Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland)

Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland)

Raimon Tolosana-Delgado, Jennifer McKinley

PII: S0883-2927(16)30075-0

DOI: 10.1016/j.apgeochem.2016.05.004

Reference: AG 3651

To appear in: *Applied Geochemistry*

Received Date: 20 January 2016

Revised Date: 6 May 2016

Accepted Date: 9 May 2016

Please cite this article as: Tolosana-Delgado, R., McKinley, J., Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland), *Applied Geochemistry* (2016), doi: 10.1016/j.apgeochem.2016.05.004.

Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland)

Raimon Tolosana-Delgado[a*], Jennifer McKinley[b]

[a]Helmholtz Zentrum Dresden Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Dept. Modelling and Valuation, Halsbrueckerstr 34 D-09599 Freiberg (Germany); r.tolosana@hzdr.de

[b]School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, BT7 1NN, UK; j.mckinley@qub.ac.uk

[*]Corresponding author: r.tolosana@hzdr.de

Abstract

The complexity of modern geochemical data sets is increasing in several aspects (number of available samples, number of elements measured, number of matrices analysed, geological-environmental variability covered, etc), hence it is becoming increasingly necessary to apply statistical methods to elucidate their structure. This paper presents an exploratory analysis of one such complex data set, the Tellus geochemical soil survey of Northern Ireland (NI). This exploratory analysis is based on one of the most fundamental exploratory tools, principal component analysis (PCA) and its graphical representation as a biplot, albeit in several variations: the set of elements included (only major oxides vs. all observed elements), the prior transformation applied to the data (none, a standardization or a logratio transformation) and the way the covariance matrix between components is estimated (classical estimation vs. robust estimation). Results show that a log-ratio PCA (robust or classical) of all available elements is the most powerful exploratory setting, providing the following insights: the first two processes controlling the whole geochemical variation in NI soils are peat coverage and a contrast between “mafic” and “felsic” background lithologies; peat covered areas are detected as outliers by a robust analysis, and can be then filtered out if required for further modelling; and peat coverage intensity can be quantified with the %Br in the subcomposition (Br, Rb, Ni).

Keywords: centered log-ratio transformation, clr, spurious correlation, compositional data analysis.

1. Introduction

Geochemical datasets are increasing, both in the number of samples routinely collected and in the number of components analysed. These datasets include elements with typical values which cover ranges of magnitude from % to ppm or even ppb. Such geochemical datasets may cover a single deposit or formation, a relatively small area or region of interest, a country or a whole continent or subcontinent, involve one or many matrices (river water, underground water, moss or other vegetal tissues, rock, soil, stream sediments, single grains of the same mineral phase, etc.), be static or imply a time evolution. It is becoming, thus, increasingly necessary to have appropriate tools to explore this potentially large geochemical variability An example of such framework is provided by any modern regional geochemistry survey (GEMAS for Europe: Reimann et al. 2014a,b; Australia: Caritat and Cooper, 2011a,b; North America: Smith et al., 2011; Drew et al., 2010; Canada: Friske et al., 2013; China: Wang et al., 2015), typically having thousands of samples analysed for several tens of elements covering diverse geological units in non-homogeneous climatic zones and landscape environments.

Until now, most practitioners in the field of geochemistry analyse such databases with a quite informal, intuitive approach. Such an approach comprises plotting the data in standard bivariate diagrams (a.k.a. Harker diagrams), trivariate diagrams (ternary diagrams) or less frequently using multivariate approaches (Schoeller diagrams, Piper diagrams, spider diagrams) that have been proposed by others, and then using these plots to identify known patterns. This approach can be tedious (as the number of existing proposed diagrams grows with time) and unfortunately, merely confirmatory in that either the expected grouping, trend or pattern is conveniently observed, otherwise analysts simply do not show the contradictory diagram in their reports. It is thus not *exploratory* (i.e. allowing a search for known as well as unexpected patterns). An alternative approach, becoming increasingly popular, is to apply an appropriate multivariate statistical analysis to the data set.

For exploratory purposes, the most appropriate tools are Principal Component Analysis (PCA) and related projection techniques (FA: Factor Analysis, PP: Projection Pursuit, DA: Discriminant Analysis, etc). All of these techniques search for a few linear combinations of the available variables (a *projection*) that contain “interesting” patterns. Each method specifies in a quantitative manner what is defined as “interesting”. Many of these techniques also allow a graphical representation of both the original variables (the chemical elements) and the observations (the samples) in the first few interesting projections, thus providing quite powerful exploratory tools (Gabriel, 1971; Grafelman and van Eeuwijk 2005; Aitchison, 1997; Pawlowsky-Glahn and Buccianti, 2011). For the sake of simplicity, this paper deals with PCA but many of the conclusions apply to other exploratory projection methods.

Underlying such statistical methods there is most often some assumption of joint normal distribution for the data. In geochemical case studies, this might be an acceptable assumption for many major components and in small carefully sampled datasets, but it becomes decreasingly reliable with increasing complexity or with trace elements. In fact trace elements are said to rather follow lognormal (or quasi-lognormal) distributions, particularly on large spatial scales (Ahrens, 1954a; 1954b).

On the other hand, existing user-friendly multivariate statistics software is typically built for a variety of applications, where often the variables analysed do not share the same units of measurement. Thus, when one wants to build a linear combination of these variables, they are typically standardized to remove units (otherwise one would be adding apples with oranges). This is an unnecessary step in most geochemical datasets, for two reasons. Firstly, all components share the same units if they relate to the same composition, even though some variables might be in % and others in ppm or ppb, therefore we can meaningfully compare them. Secondly, we *can* (and sometimes *do*) add apples and oranges, when we expect two or more elements to behave equivalently (e.g. K and Na in a Piper or a TAS diagrams).

Finally, compositional data are known to be closed, i.e. if we would consider all elements and measure them without error then they would sum to 100% (or 10^6 ppm) on each sample. This constant sum constraint was identified to induce spurious behaviour on the correlation coefficient by Chayes (1960): the so called *negative bias* (the tendency of correlation coefficients between major components to be negative) and the *spurious correlation effect* (the fact that correlation between two components unpredictably changes when considering different subcompositions). These problems do not only affect the correlation coefficients: any statistical method based on them (as all projection methods mentioned before) do suffer from the same spurious character (Butler, 1975; 1976; 1975; 1979; Chayes and Trochimczyk, 1978; Pawlowsy, 1984). These effects can be noticed even when using a few major components, where their total sum approaches 100%.

In the 80s Aitchison (1982, 1986) suggested that all these problems would be solved by realizing that compositional data only carry relative information. He showed that this implies that an appropriate statistical analysis of compositional data should be based on log-ratio transformed data, and introduced a compositional alternative to projections, called *log-contrasts*. The fact is that all of the methods mentioned before are straightforward to apply to geochemical data by using log-contrasts.

The aim of this paper is to compare the performance of a popular projection-based analysis (PCA) using a logratio approach with a non-transformation strategy, in order to: (a) show the potential of a truly exploratory analysis with these statistical methods, and (b) demonstrate the advantages of using log-ratios over more classical approaches. These aspects will be illustrated with the Tellus soil geochemical survey, completed by the Geological Survey of Northern Ireland (GSNI).

The geology of Northern Ireland (see maps SM1 in the online supplementary material) includes a stratigraphic record commencing in the Mesoproterozoic including all geological systems up to the Palaeogene (Mitchell 2004). This has created a diversity of geological bedrock across the region. The north-east is dominated by the Palaeogene basalt lava and lacustrine sedimentary rocks, whilst the north-west is dominated largely by Dalradian psammite and semipelite. Mudstone, sandstone and limestone Carboniferous in age (with a Devonian component) are found across central to south-west Northern Ireland. The southeast comprises Ordovician and Silurian marine sedimentary rocks with younger igneous complexes. Extensive Palaeogene granite bedrock constitute the Mourne mountains to the south-east, The advance of ice sheets and their meltwaters over the last 100,000 years has resulted in at least 80% of bedrock covered by superficial deposits such as glacial till and post-glacial alluvium and peat. In Northern Ireland, the total amount of carbon stored in soils such as peat is estimated to be 386Mt (Cruickshank et al. 1998; Keaney et al. 2013). This is due to the relatively high carbon density of peat and organic-rich soils. Therefore, it is very important to obtain best estimates of peat cover (as a proxy for soil carbon) to manage carbon changes over time.

2. Materials and Methods

2.1. Sampling and data acquisition

The GSNI Tellus ground based geochemical survey, completed between 2004 and 2006, comprises 13,860 soil samples taken at a 20cm depth, collected on a regular grid of one sample site every $2km^2$ (Young and Donald 2013) following the G-BASE sampling regime established by British Geological Survey (BGS). This provides a spatial dataset with an extensive suite of soil geochemical analysis. The soil samples used in this paper were analysed for 60 elements and inorganic compounds using pressed pellet X-Ray Fluorescent Spectrometry (XRF) using Wavelength Dispersive XRF Spectrometry (WD-XRF) and Energy Dispersive/Polarised XRF Spectrometry (ED-XRF). The sampling and analysis regimes for the geochemical surveys included in the Tellus Survey are detailed in Smyth (2007) and Young and Donald (2013).

A simplified bedrock classification was defined based on the scheme used by Rawlins et al. (2012). This defined the rock types: gabbro, granite, basalt, andesite, acid volcanics, dykes, psammite and semipelite, conglomerate, sandstone, lithic arenite, mudstone and limestone. A second classification defined the rock types in terms of their textural and then chemical characteristics. The last scheme defined the Quaternary superficial deposits including peat.

