Combining raw and compositional data to determine the spatial patterns of Potentially Toxic Elements in soils.

C. Boente[a], M.T.D. Albuquerque[b], A. Fernández-Braña[a], S. Gerassis[c], C. Sierra[d], J.R. Gallego[a]

[a] INDUROT and Environmental Technology, Biotechnology, and Geochemistry Group, Universidad de Oviedo, Campus de Mieres, 33600 Mieres (Asturias), Spain

[b] Instituto Politécnico de Castelo Branco, 6001-909 Castelo Branco, Portugal and CERENA/FEUP Research Center, Portugal

[c] Department of Natural Resources and Environmental Engineering, Univ. of Vigo, Lagoas Marcosende, 36310 Vigo, Spain

[d] Departamento de Transportes, Tecnología de Procesos y Proyectos, Universidad de Cantabria, Campus de Torrelavega,39300 Torrelavega (Cantabria), Spain

Abstract

When considering complex scenarios involving a multiset of attributes, such as in environmental characterization, a clearer picture of reality can be achieved through the dimensional reduction of data.

In this context, maps facilitate the visualization of spatial patterns of contaminant distribution and the identification of enriched areas. Here we measured a set of 15 Potentially Toxic Elements (PTEs) – (As, Ba, Cd, Co, Cr, Cu, Hg, Mo, Ni, Pb, Sb, Se, Tl, V, and Zn) in soil collected in the municipality of Langreo (80 Km^2), in Asturias, northern Spain, a paradigmatic industrial area.

With the aim to explore PTE dissemination trends and to define clusters of relative enrichment, we examined the mechanisms through which these contaminants are spatially distributed.

Relative enrichment (RE) is introduced here to refer to the proportion of elements present in a given context. Indeed, we provide a new approach to research into PTE fate. This

method involves studying the variability of PTE proportions throughout the study area, thereby allowing the identification of dissemination trends.

Transformations to open closed data are widely used for this purpose. As compositions are shown along with their spatial locations, spatial patterns have an indubitable interest. In this study, we used the Centered Log-ratio transformation (*clr*), followed by its back-transformation, to build a set of compositional data that, combined with raw data, allowed us to establish the sources of the PTEs and trends of spatial dissemination.

Based on our findings, we conclude that the Langreo area is deeply affected by its industrial and mining legacy. The city centre is highly enriched in Pb and Hg and As showed enrichment in a northwesterly direction. Overall, the multivariate geochemical approach presented facilitates the identification and quantification of anthropogenic impacts and consequent adequate monitoring measures required to safeguard the health of local communities.

Keywords: Soil Pollution, PTEs, Compositional Data, Ordinary Kriging, Local G-clustering, Relative Enrichment.

1. Introduction

Environmental characterization involves complex scenarios in which a multiset of attributes must be considered. A dimensional reduction of data is pivotal to gain a clear picture of reality (Moen and Ale, 1998). Maps are useful to visualize pollutant concentrations, as well as to determine zones of contaminant enrichment, whether natural or caused by anthropogenic activity. In this context, Potentially Toxic Elements (PTEs) are increasingly affecting soils all over the world, thus posing a threat to both public health and the environment (McIlwaine et al., 2016). The presence of these elements in soils can be explained by many factors (Alloway, 1990), the growth of urbanization and resulting increase in industrial activities being among the most important (Biasioli et al., 2006). Given that high concentrations of PTEs can endanger human and environmental health, it is of utmost importance to characterize their spatial distribution, determine their source, and screen for enrichment trends (Fayiga and Saha, 2016; Li et al., 2014; Boente et al., 2017; Cachada et al., 2013).

The area of Langreo (Asturias, NW Spain) (Fig. 1) is one of the regions in the Iberian Peninsula most marked by industrialization (Gallego et al., 2016). Coal mining and industries devoted to energy, metallurgy, pharmacology, and fertilizers, among others, have been operating in this region for decades, leaving a lasting imprint on the environment (Martínez et al., 2014; Megido et al., 2017). In this regard, great amounts of PTEs have been identified in soils from former industrial plots in this area (Boente et al., 2016; Gallego et al., 2016).

Here we performed a comparative study of a set of 15 chemical elements, analyzed in soils gathered in the Langreo area (80 Km^2), paradigmatic industrial area as described above. In this sort of studies the distribution of PTEs cannot be studied by merely considering the total concentrations (raw data), especially when the concentration of chemical elements in almost all datasets is compositional (Pawlowsky-Glahn., 1989; Filzmoser et al 2009), where attributes vary together with all the others. In this context, transformations that open closed data are widely used and, as compositions are recorded along with their spatial locations, spatial patterns are of interest (Pawlowsky-Glahn., 1989). The contributions of Pawlowsky-Glahn to regionalized compositions (Pawlowsky-Glahn, 1989; Pawlowsky-Glahn and Burger, 1992; Pawlowsky-Glahn et al., 1995) and their applications are widely applied (Odeh et al., 2003; Lark and Bishop, 2007). In this context, multiple log-ratio transformations are commonly used, the most common being the additive log-ratio transformation (alr), the centered log-ratio transformation (clr) (e.g. Aitchison, 1986), and the isometric log-ratio transformation (ilr) (Egozcue et al., 2003). In this study, the clr transformation and its back-transformation were performed through CoDaPack v2.02.21 software to create a set of compositional data that provides information about the comparative magnitudes of their constituents. This compositional dataset was used to map patterns of RE, thereby allowing us to identify spatial dissemination trends for PTEs.

In summary, the main goal of this study was to test a methodology that, by means of combining raw and compositional data, has the capacity to identify spatial patterns, areas of pollution risk and anthropogenic or natural sources of PTEs. All the evidence provided is supported by uni- and multi-variate statistical analysis, together with ordinary kriging and Local G clustering for the area of Langreo. Finally, core strengths and weaknesses are extrapolated to make this methodology useful and applicable to studies of a similar nature.

2. Materials and Methods

2.1. Study area

Covering 80 km^2, the municipality of Langreo (Asturias, NW Spain, Fig. 1) has a history of mining and industrial activity that dates back to the 1850s (Martínez et al., 2014). This activity left behind a legacy of polluted sites, making this zone one of the most contaminated areas in northern Spain (Gallego et al., 2016) and thus an ideal site in which to test the method presented in this study.

The region lies along the Nalón River, which is the longest and the most voluminous in Asturias. Altitudes in the area vary from 200 m (location of the urban areas and industry) to 900 m (rural environments, forests), with the presence of steep mountains. This geography gives rise to an enclosed area that facilitates the accumulation of PTEs by atmospheric deposition.

