Compositional baseline assessments to address soil pollution: An application in Langreo, Spain

C. Boente[1,2*], M.T.D. Albuquerque[3], J.R. Gallego[4], V. Pawlowsky-Glahn[5], J.J. Egozcue[6]

[1] *Department of Mining, Mechanic, Energetic and Construction Engineering, ETSI, University of Huelva, 21071 Huelva, Spain, carlos.boente@dimme.uhu.es*

[2] *CIQSO-Center for Research in Sustainable Chemistry, Associate Unit CSIC-University of Huelva "Atmospheric Pollution", Campus El Carmen s/n, 21071 Huelva, Spain*

[3] *CERNAS | QRural, Instituto Politécnico de Castelo Branco and ICT, Universidade de Évora, Portugal, teresal@ipcb.pt*

[4] *Environmental Biogeochemistry & Raw Materials Group and INDUROT. Campus de Mieres, University of Oviedo, C/Gonzalo Gutiérrez Quirós. S/N, 33600 Mieres, Spain, jgallego@uniovi.es*

[5] *Dpt. Computer Science, Applied Mathematics and Statistics, University of Girona, Spain, vera.pawlowsky@udg.edu*

[6] *Dpt. Civil and Environmental Engineering, Technical University of Catalonia, Barcelona, Spain, juan.jose.egozcue@upc.edu*

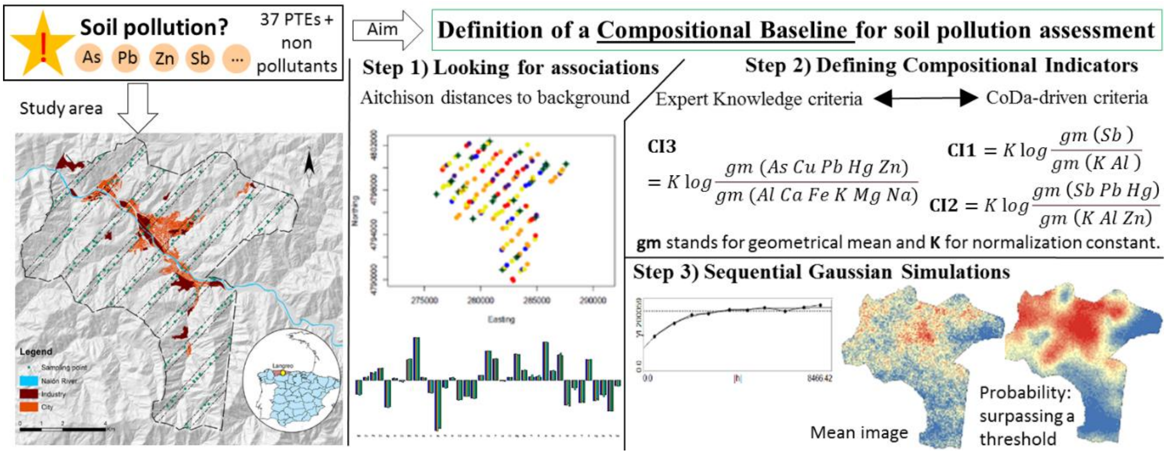
(*) Corresponding author: *carlos.boente@dimme.uhu.es*

Abstract

Potentially Toxic Elements (PTEs) are contaminants with high toxicity and complex geochemical behaviour and, therefore, high PTEs contents in soil may affect ecosystems and/or human health. However, before addressing the measurement of soil pollution, it is necessary to understand what is meant by pollution-free soil. Often, this background, or pollution baseline, is undefined or only partially known. Since the concentration of chemical elements is compositional, as the attributes vary together, here we present a novel approach to build compositional indicators based on Compositional Data (CoDa) principles. The steps of this new methodology are: 1) Exploratory data analysis through variation matrix, biplots or CoDa dendrograms; 2) Selection of geological background in terms of a trimmed subsample that can be assumed as non-pollutant; 3) Computing the spread Aitchison distance from each sample point to the trimmed sample; 4) Performing a compositional balance able to predict the Aitchison distance computed in step 3.Identifying a compositional balance, including pollutant and non-pollutant elements, with sparsity and simplicity as properties, is crucial for the construction of a Compositional Pollution Indicator (CI). Here we explored a database of 150 soil samples and 37 chemical elements from the contaminated region of Langreo, Northwestern Spain. There were obtained three Cis: the first two using elements obtained through CoDa analysis, and the third one selecting a list of pollutants and non-pollutants based on expert knowledge and previous studies. The three indicators went through a Stochastic Sequential Gaussian simulation. The results of the 100 computed simulations are summarized through mean image maps and probability maps of exceeding a given threshold, thus allowing characterization of the spatial distribution and variability of the CIs. A better understanding of the trends of relative enrichment and PTEs fate is discussed.

Keywords: Potentially Toxic Elements; Soil Pollution; Compositional Indicators; Sequential Gaussian Simulation

Graphical abstract


Soil pollution?
As
Pb
Zn
Sb
...
37 PTEs + non pollutants
Study area
Legend
Sampling point
Nalon River
Industry
City
Aim
Definition of a Compositional Baseline for soil pollution assessment
Step 1) Looking for associations
Aitchison distances to background
Step 2) Defining Compositional Indicators
Expert Knowledge criteria
CoDa-driven criteria
CI3
$= K \log \frac{gm\ (As\ Cu\ Pb\ Hg\ Zn)}{gm\ (Al\ Ca\ Fe\ K\ Mg\ Na)}$
$CI1 = K \log \frac{gm\ (Sb)}{gm\ (K\ Al)}$
$CI2 = K \log \frac{gm\ (Sb\ Pb\ Hg)}{gm\ (K\ Al\ Zn)}$
gm stands for geometrical mean and K for normalization constant.
Step 3) Sequential Gaussian Simulations
Mean image
Probability: surpassing a threshold

Highlights

- A novel method to define a baseline for non-polluted soils is proposed.
- A method to build compositional indicators to address soil pollution is proposed.
- Indicators obtained through compositional balances complement expert's criteria.
- Sequential Gaussian Simulations offer a proper visualization of the indicators.

1. Introduction

The continuous accumulation of Potentially Toxic Elements (PTEs) in distinct environmental matrices over time has compromised the health of living organisms and ecosystem quality, to the point that these substances now pose a major environmental concern worldwide (Clemens, 2006). In the case of soils, the persistence and non-biodegradability of PTEs (Kabata-Pendias, 2010), have led to a continuous increase in their concentration in soils, and, consequently, an increased risk to human and environmental health (Khanam et al., 2020; Cachada et al., 2018). The accumulation of PTEs can be explained by population growth, accompanied by the development of industrial activity and housing, which bring with them innumerable sources of pollution (Kelepertzis et al., 2020; Sánchez de la Campa et al., 2018; Juma et al., 2014; Madrid et al., 2006). In this context, in recent years, researchers have channeled considerable efforts into developing methodologies and tools able to offer an accurate characterization of the spatial distribution of PTEs in soil, as well as to identify geochemical backgrounds or baselines and their possible enrichment sources (Wang et al., 2021; McIlwaine et al., 2014; Reimann et al., 2005).

Maps are a powerful way to visually represent the spatial distribution of pollutants and they are a useful tool to support policy-making and vulnerabilities with regard to environmentally complex scenarios (Lahr and Kooistra, 2010; McKinley et al., 2016). In soil science, a common strategy to represent the distribution of PTEs consists on mapping a series of single-component contamination indices or indicators. However, they do not consider the compositional nature inherent to geochemical data (Filzmoser et al., 2009), which require to study the geochemical information by means of ratios of proportions between the chemical elements (Barceló-Vidal and Martín-Fernández, 2016; Pawlowsky-Glahn et al., 2015). In other words, these indices/indicators focus on the study of single elements, without considering that the concentration of an individual PTE depends on the concentrations of the remaining elements, as all of them belong the same whole. The use of these non-compositional indices is usual in geochemical studies, some of the most common are the Geoaccumulation Index (Muller, 1969), the Enrichment Factor (Sucharova et al., 2012), or the Single Pollution Index (SPI) (Hakanson, 1980), and others recently reviewed in Kowalska et al. (2018).

In the field of geosciences, and particularly in geochemistry, it is well known that traditional statistical methods directly applied to raw data can fail (Chayes, 1962, 1971). A solution to those problems was found by Aitchison (1982, 1986) by introducing the log-ratio approach. Since then, Compositional Data (CoDa) theories have seen a development towards a better understanding of the sample space of compositional data and their structure (Pawlowsky-Glahn and Egozcue, 2001). Representations of data in terms of pwlr (pairwise log ratios), ilr (isometric log-ratio coordinates), clr (centered log-ratio coordinates) and alr (additive log-ratio coordinates) can tackle the compositional nature of element concentration data (Pawlowsky-Glahn and Egozcue, 2001; Egozcue et al., 2003; Buccianti and Grunsky, 2014; Kynclova et al., 2017), albeit with different properties that need to be taken into account. The use of CoDa methodologies has advanced research in multiple fields of environmental science, including ecotoxicology (Mullineaux et al., 2021), city pollution (Cicchella et al., 2020), water quality control (Wei et al., 2018), dynamics (Graziano et al., 2020), and health risk assessment (Tepanosyan et al., 2020), among many others (Pawlowsky-Glahn and Buccianti, 2011; Filzmoser et al., 2021).

Moreover, CoDa techniques have shown to be a powerful tool to establish pollution indices with respect to other environmental matrices, like water (Batsaikhan et al., 2021) or air contamination (Sowden et al., 2020; Jarauta-Bragulat et al., 2016). In the case of soils, the application of the CoDa approach to tackle the pollution issue has only recently started to be explored (Boente et al., 2020b, c; Zuzolo et al., 2020). There are also few studies, specifically focusing on compositional indices or indicators, to address soil pollution by PTEs. They can be found in the literature (Petrik et al., 2018). Certainly, it is relatively simple to define geochemical backgrounds or baselines and to track the pollution when the source is clear, as it happens in areas presenting extreme concentrations of PTEs over a matrix of unaffected soil (Boente et al., 2022); Hadjipanagiotou et al., 2020). However, in largely industrialized areas, where there are a mixture of point-source and diffuse pollution sources, it is difficult to discriminate sources and other approaches to define geochemical baselines are required (Yotova et al., 2018; Peh et al., 2010). In this context, the great advantage of compositional indices that involve geochemical backgrounds, like the SPI, is that they are scale-invariant and subcompositionally coherent, implying that a change in units of the concentrations will not modify the result of the analysis (Pawlowsky-Glahn et al., 2015; Buccianti and Pawlowsky-Glahn, 2005).

