Advances in self-organizing maps for their application to compositional data

Josep A. Martín-Fernández[1] · Mark A. Engle[2] · Leslie F. Ruppert[3] · Ricardo A. Olea[4]

Abstract

A self-organizing map (SOM) is a non-linear projection of a D-dimensional data set, where the distance among observations is approximately preserved on to a lower dimensional space. The SOM arranges multivariate data based on their similarity to each other by allowing pattern recognition leading to easier interpretation of higher dimensional data. The SOM algorithm allows for selection of different map topologies, distances and parameters, which determine how the data will be organized on the map. In the particular case of compositional data (such as elemental, mineralogical, or maceral abundance), the sample space is governed by Aitchison geometry and extra steps are required prior to their SOM analysis. Following the principle of working on log-ratio coordinates, the simplicial operations and the Aitchison distance, which are appropriate elements for the SOM, are presented. With this structure developed, a SOM using Aitchison geometry is applied to properly interpret elemental data from combustion products (bottom ash, fly ash, and economizer fly ash) in a Wyoming coal-fired power plant. Results from this effort

[1] Dept. of Computer Science, Applied Mathematics and Statistics, University of Girona, Spain; ORCID: 0000-0003-2366-1592; email: josepantoni.martin@udg.edu

[2] U.S. Geological Survey, 12201 Sunrise Valley Drive, Mail Stop 956, Reston, VA 20192, USA and Dept. of Geological Sciences, University of Texas at El Paso, El Paso, TX, USA; ORCID: 0000-0001-5258-7374; email: engle@usgs.gov

[3] U.S. Geological Survey, 12201 Sunrise Valley Drive, Mail Stop 956, Reston, VA 20192, USA; ORCID: 0000-0002-7453-1061; email: lruppert@usgs.gov

[4] U.S. Geological Survey, 12201 Sunrise Valley Drive, Mail Stop 956, Reston, VA 20192, USA; ORCID: 0000-0003-4308-0808; email: rolea@usgs.gov

provide knowledge about the differences between the ash composition in the coal combustion process.

Keywords

Aitchison distance · Coal Combustion Products · Isometric logratio · Proportions · Simplex

1 Introduction

Multivariate data visualization techniques have become very important given the increasing data dimensionality in scientific studies, such as concentration data for 40+ elements routinely provided during geochemical analysis and characterization. The self-organizing map (SOM), also known as Kohonen neural network or Kohenen map (Kohonen 2001), is a type of neural network that is used as a tool to reduce data set dimensions to be able to more easily interpret the relationships. The principal aim of SOM is to project a *D*-dimensional data set into a one or two-dimensional discrete map, and to perform this non-linear projection adaptively in a *topologically* consistent way. In a SOM, the proximity among samples in the *input* space (data set) is approximately preserved in the *output* space (map) but in a lower dimension that can be more easily interpreted (Kohonen 2001). It is possible to create maps using different topologies (rectangular or hexagonal), different distances or similarity measures (Euclidean distance, Aitchison distance, Manhattan distance, Tanimoto similarity, etc.) and different parameters, such as the *learning rate* and *neighborhood size*, which determine how the units will organize themselves in the output space. Although similar to cluster analysis techniques, a SOM allows for the exploration of different distances, we follow Everitt et al. (2011, p. 69) recommendation "…the choice of measure will be guided largely by the type of variables being used and the intuition of the investigator". In the special case of compositional data (CoDa), data in which the variables consist of numeric measurements that are relative to one another (e.g., concentration, relative abundance, or distribution information) the *Aitchison distance* has been proven to be an appropriate measure for the geometry of the sample space. Palarea-Albaladejo et al. (2012) present a summary of the properties, advantages and difficulties of a number of different measures when they are used for CoDa.

CoDa conveys relative information expressed in the ratios between parts. These data are common in environmental and geochemical studies when the constituents and compounds are described in terms of their concentration in air (Jarauta-Bragulat et al. 2016), water (Olea et al. 2018), or in terms of solids and other wastes (Edjabou et al. 2017). When one decides to analyze a data set **X** ($n \times D$; rows×columns) using compositional methods, such as concentration data in coal combustion products, one is assuming that any observation **x** (a row of **X**) is a member of an equivalence class (Barceló-Vidal and Martín-Fernández 2016). That is, the information contained in **x** is the same as in $k \cdot \mathbf{x}$ for any real scalar $k>0$, a property known as scale invariance (Aitchison 2003). Importantly, CoDa occupy a quotient space (Barceló-Vidal and Martín-Fernández 2016). A representative of the quotient space is the D-part unit simplex $\mathrm{S}^D=\{\mathbf{x}\in \mathbb{R}^D / x_\mathrm{i} >0;\ \sum x_\mathrm{i} =K;\ \mathrm{i}=1, \ldots, D\}$ which is governed by Aitchison geometry (Pawlowsky-Glahn et al. 2015). In practice, for convenience, compositions are commonly expressed in terms of proportions (K=1), percentages (K=100) or parts per million (ppm; $K=10^6$). Following Barceló-Vidal and Martín-Fernández (2016), a logarithmic isomorphism between the quotient spaces S^D and $\mathrm{Z}^D=\{\mathbf{z}\in\mathbb{R}^D/\sum z_\mathrm{i} =0\}$ can be defined and a composition **x** can be expressed in terms of the vector $\mathbf{z}=\left(\ln\left(\frac{x_1}{\mathrm{g}(\mathbf{x})}\right),\ldots,\ln\left(\frac{x_D}{\mathrm{g}(\mathbf{x})}\right)\right)$, where g(**x**) is the geometrical mean of **x**. The vectors $\mathbf{z}\in\mathrm{Z}^D$, known as the *centered log-ratio* (clr) vectors (Aitchison 1986), are in a hyperplane of dimension D-1. The inner product, distance and norm in S^D can be defined via the clr variables (Barceló-Vidal and Martín-Fernández 2016). These metric elements are used to construct orthonormal log-ratio bases in S^D and a composition **x** can be expressed in terms of its corresponding *isometric log-ratio* (ilr) coordinates $\mathbf{y}=\mathrm{ilr}(\mathbf{x})=(y_1, \ldots, y_{D-1})$ (Pawlowsky-Glahn et al. 2015), where, for example

$$y_\mathrm{j}=\sqrt{\frac{D\text{-j}}{D\text{-j}+1}}\ \ln\left(\frac{x_\mathrm{j}}{\sqrt[D\text{-j}]{\prod_{\mathrm{k}=\mathrm{j}+1}^{D} x_\mathrm{k}}}\right),\ \mathrm{j}=1,\ldots,\ D\text{-}1. \qquad (1)$$

Note that ratios and logratios cannot be computed when one of the parts is zero or missing. Methods to deal with the zero problem have been described in numerous papers, and readers will find a general description in Palarea-Albaladejo and Martín-Fernández (2015). It is generally agreed upon that a statistical analysis of CoDa has to be performed on coordinates with respect to a log-ratio basis (Mateu-Figueras et al. 2011). In particular the Aitchison distance d_a between two compositions $\mathbf{x}_1$ and $\mathbf{x}_2$ can be calculated as the Euclidean distance d_e between their corresponding vectors of ilr-coordinates: $d_a(\mathbf{x}_1, \mathbf{x}_2)=d_e(\mathrm{ilr}(\mathbf{x}_1), \mathrm{ilr}(\mathbf{x}_2))$.