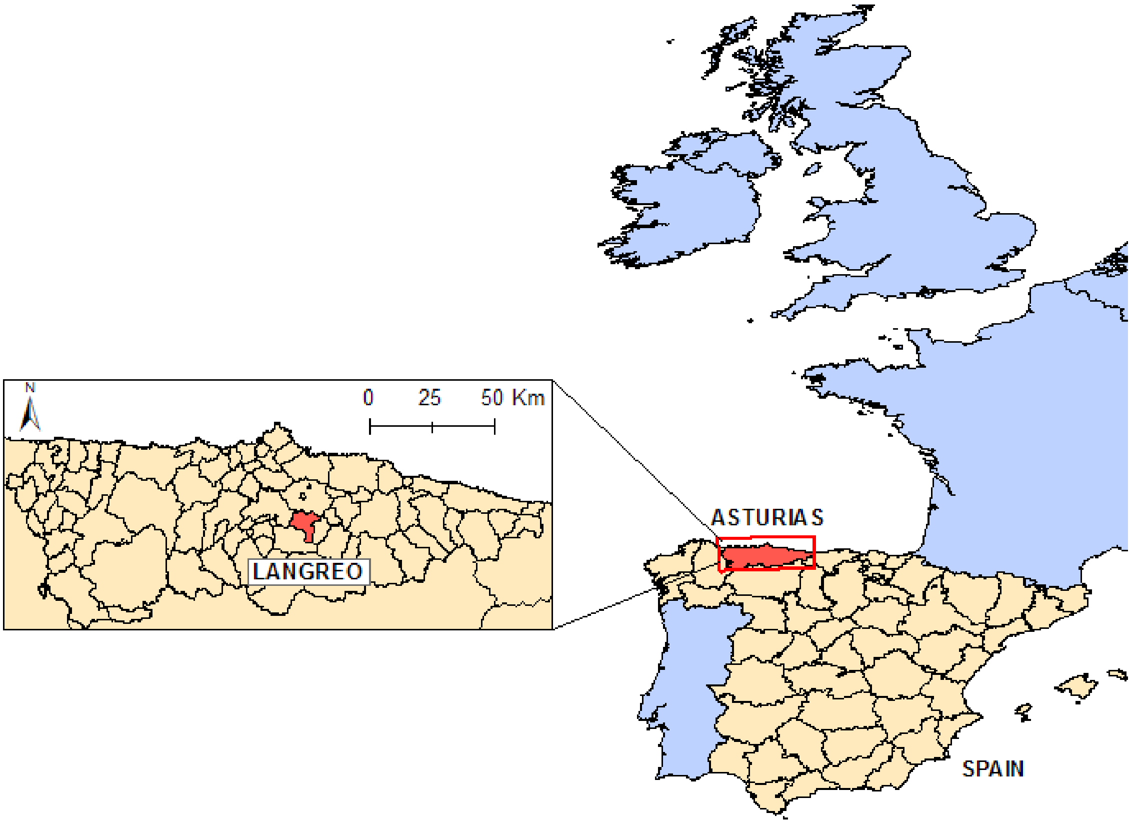


Fig. 1. Location of the study area in the municipality of Langreo in Asturias, Spain.

2.2. Data collection and chemical analyses

Samples were collected using a stratified systematic sampling method at random distances to obtain a representative set of data on the total variability of PTE content and site diversity (natural or anthropic environments, geomorphology, land uses, etc.). To this end, 10 equidistant transects, 250 m wide and each one 1000 m apart, were distributed perpendicular to the Nalón River (Fig. 2). A total of 150 samples were collected, the number *per* transect being determined proportionally to its length. The sample location within each transect was selected at random (Fig. 2).

Each sample composed of five increases taken from each vertex of a 1-m edge square and its central point from the top 20-25 cm of the soil, using an Edelman Auger. Afterwards, samples were passed through a 2-cm mesh screen *in situ* to remove large material such as organic matter, rocks and gravel. The samples were then dried in an oven at 35ºC to prevent the evaporation of volatile compounds, and finally quartered by means of a Jones riffle splitter for soil homogenization and representativeness.

These fractions were ground in an RS100 Resch mill at 400 RPM for 40 s. Then, 1-g representative sub-samples were sent to the ISO 9002-accredited Bureau Veritas Laboratories (Vancouver, Canada) and subjected to 1:1:1 “aqua regia” digestion. The total concentrations of the elements Ag, Al, As, Au, B, Ba, Bi, Ca, Cd, Co, Cr, Cu, Fe, Ga, Hg, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, S, Sb, Sc, Se, Sr, Te, Th, Ti, Tl, U, V, W and Zn in the digested material were determined by Inductively Coupled Plasma-Optical Emission Spectroscopy (ICP-OES).

A subset of the analyzed elements corresponding to PTEs was used for this study. This subset was chosen because it represented a set of typical contaminants (heavy metal(loid)s) found in environmental studies in Asturias (Albuquerque et al., 2017; Boente et al., 2016; Gallego et al., 2015), in addition the Risk Based Soil Screening Levels (RBSSLs) for these contaminants are available for this region of Spain (BOPA, 2014). Furthermore, the dispersal of the concentrations of these contaminants never exceeded three orders of magnitude and thus provided readable proportions. Therefore, of the original list of 36 elements, the following 15 were examined (PTE group): As, Ba, Cd, Co, Cr, Cu, Hg, Mo, Ni, Pb, Sb, Se, Tl, V and Zn.

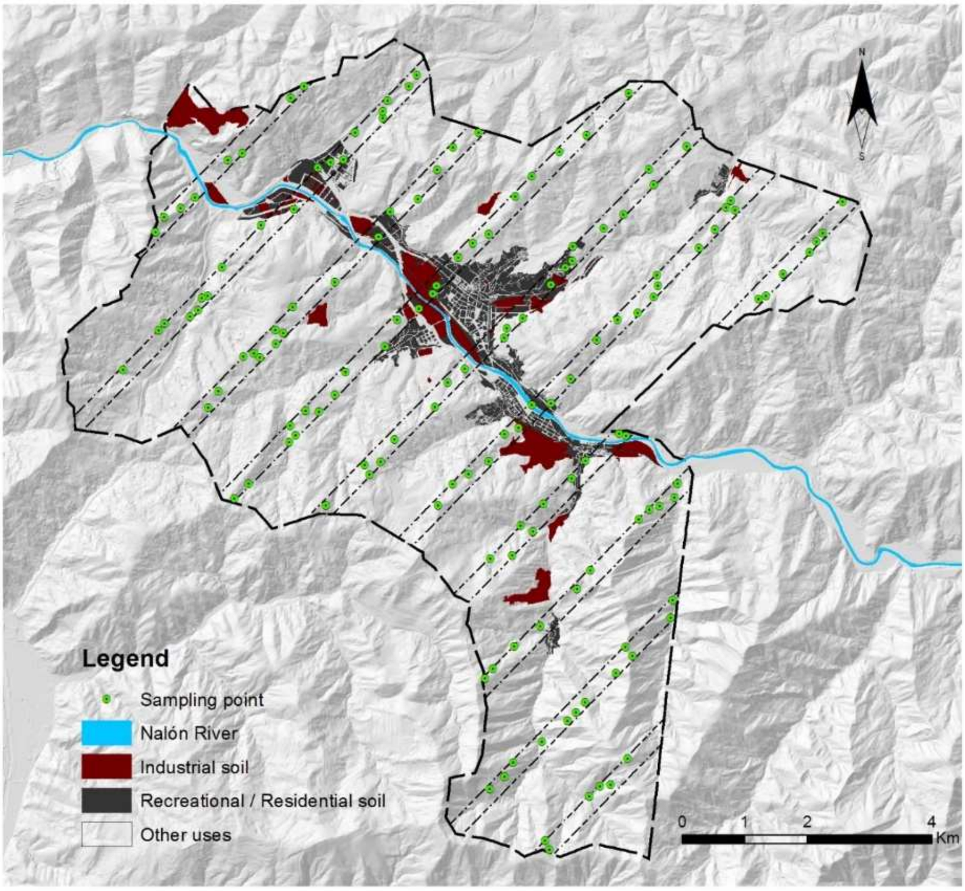


Fig. 2. Sampling design and land use categories in the study area.

