and siderophile elements in soils around the former mining area of Abbadia San Salvatore (Mt. Amiata, southern Tuscany, Italy). J. Geochem. Explor. 255, 107324.

Mondillo, N., Balassone, G., Boni, M., Rollinson, G., 2011. Karst bauxites in the Campania Apennines (southern Italy): a new approach. Period. Mineral. 80 (3), 407–432.

Nawrot, N., Wojciechowska, E., Mohsin, M., Kuittinen, S., Pappinen, A., Rezania, S., 2021. Trace metal contamination of bottom sediments: a review of assessment measures and geochemical background determination methods. Minerals 11 (8), 872.

Osman, K.T., 2013. Plant nutrients and soil fertility management. Soils 129–159.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and analysis of compositional data. In: Modeling and Analysis of Compositional Data. John Wiley & Sons, New York, NY.

Petrik, A., Thiombane, M., Lima, A., Albanese, S., Buscher, J.T., De Vivo, B., 2018. Soil contami- nation compositional index: a new approach to quantify contamination demonstrated by assessing compositional source patterns of potentially toxic elements in the Campania Region (Italy). J. Appl. Geochem. 96, 264–276.

Piana, F., Fioraso, G., Irace, A., Mosca, P., D`Atri, A., Barale, L., Falletti, P., Monegato, G., Morelli, M., Tallone, S., Vigna, G.B., 2017. Geology of Piemonte region (NW Italy, Alps–Apennines interference zone). J. Maps 13 (2), 395–405.

Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., GEMAS Project Team, 2012. The concept of compositional data analysis in practice—total major element concen- trations in agricultural and grazing land soils of Europe. Sci. Total Environ. 426, 196–210.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., 2014. Chemistry of Europe's agricultural soils. Part A: methodology and interpretation of the GEMAS data set. Schweizerbart Science Publishers. Geol. Jb. 102, 880.

Ruberti, D., Vigliotti, M., 2017. Land use and landscape pattern changes driven by land reclamation in a coastal area: the case of Volturno delta plain, Campania Region, southern Italy. Environ. Earth Sci. 76 (20), 694.

Salminen, R., Tarvainen, T., 1997. The problem of defining geochemical baselines. A case study of selected elements and geological materials in Finland. J. Geochem. Explor. 60 (1), 91–98.

Salminen, R., Batista, M.J., Bidovec, M., Demetriades, A., De Vivo, B., De Vos, W., Duris, M., Gilucis, A., Gregorauskiene, V., Halamic, J., Heitzmann, P., Lima, A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutura, J., Marsina, K., Mazreku, A., O'Connor, P.J., Olsson, S.A., Ottesen, R.T., Petersell, V., Plant, J.A., Reeder, S., Salpeteur, I., Sandström, H., Siewers, U., Steenfelt, A., Tarvainen, T., 2005. Geochemical Atlas of Europe. Part 1: Background Information, Methodology and Maps, Geological Survey of Finland, Espoo. GTK, FOREGS.

Singh, A., Maichle, R., 2015. ProUCL 5.1. User Guide: Statistical Software for Environmental Applications for Data Sets with and Without Nondetect Observations. US Environmental Protection Agency. Epa/600/R-07/041. https://www.epa.gov/sites/default/files/2016-05/documents/proucl_5.1_user-guide.pdf.

Štefelová, N., Palarea-Albaladejo, J., Hron, K., Gába, A., Dygrýn, J., 2023. Compositional PLS biplot based on pivoting balances: an application to explore the association between 24-h movement behaviours and adiposity. Comput. Stat. 1–29.

Sundaray, S.K., Nayak, B.B., Lin, S., Bhatta, D., 2011. Geochemical speciation and risk assessment of heavy metals in the river estuarine sediments-a case study: Mahanadi basin, India. J. Hazard. Mater. 186 (2–3), 1837–1846.

Tepanosyan, G., Sahakyan, L., Maghakyan, N., Saghatelyan, A., 2020. Combination of compositional data analysis and machine learning approaches to identify sources and geochemical associations of potentially toxic elements in soil and assess the associated human health risk in a mining city. Environ. Pollut. 261, 114210.

Umar, M., Rhoads, B.L., Greenberg, J.A., 2018. Use of multispectral satellite remote sensing to assess mixing of suspended sediment downstream of large river confluences. J. Hydrol. 556, 325–338.

Vitale, S., Ciarcia, S., 2013. Tectono-stratigraphic and kinematic evolution of the southern Apennines/Calabria-Peloritani Terrane system (Italy). Tectonophysics 583, 164–182.

Vitale, S., Ciarcia, S., 2018. Tectono-stratigraphic setting of the Campania region (Southern Italy). J. Maps 14 (2), 9–21.

Vitale, S., Ciarcia, S., 2022. The dismembering of the Adria platforms following the Late Cretaceous-Eocene abortive rift: a review of the tectono-stratigraphic record in the southern Apennines. Int. Geol. Rev. 64 (20), 2866–2889.

Vitale, S., Ciarcia, S., Mazzoli, S., Zaghloul, M.N., 2011. Tectonic evolution of the "Liguride" accretionary wedge in the Cilento area, southern Italy: a record of early Apennine geodynamics. J. Geodyn. 51 (1), 25–36.

Wang, Z., Chen, X., Yu, D., Zhang, L., Wang, J., Lv, J., 2021. Source apportionment and spatial distribution of potentially toxic elements in soils: a new exploration on receptor and geostatistical models. Sci. Total Environ. 759, 143428.

Zuzolo, D., Cicchella, D., Lima, A., Guagliardi, I., Cerino, P., Pizzolante, A., Thiombane, M., De Vivo, B., Albanese, S., 2020. Potentially toxic elements in soils of Campania region (Southern Italy): combining raw and compositional data. J. Geochem. Explor. 213, 106524.