The single component geochemical map: Fact or fiction?

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

Published in:
Journal of Geochemical Exploration

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2015, Elsevier B.V.
Licensed under the Creative Commons Attribution -NonCommercial-NoDerivs License (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits distribution and reproduction for non-commercial purposes, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access
This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: http://go.qub.ac.uk/oa-feedback

Download date:01. Jun. 2024

The Single Component Geochemical Map: Fact or Fiction?

Jennifer M McKinley*[1], Karel Hron[2], Eric Grunsky[3], Clemens Reimann[4], Patrice de Caritat[5,6], Peter Filzmoser[7], Karl Gerald van den Boogaart[8], Raimon Tolosana-Delgado[8]

[1]School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, BT7 1NN, UK

[2]Dept. of Math. Anal. & Appl. of Math., Palacky University Olomouc 17. listopadu 12, CZ-771 46 Olomouc, Czech Rep.

[3] Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1

[4]Geological Survey of Norway, PO Box 6315 Sluppen, N-7491 Trondheim, Norway

[5]Geoscience Australia, GPO Box 378, Canberra ACT 2601, Australia

[6]Research School of Earth Sciences, Australian National University, Canberra ACT 2601, Australia

[7]Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria

[8] Helmholtz Center Dresden Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Freiberg, Germany

* Corresponding author: j.mckinley@qub.ac.uk; hronk@seznam.cz ; egrunsky@gmail.com ; Clemens.Reimann@NGU.NO ; Patrice.deCaritat@ga.gov.au ; P.Filzmoser@tuwien.ac.at ; r.tolosana@hzdr.de ; boogaart@hzdr.de

ABSTRACT

1. Single component geochemical maps are the most basic representation of spatial elemental distributions and commonly used in environmental and exploration geochemistry. However, the compositional nature of geochemical data imposes several limitations on how the data should be presented. The problems relate to the constant sum problem (closure), and the inherently multivariate relative information conveyed by compositional data. Well known is, for instance, the tendency of all heavy metals to show lower values in soils with significant contributions of diluting elements (e.g., the quartz dilution effect); or the contrary effect, apparent enrichment in many elements due to removal of potassium during weathering. The validity of classical single component maps is thus investigated, and reasonable alternatives that honour the compositional character of geochemical concentrations are presented. The first recommended such method relies on knowledge-driven log-ratios, chosen to highlight certain geochemical relations or to filter known artefacts (e.g. dilution with SiO_2 or volatiles). This is similar to the classical normalisation approach to a single element. The second approach uses the (so called) log-contrasts, that employ suitable statistical methods (such as classification techniques, regression analysis, principal component analysis, clustering of variables, etc.) to extract potentially interesting geochemical summaries. The caution from this work is that if a compositional approach is not used, it becomes difficult to guarantee that any identified pattern, trend or anomaly is not an artefact of the constant sum constraint. In summary the authors recommend a chain of enquiry that involves searching for the appropriate statistical method that can answer the required geological or geochemical question whilst maintaining the integrity of the compositional nature of the data. The required log-ratio transformations should be applied

followed by the chosen statistical method. Interpreting the results may require a closer working relationship between statisticians, data analysts and geochemists.

Key words: soil geochemistry, compositional data analysis, log-ratios, mapping.

INTRODUCTION

Over the last decade the focus of digital soil and sediment databases at a regional, national, transnational and continent scale has increasingly become to provide information for a range of purposes including geological and soil mapping, baseline quality documentation, mineral prospecting, land and soil resource assessment, risk evaluation, environmental and educational purposes, and prediction of soil provenance for forensic purposes (Smith et al., 1997; Reimann et al., 1998, 2003; Morris et al., 2003; Lagacherie et al., 2007; McKinley, 2013). These digital soil databases are generally accompanied by geochemical atlases (e.g. Caritat and Cooper, 2011a; Young and Donald, 2013; Reimann et al., 2014a,b) showing a collection of distribution maps for individual geochemical elements. This has been matched by a corresponding increase in published studies utilising these soil geochemical surveys (e.g. Chiprés et al., 2008; Grunsky et al., 2009; Carranza, 2010; Ohta et al., 2011; Caritat and Grunsky, 2013; Cheng et al., 2014; Lancianese and Dinelli, 2014; Birke et al., 2015). The resolution of the ground-based sampling scheme used for the generation of these databases is the best compromise between the extent of the region covered, and time and resources available. Over a local to regional scale, soil sampling can be managed on a 2 km^2 grid as applied by the Geological Survey of Northern Ireland (GSNI)'s Tellus project. A regional scale dataset such as this is close to exhaustive sampling in terms of geological survey mapping, reducing the need for interpolation. It is worth noting however that this is not exhaustive sampling for mineral exploration. Generally surveys are carried out at local (1:10,000), regional (1:250,000) or continental (1:1,000,000) scales. A number of authors discuss the scale of geochemical mapping (e.g. Bølviken et al., 1992; Darnley et al., 1995; Reimann et al., 2010).