The aim of the present work is to develop a promising methodology to build compositional soil pollution indicators based on estimated soil background. Out methodology is exemplified using the composition of 37 elements, including pollutants (PTEs) and non-pollutants, for 150 topsoil samples collected in the region of Langreo (Northwestern Spain). Three main indicators (balances) for specific sub-compositions of PTEs were built and validated in terms of geochemical backgrounds. Two are data-driven balances and exclusively based on CoDa multivariate statistical analysis, thus deserving the name CoDa-driven methods. The third is a balance of elements chosen through criteria proposed by an expert geochemist (expert criteria), albeit respecting the same CoDa principles. These three balances were computed as indicators to determine whether compositional computation can provide or complement criteria proposed by expert criteria when identifying global pollution, in such a way that any inexperienced person would be able to perform a preliminary assessment of soil pollution using the methodology presented here.

2. Materials and methods

2.1 Characteristics of the data set and the study area

The data set used in this study is located in the region of Langreo, Spain. It is composed of the chemical composition of 150 samples from the top 25 cm of the soil, a very usual depth for environmental geochemistry studies as “shallow” and/or recent soils and sediments as it is a depth range that contains most of the fingerprint of common point-source and diffuse pollution effects. The distribution of the 150 samples is shown in Figure 1. All samples were categorized attending to their land use as follows: (1) Forest (54 points); (2) Farming or Agricultural plots (83 points); (3) Residential (plus recreation, 12 points); and (4) Industrial (1 point). Class (4), industrial use, containing only one point, is worthless for statistical analysis, but it is a reference point where one expects some industrial pollution. Sampling points were also classified by height above sea level into three classes: (1) valley, (2) hillside, and (3) mountain. Figure S1 in the Supplementary materials A.2 shows these classifications.

According to Baragaño et al., 2020, the parent material of the area corresponds mainly to Carboniferous and Cretaceous (conglomerates and sandstones) covered by alluvial deposits along the Nalón River, which crosses the area. Geomorphology of the area corresponds to wide valleys crossed by the mentioned Nalón River which is perpendicularly crossed by other narrow. Climatic conditions are typical interior oceanic, corresponding to abundant precipitations along the year and mild temperatures the whole year.

With respect to chemistry, the dataset includes PTEs of variable toxicity (Fabian et al., 2014). A set of 37 elements was reported in the 150 sampling points, thus giving a 37-part composition, which is assumed to represent the soil. The chemical elements considered (in parenthesis, abbreviation and detection limits in ppm) are silver (Ag; 0.002), aluminium (Al; 100), arsenic (As; 0.1), gold (Au; 0.002), boron (B; 20), barium (Ba; 0.5), bismuth (Bi; 0.02), calcium (Ca; 100), cadmium (Cd; 0.01), cobalt (Co; 0.1), copper (Cu; 0.01), chromium (Cr; 0.5), iron (Fe; 100), gallium (Ga; 0.1), mercury (Hg; 0.005), potassium (K; 100), lanthanum (La; 0.5), magnesium (Mg; 100), manganese (Mn; 1), molybdenum (Mo; 0.01), sodium (Na; 10), nickel (Ni; 0.1), phosphorus (P; 10), lead (Pb; 0.01), sulphur (S; 200), antimony (Sb; 0.02), scandium (Sc; 0.1), selenium (Se; 0.1), strontium (Sr; 0.5), tellurium (Te; 0.02), thorium (Th; 0.1), titanium (Ti; 10), thallium (Tl; 0.02), uranium (U; 0.1), vanadium (V; 2), wolfram (W; 0.1), and zinc (Zn; 0.1).

This set of elements encompasses the main pollutants identified in previous studies (Boente et al., 2020b; 2018), together with trace and major elements useful to identify both pollution sources and geogenic backgrounds. In general, the dataset contains information on soils categorized as forests (36% of total samples), farming or agricultural plots (55%), industrial (1%) and urban/recreational (8%) that were affected by a wide variety of industrial activities, such as coal mining, metalworking, and chemical factories, with special mention to those devoted to the production of fertilizers and pharmaceutical products (Martínez et al., 2014). These industries together with energy production (thermal power plants) have been operating for more than a century in the area of Langreo, which is one of the most paradigmatic examples industrialization processes all along Spain (Prada-Trigo, 2014; Gallego et al., 2016), showing also a remarkable pollution imprint in the environmental compartments comparable with similar industrial areas in Europe (Megido et al., 2017). Following these considerations the area was recently selected for a wide soil pollution study whose results dataset is used herein; in this sense a scrupulous description of the sampling campaign design, local geology, and a comprehensive pollution assessment is detailed in previous studies (Boente et al., 2020b, 2018).

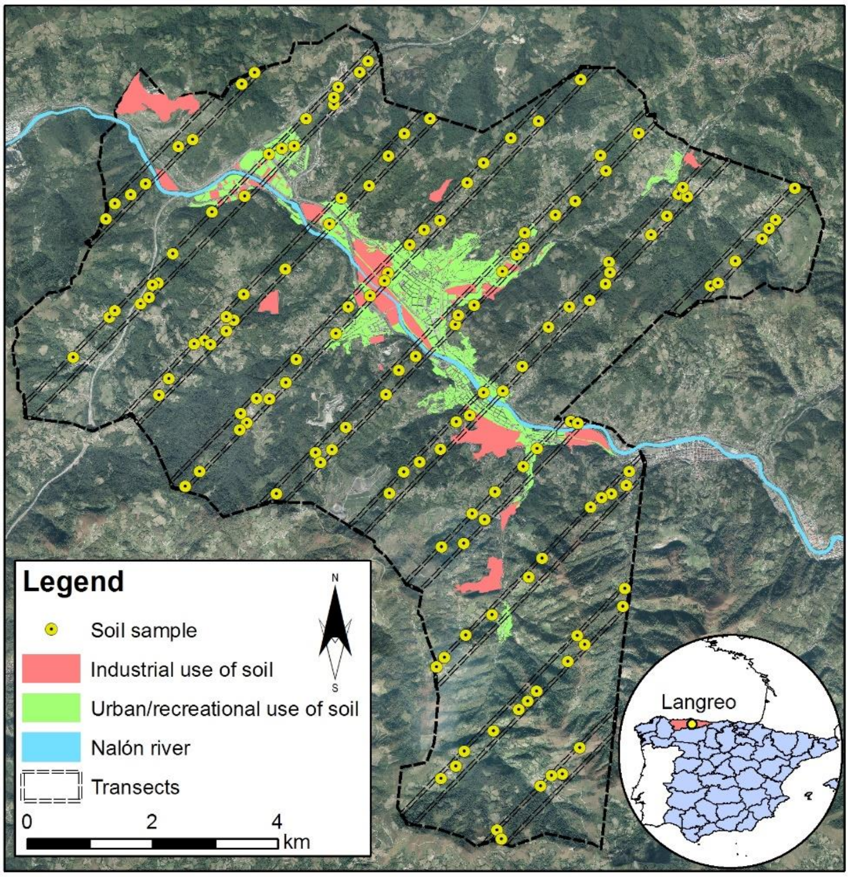


Figure 1. Location of the 150 samples of the dataset in Langreo (Asturias, Spain). Colour code indicates the land use (see legend).

2.2 Nature and requirements of the compositional soil pollution indicator

The definition of a compositional baseline for soil pollution assessment and deviations of the same requires the consideration of a set of key points:

- **Compositional character** (Aitchison, 1986; Eynatten, 2004; Parent et al., 2013; Mueller and Grunsky, 2016): Soil sample analysis usually reports on concentrations of chemical elements and/or other chemicals present. These analyses should be considered compositional, i.e., as a single composition. The indicators should be coherent with this preliminary assumption.
- **Definition of pollution**: Pollution is here defined as an anomaly (compositional difference) of the composition of one sample compared to what is considered a non-polluted, natural soil, called background. The background should include elements that experts consider pollutants, as well as non-pollutant components.
- **Spatial changes in background**: Although it is possible to define a universal background, it is a very rough estimate (Reimann et al., 2005). It is preferable to consider a spatially variable background, thus allowing removal of the efects of geological variations or other natural effects. This means that, in an analysis of pollution, natural sources of variability should be removed, and human-introduced changes should be retained. Thus, pollution is intended to account for geochemical anomalies caused by humans.
- **The indicator as a log-contrast:** As stated in Tolosana-Delgado et al. (2005), an indicator is a function of the sample composition. The main principle in compositional analysis is that summary functions should be scale-invariant, thus acknowledging the compositional character of the data. Scale-invariant linear functions on compositions are called log-contrasts. They are linear combinations of the logarithms of the parts, such that the sum of their coefficients is zero, thus assuring scale invariance. However, log-contrasts involving many elements can be difficult to interpret and might not be useful if some of the elements involved are not reported in the sample. Sparsity and simplicity are therefore desirable

properties of any indicator. Compositional balances are a general form of indicators, as they are log-ratios of the geometric means of parts. They attain simplicity and, if a small number of parts are involved, are also sparse.

- **One indicator for each sort of pollution**: There are different types of pollution and distinguishing them may be important. For instance, pollution can derive from agriculture, water from cities, industry, etc. When compositional samples are represented in coordinates, these distinct types of pollution are identified with directions in the sample space. Each of these directions can define a specific indicator associated with the type of contamination (Tolosana-Delgado et al., 2005). The study of these different types of contamination requires the availability of samples covering all these types of pollution and qualitative classification of the types, thus allowing discriminant analyses.

2.3 Compositional data

The early fundamentals on compositional data can be found in the seminal work by Aitchison (1986). These early contributions are explained and extended in works of general purpose like Pawlowsky-Glahn et al. (2015); Boogaart and Tolosana-Delgado (2013); Filzmoser and Hron (2011); Pawlowsky-Glahn and Buccianti (2011); Egozcue and Pawlowsky-Glahn (2019a). Only specific references on CoDa are cited below.

The analysis of a soil sample, given by its chemical composition, in units like mg/kg, should be conducted under the assumption that these data are compositional. Indeed, the conversion of units from mg/kg to g/kg, for instance, or the expression of units in proportions adding to 1, that is to say by multiplying all elements by 1.000 , or dividing them by the sum of all observed elements respectively, must not change the information in the sample. This is summarized in one of the principles of CoDa analysis, named Scale Invariance Principle. As a result, when performing data analysis, the functions used to describe the composition should be invariant under multiplication by a positive constant. Also, any composition can be expressed in proportions (components adding to 1) without adding or losing any information and irrespective of the units in which the data were initially reported.