2.2. Quantifying variability and dependence

Let us consider the proportions of the D elements measured on one particular sample n as a vector of D non-negative values $\mathbf{x}_n = [x_{n1}, x_{n2}, \dots, x_{nD}]$. Consider a sample of N of these vectors. The variance is the classical way of measuring the variability of each component,

$$s_j^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(x_{nj} - \bar{x}_j\right)^2,$$

where the mean value of the j-th variable is computed as

$$\bar{x}_j = \frac{1}{N}\sum_{n=1}^{N} x_{nj}.$$

The covariance between two variables i and j

$$s_{ij} = \frac{1}{N-1}\sum_{n=1}^{N}(x_{ni} - \bar{x}_i)\left(x_{nj} - \bar{x}_j\right)$$

is often used to describe the co-dependence of the two variables. But given that this quantity has units (the product of the units of variables i and j), most often the dimensionless correlation coefficient $r_{ij} = s_{ij}/\left(s_i s_j\right)$ is reported. Variance, covariance and correlation of the D components can be arranged in two matrices,

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \cdots & s_{1D} \\ \vdots & \ddots & \vdots \\ s_{D1} & \cdots & s_D^2 \end{bmatrix} \quad \text{and} \quad \mathbf{R} = \begin{bmatrix} 1 & \cdots & r_{1D} \\ \vdots & \ddots & \vdots \\ r_{D1} & \cdots & 1 \end{bmatrix},$$

called covariance matrix and correlation matrix. Note that the correlation matrix coincides with the covariance matrix of the standardized scores, $z_{ni} = (x_{ni} - \bar{x}_i)/s_{ii}$.

If we understand that each vector $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ represents a point on the D-dimensional real space R^D, then the vector of means $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_D]$ represents the centre of gravity of the cloud of points, and the covariance matrix is associated with an ellipsoid describing the shape of the data cloud. This duality between mean vector-covariance matrix and a geometric object has been exploited to define robust alternatives to the classical formulae given before. The so called minimum determinant covariance (MCD) looks for the smallest ellipsoid that contains 50% of the data, and delivers its associated mean value and covariance matrix as a robust estimator of these statistics (Rousseeuw and van Driessen, 1999; Filzmoser, Hron and Reimann, 2009). Robust statistics have the property to be resistant to arbitrary contaminations of a high proportion of observations: as it is defined, the MCD can admit contaminations in less than 50% of the data. The method is equivalent to selecting that 50% plus one data that are most probably non-contaminated, and filtering the rest of the observations by giving them a weight of zero.

2.3. Principal component analysis (PCA)

Principal component analysis can be defined in several ways. For the goals of this paper, it is convenient to understand it as a description of the size, shape and orientation of the ellipsoid associated with the covariance matrix $\mathbf{S}$ or the correlation matrix $\mathbf{R}$ (obtained with the classical formulae or with the MCD procedure). This description is obtained with the eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_D\}$ and eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_D\}$ of the matrix analysed. The matrix $\mathbf{V}$, containing the eigenvectors in columns, is called the *loadings* matrix. The first vector defines the largest principal axis of the ellipsoid, and the direction of highest variability of the data set. The second principal axis is linked to the second eigenvector, and so on. These vectors form a new reference system, onto which to project both the observations and the axes of the original variables. Because the new variables (the *principal components*) are ordered in decreasing variance, if we select the first r of them we will obtain the best rank r approximation to the true data set (*best* in the sense to produce the minimal distortion of the real distances observed in the original set of variables) (Eckart and Young, 1936). Moreover, the last eigenvectors, those related to smaller eigenvalues, define principal components the scores of which have very low variance: these principal components deserve some attention as well, because they might provide insights on quasi-constant combinations of variables, in the fashion of equilibrium constants.

2.5. Compositional data adaptions

In the case of compositional data, it has been mentioned that covariance and correlation are flawed measures of spread and codependence. Instead, one should analyse the data after a log-ratio transformation (Aitchison, 1986). Many log-ratio transformations have been proposed in the literature, each having some advantages and drawbacks (Aitchison, 1982; Aitchison, 1986; Egozcue et. al 2003; Egozcue and Pawlowsky-Glahn, 2005). For the purpose of this paper, we use Aitchison’s centered logratio transformation (clr) for PCA

$$\mathrm{clr}(\mathbf{x}) = \ln \frac{\mathbf{x}}{\sqrt[D]{x_1 x_2 \cdots x_D}},$$

where these logarithms must be applied component-wise. This is the conventional choice in exploratory applications of principal component analysis, because it allows us to keep track of the individual original components whilst preserving the multivariate relative scale of compositional data (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015; Ch. 5).

2.6. Graphical representation of PCA: the biplot

A very compact graphical representation of a data set can be obtained by plotting scatterplots of the data on the first 2 or 3 eigenvectors of the PCA, together with the original axes. This is called a *biplot* (Gabriel, 1971), and it can be based on either a robust or a classical estimate of: (a) the covariance or (b) the correlation of raw components, or alternatively (c) the covariance of the clr-transformed composition. Its most interesting feature is the analogy on a biplot between rays and variables: the length of each ray is roughly proportional to the variance of its associated variable, while the cosine of the angle between two rays is an indicator of the correlation coefficient between their associated variables. Moreover, these rules apply to the links between arrow tips, which represent then the difference between two variables.

These general rules boil down, in the case of a covariance biplot of raw data, to the following: (a.1) long rays typically represent the variables with the largest average values, because of the typical proportionality effect between the variance and mean in positive variables; (a.2) parallel rays indicate variables that are well correlated, positively if the rays point towards the same direction, negatively otherwise; and (a.3) orthogonal rays indicate that the variables are poorly correlated to uncorrelated. In the case of a correlation biplot of raw data, these criteria are: (b.1) long rays typically represent variables very well correlated with the 2 principal components of the plot; and (b.2) parallel resp. (b.3) orthogonal rays indicate that the variables are well resp. poorly correlated with each other. In the case of a covariance biplot of log-ratio transformed data, we can say that: (c.1) the length of a link is proportional to the variance of the simple log-ratio between the two components involved, hence coincident rays (with very short links) indicate variables which are highly proportional; (c.2) parallel links (hence collinear variables) suggest subcompositions dominated by a single one-dimensional process, i.e. where all simple log-ratios are well correlated among them; (c.3) orthogonal links on the other hand suggest that the two involved log-ratios (or subcompositions) are uncorrelated. On a compositional biplot, not very much attention is paid to the rays, as they represent clr-transformed variables, which should not be confounded with the original components. They still have some interpretability, as illustrated later in section 3.4.

All criteria of variability and correlation listed above are actually just hints. Being a projection, the biplot aims at best to capture all of these high-dimensional relations on a single diagram or two. This will only be a realistic representation if the proportion of total variance captured by the chosen principal components is high enough. In geological systems, the experience of the authors suggests that above 65% of explained variance, what biplot shows is typically a good approximation to the real structures in the data set. In the next sections, this proportion of explained variance is always included in the axes of each biplot.

PCA is properly defined for homogeneous, normally-distributed data from a single population, i.e. where no groups or clusters occur. Otherwise two sorts of variability are merged: the inner variability of each group, and the variability between the groups. These sources of variability are in general unrelated. A proper PCA should actually target the within-group variability, which might obscured by the between-groups variability. However, a biplot can still be used for grouped data sets, albeit in an exploratory fashion only. In this case, it is wise to track which observations correspond to each group, in order to highlight systematic differences between groups. If the groups do not appear properly merged in the biplot, then the PCs rather relate to between-group differences. Note that a PCA biplot is not tailored to highlight these differences. For this purpose, graphical representations of discriminant analysis should be more appropriate. This is however beyond the scope of this paper.

