Journal of Geochemical Exploration 159 (2015) 302–316

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Exploring topsoil geochemistry from the CoDA (Compositional Data Analysis) perspective: The multi-element data archive of the Campania Region (Southern Italy)

A. Buccianti [a,b,*], A. Lima [c], S. Albanese [c], C. Cannatelli [c,d], R. Esposito [c], B. De Vivo [c]

[a] Department of Earth Sciences, University of Florence, Via G. La Pira 4, 50121 Firenze, Italy
[b] CNR-IGG (Institute of Geosciences and Earth Resources), G. La Pira 4, 50121 Firenze, Italy
[c] Department of Earth, Environment and Resources Sciences, University of Naples Federico II, Via Mezzocannone 8, 80134 Napoli, Italy
[d] Department of Geology, Universidad de Chile, Plaza Ercilla 803, Santiago, Chile

ARTICLE INFO

Article history:
Received 7 September 2015
Revised 4 October 2015
Accepted 10 October 2015
Available online 22 October 2015

Keywords:
Soil composition
Compositional Data Analysis
Log-ratio transformation
Campania Region
Multivariate analysis
Geochemical baseline

ABSTRACT

Soil geochemistry is often investigated by considering a large number of variables, including major, minor and trace elements. Some of the variables are usually highly correlated due to coherent geochemical behaviour, but the effect of anthropic factors tends to increase data variability, sometimes obscuring natural relationships governing their distributions. In this framework it may be difficult to identify geochemical features linked to natural phenomena as well as to separate geogenic anomaly from the anthropogenic ones. Consequently the identification of background/baseline values may be seriously compromised. However, knowledge about these reference terms is fundamental to manage and protect natural resources on different scales. Moreover, adequate estimations of background/baseline values are possible only if a sufficient number of chemical analyses are stored in complex repositories.

In this contribution the multi-element data archive of the Campania Region (Southern Italy) was explored from the CoDA (Compositional Data Analysis) multivariate perspective to characterise its structure. The archive contains abundance data of Al, As, B, Ba, Ca, Co, Cr, Cu, Fe, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, Sr, Th, Ti, V and Zn (mg/kg) determined in 3535 new topsoils as well as information on coordinates, geology and land cover. Under CoDA the proportionality features of abundance data are fully taken into account enhancing their relative multivariate behaviour in the correct sample space.

Results indicate that the structure of the whole matrix appears to be constituted by a core that geographically is mainly given by topsoils developed on volcanic materials and several outlier compositions whose origin is different. Anomalous compositions can originate from the robust barycentre all around when the following conditions are present: 1) high Na–K volcanic products, 2) limestones and dolostones with their terrigenous component, 3) flysch deposits or 4) fertiliser contribution.

The ($1 \times D$) robust barycentre of the whole dataset together with the variation array of the core represents the most frequent ($1 \times D$) multi-element vector as well as the proportionality relationships among its components. It might be considered a compositional baseline.

1. Introduction

In this paper a methodology to explore the information contained in a D multi-element data archive describing regional topsoil geochemistry is proposed. The approach is based on the use of consolidated multivariate procedures for compositional data (Aitchison, 1986) that for the first time have been sequentially combined to tentatively identify background/baseline compositions. These compositions have to be discriminated from anomalous ones, which are often related to peculiar conditions affecting limited portions of the territory and associated with restricted geochemical natural or anthropogenic conditions. From this perspective the background/baseline composition has to be considered the ($1 \times D$) most frequent vector representing the result of recurrent geochemical processes.

The approach does not consider single D variables one by one but the joint behaviour of the D elements that constitute the compositions. The comparison with the results that could be obtained by the analysis of single variables has limited value from a statistical point of view; therefore it does not represent a priority and the final aim of this paper, which is instead focused on multivariate analysis.

1.1. Background

Soils are a special component of the biosphere acting as both a geochemical sink for hazardous contaminants and a natural buffer

[*] Corresponding author at: Department of Earth Sciences, University of Florence, Via G. La Pira 4, 50121 Firenze, Italy.
E-mail address: antonella.buccianti@unifi.it (A. Buccianti).

http://dx.doi.org/10.1016/j.gexplo.2015.10.006
0375-6742/© 2015 Elsevier B.V. All rights reserved.

able to moderate the flux of matter among atmosphere, hydrosphere, and biota thus affecting the surficial cycles of elements and chemical compounds. The contamination of soils at regional scales occurs mainly in industrial regions where factories, motor vehicles and municipal wastes (and others) represent the most important exogenous source of trace metals. However, elements such as As, Se, Sb and Hg may be associated with long distance contamination by aerial transport due to their affinity with volatile phases.

Investigation of soil chemistry at regional scales is often aimed to determine the geochemical background or baseline for single elements (Salminen and Gregorauskiene, 2000; Sinclair, 1974, 1991; Stanley and Sinclair, 1989). Regardless of the developments in this field of investigation for the responsible use of natural resources, a shared protocol to determine background/baseline values, from sampling to statistical analysis, has not yet been universally adopted. One of the difficulties is to relate the results of scientific research published in peer-reviewed journals to applicative investigations that represent reports of public agencies (see COM (Commission of the European Communities), 2006; FAO (Food and Agriculture Organization of the United Nations), 2015; USEPA (U.S. Environmental Protection Agency), 2002). For instance, the term baseline does not have a common accepted definition as discussed by Gałuszka (2007). In fact, the term baseline is sometimes used interchangeably with background, while at other times it is used to describe only "natural" conditions, meaning those that are a result of the local conditions like geology, climate and hydrology, without reference to the influence of anthropogenic activity (Lee and Helsel, 2005; Reimann and Garrett, 2005; Rodrigues and Nalini Júnior, 2009; Salminen and Gregorauskiene, 2000). A discussion on the use of these terms can be found in Reimann and Garrett (2005) and Salminen and Gregorauskiene (2000). Consider that the definition used in the Forum of European Geological Surveys (FOREGS) project refers to the baseline as the concentration at a specific point in time of a chemical element, species or compound in a sample of geological material (De Vos et al., 2006; Johnson and Ander, 2008; Salminen et al., 2005). Recently, Levitan et al. (2014) state that "pre-operational baseline" is the summary of important geochemical factors (including element concentrations and other parameters) that characterise conditions of one or more media (e.g., soil, sediment, ground water, surface water) prior to mining or other anthropic development. Finally, Nordstrom (2015) discusses the meaning of the term "natural background" as related to the environmental conditions that existed before any mining or anthropogenic activities.

Concerning soil geochemistry, several methods have been developed to calculate the pristine contents of major, minor and trace elements and their relationships with soil parameters and geological factors, often based on statistical calculations (Kabata-Pendias, 2011; Tidball and Ebens, 1997). In this context Pedochemical Enrichment Factors (PEF) are commonly used. These factors are ratios calculated versus chosen reference values as for example Clarke's values of mother rocks normalised to Al or Ti contents, considered as "stable" elements. However, the use of enrichment factors (EFs) as a proof of anthropogenic impact is questioned in Sucharovà et al. (2012) who consider a soil profile as an open system, where elements exchange and their turnover in the biosphere determines soil formation. In this respect, high top-/bottom-soil ratios, or EFs, may highlight the geochemical de-coupling of the lithosphere from the biosphere rather than any contamination present.

1.2. Purpose and organisation of the research

All the approaches discussed in the previous paragraph do not take into account the peculiar features of compositional data that have a proportional and multivariate nature, including a dependence on non-compositional variables such as time or space (Buccianti and Gallo, 2013; de Caritat and Grunsky, 2013). Consequently, with the objective of investigating the whole covariance structure of a geochemical dataset and tentatively identifying a baseline composition ($1 \times D$, D number of variables) able to represent frequent and potentially spatially organised processes (following the FOREGS definition), a procedure is proposed and developed under the CoDA (Compositional Data Analysis) theory (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011). The research challenges identified in this contribution are motivated by the fact that over the last 25–30 years many of the characterisations of background/baseline values have focused on the application of different types of univariate approaches to diverse problem types (Levitan et al., 2014; Nordstrom, 2015; Reimann and Garrett, 2005; Salminen and Gregorauskiene, 2000; Sinclair, 1974, 1991; Stanley and Sinclair, 1989). In the majority of these studies the approach has been neither multivariate nor compositional. Thus the state of the art research challenges and future directions associated with the improvement of our understanding of environmental phenomena, that are multivariate and therefore compositional, require the development of a new way of thinking (Buccianti and Grunsky, 2014; Buccianti, 2015a).

The availability of an original wide spatial dataset of determinations of Al, As, B, Ba, Ca, Co, Cr, Cu, Fe, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, Sr, Th, Ti, V and Zn (mg/kg) in 3535 topsoils of the Campania Region (Southern Italy) has provided us with the opportunity to take on this new challenge. The concentrations of these elements were all above analytical detection limits, making them amenable to statistical analysis without any preliminary substitutions or imputations. However, the source repository contained also determinations for several other elements (Ag, Au, Be, Bi, Cd, Ce, Cs, Ga, Ge, Hf, Hg, In, Li, Nb, Pd, Pt, Rb, Re, S, Sb, Sc, Se, Sn, Ta, Te, Tl, U, W, Y, Zr) but most of the values were below the detection limits for not corresponding samples. When the data matrix is affected by the presence of a reasonable number of observations with values below the detection limit it is advisable to adopt the procedure proposed by Palarea-Albaladejo and Martin-Fernandez (2015). A new investigation is in progress to verify how the cited method is able to recover the information for left-censored data by considering the available multivariate information.

In a first exploratory phase, the application of clustering object-oriented procedure for large datasets was applied on centred log-ratio (clr) transformed data (Aitchison, 1982) finding for natural groups (Kaufman and Rousseeuw, 1990). The clr real coordinates were used as they reduce the computation of Aitchison distances to ordinary distances (Aitchison et al., 2000). Isometric log-ratio coordinates can be also used here (Egozcue et al., 2003). The aim was to verify if samples were structured or organised in such a manner that a link with geology or other spatial features was mainly well recognisable. This analysis was also useful to understand if the sampling plan, characterised by different densities in urban and not urban areas, could have some effect in grouping cases.

After having verified that no clear discriminations were statistically significant at a regional scale, in order to justify an a priori clustering of the data, the subsequent use of consolidated robust methods for compositions was applied to explore the whole repository.

Robust methods are developed because atypical observations in a dataset heavily affect the classical estimates. Outliers can occur by typesetting errors or malfunction of the experimental equipment. Another type of atypical observations is that that belongs to different populations due to a change in experimental or natural conditions (Verboven and Hubert, 2005).

The practice of robustly seeking multivariate outliers was introduced in applied geochemistry in the 1980s (Garrett, 1989; Garrett et al., 1982; Smith et al., 1983, 1984) and subsequently refined by Filzmoser et al. (2005). In our case, concentration values (mg/kg) were transformed by using the isometric log-ratio (ilr) conversion (Egozcue et al., 2003). The transformation permits the statistical analysis in the R^{D-1} sample space compared with the S^D constrained simplex. The robust estimates of the compositional centre $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$ were obtained by using the Minimum Covariance Determinant (MCD) estimator (Filzmoser and Hron, 2008; Filzmoser et al., 2011; Rousseeuw, 1984; Verboven and Hubert, 2005). Under the normal assumption, the compositions having a robust distance RD_i larger than the cut off-value $\sqrt{\chi^2_{D-1,0.975}}$ (χ = chi-squared sample distribution) were identified and discriminated from the rest of the data. Thus two different groups of data were obtained, A (matrix without the compositions of the tails, what geochemists would refer to as a background/baseline population) and B (only compositions of the tails) datasets.

Filzmoser and Hron (2008) and Filzmoser et al. (2011) followed a similar approach in mapping outliers of a subcomposition of the Kola moss data. However, their spatial analysis concerned samples discriminated by using their position in a quadrant of the log–log plot of the robust Mahalanobis distance calculated for log and ilr-transformed data. They asserted that horizontal and vertical lines dividing the plot in quadrants were obtained through the 0.975 quantiles of the corresponding χ^2 distributions, which were used as cut-off values for outlier identification. In our approach the RD_i single values are mapped instead. In this way it was possible to visualise the compositional continuous changes from the robust barycentre towards the tails of the multivariate distribution. Subsequently, a biplot was constructed to associate the identified compositional changes with the behaviour of the chemical elements.

After this step, the variables of the A dataset were transformed by re-using the centred log-ratio (clr) transformation to apply the clustering object-oriented procedure for large datasets (Kaufman and Rousseeuw, 1990). The aim was to check the internal homogeneity of group A (Daszykowski et al., 2007; Hubert et al., 2005; Verboven and Hubert, 2005). If the clustering of dataset A is not significant, the robust compositional barycentre, variation array, clr-variances, and total variance are all CoDA tools that can help in the characterisation of a baseline composition (Aitchison, 1986).