A second assumption is known as Subcompositional Coherence Principle. When a soil composition is observed, the elements reported depend on the analytical procedure used and its accuracy. The whole periodic table is never reported, only a subset of elements is measured, and this subset can change in time and campaign. The elements observed form a composition and any subset of the same is a subcomposition, subject again to the Scale Invariance Principle. Analyses performed on the initial composition or a subcomposition should lead to consistent conclusions describing the role of common elements. Historically, the most frequent violation of these principles is the spurious correlation phenomenon: correlation between the concentrations of two elements normalized to proportions in a composition and a subcomposition can give distinct correlation values, sometimes dramatic, including change of signs. These principles were initially formulated in Aitchison (1986) and then rephrased and explained elsewhere (e.g. Barceló-Vidal and Martín-Fernández, 2016; Egozcue and Pawlowsky-Glahn, 2018).

There are cases in which some elements are given as a percentage of major oxides and trace elements in mg/kg or atomic weight. Then, it is recommended to express the concentrations in homogenous units, for instance, changing all units to mg/kg. The conversion of units consists of multiplying each element in the initial composition by a positive coefficient, which may be different for each element. This operation is called perturbation (Aitchison, 1986) and it plays the role of an addition between compositions (the coefficients for the change of units are again a composition). The simplex, complemented with an operation with real scalars, called powering, and an inner product, becomes a Euclidean vector space (Pawlowsky-Glahn and Egozcue, 2001; Billheimer et al., 2001) (see also previous references in this section). This geometry for CoDa is known as Aitchison geometry.

An important consequence of the Aitchison geometry is that compositions can be represented in Cartesian orthogonal coordinates, usually known as isometric log-ratio or orthonormal log-ratio coordinates (ilr, olr) (Egozcue et al., 2003; Martín-Fernández, 2019), which can be treated as usual in an Euclidean space (Mateu-Figueras et al., 2011). A practical way of representing compositions by their ilr coordinates is choosing a basis of the simplex by means of a contrast matrix *V*. Assume that compositions have *D* components, called parts, then *V* is a (*D;D-1*)-matrix such that

$$\mathrm{V^TV = I_{D-1}} \text{ and } \mathrm{VV^T = I_D} - \left(\frac{1}{\mathrm{D}}\right) 11^{\mathrm{T}}, \qquad (1)$$

where $(\cdot)^T$ denotes matrix transposition, I_D is the unit matrix of *D* components and 1 is a *D-vector* with all its components equal to one. An intermediate to define ilr-coordinates is to obtain the so called *centered logratio transformation*, clr, of the composition $x = (x_1, x_2, \dots, x_D)^T$ defined as

$$\mathrm{clr(x)} = \left(\ln\frac{\mathrm{x_1}}{\mathrm{g_m(x)}}, \ln\frac{\mathrm{x_2}}{\mathrm{g_m(x)}}, \dots, \ln\frac{\mathrm{x_D}}{\mathrm{g_m(x)}}\right)^{\mathrm{T}}, \mathrm{g_m(x)} = \prod_{\mathrm{i=1}}^{\mathrm{D}} \mathrm{x_i^{1/D}}. \qquad (2),$$

Then, the ilr-coordinates with respect the basis defined by the contrast matrix V are

$$\mathbf{z} = \mathrm{ilr(x) = V^T clr(x), C_x = ilr^{-1}(z) = Cexp(Vz)}, \qquad (3)$$

where the second equality is the recovery of a closed composition from its ilr-coordinates.

The Aitchison distance between compositions x and y can be computed in different ways, particularly using ilr-coordinates, or the respective clr's:

$$\mathrm{d_a(x,y)} = \left(\sum_{\mathrm{i=1}}^{\mathrm{D}} (\mathrm{clr_i(x)} - \mathrm{clr_i(y)})^2\right)^{1/2} = \left(\sum_{\mathrm{i=1}}^{\mathrm{D-1}} (\mathrm{ilr_i(x)} - \mathrm{ilr_i(y)})^2\right)^{1/2}. \quad (4)$$

In the exploratory analysis of soil samples, assumed compositional, elementary statistics change accordingly to the Aitchison geometry of the simplex. The center or compositional mean is estimated as a compositional average, which is the geometric mean along the parts of the sample, possibly closed to a constant. The total variance of the sample can be computed in at least three ways: using the variances of the pairwise log ratios, the variances of the clr coefficients, or the variances of the ilr-coordinates. Let $\mathbf{X} = [x_{ij}], i = 1,2, \dots, n, j = 1,2, \dots, D,$ be the compositional data matrix; the columns of $\mathbf{X}$, called parts in the sample, are denoted $\mathrm{X_j}$. Then, the total variance of $\mathbf{X}$ is

$$\begin{aligned} \mathrm{totVar}[\mathbf{X}] &= \frac{1}{2D}\sum_{j=1}^{D}\sum_{k=1}^{D} Var\left[\ln\left(\frac{X_j}{X_k}\right)\right] \\ &= \sum_{j=1}^{D} Var[clr_j(X)] \qquad (5) \\ &= \sum_{k=1}^{D-1} Var[ilr_k(X)], \end{aligned}$$

where $\mathrm{ilr}(\mathbf{X})$, $\mathrm{clr}(\mathbf{X})$, are matrices obtained after applying ilr, respectively clr, to the rows of $\mathbf{X}$. The $\mathrm{Var}[\mathrm{ilr_k}(\mathbf{X})]$ ($\mathrm{Var}[\mathrm{clr_k}(\mathbf{X})]$ is the variance across the sample of the *k-th* ilr-coordinate (the *k-th* clr coefficient). The (D, D)- matrix with entries $Var\left[\ln(\frac{X_i}{X_j})\right]$ is called the variation matrix and each entry compares two parts of the compositional sample. Interestingly, small values in the variation matrix indicate that the parts are near to proportionality. This is called linear association for compositional parts (Lovell et al., 2015; Egozcue et al., 2018), and it suggests that information in these parts is almost equivalent. To make variation matrices comparable, the following normalization is used

$$\mathrm{T_{jk}} = \frac{(\mathrm{D}-1)\mathrm{Var}\left[\ln(\frac{\mathrm{X_j}}{\mathrm{X_k}})\right]}{2\ \mathrm{totVar}[\mathbf{X}]}. \quad (6)$$

The idea is to compare the entry of the variation matrix with an ideal variation matrix with identical non-null entries. Then $T_{jk} \geq 1$ indicates that parts X_j and X_k are not linearly associated. Values $T_{jk} < 1$ do not exclude association, and a rule-of-thumb is that only $T_{jk} < 0.2$ suggests effective linear association (see Table S1 in supplementary material).

The CoDa-biplot is a simultaneous representation of the observations and the clr-transformed components (Aitchison, 1983; Aitchison and Greenacre, 2002). It is obtained from the singular value decomposition (svd) of the clr transformation of the centered sample, that is, a principal component analysis of clr(X) after centering, also known as CoDa-PCA. The loading matrix is a contrast matrix and the principal components are ilr coordinates. Compared to the principal component analysis applied to raw data and its biplots, the interpretation of the CoDa-biplot differs in the sense that attention is paid to the links between the rays corresponding to the clr variables. Some examples are given in Section 3.1.

The CoDa-PCA is not the only way to obtain an orthogonal basis and its ilr coordinates. A sequential binary partition (SBP) of the composition (Egozcue and Pawlowsky-Glahn, 2005, 2006) also provides an orthogonal basis. The corresponding ilr coordinates are a special type of log ratio called balances. For composition x, a balance is of the form

$$B\left(\frac{G}{H}\right) = \sqrt{\frac{N_G N_H}{N_G + N_H}} \ln \frac{g_m(G)}{g_m(H)}, \quad (7)$$

where G and H are two non-overlapping groups of parts included in x, and N_g, N_h are the number of parts included in G and H, respectively. Recall that $g_m(\cdot)$ stands for the geometric mean as defined in Equation (2). The square root in front of the balance is a normalizing constant. In this way, the norm of the element of the basis is unitary, thus accounting for the number of elements in each group. Balances are important because they are simple, as parts in each group are treated in a homogeneous way, and, when the groups G and H include a small number of elements, they are also sparse. Principal balances (Martín-Fernández et al., 2018) are techniques that attempt to approximate CoDa-PCA by balances which constitute an ilr basis. The result is an SBP that can be represented by a tree structure in a dendrogram. In addition to the structure of the SBP, the CoDa dendrogram shows the decomposition (vertical bars) of the total variance in variances of ilr coordinates (Eq. 5), and the mean values of the ilr balances, which are represented by the fulcrum of each vertical bar. If there are two or more classes of samples, vertical bars corresponding to each class compare the mean and variance of each balance with the mean and variance of the whole sample. This approach allows an intuitive comparison of classes of samples. All balances performed and their predictions were evaluated through linear regression. Statistical applications and CoDa analysis were performed using R software (R Development Core Team, 2009) and R-package compositions (Boogaart et al., 2009).

2.4 Spatial modelling – geostatistical approach

The three indicators (CI_1, CI_2 and CI_3), as regionalized variables, were computed following a two-step geostatistical modelling methodology:

1. The three indicators went through structural analysis and experimental variograms were then computed. The variogram is a directional function used to compute the spatial variation structure of regionalized variables (Matheron, 1971; Journel and Huijbregts, 1978; Pawlowsky-Glahn and Serra, 2019).
2. Sequential Gaussian Simulation (SGS) was used as a stochastic simulation algorithm over a 100x100 m grid mesh. SGS starts by computing the univariate

experimental distribution of values and performing a normal score transformation of the original values to a standard normal distribution. Normal scores at grid node locations are then simulated sequentially using normal score data through simple kriging (SK) with zero mean, assessed by a leaving out cross-validation, as specified in Goovaerts (1997). Once all normal scores have been simulated they were back-transformed to their original units. For the computation, the Space-Stat Software V. 4.0.18, Biomedwere, was used (Albuquerque et al., 2014).