2.3. Data transformation – compositional data and the closure problem

In geochemistry, compositional data is obtained by transforming each original raw concentration (i.e. mg/kg of an element in a sample) into proportions of a whole whose elements sum one or 100% (Pawlowsky-Glahn and Egozcue, 2006). However, the unfeasibility of analyzing all the elements in a given soil hinders the consideration of proportions. Indeed, this issue has been heavily debated and is referred to by researchers as the closure problem (Filzmoser et al., 2009b). In environmental science studies, it is generally accepted that the elements analyzed make up the entirety of the soil on the condition that a suitable number of such elements is included in the study (Campbell et al., 2009; Reimann et al., 2012). Moreover, other authors work with subcompositions, defined as a subset of components of parts of a composition (Pawlowsky-Glahn and Buccianti, 2011). Subcompositions are feasible when they respect the principles of compositional data (Greenacre and Lewi, 2009), including the subcompositional coherence principle (Aitchison, 1986).

The most frequently used log-ratio transform functions (*alr*; *clr* and *ilr*) have both advantages and disadvantages, which are widely discussed in the literature. The *clr* transformation is the prevailing function in geochemical studies as it uses the geometric mean as normalizer parameter and it was chosen for the purposes of the present study.

The centred log-ratio transformation (*clr*) equation was adapted from (Aitchison, 1986):

$$clr(x) = \ln\left(\frac{C_j}{\sqrt[D]{\prod_{j=1}^{D} C_j}}\right) \tag{1}$$

where C_j is the concentration of pollutant j and D is the number of parts into which the composition is divided (in this case, the number of pollutants considered).

The back-transformation equation is computed as:

$$\overline{clr}(x) = \frac{e^{clr\,(x)}}{\sum_{j=1}^{D} e^{clr\,(x)}} \tag{2}$$

This equation allows representation of the *clr*-transformed data as compositional data (proportions). This means that the sum of all the elements after back-transformation is equal to 1. The *clr* transformation and the calculation of its back-transformation was performed using CoDaPack v2.02.21 software (http://www.compositionaldata.com/codapack.php).

2.4. Spatial modeling

The spatial characterization of PTE distribution was performed with the following two complementary objectives in mind. First, we sought to define spatial clusters of PTE concentration. To accomplish this, the raw dataset was used, allowing us to interpret contamination outbreaks and therefore locate the main sources of PTEs. Second, we aimed to define RE spatial cluster spots. The RE is used to assess the elements proportions evaluation. Thus, rather than simply looking at PTE content enrichment, we sought to develop a new approach to study PTE fate by examining the changes in their proportions throughout the study area, thus allowing us to define trends of dissemination. The compositional dataset was used to tackle this issue, and spatial clusters of RE were computed.

A four-step methodology was adopted as follows:

- Principal Components Analysis (PCA) for reducing dimensionality and for evaluating variable association was performed. PCA is one of the most important multivariate statistical methods and it is widely used for data preprocessing and dimension reduction (raw and compositional data). The aim of PCA is to reduce the dimensionality of data while simultaneously preserving the within variability structure (variance–covariance) (e.g. Zuo et al., 2016). The analysis starts with p random attributes $X_1, X_2, \dots, X_p$, where no assumption of multivariate normality is required. The axes of the constant ellipsoids correspond to the new synthesis variables, the principal components. The XlStat 2013.1.01 software (https://www.xlstat.com/en/) was used for computational purposes.

- Selected attributes were subjected to a structural analysis, and experimental variograms were computed for both raw and compositional data. The variogram is a vector function used to calculate the spatial variation structure of regionalized variables (Matheron, 1971; Journel and Huijbregts, 1978; Gringarten and Deutsch, 2001).

- Spatial prediction through Ordinary Kriging (OK) aiming to predict the values for the variables at any arbitrary spatial location within the study region was performed. The raw dataset was used to infer the concentration and PTE origin, as the compositional dataset was used for dissemination trend detection and local RE evaluation. Of note, geostatistics are a reference approach for the characterization of environmental hazards in contexts in which the information available is scarce. The primary application of geostatistics is to estimate and map environmental attributes in unsampled areas where Kriging is a generic name for a set of generalized least-squares regression algorithms. OK accounts for local fluctuations of the mean by limiting the field of stationary of the mean to the local neighborhood (Goovaerts 1997). For the computation, the Space-Stat Software V. 4.0.18, Biomedwere was used (Albuquerque et al., 2014) (Fig. 6).
- Finally, Local G clustering was performed. This technique allows measurement of the degree of association that results from the concentration of weighted points (or region represented by a weighted point) and all other weighted points included within a radius of distance from the original and defining clusters of high (high-ring) and low (low-ring) significance. For computation, the SpaceStat V. 4.0-.18. software (https://www.biomedware.com/) was used.

3. Results and discussion

3.1. Descriptive statistics

Descriptive statistics for raw and *clr*-transformed data were computed (Table 1). The raw data revealed considerable variability for some elements, which was of particular concern for As, Cd, Cu, Pb, Sb and Zn, whose maximum values surpassed the RBSSLs (BOPA, 2014). The 5% trimmed mean allowed us to conclude that extreme values were concentrated mainly in the upper 2.5% intervals, as the remaining 97.5% can be approximated by the normal distribution. Once the *clr*-transformed data were applied, the associated standard deviation was clearly reduced and the mean, median and 5% trimmed mean tended to be similar. Indeed, the *clr* data showed a normal distribution as a result of diminishing the weight of outliers. This diminished weight enhanced the prediction of data proportions after the back-transformation of *clr* data, and compositional data were obtained.

Table 1. Descriptive statistics for 15 PTEs: Range, Mean, Median, Standard Deviation (SD), and Trimmed Mean (T.Mean 5%) are expressed in mg·kg^{-1}, Relative Standard Deviation (RSD) is expressed in %.

	Raw Data						Clr-Transformed Data				
PTE	Range	Mean	Median	SD	RSD	T.Mean 5%	Mean	Median	SD	RSD	T.Mean 5%
As	6.4 - 91.1	21.8	18.5	10.9	49.8	21.0	21.9	20.9	6.3	28.9	21.7
Ba	11.0 - 1747.1	107.9	66.9	168.7	156.3	90.2	79.2	74.6	16.7	21.1	78.3
Cd	0.02 - 26.9	0.6	0.3	2.2	382.6	0.4	0.4	0.3	0.1	19.2	0.3
Co	1.1 - 34.0	10.0	9.8	5.0	49.8	9.9	9.4	10.2	11.4	121.7	9.4
Cr (III)	5.7 - 69.0	18.9	18.6	6.7	35.6	18.5	19.6	20.1	4.8	24.5	19.6
Cu	3.0 - 2022.2	39.0	22.7	163.6	419.2	24.6	24.4	24.2	7.3	29.7	24.1
Hg	0.1 - 2.6	0.4	0.3	0.4	95.5	0.4	0.3	0.3	0.1	21.3	0.3
Mo	0.4 - 4.6	1.0	0.9	0.6	53.6	1.0	1.0	1.0	0.2	16.0	1.0
Ni	1.4 - 52.8	18.3	16.5	9.1	49.7	18.0	17.5	17.5	7.2	41.1	17.5
Pb	10.5 - 3729.5	91.6	52.2	302.7	330.6	64.0	62.8	60.7	11.1	17.7	61.8
Sb	0.3 - 256.6	2.5	0.6	20.8	821.8	0.8	0.8	0.7	0.2	26.6	0.8
Se	0.1 - 1.9	0.9	0.8	0.4	45.1	0.8	0.8	0.9	0.3	30.3	0.8
Tl	0.0 - 0.5	0.2	0.2	0.1	33.8	0.2	0.2	0.2	0.0	10.6	0.2
V	7.0-56.0	27.9	27.0	6.9	24.8	27.8	29.6	29.8	6.3	21.2	29.8
Zn	16.9-2161.0	136.2	107.2	179.4	131.7	120.8	119.8	120.8	11.8	9.9	120.1

On the basis of comparison of the histograms (Fig. 3) of the raw and compositional datasets, it is possible to reason that: a) when considering the raw dataset, asymmetric distributions are found for almost all the PTEs, and these distributions are biased mainly by the presence of outliers; b) the *clr*-transformed dataset shows an important feature as it allows the assumption of normality. Therefore, we conclude that the *clr*-transformed dataset and the compositional dataset (after *clr* back-transform) have two principal advantages, namely they allow work with proportions and also improved data normalization.