All the analyses were performed using routines developed in Matlab_R2014b and R (Everitt and Hothorn, 2011; R Foundation for Statistical Computing, 2014; Templ et al., 2011; van den Boogaart and Tolosana-Delgado, 2013; Verboven and Hubert, 2005; Wehrens, 2011).

2. Material and methods

2.1. Study area: Campania Region (Southern Italy) geomorphology and geology

Campania Region (Southern Italy, Fig. 1) presents different morphologies. The Eastern hilly and mountainous areas were formed by the Apennine orogen and are characterised by a series of peaks, plateau and highlands where the two main rivers (Volturno and Sele) flow. The western coastal area, developed in the NW–SE direction, is about 15% of the study area and represents a large structural depression now occupied by the Campanian and Sele plains and three volcanic complexes: Roccamonfina, Campi Flegrei, and Somma–Vesuvius. The plains, developed in subsiding grabens, have been filled by sediments originating from the erosion of the Apennine ridge and also from products of intensive volcanic activity.

Most of the Campania Region is occupied by the Apennine chain closely related to the structural events that formed the Italian peninsula and consisting of a salient north-east verging thrust and fold belt, interposed between the back-arc Tyrrhenian basin to the West and the undeformed Apulian–Adriatic foreland to the East (Bonardi et al., 2009). The Apennine orogen, beginning in the early Neogene (Lavecchia, 1987) and continuing today, formed a chain consisting of a series of nappes over-thrusting towards the N–NE, and structured in

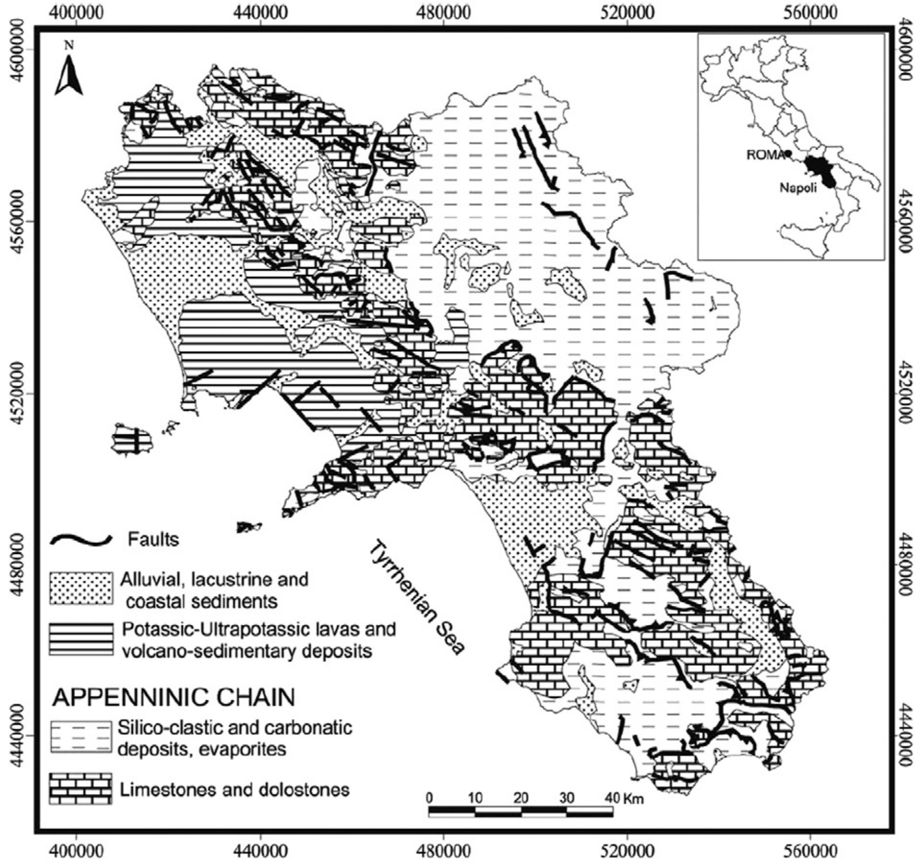
400000
440000
480000
520000
560000
4600000
4560000
4520000
4480000
4440000
N
ROMA
Napoli
Tyrrhenian Sea
Faults
Alluvial, lacustrine and coastal sediments
Potassic-Ultrapotassic lavas and volcano-sedimentary deposits
APPENNINIC CHAIN
Silico-clastic and carbonatic deposits, evaporites
Limestones and dolostones
0 10 20 30 40 Km

Fig. 1. Simplified geological map of Campania region, Italy.
From Lima et al. (2003).

several contiguous blocks along tectonic discontinuities (Patacca and Scandone, 2007). The current structure is the result of complex tectonic settings (Steckler et al., 2008). The east side of the Apennine fold and thrust belt was raised by compressional forces acting under the Adriatic Sea (Apennines Convergence Zone), while on the west side fault-block mountains prevail, created by spreading or extension of the crust under the Tyrrhenian Sea (Tyrrhenian Extensional Zone). The alkaline magmatism, characterised by high content of K_2O (De Vivo et al., 2010; Peccerillo, 2005), that extends and gets younger from Tuscany to Campania formed as the result of the rollback of a west dipping subduction zone (Rosenbaum and Lister, 2004 and references therein). The famous *Roman Co-magmatic Province* of Washington (1906) represents the subduction-related magmatism, which includes the magmatism in the Campanian Plain (Fig. 1). The latter is a graben bordered by Mesozoic carbonates that subsided during the Pliocene–Pleistocene to perhaps as much as 5 km (Ippolito et al., 1973) and place of fissural volcanic activities, with different ignimbrite eruptions (Campanian Ignimbrites), as well as of volcanic complex formations (Roccamonfina, Campi Flegrei, Vesuvio) (De Vivo, 2006, and ref. therein; De Vivo et al., 2001; De Vivo et al., 2010; Rolandi et al., 2003 and ref. therein; Milia and Torrente, 2000; Torrente and Milia, 2013). Outcropping lithologies, shown in Fig. 1, consist mostly of sedimentary and volcanic rocks, spanning from the Triassic to recent age. The sedimentary rocks include: i) stratigraphic Mesozoic Units, made up of limestones, dolostones, siliceous schists and terrigenous sediments (clays, siltstones, sandstones, conglomerates) that characterise mostly the external Apennine domains and, ii) the Neogene Units, made up mostly of silico-clastic, carbonatic and evaporitic sediments and Quaternary sediments represented by alluvial, lacustrine and coastal lake sediments and by pyroclastic fall and flow deposits.

2.2. Economy

Agriculture represents one of the most important economic activities of the Campania Region and in the Northern Territory agricultural activities cover more than 50% of the total available land surface (see electronic Supplementary material SM1). Farming is mostly located around coastal areas, due to the occurrence of mountainous areas and the lack of significant superficial drainage in the hinterland. Productivity is enhanced by the soils that are mainly volcanic and rich in nutrients. Tomatoes, potatoes, aubergines, peppers, peas, tobacco and citrus fruits are mainly cultivated in the plain areas, olive trees and vineyards mostly in the hilly sector. Despite the natural fertility of the soil, Campania is the foremost consumer of fertilisers in Southern Italy and azotic fertilisers represent 50% of the regional consumption per year.

Industries present a scattered spatial distribution and are mainly concentrated in the northern half of the region. The majority of industries have been developed next to large cities and around agricultural areas. They are mainly devoted to vegetable preserving processes, textile-apparel, clothes production and tannery. While textile industries have a low environmental impact (raw materials are produced and imported from foreign countries), inadequacies in waste-water systems in the tannery and vegetable preserving industries result in considerable pollution of stream and ground waters and sediments.

Transport communication networks (see electronic Supplementary material SM2) are highly developed in coastal areas and in the central-northern part of the region, due to strong economic pressure. Highways cross the whole territory but, especially during the summer, highway A3 in the Southern sector of Campania (Vallo di Diano) is not able to sustain the intense traffic that moves to Southern regions of Italy (Calabria and Sicily). No economic mineral deposits have been found in Campania; only a few minor bauxite mineral occurrences – of no economic relevance – occur in the Mesozoic rocks of Mt. Matese in the Apennine Mountains. The karst bauxite deposits are at present uneconomic, due to their very small dimensions and scattered distribution (Boni et al., 2013).

2.3. Sampling and experimental

From 2013 to 2014, 3535 new surface soil (topsoil) samples were collected from the Campania Region (13,600 km^2) on a 16 km^2 grid in the suburban and agricultural areas and on a smaller grid in urban areas (about 4 km^2) with a nominal density of 1 sample per about 3.8 km^2. Around 1.5 kg soil samples were collected at a depth between 5 and 15 cm under the ground surface after the removal of the vegetation cover, following the protocol according to the FOREGS sampling procedures (Plant et al., 1996; Salminen et al., 1998). Every 20 sampling sites a duplicate sample was collected in the same cell to allow the blind control of the cell sampling variability combined with analytical variability. At each sampling spatial coordinates, topography, local geology, type and main properties of soils, land use, and any additional detail related to anthropic activities in the surroundings were recorded. Every cell was uniquely identified by an alphanumerical code. After being dried with infra-red lamps at a temperature below 35 °C, the 3535 samples were pulverised in a ceramic mortar and then sieved to retain the <2 mm fraction. The pulps were stored in small plastic bags containing at least 30 g of samples, and then sent to ACME Analytical Laboratories Ltd. (Vancouver, Canada), where they were analysed by ICP-MS after aqua regia digestion for the determination of 53 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, Ta, Te, Th, Ti, Tl, U, V, W, Y, Zn and Zr. Precision of the analysis was calculated using 29 in-house replicates, and 5 blind duplicates submitted by the authors (median value of the Relative Percentage Difference, RPD = 1.3%). Accuracy was determined using in-house (ACME) reference materials (STD DS9, STD DS10, STD DS11, STD OXC109, median value 2.2%). Method detection limits for Al, Ca, Fe, K, Mg, Na, P, S and Ti range from 0.001 wt.% (Na, P, Ti) to 0.02 wt.% (S) and from 0.01 mg/kg to 5 mg/kg for all the other elements.

3. Statistical methods

In this paper our objective is to introduce a procedure that uses the theory of Compositional Data Analysis (CoDA) to investigate complex topsoil D multi-element databases. The aim is to possibly identify multi-element vectors ($1 \times D$) to be interpreted as baseline compositions instead of considering single elements. The motivations are in the relative nature of compositional data and in the complexity of natural phenomena, both of which require the joint investigation of several variables. The term baseline is used instead of background due to the presence of anthropogenic influences in most parts of the investigated region (Angelone et al., 2002; Cicchella et al., 2003; Cicchella et al., 2008; Vitrone, 2003). However, the approach can also be applied in pristine conditions to evaluate the background composition. From this perspective the background/baseline compositions have to be considered the ($1 \times D$) most frequent vector representing the result of recurrent geochemical processes.

To achieve this target, consolidated principles of the CoDA (Compositional Data Analysis) theory were followed (Aitchison, 1982, 1986), implemented by the application of multivariate robust methodologies (Daszykowski et al., 2007). The presented strategy can be applied to different environmental matrices and wishes to contribute to a change in the current way of approaching background or baseline estimation, generally neither multivariate nor compositional.

Compositional data are vectors of positive values quantitatively describing the contribution of D parts (variables) of some whole, which carry only relative information (Aitchison, 1982, 1986). Due to these features, the univariate Euclidean geometrical approach to the statistical analysis of compositions may give misleading results for two reasons: 1) compositional data pertain to the simplex sample space S^D (D constrained dimensions) and not to the real one R^D (Buccianti and Magli, 2011; Buccianti, 2013; Egozcue and Pawlowsky-Glahn, 2006; Pawlowsky-Glahn and Egozcue, 2015) and 2) only a multivariate approach permits an investigation of the relative behaviour of all the variables of the composition (Pawlowsky-Glahn and Egozcue, 2015).

The simplex sample space S^D is governed by the Aitchison geometry, and has all the properties of a $(D-1)$ dimensional Euclidean space (Egozcue and Pawlowsky-Glahn, 2006). To work in the unconstrained conditions of the real space, compositions need to be expressed as vectors of values that belong to such a space. To obtain these new vectors a family of log-ratio transformations can be applied. Aitchison (1982) proposed two types of transformation both based on logarithms of ratios of parts as a natural way of representing compositions, the additive log-ratio (alr) and the centred log-ratio (clr) transformation. The alr transformation from the simplex S^D to the real space R^{D-1} is defined as:

$$y = \mathrm{alr}(x) = \left[\ln\left(\frac{x_1}{x_D}\right), \ln\left(\frac{x_2}{x_D}\right), \cdots, \ln\left(\frac{x_{D-1}}{x_D}\right)\right]. \tag{1}$$

On the other hand the clr transformation from S^D to R^D is defined as:

$$y = \mathrm{clr}(x) = \left[\ln\left(\frac{x_1}{\mathrm{g}(\mathbf{x})}\right), \ln\left(\frac{x_2}{\mathrm{g}(\mathbf{x})}\right), \cdots, \ln\left(\frac{x_{D-1}}{\mathrm{g}(\mathbf{x})}\right)\right], \tag{2}$$

where $\mathrm{g}(\mathbf{x}) = [x_1 \bullet x_2 \bullet \ldots \bullet x_D]^{1/D}$ is the geometric mean of the composition $\mathbf{x}$.