The outcome of a simulation is always a random version of the estimation process, reproducing the statistics of the known data and building a realistic picture of reality. The associated spatial uncertainty is visualized through the construction of probability maps and validated overlapping the geochemical results obtained in each collected point sample. If multiple sequences of simulation are computed, it is possible to obtain reliable probabilistic maps. The mean image (MI), together with the representation of the probability of exceeding a previously defined threshold, allows broad discussion of the spatial patterns of indicators and the identification of hazard clustering. The Jenks natural break classification (Jenks, 1967) was used to create ten distinct classes to determine the best arrangement of values, seeking a reduction in the variance within classes and maximization of the variance between classes.

3. Results and discussion

3.1 Variation matrix: Looking for associations

The definition of a compositional baseline for soil pollution assessment and deviations of the same requires the consideration of a set of key points: Table S1 shows the normalized variation matrix (Egozcue and Pawlowsky-Glahn, 2019b; Egozcue et al., 2018; Pawlowsky-Glahn et al., 2015) for the chemical parts. Variations larger than 1.0 indicate a lack of linear association between the elements. Only values smaller than 0.2 (marked in blue) suggest a linear association or proportionality. Clear proportionality normally corresponds to values less than 0.1. Examination of this table reveals that the minimum value is 0.09 for the association between Fe and Cr. This implies that linear associations between chemical elements are, in general, weak in this data set. The larger variability comes from the relation of Ca relative to most elements. The sum of the elements of the variation matrix over *2D*, *D = 37* being the number of chemical elements, is the total variance of the data set, which is 9.77. The lack of strong associations between elements indicates that it is difficult to identify distinct types of pollution.

3.2 Exploratory analysis

The sampling points shown in Figure 1 were classified according to described in section 2.1. Their spatial distribution does not show any interesting feature, thus suggesting predominant air transport of contaminants rather than direct deposition. After a CoDa-PCA, Figure 2 shows the covariance and form biplots of the chemical data set. The larger relative variability of the clr component of Ca is visible in the length of the ray corresponding to the clr-Ca component, labeled Ca for readability in Figure 2. In fact, all links from Ca to those of other elements are large in the covariance biplot. The first and second principal components (ilr coordinates) are log-contrasts whose loadings are shown in Table 1. For the first principal coordinate, Ca participates with the largest loading, but many other elements are positively and negatively involved, thereby hindering the interpretation. A more complex situation appears with the second principal coordinate. The larger loadings correspond to Th (positive) and Sb (negative), but many other elements participate with comparable loadings (see Table 1). For the first principal coordinate, Ca participates with the largest loading, but many other elements are positively and negatively involved, making the interpretation difficult. Remember that the sum of all loadings is necessarily null. A more complex situation appears with the second principal coordinate. The larger loadings correspond to Th (positive) and Sb (negative), but many other elements participate with comparable loadings (see Table 1).

FIGURE 2

Table 1. Loadings of the two principal coordinates in the CoDa-PCA, explaining 49.3% of the total variance. They are the clr components of the principal element of the ilr-basis. As clr representations of compositions, the sum of these coefficients is zero. The difficulty to interpret the data is obvious in this case, as many of the loadings are of a similar magnitude.

	pc1	**pc2**		**pc1**	**pc2**		**pc1**	**pc2**
Ag	-0.15	-0.14	**Ga**	-0.15	0.12	**Sc**	0.04	0.19
Al	-0.07	0.19	**Hg**	-0.16	-0.15	**Se**	-0.21	0.06
As	-0.12	-0.05	**K**	0.01	0.12	**Sr**	0.31	-0.12
Au	-0.11	-0.46	**La**	-0.08	0.07	**Te**	-0.08	0.06
B	-0.09	0.08	**Mg**	0.22	0.24	**Th**	-0.05	0.32
Ba	0.17	-0.13	**Mn**	0.17	0.17	**Ti**	-0.05	-0.25
Bi	-0.08	-0.03	**Mo**	-0.13	-0.03	**Tl**	-0.15	0.05
Ca	0.67	-0.17	**Na**	0.01	0.05	**U**	0.00	0.05
Cd	0.10	-0.10	**Ni**	0.10	0.17	**V**	-0.13	0.08
Co	0.15	0.21	**P**	0.13	-0.05	**W**	-0.12	-0.15
Cr	-0.06	0.12	**Pb**	-0.11	-0.19	**Zn**	0.07	-0.06
Cu	0.10	-0.08	**S**	-0.01	-0.04			
Fe	-0.06	0.15	**Sb**	-0.07	-0.32			

The most appealing feature of the biplots is that the first principal component seems to separate the class of forest sample points (green) from the residential plot points (orange). However, the separation is not clear enough to discriminate every individual point, as some orange/green points are intercalated. This observation suggests that large ratios of Ca over other elements is a differential feature between the mentioned classes colored in green (forest) and yellow/violet (plots/residential). Other features like the association between Fe and Cr, visible in Table S1, are also discernible in the covariance biplot (Figure 2).

The difficulties encountered when interpreting principal coordinates suggest that principal balances (Martín-Fernández et al., 2018) would be useful to identify simple and sparse balances approaching principal coordinates and linearly associated elements. A clustering of the chemical elements based on the variation matrix provides a sequential binary partition which is visualized in the CoDa-dendrogram in Figure 3 (Pawlowsky-Glahn and Egozcue, 2011). The clustering of variables is seen (short vertical bars correspond to linear associations). Moreover, the colored bars correspond to different populations, classified as forest (green), non-residential plots (yellow), and residential plus recreational-leisure areas (violet). The CoDa-dendrogram in Figure 3 shows the differences in the mean of the balances for these three classes. Discrimination of the forest class seems quite reasonable based on some balances shown in Figure 3. Again, Ca is involved in two balances, placed on the right of the dendrogram, that distinguish between forest and the other two classes.

A relatively complex balance seems to separate the class corresponding to residential-recreational areas. This balance can be identified in Figure 3 as two groups of elements: Group A, including elements starting at Th and running to Te, which includes major non-toxic elements, or not highly toxic elements like K, Na, Al, Fe, associated to the geogenic elements of the area; and Group B, running from Au to W, which includes PTEs like Hg, Pb, As and Sb, which are more abundant in the residential-recreational sample points as reported in previous studies (Boente et al., 2018). This observation again suggests the predominance of air transportation of major PTEs.

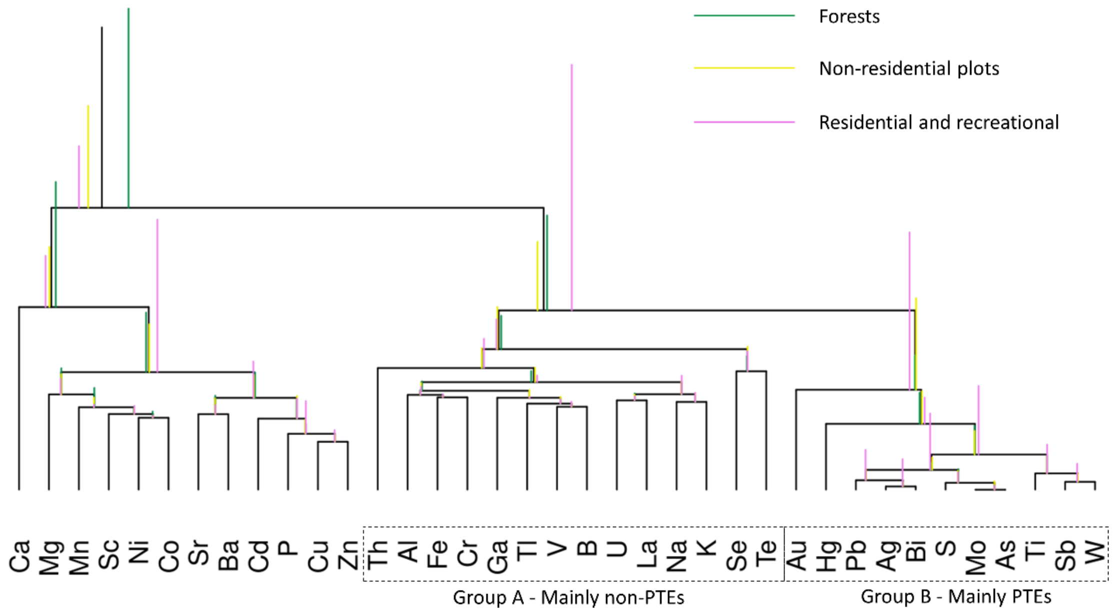


Figure 3. CoDa-dendrogram corresponding to (approximate) principal balances. It was obtained by clustering parts (chemical elements). The length of vertical bars over horizontal bars is proportional to the fraction of total variance associated with the split in the sequential binary partition defining the basis of balances.

3.3 Looking for background for pollution assessment

Quantifying the pollution of soils, or other media like air or water requires a full understanding of the term pollution-free soil. This background is commonly undefined or only partially known. An idea of the background in the Langreo case could be achieved as follows. As an external assessment of pollutants, the official admissibility thresholds for some chemical elements in soils (BOPA, 2014) were considered. These thresholds for some PTEs are given as an upper limit admissible value (in mg/kg). Moreover, the thresholds are specified depending on the land use. Table 2 shows these values in the columns on the left-hand side. Thresholds for other (Oth.) land uses are, in general, the most restrictive.

Table 2. Official thresholds (mg/kg) for some PTEs depending on the land use (labelled Ind. (Industrial), Urb. (Urban), Oth. (Other), and Recr. (Recreational)). On the left part of the Table, backgrounds (mg/kg) obtained: the column med is the element-wise median along the whole sample; columns labeled with a value correspond to the center of the sample trimmed to different values of the reduction coefficient. Non-available values are marked with - .