The sampling strategy generally follows standard protocols (e.g. UNESCO's IGCP Global Geochemical Database – Darnley et al., 1995; G-BASE standard established by BGS - Johnson 2005; FOREGS field handbook – Salminen et al., 1998; GEMAS field handbook – EGS, 2008; North American Soil Geochemical Landscapes Project – Smith et al. 2011; National Geochemical Survey of Australia field manual – Lech et al., 2007; Canadian component of the North American Soil Geochemical Landscapes Project – Friske et al., 2013; China Geochemical Baseline Project – Wang et al., 2015).

Geochemical survey data are typically represented and interpreted using single element geochemical maps. The interpretability and validity of these single components have repeatedly been challenged because they are prone to several artefacts: spurious negative bias on correlations (Chayes, 1960), dependence of interpretation on other (potentially non-reported) components (Aitchison, 1986), dependence on units (e.g. mass, molar), and dependence on processes acting on some components (e.g. weathering, dilution) but influencing all of them (van den Boogaart and Tolosana-Delgado, 2013). All the issues mentioned are due to the fact that geochemical data constitute amounts of components with relative portions of a total even if this total is unknown. The components may be reported in different physical units (ppm, mg/kg or as percentages) and all the components may not be reported or measured. However, each component has an amount which represents its importance as part of the whole composition. The constraints of constant sum or the closed nature of the relative amounts of components have implications for the analysis of geochemical data. In statistics and mathematical geosciences, powerful solutions to deal with these issues have been developed in a field known as Compositional Data Analysis (CoDA) (Aitchison, 1986; Grunsky, 2010; Pawlowsky-Glahn and Buccianti, 2011; van den Boogaart and Tolosana-Delgado, 2013; Templ et al., 2011). This paper investigates the question of what the compositional nature of geochemical data means in regional geochemical mapping, specifically with regard to single element (univariate) distribution maps.

The collaborative research presented stems from the first GeoMap Workshop (held in Olomouc, Czech Republic, 17-20 June 2014) that discussed the consequences of these challenges and the usefulness of CoDA for regional geochemistry. Present were representatives from regional geochemical surveys (Tellus Survey covering Northern Ireland and Tellus Border Survey covering the border counties of the Republic of Ireland (Young and Donald, 2013)), the GEMAS project covering a large part of Europe (Reimann et al. 2014a,b), the soil geochemical survey of the conterminous USA (Smith et al., 2011; Drew et al., 2010) and the continental scale National Geochemical Survey of Australia (Caritat and Cooper, 2011a,b). This paper, while acknowledging the historically important role of single component maps, aims to: (1) discuss their appropriateness, (2) provide some examples to highlight the problems raised above, and (3) offer some solutions to present interpretable maps free of the issues arising from the compositional nature of geochemical data.

PROBLEMS RELATED TO SINGLE COMPONENT MAPS

Geochemical surveys generate datasets with several tens of components (between 50 and up to 70 elements are commonly reported), obtained from different sample materials (soil horizons, size fractions, vegetal tissues, sediments, water, etc.) and with different analytical techniques (total analyses, partial or selective digests). For regional geochemistry, the key applications of the data are generally either to produce and use elemental concentration maps (i.e. one-component regional distribution maps) or to explore associations between elements affected by geological/geochemical processes, which can also be mapped (e.g. principal components). For the first of these tasks, standard practice has included producing a single component map thought to represent the raw or “absolute” input data in the form of dot (or point) maps, but also as interpolated maps of these raw concentrations. Reimann (2005) defines the purpose of such geochemical maps as ‘to display different processes in a map form and to detect local deviations from the dominant process in any one sub-area’. The problems discussed in this section result from the closure property of geochemical compositions, i.e. the unavoidable fact that samples are *aliquots* of the geological bodies we want to investigate and therefore do not really convey information about the element mass distributions (Aitchison, 1986). This has implications for baseline quality mapping, mineral prospecting, land and soil resource assessment or risk evaluation, though these issues are beyond the scope of this contribution. Nevertheless, in the following we show that the traditional meaning of closure effect (linked to closed data with unit, or any other fixed sum constraint of components, and the resulting distortion of the correlation structure) should be considered also in a broader sense. Namely, compositional data are primarily observations that contain quantitatively expressed relative contributions of parts on a whole. From this perspective, the unit sum constraint is just a representation, obtained without altering the source information, conveyed by ratios between the components. Therefore, even with a variable sum of geochemical concentrations, resulting, e.g., from designed omitting of some components, one should be aware that the relative nature of data is still present and needs to be taken into account by proper statistical processing, as exemplified below by CoDA (Pawlowsky-Glahn et al., 2015).