A further transformation is the isometric log-ratio (ilr) one, proposed by Egozcue et al. (2003), but despite its theoretical advantages and practical properties, it leads to coordinates difficult to interpret from a geochemical point of view. However, if the concept of balance between groups of parts originated by a sequential binary partition is considered (Borgheresi et al., 2013; Egozcue and Pawlowsky-Glahn, 2005), the geochemical interpretation may be highly simplified. In a sequential binary partition in each of the $D-1$ steps of the procedure the compositional parts are divided into two non-overlapping groups; the resulting $D-1$ ilr variables represent balances between these groups in R^{D-1}:

$$\mathrm{ilr}_i = \sqrt{\frac{r \times s}{r+s}} \log \frac{\mathrm{g}(\mathrm{c}_+)}{\mathrm{g}(\mathrm{c}_-)} \tag{3}$$

with $i = 1, 2, \ldots, D-1$ and where $\mathrm{g}(\mathrm{c}_+)$ represents the geometric mean of the r variables of the numerator of the balance, $\mathrm{g}(\mathrm{c}_-)$ the geometric mean of the s variables of the denominator.

All the transformations alr, clr and ilr are representations able to enhance the geometric aspects involved (Egozcue et al., 2003); both the alr and ilr transformations are coordinates of a composition with respect to a basis, an oblique basis in the case of the alr transformation and an orthonormal one in the case of the ilr transformation. The clr transformation is also one-to-one representation but the clr coefficients are not coordinates with respect to a basis. Metric properties, particularly distances, are not easy to handle with alr coordinates in contrast with both clr coefficients and ilr coordinates. For this reason only the clr and ilr conversions were used in our analysis.

The abundance of Al, As, B, Ba, Ca, Co, Cr, Cu, Fe, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, Sr, Th, Ti, V and Zn (mg/kg), with all the values above the experimental detection limits, determined on 3535 new topsoil samples in the Campania Region, was first analysed using classical descriptive statistics (minimum, maximum, 1st and 3rd quartiles, median).

Subsequently, the clr and ilr transformations were applied to build up matrices of real coordinates (Egozcue et al., 2003). The clr-matrix was analysed by robust clustering methods with the aim to identify at first the presence of natural groups clearly discriminated and possibly related to the nature of the substrate. The procedure of clustering for large applications *CLARA*, proposed by Kaufman and Rousseeuw (1990) and implemented by Struyf et al. (1996), was applied (R package "cluster", 2015). It is an algorithm of the partitioning type that divides the dataset into k clusters, where the integer k needs to be specified by the user. Typically the algorithm is tested for a range of k-values and for each one a quality index, called the *average silhouette width* is determined so that a meaningful organisation structure of the data can be found when its values approximate 1.

We applied a procedure to verify that no anomalous compositions were present in the tails of the multivariate distribution, after to have controlled that data were not naturally clustered in different groups, so that an a priori discrimination was not justified. For this purpose, the robust Mahalanobis distance, labelled RD_i, was calculated on ilr coordinates. It was defined as:

$$RD_i = \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_{MCD})^T \boldsymbol{\Sigma}_{MCD}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_{MCD})}$$

with $\boldsymbol{\mu}_{MCD}$ and $\boldsymbol{\Sigma}_{MCD}$ the MCD location and scatter estimates. Robust estimates of the compositional centre $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$ can be obtained by using the Minimum Covariance Determinant (MCD) estimator (Filzmoser and Hron, 2008; Filzmoser et al., 2011; Rousseeuw, 1984; Verboven and Hubert, 2005). The proposed robust distance is a robustification of the Mahalanobis one where classical mean and empirical covariance matrix are used as estimates of location and scatter. Under the normal assumption, the compositions of the tails are those compositions having a robust distance larger than the cut-off value $\sqrt{\chi^2_{D-1,0.975}}$. If this statistical threshold is considered, the analysed matrix can be divided into two groups, from now on called dataset A (compositions below the cut-off) and dataset B (compositions above the cut-off). The regionalised structure of the RD_i values can be investigated to visualise how compositions move from the robust barycentre towards the tails. This phase of the analysis represents an improvement on the approach proposed by Filzmoser and Hron (2008) and Filzmoser et al. (2011) for the Kola moss data.

Subsequently, with the aim to associate the presence of the two datasets A and B with the behaviour of chemical elements, a biplot analysis was performed. A biplot is a graphical display used to represent simultaneously the rows and columns of any matrix by means of a rank-2 approximation (Gabriel, 1971). Aitchison and Greenacre (2002) adapted it for compositional data and proved it to be a useful exploratory tool. The philosophy and mathematics of this technique are well summarised in Pawlowsky-Glahn et al. (2015).

The association of the map of the RD_i values with biplot analysis can indicate not only where anomalous compositions are located in the regional territory, but also which variables are able to mainly characterise and determine the anomalous conditions.

Finally, with the aim to check for the presence of some natural grouping structure in the A dataset, the procedure of clustering for large applications *CLARA*, was again applied on clr-transformed variables (Kaufman and Rousseeuw, 1990; R package "cluster", 2015; Struyf et al., 1996). The check of the internal structure of the dataset A is fundamental for the potential identification of a representative baseline composition. If the clustering structure of the dataset A is not statistically significant, the robust compositional barycentre, variation array, clr-variances, total variance and variation coefficients of dataset A (Aitchison, 1982) may all be useful CoDA tools to fully characterise the baseline composition and its variability. All the analyses were performed using routines developed in Matlab_R2014b and R (Everitt and Hothorn, 2011; R Foundation for Statistical Computing, 2014; Templ et al., 2011; van den Boogaart and Tolosana-Delgado, 2013; Verboven and Hubert, 2005; Wehrens, 2011). The complete sequence of data analysis is reported in the flow chart of Fig. 2.

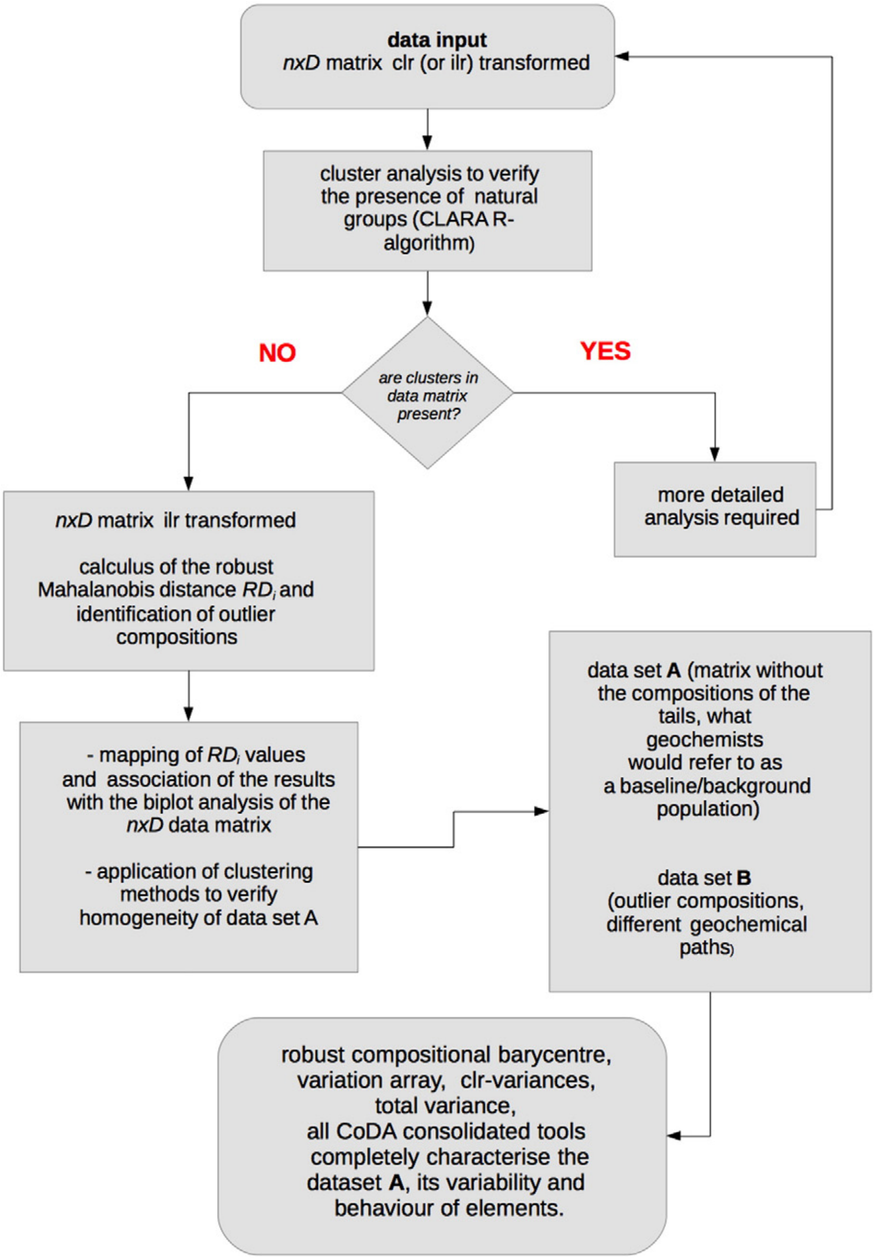
data input
nxD matrix clr (or ilr) transformed
cluster analysis to verify the presence of natural groups (CLARA R-algorithm)
NO
YES
are clusters in data matrix present?
more detailed analysis required
nxD matrix ilr transformed
calculus of the robust Mahalanobis distance RD_i and identification of outlier compositions
- mapping of RD_i values and association of the results with the biplot analysis of the nxD data matrix
- application of clustering methods to verify homogeneity of data set A
data set **A** (matrix without the compositions of the tails, what geochemists would refer to as a baseline/background population)
data set **B** (outlier compositions, different geochemical paths)
robust compositional barycentre, variation array, clr-variances, total variance, all CoDA consolidated tools completely characterise the dataset **A**, its variability and behaviour of elements.

Fig. 2. Flow chart illustrating the CoDA tools used to completely characterise the data structure of a multi-element database and to possibly identify a $(1 \times D)$ compositional baseline.

4. Results

4.1. Descriptive statistics

As a first explorative analysis, information for each variable of this study is given as a range of values in mg/kg, including the minimum and maximum values, the 1st and 3rd quartiles and the median (Table 1). For comparison the median values of the GEMAS (Geochemical Mapping of Agricultural and grazing land Soils, Reimann et al., 2014a, 2014b) databases are also reported. By considering the nature of the samples collected in the Campania Region and the used analytical methods, the comparison with GEMAS data appears to be appropriate.

4.2. Identification of the outlier compositions

The identification of outlier compositions often located in the tails of the multivariate distribution of the whole dataset was performed on the ilr coordinates. This step was applied after having verified that the whole matrix does not contain statistically significant natural groups, a condition that justifies an a priori discrimination eventually due to the parent material or to the sampling strategy (higher density in urban areas). In fact the *average silhouette width* associable with a meaningful organisation structure of the data was lower than 0.5 for different cluster solutions (Kaufman and Rousseeuw, 1990; Struyf et al., 1996).

In Fig. 3a the proportionality between the robust distance values and the square root of the quantiles of the chi-squared distribution is reported. A linear pattern allows the enforcing of the hypothesis of multivariate normal assumption. In Fig. 3b, the relationship between the Mahalanobis (classical) distance and the robust one is visualised. Horizontal and vertical lines are drawn at the cutoff-value $\sqrt{\chi^2_{D-1,0.975}}$ for both. Classical and robust analyses identify the same atypical observations as they exceed the horizontal and vertical cut-off lines (Verboven and Hubert, 2005). Observations with a large robust but a

Table 1
Descriptive statistics (values in mg/kg) of the whole data. For comparison the median values of the GEMAS databases are also reported.