The main goal of this paper is to provide the methodological framework to perform SOM analysis to CoDa (CoDa-SOM). The techniques used in this work are introduced in Section 2, which is a refinement of an earlier formulation that was presented in Cortés and Palma (2013). A geochemical example is presented in Section 3, where all the techniques introduced are applied and the results are interpreted. Finally, in Section 4, some concluding remarks are presented. The programming of the techniques discussed in this article was carried out using the *kohonen* package of the open source R statistical programming language and software (Wehrens and Buydens 2007).

2 Self-organizing maps for compositional data

2.1 PCA and MDS versus SOM

In addition to the SOM, there are many other well-known approaches for dimension reduction of data in a Real space, for which principal component analysis (PCA, Jolliffe 2002) and multi-dimensional scaling (MDS, Cox and Cox 2001) are the most popular. Both PCA and MDS map the objects in a continuous output space; in contrast, the SOM uses a regular grid of nodes onto which samples are projected. In many cases, particularly for high-dimensional data sets, PCA requires more than two axes to reduce the dimension with a reasonably quality. Moreover, due its own design PCA might provide misleading interpretations when there are groups in the data (Martín-Fernández et al. 2015). Because of this, MDS and the SOM are more accurate because they start from an *appropriate* distance function for the input space, that is, a distance consistent with the geometry of the sample space of the data. MDS provides a two-dimensional continuous map of points where the Euclidean distance matrix approaches the original distance matrix in some optimal sense so that the distances in the output space can be interpreted as an estimate of the distances in the input space. However, the algorithm for MDS can present difficulties when local optima are present, which can be problematic when the sample size is large (Wehrens and Buydens 2007). In contrast, the SOM is less rigid in terms of its data requirements; unlike PCA and MDS, the SOM does not require data to be available for every parameter for every sample (Dickson and

Giblin 2007). Moreover, the SOM algorithm is very simple and allows for a more flexible exploratory analysis (Kohonen 2001).

2.2 Compositional data analysis formulation for SOM

In this section, some basic concepts of the CoDa-SOM are introduced. Interested readers may consult Kohonen (2001) for an in-depth discussion of basic concepts. The stages of a CoDa-SOM can be summarized in the steps that follow.

Step 0: Map definition in the output space: The user chooses the m number of nodes for the map in the discrete output space, the topology of which is formed by arranging the set of nodes in a grid. Although determining the number of empty nodes can be a trial and error process, Vesanto (2000) recommends $m=5\cdot\sqrt{n}$, where n is the sample size. Despite a map with too many empty nodes may not seem useful, some empty nodes (e.g., nodes serving as a best matching unit for no samples at Step 2) are needed to facilitate cluster interpretation. In addition, there is no need for the dimensions of the map to match; frequently the map is rectangular and in some cases scaled to match the relative variance of the first two components from PCA of the sample set (Akinduko et al. 2016). For a two-dimensional map, the hexagonal topology, where each node has six neighborhoods, is the most popular option. Another common option is a rectangular grid, where each node is directly connected with four other nodes.

Step 1: Initialization of weight vector of nodes: A D-component vector $\mathbf{w}_j$, j=1,..,m must be assigned to each node. The vector $\mathbf{w}_j$, which belongs to the simplex S^D in the case of CoDa, is commonly randomly generated from the input space but it can also be randomly selected from the data set or constructed using the scores from PCA (Akinduko et al. 2016) applied to clr-transformed data.

Step 2: Matching a sample with its best-matching unit (BMU): The BMU is chosen by randomly drawing a sample vector $\mathbf{x}_i$ from the data set and finding the node with coordinates "**k**" in the output space where its weight vector $\mathbf{w}_k$ is the most similar vector (BMU) along the output space. That is, the node that has the shortest distance to the selected sample vector $\mathbf{x}_i$. The most common distance measure used is the Euclidean distance, although in the case of compositional data sets, Aitchison distance is the most appropriate, which is the Euclidean distance between the corresponding log-ratio coordinates.

Step 3: Weight vectors updating: For CoDa, the SOM algorithm can be performed on coordinates with respect to a log-ratio basis (Mateu-Figueras et al. 2011). That is, the vector $\mathbf{w}_k$ of the BMU and of its neighborhood vectors $\mathbf{w}_j$ is modified in terms of the simplicial operations perturbation "$\oplus$" and powering "$\odot$" (Aitchison 1986)

$$\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} \oplus (\alpha(t) \cdot N(t, \mathbf{k}, \mathbf{j})) \odot (\mathbf{x}_i \ominus \mathbf{w}_j^{(t)}),$$

and its expression on coordinates is

$$\mathbf{v}_j^{(t+1)} = \mathbf{v}_j^{(t)} + \alpha(t) \cdot N(t, \mathbf{k}, \mathbf{j}) \cdot (\mathbf{y}_i - \mathbf{v}_j^{(t)}), \qquad (2)$$

where $\mathbf{v}_j = \mathrm{ilr}(\mathbf{w}_j)$ and:

- $t \geq 0$ is an integer, the number of iterations made or the discrete-time coordinate; the function $\alpha(\cdot)$, known as the *learning-rate* factor ($0 < \alpha(\cdot) < 1$), should initially have reasonably high values (close to unity), thereafter decreasing monotonically. It serves as a weight for the magnitude that the BMU and the neighborhood nodes are modified, and decreases with each iteration (i.e., the magnitude of the updating the vectors decreases with time). Although Kohonen (2001) states that "An accurate time function is not important", he suggested that the most recommended functions are exponential, inversely proportional to *t*, or linear. Wehrens and Buydens (2007) programmed the latter option of Kohonen (2001) to $\alpha(t) = \alpha_0 \cdot (1 - t/\alpha_r)$, where α_0 and α_r are parameters; and
- the function $N(\cdot)$, called the *neighborhood function*, is a smoothing kernel defined over the grid points. This factor regulates the rate at which the size of the neighborhood is impacted by diminishing the distance between the BMU "**k**" and its neighborhoods "**j**" (i.e., the function describes the rate at which the neighborhood size around each BMU shrinks with each iteration). A popular choice is the Gaussian kernel $N(t, \mathbf{k}, \mathbf{j}) = \exp(-d_e(\mathbf{k}, \mathbf{j})^2/(2\sigma^2(t)))$, where $\sigma(t) = \sigma_0 \cdot \exp(-t/\sigma_r)$ and σ_0 and σ_r are parameters. Note that here the Euclidean distance is applied between two elements of the output space, usually R^2.