3. Results

3.1. PCA of raw data

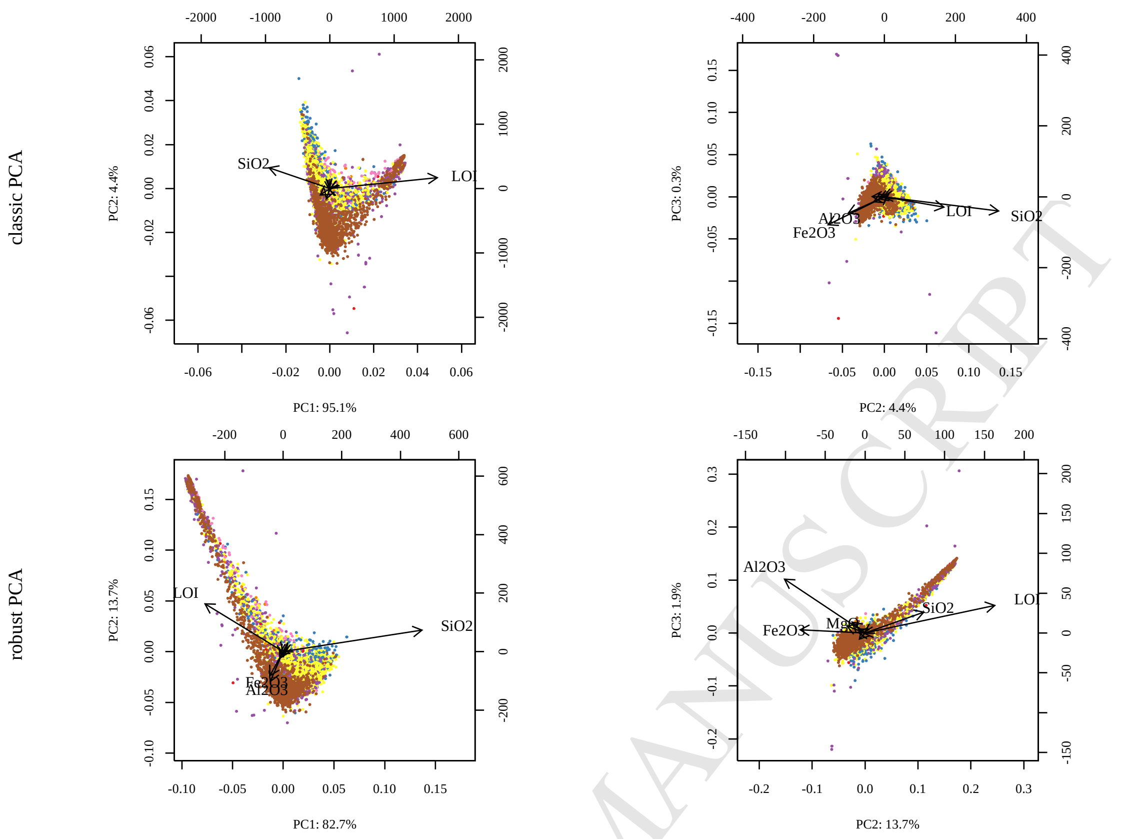

Figure 1 Comparison of biplots of the first 3 classical and robust principal components obtained for the raw composition of major components, with an indication of the proportion of explained variance.

The biplots of the first three PCs (Figure 1), both from a classical perspective and with a robust approach, show a clear dominance of LOI, SiO_2, Al_2O_3 and Fe_2O_3 on the geochemical variability within the subcomposition of major components. This is related to the well-known *proportionality effect*, an (undesirable) positive correlation between the mean and the variance of a positive variable (Figure 2). In both the classical and robust approaches, PC1 compares LOI to SiO_2, while PC2 compares these two against Al_2O_3 and Fe_2O_3. No clear structure can be distinguished in PC3. With regard to the data cloud, a clear cut V-shape appears in all cases, with the two sides of the V roughly orthogonal to the two most dominant variables in each diagram: a composition would have to have negative SiO_2 or LOI to fall beyond these alignments.

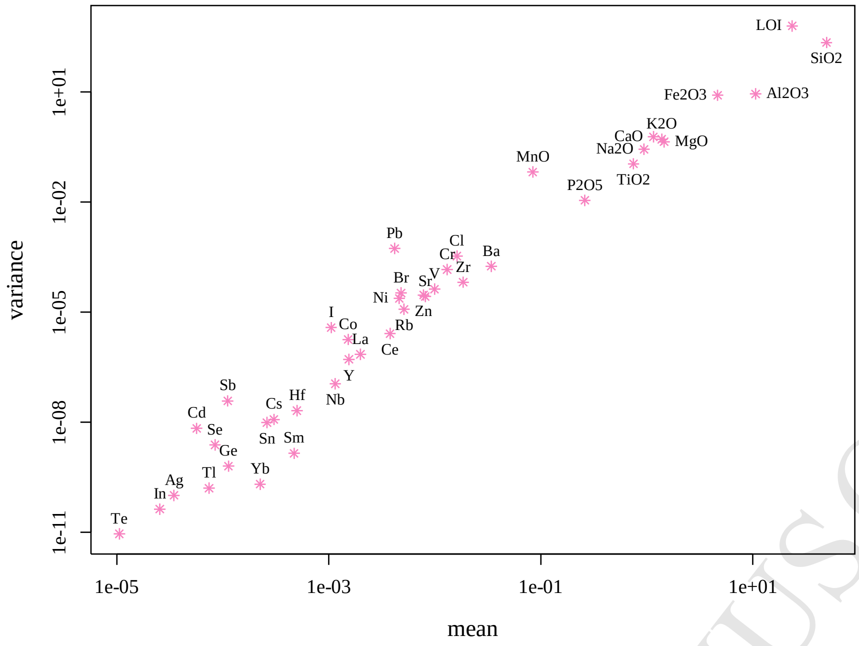

Figure 2 Relationship between the classical mean and variance of raw components.

3.2. PCA of standardized data

The proportionality effect can be removed by standardizing each variable. Figure 3 shows the resulting biplots of the two first classical and robust PCs. The most striking difference with the raw biplots is that almost all variables involved show up now with significant contributions to the biplot. In the classical case, PC1 is a sort of contrast between LOI and the rest of the major oxides, while PC2 is a contrast between CaO-Fe_2O_3-TiO_2-MgO and SiO_2-K_2O-Na_2O. Looking at the data cloud, and in particular to certain subsets of data, it appears that samples from peat-covered areas tend to fall along the link between LOI and SiO_2. Quite clear differences can also be seen between the positions of samples on basalts against those from acidic magmatic rocks (granites, granodiorites and acidic volcanic materials): basalts concentrate along the axis formed by CaO-Fe_2O_3-TiO_2-MgO, while acidic materials (especially granites and granodiorites) rather fall near the tips of SiO_2-K_2O-Na_2O. Other rocks of typical acidic fingerprint (psammites and metapelites, arenite-rich siliciclastics) fall as well in positions near to the axes SiO_2-K_2O-Na_2O and LOI.

The biplot of a robust PCA shows a very similar picture, although this time PC1 appears to be dominated by SiO_2 against the rest of major oxides; no relevant differences exist between the robust and classical versions of PC2. In effect, a similar preferential distribution of data samples according to groups can be observed between the two sets of biplots, namely with peat concentrating on positive values of PC1 (high LOI), basalts on negative values of PC2 (high CaO, Fe_2O_3, TiO_2) and granites, granodiorites, psammites, metapellites and siliciclastics on positive PC2 (Na_2O, K_2O, SiO_2).

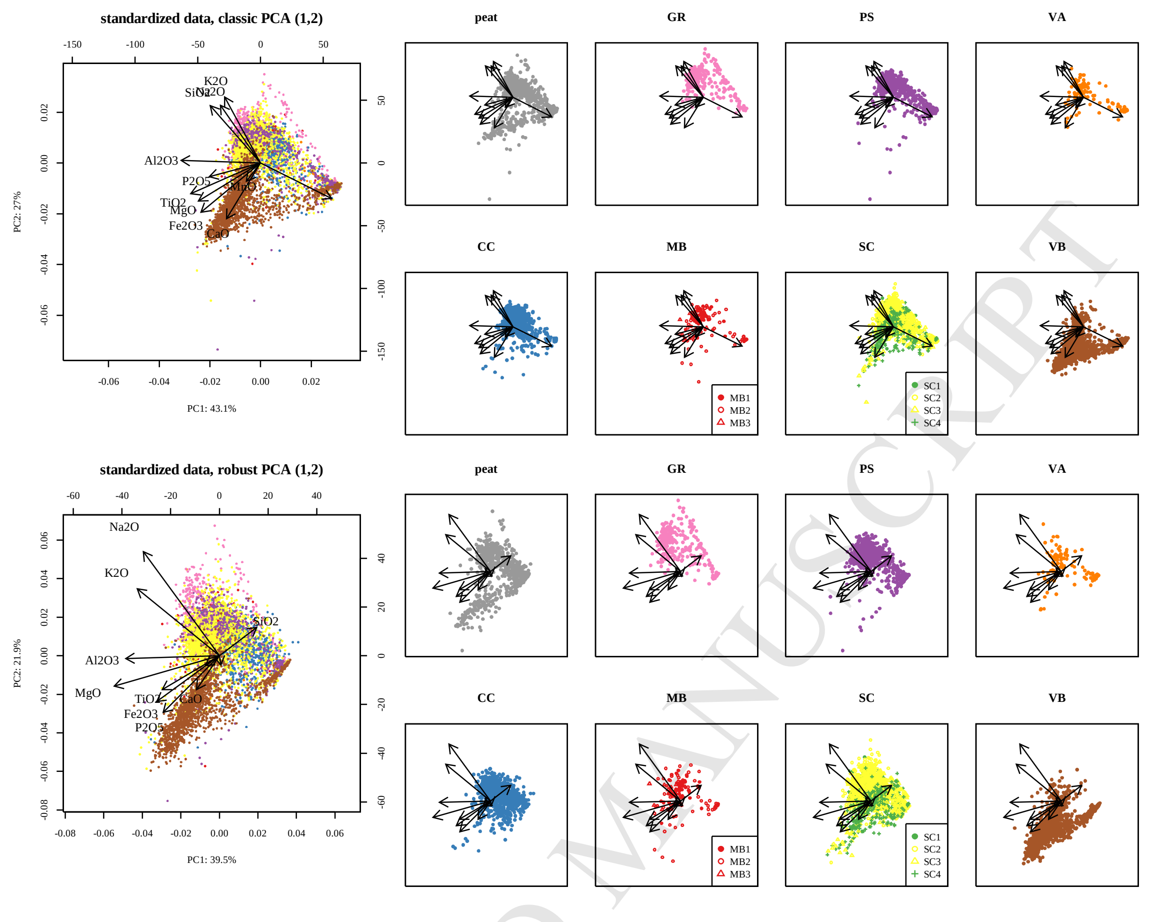

Figure 3 Biplots of the first two classical and robust principal components (left diagrams) for the standardized major components, with 8 parallel plots for several groups of samples: peat covered areas; CC=limestone and calcareous landscapes; GR = acidic magmatic rocks (granites and granodiorites); MB = basic magmatic rocks; PS = psammites and metapelites; SC = siliciclastic rocks; VA = acidic volcanic rocks; VB = basic volcanic rocks (basalts). Extended legend and color versions of these figures in the online supplementary material.