	Minimum	1st quartile	Median	3rd quartile	Maximum	GEMAS median
Al	2100	26,300	40,900	53,600	94,700	10,993
As	1.00	8.00	12.00	15.00	164	5.48
B	0.50	8.00	11.00	19.00	98	2.42
Ba	9.00	217.80	367	541	2631	61.70
Ca	800	12,700	22,500	44,200	295,200	3035
Co	0.50	6.70	10.20	13.50	79.00	7.46
Cr	0.25	9.05	14.00	20.55	808	20.2
Cu	2.51	32.02	63.09	129	2394	14.50
Fe	1600	19,300	24,800	30,500	154,600	17,199
K	400	4700	9500	18,500	68,200	1250
La	0.90	31.00	41.00	50.55	162	14.35
Mg	700	3800	5700	8100	104,600	2861
Mn	77.00	646	775	965	7965	445
Mo	0.06	0.85	1.24	1.88	62.15	0.42
Na	20	700	2660	5920	29,490	48
Ni	0.50	9.80	14.70	18.20	101	14.72
P	50	720	1260	2380	16,620	653
Pb	3.12	36.34	54.37	79.46	2052	15.81
Sr	4.60	99.20	153	240	1153	18.10
Th	0.30	8.00	12.30	16.00	59.10	2.89
Ti	5.00	735	1210	1600	2900	85.70
V	5.00	43.00	60.00	87.00	224	25.35
Zn	11.40	69.45	91.10	129	3211	45.05

small Mahalanobis distance are not recognised with a classical approach (points on the left of the vertical line). In our case, on the total of 3535 cases, 1129 (31.9%) have a robust distance higher than the statistical threshold thus constituting the dataset B, discriminated from the dataset A (2406 cases, 68.1%) whose compositions are below the chosen cut-off value ($RD_i < 6.1$). The application of the statistical tests contained in the MVN R-package (Korkmaz et al., 2015) permitted the evaluation of the multivariate normality assumption for dataset A.

Table 2 reports the cross-table between the membership for A and B datasets and the class that identifies the geological nature of the terrains from which the topsoil samples originated. Results are significant ($p < 0.01$, contingency coefficient and Phi and Cramer's V), indicating that the discrimination between the core of the distribution and its tails are affected by the nature of the bedrock, with some lithologies more represented in dataset A (volcano-sedimentary deposits, alluvial materials, limestones and dolostones) than in B (silico-clastic and carbonate deposits, ultrapotassic volcanic compositions). However, the wide overlapping of topsoils collected on different parent materials confirms the results of the previous clustering analysis.

4.3. Biplot analysis and geochemical evaluation

If the values of the robust distance RD_i are spatially analysed, discriminating between datasets A and B, we obtain the map in Fig. 4. As we can see, values lower than the threshold mainly characterise the central-northwestern sector of the region principally following the outcrop of volcano-sedimentary deposits and alluvial materials, including some limestone and dolostone outcrops. On the other hand, higher values mainly follow the outcrop of sedimentary rocks (principally flysch deposits) as well as some volcanic deposits characterised by ultra-potassic composition (Ischia island and Vesuvius volcano area), limestones and dolostones.

In order to link the RD_i spatial behaviour with that of the chemical elements, the biplots reported in Fig. 5a and b have to be analysed. The graphical display is able to represent about 65% of the data variability. In Fig. 5a, the position of the elements in the space originated by the two extracted principal axes is reported. The distance from the (0,0) compositional barycentre is related to the variance of the clr-variables with comparison to the full composition. Consequently Ti, Na, Cr, Ni and Ca are expected to have higher clr-variances ($\sum_{i=1}^{D} var[\mathrm{clr}_i](\mathbf{x})$) thus governing the variability of the whole dataset. On the other hand, Ba, Mo, Pb, Zn and V are the elements with the lower clr variances. On the whole, the position of the elements characterises the different quadrants of the biplot. Moving from the upper part clockwise, the passage from more felsic associations (Th, La, As, Al) towards the more femic ones (Fe, Mn, Co, Cr, Ni) can be observed. Moving again clockwise, the dominant elements are Ca and Mg, followed by Cu, Sr, P and B, then Na and K and finally Ti. Vanadium, Mo, Zn, Ba and Pb occupy the central part of the biplot, near to the compositional barycentre, due to their lower clr-variance.

In Fig. 5b the positions of the samples are reported in the same space. The wide overlapping between cases pertaining to datasets A and B is evident. However, most of the topsoils of dataset A are positioned on the left while topsoils collected on flysch, clay, arenaceous formations and sand and pebbles are mainly located on the right. Exceptions are the ultra-potassic lavas of dataset B, which overlap the cases of dataset A on the left.

The application of the cluster object-oriented analysis on the clr-variables for dataset A has revealed only the presence of a weak grouping structure, with the quality index again below 0.5 for several solutions of clustering. Consequently, we can hypothesise the identification of a sufficiently homogeneous set of data (Hubert et al., 2002, 2005; Kaufman and Rousseeuw, 1990) so that the robust barycentre could be the ($1 \times D$) compositional baseline.

The robust barycentre is characterised by the following values for single elements (mg/kg): Al = 45,623; As = 13.52; B = 16.04; Ba = 459; Ca = 29,577; Co = 10.38; Cr = 14.27; Cu = 124; Fe = 25,932; K = 16,782; La = 44.74; Mg = 6792; Mn = 812; Mo = 1.52; Na = 4207; Ni = 13.74; P = 1859; Pb = 76.33; Sr = 195; Th = 13.87; Ti = 1355; V = 73.31; and Zn = 120. These values were obtained by back-transforming the ilr barycentre values of the Matlab routine of Verboven and Hubert (2005), knowing the adopted original sequential binary partition (Egozcue and Pawlowsky-Glahn, 2005).

With the aim to have a complete characterisation of the compositional baseline, some information about the variability of dataset A is needed. A way to have some insight into this item is to inspect the variation array of dataset A (Aitchison, 1986). The variation array is a tool to describe dispersion in a compositional matrix and contains all the values $var[\log(x_i/x_j)]$ representing the usual variance of the log-ratio of parts i and j of the composition. When these values are multiplied by $1/\sqrt{2}$ the usual variance of the balances of parts i and j is obtained (Pawlowsky-Glahn and Egozcue, 2015). These relative variances substitute the use of correlation interpretation that for compositional data may be misleading. In fact $var[\log(x_i/x_j)] = 0$ means a perfect relationship between x_i and x_j in the sense that the ratio x_i/x_j is constant replacing the idea of perfect positive correlation with that of perfect proportionality (Aitchison, 1997). Moreover, the larger the value of $var[\log(x_i/x_j)]$ the more the departure from proportionality with $var[\log(x_i/x_j)] = \infty$ replacing the idea of zero correlation or independence between x_i and x_j.

The inspection of the variation array for dataset A (see electronic Supplementary material SM3) permits us to obtain, through the clr-variances, the following sequence of decreasing contributions to the total variance (equal to 4.65): Na > Ca > Cu > Cr > Pb > K > Ni > Th > P > Mg > Ti > Zn > As > La > Mn > Sr > B > Mo > Co > Al > Ba > Fe > V. As a comparison consider that for the dataset B the total variance is 14.16 while the elements characterised by the higher clr-variances are Na, Ti, Ca, Cr and Ni (see electronic Supplementary material SM3), all the elements showing the longer vectors in the biplot of Fig. 5. In searching for proportionality in dataset A ($var[\log(x_i/x_j)] \rightarrow 0$) geochemical relationships are expected for elements involved in the following log-ratios: 1) Al/As, Al/Ba, Al/Fe, Al/La, Al/Th, Al/Ti and Al/V; 2) As/La and As/Th; 3) Ba/V; 4) Co/Fe, Co/Ni and Co/V; 5) Fe/La, Fe/Mn and Fe/V; and 6) La/Mn and La/Th. The not-opposite position of the vectors in the biplot of Fig. 5a indicates

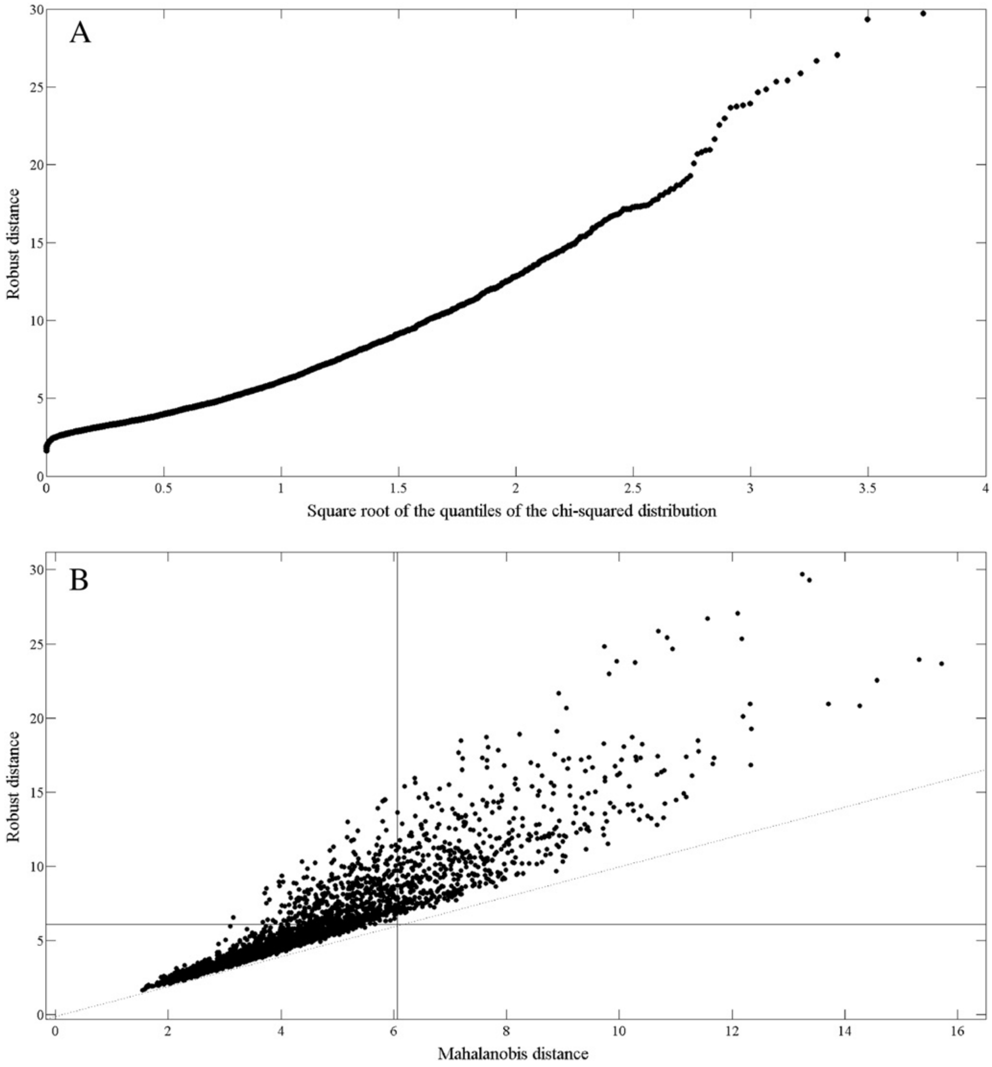

Fig. 3. a. Proportionality between the values of robust Mahalanobis distance and the square root of the quantiles of the chi-squared distribution. b. Relationship between the Mahalanobis (classical) distance and its robust version.

that the relationships are also mainly positive, as they point out a coherent geochemical behaviour.

A way to characterise the variability of the robust barycentre at the level of its *D* components, sometimes more useful for practical purposes, is to consider simple balances of the ratios x_i/x_j $(1/\sqrt{2} \times \log(x_i/x_j))$ of dataset A (van den Boogaart and Tolosana-Delgado, 2013). The previous results about proportionality (ratios for which $var[\log(x_i/x_j)] \rightarrow 0$) can be considered in order to choose appropriately the terms of the ratio.

Table 2
Cross-table between datasets A and B and the lithology of the sampling sites.

		Dataset		
		A	B	Total
Lithology	(1) Alluvials and mixed sediments	407	137	544
	(2) Clays	13	27	40
	(3) Limestone and dolostones	342	172	514
	(4) Evaporites	35	21	56
	(5) Flysch	100	269	369
	(6) Arenaceous formations	18	33	51
	(7) Lavas, pyroclastic rocks, ignimbrites	1454	435	1889
	(8) Sands and clays	1	6	7
	(9) Sands and pebbles	33	28	61
	(10) Travertines	3	1	4
Total		2406	1129	3535

If the balance follows a normal distribution, then the 95% confidence interval for mean can be determined. On the contrary robust statistics can be used as median and interquartile range. Consider, for example, the balance between Al and Fe characterised by the lower value for $var[\log(x_i/x_j)]$. The 95% of confidence interval for mean (0.382) is equal to (0.376, 0.388). The value of the balance for GEMAS data is equal to −0.32 while for the average continental crust is equal to 0.27. A comparison with ratios obtained from legal references used in different countries could also be possible (Kabata-Pendias, 2011; Kabata-Pendias and Sadurski, 2004). Finally, the chosen balance of parts *i* and *j* can be also mapped (Fig. 6) and geostatistical analysis applied without bias (Pawlowsky-Glahn et al., 2015). As we can see, higher values of the balance characterise different lithologies both in the western and eastern parts of the region, confirming the wide overlapping among topsoils originated from different source materials and the results of the clustering procedure.