Point maps – the data “as is”

It is often thought that raw one-component maps report “what is there”, that they report a sort of “objective ground truth”. However, Reimann (2005) and others (Reimann and Filzmoser, 2000; Reimann and Garrett, 2005; Reimann et al., 2005; Reimann et al., 2008) highlight that to give sense to that set of spatially dispersed values, space-dependent geochemical processes must be interpreted, and that these are highlighted with a proper representation. Reimann (2005) discusses the advantages and disadvantages of such different representations. Reimann (2005) concludes that, actually, the most important issue becomes the scaling chosen to define the points (colour, size and symbol). Splitting the data into groups (classes) on the basis of order statistics in exploratory data analysis (such as the quartiles of boxplots, or other percentiles), Reimann suggests, may shed light on the spatial structure that reflects at least a number of these processes in a map. Figure 1 shows how impactful this choice of symbol/colour scaling can be with a real dataset. Obviously, equidistant colour scales do not necessarily yield the most easily interpretable maps, neither in raw (or “absolute”) nor in logarithmic values. The proportion of the entire study area actually measured by the dataset is largely exaggerated by the size of the dots. The conclusion is that it may be more appropriate to use quantile-based intervals (following the guidelines on scaling as discussed in Reimann, 2005) to present geochemical data to give an initial assessment of the distribution of elemental concentrations. Readers are referred to Reimann (2005) for further details on methods used to establish intervals in data scaling. It should be pointed out that this choice of scaling does not address the problems resulting from the relative character of elemental concentrations, an aspect that will be explored later.

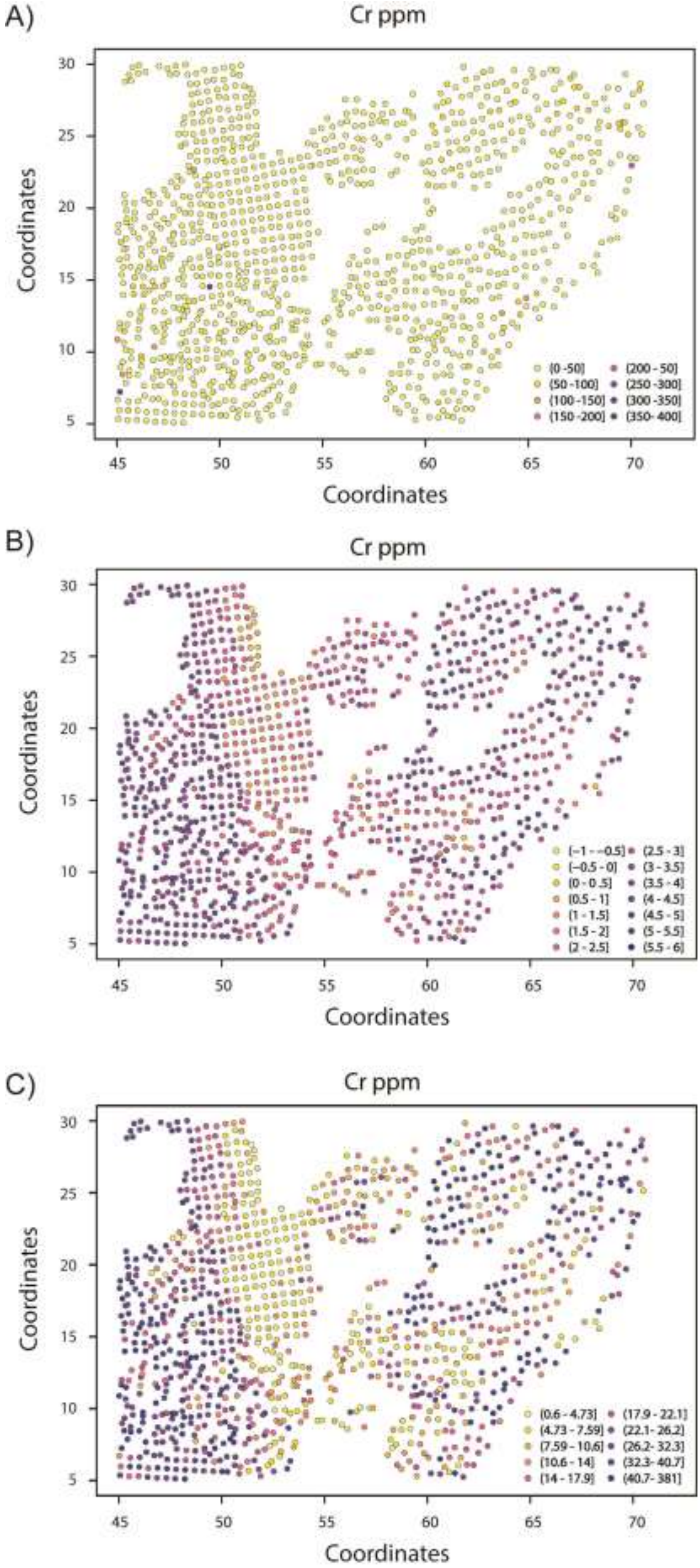
A)
Cr ppm
Coordinates
Coordinates
30
25
20
15
10
5
45
50
55
60
65
70
[0 -50]
(50 -100]
(100 -150]
(150 -200]
(200 - 50]
(250 -300]
(300 -350]
(350- 400]
B)
Cr ppm
Coordinates
Coordinates
30
25
20
15
10
5
45
50
55
60
65
70
[−1 - −0.5]
(−0.5 - 0]
(0 - 0 .5]
(0.5 - 1]
(1 - 1.5]
(1.5 - 2]
(2 - 2.5]
(2.5 - 3]
(3 - 3.5]
(3.5 - 4]
(4 - 4.5]
(4.5 - 5]
(5 - 5.5]
(5.5 - 6]
C)
Cr ppm
Coordinates
Coordinates
30
25
20
15
10
5
45
50
55
60
65
70
[0.6 - 4.73]
(4.73 - 7.59]
(7.59 – 10.6]
(10.6 - 14]
(14 - 17.9]
(17.9 - 22.1]
(22.1- 26.2]
(26.2- 32.3]
(32.3- 40.7]
(40.7- 381]