The results in the SOM are invariant under change of basis. Indeed, because a change of orthonormal coordinates system is a linear transformation (rotation), the new log-ratio coordinates are $\mathbf{y^*}_j = \mathbf{A}\cdot\mathbf{y}_j$, where $\mathbf{A}^t\cdot\mathbf{A}=\mathbf{I}$, and $\mathbf{I}$ is the identity matrix. When applied to Eq. (2)

$$\mathbf{v^*}_j^{(t+1)}= \mathbf{v^*}_j^{(t)} + \alpha(t)\cdot N(t, \mathbf{k}, \mathbf{j})\cdot(\mathbf{y^*}_i - \mathbf{v^*}_j^{(t)}),$$

$$\mathbf{A}\cdot\mathbf{v}_j^{(t+1)}= \mathbf{A}\cdot\mathbf{v}_j^{(t)} + \alpha(t)\cdot N(t, \mathbf{k}, \mathbf{j})\cdot(\mathbf{A}\cdot\mathbf{y}_i - \mathbf{A}\cdot\mathbf{v}_j^{(t)}),$$

$$\mathbf{v}_j^{(t+1)}= \mathbf{v}_j^{(t)} + \alpha(t)\cdot N(t, \mathbf{k}, \mathbf{j})\cdot(\mathbf{y}_i - \mathbf{v}_j^{(t)}).$$

This property is of critical importance because it allows the analyst to select the log-ratio orthonormal basis to improve the interpretation of the results. In addition, the property shows that the CoDa-SOM is perfectly compatible with other techniques such as the principal balances algorithm (Martín-Fernández et al 2018b) that allow for the selection of log-ratio coordinates taking into account the variability of the data set.

Step 4: Iteration: The process is repeated from Step 2 until either the maps stop changing significantly or a fixed number of iterations is reached. Kohonen (2001) recommends “A rule of thumb is that, for good statistical accuracy, the number of steps must be at least 500 times the number of network units.” In contrast, Wehrens and Buydens (2007) propose 100 as the number of iterations, that is, the number of times the data set will be presented to the map. Regardless of the exact number of iterations performed, the size of the neighborhood and the learning-rate factor both decrease with each iteration, allowing for a more stable structure to develop.

The unsupervised SOM can be extended to a supervised version for classification problems (Melssen et al. 2006). In the supervised CoDa-SOM a fused distance (d_{Fused}), from both composition and categorical variables, is proposed for step 2. Following Wehrens and Buydens (2007), we propose a distance based on a weighted combination of the distances between the compositions (d_a) and a distance between the corresponding categories or classes (d_c)

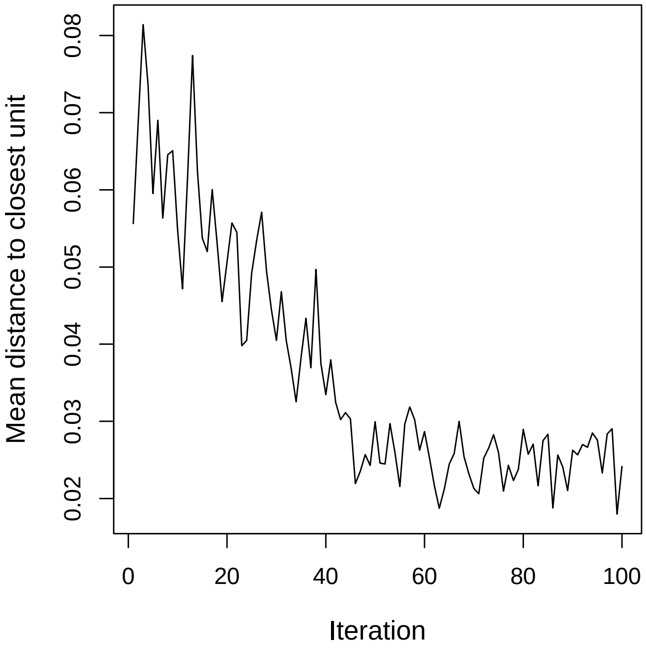
$$d_{\text{Fused}}(\mathbf{x}_i,\mathbf{w}_i)= \partial(t)\cdot d_a(\mathbf{x}_i,\mathbf{w}_i)+(1- \partial(t))\cdot d_c(\mathbf{x}_i,\mathbf{w}_i),$$

where the parameter $\partial(t)$ regulates the relative weight between the distances, decreasing linearly in time. Initially, the distance between the compositional part will dominate the determination of the BMU. At the end, both distances d_a and d_c contribute equally to the determination of the BMU (Wehrens and Buydens 2007).

2.3 A simple example for CoDa-SOM

To illustrate the performance of an unsupervised SOM for CoDa, the results of a one-dimensional SOM for a typical environmental data set in S^3 are shown in Figure 1. The data (Aitchison 1986), comprised of 39 samples, contain information on sediment grain size distribution [sand, silt, clay] as a function of depth in an Arctic lake. The parameters for the SOM algorithm used the default values provided in Wehrens and Buydens (2007). Figure 1(a) shows the evolution along the iterations of the distance, on average, between the samples and their BMUs. The convergence of the algorithm is achieved after 50 iterations when the mean distance stabilizes. The samples are represented by circles in the Fig 1(b), where the squares represent the final value of the nine weight vectors. The few nodes that are present are enough to fit the trend of the data set reasonably well. The labels for the nodes show the average of the depth of the samples assigned to each BTU. At the first glance, the ternary diagram suggests that the sand percentage diminishes with depth. The linear regression model, which is described in Aitchison (2008), fits the composition using the log-transformation of the depth as predictor using the direction [sand, silt, clay]=[4.6%, 23.89%, 71.51%] where the proportion of the variability explained, that is, the R-squared (Egozcue et al 2012) is 71%. Using this model one obtains the predicted composition [7.90, 52.59%, 39.50%] for depth= 61.7, which is reasonably consistent with the composition [6.73%, 49.64%, 43.63%] in the corresponding node created by the CoDa-SOM (Fig. 1).

