

Article

Accounting for the Compositional Nature of Geochemical Data to Improve the Interpretation of Their Univariate and Multivariate Spatial Patterns: A Case Study from the Campania Region (Italy)

Lucia Rita Pacifico , Annalise Guarino , Antonio Iannone and Stefano Albanese 

Department of Earth Sciences, Environment and Resources (DiSTAR), University of Naples Federico II, 80126 Naples, Italy; luciarita.pacifico@unina.it (L.R.P.); antonio.iannone2@unina.it (A.I.)

* Correspondence: annalise.guarino@unina.it (A.G.); stefano.albanese@unina.it (S.A.)

Abstract: This study investigates the application of Compositional Data Analysis (CoDA) and multivariate statistical techniques to geochemical data from the soils of the Campania region. The dataset examined includes 3571 soil samples analyzed for 37 chemical elements. Principal Component Analysis (PCA) was employed to reduce the dataset's dimensionality and identify key relationships between elements. The first PCA identified groups of highly correlated variables, which were then reduced to 20 representative elements for a second PCA. The three most significant principal components (PC1, PC2, and PC3) explained approximately 65% of the total variability. PC1 (accounting for 29.97% of variability) revealed an anticorrelation between Ti, La, and Sc with Au, Hg, and Ag, with positive scores primarily located in the inland Apennine areas. PC2 (21.8%) was dominated by Na, K, and Cu, with positive scores corresponding to volcanic deposits, aligning with the dispersion patterns of historical Vesuvian eruption products. PC3 (11%) was associated with Ca and S, with higher scores found in the alluvial plains and inland areas. These results demonstrate the efficacy of CoDA in minimizing spurious correlations and uncovering latent relationships between elements, thereby enhancing the interpretation of natural and anthropogenic processes influencing soil variability in the region.



Academic Editor: Zoltán Kern

Received: 4 November 2024

Revised: 4 January 2025

Accepted: 6 January 2025

Published: 9 January 2025

Citation: Pacifico, L.R.; Guarino, A.; Iannone, A.; Albanese, S. Accounting for the Compositional Nature of Geochemical Data to Improve the Interpretation of Their Univariate and Multivariate Spatial Patterns: A Case Study from the Campania Region (Italy). *Geosciences* **2025**, *15*, 20. <https://doi.org/10.3390/geosciences15010020>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: environmental geochemistry; compositional data analysis; geochemical mapping; multivariate statistics

1. Introduction

The understanding of geochemical processes occurring in both the exogenous and endogenous environment improves when considering the synergistic behavior (in both space and time) of multiple chemical variables rather than analyzing a single element or compound. In fact, geochemical associations that comprise groups of major and/or trace elements can be used to delineate the spatial extent of environmental impacts arising from a specific natural source (e.g., mineralization or lithology) or from an anthropogenic contamination source (whether point or diffuse), capable of releasing specific mixtures of chemicals into the environment (Table 1) [1].

However, geochemists have not yet codified all potential associations, and the analysis of the correlations that exist between the variables in a geochemical dataset, particularly in the context of multivariate analysis, can facilitate the identification of the processes that predominantly influenced the compositional variations observed in the studied environmental media (e.g., soil, water, and sediments).

Table 1. Major associations of chemical elements that can be found in environmental media (soil, sediments, and waters) and whose presence can serve as markers of specific anthropogenic processes [1].

Source	Cu	Pb	Zn	Sn	Cd	Hg	Ni	V	Cr	As	Sb	Other Elements
Urban areas	+	+	+	+		+						
Mining activities		+		+			+					
Foundries	+	+	+	+		+	+			+	+	
Steelworks	+			+	+		+	+	+	+		Ca, P ₂ O ₅
Heavy engineering, tool construction	+		+				+	+	+			Mn, Mo, W
Metal plating and finishing	+		+	+			+		+			
Manufacturing of electronic components	+				+	+	+			+	+	Rare Earth Element
Production/processing of ceramics and glass		+		+	+						+	
Incinerators						+				+		
Coal-fired power plants		+		+				+			+	
Vehicles and transportation	+	+	+	+	+			+			+	Ba, Mn, Pt, Pd
Cremation furnaces						+						
Agriculture (vineyards)	+									+		P

Geochemical data are expressed as compositions represented by a quantity relative to a total, and, within a sample, the sum value of these data is constant (e.g., 1 kg, 100%), implying that geochemical data are characterized by a “closed” structure, wherein knowledge of D-1 components suffices to gain a complete understanding of a composition constituted by D components [2–4].

Thus, the geochemical composition of a sample (such as soil, water, and rock) can be conceptualized as the projection of a vector, defined within a real-space R of D dimensions (corresponding to the number of variables) onto a finite surface, defined Simplex (S), constrained by vertices that coincide with the axes of a D-dimensional Cartesian system reflecting a unit value [5,6].

Even soil particle size distribution is expressed as a composition of the percentage quantity of sand, silt, and clay (D = 3). In this case, S is comparable to an equilateral triangle, and any point has a composition characterized by a sum of its components that always equals 100. If we consider that this sum must, by convention, remain constant, and if we add sand to the system, the percentage increase in the content of sand will correspond to a proportional decrease in the other two components that is only apparent if the actual quantities of clay or silt have not been modified.

The phenomenon described for particle size distribution is valid for any type of compositional data (including geochemical data). This condition can produce false correlations among the components, which Pearson in 1897 defined as “spurious” [7]. Additionally, it should be considered that every composition we observe is inevitably a subcomposition of a potentially larger composition, of which we do not necessarily know all the parts.

To give an example, let us suppose that a number of soil samples are analyzed using two independent laboratories. Due to the different detection limits of the measuring instruments used by the laboratories, the first laboratory detected 20 chemical elements in the soil samples, while only 10 were detected by the second. The different analytical capabilities of the two laboratories will result in datasets that turn out to be closed with respect to different compositions (or subcompositions). Therefore, the concentration values of the elements detected by both laboratories could appear artificially different.

To address the issue of spurious correlations and the consequences arising from the use of subcompositions consisting of a different number of parts, over time, various researchers

have proposed mathematical solutions aimed at making the variability of individual subsets independent of the overall variability of the composition to which they belong [3,8].

Aitchison [2,9,10] proposed the use of ratios between the parts of a composition to overcome the issue, considering that the ratio between two parts remains constant regardless of the other parts present (subcompositional coherence) and is not affected by the removal or addition of parts and the reclosure of the dataset. The application of logarithms to the ratios (log-ratio) transfers the data out of S and into R , allowing their use with many standard univariate and multivariate statistical analysis techniques.