Of note were the anomalous As, Cd, Cu, Pb, Sb and Zn concentrations, which greatly exceeded the RBSSLs (BOPA, 2014) (Table 1). These elements are classic fingerprints of heavy industrial activity. However, the presence of Ba, Co, Cr, Hg, Mo, Ni, Se, Tl and V did not constitute an immediate risk for human health or the environment.

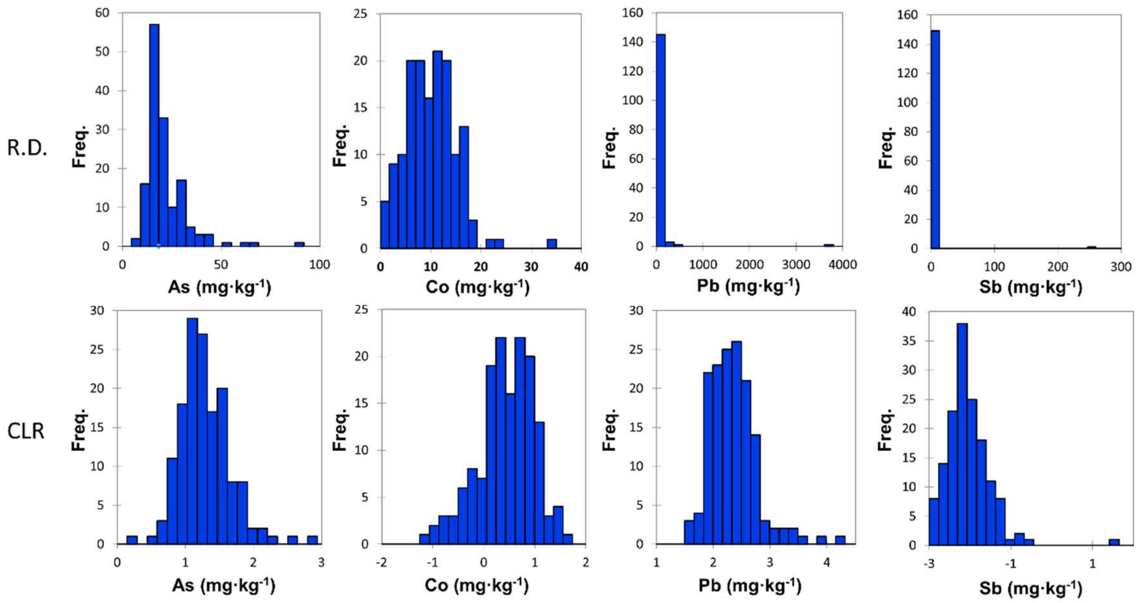


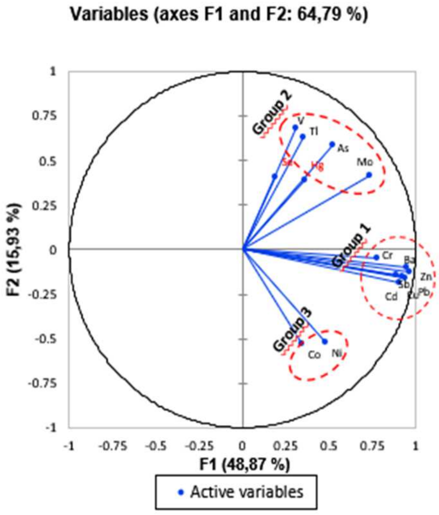
Fig. 3. As, Co, Pb and Sb histograms for raw data (R.D.) and *clr*-transformed data (CLR).

3.2. Multivariate statistics – Principal Components Analysis

When running the raw dataset, PCA results revealed three groups (Fig. 4 a)): a) the first formed by Ba, Cd, Cr, Cu, Pb, Sb and Zn—a typical association of heavy metals; b) the second composed by As, Mo, Tl and V; and c) the third representing Co and Ni. Finally, Hg and Se showed independent behaviors, thereby possibly indicating different sources. On the other hand, when considering the compositional dataset, slight differences in the results were observed (Fig. 4 b). The first-mentioned group (Ba, Cd, Cr, Cu, Pb, Sb and Zn) was split in two: a) the first comprising Cd and Zn; b) the second Cu and Sb. Furthermore, two more groups were identified, c) the third comprising As, V, Tl Mo, Se and Cr; and d) the fourth Ni and Co. Mercury (Hg) and Pb were found to be independent. On the basis of the PCAs, we conclude that the compositional dataset provides a fuller recognition of relevant contaminant associations. When setting a dependence on weight between elements, those which increase or decrease proportionally tend to be associated.

a)

	F1	F2	F3
Eigenvalue	7.330	2.389	1.572
Variability (%)	48.865	15.928	10.477
Cumulative %	48.865	64.793	75.271



b)

	F1	F2	F3	F4	F5
Eigenvalue	4.402	2.825	1.733	1.279	0.939
Variability (%)	29.349	18.832	11.553	8.526	6.260
Cumulative %	29.349	48.181	59.734	68.259	74.519

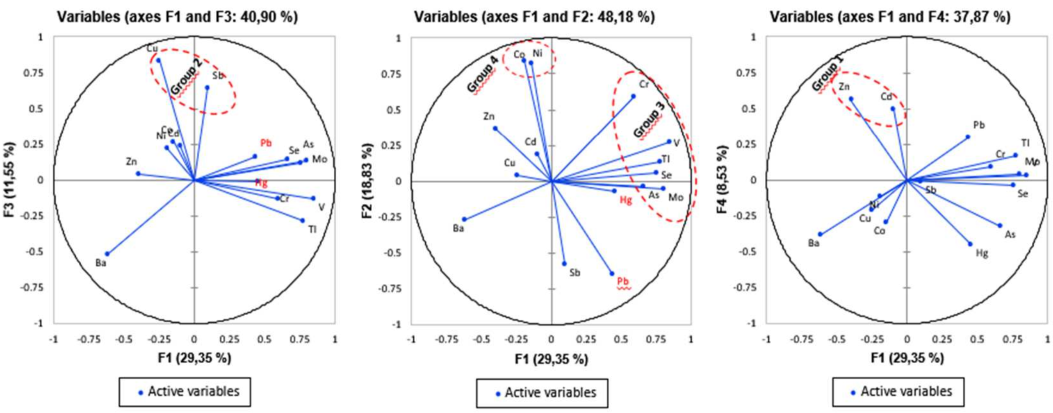


Fig. 4. a) PCA - Raw dataset; b) PCA - Compositional data.

3.3. Spatial modeling – geostatistical approach

At this point, As, Cu, Hg, Pb, and Zn were chosen for spatial modeling purposes as they are core PTEs in contamination forecasts and also representative of the most important groups identified (Fig. 4).