Element	Ind.	Urb.	Oth.	Recr.	med	α = 1	α = 0.8	α = 0.6
Ag	200	20	2	20	0.10	0.10	0.10	0.10
Al	-	-	-	-	11400	10913	10598	11084
As	200	40	40	40	18.40	17.30	16.90	15.40
Au	-	-	-	-	0.00	0.00	0.00	0.00
B	-	-	-	-	20.00	20.00	20.00	20.00
Ba	10000	10000	1540	10000	66.80	60.10	55.60	59.30
Be	205	30	20	140	-	-	-	-
Bi	-	-	-	-	0.40	0.40	0.40	0.30
Ca	-	-	-	-	2500	2212	2133	2029
Cd	200	20	2	20	0.30	0.30	0.30	0.20
Co	300	25	25	105	9.80	8.30	8.20	8.00
Cu	4000	400	55	400	22.70	18.30	17.10	16.40
Cr	10000	10000	10000	10000	18.60	17.10	16.50	16.60
Fe	-	-	-	-	27150	25719	25391	24120
Ga	-	-	-	-	4.10	3.80	3.70	3.50
Hg	100	10	1	10	0.30	0.30	0.20	0.20
K	-	-	-	-	1100	1073	1100	1213
La	-	-	-	-	9.50	9.00	8.90	9.10
Mg	-	-	-	-	1300	1237	1197	1239
Mn	9635	2135	2135	4970	545	442	436	414
Mo	600	60	6	60	0.90	0.80	0.70	0.70
Na	-	-	-	-	60	56	54	56
Ni	6500	650	65	4150	16.40	15.20	14.60	14.30
P	-	-	-	-	590	532	508	493
Pb	800	400	70	400	52.20	43.10	37.90	32.30
S	-	-	-	-	500	446	424	376
Sb	295	25	5	120	0.60	0.50	0.50	0.50
Sc	-	-	-	-	2.90	2.60	2.50	2.40
Se	2500	250	25	1740	0.80	0.70	0.70	0.60
Sn	10000	10000	4360	10000	-	-	-	-
Sr	-	-	-	-	16.60	15.40	14.90	15.00
Te	-	-	-	-	0.04	0.04	0.04	0.03
Th	-	-	-	-	2.90	2.90	3.00	2.90
Ti	10	1	1	3	20.00	20.10	18.60	19.60
Tl	-	-	-	-	0.20	0.20	0.20	0.20
U	-	-	-	-	1.10	1.00	1.00	1.00
V	1505	190	50	845	27.00	25.30	24.40	23.60
W	-	-	-	-	0.10	0.10	0.10	0.10
Zn	10000	4550	455	4550	107	92	83	77

Since we are looking for non-contaminated soil, it would be reasonable to take the Other land use thresholds (column Oth. In Table 2) as a reference. This set of thresholds for each element is denoted t_1. The non-available thresholds for elements for each element is denoted t_1. The non-available thresholds for elements in the Table (marked with -) are set to 10^6 mg/kg, thus meaning that everything is admissible. We can be more restrictive by multiplying these thresholds by a reduction coefficient like 0.9, 0.6 or similar. The procedure to find a background consists of filtering out samples that have one element or more over the selected threshold, thus extracting a reduced or trimmed sample.

Considering $t_\alpha = \alpha \cdot t_1$ for α = 1.00, 0.95, 0.90; … ; 0.50 (11 α values) the corresponding trimmed samples are obtained. The number of remaining samples after filtering is 95, 85, 81, 76, 71, 60, 49, 35, 25, 13, 6, out of the 150 initial samples, respectively. The compositional center (geometric mean for each element in mg/kg) can then be taken as representative for each trimmed sample. The element-wise median value of the concentrations in the sample is labelled med and is reported in Table 2. The center of the trimmed sample for some values (left columns, labelled with the value) is also shown in Table 2. The compositional center of each trimmed sample can then be taken as representative of a non-polluted background.

To visualize the backgrounds in Table 2, the centers of the trimmed samples were considered as a compositional sample and the corresponding biplots are shown in Figure S2 in the Supplementary materials. Note that the origin of rays in the plot corresponds to the center of the different backgrounds used in the plot and has no particular interest. Note also that these sets of thresholds, here called backgrounds, are not comparable to soil compositions and are considered here for their visualization. These biplots support discussion on the selection of a trimmed sample; see Supplementary materials.

The backgrounds obtained for different αs can also be compared jointly plotting their clr. Figure S3 in Section A.4 in Supplementary materials shows this comparison, which does not provide further insight into the characteristics of the backgrounds. After examining Figure S2 and based on the discussion of it in the Supplementary materials, $\alpha = 0.6$ was selected to choose a convenient background representing non-polluted soil.

3.4 Aitchison distance to background spread sample

Once a trimmed sample and its center are available, a first approach consists of computing the Aitchison distances of each point in the whole sample to the center of the (non-polluted) background. These distances define a preliminary contamination indicator: zero corresponds to the center of the background while large distances correspond to increasingly more polluted sites. These distances can then be transformed monotonically to obtain more scalable values. However, the mentioned Aitchison distances do not behave as expected. There are points within the reference trimmed sample whose Aitchison distance to the center is in the third quartile of distances in the whole sample. This finding is somewhat disappointing: samples in the trimmed sample assumed not to be polluted show distances of the order of other samples considered polluted. This is possible if the trimmed sample is compositionally dispersed. Figure S4 in Supplementary materials shows the geographical locations of the trimmed sample for $\alpha = 0.6$ marked with a plus sign. The crosses are spread over the whole region where fluctuations in geology are expected. The alternative is to consider that the background is not defined by the center of the trimmed sample, which is a single composition, but rather by the whole trimmed sample. In this way, the background can be thought of as a geological fluctuation described by the trimmed sample.

Then, the Spread Aitchison distance or pollution size is defined as

$$S_a(x_i) = \min_{x_{tr}} d_a(x_i, x_{tr}), \tag{8}$$

Where x_{tr} spans all the points in the trimmed sample and x_i moves over the available sample. When x_i belongs to the trimmed sample $S_a(x_i) = 0$ is, the point is considered not polluted. Figure S4 (Supplementary materials) shows the sampling points coloured following the quantiles of S_a (see caption). All points in the trimmed sample, marked with a plus sign, correspond to the first quartile of S_a (green points).

3.5 Balances as proxies of S_a: Compositional Pollution Indicators

The major inconvenience of S_a as pollution size is that it depends on all elements reported in the sample and also on the selection of the trimmed sample. It is therefore convenient to simplify the expression of S_a so that the selected proxy contains only a few elements commonly reported in samples and corresponding to the requirements enumerated in Section 2.

Three approaches were explored for the chemical sample: the first taking into account the whole observed composition; the second only a subcomposition, as suggested in Boente et al. (2018), which reports elements such as Na, K, Ca, Al, Mg, Fe as non-pollutant elements (mainly natural sources), and Cu, Pb, Zn, As, Sb, Hg as pollutants (mainly anthropogenic sources); and the third based on expert opinion using elements in the above-mentioned subcomposition. These approaches provide balances (Eq. 7) as Compositional Pollution Indicators (CIs), thus satisfying the requirements for Cis explained in Section 2.1. However, the characteristics of the Langreo region and the available data set do not allow distinctions between different sources of pollution. For the first indicator, CI_1, the strategy is to look for a balance optimally predicting S_a based on the whole observed composition. This can be done using the selbal procedure for the prediction of S_a as a continuous response (Rivera-Pinto et al., 2018). In the analysis of the complete composition, the result obtained was the balance

$$CI_1 = \sqrt{\frac{2}{3}}\left(\ln\frac{Sb}{(K\cdot Al)^{\frac{1}{2}}}\right), \tag{9}$$

which optimally predicts S_a after excluding the zero-distances corresponding to the trimmed sample. The linear regression gives $R^2 = 0.6$, which is not very high but still large enough to consider CI_1 a good proxy for pollution size. A predicting balance can be selected in several ways. For instance, taking logarithms on S_a after removing the zeros; not removing zeros of S_a; and not taking logs on Sa. In all cases, Sb appears in the numerator of the balance, and in the denominator, there is K or Al, or both. Figure 4 (left panel) shows the regression line when CI_1 is used to predict the spread Aitchison distance to the trimmed sample representing the background denoted S_a. In the analysis of the subcomposition, the balance considered optimal after cross-validation in the selbal procedure is different, but it includes Sb in the numerator and (Al; K) in the denominator. The optimal balance, using the subcomposition, is then

$$CI_2 = \sqrt{\frac{9}{6}}\ln\left(\frac{(Sb\cdot Pb\cdot Hg)^{1/3}}{(K\cdot Al\cdot Zn)^{\frac{1}{3}}}\right), \tag{10}$$

This balance was obtained after removing the zero distances to elements of the trimmed sample and predicting $\ln(S_a)$. When predicting S_a, without logarithm, the balance obtained is the same but removing Hg from the numerator.

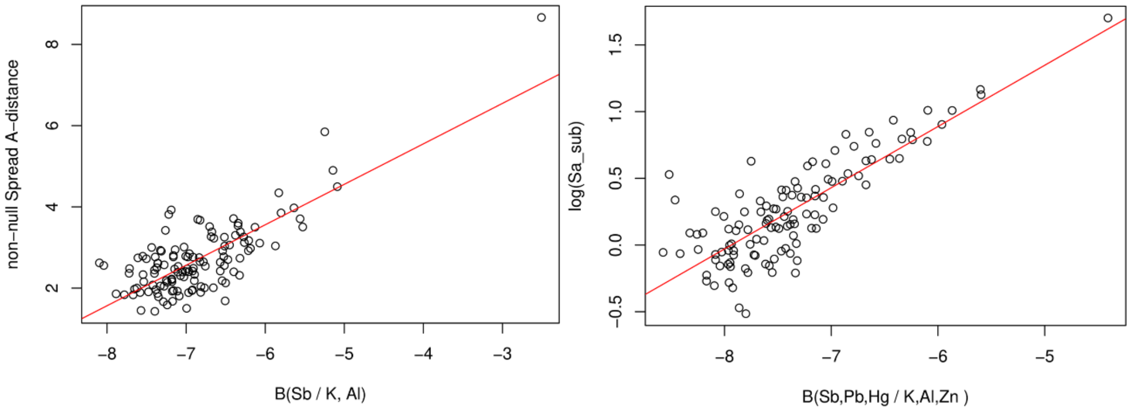


Figure 4. Regression lines of spread Aitchison distance, S_a, on the balance $CI_1 = B\left(\frac{Sb}{K}, Al\right)$; R^2 = 0.6, using the whole sample, left panel. Right panel: Regression of ln(S_a) on the balances $CI_2 = B\left(Sb; Pb; \frac{Hg}{K}; Al; Zn\right)$; R^2= 0.68 in the analysis. Points for which $S_a = 0$ were excluded.