The simplest logarithmic ratio is the “Additive Log-Ratio” (alr) transformation, which transfers the original composition from S^D into R^{D-1} . In the alr transformation, one part of the composition (arbitrarily chosen) is used as the denominator [6]. The formula that summarizes the alr transformation is as follows:

$$alr(x) = \left[\log \frac{x_1}{x_D} \cdots \log \frac{x_{D-1}}{x_D} \right] \quad (1)$$

where x_n is the concentration value of the chemical element; and x_D is the concentration of the chemical element, which is part of the original composition, chosen to open the dataset.

The “Centered Log-Ratio” (clr) transformation transfers the compositional values from S^D into R^D . Specifically, in a geochemical dataset where the columns represent the variables (chemical elements, oxides, etc.) and the rows represent the different samples (observations) analyzed, the clr transformation is obtained by calculating, for each cell, the logarithm of the ratio between the original value and the geometric mean of all the values in the row, according to the following formula [6]:

$$clr(x) = \left[\log \frac{x_1}{g(x)} \cdots \log \frac{x_D}{g(x)} \right] \quad (2)$$

where x_n is the concentration value of individual elements, and $g(x)$ is the geometric mean of all variables contributing to the composition.

The “Isometric Log-Ratio” (ilr) transformation, like the alr, transfers the original composition from S^D into R^{D-1} and is based on the logarithm of the geometric means of two subsets of parts [11].

In this study, we sought to evaluate the effectiveness of log-ratio transformations, specifically the centered log-ratio (clr) and isometric log-ratio (ilr) transformations, in addressing the compositional nature of geochemical data. We utilized a comprehensive dataset derived from soils in the Campania region. Our objectives were twofold: first, to assess how Compositional Data Analysis (CoDA) can enhance the interpretation of geochemical patterns for individual elements; and second, to evaluate the performance of a proposed operational sequence that applies multivariate statistics to the transformed data. This method aims to uncover associations among geochemical elements that represent distinct processes, which are often obscured by influences related to primary sources.

2. Geological and Pedological Setting of the Study Area

The Campania region covers a total area of approximately 13,600 km². The regional territory is crossed longitudinally by a portion of the Southern Apennines, oriented NW-SE, whose highest peaks correspond to Mount Matese in the north, Mount Taburno and the Picentini Mountains in the center, and the Alburni Mountains in the southeast. These reliefs are mainly composed of sedimentary rocks (i.e., limestone and dolomite), while the innermost domains consist of siliceous schists and terrigenous sediments (i.e., clays, siltstones, sandstones, and conglomerates) [4,12]. Volcanic lithotypes are widely present in the coastal areas, rocks, and sediments with predominantly potassic and ultrapotassic

chemistry produced by the Quaternary volcanic activity of the Roccamonfina in the north-west, and the Somma–Vesuvius, Phlegraean Fields, and Mt. Epomeo (Ischia Island) in the central coastal zone [13,14]. The coastal areas and plains are mainly composed of alluvial sediments, primarily derived from the reworking of pyroclastic deposits that cover a large part of the Apennine carbonatic reliefs [12,15]. These reliefs are part of the structural horst bordering the graben formed by the Campanian Plain [15] (Figure 1A).

The average rainfall amount of the region is about 800 mm/year, with higher values characterizing the plains and coastal areas, whereas the eastern sectors receive lower amounts [16]. Thanks to their predominantly volcanic origin and favorable climate, regional soils, particularly those in the plains, are characterized by a significant abundance of water and exceptional agricultural yield. Campania is, in fact, one of the Italian regions where agriculture is thriving, both in terms of the quantity and quality of its products. The land is mostly dedicated to herbaceous crops (e.g., wheat, cereals, weeds, pastures, ornamental plants, and industrial crops), horticultural crops (e.g., potatoes and tomatoes), and arboricultural crops (e.g., fruit and nut trees). The forests, predominantly concentrated in the Apennine reliefs, are characterized by the presence of conifers, broadleaf trees, sclerophyllous plants, bushes, and shrubs (Figure 1B).

The principal urban areas, including the metropolitan area of Naples, are located along the northern and central Tyrrhenian coastal margin of the region. In the hinterland, the cities of Benevento, Caserta, and Avellino are the capitals of their respective provinces. The region, with a total population of 5.6 million inhabitants in 2022, has an average population density of about 424 inhabitants per km², with values up to 11.818 inhabitants per km². Just over half of the population lives in the province of Naples (53.1%), which is close to 3 million inhabitants. The province of Salerno, with over 1 million residents, contains 18.9% of the region's residents, while the other three provinces contain 28.0%.

The main industrial areas are situated near the major cities; the more recent ones are concentrated in industrial zones often placed in easily accessible plains, surrounded by predominantly agricultural lands, and crossed by high-speed roadways.

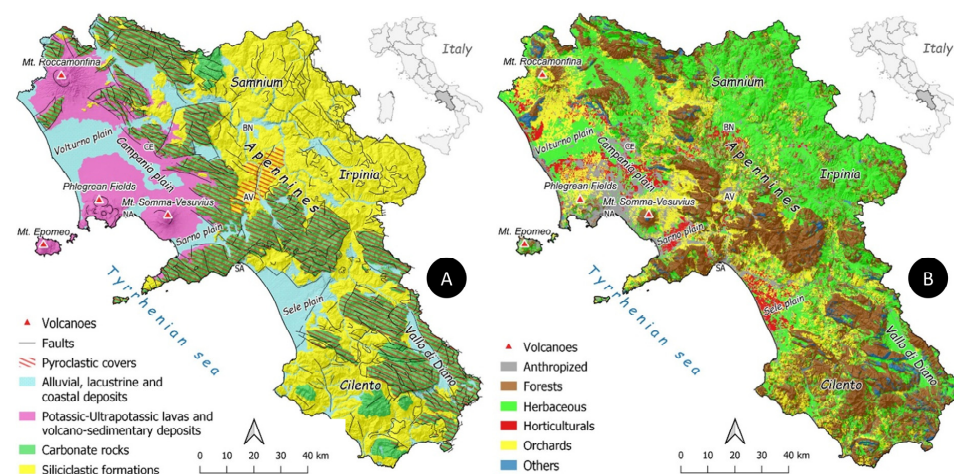


Figure 1. (A) Geolithological map indicating the extent of the main pyroclastic covers [17]. (B) Land use map derived from CORINE Land Cover 2018 data (source: <https://land.copernicus.eu/en/products/corine-land-cover/clc2018> (accessed on 5 January 2025)).

3. Materials and Methods

3.1. Sampling and Chemical Analysis

A huge geochemical dataset consisting of 3571 sampling points is available for the entire Campania region. Soil samples were collected over several years with a spatial density ranging from one sample per 16 km² in rural and/or agricultural areas to

one sample per 4 km² in major urban areas [18]. At each sampling point, samples were collected at a depth of 10–30 cm, following the international sampling procedures guidelines established by the FOREGS Geochemistry Group [19].