5. Discussion

Generally, the spatial distribution of values for chemical elements at the regional scale is mostly controlled by lithology of the area, surficial natural processes, mineralisations, if present, or anthropic activity. As shown in Fig. 1, the lithology of the Campania Region is characterised by a high variability, even if it can be roughly divided into three main

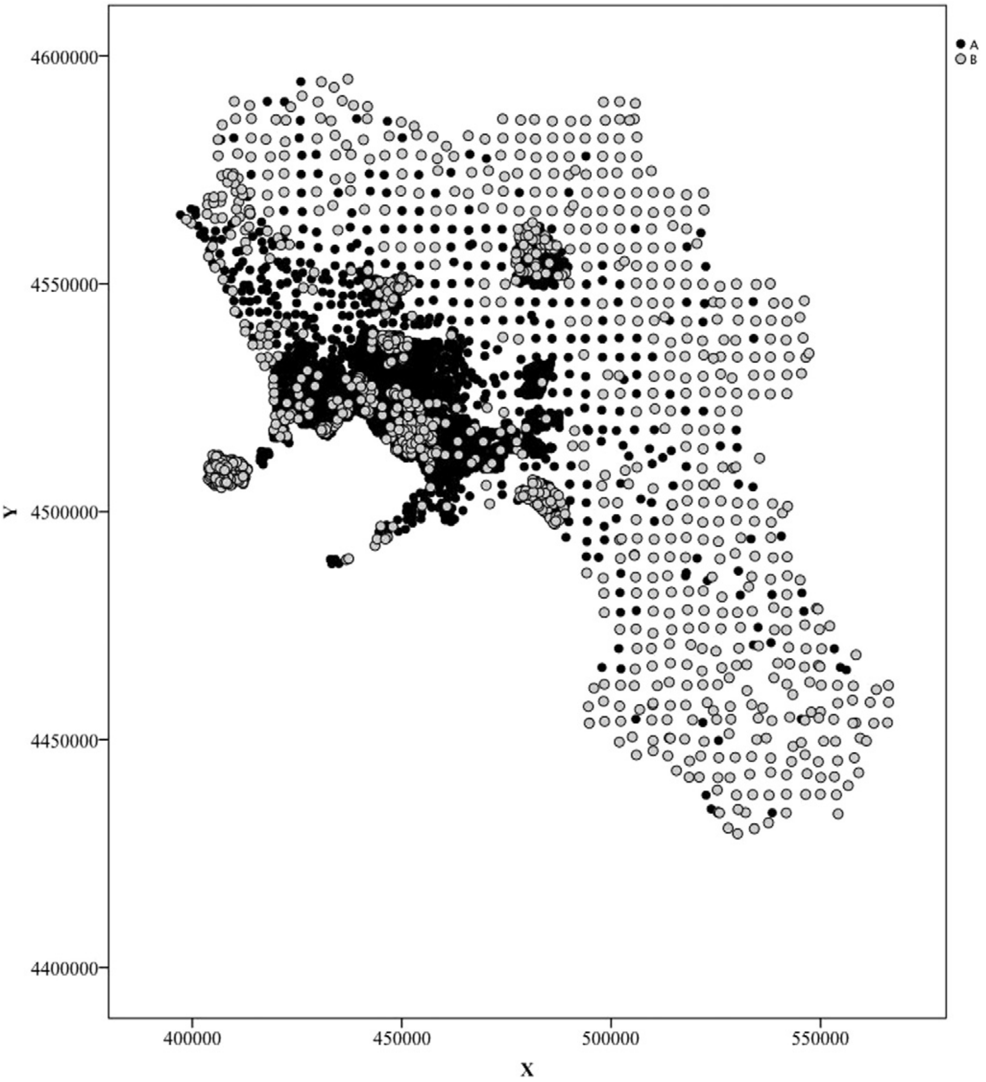

Fig. 4. Map of the robust Mahalanobis distance RD_i, discriminating datasets A and B ($RD_i = 6.1$ threshold value).

areas: the eastern one made up of silico-clastic deposits, the central one made up mostly of limestones and dolostones, and the western part of volcanoclastic, volcano-sedimentary and alluvial and coastal sediments. For some elements, the influence of anthropic activity cannot be ruled out as in the case of Cu and P, sometimes influenced by the agricultural practices in cultivated areas (see Supplementary material SM1), or hazardous metals due to industrial activity or illegal waste disposal (see Supplementary material SM2; Angelone et al., 2002; Cicchella et al., 2003; Cicchella et al., 2008; Vitrone, 2003).

Major factors in determining rock weathering and soil formation are chemical and mineralogical composition, grain size, landscape position and the properties of the circulating waters (Macias and Chesworth, 1992). Phases inherited by the parental materials are predominantly silicates with a secondary important group given by carbonates (calcite and dolomite). Secondary phases are mainly given by Al–Si clay minerals, oxides and hydroxides with sulphides, sulphates and chlorides occurring under special conditions such as abandoned mines (Bini, 2011) or evaporite outcrops (Dazzi and Monteleone, 2001). In the soils of Italy the nature of the parental rock can often be masked by the effects of climate and vegetation, that play a fundamental role in determining soil properties and development (Costantini and Dazzi, 2013). For example, the heterogeneous flysch complexes with clayey matrix that are diffused in the southern Apennines of Campania present active morphological dynamics with landslides and strong erosion phenomena, thus giving the formation of thin soils. The Somma–Vesuvius complex, near Naples, includes soils with weak differentiation as well as calcareous soils developed on pyroclastic–calcareous deposits and weakly developed soils on the lapilli from the last eruption of 1944 (Costantini and Dazzi, 2013). At Roccamonfina parental material is mainly pyroclastic with different degrees of weathering so that the most significant feature is the clay translocation (Lulli, 2007).

The result of cluster analysis on clr-variables confirms the presence of wide overlapping in the chemical composition of topsoils, notwithstanding the different source materials. The fingerprint of the parent bedrock can be sufficiently recognised but it is not able to clearly discriminate groups of samples spatially well located and related to the geology of the sampling sites (Table 2). The combined effect of climate, vegetation, use of the soils and their scarce development as well as anthropic contribution partially mask the multivariate fingerprint of the source and tend to increase the chemical variability of the different lithologies.

Descriptive classical numerical statistics determined for Al, As, B, Ba, Ca, Co, Cr, Cu, Fe, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, Sr, Th, Ti, V, and Zn concentrations (mg/kg) for 3535 topsoils in the Campania Region are reported in Table 1 together with the median of the GEMAS project (Reimann et al., 2014a; Reimann et al., 2014b) obtained for very similar geological media and experimental methods of analysis. All our data (Table 1) show higher element contents relative to the median values

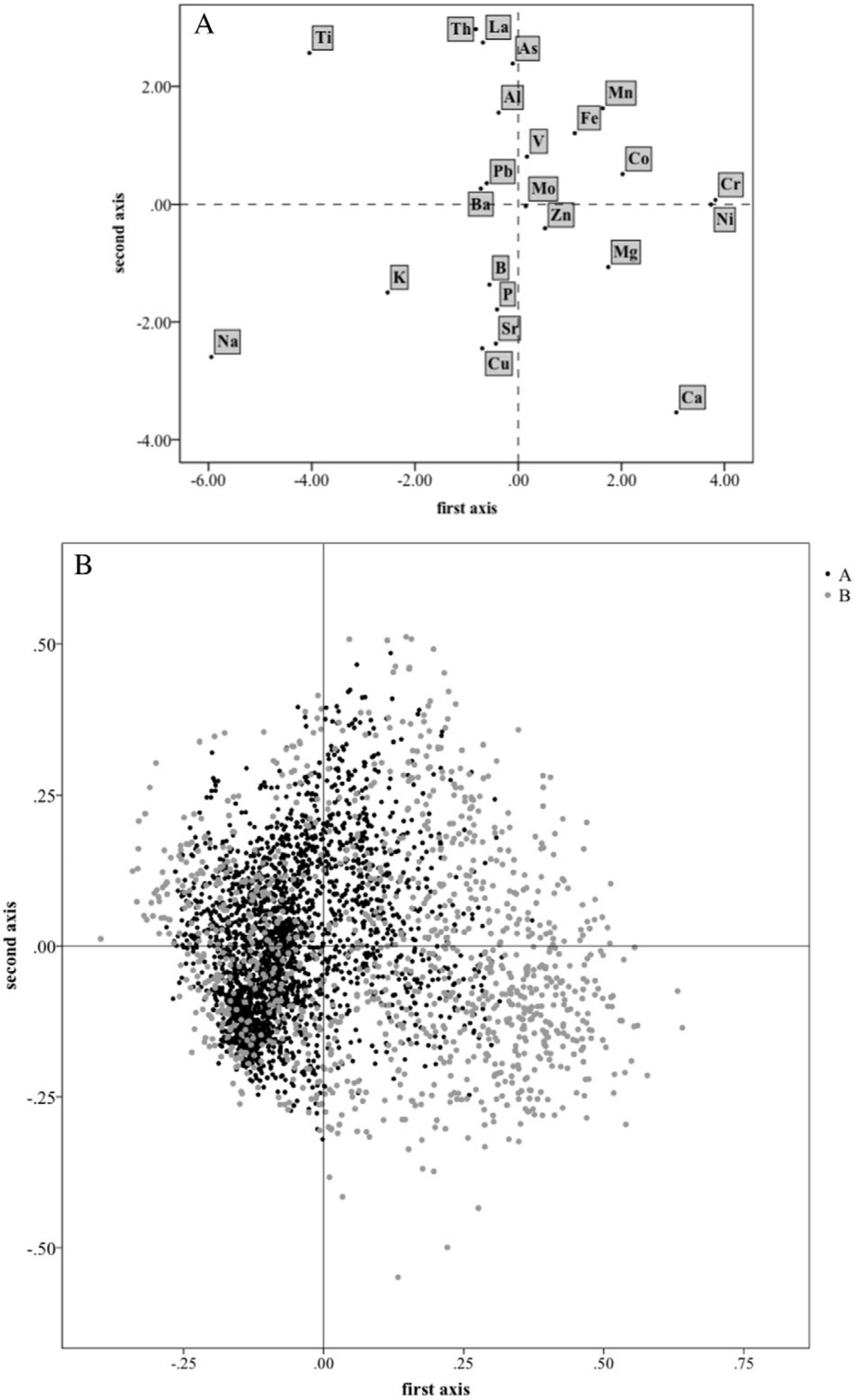

Fig. 5. a. Biplot analysis for the Campania multi-element dataset. Position of the elements clr-transformed in the space originated by the two principal axes. The graphical display explains about 65% of the total variability. b. Biplot analysis for the Campania multi-element dataset. Position of the cases of datasets A (black points, core of the distribution) and B (grey points, anomalous compositions) in the space originated by the two principal axes. The graphical display explains about 65% of the total variability.

of the GEMAS with Cr and Ni abundance representing the only two exceptions. These elements on a European level are associated with the presence of ophiolitic deposits and mineralisations. Marked differences characterise Na and Ti, elements typically present in the geological outcrops of the western and central part of the Campania Region where outcrops of lavas, pyroclastic deposits, ignimbrites, limestone and dolostone dominate.