The third balance was obtained based on expert criteria after conventional examination of the geochemical data set using multivariate procedures. Unlike the previous approaches, these criteria attend a selection of elements, of which some are considered pollutants while others are not. In the case of Langreo, the identification of the main pollutants was addressed in a previous study (Boente et al., 2018), where the authors stated that the main contaminants were typical pollutants such as As, Hg or Pb. while the main natural-source elements (or non-pollutants) were several major elements (i.e., Al, Ca, Fe, K, Mg, and Na). Based on this previous study, the selected balance, CI$_3$, was

$$CI_3 = \sqrt{\frac{30}{11}} \ln\left(\frac{(As \cdot Cu \cdot Hg \cdot Pb \cdot Zn)^{1/5}}{(Al \cdot Ca \cdot Fe \cdot K \cdot Mg \cdot Na)^{\frac{1}{6}}}\right), \qquad (11)$$

In conclusion, the compositional analysis revealed that overall pollution in the Langreo area is related to the relative content of Sb. Note that the chemistry of this element is similar to that of As, as both are metalloids that present a high geochemical affinity and are commonly enriched together in soils (Casiot et al., 2007; Wilson et al., 2010). In fact, As (and also Sb) are well-known soil contaminants in regions that host heavy industry, power stations and coal mining (Woon et al., 2021; Rodriguez-Iruretagoiena et al., 2015), like Langreo (Boente et al., 2020a). However, the association between As and Sb is not confirmed in the Langreo data set, as can be seen in the normalized variation matrix in Table S1.

The balance CI$_1$ is a log-contrast between a contaminant, Sb, over other non-contaminant elements such as K or Al, which are lithogenic and usually linked to natural clays and other soil minerals. When few elements are considered, as in the CI$_2$ analysis, Sb still appears in the balance and is complemented by two typical pollutants like Pb and Hg (also abundant in the Langreo area). The denominator has elements that are not usually considered pollutants and that are stable (compositional relative scale) across the study area, like Al, K, and Zn. The idea that CI_1 and CI_2 are suitable measures of the pollution size is reinforced by the fact that, of the 37 elements studied, these few elements are included within those considered pollutants and non-pollutants, respectively, according to the expert criteria in the construction of CI_3.

The configuration of the three CIs proposed, pollutants in the numerator and non-pollutants in the denominator, implies that the larger the value of the CI, the larger the relative pollution in the studied point. Some values of CIs evaluated on the trimmed sample (background) illustrate the scales of the three CIs. Reference thresholds for the CIs were chosen as explained in Supplementary materials, Section A.5. The reference values were -6.96, -7.52, and -7.91 for CI_1, CI_2 and CI_3 respectively. When finding values over these thresholds, one expects an approximately 70 - 75% probability of exceeding some official threshold of admissibility. See Table S2 in the Supplementary material for further details.

3.6 Spatial distribution: significant clusters definition

Isotropic variograms computed and corresponding models fitted are shown in Figure 5 for each of the selected indicators (CI_1, CI_2 and CI_3). No clear evidence of anisotropies was found. Cross-validation correlation indices of the observed and estimated CIs ranged between 0.70 and 0.88 and, therefore, results were considered satisfactory for the selected models. At first sight, all three indicators show a similar distribution over the study area. They are also similar to the maps presented in Boente et al. (2018), thus validating previous results. However, some differences call for discussion.

Visual comparison of Figures 1 and 5 reveals that the balance obtained by means of expert criteria (CI_3) presents a good representation of hot points, specially of the city and industrial areas, thereby confirming the larger pollution detected in previous studies (Boente et al., 2018; Martínez et al., 2014), while the areas to the east and south of Langreo appear to have predominantly low contamination, as corresponds to natural soils and forests. The northwestern area of Langreo appears partially with high values of the indicators, specially CI_2, CI_3, because it is enriched in Hg, as previously identified given the presence of old Hg-mining activities in the surroundings (González-Fernández et al., 2018), whereas the northern area of the municipality is also partially red. This observation is attributable to the preferential wind direction according to a study of the air quality in Langreo (Martínez et al., 2014). In general, CI_3 presents sharper contours, probably because more elements, pollutants or not, are explicitly involved in its expression. The design of an indicator like CI_3 has the inconvenience that it requires the hand of an expert using geochemical tools to manually define elements that are dangerous and those that better represent the geology of the area.

The indicators constructed using *selbal*, namely CI_1 and CI_2, both contain Sb as a driving pollutant. This finding is consistent with the fact that Sb has a similar chemistry to that of As, which has been reported to be enriched in the area (Boente et al., 2018). However, the agreement with the underlying assumptions on sample space and the scale, as well as the absence of outliers, provides higher robustness for the compositional analysis, focusing on the compositional criteria indicators. For this case study, the selection of Zn as part of the compositional baseline (but not in the group of pollutants for CI_2) indicates a partial relationship with geogenic elements like K and Al (Boente et al., 2018).

Regarding the results, CI_1 and CI_2 show similar distributions. In this context, both highlight the city and its surroundings as the main area affected by pollution.

Nevertheless, the absence of other PTEs enriched in soils like Cu or As, or even the inclusion of Zn in the denominator in the case of CI_2, leads to a less sharp definition of other hot points and blurs the maps, as can be seen particularly for CI_1 in Figure 5. In global terms, both CoDa-driven CIs are suitable to indicate the location of major pollution.

Attending to the definition of red/blue shapes, it seems easier to identify polluted areas in Figure 5(b) than in SGS presented in Figure 5(a). These Figure 5(b) maps predict the probability of exceeding thresholds for each CI: -6.96, -7.52 and -7.91 for CI_1, CI_2 and CI_3, respectively. They are roughly similar to the spatial interpolation of the CIs themselves, but here a smoothing effect can be appreciated that induces a sharper definition of the principal hazardous areas, as well as other minor locations, thus providing greater robustness to the predictions. Here, once again, the effect of considering a lower number of pollutants in CI_1, particularly the role of Sb, is visible as there are areas that do not appear in red, such as the occidental one. In this respect, the mathematically obtained CI_2 and the manual selection of CI_3 seem to be more accurate and closer to reality.

Finally, the spatial distribution of CI_1 is more complex to interpret, as the areas of high/low values appear to be mixed. Nevertheless, the spatial patterns obtained are consistent with the other two indicators, showing a northern hot-spot and a southern cold-spot. These results thus evidence that, when using K and Al as a reference of natural sourcing, Sb alone is a suitable predictor of pollution in the area.

$$CI_1 = \sqrt{\frac{2}{3}} \cdot ln\left[\frac{Sb}{(K\ Al)^{1/2}}\right] \quad CI_2 = \sqrt{\frac{3}{2}} \cdot ln\left[\frac{(Hg\ Pb\ Sb)^{1/3}}{(K\ Al\ Zn)^{1/3}}\right] CI_3 = \sqrt{\frac{30}{11}} \cdot ln\left[\frac{(Cu\ Pb\ As\ Hg\ Zn)^{1/5}}{(Fe\ Al\ Ca\ Mg\ K\ Na)^{1/6}}\right]$$

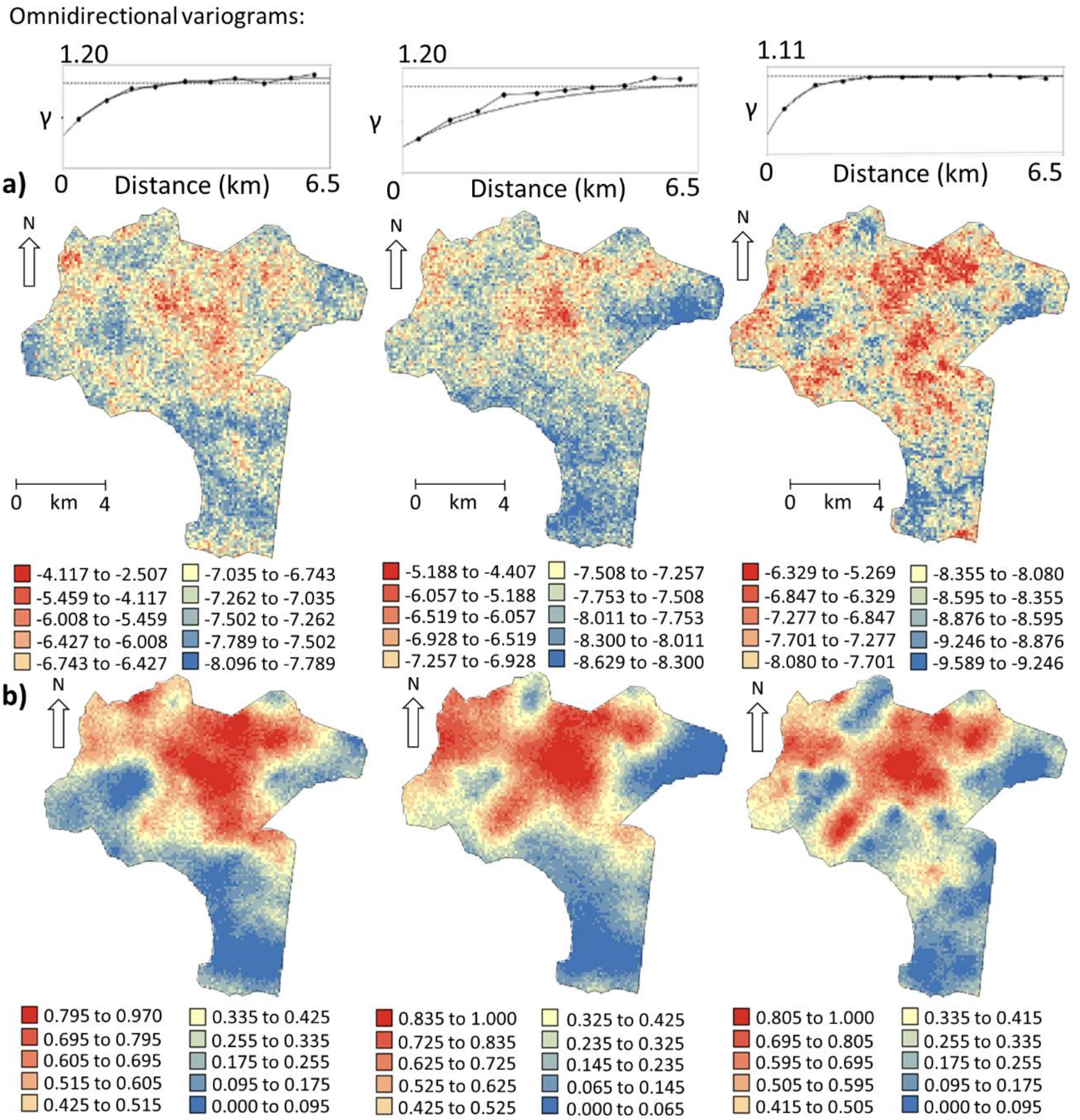


Figure 5. (a) SGS average images (MI) and (b) probability maps of exceeding the defined threshold for CI1 (left, threshold -6.96), CI2 (middle, threshold -7.52), CI3 (right, threshold -7.91). Fitted omnidirectional variograms are also shown. The colour scales correspond to Jenks natural breaks classification.