Topsoil samples were properly dried by infrared lamps ($T < 35\text{ }^{\circ}\text{C}$), disaggregated, and sieved to retain the 2 mm fraction for analysis. In all, 15 g of the treated sample was digested at 95 °C for 1 h in 90 mL of an Aqua Regia solution. The cooled solution had a final volume of 300 mL with 5% HCl. The final solutions were analyzed by inductively coupled plasma–mass spectrometry (ICP-MS) and inductively coupled plasma–atomic emission spectrometry (ICP-AES) at the Bureau Veritas (formerly Acme Lab) Analytical Laboratories Ltd. (Vancouver, Canada) [4,12]. The chemical analyses conducted provided information on the quasi-total concentration of 52 elements (Ag, Al, As, Au, B, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Fe, Ga, Ge, Hf, Hg, In, K, La, Li, Mg, Mn, Mo, Na, Nb, Ni, P, Pb, Pd, Pt, Rb, Re, S, Sb, Sc, Se, Sn, Sr, Te, Th, Ti, Tl, U, V, W, Y, Zn, and Zr). Precision of the analysis was calculated using in-house replicates and blind duplicates (median value of the Relative Percentage Difference = 1.3%), and accuracy was determined using in-house reference materials (median value 2.2%).

3.2. Data Preparation and Preliminary Analysis of Spatial Distribution

Chemical elements with more than 40% of missing values, either because they were below the detection limit (BDL) or not analyzed, were excluded, resulting in a final dataset that includes only 37 variables (Ag, Al, As, Au, B, Ba, Bi, Ca, Cd, Co, Cr, Cu, Fe, Ga, Hg, La, Mg, Mn, Mo, Na, Ni, P, Pb, S, Sb, Sc, Se, Sr, Te, Th, Ti, Tl, U, V, W, and Zn). For the elements included in the dataset that presented less than 40% of missing values, the k-Nearest Neighbor (kNN) approach was successfully used to determine the most probable value for each missing data point.

Aiming to check for potential differences in the spatial distribution of the log-ratio values compared to the original compositional data, the subcomposition made up of the 37 selected variables was “opened” using the clr transformation of Equation (2), where x_n is the concentration value of the individual element, and $g(x)$ is the geometric mean of all variables contributing to the composition.

The optimized dataset was integrated with the coordinates of the sampling points, and both the raw compositional data and the transformed data were interpolated using the Multifractal Inverse Distance Weighted (MIDW) method [20,21] and mapped in ArcMap 10.3.1. The interpolated grids were classified using the Concentration–Area (C-A) plots (Figures 2 and 3), identifying the main inflexion points on the curves with the help of the ArcFractal add-in [22].

3.3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate statistical technique employed to analyze the correlation between the parts of a (sub)composition through a reduced number of linearly uncorrelated variables, known as principal components (PCs) [8,23,24]. Generally, the PCs are determined and ranked based on their contribution to the total data variability. Those that most influence the overall variance represent the processes most active in the study domain [23,25].

In the context of geochemical studies applied to mining exploration and environmental research, the new variables produced are a direct consequence of the processes (whether natural or anthropogenic) that control the statistical variability of the multiple chemical elements present within the analyzed media (e.g., soil, sediments, and water) [8].

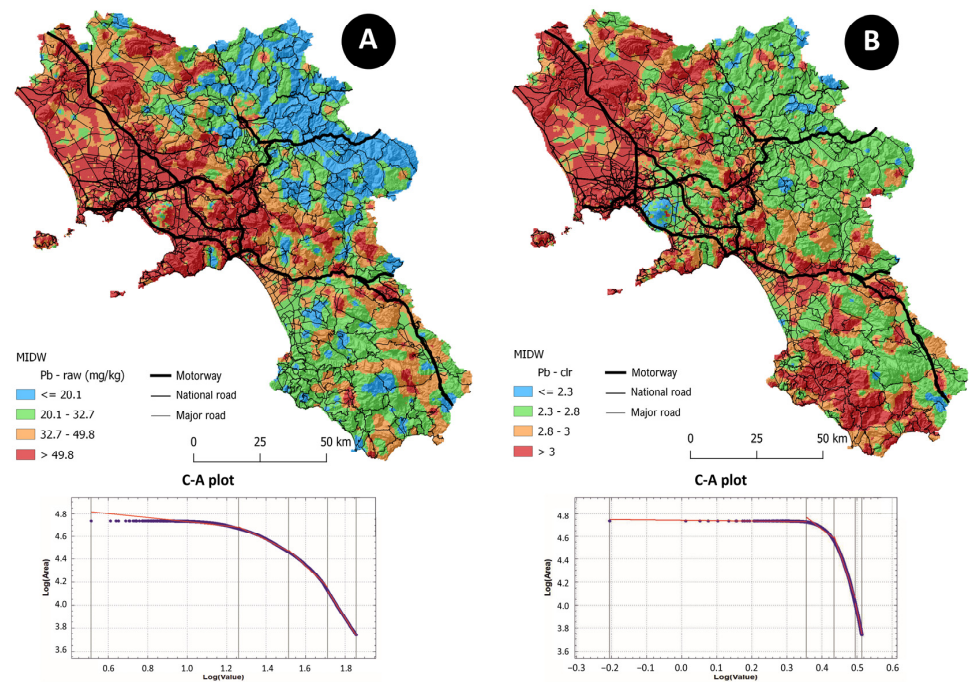


Figure 2. Interpolated Pb value maps and Concentration–Area (C–A) plots used for the definition of the value ranges in the legends: (A) original (raw) soil concentration data in mg/kg; and (B) data transformed using the Centered Log-Ratio (clr) method. In each map, raster data are overlaid with the principal elements of the regional road network.