The analysis of Fig. 3a indicates that the ilr-transformed abundance joint distribution is not multivariate normal, mainly due to the presence of anomalous compositions located in its tails. Consequently, simple lognormal processes generated by the product of many independent random compositions multiplied together are not able to reliably describe the observed geochemical behaviour (Ott, 1990, 1994). The synthesis is that the superimposition of several enrichment/depletion

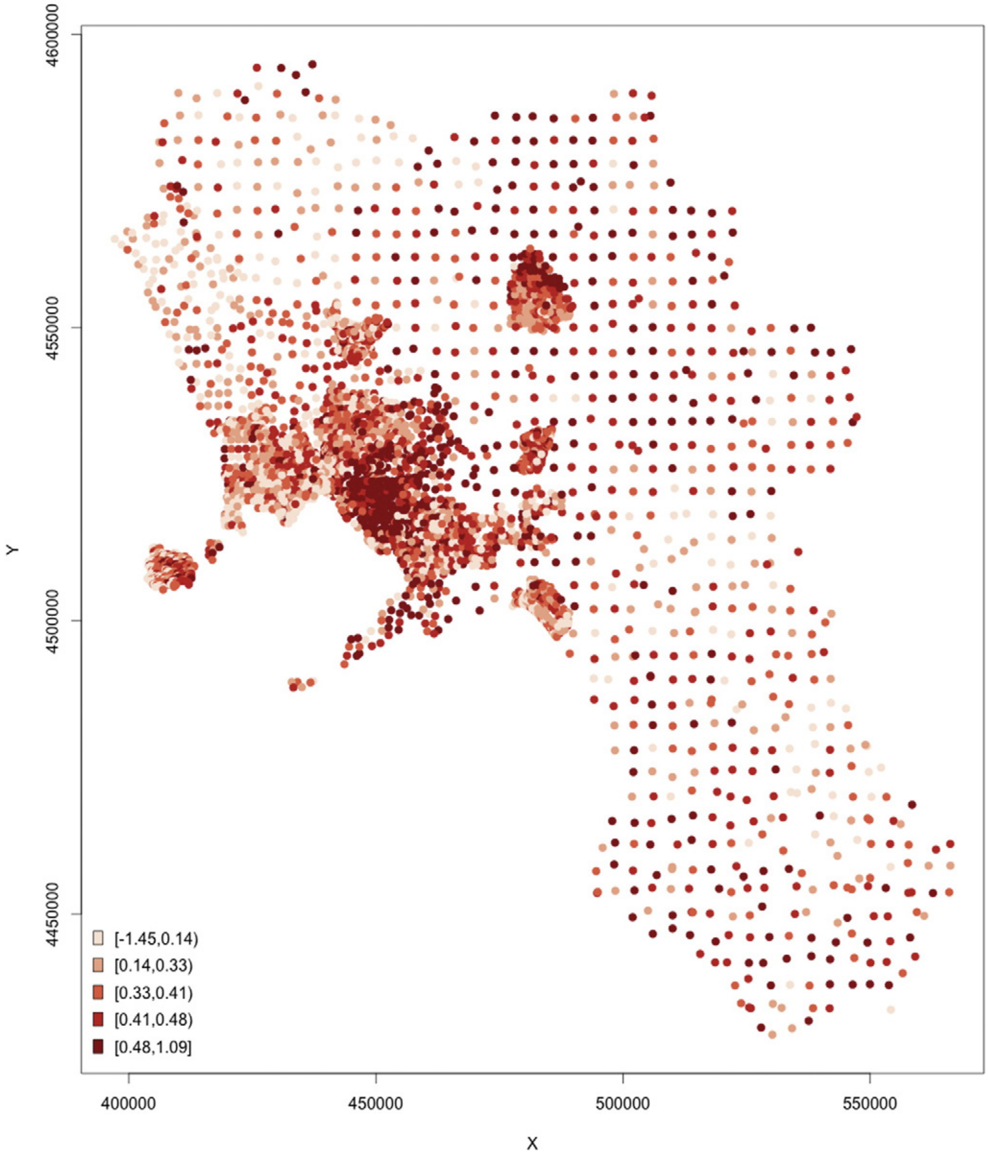

Fig. 6. Map of the balance between Al and Fe, that is 1/√2 log(Al/Fe). The values have been divided into five classes by using quantiles. The balance for GEMAS data is equal to −0.32 while for the average continental crust the value is equal to 0.27.

episodes seems insufficient to explain the transfer mechanisms of the elements from parent material to topsoils. Thus, multimodality, due to the different source materials, is partially evident again (Allegre and Lewin, 1996; Buccianti, 2011, 2015b).

The map of the RD_i values (Fig. 4) indicates that the core of the distribution (dataset A) is mainly located in the north-western part of the region and that all around a transition towards the tails can be observed with overlapping of the datasets A and B in several areas. This condition is well described by the biplots of Fig. 5a and b, indicating that the tails can be generated both by high values of Na, Ti and K as well as by that of Cr, Ni and Ca.

Among all these elements, Ca and K are considered macronutrients and their high variability in topsoils is related to the different source lithologies. For Ca values range from 800 mg/kg in topsoils on flysch to 295,200 mg/kg on limestones and dolostones. The range in agriculture soils is 7000–500,000 with the world average equal to 13,700 (Alloway, 2005). For K values range from 400 mg/kg on limestones and dolostones to 68,200 mg/kg on lavas, pyroclastic deposits, and ignimbrites. The range in agriculture soils is 400–30,000 with the world average equal to 8300 (Alloway, 2005). Considered a common constituent in plants, Na is not an essential element and can be dangerous in high concentrations with detrimental effects on plant growth (Tavakkoli et al., 2010). Sodium presents the minimum content in topsoils on flysch deposits (20 mg/kg) and the maximum on volcanic deposits (29,490 mg/kg). The range in agriculture soil is 750–7500 mg/kg while the world average in agriculture soils is 6300 (Alloway, 2005).

A similar behaviour is shown by Ni, considered a micronutrient, and Cr, both displaying their lower content (0.3 and 0.5 mg/kg respectively) in topsoils on lavas, pyroclastic deposits, and ignimbrites and their maximum content in topsoils on limestones and dolostones (Cr, 808 mg/kg) and in alluvial and mixed sediments (Ni, 101 mg/kg). Chromium and Ni tend to coprecipitate with Fe and Mn oxides after weathering or to associate with carbonates, phosphates and silicates.

Titanium is considered a stable element that generally does not give any environmental problem due to a low mobility (Kabata-Pendias and Szteke, 2015). Its lower contents are in topsoils on flysch and arenaceous formations (5 mg/kg) while the higher ones are in volcanoclastic materials and limestones and dolostones (2900 mg/kg). The abundance of Ti in topsoils developed on calcareous parent material suggests the incorporation of a terrigenous component.

Summarising, the variability of Na, K, and Ti mainly marks the presence of extreme volcanic compositions outcropping in the areas of Ischia, Vesuvius and the Salerno gulf, while Ca marks the presence of the carbonatic component that locally may or may not have incorporated terrigenous material. Cr and Ni behaviour, besides their origin from a mafic bedrock, is instead highly influenced by dispersion mechanisms that characterise sedimentary material, particularly flysch deposits. They participate in redox (with Fe and Mn, in the same part of the biplot) and pH-mediated adsorption/desorption (with Co, in the same part of the biplot) reactions (Gurumurthy et al., 2014) and their abundance can be conditioned by the presence of organically bound forms (Kabata-Pendias, 2011). Thus dataset A represents the core of the multivariate distribution of the topsoils whose parental material is mainly represented by volcanic material while dataset B includes compositional changes towards different geochemical compositions for topsoil chemistry as deduced by the dominant element association (Fig. 5a, b).

A deep inspection of the biplot permits us to obtain further interesting information about the behaviour of the elements during weathering processes that have affected the parent material. Na and K show their higher contents in topsoils from volcanic products while Ti, besides these, is highly associated also with topsoils from limestone and dolostone. This condition may be associated with its limited mobility due to adsorption by minerals of the terrigenous component mixed with the calcareous one. Th, La and As present their higher content in topsoils first from limestone, then from lavas and flysch deposits (La). This indicates that the most important processes, affecting their mobility during weathering, are related to the formation of slightly soluble compounds, adsorption by clays, coprecipitation with Fe–Mn oxides and presence of organic matter. Aluminium, a main component of several common minerals, is in fact present in the same part of the biplot showing its higher values in topsoils from alluvial and mixed terrain, limestones and lavas. Similar considerations can be reported for V, Fe, Mn, Mo, Zn, and Co, whose higher contents are found in topsoils from limestones and lavas (V), lavas (Fe), flysch (Mn), arenaceous material (Mo) or flysch and clay (Co). The low clr-variance of Mo, Zn and V indicates that they maintain similar abundance in the entire region.

Lead and Ba present their higher contents in topsoils from volcanic material. During weathering processes they are strongly adsorbed by clays, oxides and hydroxides and soluble organic matter (Pb). Their low clr-variance indicates a similarity with the behaviour of Mo, Zn and V.

Boron and P show their higher contents in topsoils from volcanic material even if they may have a common origin from fertilisers. Boron during weathering processes is solubilised and then entrapped in the clay lattice forming B-silicate compounds. Adsorbed B on soil minerals is leachable but the irreversibility of B sorption is documented (Kabata-Pendias, 2011). The solubility of apatite and Fe–Al phosphate and the absorption of phosphate on clay minerals are the two most important factors in the weathering of minerals containing P. In soil P also forms low solubility minerals with Pb and Ca while the sorption on alumina-silicate clays and hydrous oxide of Fe and Al depends on soil pH.

Strontium and Cu present their higher abundance in topsoils on flysch deposits. Strontium is moderately mobile in soils and it is likely to be sorbed in hydrated form by clay minerals and Fe oxides and hydroxides, a fate similar to Cu for which bio-accumulation may be also an important phenomenon. The proximity of Sr and Cu with P (and B) in the biplot (Fig. 5a) suggests for a role played by the use of fertilisers (Kabata-Pendias and Szteke, 2015).

Magnesium and Ca show their higher content in topsoils from limestones and dolostones (Mg and Ca), lavas (Mg), evaporates, flysch and arenaceous material (Ca). The opposite position in the biplot of Ca and Mg compared with Ti may here indicate that some limestone deposits are mixed with terrigenous material.

Higher values for Cr and Ni are found in topsoils from limestones, lavas and alluvial material (Cr), flysch, sandstone and clay (Ni). Their association in the same part of the biplot and their high clr-variance indicate the presence of complex geochemical processes of re-distribution. Their occurrence in soils ranges from highly mobile species to ones that have no reactivity due to organic matter, clay fractions, pH control, presence of Fe–Mn hydroxides, and organic acids (Kabata-Pendias and Sadurski, 2004).

Summarising the results obtained by biplot analysis (Fig. 5a) the transition from Na and K towards Fe, Mn and Co (through Ba, Pb, V, Mo, and Zn) represents the compositional shift of topsoils from the volcanic component towards the sedimentary one (lavas ➔ flysch). In this framework the higher clr-variances move B, P, Sr and Cu downwards, probably due to a perturbation effect on compositions attributable to fertiliser use. If compositions move upwards (Al, Th, La, As and also Ti) or downwards to the right (Mg and Ca), a possible mixing of processes also affecting the calcareous component is present. Finally, if compositions move towards the right a complex scenario could have affected the behaviour of Ni and Cr, since high abundance characterises different source lithologies and variability increases.

As previously reported, the robust barycentre of the whole dataset (in mg/kg: Al = 45,623; As = 13.52; B = 16.04; Ba = 459; Ca = 29,577; Co = 10.38; Cr = 14.27; Cu = 124; Fe = 25,932; K = 16,782; La = 44.74; Mg = 6792; Mn = 812; Mo = 1.52; Na = 4207; Ni = 13.74; P = 1859; Pb = 76.33; Sr = 195; Th = 13.87; Ti = 1355; V = 73.31; Zn = 120) may represent a $(1 \times D)$ compositional baseline mainly given by topsoil developed, but not only, on volcanic materials (Table 2). The barycentre, together with the variation array of dataset A represents the most frequent $(1 \times D)$ multi-element vector and the proportionality relationships among its components. Anomalous compositions can originate from all around this barycentre due to the processes discussed above, able also to mask the contribution of the parental material. By choosing couples of elements of the compositional baseline to be involved in x_i/x_j balances, the 95% confidence interval can be calculated for each of them, giving a measure of the variability of the ratio among the D components of the composition. The balances and their confidence interval can be used to make comparisons with values obtained by considering some natural reservoirs (for example average crust, world average soils composition), or legal references (maximum allowable concentrations or trigger action value, Kabata-Pendias, 2011). They can also be visualised in continuous coloured maps, since geostatistical analysis can be applied without bias (Fig. 6 for the Al/Fe balance).

When compared with the average crust composition (Faure, 1998) the compositional baseline presents an Aitchison distance (d_a) equal to 4.85 while a value of 6.90 characterises the comparison of the average crust with GEMAS data (Reimann et al., 2014a, 2014b). This metric is given by:

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

where $\mathrm{g_m}(\cdot)$ is the geometric mean of the components of the compositions $\mathbf{x}$ and $\mathbf{y}$. It represents another useful index to understand the intensity of the action of the weathering and re-distribution processes from a given starting point as the average crust.