4. Conclusions

Geochemical data are compositional data, as the concentrations of elements in any environmental matrix are commonly expressed as parts of a whole and vary together. Once established this feature, it is possible to apply Compositional Data procedures to obtain indicators that address pollution, for instance, in soils.

Here, we presented a novel methodology to address soil pollution basing on compositional principles. The strength of this methodology is that it allows to build compositional-based, non-polluted background and indicators measuring the deviation from the background to obtain a wide view of PTEs pollution. The indicators produced are easily programmable in R packages, and allow an easy and intuitive identification of the most polluted subareas, offering a proper overview of pollution for both large and small scales for both experienced and unexperienced users. An additional possibility we have checked here to enhance the interpretation of pollution is to build maps showing the probability of exceeding defined thresholds through SGS.

With respect to the weaknesses, one of the most important is that, unlike other classical single-component indices, the indicators obtained in this work are only valid for the example of Langreo, whereas the novel methodology proposed must be computed for each case study. Moreover, as indicators are based on concentration data, they are useful as they offer a global map of pollution, but this approach cannot use other geochemical variables such as the bioavailability of elements, the abundance of toxic species, or a precise assessment of pollution sources that should require forensic techniques. Thus, in further studies, it would be interesting to face these limitations by exploring whether other geochemical variables different to concentrations might be also expressed in a compositional way, and also if a complementary, specific, pollution sources study may complement CoDa results.

All things considered, the methodology presented constitutes a powerful tool for non-proficient users in the topic of soil pollution, public administration, or private companies. We encourage researchers to apply it in pollution prevention and effective environmental quality management, as it can be very useful for decision making and assessment of the variability through geostatistical analysis.

CRediT author statement

C. Boente: conceptualization, resources, data curation, formal analysis, writing - original draft; **M.T.D. Albuquerque**: software, formal analysis, visualization, writing - original draft; **J.R. Gallego**: Funding acquisition, supervision, validation, writing - review & editing; **V.Pawlowsky-Glahn**: methodology, visualization, writing - review & editing, validation, software; **J.J. Egozcue**: methodology, conceptualization, formal analysis, data curation, supervision, writing - original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

CB obtained a post-doctoral contract within the PAIDI 2020 program (Ref 707 DOC 01097), co-financed by the Junta de Andaluca (Andalusian Government) and the EU. JJE and VPG were supported by the Spanish Ministry of Science, Innovation and Universities and the European Regional Development Fund through grant RTI2018-095518-B-C21 (C22) (MCIU/AEI/FEDER).

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). J R STAT SOC B 44 (2): 139-177.

Aitchison, J. (1983). Principal component analysis of compositional data. BIOMETRIKA 70 (1): 57-65.

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. and M. Greenacre (2002). Biplots for compositional data. J R STAT SOC C 51 (4): 375-392.

Albuquerque, M., I. Antunes, M. Seco, N. Roque, and G. Sanz (2014). Sequential Gaussian simulation of uranium spatial distribution - a transboundary watershed case study. Procedia Earth Planet. Sci. 8: 2-6.

Baragaño, D., C. Boente, E. Rodríguez-Valdés, A. Fernández-Brana, A. Jiménez, J.R. Gallego, B. González-Fernández (2020). Arsenic release from pyrite ash waste over an active hydrogeological system and its effects on water quality. Environmental Science and Pollution Research 27: 10672-10684.

Barceló-Vidal, C. and J.-A. Martín-Fernández (2016). The mathematics of compositional analysis. Austrian J Stat 45: 57-71.

Batsaikhan, B., S.-T. Yun, K.-H. Kim, S. Yu, K.-J. Lee, Y.-J. Lee, and J. Namjil (2021). Groundwater contamination assessment in Ulaanbaatar city, Mongolia, with combined use of hydrochemical, environmental isotopic, and statistical approaches. SCI TOTAL ENVIRON 765: 14279.

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. J AM STAT ASSOC 96 (456): 1205-1214.

Boente, C., M. T. D. Albuquerque, A. Fernandez-Brana, S. Gerassis, C. Sierra, and J. R. Gallego (2018). Combining raw and compositional data to determine the spatial patterns of potentially toxic elements in soils. SCI TOTAL ENVIRON 632-631: 1117-1126.

Boente, C., D. Baragao, and J. Gallego (2020a). Benzo[a]pyrene sourcing and abundance in a coal region in transition reveals historical pollution, rendering soil screening levels impractical. Environ. Pollut. 266: 115341.

Boente, C., S. Gerassis, M. T. D. Albuquerque, J. Taboada, and J. R. Gallego (2020b). Local versus regional soil screening levels to identify potentially polluted areas. MATH GEOSCI 52: 381-396.

Boente, C., I. Martín-Méndez, A. Bel-Lan, and J. R. Gallego (2020c). A novel and synergistic geostatistical approach to identify sources and cores of potentially toxic elements in soils: An application in the region of cantabria (northern spain). J. Geochemical Explor. 208 (10639):7.

Boente, C., D. Baragaño, R. Forjan, N. García-González, A. Colina, and J.R. Gallego (2022). A holistic methodology to study geochemical and geomorphological control of the distribution of potentially toxic elements in soil. CATENA 208, 105730.

Boogaart van den, K. G., R. Tolosana-Delgado, and M. Bren (2009). compositions:

Compositional Data Analysis. R package version 1.02-1.

Boogaart, van den, K. G. and R. Tolosana-Delgado (2013). Analysing Compositional Data with R. Springer-Verlag, Berlin. 258 p.

BOPA (2014). Generic reference levels for heavy metals in soils from Principality of Asturias, spain. Boletín Oficial del Principado de Asturias. Accessed August 2021.

Buccianti, A. and E. Grunsky (2014). Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes. J. Geochemical Explor. 141: 1-5.

Buccianti, A., B. Nisi, and B. Raco (2016). Towards the Concept of Background/baseline Compositions: A Practicable Path? In: Compositional Data Analysis. CoDaWork 2015, Springer Proceedings in Mathematics & Statistics 187: 31-43. Springer, Cham.

Buccianti, A. and V. Pawlowsky-Glahn (2005). New Perspectives on Water Chemistry and Compositional Data Analysis. MATH GEOL 37 (7): 703-727.

Cachada, A., T. Rocha-Santos, and A. C. Duarte (2018). Soil and Pollution, in: Soil Pollution. Elsevier.

Casiot, C., M. Ujevic, M. Munoz, J. Seidel, and F. Elbaz-Poulichet (2007). Antimony and arsenic mobility in a creek draining an antimony mine abandoned 85 years ago (upper Orb basin, France). Appl. Geochemistry 22:788-798.

Chayes, F. (1962). Numerical correlation and petrographic variation. The Journal of Geology 70(4), 440-452.

Chayes, F. (1971). Ratio Correlation. University of Chicago Press, Chicago, IL (USA). 99p.

Cicchella, D., D. Zuzolo, S. Albanese, L. Fedele, D. Tota, G. I., T. I., D. V. M., and L. B. (2020). Urban soil contamination in salerno (italy): Concentrations and patterns of major, minor, trace and ultra-trace elements in soils. J. Geochemical Explor. 213: 106519.

Clemens, S. (2006). Toxic metal accumulation, responses to exposure and mechanisms of tolerance in plants. Biochimie 88: 1707-1719.

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

Egozcue, J. J. and V. Pawlowsky-Glahn (2006). Simplicial geometry for compositional data. In Compositional Data Analysis in the Geosciences: From Theory to Practice, Volume 264 of Special Publications: 145-159. Geol. Soc., London.

Egozcue, J. J. and V. Pawlowsky-Glahn (2018). Modelling compositional data. the sample space approach. In B. S. Daya Sagar, Q. Cheng, and F. Agterberg (Eds.), Handbook of Mathematical Geosciences - Fifty Years of IAMG, pp. XXV, 875. Springer International Publishing.

Egozcue, J. J. and V. Pawlowsky-Glahn (2019a). Compositional data: the sample space and its structure. TEST 28 (3): 599-638.

Egozcue, J. J. and V. Pawlowsky-Glahn (2019b). Compositional data: the sample space and its structure (with discussion). TEST 28 (3): 599-638. doi.org/10.1007/s11749-019-00670-6.

Egozcue, J. J., V. Pawlowsky-Glahn, and G. B. Gloor (2018). Linear association in compositional data analysis. Austrian J Stat 47 (1): 3-31.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. MATH GEOL 35 (3): 279-300.

Eynatten, H. von (2004). Statistical modelling of compositional trends in sediments. Sedimentary Geology 171: 79-89.

Fabian, C., C. Reimann, K. Fabian, M. Birke, R. Baritz, and E. Haslinger (2014). Gemas: Spatial distribution of the ph of european agricultural and grazing land soil. Appl. Geochemistry 48: 207-216.

Filzmoser, P. and K. Hron (2011). Compositional data analysis: Theory and applications. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis: Theory and Applications: 59-72. John Wiley & Sons.

Filzmoser, P., K. Hron, J. Martín-Fernández, and J. Palarea-Albaladejo (2021). Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn. Springer International Publishing.

Filzmoser, P., K. Hron, and C. Reimann (2009). Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. SCI TOTAL ENVIRON 407 (23): 6100-6108.

Gallego, J., E. Rodríguez-Valdés, N. Esquinas, A. Fernández-Braña, and E. Afif (2016). Insights into a 20-ha multi-contaminated brownfield megasite: An environmental forensics approach. SCI TOTAL ENVIRON 563-564: 683-692.

González-Fernández, B., E. Rodríguez-Valdés, C. Boente, E. Menéndez-Casares, A. Fernández-Brana, and J. Gallego (2018). Long-term ongoing impact of arsenic contamination on the environmental compartments of a former mining-metallurgy area. SCI TOTAL ENVIRON 610: 820-830.

Goovaerts, P. (1997). Geostatistics for Natural Resources Evaluation. Applied Geostatistics Series. Oxford University Press, New York, NY (USA). 483 p.

Graziano, S., G. R., and B. C. (2020). Is compositional data analysis (coda) a theory able to discover complex dynamics in aqueous geochemical systems? J. Geochemical Explor. 211: 106465.