Figure 1: Three representations of Cr ppm concentrations from the GEMAS European soil dataset (Reimann et al. 2014a,b), changing the dependence of the colour scale: (A) using equidistant classes in the raw value scale, (B) using equidistant classes in the log-transformed value scale, and (C) using quantile based (equal probability) classes in the raw value scale. Note that the units are latitude and longitude coordinates but the orientation of the image has been changed to highlight changes that are dependent on the colour scale used.

Geospatial continuity – the interpolated map

It is becoming increasingly common for geochemical atlases (online and published printed versions; Reimann et al., 2003, 2014a,b; Young and Donald, 2013) and research publications to produce smoothed (colour) surface maps because most end-users have come to expect a "nice looking" output map. Essential to the generation of an interpolated map, is the demonstration of a spatially coherent pattern through the generation of semi-variograms and other available geostatistical tools. The issues relating to spatial interpolation are well documented elsewhere (e.g. Li and Heap, 2011) and are distinct from the question of the compositional justification for single component maps, which is the focus here. Nevertheless, it should be noted that when using raw 'point data' the interpolation method should be chosen to ensure several aspects relevant to compositional data, in as much as the interpolated values should (1) remain in the real positive space, and (2) not sum up to more than 100%. Thus interpolated maps should not be obtained separately component by component, as this cannot guarantee adherence to CoDA constraints (as outlined earlier).

Joint interpolation methods able to cope with these constraints should be preferred (for example Tolosana-Delgado and van den Boogaart, 2013).

Do compositions actually represent absolute or relative abundances?

The central question of this paper is whether a single component map can support an accurate interpretation on its own. As mentioned before, it is often thought that these geochemical maps do actually represent absolute abundances. However, geochemical data are just a particular case of compositional data, and compositional variables carry only relative information of a part to a whole. This can be seen in the conventional concentration units, such as mg/kg, being essentially a constant multiplied by a dimensionless unit: 1 mg/kg = 1 mg/1,000,000 mg, or 10^{-6} mg/mg, or simply 10^{-6}. At every sampled or interpolated point, a concentration value provides information only on the relative weight of one particular element to the total (Tolosana-Delgado & van den Boogaart, 2013), regardless of whether we know all, some or none of the other parts of the whole. This is illustrated in Figure 2 where a lower chromium (Cr) east-west zone is clearly seen in the quantile scaling plot (Fig. 2a). However a corresponding increase in silica oxide (SiO_2) in the same area, (as shown in Fig.2b) shows that any sensible interpretation must consider some sort of possible dilution by SiO_2 that may be related to a change in geology or soil type for example . This is a consequence of the constant sum (closure) effect (Chayes, 1960): if one component increases (due to the addition of a pollutant or a diluent) the mass of all other components remains unaffected, and their apparent proportions (concentrations) must decrease. One solution to this problem is to use element ratios, rather than raw concentrations, because ratios are free of any such effect.