In our case the d_a value can also be useful to verify some bias due to the different sampling densities for urban and suburban — agricultural areas. In fact dataset A is also characterised by the presence of topsoils from lavas, pyroclastic rocks and ignimbrites of urban areas where the sampling density is higher. There is the possibility that the established compositional baseline might represent the urban conditions compared with the suburban and agricultural ones. However, the d_a compositional difference between the identified baseline and the robust barycentre of

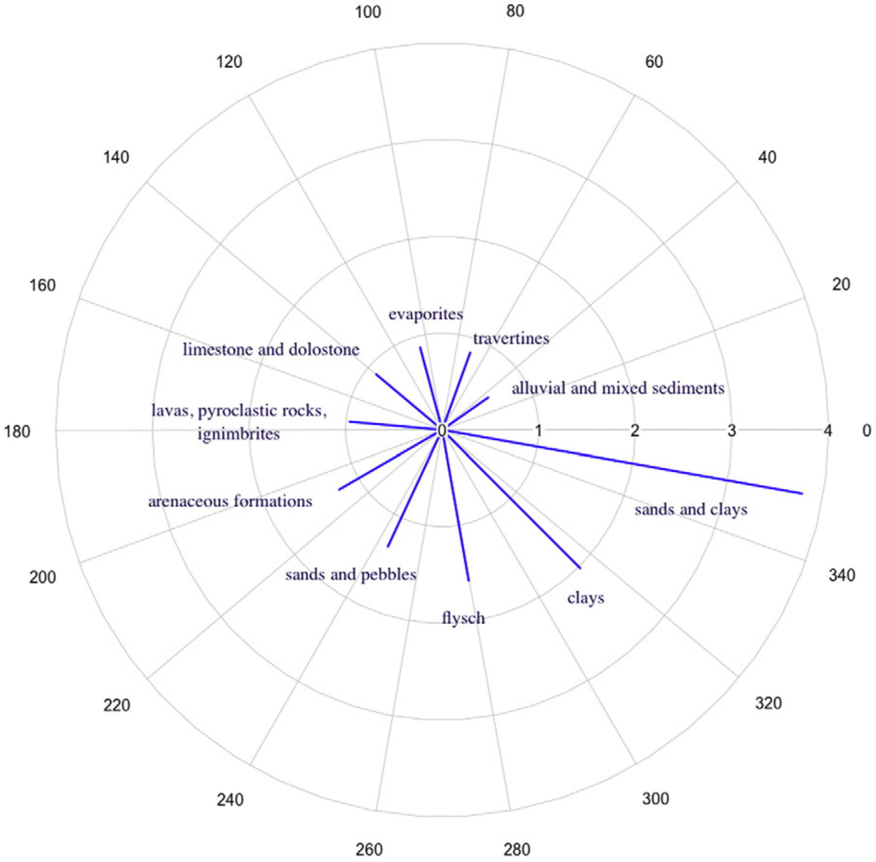

Fig. 7. Representation of the Aitchison distance d_a by a plot of radial lines related to the value of different lithologies for dataset A. The centre represents the $(1 \times D)$ compositional baseline while the d_a value increases anticlockwise.

the volcanic material alone (lavas, pyroclastic rocks, ignimbrites, lithology 7 for group A, Table 2) is equal to 0.96. For the same dataset A the difference from the compositional baseline is equal to 0.58 for alluvials and mixed sediments, a lithology that in group A does not only characterise urban areas. The value increases to 0.85 for travertines, 0.88 for evaporites, 0.89 for limestone and dolostone, 1.23 for arenaceous formations, 1.33 for sands and pebble, 1.58 for flysch, 2.02 for clays and 3.78 for sand and clays (Fig. 7). As observed in Fig. 7, notwithstanding the different sampling strategies, the intercepted $(1 \times D)$ compositional baseline for topsoils tends to represent wider conditions of the sole reference to the parental material and density sampling.

6. Conclusions

The $n \times D$ multi-element archive (n number of samples, D number of elements) of the Campanian Region (southern Italy) containing the abundance (mg/kg) of Al, As, B, Ba, Ca, Co, Cr, Cu, Fe, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, Sr, Th, Ti, V and Zn determined on 3535 new topsoils has given us the opportunity to explore the data structure with CoDA tools and to try to identify a $(1 \times D)$ compositional baseline. The motivations were in the relative nature of compositional data (concentrations) and in the complexity of natural phenomena both requiring the joint investigation of several variables. Moreover, since often the frequency distributions of the chemical elements are asymmetrical, bimodal or present heavy tails or anomalous values, robust methods were applied.

Notwithstanding the area is characterised by high geological heterogeneity, the fingerprint of the parental material is recognisable even if complex processes related to the soil formation and involving different lithologies have generated wide overlaps.

The structure of the whole matrix appears to be constituted by a core that geographically is mainly given by topsoils developed on volcanic materials and several anomalous compositions whose origin is different. However, the robust barycentre of the whole dataset is more compositionally similar to alluvials and mixed sediments with respect to all the other lithologies (lower Aitchison distance d_a).

Anomalous compositions can originate from the robust barycentre all around when the following conditions are present: 1) high Na–K volcanic products, 2) limestones and dolostones with their terrigenous component, 3) flysch deposits or 4) fertiliser contribution.

The $(1 \times D)$ robust barycentre of the whole dataset together with the variation array of the core represents the most frequent $(1 \times D)$ multi-element vector (the compositional baseline) and the proportionality relationships among its components. By using the information contained in the variation array of the core (dataset without outlier compositions), a measure of the variability of the $(1 \times D)$ robust barycentre can in fact be obtained. By choosing the values of couples of elements of the compositional baseline to be involved in x_i/x_j balances, the 95% confidence interval can be calculated for each of them. The ratios and their confidence interval can be used to make comparisons with different reference terms, also including legal ones.

While further investigations will address these topics, our results indicate that the compositional approach could already be applied in the identification of the compositional baseline in a wide array of environmental contexts by applying a multivariate point of view. The possibility to extract values related to single balances, moving from multivariate to bivariate conditions, would also permit a profitable use for public decision makers who need a specific value to be compared with legal references.

Acknowledgements

This work was supported by the grant from the Ministero dell'Università e della Ricerca Scientifica — Industrial Research Project “Integrated agro-industrial chains with high energy efficiency for the development of ecocompatible processes of energy and biochemicals production from renewable sources and for the land valorization (EnerbioChem)” PON01_01966, funded (Prof. B. De Vivo) in the frame of the Operative National Programme Research and Competitiveness 2007–2013 D. D. Prot. n. 01/Ric. 18.1.2010. The University of Florence (I) is also thanked for 2014 funds. The authors also wish to thank the Editor-in-Chief R.A. Ayuso and reviewers for useful comments and suggestions.

References

Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). J. R. Stat. Soc. Ser. B (Stat Methodol.) 44 (2), 139–177.

Aitchison, J., 1986. The statistical analysis of compositional data. Monographs on Statistics and Applied Probability. Chapman and Hall, Ltd., London, UK (416 pp. (reprinted 2003 with additional material by The Balckburn Press)).

Aitchison, J., 1997. The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn, V. (Ed.), Proceedings of IAMG'97 — The Third Annual Conference of the International Association for Mathematical Geology. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), pp. 3–35 (Vol. I and II and addendum).

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V., 2000. Logratio analysis and compositional distance. Math. Geol. 32 (3), 271–275.

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

Allegre, C., Lewin, E., 1996. Scaling laws and geochemical distributions. Earth Planet. Sci. Lett. 132, 1–13.

Alloway, B.J., 2005. Bioavailability of elements in soil. In: Selenius, O. (Ed.), Essential of Medical Geology, Impacts of the Natural Environment on Public Health. Elsevier Academic Press, Amsterdam, pp. 347–372.

Angelone, M., Armiento, G., Cinti, D., Somma, R., Trocciola, A., 2002. Platinum and heavy metal concentration levels in urban soils of Naples (Italy). Fresenius Environ. Bull. 11 (8).

Bini, C., 2011. Environmental Impact of Abandoned Mine Sites: A Review. Nova Science Publishers, New York (92 pp.).

Bonardi, G., Ciarcia, S., Di Nocera, S., Matano, F., Sgrosso, I., Torre, M., 2009. Carta delle principali unità cinematiche dell'appennino meridionale. Nota illustrativa. Ital. J. Geosci. 128 (1), 47–60.

Boni, M., Rollinson, G., Mondillo, N., Balassone, G., Santoro, L., 2013. Quantitative mineralogical characterization of karst bauxite deposits in the Southern Apennines, Italy. Econ. Geol. 108, 813–833.

Borgheresi, M., Buccianti, A., Di Benedetto, F., Vaughan, D.J., 2013. Applications of compositional techniques in the field of crystal chemistry: a case study of luzonite, aSn-bearing mineral. Math. Geosci. 45, 183–206.

Buccianti, A., 2011. Natural laws governing the distribution of the elements in geochemistry: the role of the log-ratio approach. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. John Wiley & Sons, pp. 255–266.

Buccianti, A., 2013. Is compositional data analysis a way to see beyond the illusion? Comput. Geosci. 50, 165–173.

Buccianti, A., 2015a. The FOREGS repository: modelling variability in stream waters on a continental scale revising classical diagrams from CoDA (compositional data analysis) perspective. J. Geochem. Explor. 154, 94–104.

Buccianti, A., 2015b. Frequency distributions of geochemical data, scaling laws and properties of compositions. Pure Appl. Geophys. 172 (7), 1851–1863.

Buccianti, A., Gallo, M., 2013. Weighted principal component analysis for compositional data: application example for the water chemistry of the Arno river (Tuscany, central Italy). Environmetrics 24, 269–277.

Buccianti, A., Grunsky, E., 2014. Compositional data analysis in geochemistry: are we sure to see what really occurs during natural processes? J. Geochem. Explor. 141, 1–5.

Buccianti, A., Magli, R., 2011. Metric concepts and implications in describing compositional changes for world river's chemistry. Comput. Geosci. 37 (5), 670–676.

Cicchella, D., De Vivo, B., Lima, A., 2003. Palladium and platinum concentration in soils from the Napoli metropolitan area, Italy: possible effects of catalytic exhausts. Sci. Total Environ. 293, 47–57.

Cicchella, D., De Vivo, B., Lima, A., Albanese, S., Fedele, L., 2008. Urban geochemical mapping in the Campanian region (Italy). Geochem. Explor. Environ. Anal. 8, 19–29.

COM (Commission of the European Communities), 2006. Thematic strategy for soil protection. (http://eusoils.jrc.ec.europa.eu/ESDB_Archive/Policies/Directive/com_2006_0231_en.pdf).

Costantini, E.A.C., Dazzi, C., 2013. The Soils of Italy. Springer (354 pp.).

Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., Walczak, B., 2007. Robust statistics in data analysis. A review. Basic concepts. Chemom. Intell. Lab. Syst. 85, 203–219.

Dazzi, C., Monteleone, S., 2001. Soil and soil–landform relationships along an elevation transect in a gypsiferous hilly area in central Sicily, Italy. Proceedings of the 7th International Conference on Soils with Mediterranean Type of Climate. Options Méditerranéennes 50, pp. 73–86.

de Caritat, P., Grunsky, E., 2013. Defining element associations and inferring geological processes from total element concentrations in Australia catchment outlet sediments: multivariate analysis of continental-scale geochemical data. Appl. Geochem. 33, 104–126.

De Vivo, B., 2006. Volcanism in the Campania Plain. Vesuvius, Campi Flegrei and IgnimbritesDevelopments in Volcanology vol. 9. Elsevier, pp. VII–XII.

De Vivo, B., Rolandi, G., Gans, P.B., Calvert, A., Bohroson, W.A., Spera, F.J., Belkin, H.E., 2001. New constraints on the pyroclastic eruptive history of the Campanian volcanic Plain (Italy). Mineral. Petrol. 73, 47–65.

De Vivo, B., Petrosino, P., Lima, A., Rolandi, Belkin, H.E., 2010. Research progress in volcanology in Neapolitan area, southern Italy: a review and alternative views. Mineral. Petrol. 99, 1–28.

De Vos, W., Tarvainen, T., Salminen, R., Reeder, S., De Vivo, B., Demetriades, A., Pirc, S., Batista, M.J., Marsina, K., Ottesen, R.T., O'Connor, P.J., Bidovec, M., Lima, A., Siewers, U., Smith, B., Taylor, H., Shaw, R., Salpeteur, I., Gregorauskiene, V., Halamic, J., Slaninka, I., Lax, K., Gravese, P., Birke, M., Breward, N., Ander, E.L., Jordan, G., Duris, M., Klein, P., Locutura, J., Bel-Lan, A., Pasieczna, A., Lis, J., Mazreku, A., Gilucis, A., Heitzmann, P., Klaver, G., Petersell, V., 2006. Geochemical atlas of Europe. Part 2. Interpretation of Geochemical Maps, Additional Tables, Figures, Maps, and Related Publications. Geological Survey of Finland, Espoo (690 pp., ISBN 951–690-956-6).

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.

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

Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. In: Buccianti, A., Mateu-Figuera, G., Pawlowsky-Glahn, V. (Eds.), Compositional Data Analysis in the Geosciences: From Theory to PracticeSpecial Publications 264. Geological Society, London, pp. 12–28.

Everitt, B., Hothorn, T., 2011. An Introduction to Applied Multivariate Analysis with R. Springer-Verlag, Berlin Heidelberg (274 pp.).

FAO (Food and Agriculture Organization of the United Nations), 2015,. Understanding mountain soils: a contribution from mountain areas to the International Year of Soils 2015. (Job Number I4704, 170 pp., http://www.fao.org/publications/card/en/c/6c2daffb-b253-44e6-9a5d-352fc9cf9f67/).

Faure, G., 1998. Principles and Applications of Geochemistry. Prentice Hall, Upper Saddle River, New Jersey 07458, 600.

Filzmoser, P., Reimann, C., Garrett, R.G., 2005. Multivariate outlier detection in exploration geochemistry. Comput. Geosci. 31, 579–587.

Filzmoser, P., Hron, K., 2008. Outlier detection for compositional data using robust methods. Math. Geosci. 40 (3), 233–248.

Filzmoser, P., Hron, K., Reimann, C., 2011. Interpretation of multivariate outliers for compositional data. Comput. Geosci. 39, 77–85.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58 (3), 453–467.