Biplots of standardized components can be extended to include (standardized) trace elements. These are shown in Figure 4. In this case, no significant difference can be observed between the robust and classical PCA results (except an irrelevant mirroring of the PC1 axis), and actually both diagrams are pretty much expansions of those of Figure 3 by including more elements following known (Rollinson, 1993) geochemical associations: volatiles (Cl, Br) with LOI, “mafic” elements (Ni, Co, Cr, V, Zn, Y, Yb) with TiO_2, MgO and Fe_2O_3, and “felsic” elements (Rb, Ba, Cs, Sm, Ce, La) with Na_2O and K_2O. CaO (and Sr) appears related to the mafic component and strongly controlled by basalt compositional variation.

It is as well worth mentioning that the standardization process did not remove the boundary effects related to negative components: in the biplots of Figure 3 one can still observe clear cut alignments of samples orthogonal to certain arrows and in the direction opposite to the arrow (TiO_2, MgO or Fe_2O_3 in granites; Na_2O or LOI in basalts; K_2O in limestones). One of the inconveniences of boundary effects is that the resulting cloud of dots does not bring further insights beyond those provided by a ternary diagram or a scatterplot of the variables affected. This can be seen by comparing the preceding biplots (Figure 1 and Figure 3) with the color ternary diagrams of the online supplementary materials (SM2).

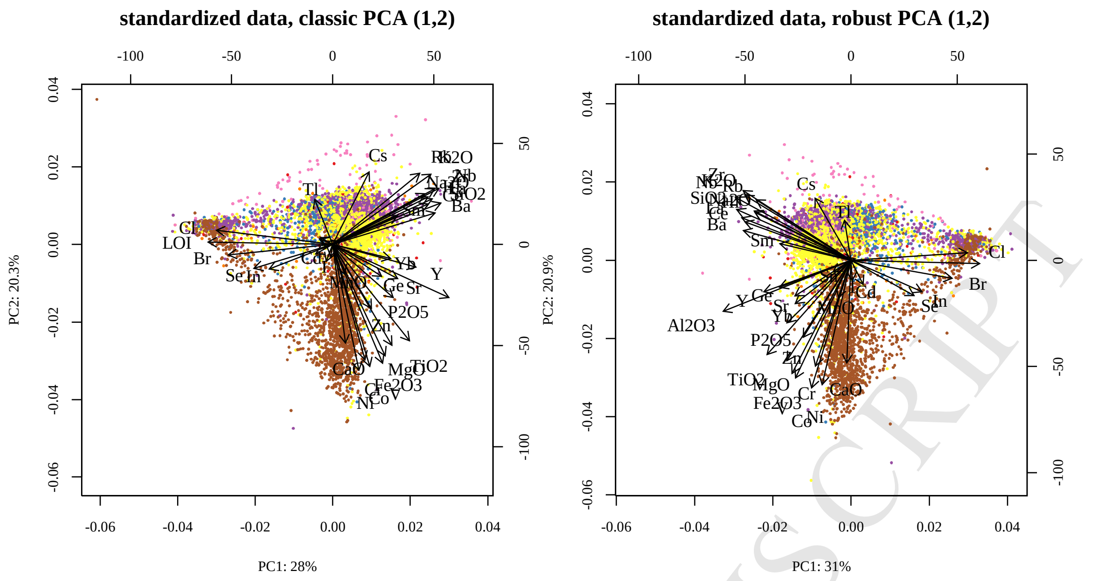

Figure 4 Biplots of the first two classical (left) and robust (right) principal components, for the standardized composition of major components and trace elements. Color versions of these figures available as online supplementary material.

3.3. PCA of log-ratio transformed data

Applying a log-ratio transformation to the compositional data set has the advantage of removing simultaneously the proportionality effect and the boundary effects, as can be seen in Figure 5. In the case of a biplot based on clr-transformed data, one should look for groups of variables falling together (e.g. K_2O, Na_2O, SiO_2) and sets of collinear variables (CaO, MnO, Fe_2O_3, LOI). These patterns can be seen on the biplots of both classical and robust PCA. Nearby arrows (K_2O, Na_2O, SiO_2) suggest highly proportional variables: if two variables are proportional, their log-ratio should be quasi constant, i.e. show a very low variance. The normalized log-ratio of SiO_2 to Na_2O shows a variance of 0.09, while the normalized log-ratio of K_2O to Na_2O is 0.07. Whether these are "small enough" values can be judged by comparing these with other variance values of the same data set: for instance, PC1 and PC2 show variances of resp. 1.69 and 0.99 (Figure 6), i.e. between 10 and 24 larger that the log-ratios between Na_2O, K_2O or SiO_2. If several arrows are collinear, the variability within the subcomposition formed by these variables is highly concentrated along one single direction, i.e. one single process (or constant combination of several processes) acts on those components. This is the case of the subcomposition [CaO, MnO, Fe_2O_3, LOI]. That "single process" is approximately described by the first PC of the clr-data recomputed within that subcomposition (Table 1, Ternary diagrams SM3 in online supplementary materials),

$$\mathrm{subPC1} = 0.754\ln(\mathrm{LOI}) - 0.603\ln(\mathrm{MgO}) - 0.244\ln(\mathrm{Fe_2O_3}) + 0.093\ln(\mathrm{CaO}).$$

Looking at the coefficients of this subcompositional PC1, the histograms of the scores of PC1, subPC1 and the normalized log-ratio of MnO to LOI (Figure 6) makes it evident that these three quantities are very similar, namely

$$\mathrm{PC1} \approx \ \mathrm{subPC1} \approx 0.707\ln(\mathrm{LOI}) - 0.707\ln(\mathrm{MnO}) = \frac{1}{\sqrt{2}}\ln\frac{\mathrm{LOI}}{\mathrm{MnO}}.$$

Interestingly, the link between LOI and MnO is roughly horizontal (Figure 5), i.e. quasi-parallel to the classical PC1, and even quite similar to the robust PC2. As shown in Figure 6, high values of that log-ratio tend to appear in peat covered areas. A similar relationship can be found between the classical PC2, the robust PC1 and the log-ratio

$$\mathrm{PC2(classic)} \approx \mathrm{PC1(robust)} \approx \frac{1}{\sqrt{2}} \ln \frac{\mathrm{Na_2O}}{\mathrm{CaO}},$$

which appears to be controlled by the lithology, basalts show negative values and more quartz-rich rocks (granites and granodiorites, psammites and metapelites, arenite rich siliciclastic materials) show positive values.

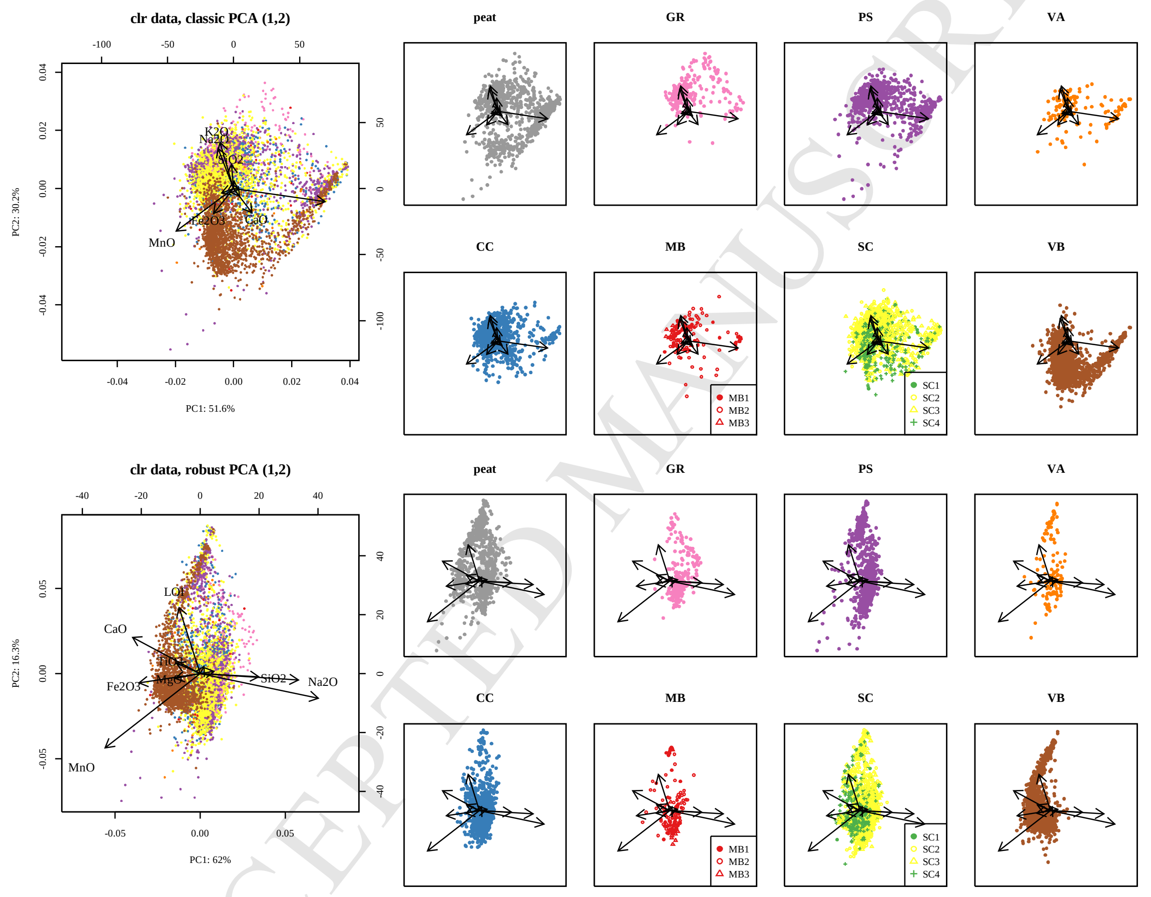

Figure 5 Biplots of the first two classical and robust principal components (left diagrams) for the clr-transformed major components, with 8 parallel plots for several groups of samples: peat covered areas; CC=limestone and calcareous landscapes; GR = acidic magmatic rocks (granites and granodiorites); MB = basic magmatic rocks; PS = psammites and metapelites; SC = siliciclastic rocks; VA = acidic volcanic rocks; VB = basic volcanic rocks (basalts). Extended legend and color versions of these figures in the online supplementary material.