The spatial stochastic patterns of the five PTEs were constructed following a three-step geostatistical modeling method.

3.3.1 Structural analysis and experimental variograms

The selected variables were subjected to a structural analysis, and experimental variograms were computed. The variogram is a vector function used to calculate the spatial variability of regionalized variables defined by the following equation (Matheron, 1971; Journel and Huijbregts, 1978):

$$\gamma(h) = \frac{1}{2N(h)} \sum_{2N(h)}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \qquad (3)$$

Its argument is *h* (distance), where $Z(x_i)$ and $Z(x_i+h)$ are the numerical values of the observed variable at points x_i, and x_i+h. The number of pairs forming for a *h* distance is *N(h)*. Thus, it is the median value of the square of the differences between all pairs of points in the geometric field spaced at a *h* distance. The graphic study of the variograms obtained provides an overview of the spatial structure of the variable. One of the parameters that provide such information is the nugget effect (C_o), which shows the behavior at the origin. The other two parameters are the sill (C_1) and the amplitude (*a*) which define the inertia used in the interpolation process and the influence radius of the variable, respectively (Table 2).

The experimental variograms $\gamma_{(h)}$ were then fitted to a theoretical model, $\hat{\gamma}_{(h)}$ (Isaaks and Srivastava 1989). The adjusted parameters for the five PTEs of the theoretical variograms (raw and compositional datasets) (Fig. 5) allowed us to observe that the isotropic variograms obtained generally showed a better fit for the compositional dataset. Indeed, the attributes showed a nugget effect below 40% of the total variance of all the attributes (Table 2). The error associated with the interpolation procedure, OK, is therefore minimized when using the compositional dataset.

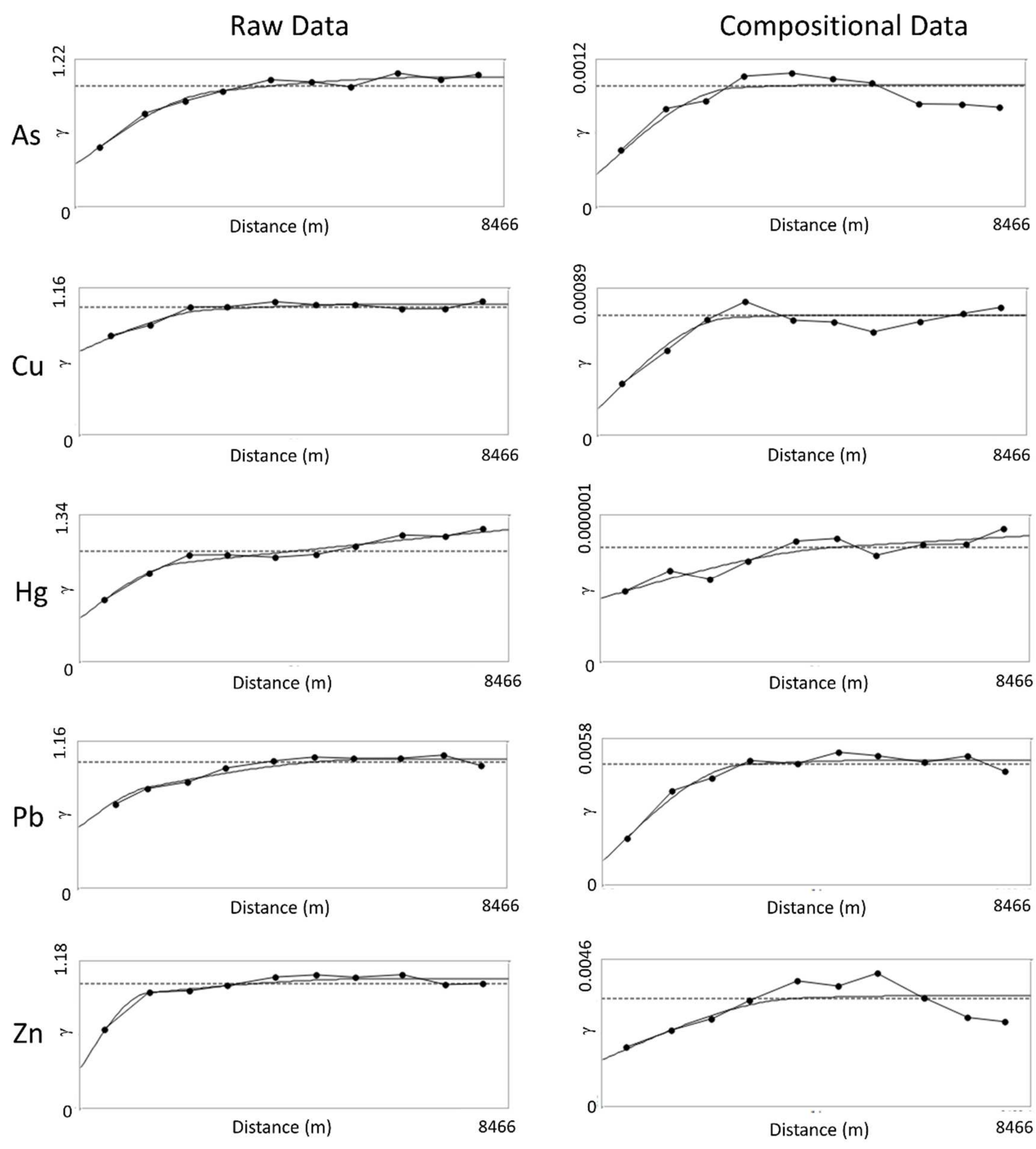


Fig. 5. Isotropic experimental variograms and fitted models for the raw and compositional datasets.

Table 2. Experimental variogram parameters for the raw and compositional datasets: *a* (m) is the amplitude; C_o represents the value of the nugget effect; C_1 and C_2, the value of the sill of the first and the second spherical structure respectively, and C_0(%Var) and C_1+C_2 (%Var) the mutual variances weighing for nugget and sill respectively.

	Parameters	As	Cu	Hg	Pb	Zn
Raw Data	A	2738	2575	1997	1376	1327
	C_o	0.356	0.664	0.401	0.488	0.330
	C_1	0.465	0.256	0.411	0.201	0.544
	C_2	0.260	0.110	1.17	0.339	0.172
	C_o (%Var)	33	64	20	47	32
	C_1+C_2 (%Var)	67	36	80	53	68
Comp. Data	A	2700	2569	4758	2808	3903
	C_o	$2.77 \cdot 10^{-4}$	$1.63 \cdot 10^{-4}$	$5.93 \cdot 10^{-7}$	$9.90 \cdot 10^{-4}$	$1.47 \cdot 10^{-3}$
	C_1	$6.45 \cdot 10^{-4}$	$4.83 \cdot 10^{-4}$	$3.50 \cdot 10^{-7}$	$3.52 \cdot 10^{-3}$	$1.58 \cdot 10^{-3}$
	C_2	$1.14 \cdot 10^{-4}$	$8.11 \cdot 10^{-5}$	$6.90 \cdot 10^{-7}$	$5.35 \cdot 10^{-4}$	$4.18 \cdot 10^{-4}$
	C_o (%Var)	27	22	36	20	42
	C_1+C_2 (%Var)	73	78	64	80	58