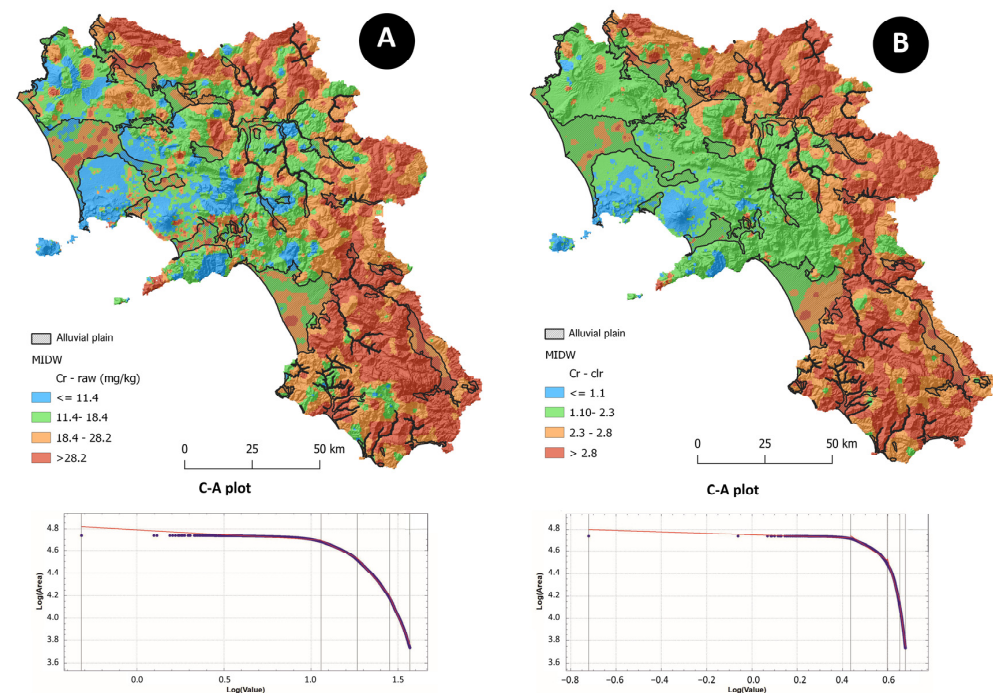


Figure 3. Interpolated Cr value maps and C–A plots used for the definition of the value ranges in the legends: (A) original (raw) soil concentration data in mg/kg and (B) data transformed using the clr method. In each map, raster data are overlaid with polygons representing the alluvial deposits shown in the geolithological map of Figure 1A.

Given that PCA is a technique developed to analyze relationships between non-compositional variables and relies on the analysis of correlations between input variables, it is crucial for its application in geochemistry that the input data be “opened”, with the aim of reducing the presence of spurious correlations.

A PCA was performed on the optimized dataset following the procedure proposed by Filzmoser et al. [26], as it provides more robust results compared to other approaches, since the input compositional dataset is first transformed into ilr coordinates. The results of the PCA were graphically represented through biplots, which are graphs that display the relationships between the input variables and their loading within the individual components. In the biplots, the individual observations (samples) are represented as points, which are located at a certain distance from the component axes according to the extent to which each component influences their variability. Scores are assigned to each observation according to its relationship with the individual components.

PCA was performed using the “pcaCoDa” function available in the “RobComposition” package in the R environment (<https://www.r-project.org/>).

4. Results and Discussion

4.1. Univariate Spatial Distribution

The application of the MIDW algorithm to both the original raw geochemical data and to the same transformed using the clr transformation has shown variations in the distribution patterns of certain environmentally relevant elements. These observations certainly offer insights into the usefulness of these types of transformations. For the sake of brevity, we present only the maps produced for Pb and Cr (Figures 2 and 3).

Regarding Pb, it can be observed that, in the raw data (Figure 2A), areas with the highest values are mainly located in the soils of the Campanian Plain, from the north (where the Roccamonfina volcanic complex is located), extending southwards to the Sorrento Peninsula and including the territory of the Sarno River basin towards the interior of the region. Lead gasolines have no longer been produced since 1 January 2002, but the high concentration values also appear to roughly follow the main road axes, including the regional section of the A2 motorway (commonly known as the “Salerno–Reggio Calabria”), which represents the main road link to the Calabria region. Mainly in the summer months, this motorway experiences extremely high levels of vehicular traffic.

The clr transformed data (Figure 2B) show a distribution pattern not entirely different from that produced by the raw data but with a few noteworthy peculiarities:

- The overall extent of areas marked by the highest values is reduced, especially in the inland territories of the Campanian Plain and the less urbanized areas of the Vesuvius volcanic complex;
- The western sector of the Cilento, in the southern part of the region, displays transformed values within a significantly higher range compared to the raw data;
- The inland areas of the region, generally characterized by siliciclastic formation, show transformed values that belong to a higher range than the original compositional data;
- The distribution of the clr transformed values appears to positively correlate with the road network distribution, with higher values corresponding to areas where the network is denser and associated with the presence of the principal roads.

Regarding Cr, the highest values of the raw data are predominantly found in the inland areas of the region (in correspondence with the soils developed on siliciclastic deposits), as well as the Cilento area (Figure 3A). Relatively high values can also be observed roughly in correspondence with the alluvial soils of the main regional rivers’ plains (i.e., Volturno, Sarno and Sele) and of some smaller watercourses.

The spatial distribution of the clr transformed data shows more homogeneous patterns: higher values correspond to the inland siliciclastic domains; intermediate values characterize soils developed on loose volcanic deposits that cover the carbonate reliefs and fill the Campanian Plain; and lower values correspond to the igneous lithologies of the Somma–Vesuvius complex, the Phlegraean Fields, and Mt. Epomeo (Ischia).

In contrast to the compositional data, the transformed chromium (Cr) data do not exhibit significantly elevated values in the alluvial regions, aligning with the neighboring volcanic soils of the Campanian Plain (Figure 3B) that serve as their source materials. The highest reported concentration values of Cr in the alluvial plain soils, based on the raw data, likely indicate relative enrichment resulting from absorption processes that influence the mobility of trace metals in this environment.

The results for both selected variables highlight how applying the clr transformation enhances the distribution patterns associated with the main geological and spatially diffuse anthropic processes influencing the overall variability of chemical elements.

Specifically, as for Pb, the variability observed in the soil of Campania appears to be associated, as in other regions, with widespread contamination linked to atmospheric fallout from vehicular traffic emissions. In fact, until the early 2000s, significant amounts of Pb were released into the environment due to its use as a petrol additive in its organic form (tetraethyl lead— $\text{Pb}(\text{C}_2\text{H}_5)_4$).

In the case of Cr, the variability seems to depend mainly on the geological features since it shows a particular affinity for the clay fractions that characterize the soils developed on the siliciclastic deposits of inland Campania and the Cilento, and, to a lesser extent, for the regional volcanic soils.

The advancements conferred by the clr transformation on geochemical patterns do not contradict the findings derived from untransformed data; rather, they enhance the efficacy of information utilization concerning the scale and extent of the study area. This data transformation facilitates a greater emphasis on active enrichment processes associated with the substantial contributions of elements from specific geological units or influenced by anthropogenic activities.

Conversely, local geochemical processes, which are critical for assessing the fate and transport of elements on a small scale, may become less significant and even pose a disturbance when the aim of the geochemical investigation is to ascertain the source of natural enrichment, such as mineralization, or to determine the origin of human-induced contamination.

Compositional Data Analysis (CoDA), when applied to various geochemical datasets, particularly in contexts characterized by multiple active sources and processes, can serve as a robust tool. This approach prioritizes the identification of principal sources over secondary influences, thereby enhancing the management of initiatives with economic implications, such as mineral resource exploitation, as well as those focused on environmental recovery efforts.