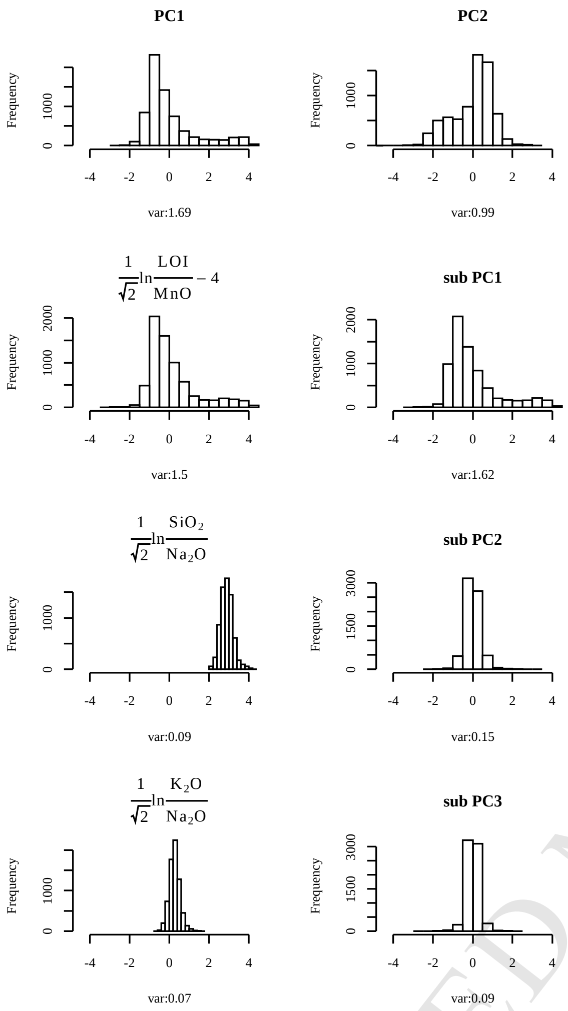

Figure 6 Histograms of the scores of the first two PCs, of the three PCs within the (quasi one-dimensional) subcomposition [LOI, MnO, Fe_2O_3, CaO] and of the normalized log-ratios of SiO_2 and K_2O to Na_2O, all on the same scale and with indication of variance of each case.

Table 1 Loadings (eigenvectors) and variances (eigenvalues) of a classical clr-PCA within the subcomposition [CaO, MnO, Fe_2O_3, LOI].

	PC1	PC2	PC3
LOI	0.754	0.425	0.037
MnO	-0.603	0.444	0.435
Fe_2O_3	-0.244	-0.084	-0.827
CaO	0.093	-0.785	0.355
variance	1.622	0.148	0.092
	87 %	8%	5%

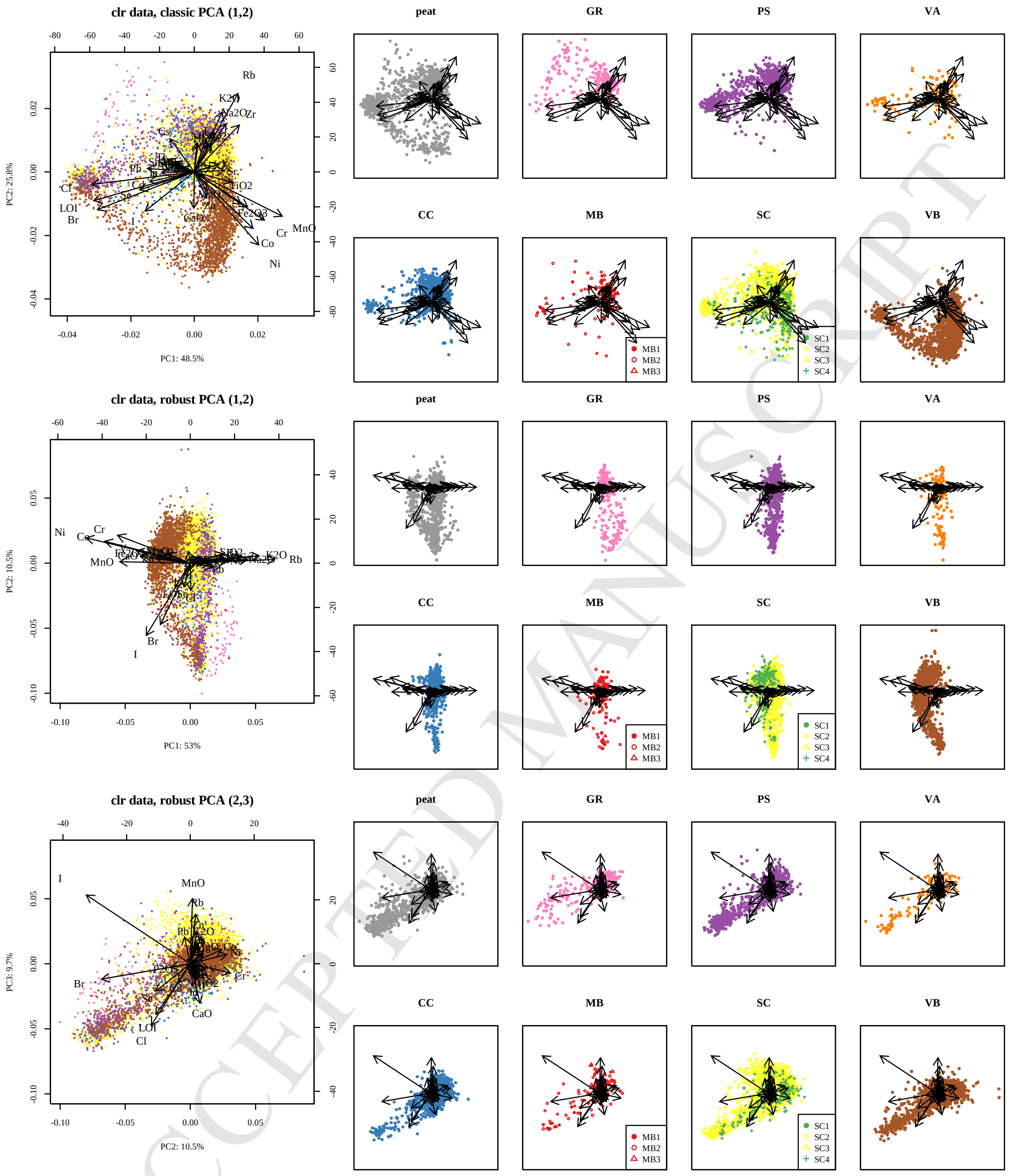

Figure 7 Biplots of the classical and robust principal components (left diagrams, PC1-PC2 for the classical PCA, PC1-PC2-PC3 for the robust PCA) for the clr-transformed whole composition, with 8 parallel plots for several groups of samples: peat covered areas; CC=limestone and calcareous landscapes; GR = acidic magmatic rocks (granites and granodiorites); MB = basic magmatic rocks; PS = psammites and metapelites; SC = siliciclastic rocks; VA = acidic volcanic rocks; VB = basic volcanic rocks (basalts). Extended legend and color versions of these figures in the online supplementary material.

3.4. PCA of log-ratio transformed major and trace elements

Having seen the potential of using the centered log-ratio transformation for PCA of major components, we will proceed to study simultaneously all variables included in Figure 2, which span six orders of magnitude. The same clr transformation can be used, without treating the trace elements separately from the major components. Figure 7 summarizes the biplots obtained with classical and robust PCA, and shows similar patterns to the biplots of major components. In the classical case, PC1 appears to be controlled by the presence/absence of peat coverage and PC2 by lithology, while in the robust case these switch roles and PC1 follows lithology while PC2 indicates peat coverage. Moreover, we clearly see three groups of variables,

- group A: Rb, K_2O, Na_2O, SiO_2, Zr, La, Ba, Ce, Nb, Hf
- group B: Ni, MnO, Co, Cr, Fe_2O_3, V, MnO, Zn, TiO_2
- group C: Cl, LOI, Br, I, Se, Cd

each group formed by relatively long arrows, consistently pointing towards diverging directions. The robust PC1 (and the classical PC2) can be seen as a contrast between components of groups A and B, while robust PC2 (and the classical PC1) is dominated by the elements of group C. When such configuration is found, a ternary diagram of the subcomposition formed by the longest arrow of each group usually describes the whole variability exceptionally well. This is the case of the subcomposition [Br, Ni, Rb] (Figure 8), showing an increasing peat influence with increasing Br mg/kg, and a lithological influence along the log-ratio of Rb mg/kg to Ni mg/kg.