3.3.2 Spatial prediction: Ordinary kriging

Analysis of the outputs obtained (Fig. 6) revealed evident contrasts between the raw and the compositional dataset representations. In reality, care must be taken when interpreting representations as they reflect distinct data. In this regard, the raw dataset mapping shows the estimated picture of PTE concentration distribution, thus indicating possible sources of these contaminants. In contrast, the compositional dataset mapping shows the spatial variability of PTE proportion, thus reflecting PTE RE and providing crucial information about the fate of these compounds within the study area. To facilitate understanding of the results, the study area was divided into various zones of interest (Fig. 6) and interpreted as follows:

a) Considering the maps of the raw data set (Fig. 6 -R.D.), OK revealed high concentrations for all PTEs (Zn, Hg, As, Pb and Cu) in the central zone (zone *A*), which coincides with the city of Langreo (Fig. 6). Moreover, Cu and Zn showed notable presence in the southern area (zone *B*), where the mining industry (coal mines and processing) were located (Fig. 1). The Cu map shows a north-eastern red-colored site (zone *C*) coinciding with a former coal-mining area. On the other hand, high concentrations of Hg and As were observed in the western (zone *D*) and northern (zone *E*) areas, which may be explained by the proximity to a derelict Hg mine (El Terronal site) whose impact has been widely discussed (e.g. Gallego et al., 2015, González-Fernández et al., 2018);

b) Concerning the compositional dataset (Fig. 6-C.D.), RE in Cu, Pb and Zn was identified towards south (zone *F*) and northeast (zone *C*) of the area (Fig. 6), where the corresponding distribution was at its lowest level when using the raw data. Cu, Pb and Zn showed a significant distribution throughout the area and therefore marked RE.

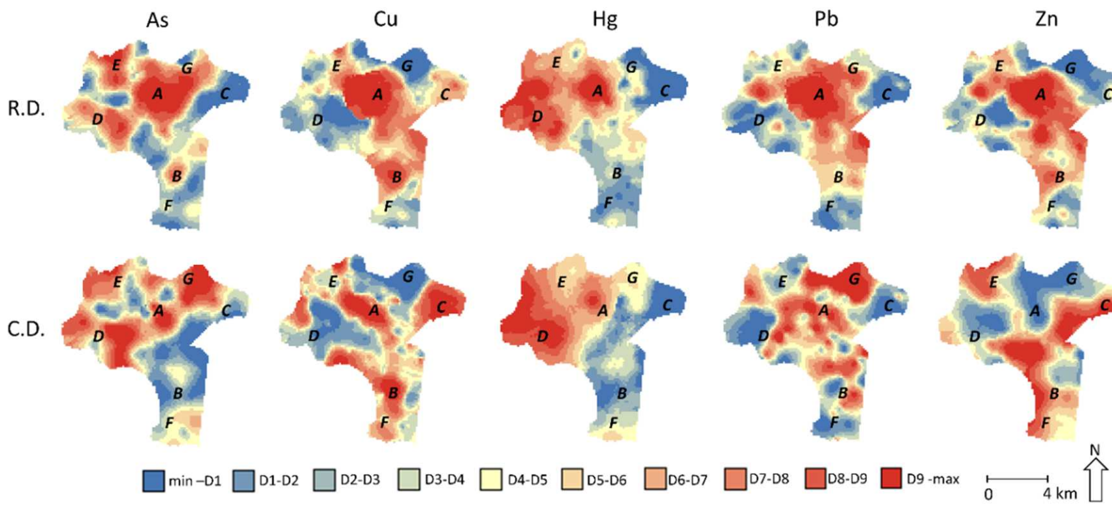


Fig. 6. Ordinary kriging results. Raw data (R.D) and compositional data (CD) respectively. Scale is expressed in deciles (D_i) of mg·kg^{-1} (R.D). and of % (C.D).

3.3.3 Spatial prediction: Local G clustering

To reinforce the findings of the previous section, a Local G clustering was conducted to assess the level of association resulting from the concentration of weighted points (or region represented by a weighted point) and all other weighted points included within a radius from the original point. In this regard, a given zone was subdivided into n regions, I =1, 2,…, n, where each neighborhood is distinguished with a point whose Cartesian coordinates are known. Each *i* has a value *x* (a weight) taken from a variable *X* associated with it. The variable holds a natural origin and it is positive. The *G(i)* statistic developed below allows the testing of hypotheses concerning the spatial concentration of the sum of *x* values associated with the *j* points within *d* of the i^{th} point. The following statistic is obtained:

$$G_i(d) = \frac{\sum_{j=1}^{n} W_{ij}(d)X_i}{\sum_{j}^{n} x_j} \quad (4)$$

where W_{ij} is a symmetric one/zero spatial weight matrix with a value of 1 for all links defined as being within distance *d* of a given *i*; all other links are zero, including the link of point *i* to itself. The numerator is the sum of all x_j within *d* of *i* but not including x_i. The denominator is the sum of all x_j, excluding x_i (Getis and Ord, 1992).

The maps obtained (Fig. 7) provide a faster and more intuitive way to verify whether the problematic zones detected previously are indeed of concern. Thus, red areas (high ring) show the sites with the greatest accumulation of the PTEs, while the blue areas (low ring) represent zones with low accumulation (Fig. 7). The highest accumulation of PTEs, when considering the raw data clusters, was in the city center (high ring-zone A). The soils in this area were clearly affected by PTE deposition, presumably due to heavy industry and/or the transport of pollutants. However, examination of the significance of the spatial clusters obtained using the compositional data shows several differences. The central high ring (high significance) is now smaller, showing that the areas with the highest concentration of these PTEs (Zn, Hg, As, Pb and Cu) do not totally overlap with the corresponding higher proportions and indicating that PTE transport and RE occurs in a westerly and southerly direction.

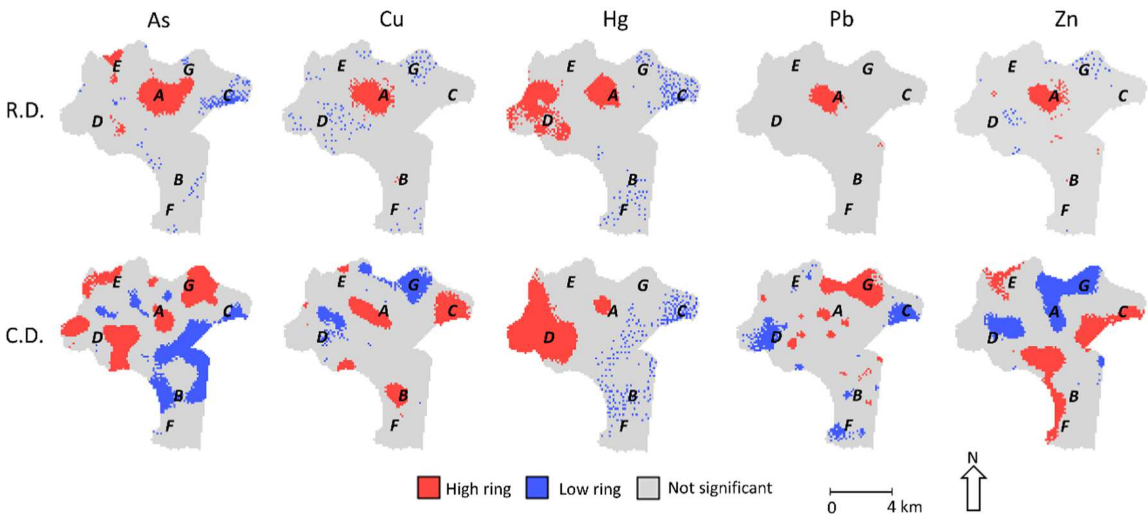


Fig. 7. Local G clusters. Raw data (R.D.) and compositional data (C.D.) respectively.