(a)


Mean distance to closest unit
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0
20
40
60
80
100
Iteration

(b)

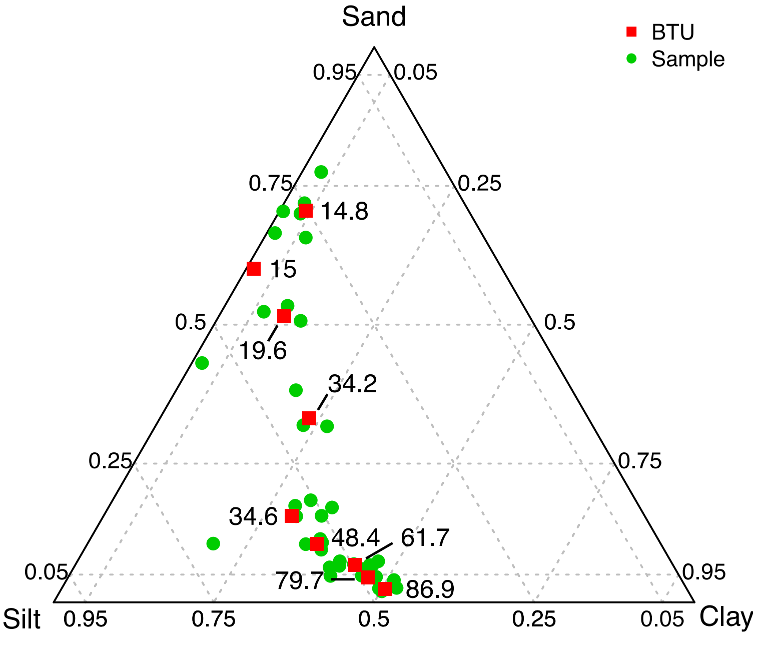

Sand
Silt
Clay
■ BTU
● Sample
0.95
0.05
0.75
0.25
0.5
0.5
0.25
0.75
0.05
0.95
0.95
0.75
0.5
0.25
0.05
14.8
15
19.6
34.2
34.6
48.4
61.7
79.7
86.9

Fig 1 Results of the SOM for a CoDa in S^3 (Aitchison 1986): (a) error function evolution using Aitchison distance; (b) data set (circles) and weight vectors of BTU (squares) for a one-dimensional SOM with nine nodes in a ternary diagram. Labels assigned to the BTU denote the average of the depth of the corresponding assigned samples.

3 The Wyoming coal power plant case study

3.1 The data and some previous analysis

The United States produces more than 50 million tons of coal combustion products (CCPs) per year, which contain a wide variety of potentially hazardous trace elements. Improper CCP management has led to deleterious environmental impacts, such as those linked to large coal spills near Eden, North Carolina in 2014 and Kingston, Tennessee in 2008, partly due to their trace element load (Ruhl et al. 2009). Moreover, knowledge about elemental abundances in CCPs is needed to investigate possible health effects or potential mineral sources for certain critical elements, such as rare earth elements (Kolker et al. 2017). To understand the occurrence and abundance of elements within CCPs and to examine elemental redistribution during coal combustion, the U.S. Geological Survey collected pulverized coal and CCPs from several coal burning power plants (Affolter et al. 2011). Here we focus on results from a single power plant in Wyoming, USA, utilizing subbituminous (Fruitland Formation) pulverized coal. The pulverized feed coal was collected just before it entered the boiler and CCPs that were representative of that feed coal were collected based on plant operating conditions and estimations of the length of time it would take the pulverized coal (PC) to reach the collection points: bottom ash (BA) collected from the boiler, economizer fly ash (EFA) captured at the economizer unit, and fly ash (FA) collected from an electrostatic precipitator. Affolter et al. (2011), Swanson et al. (2013), and Martín-Fernández et al. (2018a) provide more details of the sampling process and the analytical methodologies used to determine oxide and elemental compositions. In total, five full sets of 15 samples were collected from the plant, including of feed coal, PC, BA, EFA, and FA. Following Martín-Fernández et al. (2018a), feed coal was not included in this current study as it was not introduced into the boiler. Each sample includes the proportions of 10 major elements expressed as percent oxides (Al_2O_3, CaO, Fe_2O_3, K_2O, MgO, Na_2O, P_2O_5, SiO_2, SO_3, TiO_2) and 30 minor and trace-elements in ppm (As, Ba, Be, Bi, Cd, Cl, Co, Cr, Cs, Cu, Ga, Ge, Hg, Li, Mn, Mo, Nb, Ni, Pb, Rb, Sb, Sc, Se, Sr, Th, Tl, U, V, Y, Zn). Because the focus was on the change of elemental composition during the combustion process, following Martín-Fernández et al. (2018a), we focused on a paired analysis between the compositions of two consecutive sampling points: BA with respect to PC; EFA compared to BA; and FA compared to EFA. For simplicity we denote $\mathbf{X}_j/\mathbf{X}_k$ for the paired-data set obtained (perturbation differences). For instance, $\mathbf{X}_{FA}/\mathbf{X}_{EFA}$ represents the paired-data for the comparison between the multivariate compositions in FA with respect to EFA. As a consequence, our data set is formed of three full sets of 15 perturbations difference (5 sets x 3 perturbation differences per set).

Martín-Fernández et al. (2018a) confirm that, on average, the three perturbation difference sets are different from the neutral perturbation: for any two consecutive sampling points the compositions of the ash "before" and "after" are, on average, significantly different. This current study extends the analysis of Martín-Fernández et al. (2018a) and demonstrates that there are differences among the three perturbations difference sets that show that there are differences among the variation of the CCP compositions in the sampling points. The results from CoDa-SOM provide information to characterize the combustion process in the points BA, EFA, and FA.

3.2 A separate analysis for the major, minor and trace elements?

Major, minor, and trace elements typically are separately interpreted because of the large differences between their abundances. However, because the data sets analyzed in this study are formed of perturbation differences, the scale of all the elements is similar. Because the SOM are based on distances as a measure of similarity, we analyzed the Aitchison distances of the dataset under three different conditions: the full composition (major, minor and trace elements), major-only elements, and minor and trace elements only. Table 1 summarizes the corresponding Aitchison distance matrices for the three conditions.

Table 1 Basic statistics for the Aitchison distance matrix for the perturbation differences: full composition (all elements); major elements; and minor and trace elements.

	minimum	Q_1	median	Q_3	maximum
Full composition	0.96	4.24	12.13	13.33	17.72
Major elements	0.04	1.20	3.10	6.19	9.30
Minor and trace elements	0.85	4.09	10.55	11.97	17.53

Consistently, the Aitchison distances of the subcompositions formed by perturbation differences of the major elements and of the minor and trace elements are lower than those of the full composition. Notably, the values for the minor and trace elements are more similar to the values for the full composition than those of the major elements. This relationship was expected because the minor and trace element subcomposition is formed by 30 parts of the 40 parts that constitute the full composition.

Table 2 Pearson correlation coefficient (r) and mean square error (mse) between the Aitchison distance matrices for the perturbation differences: full composition (all elements); major elements; and minor and trace elements.