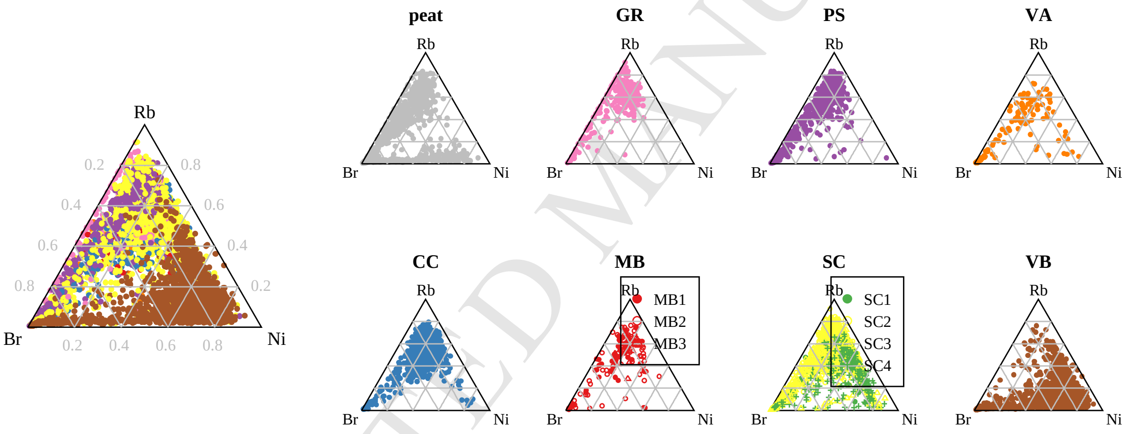

Figure 8 Ternary diagrams of the subcomposition [Br, Ni, Rb], for the whole data set and for subsamples. Compare with the biplots of PC1 and PC2 on Figure 7.

Another useful characteristic of clr biplots is the identification of linear dependence with parallelism: the existence of two parallel links suggests that the log-ratios of the two pairs of variables may be highly correlated; two orthogonal links suggest as well a lack of correlation between the corresponding log-ratios. The links between pairs Br-Ni, Cl-Cr, Sb-TiO2 and Cd-MnO are quite parallel in all diagrams, and correspondingly, the log-ratios of these pairs are highly correlated (Figure 9); the same can be said of Rb-Zn, Cs-I and Cr-Ni. On the other hand, Rb-Zn and Br-Ni are orthogonal in the biplots, and their log-ratios show a quasi-zero correlation (p-value=0.03). Moreover, the 4 log-ratios of the first group appear to increase towards peat-rich areas, while those of the second log-ratio seem more related to lithology.

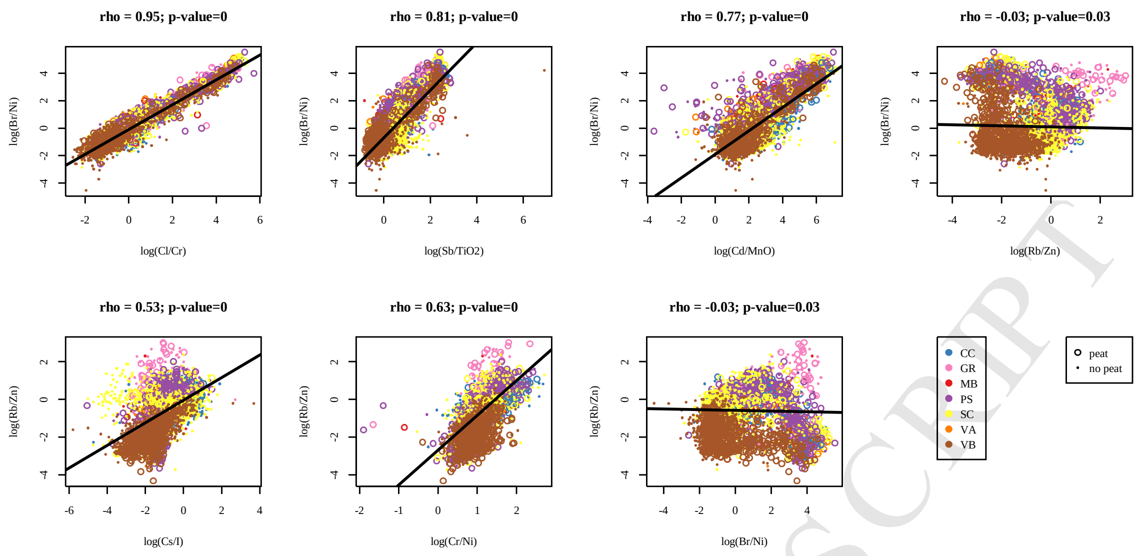

Figure 9 Scatterplots of some log-ratios of the composition of major and trace elements, selected because of showing parallel or orthogonal links in the biplots on Figure 7 and Figure 8. Color versions available in the online version and supplementary materials. Correlation coefficients and their associated p-value for the hypothesis of null correlation are included.

4. Discussion

4.1. The potential of exploratory tools

In multivariate analysis, we can consider an individual observation containing D variables as a point in the D-dimensional real space R^D. Then, a biplot is just a projection of the whole cloud of data points (and of the D axes corresponding to each variable) onto a certain plane: in particular, PCA biplots are chosen to maximize the variance of the projected data cloud. Along this idea, Harker diagrams are also projections onto a certain plane, chosen to filter out (ignore, remove) all information regarding the other $D-2$ variables. Thus, if we do not have strong prior knowledge about the probable most important projections, exploring a data set would require visualizing $D(D-1)/2$ Harker diagrams, or just a handful of biplots. For the data set considered here ($D=41$), this could have required visualizing 820 Harker diagrams, instead of the 11 biplots of Figs. 1, 3, 5 and 8, to have a first look at a data set without being guided (and potentially biased) by prior conceptions.

The analysis of these biplots suggest that most of the variability is controlled by two factors, lithology and peat coverage, with a relative importance that depends on the exact setting of the analysis: the principal components clearly associated with peat coverage represent between 10% and 50% of the total variability (depending on whether a robust or a non-robust approach is taken); those clearly linked to lithology have a contribution to variability between 25%-50%.

That peat coverage exerts such a generalized and strong influence on the soil geochemistry is not a surprise. Besides LOI (associated with water and organic matter), peat coverage is associated with relative enrichments in Cl and Br (but not in I), and to a second order with Se (but not in heavy metal contaminants), as shown in the robust biplots of PC2 and PC3 (Figure 7). It has been suggested that peat bogs fed solely by atmospheric deposition (ombrotrophic) act as archives for many types of atmospheric constituents (Shotyk 1996). Depletion of Cr, Cu, Ni and V in peat covered areas in Northern Ireland, as found by McIlwaine et al. (2014) has been linked to biogeochemical cycling of potentially harmful elements within peat (Novak et al. 2011).

It is to be expected that the influence of lithology is as important as peat coverage. It is more striking that this influence can be summarized in one single principal component, mainly a contrast between mafic- and felsic-related elements, particularly clearly seen in the robust PC1 of Figure 7 explaining >50% of the robust compositional variability of 41 elements.

In addition, a clr biplot can be used to extract some interesting ternary diagrams (Br-Ni-Rb, CaO-MnO-Fe_2O_3-LOI) and log-ratio scatter plots, showing either good correlation (e.g., Br/Ni, Cl/Cr, Sb/TiO_2 and Cd/MnO) or lack of it (e.g., Rb/Zn vs. Br/Ni). These ratios highlight again the interplay between variable peat coverage and a lithology changing between felsic-dominated and mafic-dominated.

4.2. Comparison of raw vs. log-ratio approaches

Data transformations play a determinant role in the structures that biplots can show. A blindfold application of conventional software will most often produce a PCA on standardized data, or equivalently, a PCA of the correlation matrix (Figure 3). A “no transformation” option can be produced if the software allows treating the covariance matrix of raw data, although the information that can be extracted from it is extremely limited (Figure 1). This is because often the variance (and covariances) of a positive variable scales with its mean value (Figure 2). Thus, the co-dependence structure between elements will be obscured by their abundances. One could say that standardization cannot be avoided if the relations between major and trace elements are sought. However, this also removes the contrast between highly varying variables (i.e. strongly controlled by processes acting within the studied area) and roughly constant ones (i.e. stable with respect to those processes), as all elements are scaled by the standardization to have a variance of one unit. For instance, we would lose the fact that, having the same mean, Pb is 10 times more variable than Ce as shown in Figure 2.

On the other hand, in the preceding sections we have reproduced some of the arguments in favour of applying the log-ratio framework of Aitchison (1986) and co-workers. If the analyst considers them relevant, the centered log-ratio transformation should be applied prior to the calculations of PCA. This transformation will remove the proportionality effect, allowing for a covariance-based PCA without losing the different, relative variances of each element.

In the application presented, the exploratory power of the resulting clr biplots was clearly superior to that of raw or standardized biplots. Nothing relevant could be seen from raw biplots, which were strongly affected by the proportionality effect. In the standardized biplots with only major elements, good relations between Na-K and between Fe-Mn-Ti, and an orthogonality of these two groups were apparent. Extending the analysis to include trace elements provided a picture where these elements associate following known rules of geochemical affinities, but no further insights can be drawn from them. The biplot (Figure 4) shows a clear triangular structure where each observation can be identified as a mixture of three groups of elements: one related to LOI and volatiles, one to elements typically enriched in mafic rocks and one of elements associated with felsic rocks.