4.2. Elementary Associations

For the multivariate analysis of the geochemical data of soils in the Campania region, two PCAs were conducted. The first PCA included all 37 available variables. The analysis of the results, supported by the biplot of the first two PCs (accounting for 49.9% of the total data variability) (Figure 4), allowed for identifying groups of variables strongly correlated with each other. Thus, variables whose vectors were nearly overlapping in the biplot were reduced to a single representative element, selected based on the vector magnitude. This reduction brought us to the selection of twenty variables (Cu, Pb, Ag, Ni, U, Au, Cd, Sb, Bi, Ca, La, Cr, Ti, Na, K, Sc, Tl, S, Hg, and Se), which were extracted from the original dataset and used as input for the second PCA.

The output of the PCA conducted on the reduced dataset identified three significant principal components (PC1, PC2, and PC3), which, despite the substantial reduction in input variables compared to the first analysis, account for approximately 65% of the total data variability (Table 2).

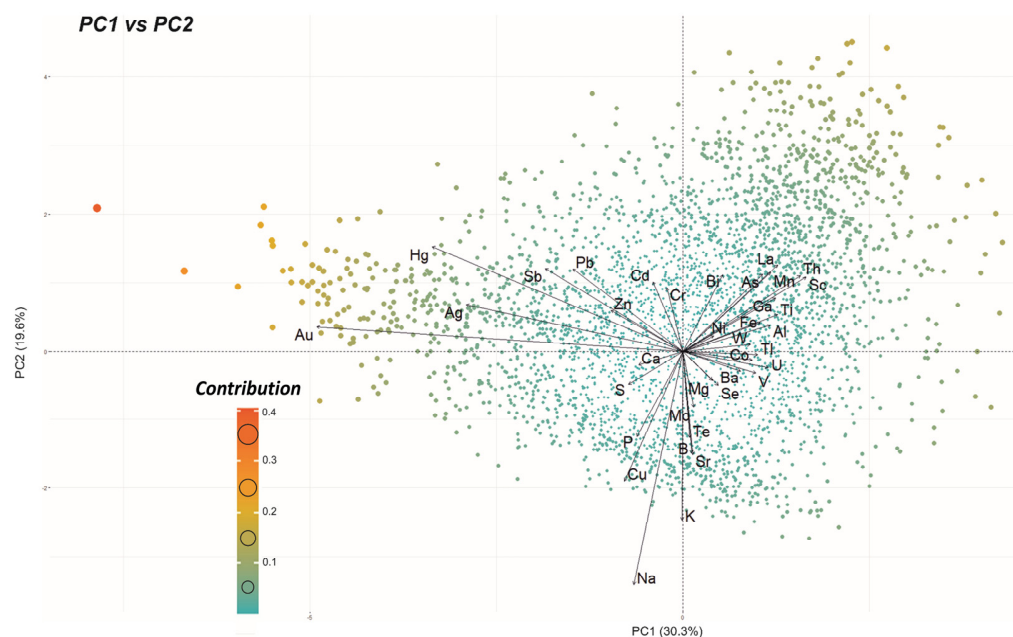


Figure 4. Biplot of the first two principal components (PCs), obtained from the PCA conducted including all the 37 available compositional variables (chemical elements). Observations (samples) are represented as points whose size and color refer to their contribution to the overall data variability in the represented components.

Table 2. Loadings of individual variables concerning the first three PCs (PC1, PC2, and PC3) determined by the PCA applied to the reduced database. The values reported in red represent the positive loadings of the most relevant elements for each component, and those in blue represent the negative ones. The percentage values of the individual and cumulative contribution of the PCs to the overall variance are also provided.

Element	Loadings		
	PC1	PC2	PC3
Cu	−0.07	0.34	0.10
Pb	−0.12	−0.15	−0.13
Ag	−0.31	−0.09	−0.14
Ni	0.15	−0.08	0.11
U	0.20	0.13	−0.24
Au	−0.58	−0.06	−0.07
Cd	0.06	−0.21	0.16
Sb	−0.16	−0.20	−0.04
Bi	0.14	−0.11	−0.22
Ca	−0.03	−0.09	0.61
La	0.25	−0.13	−0.23
Cr	0.09	−0.20	0.15
Ti	0.24	−0.03	−0.17
Na	−0.12	0.63	−0.05
K	0.01	0.45	−0.02
Sc	0.31	−0.15	−0.09
Tl	0.19	0.07	−0.14
S	−0.02	−0.02	0.44
Hg	−0.37	−0.19	−0.24
Se	0.14	0.09	0.21
Variance %	29.97	21.74	13.22
Variance cum. %	29.97	51.71	64.93

The results of the PCA were presented in biplots in pairs of components (PC1 vs. PC2 and PC2 vs. PC3), in which the observations (samples) are classified based on both the geolithological characteristics (Figure 5) and the land use (Figure 6).