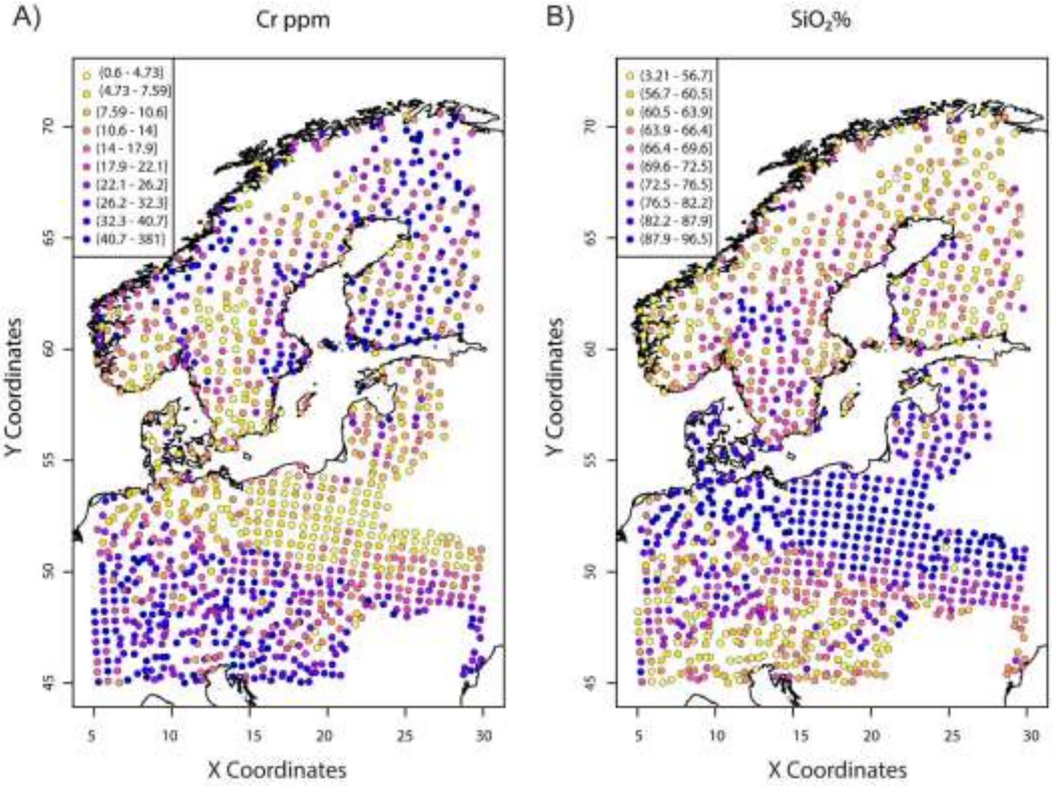

Figure 2: Illustrative maps from the GEMAS European soil dataset (Reimann et al. 2014a,b): A) (left) map of Cr ppm, B) (right) map of SiO_2 %, showing that the low Cr ppm central band is due to dilution with SiO_2 %. The units are latitude and longitude coordinates.

Despite this clear limitation, individual raw geochemical maps are deemed fundamental to several applications of geochemical maps, environmental assessments for instance. Thus potential toxicity to humans is typically given as a threshold or a range above which some kind of action is required by law. For example, the Canadian Council for Ministers of the Environment (CCME, 2014) states that arsenic (As) is considered to be potentially hazardous to humans if it exceeds 12-50 mg/kg in soils. The range in tolerance is based on whether the soils are in residential or industrial sites. In Europe a large variety of soil guideline values for agricultural and grazing land soils exists (see Appendix A in Reimann et al., 2014b).

Another way of looking at the relative or absolute scale is to consider a simple process affecting a material. Let us imagine a fresh rock with 20% potassium oxide (K_2O) and its in-situ weathered product having 10% of this oxide: what is relevant, that it lost 10% of K_2O or half of it? What would be the most reasonable percentage of K_2O in the weathered material if that original concentration would have been 12%? The correct answer is 6%, and not 2%. These considerations are nevertheless often clearer for trace elements: if a sample has 20 ppm of element A and another sample, genetically derived from the first, has 10 ppm of it, half of element A has been removed in the process of forming the second sample. The relative scale effect applies mostly for components with lower concentrations, e.g., for trace elements.