This structure of three groups of elements is reproduced again in the biplots generated with the clr transformed data (Figure 5 and Figure 7), though the associated triangular shape of the cloud of samples is not easily recognized. Actually, the clr biplots do not show so clearly the boundary effects of the preceding biplots. Instead, two groups can be seen: one dominated by samples from peat-covered areas, and the other by samples from areas free of peat. Within the (mostly) peat-free samples, no clear groupings can be seen between the several lithologies, rather a continuum of compositions along the felsic/mafic axis is present (represented by a log-ratio Rb/Ni). The biplots by lithologies show converging “peatification” paths from soils that entirely reproduce the background lithology to soils which do not keep any trace of this background. This global picture as conveyed by the multi-elemental biplots can be summarized as well in the ternary diagram Ni-Rb-Br. Thus, the %Br in this subcomposition might be taken as a sort of peat penetration index.

According to the authors' experience clr biplots and standardized biplots almost always resemble each other, especially for geochemical data sets with low to moderate variability. However, the real power of clr biplots lies in the ability to suggest particular bivariate log-ratio scatterplots and ternary diagrams where high correlation or absence of it can be observed, and might be worth modelling and interpreting. This ability is very limited for standardized biplots. In other words, the authors always prefer clr biplots. Nevertheless, producing one or two biplots for the two transformations is nowadays an easy task. Sceptical readers are invited to explore their data sets with both approaches and compare results.

4.3. The role of robust analyses

An analysis is said to be robust if it is not sensitive to the presence of a small to moderate number (less than 50%) of contaminated or erroneous samples. The robust methods used in this contribution detect which samples are most probably non-contaminated and derive estimates of mean values, variances and covariances from this subset of the data. In the case study presented here, it is appealing to consider "peatification" as one such contamination process, in which case the non-contaminated samples could be seen as reflections of the lithological background. To assess the appropriateness of this interpretation, the weights allocated to each sample by a MCD covariance estimator were compared with the information available on peat coverage. Table 2 summarizes the absolute and relative frequencies of samples from peat-covered areas that are considered contaminated (weight 0, outlier) or uncontaminated (weight 1, regular observation) by a robust PCA analysis. The same information is graphically conveyed in mosaic plot SM4 of the online supplementary materials. A robust analysis of the clr-transformed data set considers 80% of the 931 samples from peat regions to be anomalous. The 40% of the samples from peat-free areas that are considered anomalous as well might be related to other processes different from peat forming mechanisms. Note that an analysis based on the standardized data does not deliver such a clear cut picture of the association between zero weights and peat coverage: actually, one would even say that the weighting is rather independent of the peat coverage in that case.

Table 2 Absolute frequency of co-occurrence of peat coverage with weights of 0 or 1 in a robust analysis of the data set, in the cases of clr-transformed data and standardized data. Percentages are given by rows.

	peat	robustness weights			
	coverage	0	1	0 (%)	1 (%)
clr	0	2371	3620	39,6	60,4
	1	739	192	79,4	20,6
standardized	0	2166	3825	36,2	63,8
	1	417	514	44,8	55,2

In the case study presented here, the preceding considerations imply that the robust PCA would focus on lithology controlled variations. Note as well that outliers (samples with weight 0) are very apparent in the robust biplots (Figure 10). Since they are not involved in the choice of the projection, they tend to show very marked patterns, while regular samples concentrate on a more or less circular cloud in the middle of the diagram. Figure 10 (central and right columns) display outliers (quite well corresponding to peat samples in Figure 7) which appear to follow some sort of mixture law between the non-contaminated part of the data set and a hypothetical pure peat composition (e.g. thick peat cover), which would be very rich in LOI, Br and Cl. This concentrates on PC2 and PC3, given that PC1 is mostly a lithological signal. This indicates the potential to develop these results further to provide an improved method for estimating the extent of high carbon peat areas and organic-rich soils using LOI, Br and Cl rather than based solely on LOI. Analysed Br, V and U from dry peat cores, have been suggested as useful indicators of peat accumulation rate (Davis and Wickham 1987). Chagué-Goff and Fyfe, (1996) suggested that although Cl, I and Br were probably first deposited as salts from sea spray, they were subsequently incorporated within the organic fraction of the peat. Biester et al. (2004) showed that Cl, I and Br are transformed to organohalogens during peat decomposition, and are therefore retained.

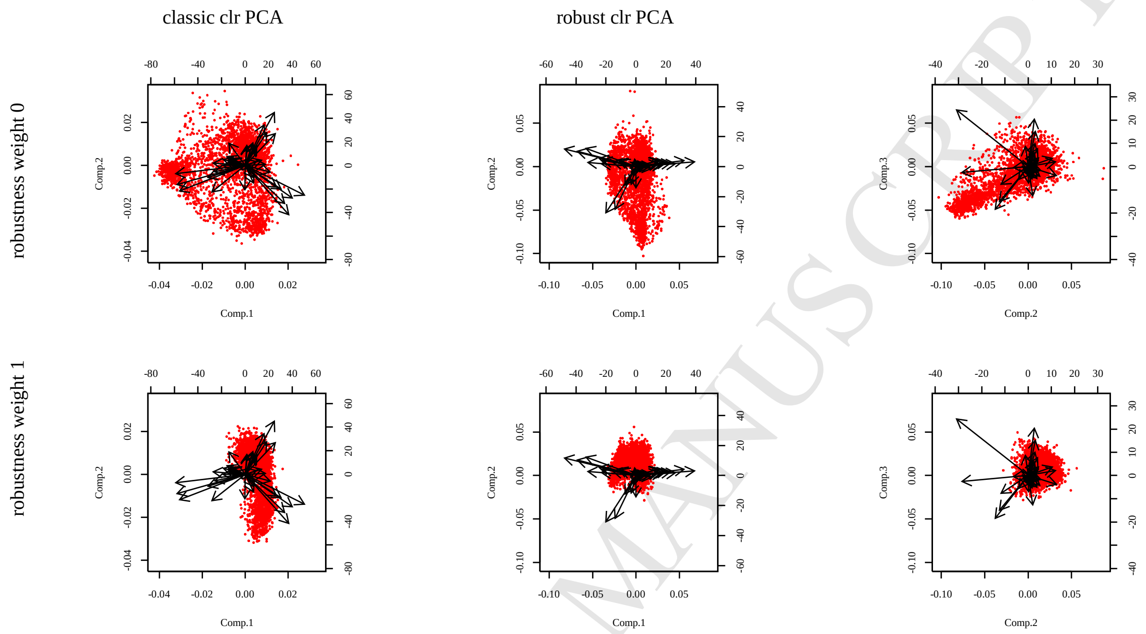

Figure 10 clr biplots of Figure 7, distinguishing between outliers and regular samples as classified by the robust PCA.

4.4. About zeroes and values below the detection limit

The application of log-ratio methods suffer from a fundamental problem when the data set has values below the detection limit (BDLs). Numerically, quite often these BDLs are replaced by zero or by a small value: this practice does not imply any problem for the raw and standardized approaches, but the logarithm of zero is $-\infty$, thus intractable. Conceptually, the log-ratio methodology is based on assuming that the relevant changes within the data set occur in terms of orders of magnitude (and not in absolute terms). Replacement by small numbers becomes extremely dangerous then, as the variability induced by the replaced values can be larger than the variability from the part of the data set that has been observed. The formal treatment of zeroes in log-ratio methods is a field of ongoing research, and it will not be discussed here. Interested readers can refer to Martín-Fernàndez, Barceló-Vidal and Pawlowsky-Glahn (2003), Palarea-Albadalejo, Martín-Fernàndez and Gómez-Garcia (2007), van den Boogaart and Tolosana-Delgado (2013, Ch. 7) or Palarea-Albaladejo and Martín-Fernández (2015). From a practical point of view, for the sake of an exploratory analysis, the following rule can be considered. For each element, if the number of zeros and BDLs is large (say, around 50% of the sample or more), then it is probably wiser to simply remove that variable; on the other hand, if the proportion of BDLs is smal, then these values can be replaced *by the detection limit itself* (van den Boogaart, Tolosana-Delgado and Bren, 2011); the resulting replaced sample should be preferentially analysed with robust methods. No hard threshold can be given on what is a “small proportion” of BDLs: practice, caution and common sense are always required when dealing with replaced data sets, no matter whether it is one single lost value or half of the data set.

5. Conclusions

The biplot, a data-driven graphical representation of the whole variability of a data set onto two dimensions, has been shown to be a powerful exploratory tool to understand the factors controlling the variability in a data set. Such biplots are constructed on the basis of a PCA (principal component analysis), either using the raw data, or else standardized or centered logratio (clr)-transformed data sets. Raw data are prone to both proportionality effects and boundary effects, and should be utterly avoided. Standardized data are only prone to boundary effects, while clr data are free from these distorting effects. Furthermore, insights from this study suggest that clr biplots and standardized biplots will be often quite similar, although clr biplots convey more information about the different spread of the variables considered and can be used to select simple, powerful scatterplots and ternary diagrams for exploratory analysis. In any case, both robust and classical methods could be used and results compared.