Gałuszka, A., 2007. A review of geochemical background concepts and an example using data from Poland. Environ. Geol. 52 (5), 861–870.

Garrett, R.G., 1989. The chi-square plot: a tool for multivariate outlier recognition. J. Geochem. Explor. 32, 319–341.

Garrett, R.G., Goss, T.I., Poirier, P.R., 1982. Multivariate outlier detection — an application of robust regression in the earth sciences. Am. Stat. Assoc. Ann. Mtg. (abstract 101).

Gurumurthy, G.P., Balakrishna, K., Tripti, M., Audry, S., Riotte, J., Brraun, J.J., Udaya Shankar, H.N., 2014. Geochemical behaviour of dissolved trace elements in a monsoon-dominated tropical river basin, Southwestern India. Environ. Sci. Pollut. Res. Int. 21 (7), 5098–5120.

Hubert, M., Rousseeuw, P.J., Verboven, S., 2002. A fast method for robust principal components with applications to chemometrics. Chemom. Intell. Lab. Syst. 60, 101–111.

Hubert, M., Rousseeuw, P.J., Vanden, B.K., 2005. ROBPCA: a new approach to robust component analysis. Technometrics 47 (1), 64–79.

Ippolito, F., Ortolani, F., Russo, M., 1973. Struttura marginale dell'appennino campano: reinterpretazioni di dati di antiche ricerche di idrocarburi. Mem. Soc. Geol. Ital. 12, 127–250.

Johnson, C.J., Ander, E.L., 2008. Urban geochemical mapping: how and why we do them. Environ. Geochem. Health 30, 511–530.

Kabata-Pendias, A., 2011. Trace Elements in Soils and Plants. Fourth edition. CRC Press, Taylor & Francis Group, Boca Radon, Florida (505 pp.).

Kabata-Pendias, A., Sadurski, W., 2004. Elements and Their Compounds in the Environment. Wiley-VCH, Weinheim.

Kabata-Pendias, A., Szteke, B., 2015. Trace Elements in Abiotic and Biotic Environments. CrC Press, Taylor & Francis Group (440 pp.).

Kaufman, L., Rousseeuw, P.J., 1990. Clustering large data set (with discussion). In: Glesema, E.S., Kanal, L.N. (Eds.), Pattern Recognition in Practice II. North-Holland, Amsterdam, pp. 425–437.

Korkmaz, S., Goksuluk, D., Zararsiz, G., 2015. MVN: an R package for assessing multivariate normality. (https://cran.r-project.org/web/packages/MVN/vignettes/MVN.pdf).

Lavecchia, G., 1987. The Tyrrhenian–Apennines system: structural setting and seismotectogenesis. Tectonophysics 147, 263–296.

Lee, L., Helsel, D., 2005. Baseline models of trace elements in major aquifers of the United States. Appl. Geochem. 20 (8), 1560–1570.

Levitan, D.M., Schreiber, M.E., Seal II, R.R.I.I., Bodnar, R.J., Aylor Jr., J.G., 2014. Developing protocols for geochemical baseline studies: an example from the Coles Hill uranium deposit, Virginia, USA. Appl. Geochem. 43, 88–100.

Lima, A., De Vivo, B., Cicchella, D., Cortini, M., Albanese, S., 2003. Multifractal IDW interpolation and fractal filtering method in environmental studies: an application on regional stream sediments of (Italy), Campania region. Appl. Geochem. 18, 1853–1865.

Lulli, L., 2007. Italian volcanic soils. In: Arnalds, O., Bartoli, F., Buurman, P., Oskarsson, H., Stoops, G., Garcia-Rodeja, E. (Eds.), Soils of Volcanic Regions in Europe. Springer-Verlag, pp. 51–67.

Macias, F., Chesworth, W., 1992. Weathering in humid regions, with emphasis on igneous rocks and their metamorphic equivalents. In: Martini, P., Chesworth, W. (Eds.), Weathering, Soils and Paleosols. Elsevier, Amsterdam, pp. 284–303.

Milia, A., Torrente, M.M., 2000. Fold uplift and syn-kinematic stratal architectures in a region of active transtensional tectonics and volcanism, Eastern Tyrrhenian Sea. Geol. Soc. Am. Bull. 112, 1531–1542.

Nordstrom, D.K., 2015. Baseline and premining geochemical characterization of mined sites. Appl. Geochem. 57, 17–34.

Ott, W.R., 1990. A physical explanation of the lognormality of agent concentrations. J. Air Waste Manage. Assoc. 40, 1378–1383.

Ott, W.R., 1994. Environmental Statistics and Data Analysis. CRC Press, LLC, Lewis Publishers, USA (336 pp.).

Palarea-Albaladejo, J., Martin-Fernandez, J.A., 2015. ZCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemom. Intell. Lab. Syst. 143, 85–96.

Patacca, E., Scandone, P., 2007. Geology of Southern Apennines. CROP-04. In: Mazzotti, A., Patacca, E., Scandone, P. (Eds.), Boll. Soc. Geol. It. 7, pp. 75–119

Pawlowsky-Glahn, V., Buccianti, A., 2011. Compositional data analysis. Theory and Applications. John Wiley & Sons, Ltd., London (400 pp.).

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and analysis of compositional data. Statistics in Practice, Wiley, (247 pp.).

Peccerillo, A., 2005. Plio-Quaternary volcanism in Italy. Petrology, Geochemistry, Geodynamics. Springer-Verlag, Berlin Heidelberg (ISBN 978-3-540-29,092-6, 365 pp.).

Plant, J.A., Klaver, G., Locutura, J., Salminen, R., Vrana, K., Fordyce, F.M., 1996. Forum of European Geological Surveys (FOREGS) Geochemistry Task Group 1994–1996. British Geological Survey (BGS) Technical Report WP/95/14.

R Foundation for Statistical Computing, 2014. version 3.1.0 (2014-04-10), "Spring Dance". (https://cran.r-project.org/bin/macosx/).

R package "cluster", 2015. Version 2.0.3, cran. r-project.org/web/packages/cluster/index.html.

Reimann, C., Garrett, R.G., 2005. Geochemical background — concept and reality. Sci. Total Environ. 350 (1–3), 12–27.

Reimann C., Birke M., Demetriades A., Filzmoser P., O'Connor P. (Editors) and GEMAS Team, 2014a. Chemistry of Europe's agricultural soils — part A: methodology and interpretation of the GEMAS data set. Geologisches Jahrbuch (Reihe B), Schweizerbarth: Hannover, (528 pp.).

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor P. (Editors) and GEMAS Project Team, 2014b. Chemistry of Europe's agricultural soils — part B: general background information and further analysis of the GEMAS data set. Geologisches Jahrbuch (Reihe B), Schweizerbarth: Hannover, (352 pp.)

Rodrigues, A.S.L., Nalini Júnior, H.A., 2009. Geochemical background values and its implications in environmental studies. Rem: Rev. Esc. Minas 62 (2), 155–165.

Rolandi, G., Bellucci, F., Heizler, M.T., Belkin, H.E., De Vivo, B., 2003. Tectonic controls on the genesis of ignimbrites from the Campanian Volcanic Zone, southern Italy. Mineral. Petrol. 79 (1–2), 3–31.

Rosenbaum, G., Lister, G.S., 2004. Neogene and Quaternary rollback evolution of the Tyrrhenian Sea, the Apennines, and the Sicilian Maghrebides. Tectonics 23, 23, TC1013. http://dx.doi.org/10.1029/2003TC001518.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Am. Stat. Assoc. 79, 871–880.

Salminen, R., Gregorauskiene, V., 2000. Considerations regarding the definition of a geochemical baseline of elements in the surficial materials in areas differing in basic geology. Appl. Geochem. 15 (5), 647–653.

Salminen, R., Tarvainen, T., Demetriades, A., Duris, M., Fordyce, F.M., Gregorauskiene, V., Kahelin, H., Kivisilla, J., Klaver, G., Klein, H., Larson, J.O., Lis, J., Locutura, J., Marsina, K., Mjartanova, H., Mouvet, C., O'Connor, P., Odor, L., Ottonello, G., Paukola, T., Plant, J.A., Reimann, C., Schermann, O., Siewers, U., Steenfelt, A., Van der Sluys, J., De Vivo, B., Williams, L., 1998. FOREGS geochemical mapping field manual. Geological Survey of Finland, Espoo, Guide 47 (http://www.gtk.fi/foregs/geochem/ fieldmanan.pdf).

Salminen, R., Batista, M.J., Bidovec, M., Demetriades, A., De Vivo, B., De Vos, W., Gilucis, A., Gregorauskiene, V., Halamic, J., Heitzmann, P., Lima, A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutura, J., Marsina, K., Mazreku, A., Mrnkova, J., O'Connor, P.J., Olsson, S., Ottesen, R.T., Petersell, V., Plant, J.A., Reeder, S., Salpeteu, I., Sandström, H., Siewers, U., Steenfelt, A., Tarvaine, T., 2005. FOREGS geochemical atlas of Europe. Part 1. Background Information, Methodology, and Maps. Geological Survey of Finland, Espoo (525 pp. ISBN 951-690-913-2, http//gtk/publ/foregsatlas).

Sinclair, A.J., 1974. Selection of threshold values in geochemical data using probability graphs. J. Geochem. Explor. 3, 129–149.

Sinclair, A.J., 1991. A fundamental approach to threshold estimation in exploration geochemistry: probability plots revisited. J. Geochem. Explor. 41, 1–22.

Smith, R.E., Campbell, N.A., Perdrix, J.L., 1983. Identification of some Western Australian Cu–Zn and Pb–Zn gossans by multi-element geochemistry. In: Smith, R.E. (Ed.), Geochemical Exploration in Deeply Weathered Terrain. CSIRO Inst. of Energy & Earth Resources, Wembley, Western Australia, pp. 109–126.

Smith, R.E., Campbell, N.A., Litchfield, R., 1984. Multivariate statistical techniques applied to pisolitic laterite geochemistry at Golden Grove, Western Australia. J. Geochem. Explor. 22 (1–3), 193–216.

Stanley, C.R., Sinclair, A.J., 1989. Comparison of probability plots and the gap statistic in the selection of thresholds for exploration geochemistry data. J. Geochem. Explor. 32, 355–357.

Steckler, M.S., Agostinetti, N.P., Wilson, C.K., Roselli, P., Seeber, L., Amato, A., Lemer-Lam, A., 2008. Crustal structure in the Southern Apennines from teleseismic receiver functions. Geology 2, 155–158.

Struyf, A., Hubert, M., Rousseeuw, P.J., 1996. Clustering in an object-oriented environment. J. Stat. Softw. 1 (4) (http://www.jstatsoft.org/v01/i04).

Sucharovà, J., Suchara, I., Hola, M., Marikova, S., Reimann, C., Boyd, R., Filzmoser, P., Englmaier, P., 2012. Top-/bottom-soil ratios and enrichment factors: what do they really show? Appl. Geochem. 27 (1), 138–145.

Tavakkoli, E., Rengasamy, P., McDonald, K., 2010. High concentrations of Na^+ and Cl^- ions in soil solution have simultaneous detrimental effects on growth of faba bean under salinity stress. J. Exp. Bot. 61 (15), 4449–4459.

Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: an R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data AnalysisTheory and Applications. Wiley & Sons Ltd., UK, pp. 341–355.

Tidball, R.R., Ebens, R.J., 1997. Regional geochemical baselines in soils of the Power River Basin, Montana-Wyoming. In: Laudon, R.B. (Ed.), Geology and Energy Resources of the Powder RiverWyoming Geological Association, 28th Annual Field Conference. Wyoming Geological Association, Casper Wyo., pp. 299–310.

Torrente, M.M., Milia, A., 2013. Volcanism and faulting of the Campania margin (Eastern Tyrrhenian Sea, Italy): a three-dimensional visualization of a new volcanic field off Campi Flegrei. Bull. Volcanol. 75–719.

USEPA (U.S. Environmental Protection Agency), 2002. Guidance for comparing background and chemical concentrations in soil for CERCLA sites. Office of Emergency and Remedial Response U.S. Environmental Protection Agency Washington, EPA 540-R-01-003, DC 20460.

Van den Boogaart, K.G., Tolosana-Delgado, R., 2013. Analyzing Compositional Data with R. Springer-Verlag, Berlin Heidelberg.

Verboven, S., Hubert, M., 2005. LIBRA: a Matlab library for robust analysis. Chemom. Intell. Lab. Syst. 75, 127–136.

Vitrone, G., 2003. Valutazione del grado di inquinamento da metalli pesanti nei suoli del commune di Caserta. Università di Napoli Federico II, Tesi di Laurea.

Washington, H.S., 1906. The Roman Comagmatic Region. Carnegie Institution of Washington (Publication 57, 199 pp.).

Wehrens, R., 2011. Chemometrics with R. Multivariate Analysis in the Natural Sciences and Life Sciences. Springer-Verlag, Berlin Heidelberg (285 pp.).