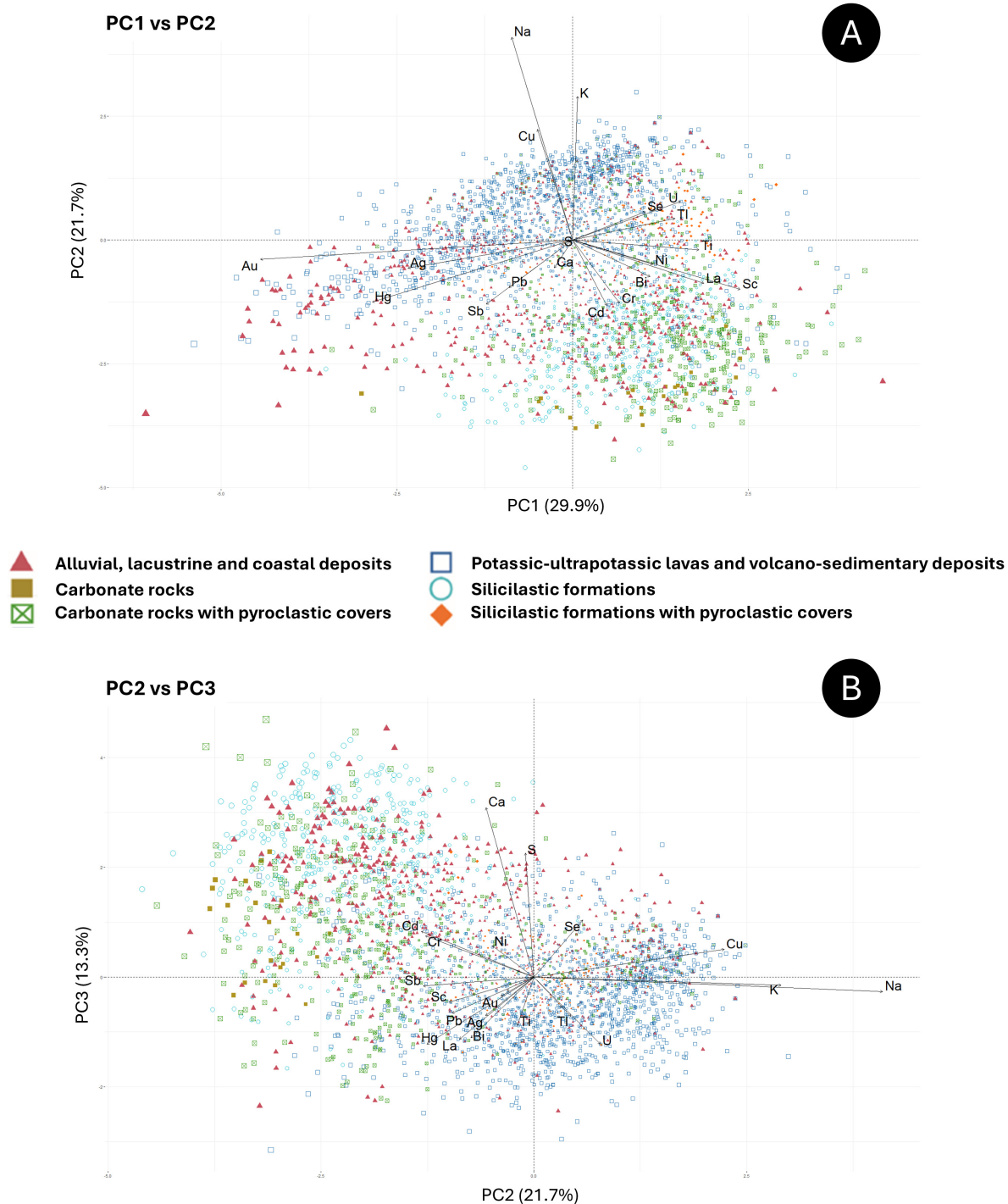


Figure 5. Biplot of (A) PC1 vs. PC2 and (B) PC2 vs. PC3, obtained from the PCA applied to the reduced dataset. The observations are represented by symbols based on the different geolithological backgrounds (Figure 1A) of the soil sample collection sites.

The scores of the different PCs were associated with the coordinates of each sample. Interpolated maps were then produced (again using the MIDW method), showing for each PC the areas of greatest influence on local geochemical variability (Figure 7).



Figure 6. Biplot of (A) PC1 vs. PC2 and (B) PC2 vs. PC3, obtained from the PCA applied to the reduced dataset. The observations are represented by symbols based on the land use (Figure 1B) of the soil sample collection sites.

The variables influencing PC1, which accounts for 29.97% of the total variability, are scandium (Sc), lanthanum (La), titanium (Ti), gold (Au), mercury (Hg), and silver (Ag). These elements individually contribute more than 5% to the overall variability of the component (Figure 7A), and all have an absolute value of loadings > 0.2 (Table 2). Specifically, in this component, La, Ti, and Sc (with positive loadings) contrast Au, Hg, and Ag (with negative loadings), defining a clear anticorrelation between these two groups of elements, as can be easily observed in the corresponding biplots (Figures 5A and 6A). The spatial distribution of the positive scores predominantly characterizes the inland Apennine areas.

In contrast, the negative values characterize large portions of the plains in proximity to the main urban and industrial settlements (Figure 7A). The interpretation of the results is aided by the analysis of the biplots, which show that, in general, the samples with high positive scores are mainly characterized by carbonate rocks or siliciclastic formations covered by pyroclastic deposits (Figure 5A), and by varied land use that includes forests, cultivated areas, and, to a lesser extent, other uses; as for the negative scores, as observed on the map, they are mainly associated with alluvial and coastal deposits and with predominantly urban/residential land use (Figure 6A).

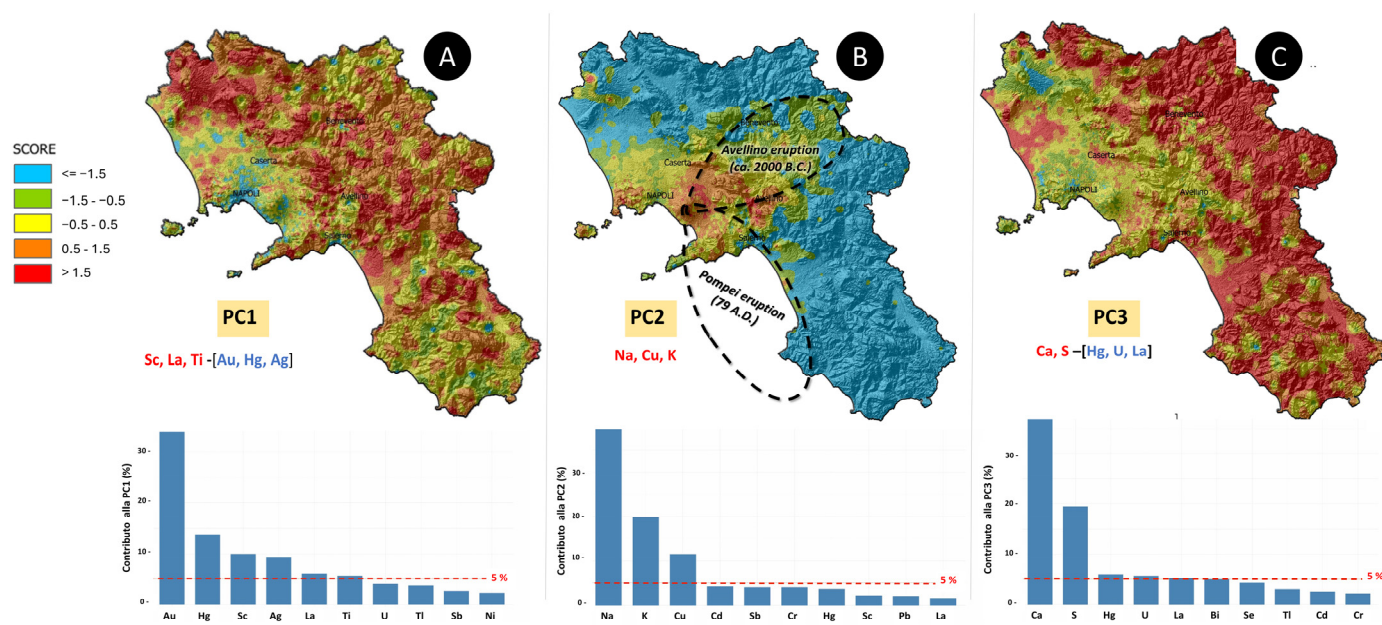


Figure 7. Raster maps of the distribution of scores for (A) PC1, (B) PC2, and (C) PC3, obtained from the PCA applied to the reduced database, indicating the variables that contribute most significantly (contribution > 5%) to the variability of each component. The legend intervals are determined according to a principle of symmetry with respect to the value 0, which corresponds to the total absence of contribution of the components to the variability of the data. Polygons representing accumulation areas of distal fall deposits from the Vesuvian eruptions of Avellino [27] and Pompei [28] were superimposed on the raster data of the PC2 map.