In the case study presented here, analysing clr-transformed data delivered far more powerful insights than analysing raw or standardized data. The robust PC1 can be taken as a lithological index, describing the contrast between lithologies of mafic affinity (rich in Ca, Fe, Ti, Mg, Ni, Co, Cr; basalts, gabbro and certain siliciclastic materials) and those of felsic affinity (rich in Si, K, Na, Al, Rb, Zr, La, Cs, Ba; all other lithologies). Unsurprisingly, this represents the most important control on the geochemical variability of the soils of the data set if the influence of peat is removed. Following insights from both the classical and robust clr PCA, the influence of peat might be quantifiable by the %Br within the subcomposition (Br, Rb, Ni). Finally, it was found that a robust analysis is quite capable of detecting and filtering out the samples from peat-covered areas (among other samples not from peat areas), as peat-affected samples are considered outliers by clr robust methods.

Acknowledgements

The Geological Survey of Northern Ireland (GSNI) is thanked for the use of the Tellus dataset. The Tellus Project was carried out by GSNI and funded by The Department for Enterprise, Trade and Investment (DETINI) and The Rural Development Programme through the Northern Ireland Programme for Building Sustainable Prosperity. The authors thank and acknowledge the participants of the first GeoMap Workshop (held in Olomouc, Czech Republic, 17-20 June 2014) from which this research stems. This included representatives from several geochemical surveys (Tellus Survey Northern Ireland, Tellus Border Survey, Republic of Ireland, the GEMAS project, North American Soil Geochemical Landscapes Project, and National Geochemical Survey of Australia).

References

Ahrens, L.H., 1954a. The lognormal distribution of the elements (a fundamental law of geochemistry and its subsidiary). Geochimica et Cosmochimica Acta 5 (2), 49-73

Ahrens, L.H., 1954b. The lognormal distribution of the elements (part 2). Geochimica et Cosmochimica Acta 6 (2-3), 121-131

Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology), 44, 139–177.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press).

Aitchison, J., 1997. The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn, V. (Ed.) Proceedings of IAMG'97 - The 3rd Annual Conference of the Int. Association for Mathematical Geology, International Center for Numerical Methods in Engineering (CIMNE), Barcelona (Spain)

Biester, H., Keppler, F., Putschew, A., Martinez-Cortizas, A., Petri, M., 2004. Halogen retention, organohalogens, and the role of organic matter decomposition on halogen enrichment in two Chilean peat bogs. Environmental Science and Technology 38, 1984–1991.

Butler, J.C., 1975. Occurrence of negative open variances in ternary diagrams. Mathematical Geology, 7 (1), 31-45

Butler, J.C., 1976. Principal components analysis using the hypothetical closed array. Mathematical Geology, 10 (2), 243-252

Butler, J.C., 1978. Visual bias in R-mode dendrograms due to the effect of closure. Mathematical Geology, 10 (2), 243-252

Butler, J.C., 1979. The effects of closure on the moments of a distribution. Mathematical Geology, 11 (1), 75-84

Caritat, P. de, Cooper, M., 2011a. National Geochemical Survey of Australia: The Geochemical Atlas of Australia. Geoscience Australia Record, 2011/20 (2 Volumes), 557.

Caritat, P. de, Cooper, M., 2011b. National Geochemical Survey of Australia: Data Quality Assessment. Geoscience Australia Record, 2011/21 (2 Volumes).

Chagué-Goff, C., Fyfe, W.S., 1996. Geochemical and petrographical characteristics of a domed bog, Nova Scotia: a modern analogue for temperate coal deposits. Organic Geochemistry 24 (2), 141–158.

Chayes, F., 1960. On correlation between variables of constant sum. Journal of Geophysical Research 65 (12), 4185–4193.

Chayes, F., Trochimczyk, J., 1978. An effect of closure on the structure of principal components. Mathematical Geology 10 (4), 323-333

Cruickshank, M.M., Tomlinson, R.W., Devine, P.M., Milne, R., 1998. Carbon in the vegetation and soils of Northern Ireland. Biology and Environment: Proceedings of the Royal Irish Academy 98B (1), 9-21.

Davis A.M., Wickham, S.M., 1987. The Microstratigraphy of Two Peat Sequences from Northeastern Newfoundland. Géographie physique et Quaternaire 41 (3), 355-364.

Drew, L.D., Grunsky, E.C., Sutphin, D.M., Woodruff, L.G., 2010. Multivariate analysis of the geochemistry and mineralogy of soils along two continental-scale transects in North America, Science of the Total Environment, 409, 218-227

Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika, 1 (3), 211-218

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

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

Filzmoser, P., Hron, K., Reimann, C., 2009. Principal component analysis for compositional data with outliers. Environmetrics 20, 621-632

Friske, P.W.B., Rencz, A.N., Ford, K.L., Kettles, I.M., Garrett, R.G., Grunsky, E.C., McNeil, R.J., Klassen, R.J., 2013. Overview of the Canadian component of the North American Soil Geochemical Landscapes Project with recommendations for acquiring soil geochemical data for environmental and human health risk assessments. Geochemistry: Exploration, Environment, Analysis 13, 267-283

Gabriel, K.R., 1971. The biplot graphical display of matrices with application to principal component analysis. Biometrika 58, 453-467

Graffelman, J., van Eeuwijk, F., 2005. Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research. Biometrical Journal 47 (6), 863-879.

Keaney, A., McKinley, J.M., Graham, C., Robinson, M., Ruffell, A., 2013 Spatial statistics to estimate peat thickness using airborne radiometric data. Spatial Statistics 5, 3-24

Martín-Fernández, J. A., Barceló-Vidal, C., Pawlowsky-Glahn, V., 2003. Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation Mathematical Geology 35 (3), 253-278

McIlwaine, R., Cox, S., Doherty, R., Palmer, S., Ofterdinger, U. McKinley, J.M., 2014. Comparison of methods used to calculate typical threshold values for potentially toxic elements in soil. Environmental Geochemistry and Health 36 (5), 953-971

Mitchell, W.I., 2004 The Geology of Northern Ireland- Our Natural Foundation. Geological Survey of Northern Ireland, Belfast (UK), 318p.

Novak, M., Zemanova, L., Voldrichova, P., Stepanova, M.,Adamova, M., Pacherova, P., Komarek, A., Krachler, M., Prechova, E. (2011). Experimental evidence for mobility/ immobility of metals in peat. Environmental Science & Technology 45, 7180–7187. http://www.ncbi.nlm.nih.gov/pubmed/21761934.

Palarea-Albaladejo, J., Martín-Fernández, J.A., Gómez-García, J.A., 2007. Parametric Approach for Dealing with Compositional Rounded Zeros. Mathematical Geology 39 (7), 625-645

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

Pawlowsky, V., 1984. On spurious spatial covariance between variables of constant sum. Science de la Terre Série Informatique 21, 107-113.

Pawlowsky-Glahn, V., Buccianti, A. (Eds.) 2011. Compositional Data Analysis, Theory and Applications. Wiley, Chichester (UK) 378p.

Pawlowsky-Glahn, V.; Egozcue, J.J.; Tolosana-Delgado, R. (2015) Modeling and Analysis of Compositional Data. Wiley, Chichester (UK). 247p.

Rawlins, B. G., Scheib, C., Tyler, A.N., Beamish D., 2012. Optimal mapping of terrestrial gamma dose rates using geological parent material and aerogeophysical survey data. J. Environ. Monit., 14, 3086

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), 2014a. Chemistry of Europe's agricultural soils – Part A: Methodology and interpretation of the GEMAS data set. Geologisches Jahrbuch (Reihe B 102 + DVD), Schweizerbarth, Hannover (Germany).

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), 2014b.

Chemistry of Europe's agricultural soils – Part B: General background information and further analysis of the GEMAS data set. Geologisches Jahrbuch (Reihe B 103), Schweizerbarth, Hannover (Germany).

Rollinson, H.R., 1993 Using Geochemical Data: Evaluation, Presentation, Interpretation. Taylor and Francis, New York (US).

Rousseeuw, P., van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212-223

Shotyk, W., 1996. Peat bog archives of atmospheric metal deposition: Geochemical evaluation of peat profiles, natural variations in metal concentrations, and metal enrichment factors. Environmental Reviews 4 (2), 149–183.

Smyth, D., 2007. Methods used in the Tellus Geochemical Mapping of Northern Ireland. OR/02/022, 2007.

van den Boogaart, K.G., Tolosana-Delgado, R., Bren, M., 2011. The compositional meaning of a detection limit. In: Egozcue, J. J.; Tolosana-Delgado, R., Ortego, M. I. (Eds.) Proceedings of the 4th International Workshop on Compositional Data Analysis (2011), CIMNE, Barcelona (Spain)

van den Boogaart, K.G., Tolosana-Delgado, R., 2013. Analyzing compositional data with R. Springer, Heidelberg (Germany) 258p.

Wang, X. and the CGB Sampling Team, 2015. China geochemical baselines: Sampling methodology. Journal of Geochemical Exploration 148, 25-39

Young, M.E., Donald, A.E., 2013. A Guide to the Tellus Data. Belfast: Geological Survey of Northern Ireland, Belfast (UK).

List of Figures

List of Tables

Highlights:

- we get and interpret a multivariate exploratory analysis for a geochemical survey
- we compare classical and robust covariances of raw, standardized and log-ratio data
- log-ratio robust principal component (rPC) analysis offers the most insights
- background rock geochemistry controls rPC1, simplified as logratios Na/Ca or Rb/Ni
- peat coverage is related to rPC2, to Br/(Br+Ni+Rb) and to the robustness weights