Compositional data are always multivariate

In the preceding example where the distribution of Cr was mapped (Figs. 1 and 2) it became evident that any interpretation of the variation in Cr should include an awareness of the dilution effect of other elements and should be done on the basis of ratios such as Cr/ SiO_2. The component increase considered above and its effect on all other components may appear in many other one-component geochemical maps as (high or low) outlier(s) or anomalous value(s). The issue is that geochemical components form a multivariate system that should be analysed as a whole, not component-wise. The following examples from the regional Tellus Survey (Young and Donald, 2013) highlight this aspect. Figure 3 shows the north east region of Northern Ireland where Dalradian metasediments are covered with

Palaeogene flood basalts. The basalt bedrock shows high nickel (Ni) concentrations (Fig. 3a) with a low Ni area in the uppermost north east corner related to metasedimentary bedrock. However, lower Ni values are also found in areas where the basaltic bedrock is covered by peat rather than conventional soils (Fig. 3b). A similar effect is seen for zirconium (Zr), though it is enriched in the metasedimentary rocks (Fig. 3c). In areas of consistent bedrock, these low areas of both Ni and Zr are dilution signals formed by high volatiles (represented by Loss On Ignition, LOI %) in blanket peat areas. The variation in geochemical maps can therefore be influenced by dilution, enrichment or depletion in other elements, rather than just by the processes specific to the element mapped. We will show later that maps based on (log)-ratios can avoid this weakness.