Hadjipanagiotou, C., A. Christou, A. M. Zissimos, E. Chatzitheodoridis, and S.P. Varnavas (2020). Contamination of stream waters, sediments and agricultural soil in the surroundings of an abandoned copper mine by potentially toxic elements and associated environmental and potential human healthderived risks: a case study from agrokipia, Cyprus. Environmental Science and Pollution Research 27, 41279-41298.

Hakanson, L. (1980). An ecological risk index for aquatic pollution control.a

sedimentological approach. Water Res 14: 975{1001.

Jarauta-Bragulat, E., C. Hervada-Sala, and J. J. Egozcue (2016). Air quality index revisited from a compositional point of view. MATH GEOSCI 48:581-593.

Jenks, G. F. (1967). The data model concept in statistical mapping. International Yearbook of Cartography 7: 186-190.

Journel, A. G. and C. J. Huijbregts (1978). Mining Geostatistics. Academic Press, London (UK). 600 p.

Juma, D. W., H. Wang, and F. Li (2014). Impacts of population growth and economic development on water quality of a lake: case study of lakebvictoria kenya water. Environ. Sci. Pollut. Res. 21: 5737-5746.

Kabata-Pendias, A. (2010). Trace Elements in Soils and Plants. CRC Press, Boca Raton, USA. 548 p.

Kelepertzis, E., A. Argyraki, V. Chrastny, F. Botsou, K. Skordas, M. Komarek, and A. Fouskas (2020). Metal(loid) and isotopic tracing of pb in soils, road and house dusts from the industrial area of volos (central greece). SCI TOTAL ENVIRON 725: 13830.

Khanam, R., A. Kumar, A. Nayak, M. Shahid, R. Tripathi, S. Vijayakumar, D. Bhaduri, U. Kumar, S. Mohanty, P. Panneerselvam, D. Chatterjee, B. Satapathy, and H. Pathak (2020). Metal(loid)s (as, hg, se, pb and cd) in paddy soil: Bioavailability and potential risk to human health. SCI TOTAL ENVIRON 699: 13433.

Kowalska, J. B., R. Mazurek, M. Gasiorek, and T. Zaleski (2018). Pollution indices as useful tools for the comprehensive evaluation of the degree of soil contamination: A review. Environ. Geochem. Health 40: 2395-2420.

Kynclova, P., K. Hron, and P. Filzmoser (2017). Correlation between compositional parts based on symmetric balances. MATH GEOSCI 49: 777-796. doi 10.1007/s11004-016-9669-3.

Lahr, J. and L. Kooistra (2010). Environmental risk mapping of pollutants: State of the art and communication aspects. SCI TOTAL ENVIRON 408: 3899-3907.

Lovell, D., V. Pawlowsky-Glahn, J. J. Egozcue, S. Marguerat, and J. Bahler

(2015). Proportionality: A valid alternative to correlation for relative data. PLoS Comput Biol 11 (3): e1004075.

Madrid, L., E. Diaz-Barrientos, E. Ruiz-Cortes, R. Reinoso, M. Biasioli, C. M. Davidson, A. C. Duarte, H. Grcman, I. Hossack, A. S. Hursthouse, T. Kralj, K. Ljung, E. Otabbong, S. Rodrigues, G. J. Urquhart, and F. Ajmone-Marsan (2006). Variability in concentrations of potentially toxic elements in urban parks from six european cities. J. Environ. Monit. 8: 1158-1165.

Martín-Fernández, J. A. (2019). Comments on: Compositional data: the sample space and its structure, by egozcue and pawlowsky-glahn. TEST 28 (3): 653-657.

Martín-Fernández, J. A., V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosona-Delgado (2018). Advances in principal balances for compositional data. MATH GEOSCI 50: 273-298.

Martínez, J., J. Pineiro, C. Iglesias, J. Tabiada, J. Sancho, J. Pastor, A. Saavedra, and P. García-Nieto (2014). Air quality parameters outliers detection using functional data analysis in the langreo urban area (northern spain). Appl. Math. Comput. 241: 1{10.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. J. Egozcue (2011). The principle of working on coordinates. In Pawlowsky-Glahn and Buccianti (2011): 31-42.

Matheron, G. (1971). The Theory of Regionalized Variables and Its Applications. Les Cahiers du Centre de Morphologie Mathématique 5, Ecole des Mines de Paris. 211 p.

McIlwaine, R., S. F. Cox, R. Doherty, S. Palmer, U. Ofterdinger, and J. M. McKinley (2014). Comparison of methods used to calculate typical threshold values for potentially toxic elements in soil. Environ. Geochem. Health 36: 953-971.

McKinley, J. M., K. Hron, E. C. Grunsky, C. Reimann, P. de Caritat, P. Filzmoser, K. G. van den Boogaart, and R. Tolosana-Delgado (2016). The single component geochemical map: Fact or fiction? J. Geochem. Explor. 162: 16-28.

Megido, L., B. Suárez-Peña, L. Negra, L. Castrillón and Y. Fernández-Nava (2017). Suburban air quality: Human health hazard assessment of potentially toxic elements in pm10. Chemosphere 177, 284-291.

Mueller, U. A. and Grunsky, E. C. (2016). Multivariate Spatial Analysis of Lake Sediment Geochemical Data; Melville Peninsula, Nunavut, Canada. Applied Geochemistry 75(1): 247-262. Doi:10.1016/j.apgeochem.2016.02.007.

Muller, G. (1969). Index of geoaccumulation in sediments of the rhine river. Geol. J. 2: 108-118.

Mullineaux, S. T., J. M. McKinley, N. J. Marks, D. M. Scantlebury, and R. Doherty (2021). Heavy metal (pte) ecotoxicology, data review: Traditional vs. a compositional approach. SCI TOTAL ENVIRON 769 (14524): 6.

Olea, R. A., J. A. Luppens, J. J. Egozcue, and V. Pawlowsky-Glahn (2016). Caloric value and compositional ultimate analysis with a case study of a texas lignite. Int. J. Coal Geol. 162: 27-33.

Parent, S. E., Parent, L. E., Egozcue, J. J., Rozane, D. E., Hernandes, A., Lapointe, L., Hebert-Gentile, V., Naess, K., Marchand, S., Lafond J., Mattos, D. Jr., Barlow, P. and Natale, W. (2013). The plant ionome revisited by the nutrient balance concept. Frontiers in Plant Science 4: 1-10.

Pawlowsky-Glahn, V. and A. Buccianti (Eds.) (2011a). Compositional Data Analysis: Theory and Applications. John Wiley & Sons. 378 p.

Pawlowsky-Glahn, V. and J. Egozcue (2011). Exploring Compositional Data with the Coda-Dendrogram. Austrian J Stat 40 (1 & 2): 103-113.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. Stoch. Environ. Res. Risk Assess. 15 (5): 384-398.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). Modeling and analysis of compositional data. Statistics in practice. John Wiley & Sons, Chichester UK. 272 p.

Pawlowsky-Glahn, V. and J. Serra (Eds.) (2019). Matheron's Theory of Regionalised Variables. Oxford University Press. 208 p.

Peh, Z., S. Miko, and O. Hasan (2010). Geochemical background in soils: a linear process domain? An example from Istria (Croatia). Earth. Sci.Environ. 59: 1367-1383.

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

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Reimann, C., P. Filzmoser, and R. G. Garrett (2005). Background and threshold: critical comparison of methods of determination. SCI TOTAL ENVIRON 346: 1-16.

Rivera-Pinto, J., J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle (2018). Balances: a new perspective for microbiome analysis. mSystems 3 (4).

Rodriguez-Iruretagoiena, A., S. Fdez-Ortiz de Vallejuelo, A. Gredilla, C. G. Ramos, M. L. Oliveira, G. Arana, A. de Diego, J. M. Madariaga, and L. F. Silva (2015). Fate of hazardous elements in agricultural soils surrounding a coal power plant complex from santa catarina (brazil). SCI TOTAL ENVIRON 508: 374{382.

Sánchez de la Campa, A. M., D. Sánchez-Rodas, L. Alsiou, A. Alastuey, X. Querol, and J. D. de la Rosa (2018). Air quality trends in an industrialised area of sw spain. J. Clean. Prod. 186: 465{474.

Sowden, M., D. Blake, D. Cohen, A. Atanacio, and U. Mueller (2020). Development of an infrared pollution index to identify ground-level compositional, particle size, and humidity changes using himawari-8. Atmos. Environ 229 (11743): 5.

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

Tepanosyan, G., L. Sahakyan, N. Maghakyan, and A. Saghatelyan (2020). Combination of compositional data analysis and machine learning approaches to identify sources and geochemical associations of potentially toxic elements in soil and assess the associated human health risk in a mining city. Environ. Pollut. 261: 11421.

Tolosana-Delgado, R., Otero, N., Pawlowsky-Glahn, V., Soler, A. (2005). Latent composicional factors in the Llobregat river basin (Spain) hydrogeochemistry. Mathematical Geology 37(7): 681-702.

Wang, Z., X. Chen, D. Yu, L. Zhang, J. Wang, and J. Lv (2021). Source apportionment and spatial distribution of potentially toxic elements in soils: A new exploration on receptor and geostatistical models. SCI TOTAL ENVIRON 759 (14342): 8.

Wei, Y., Z. Wang, H. Wang, T. Yao, and Y. Li (2018). Promoting inclusive water governance and forecasting the structure of water consumption based on compositional data: A case study of beijing. SCI TOTAL ENVIRON 634: 407-416.

Wilson, S., P. Lockwood, P. Ashley, and M. Tighe (2010). The chemistry and behaviour of antimony in the soil environment with comparisons to arsenic: A critical review. Environ. Pollut. 158: 1169-1181.

Woon, S., K. Srinuansom, C. Chuah, S. J. Ramchunder, J. Promya, and A. Ziegler (2021). Pre-closure assessment of elevated arsenic and other potential environmental constraints to developing aquaculture and fisheries: The case of the mae moh mine and power plant, lampang, thailand. Chemosphere 269: 128682.

Yotova, G., M. Padareva, M. Hristova, A. Astel, M. Georgieva, N. Dinev, and S. Tsakovski (2018). Establishment of geochemical background and threshold values for 8 potential toxic elements in the bulgarian soil quality monitoring network. SCI TOTAL ENVIRON 643: 1297-1303.

Zuzolo, D., D. Cicchella, A. Lima, I. Guagliardi, P. Cerino, A. Pizzolante, M. Thiombane, B. De Vivo, and S. Albanese (2020). Potentially toxic elements in soils of campania region (southern italy): Combining raw and compositional data. J. Geochemical Explor. 213 (10652): 4.