4. Conclusions

The degree of PTE contamination in the soil of an industrial area can be characterized using two datasets, namely raw and compositional (*clr*-transformed followed by the back-transformation function). To exemplify the complementary attributes of these two types of dataset, 150 soil samples were collected and 36 elements were analyzed in the area of Langreo (80 km^2), a paradigmatic example of an industrial area affected by heavy metal and metalloid contamination. Univariate statistics allowed recognition of redundant information and the identification of outliers. The space of analysis was then reduced for both datasets by building the synthesis variables held by PCA. Five PTEs, namely Zn, Hg, As, Pb and Cu, were retained for spatial modeling due to their significance in the contamination forecast. OK and Local G clustering allowed the construction of hazard maps, which facilitate the evaluation of probable origin of PTEs (raw data) and their possible RE (compositional data).

Regarding the Langreo area, it is extensively affected by its industrial and mining history. The following observations support this conclusion: 1. The city centre is highly enriched in PTEs, which can be explained by heavy industry and pollutant transport, Pb being the main contaminant; 2. The spatial distribution of Cu indicates a strong association with coal mining and processing; and 3. Hg and As show enrichment in a northwesterly direction, which is linked to natural mineralization and former Hg mining and metallurgy. Future work would require an exhaustive study of covariates to shed light on PTE dynamics and to clarify the main sources of PTEs, as well as their RE throughout the study area.

The information gathered provides a basis for delimiting the polluted zones and the sources of pollutants, thus facilitating the development of specific air and soil monitoring activities, urban planning and environmental policies.

Acknowledgements

C. Boente obtained a grant from the “Formación del Profesorado Universitario” program, financed by the “Ministerio de Educación, Cultura y Deporte de España”.

M.T.D Albuquerque acknowledges a scholarship 567 SFRH/BSAB/127907/2016 from the Foundation for Science and Technology (Portugal).

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. J. R. Stat. Soc. 44, 139–177. doi:10.1007/978-94-009-4109-0.

Albuquerque, M.T.D., Gerassis, S., Sierra, C., Taboada, J., Martín, J.E., Antunes, I.M.H.R., Gallego, J.R., 2017. Developing a new Bayesian Risk Index for risk evaluation of soil contamination. Sci. Total Environ. 603–604, 167–177.

doi:10.1016/j.scitotenv.2017.06.068.

Alloway, B.J., 1990. The origins of heavy metals in soils, in: Heavy Metals in Soils. p. 339.

Antunes, I.M.H.R., Albuquerque, M.T.D., 2013. Using indicator kriging for the evaluation of arsenic potential contamination in an abandoned mining area (Portugal). Sci. Total Environ. 442, 545–552. doi:10.1016/j.scitotenv.2012.10.010.

Biasioli, M., Barberis, R., Ajmone-Marsan, F., 2006. The influence of a large city on some soil properties and metals content. Sci. Total Environ. 356, 154–164. doi:10.1016/j.scitotenv.2005.04.033

Bishop, T.F.A., McBratney, A.B., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103, 149–160. doi:10.1016/S0016-7061(01)00074-X.

Boente, C., Matanzas, N., García-González, N., Rodríguez-Valdés, E., Gallego, J.R., 2017. Trace elements of concern affecting urban agriculture in industrialized areas: A multivariate approach. Chemosphere 183, 546–556. doi:10.1016/j.chemosphere.2017.05.129.

Boente, C., Sierra, C., Rodríguez-Valdés, E., Menéndez-Aguado, J.M., Gallego, J.R., 2016. Soil washing optimization by means of attributive analysis: Case study for the removal of potentially toxic elements from soil contaminated with pyrite ash. J. Clean. Prod. doi:10.1016/j.jclepro.2016.11.007.

BOPA, Boletín Oficial del Principado de Asturias, 91, April 21, 2014. Generic Reference Levels for Heavy Metals in Soils from the Principality of Asturias, Spain, https://sede.asturias.es/portal/site/Asturias/menuitem.1003733838db7342ebc4e191100000f7/?vgnextoid=d7d79d16b61ee010VgnVCM1000000100007fRCRD&fecha=21/04/2014&refArticulo=2014-06617&i18n.http.lang=es (Accessed December 2017).

Buccianti, A., Grunsky, E., 2014. Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes? J. Geochemical Explor. 141, 1–5. doi:10.1016/j.gexplo.2014.03.022.

Cachada, A., Dias, A.C., Pato, P., Mieiro, C., Rocha-Santos, T., Pereira, M.E., Da Silva, E.F., Duarte, A.C., 2013. Major inputs and mobility of potentially toxic elements contamination in urban areas. Environ. Monit. Assess. 185, 279–294. doi:10.1007/s10661-012-2553-9.

Campbell, G.P., Curran, J.M., Miskelly, G.M., Coulson, S., Yaxley, G.M., Grunsky, E.C., Cox, S.C., 2009. Compositional data analysis for elemental data in forensic science. Forensic Sci. Int. 188, 81–90. doi:10.1016/j.forsciint.2009.03.018.

Candeias, C., da Silva, E.F., Salgueiro, A.R., Pereira, H.G., Reis, A.P., Patinha, C., Matos, J.X., Ávila, P.H., 2011. The use of multivariate statistical analysis of geochemical data for assessing the spatial distribution of soil contamination by potentially toxic elements in the Aljustrel mining area (Iberian Pyrite Belt, Portugal). Environ. Earth Sci. 62, 1461–1479. doi:10.1007/s12665-010-0631-2.

Dalla Libera, N., Fabbri, P., Mason, L., Piccinini, L., Pola, M., 2017. Geostatistics as a tool to improve the natural background level definition: An application in groundwater. Sci. Total Environ. 598, 330–340. doi:10.1016/j.scitotenv.2017.04.018.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric Logratio Transformations for Compositional Data Analysis. Math. Geol. 35, 279–300. doi:10.1023/A:1023818214614.

Fayiga, A.O., Saha, U.K., 2016. Soil pollution at outdoor shooting ranges: Health effects, bioavailability and best management practices. Environ. Pollut. 216, 135–145. doi:10.1016/j.envpol.2016.05.062.

Filzmoser, P., Hron, K., Reimann, C., 2009a. Principal component analysis for compositional data with outliers, in: Environmetrics. pp. 621–632. doi:10.1002/env.966.

Filzmoser, P., Hron, K., Reimann, C., 2009b. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. Sci. Total Environ. 407, 6100–6108. doi:10.1016/j.scitotenv.2009.08.008.

Gallego, J.R., Esquinas, N., Rodríguez-Valdés, E., Menéndez-Aguado, J.M., Sierra, C., 2015. Comprehensive waste characterization and organic pollution co-occurrence in a Hg and As mining and metallurgy brownfield. J. Hazard. Mater. 300, 561–571. doi:10.1016/j.jhazmat.2015.07.029.

Gallego, J.R., Rodríguez-Valdés, E., Esquinas, N., Fernández-Braña, A., Afif, E., 2016. Insights into a 20-ha multi-contaminated brownfield megasite: An environmental forensics approach. Sci. Total Environ. 563–564, 683–692. doi:10.1016/j.scitotenv.2015.09.153.

Getis, A., Ord, J.K., 1992. The Analysis of Spatial Association by Use of Distance Statistics. Geogr. Anal. 24, 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x.

González-Fernández, B., Rodríguez-Valdés, E., Boente, C., Menéndez-Casares, E., Fernández-Braña, A., Gallego, J.R., 2018. Long-term ongoing impact of arsenic contamination on the environmental compartments of a former mining-metallurgy area. Sci. Total Environ. 610–611, 820–830. doi:10.1016/j.scitotenv.2017.08.135.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. University Press, New York: Oxford.