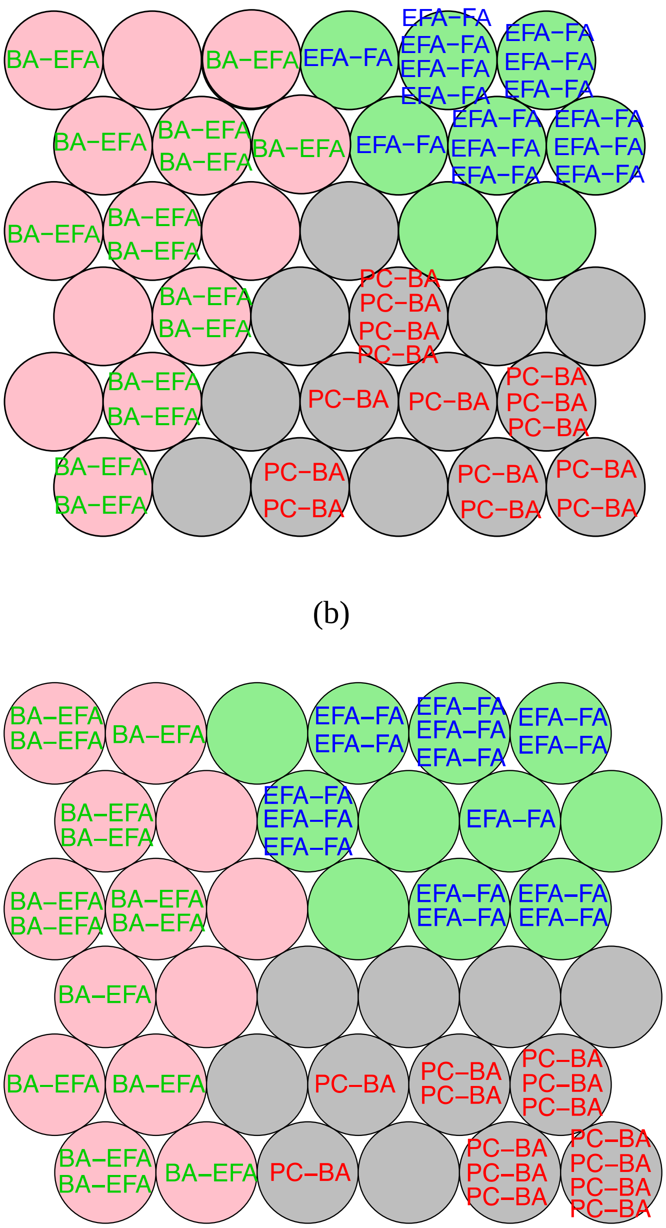
	r	mse
Full composition vs Major elements	0.748	51.07
Full composition vs Minor and trace elements	0.981	1.96
Major elements vs Minor and trace elements	0.612	39.75

The Pearson correlation coefficient between the corresponding distances (Table 2) is used to describe how similar are two distance matrices (Sokal and Rohlf 1962). Table 2 suggests a similar relationship to the Aitchison distance statistics (Table 1) between the distance matrices. Indeed, the correlation between the Aitchison distance of the perturbation differences for full composition and the minor and trace elements is equal to 0.981, where it is equal to 0.748 for the major elements with the full composition. The correlation coefficient between the two subcompositions is only 0.612 (Table 2). In addition, the square error mean between the corresponding distance matrices exhibits the same behavior. This error is 51.07 for the matrices of the full composition and the major elements, but much smaller (1.96) for the minor and trace elements. The mean square error between the matrices of the two subcompositions is equal to 39.75, suggesting notable differences between major elements and the minor- and trace elements. All these indices suggest that SOM applied to the full composition will be very similar to the SOM for the minor and trace elements. In consequence, for geological interpretation purposes, it is worth analyzing the major elements separately. This conclusion also applies to other statistical techniques using distances, such as cluster analysis or MDS.

3.3 Results of from application of CoDa-SOM

Both data sets (major elements vs. minor and trace elements) are formed by 45 perturbation differences and the categorical variable that informs the sampling points BA, EFA, and FA. Wehrens and Buydens (2007) introduce the SOM for supervised pattern recognition to model the categorical variable as a dependent variable for which predictions can be obtained (as opposed to unsupervised approaches wherein categorical or data information on group association are excluded, even if known). Figure 2 shows the resulting supervised SOM for a hexagonal topology with 36 nodes (6 x 6). Default values of the program (Wehrens and Buydens 2007) were used for all other parameters In both cases, the convergence of the algorithm is achieved in 50 or less iterations as indicated by stabilization of the mean distances, following a pattern similar to Fig 1a. The labels “PC-BA”, “BA-EFA”, and “EFA-FA” indicate that this node is the BMU of a perturbation difference in the corresponding sampling point. Positions within a node in a SOM are without meaning, they simply indicate that the BMU was determined from multiple perturbation differences. Intuitively these perturbations capture the changes in chemical composition during coal combustion and the myriad processes which occur in the flue gas. The BMU for both major oxides (Fig. 2a) and minor and trace elements (Fig. 2b) form non-overlapping groups. The colors gray (PC-BA), pink (BA-EFA) and green (EFA-FA) indicate the predicted group for the node and were assigned strictly, because a supervised SOM algorithm was performed. The predictions define clear borders between the three groups, suggesting that the chemical processes occurring between the different stages in the power plant are distinct. The majority of the nodes close to the borders do not correspond to any BMU, suggesting that these nodes are transitional nodes between groups. This clear separation between the groups was corroborated with a leave-one-out cross-validation procedure where the resulting misclassification error rate was respectively 2.22% and 0% for the major oxides and for the minor and trace elements, respectively. These results are fully comparable with linear discriminant analysis misclassification rates of 0% and 4.44%, respectively, reported by Vasighi and Kompany-Zareh (2013).

(a)

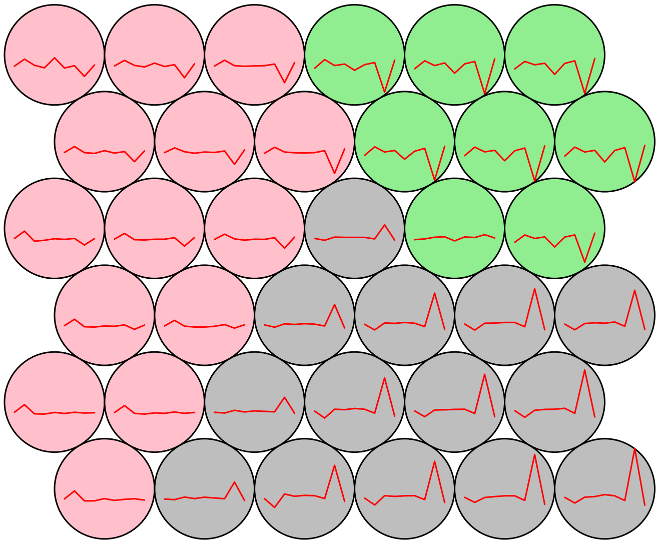


(b)