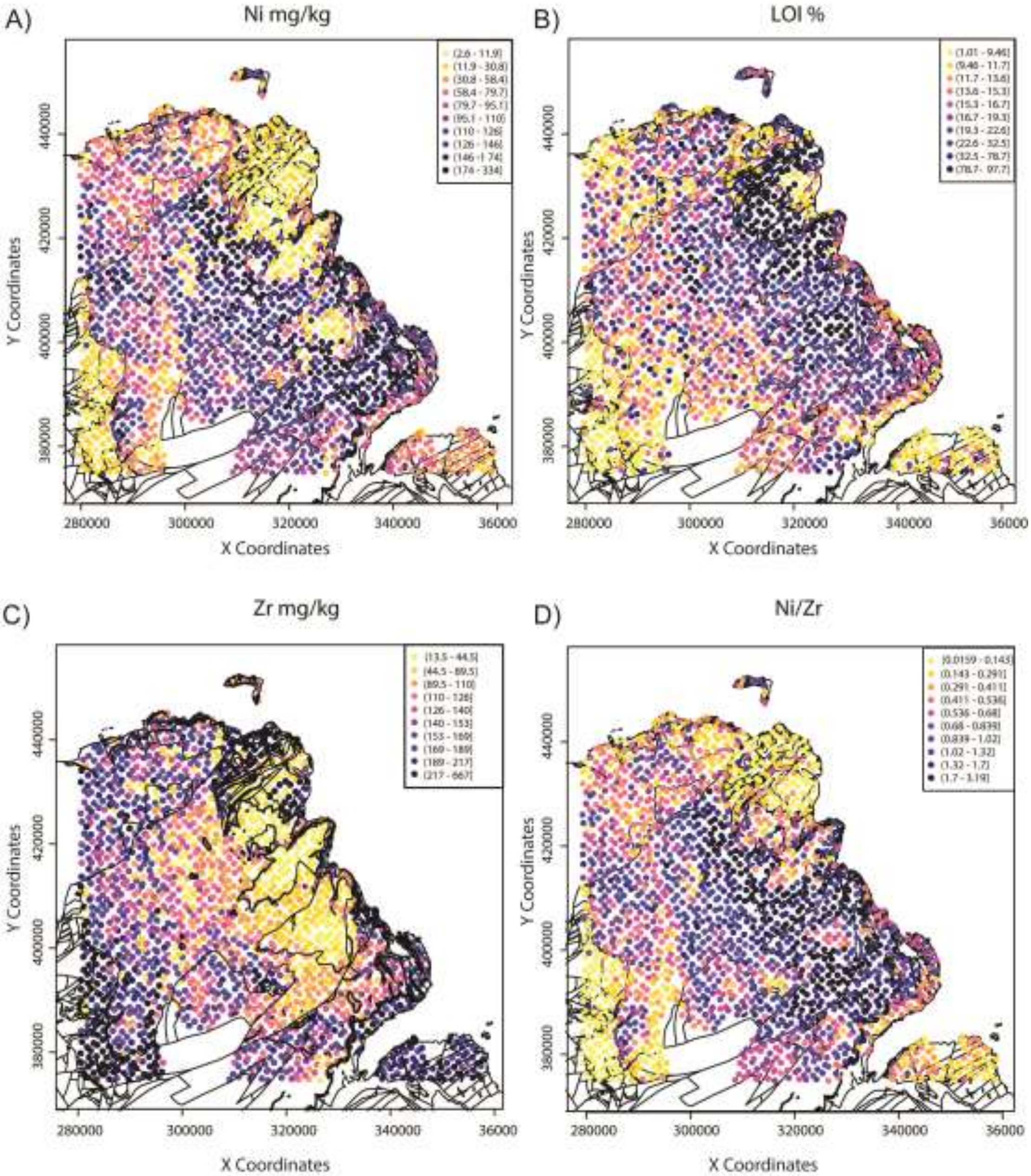

Figure 3: Four map representations using the Tellus soil survey (Young and Donald 2013) showing the north east region of Northern Ireland: A) Ni mg/kg, B) Loss on Ignition (LOI %), C) Zr mg/kg and D) Ni/Zr. The units are coordinates in the Irish Transverse Mercator Grid (EPSG:29903).

BASICS OF COMPOSITIONAL ANALYSIS

Generalities

A common way to treat the compositional nature of geochemical data has been the application of log-ratio techniques (Aitchison, 1986; Filzmoser and Hron, 2011; Pawlowsky-Glahn and Buccianti, 2011; Reimann et al., 2012; van den Boogaart and Tolosana-Delgado, 2013), due to the fact that the ratios between the components are unaffected by constant sum closure effects related to the relative nature of data (Egozcue and Pawlowsky-Glahn, 2011).

It is important to note that inherent compositional data properties apply to major, minor and trace elements, so there is no reason to distinguish them methodologically, although the effects of closure and other compositional properties (such as the relative scale) depend on abundance and the number of analysed components. In particular, the closure effect can be expected to be higher with a lower number of components and for the major components, while the relative scale will apply mostly in the case of trace elements (Pawlowsky-Glahn and Buccianti, 2011; Egozcue and Pawlowsky-Glahn, 2011; van den Boogaart and Tolosana-Delgado, 2013). Indeed, the closure effect results from distortion of the covariance structure of closed data (Aitchison, 1986), where the sum of covariances of an element to the other components is equal to the negative value of its variance; accordingly, for the extreme case of two-part (closed) compositions the correlation would always equal -1.

Advancing towards the conclusions of this paper, the compositional alternative to single component maps will be maps of certain well-chosen log-ratios of two or several components. The next subsections explain the basics of these transformations, followed by the geochemical and statistical rationale behind each proposed strategy.

The log-ratio transformations

Standard logarithmic transformations are not sufficient because it is the ratio to another element that deals with the closure restraint. The use of one to one ratios does eliminate the initial closure effect but ratio correlations are known to be subject to restrictions or "spurious" correlations. Several families of log-ratio transformations exist in the literature. Aitchison (1986) introduced the pairwise log-ratio transformation (pwlr), the additive log-ratio transformation (alr) and the centred log-ratio transformation (clr), along with other transformations which are not relevant for geochemical purposes. Later, Egozcue et al. (2003) proposed the isometric log-ratio (ilr) transformation. None is inherently better than the other, each has advantages and disadvantages. The following explanations assume that the composition $\mathbf{x} = [x_1, x_2, \dots, x_D]$ involves *D* elements (with boldface indicating a vector and the square brackets enclosing the concentrations x of elements *1, 2, … ,D*).

The pairwise log-ratio transformation (pwlr) takes *all* possible pairs of elements and computes their log-ratios.

$$\mathrm{pwlr}(\mathbf{x}) = \begin{bmatrix} 0 & \ln\frac{x_1}{x_2} & \cdots & \ln\frac{x_1}{x_D} \\ \ln\frac{x_2}{x_1} & 0 & & \ln\frac{x_2}{x_D} \\ \vdots & & \ddots & \vdots \\ \ln\frac{x_D}{x_1} & \ln\frac{x_D}{x_2} & \cdots & 0 \end{bmatrix} = \left[\ln\frac{x_i}{x_j}\right] = [\xi_{ij}]. \tag{1}$$