Sodium (Na), potassium (K), and copper (Cu) are associated with PC2, accounting for 21.8% of the total data variability, all with positive loadings (Table 2). The positive scores of this association mark a broad strip of the territory, spanning the central sector of the region, following a transverse axis from the volcanic coastal areas (including the Sorrento Peninsula) towards the Irpinia area, part of the inland Apennine domains. Positive signals are also observed on the western slopes of the Roccamonfina volcanic complex (to the north) and in some portions of the Sele River alluvial plain (to the south), while strongly negative score values characterize the remaining part of the territory. This well-defined spatial distribution of Na, Cu, and K suggests the existence of a source process for the association (Figure 7B). The biplots (Figures 5 and 6) show that observations with positive scores are predominantly associated with soils developed on alkaline lavas and volcano–sedimentary deposits, with a marginal presence of samples collected from alluvial or carbonate areas covered by pyroclastic deposits (Figure 5A,B), and with land dedicated, in order of abundance, to orchards and vineyards, agricultural areas, and wood (Figure 6A,B). Na and K are elements that unmistakably characterize the chemistry of most Vesuvian volcanoclastic deposits. An investigation of the main dispersion dynamics and the chemistry of products from major historical Vesuvian eruptions revealed an interesting overlap between the positive PC2 score values distribution and the main dispersion axes of the distal deposits from the

third (hydromagmatic) phase of the 2000 B.C. Vesuvian eruption (“Avellino Eruption”) [27]. Additionally, the positive signals characterizing the soils of the Sele plain can be associated with the main dispersion direction of the A.D. 79 Plinian eruption [28]. The presence of Cu in the association is interpreted as a consequence of the predominant horticultural use of these highly fertile lands; in fact, Cu is widely used in Campania as a key element in many phytosanitary strategies [29,30]. The PCA supports this interpretation performed on the original dataset (37 variables), as the Cu vector overlaps with that of phosphorus (P) (Figure 4), another element used in agriculture as a fundamental component of highly efficient organo-mineral fertilizers (such as NPK).

The elements correlated with PC3, which accounts for a smaller proportion (11%) of the overall data variability compared to the first two components, are calcium (Ca), sulfur (S), mercury (Hg), uranium (U), and lanthanum (La) (Table 2). Positive loadings individuate Ca and S, which exhibit higher absolute values than the other three elements, which have negative loadings. The highest positive scores characterize much of the inland areas, the southern part of the region, the soils of the alluvial plains of the Volturno and Sele Rivers, and, to a lesser extent, the Sarno River (Figure 7C). The biplots (Figures 5B and 6B) show that soils with high positive scores are predominantly developed on siliciclastic formations and, secondarily, on carbonates with pyroclastics and alluvial deposits (Figure 5B), and dedicated mainly to cultivated areas (Figure 6B). The high positive scores of Ca associated to the siliciclastic deposits are consistent with the general ability of clays to adsorb free Ca ions through ion exchange mechanisms. The elevated positive values found in the plains can also be linked to precipitation and adsorption mechanisms in the fine soil fractions of the Ca carried in solution by surface waters flowing from the carbonate reliefs at the edges of the Campanian Plain. Associated with Ca in the more clayey soils, we find S, probably because it was already present in higher concentrations and because clayey soils, compared to sandy ones, have a greater capacity to retain sulfates.

The results obtained by the application of the “two-step” PCA (using the approach proposed by Filzmoser et al. [26]) showed a two-fold advantage: the obvious reduction in the involved variables’ number and the highlighting of latent geochemical processes that are often not visible in the presence of associations of elements that are “redundant” and capable of conditioning most of the observed variability. The proposed approach made it possible to take into account, for the selection of the variable to be used in the second PCA, chemical elements that presented a good level of uniqueness in terms of modulus and direction of the relative vectors. Selecting only a representative variable for each group of overlapping vectors in the first PCA allowed us to diminish the influence of the prominent geological and anthropogenic sources allowing for a focus on lesser-known element associations that are indicative of more latent processes.

Furthermore, the interpretation of the results was significantly enhanced by classifying the samples according to their lithology and land use. While the methods employed are well-established in the literature, the proposed operational sequence introduces a novel approach that offers clear advantages in understanding the geochemical patterns of the Campania region. It is important to note that the findings relevant to one area cannot be directly applied to other regions with different geological characteristics and varying human impacts. Nevertheless, the proposed procedure is flexible and can be adapted to geochemical data from areas of varying size and nature without any inherent constraints.

5. Conclusions

The results obtained demonstrated how a better understanding of data characteristics can lead to more sound univariate and multivariate statistical analyses, reveal-

ing information that significantly contributes to an improved understanding of complex natural systems.

Furthermore, addressing the constraints imposed by the compositional nature of geochemical data by applying linear transformation techniques through Compositional Data Analysis (CoDA) can reduce the impact of potential spurious correlations among variables. This approach has enabled the identification of relationships between elements that would otherwise remain concealed, thereby highlighting the natural or anthropogenic processes that have occurred in the study area. Gaining insight into these processes is essential for assessing geological and environmental risks.

The use of the CoDA approach is certainly advantageous, especially when applying multivariate statistical techniques (e.g., PCA, Cluster Analysis, etc.), since these methods, originally developed to be used in social sciences, are based on the analysis of relationships between multiple variables featuring non-compositional data.