Fig 2 SOM for the Wyoming coal power plant case study: BMU codes for the perturbation differences in the sampling points: bottom ash (PC-BA); economic fly ash (BA-EFA); and fly ash (EFA-FA). Colors gray, pink and green represent respectively the assigned group PC-BA, BA-EFA, and EFA-FA: (a) major oxides; (b) minor and trace-elements

Figure 3 shows the value of the ilr-coordinates (Eq. (1)) for the nodes in the major oxides data set (Fig. 3a) and in the minor and trace elements (Fig. 3b), separately for the same SOMs shown in Figure 2. Much like PCA (but unlike cluster analysis and MDS), information can be gleaned from a single SOM about both the relationships between the samples and the variables. For each node of Fig. 3a, a *line* plot was created showing the value of the nine ilr-coordinates created from the 10 major elements (Martín-Fernández et al 2018a), ordered in the same order of the columns in the data matrix (ilr$_1$, ilr$_2$, ilr$_3$, ilr$_4$, ilr$_5$, ilr$_6$, ilr$_7$, ilr$_8$, ilr$_9$). The nodes of the group PC-BA have large negative values in the coordinates ilr$_2$, ilr$_7$, and ilr$_9$, and positive in ilr$_8$. The group BA-EFA shows large positive values in ilr$_2$. In contrast, ilr$_5$ and ilr$_8$ are the coordinates with the highest negative values for the group EFA-FA. In Fig. 3b, the differences between the patterns of the lines for the three groups are relevant. For example, the group BA-EFA (pink color) has a different pattern from the group PC-BA (gray color) because the former takes high values in the first coordinates and values close to zero in the rest. In contrast, the nodes of the latter group have high values (positive and negative) for several coordinates. The group BA-EFA is noted for its high negative values in a few ilr-coordinates. To complete the interpretations these plots can be produced using the raw data of the perturbation differences, by applying inverse of the ilr transformation.

(a)



(b)

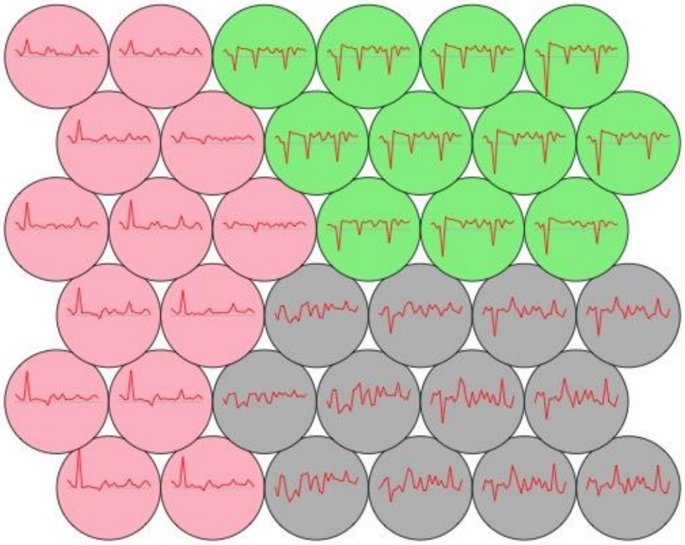
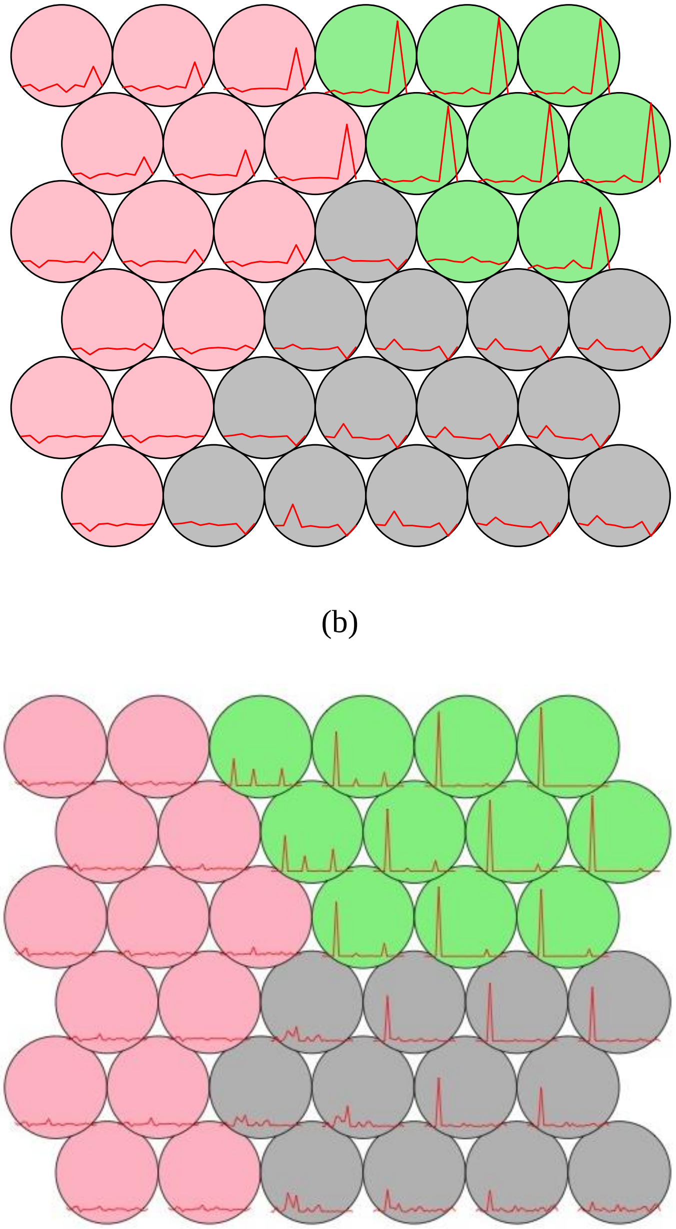


Fig 3 SOM for the Wyoming coal power plant case study: ilr-coordinates of the codes. The colors gray, green and pink represent the predicted group PC-BA, BA-EFA, and FA-EFA, respectively: (a) major oxides; (b) minor and trace-elements