Goovaerts, P., 1999. Geostatistics in soil science: State-of-the-art and perspectives. Geoderma 89, 1–45. doi:10.1016/S0016-7061(98)00078-0.

Greenacre, M., Lewi, P., 2009. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. J. Classif. 26, 29–54. doi:10.1007/s00357-009-9027-y.

Gringarten, E., Deutsch, C. V, 2001. Teacher's Aide Variogram Interpretation and Modeling. Math. Geol. 33, 507–534. doi:10.1023/a:1011093014141.

Guagliardi, I., Cicchella, D., De Rosa, R., 2012. A geostatistical approach to assess concentration and spatial distribution of heavy metals in urban soils. Water. Air. Soil Pollut. 223, 5983–5998. doi:10.1007/s11270-012-1333-z.

Isaaks, E.H., Srivastava, R.M., 1989. An introduction to applied geostatistics. University Press, New York: Oxford. pp. 278–322.

Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press, San Diego.

Lark, R.M., Bishop, T.F., 2007. Cokriging particle size fractions of the soil. Eur. J. Soil Sci. 58, 763–774. doi:10.1111/j.1365-2389.2006.00866.x.

Li, Z., Ma, Z., van der Kuijp, T.J., Yuan, Z., Huang, L., 2014. A review of soil heavy metal pollution from mines in China: Pollution and health risk assessment. Sci. Total Environ. doi:10.1016/j.scitotenv.2013.08.090.

Lu, A., Wang, J., Qin, X., Wang, K., Han, P., Zhang, S., 2012. Multivariate and geostatistical analyses of the spatial distribution and origin of heavy metals in the agricultural soils in Shunyi, Beijing, China. Sci. Total Environ. 425, 66–74. doi:10.1016/j.scitotenv.2012.03.003.

Martínez, J., Saavedra, Á., García-Nieto, P.J., Piñeiro, J.I., Iglesias, C., Taboada, J., Sancho, J., Pastor, J., 2014. Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). Appl. Math. Comput. 241, 1–10. doi:10.1016/j.amc.2014.05.004.

Mateu-Figueras, G., Pawlowsky-Glahn, V., 2008. A critical approach to probability laws in geochemistry, in: Progress in Geomathematics. pp. 39–52. doi:10.1007/978-3-540-69496-0_4.

Matheron, G., 1963. Principles of geostatistics. Econ. Geol. 58, 1246–1266. doi:10.2113/gsecongeo.58.8.1246.

McIlwaine, R., Doherty, R., Cox, S.F., Cave, M., 2016. The relationship between historical development and potentially toxic element concentrations in urban soils. Environ. Pollut. 220, 1036–1049. doi:10.1016/j.envpol.2016.11.040.

McKinley, J.M., Hron, K., Grunsky, E.C., Reimann, C., de Caritat, P., Filzmoser, P., van den Boogaart, K.G., Tolosana-Delgado, R., 2016. The single component geochemical map: Fact or fiction? J. Geochemical Explor. 162, 16–28. doi:10.1016/j.gexplo.2015.12.005.

Megido, L., Suárez-Peña, B., Negral, L., Castrillón, L., Fernández-Nava, Y., 2017. Suburban air quality: Human health hazard assessment of potentially toxic elements in PM10. Chemosphere 177, 284–291. doi:10.1016/j.chemosphere.2017.03.009.

Moen, J., Ale, B.J.M., 1998. Risk maps and communication. J. Hazard. Mater. 61, 271–278.

Odeh, I.O.A., Todd, A.J., Triantafilis, J., 2003. Spatial prediction of soil particle size fractions as compositional data. Soil Sci. 168, 501–515.

Pawlowsky, V., 1989. Cokriging of regionalized compositions. Math. Geol. 21, 513–521. doi:10.1007/BF00894666.

Pawlowsky, V., Burger, H., 1992. Spatial Structure-analysis of Regionalized Compositions. Math. Geol. 24, 675–691. doi:10.1007/BF00894233.

Pawlowsky, V., Olea, R.A., 2004. Geostatistical Analysis of Compositional Data, Chapman & Hall, Ltd., London, UK.

Pawlowsky, V., Olea, R.A., Davis, J., 1995. Estimation of regionalize compositions: A comparison of three methods. Math. Geol. 27, 105-127.

Pawlowsky-Glahn, V., Egozcue, J.J., 2006. Compositional data and their analysis: an introduction. Geol. Soc. London, Spec. Publ. 264, 1–10. doi:10.1144/GSL.SP.2006.264.01.01.

Pawlowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. Stoch. Environ. Res. Risk Assess. 15, 384–398. doi:10.1007/s004770100077.

Pawlowsky-Glahn, V., Buccianti, A., 2011. Compositional Data Analysis: Theory and Applications. Wiley.

Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., Albanese, S., Andersson, M., Arnoldussen, A., Baritz, R., Batista, M.J., Bel-lan, A., Cicchella, D., De Vivo, B., De Vos, W., Duris, M., Dusza-Dobek, A., Eggen, O.A., Eklund, M., Ernstsen, V., Finne, T.E., Flight, D., Forrester, S., Fuchs, M., Fugedi, U., Gilucis, A., Gosar, M., Gregorauskiene, V., Gulan, A., Halamic, J., Haslinger, E., Hayoz, P., Hobiger, G., Hoffmann, R., Hoogewerff, J., Hrvatovic, H., Husnjak, S., Janik, L., Johnson, C.C., Jordan, G., Kirby, J., Kivisilla, J., Klos, V., Krone, F., Kwecko, P., Kuti, L., Lima, A., Locutura, J., Lucivjansky, P., Mackovych, D., Malyuk, B.I., Maquil, R., McLaughlin, M.J., Meuli, R.G., Miosic, N., Mol, G., Négrel, P., O'Connor, P., Oorts, K., Ottesen, R.T., Pasieczna, A., Petersell, V., Pfleiderer, S., Ponavic, M., Prazeres, C., Rauch, U., Salpeteur, Schedl, A., Scheib, A., Schoeters, I., Sefcik, P., Sellersjö, E., Skopljak, F., Slaninka, I., Šorša, A., Srvkota, R., Stafilov, T., Tarvainen, T., Trendavilov, V., Valera, P., Verougstraete, V., Vidojevic, D., Zissimos, A.M., Zomeni, Z., 2012. The concept of compositional data analysis in practice - Total major element concentrations in agricultural and grazing land soils of Europe. Sci. Total Environ. 426, 196–210. doi:10.1016/j.scitotenv.2012.02.032.

Tepanosyan, G., Maghakyan, N., Sahakyan, L., Saghatelyan, A., 2017. Heavy metals pollution levels and children health risk assessment of Yerevan kindergartens soils. Ecotoxicol. Environ. Saf. 142, 257–265. doi:10.1016/j.ecoenv.2017.04.013.

Venkatramanan, S., Chung, S.Y., Kim, T.H., Kim, B.W., Selvam, S., 2016. Geostatistical techniques to evaluate groundwater contamination and its sources in Miryang City, Korea. Environ. Earth Sci. 75. doi:10.1007/s12665-016-5813-0.

Zuo, R., Carranza, E.J.M., Wang, J., 2016. Spatial analysis and visualization of exploration geochemical data. Earth-Science Rev. 158, 9–18. doi:10.1016/j.earscirev.2016.04.006.