Author Contributions: Conceptualization, S.A. and L.R.P.; methodology, S.A., A.G. and L.R.P.; software, L.R.P.; validation, S.A. and A.G.; formal analysis, A.I., L.R.P., S.A. and A.G.; investigation, L.R.P.; resources, S.A.; data curation, A.G.; data collection and analysis, L.R.P., S.A. and A.G.; writing—original draft preparation, L.R.P., S.A. and A.G.; writing—review and editing, A.I. and A.G.; visualization, L.R.P. and A.G.; supervision, S.A.; project administration, S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study was carried out within the RETURN Extended Partnership and received funding from the European Union NextGenerationEU (National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3—D.D. 1243 2/8/2022, PE0000005).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Reimann, C.; De Caritat, P. *Chemical Elements in the Environment. Factsheets for the Geochemist and Environmental Scientist*; Springer: Berlin, Germany, 1998; p. 398.
2. Aitchison, J. *The Statistical Analysis of Compositional Data*; Chapman and Hall: London, UK, 1986.
3. Pawlowsky-Glahn, V.; Buccianti, A. *Compositional Data Analysis: Theory and Applications*; Wiley: Chichester, UK, 2011; p. 378.
4. Buccianti, A.; Lima, A.; Albanese, S.; Cannatelli, C.; Esposito, R.; De Vivo, B. Exploring topsoil geochemistry from the CoDA (Compositional Data Analysis) perspective: The multi-element data archive of the Campania Region (Southern Italy). *J. Geochem. Explor.* **2015**, *159*, 302–316. [CrossRef]
5. Egozcue, J.J.; Pawlowsky-Glahn, V. Simplicial Geometry for Compositional Data. In *Compositional Data Analysis in the Geosciences: From Theory to Practice*; Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V., Eds.; Geological Society Special Publications: London, UK, 2006; Volume 264, pp. 145–159.
6. Buccianti, A.; Grunsky, E. Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes? *J. Geochem. Explor.* **2014**, *141*, 1–5. [CrossRef]
7. Pearson, K. Mathematical contributions to the theory of evolution—On a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proc. R. Soc.* **1897**, *60*, 489–498.
8. Buccianti, A.; Mateu-Figueras, G.; Pawlowsky-Glahn, V. *Compositional Data Analysis in the Geosciences: From Theory to Practice*; Geological Society Special Publications: London, UK, 2006; Volume 264.
9. Aitchison, J. The statistical analysis of compositional data (with discussion). *J. R. Stat. Soc. Ser. B* **1982**, *44*, 139–177. [CrossRef]
10. Aitchison, J. The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. In Proceedings of the Keynote Address Presented at CoDaWork08, Girona, Spain, 27 May 2008; pp. 27–30. Available online: <https://core.ac.uk/download/pdf/132548276.pdf> (accessed on 5 January 2025).
11. Egozcue, J.J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barceló-Vidal, C. Isometric logratio transformations for compositional data analysis. *Math. Geol.* **2003**, *35*, 279–300. [CrossRef]
12. Minolfi, G.; Albanese, S.; Lima, A.; Tarvainen, T.; Fortelli, A.; De Vivo, B. A regional approach to the environmental risk assessment—Human health risk assessment case study in the Campania region. *J. Geochem. Explor.* **2018**, *184*, 400–416. [CrossRef]

13. De Vivo, B.; Petrosino, P.; Lima, A.; Rolandi Belkin, H.E. Research progress in volcanology in Neapolitan area, southern Italy: A review and alternative views. *Mineral. Petrol.* **2010**, *99*, 1–28. [\[CrossRef\]](#)
14. Peccerillo, A. *Plio-Quaternary Volcanism in Italy. Petrology, Geochemistry, Geodynamics*; Springer: Berlin/Heidelberg, Germany, 2005; ISBN 978-3-540-29,092-6.
15. Vitale, S.; Ciarcia, S. Tectono-stratigraphic setting of the Campania region (southern Italy). *J. Maps* **2018**, *14*, 9–21. [\[CrossRef\]](#)
16. Capozzi, V.; Rocco, A.; Annella, C.; Cretella, V.; Fusco, G.; Budillon, G. Signals of change in the Campania region rainfall regime: An analysis of extreme precipitation indices (2002–2021). *Meteorol. Appl.* **2023**, *30*, e2168. [\[CrossRef\]](#)
17. Guarino, A.; Albanese, S.; Cicchella, D.; Ebrahimi, P.; Dominech, S.; Pacifico, L.R.; Rofrano, G.; Nicodemo, F.; Pizzolante, A.; Allocca, C.; et al. Factors influencing the bioavailability of some selected elements in the agricultural soil of a geologically varied territory: The Campania region (Italy) case study. *Geoderma* **2022**, *428*, 116207. [\[CrossRef\]](#)
18. De Vivo, B.; Lima, A.; Albanese, S.; Cicchella, D.; Rezza, C.; Civitillo, D.; Minolfi, G.; Zuzolo, D. *Atlante Geochimico–Ambientale dei Suoli Della Campania*; Aracne Editrice: Roma, Italy, 2016; p. 364.
19. Salminen, R.; Tarvainen, T.; Demetriades, A.; Duris, M.; Fordyce, F.M.; Gregorauskiene, V.; Kahelin, H.; Kivisilla, J.; Klaver, G.; Klein, H.; et al. *FOREGS Geochemical Mapping Field Manual*; Geological Survey of Finland: Espoo, Finland, 1998; Guide 47.
20. Albanese, S.; De Vivo, B.; Lima, A.; Cicchella, D. Geochemical background and baseline values of toxic elements in stream sediments of Campania region (Italy). *J. Geochem. Explor.* **2007**, *93*, 21–34. [\[CrossRef\]](#)
21. Cheng, Q.; Bonham-Carter, G.F.; Raines, G.L. GeoDAS: A new GIS system for spatial analysis of geochemical data sets for mineral exploration and environmental assessment. In Proceedings of the 20th International Geochemical Exploration Symposium (IGES), Santiago, Chile, 6–10 May 2001; pp. 42–43.
22. Zuo, R.; Wang, J. ArcFractal: An ArcGIS add-in for processing geoscience data using fractal/multifractal models. *Nat. Resour. Res.* **2020**, *29*, 3–12. [\[CrossRef\]](#)
23. Jolliffe, I.T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2002. [\[CrossRef\]](#)
25. Peres-Neto, P.R.; Jackson, D.A.; Somers, K.M. Giving meaningful interpretation to ordination axes: Assessing loading significance in principal component analysis. *Ecology* **2003**, *84*, 2347–2363. [\[CrossRef\]](#)
26. Filzmoser, P.; Hron, K.; Reimann, C. Principal component analysis for compositional data with outliers. *Environmetrics* **2009**, *20*, 621–632. [\[CrossRef\]](#)
27. Sulpizio, R.; Cioni, R.; Di Vito, M.A.; Mele, D.; Bonasia, R.; Dellino, P. The Pomici di Avellino eruption of Somma–Vesuvius (3.9 ka BP). Part I: Stratigraphy, compositional variability and eruptive dynamics. *Bull. Volcanol.* **2010**, *72*, 539–558. [\[CrossRef\]](#)
28. Cioni, R.; Sbrana, A.; Gurioli, L. The Deposits of A.D. 79 Eruption. In *Vesuvius Decade Volcano Workshop Handbook*; Santacroce, R., Rosi, M., Sbrana, A., Cioni, R., Civetta, L., Eds.; Consiglio Nazionale delle Ricerche: Rome, Italy, 1996; pp. E1–E31.
29. Fagnano, M.; Agrelli, D.; Pascale, A.; Adamo, P.; Fiorentino, N.; Rocco, C.; Pepe, O.; Ventorino, V. Copper accumulation in agricultural soils: Risks for the food chain and soil microbial populations. *Sci. Total Environ.* **2020**, *734*, 139434. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Roviello, V.; Caruso, U.; Dal Poggetto, G.; Naviglio, D. Assessment of Copper and Heavy Metals in Family-Run Vineyard Soils and Wines of Campania Region, South Italy. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8465. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.