Figure 4 shows the back-transformed raw values (in proportions) of the nodes for the major oxides (Fig. 4a) and the minor and trace-elements (Fig. 4b). The line plots show the value of the parts ordered in the same order of the columns in the data matrix: (Al_2O_3, CaO, Fe_2O_3, K_2O, MgO, Na_2O, P_2O_5, SiO_2, SO_3, TiO_2) and (As, Ba, Be, Bi, Cd, Cl, Co, Cr, Cs, Cu, Ga, Ge, Hg, Li, Mn, Mo, Nb, Ni, Pb, Rb, Sb, Sc, Se, Sr, Th, Tl, U, V, Y, Zn). This plot can be helpful to improve the interpretation of the ilr-coordinates plot because one can analyse which parts are responsible of the characterization of the ratios. For example, examination of Fig. 4a suggests that relative to the other major oxides, the abundance of SO_3 changed more between EFA and FA (green nodes) than any other chemical constituent, suggesting a reduction in the stability of SO_3 between those points in the plant. The line plots of the nodes in the groups PC-BA and BA-EFA have similar shapes with some relevant differences. For example, the perturbation difference for Fe_2O_3 has higher values in the group PC-BA than for the other groups. This is likely because iron has relatively high melting (~1500 ^{o}C) and boiling points (~2900 ^{o}C) thus it is enriched in the bottom ash relative to the pulverized coal. The most relevant feature in Fig 4b is that the perturbation differences for group BA-EFA in the major elements have very small values, close to the neutral element (1/30), whereas the other two groups take high values in the first chemical elements. This finding suggests that the coal in the boiler goes through a substantial re-distribution of the elements during combustion (PC-BA), with those major elements that are most mobile (i.e., SO_3) being fractionated into the flue gas and those that are difficult to mobilize being concentrated in the bottom ash (e.g., Al_2O_3, Fe_2O_3, TiO_2). Once the flue gas leaves the boiler and travels through the economizer (BA-EFA; pink nodes) there appears to be little substantial change in the major element composition of the particles, minus the most volatile major elements (Na_2O and SO_3), as most perturbation values are near zero for the major element subset. Conversely, in the SOM for the trace element results, we can see that the trace element distributions are more impacted for this perturbation, as shown by the prevalence of high values for several constituents (Fig. 4b). In the final stage (EFA-FA; green nodes) between the economizer and the electrostatic precipitation (the device which captures the fly ash), there appears to be an additional redistribution of the major and trace elements. The assignment of the samples to the nodes and their relative position (indicated in Figs. 2a and 2b) also provides information about patterns and trends in the chemical data (indicated in Figs. 4a and 4b). For instance, in the EFA-FA perturbation (pink nodes) of the major elements, samples with BMU in the left half of the SOM show little difference in the relative composition of Na_2O, while those to the right show substantial differences in the relative abundance of Na_2O (Figs. 2a and 4a). A nearly opposite relationship with respect to SO_3 is observed along the same trend suggesting an inverse relationship between Na and S as the flue gas moves between the economizer and the electrostatic precipitators. Similar interpretations can be made for the other constituents and other samples.

(a)



(b)

Fig 4 SOM for the Wyoming coal power plant case study: raw values for the perturbation differences of the codes. Gray, green and pink colors represent the predicted groups PC-BA, EFA-BA, and FA-EFA, respectively: (a) major oxides; (b) minor and trace-elements

Beyond this simple case study, one can consider the exceptionally flexible ways the SOMs could be adapted for variations of this study. For instance, one can map new samples to pre-existing SOM (Wehrens and Buydens 2007). Consider a scenario where power plant personnel changed the operating conditions of the plant (e.g., combustion temperature, cooling rates of flue gas, etc.) and they were interested in how that affected the distribution of the trace elements. Coal and coal combustion products collected and analyzed under the new operating condition, could be mapped to the BMU in the pre-existing SOMs. Comparison of where the new samples plot in the SOM in comparison to the original sample set (e.g., Figs. 2a and 2b) and the values for the specific perturbations (Figs. 4a and 4b) can provide information on changes in chemical speciation within the boiler. To simulate this situation, we selected five samples from the minor and trace elements in the original data set. We artificially contaminated these samples in all the sampling points adding a random number (between -3 and +3) of standard deviations in each ilr-variable. As a result, we generated 15 new samples (five in each sampling point) adding random noise to its counterpart original samples. Figure 5 shows how these new samples are mapped in the original SOM. The color gray, green and pink represent respectively the predicted group PC-BA, BA-EFA, and EFA-FA for the nodes using the original data (Fig. 2b). The labels bottom ash (PC-BA), economic fly ash (BA-EFA), and fly ash (EFA-FA) correspond to the BMU for the 15 perturbation differences of the new data set in the sampling points. SOM is able to show that, as expected, the random noise added to the original samples caused some of the new samples to move from one BMU to a different BMU suggesting that the new samples have a different perturbation differences distribution and perhaps become more like samples from different categories. Alternatively, if some of the data for a subset of samples was missing or lost, it can still be utilized (Dickson and Giblin 2007), as long as a distance or similarly measurement (e.g., Aitchison distance) can be calculated for each data point and the nodes, such than a BMU can be assigned. This simple behavior makes SOM extremely useful in situations where the analysis of every sample for every constituent is cost prohibitive or when constructing a database of multiple samples that have analyzed with different methodologies.

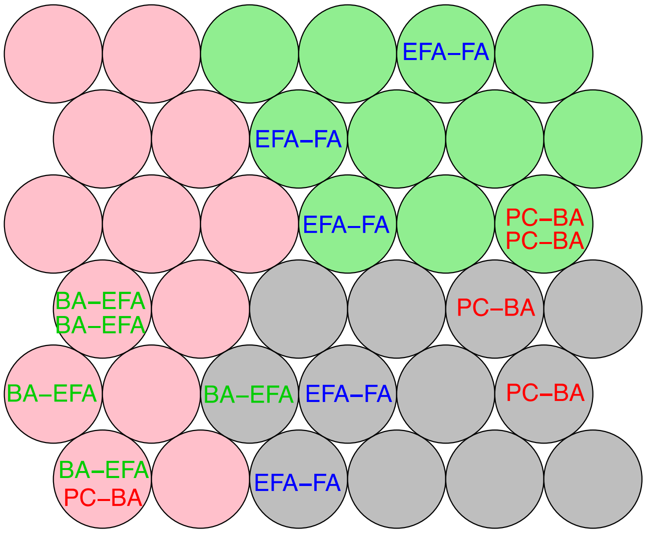


Fig 5 SOM for the minor and trace elements in the Wyoming coal power plant case study with 15 new samples generated by adding random noise to pre-existing data and mapping them to the original SOM. Colors gray, pink and green represent, respectively, the predicted group PC-BA, BA-EFA, and EFA-FA using the original data set. BMU codes for the five perturbation differences of the new data set in the sampling points: bottom ash (PC-BA); economic fly ash (BA-EFA); and fly ash (EFA-FA)

4 Conclusions

A collection of techniques is proposed to create a methodology for application of the SOM to compositional data. The Aitchison distance is an appropriate measure to find the BMU of a sample. The simplicial operations perturbation and powering allow for the construction of the SOM in a topologically consistent way. Following the principle of working on coordinates, the SOM algorithm can be applied using the classical elements: Euclidean distance and basic operations. Importantly, the interpretation of the maps should be given in terms of the log-ratio coordinates and complemented with the maps in terms of the raw data. These features have been illustrated in the analysis of CCPs that are residuals to be treated according its composition, with relevant importance of the minor and trace elements. The supervised SOM applied to the coal ashes data set provided a characterization of the changes in the composition of the CCPs. Moreover, we demonstrated the ability to map new data onto the original SOM, allowing for non-linear discrimination or other analyses involves change in operating conditions of the plant. Thus, beyond basic understand on element distribution during coal combustion, this interpretation is relevant understanding potential changes related to design and coal combustion plant management

Acknowledgments This work has been supported by the project “CODA-RETOS” (Spanish Ministry of Economy and Competitiveness; Ref: MTM2015-65016-C2-1-R) and the project “Compositional Data Analysis Related to Energy Resources Modeling” (“Salvador de Madariaga” program; “Fulbright” distinction; MECD; Ref.: PRX16/00258). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. We are grateful to C.Ö. Karacan (USGS) and G. Mateu-Figueras (U. de Girona) for their insightful review of a previous version of the paper.

References

Affolter RH, Groves S, Betterton W, Benzel W, Conrad KL, Swanson SM, Ruppert LF, Clough JG, Belkin HE, Kolker A, Hower JC (2011) Geochemical database of feed coal and coal combustion products (CCPs) from five power plants in the United States. U.S. Geological Survey Data Series 635, pamphlet, 19 pp

Akinduko AA, Mirkes EM, Gorban AN (2016) SOM: Stochastic initialization versus principal components. Information Sciences 364–365:213–221

Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability, Chapman & Hall/CRC. Reprinted in 2003 by The Blackburn Press, Caldwell, NJ

Aitchison J (2008) The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. In: Daunis-i-Estadella, J. and Martín-Fernández, J.A. (Eds.), Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop, May 27-30, University of Girona, Girona (Spain), CD-ROM (ISBN: 978-84-8458-272-4, http://hdl.handle.net/10256/706)

Barceló-Vidal C, Martín-Fernández JA (2016) The mathematics of compositional analysis. Austrian Journal of Statistics 45(4):57–71

Cortés JA, Palma JL (2013) Geological applications of self-organizing maps to multidimensional compositional data. Pioneer Journal of Advances in Applied Mathematics 7(2):17–49

Cox TF, Cox MAA (2001) Multidimensional scaling. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition, 308 pp

Dickson BL, Giblin AM (2007) An evaluation of methods for imputation of missing trace element data in groundwaters. Geochemistry: Exploration, Environment, Analysis 7:173–178

Edjabou ME, Martín-Fernández JA, Scheutz C, Astrup TF (2017) Statistical analysis of solid waste composition data: Arithmetic mean, standard deviation and correlation coefficients. Waste Management 69:13–23

Egozcue JJ, Daunis-i-Estadella J, Pawlowsky-Glahn V, Hron K, Filzmoser P (2012) Simplicial regression. The normal model. Journal of Applied Probability and Statistics (JAPS) 6(1& 2): 87-108

Everitt BS, Landau S, Leese M, Stahl D (2011) Cluster analysis. John Wiley & Sons, Ltd, Chichester, United Kingdom, 5th Edition, 330 pp

Kohonen T (2001) Self-organizing maps. Number 30 in Springer Series in Information Sciences. Springer-Verlag, Berlin, 3rd edition, 501 pp

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

Jolliffe IT (2002) Principal component analysis. Springer Series in Statistics. Springer-Verlag New York, 2nd edition, 487 pp

Kolker A, Scott C, Hower J C, Vazquez J A, Lopano C L, Dai S (2017) Distribution of rare earth elements in coal combustion fly ash, determined by SHRIMP-RG ion microprobe. International Journal of Coal Geology 184:1–10

Martín-Fernández JA, Daunis-i-Estadella J, Mateu-Figueras G (2015) On the interpretation of differences between groups for compositional data. SORT 39(2):231-252

Martín-Fernández JA, Olea RA, Ruppert LF (2018a) Compositional data analysis of coal combustion products with an application to a Wyoming power plant. Mathematical Geosciences 50(6):639-657

Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2018b) Principal balances for compositional data. Mathematical Geosciences 50(3):273–298

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue, JJ (2011) The principle of working on coordinates, in Compositional data analysis: theory and applications Pawlowsky-Glahn V, Buccianti A, eds. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/9781119976462.ch3

Melssen W, Wehrens R, Buydens L (2006) Supervised Kohonen networks for classification problems. Chemom. Intl. Lab. Sys. 83:99–113

Olea RA, Janardhana Raju N, Egozcue JJ, Pawlowsky-Glahn V, Shubhra Singh (2018) Advancements in hydrochemistry mapping: application to groundwater arsenic and iron concentrations in Varanasi, Uttar Pradesh, India. Stochastic Environmental Research and Risk Assessment 32(1):241–259

Palarea-Albaladejo J, Martín-Fernández JA, Soto JA (2012) Dealing with distances and transformations for fuzzy C-means clustering of compositional data. Journal of Classification 29:144–169

Palarea-Albaladejo J, Martín-Fernández JA (2015) zCompositions - R package for multivariate imputation of nondetects and zeros in compositional data sets. Chemometrics and Intelligent Laboratory Systems 143:85−96

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. John Wiley & Sons, Chichester, 378 pp

Ruhl L, Vengosh A, Dwyer G S, Hsu-Kim H, Deonarine A, Bergin M, Kravchenko J (2009). Survey of the potential environmental and health impacts in the immediate aftermath of the coal ash spill in Kingston, Tennessee. Environmental Science and Technology 43: 6326–6333

Sokal RR, Rohlf FJ (1962). The comparison of dendrograms by objective methods. Taxon, 11:33-40

Swanson SM, Engle MA, Ruppert LF, Affolter RH, Jones KB (2013) Partitioning of selected trace elements in coal combustion products from two coal-burning power plants in the United States. International Journal of Coal Geology 113:116–126

Vesanto J, Alhoniemi E (2000) Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11(3):586−600

Vasighi M, Kompany-Zareh M (2013) Classification ability of self-organizing maps in comparison with other classification methods. Communications in Mathematical and in Computer Chemistry 70: 29−44

Wehrens R, Buydens LMC (2007) Self- and Super-organizing maps in R: The kohonen package. Journal of Statistical Software 21(5):1−19

Zellmer G, Turner S, Hawkesworh, C (2000) Timescales of destructive plate margin magmatism: new insights from Santorini, Aegean volcanic arc. Earth and Planetary Science Letters 174:265−281