The pwlr contains all the elemental information in compositional data. The treatment of below detection values and zeros in compositional data is beyond the scope of this paper but readers are directed to Pawlowsky-Glahn and Buccianti (2011), van den Boogaart and Tolosana-Delgado (2013) and Filzmoser and Hron (2011) for further information on this subject. This baseline can be used to construct more sophisticated log-ratios, as will be shown later. The pwlr has the advantage that each transformed coefficient is quite simple and typically interpretable on its own. Its main inconvenience is the large number of resulting coefficients, and their high redundancy. For instance even after keeping only one ratio of each pair of inverse ratios $\ln(x_i/x_j)$ and $\ln(x_j/x_i)$, for *D* = 40 elements, we still need to consider $D \cdot (D-1)/2 = 780$ pairwise log-ratios. With regard to redundancy, it can be seen that any row or column of that matrix can be obtained from additions and subtractions from the elements of any other row or column, e.g. $\xi_{ij} = \xi_{ik} + \xi_{kj}$ for any three components i, j, k. Thus, for certain tasks it will be sufficient to take one of the rows or columns of the pwlr.

The pwlr allows us also to introduce an alternative to correlation to be used as a criterion of statistical association, given the spurious character of raw Pearson correlation coefficients in a compositional context. This is called the variation matrix, i.e. the variance of each of the entries of the pwlr matrix: $t_{ij} = \mathrm{Var}\left[\xi_{ij}\right]$. The variation is small for pairs of components which are highly proportional, and it increases as the two components are less and less dependent on each other.

The additive log-ratio transformation (alr) takes just one of the rows or columns of the pwlr (and removes the constant zero), typically the last column:

$$\mathrm{alr}(\mathbf{x}) = \left[\ln\frac{x_1}{x_D} \quad \ln\frac{x_2}{x_D} \quad \cdots \quad \ln\frac{x_{D-1}}{x_D}\right] = [\xi_{iD}],$$

here arranged in a row to save space. Its advantages are the low number of log-ratios capturing all the information, and the simplicity of its interpretation. Its disadvantage is the fact that some results might depend on the chosen common denominator x_D, and therefore this must be chosen carefully. The choice usually depends on geochemical context (e.g., Thomas and Aitchison, 2005), but subjectivity cannot be completely avoided. Another consideration is that the denominator should be chosen with a minimum number of, and preferably no, censored values. This ensures that the log-ratio represents a realistic range of values without interference of instrumentation or limits in the analytical protocols. The alr transformation has as well certain geometric inconveniences (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2006) not relevant for this contribution.

Another commonly used transformation is the centred log-ratio (clr) transformation. Instead of choosing a common denominator, this log-ratio transformation represents each element component as a ratio to a central value:

$$\mathrm{clr}(\mathbf{x}) = \left[\ln\frac{x_1}{g(\mathbf{x})} \quad \ln\frac{x_2}{g(\mathbf{x})} \quad \cdots \quad \ln\frac{x_D}{g(\mathbf{x})}\right],$$

corresponding to the geometric mean of *all considered* components,

$$g(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i} = \exp\left(\frac{1}{D}\sum_{i=1}^{D} \ln x_i\right).$$

The somewhat artificial role of the geometric mean becomes clear if we consider that each clr variable represents nothing more than the average of all pairwise log-ratios with one of the elements, i.e. the averaged sum of a column (or a row) of the pwlr matrix,

$$\ln\frac{x_i}{g(\mathbf{x})} = \frac{1}{D}(\xi_{i1} + \xi_{i2} + \cdots + \xi_{iD}).$$

Note that this averaged sum of all pairwise log-ratios contains the totality of the variability of the elemental composition with respect to the set of components that was considered. Its main advantage is that each transformed score can be related to one component, which might have some interpretability. However, the clr scores always depend on the geometric mean in the denominator, so they cannot be simply interpreted as the original element of interest independently of the other elements. Thus, just as for the individual components, any single component of the clr transform will depend on the notion of the "total" (meaning, the set of all components considered) and can be influenced by dilution, enrichment and depletion in these components. This is the greatest disadvantage of the clr transform. It should be noted that the clr has some desirable geometric characteristics not relevant for this contribution (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2006). It is important to note that the pwlr, alr and clr transforms do not form orthogonal bases. In the case of the alr transform, the resulting ratios may not reveal the same patterns when there is a change of denominator. For the clr transform, scatterplot matrices will have an intrinsic distortion that may or may not reveal patterns that are meaningful.

Log-contrasts and the isometric log-ratio transformation

All transformations mentioned before can be seen as vectors of some scores that are computed as *log-contrasts*, i.e. linear combination of components in log-scale:

$$\xi_i = \omega_{i1} \ln x_1 + \omega_{i2} \ln x_2 + \cdots + \omega_{iD} \ln x_D,$$

the coefficients of which are subject to $\omega_{i1} + \omega_{i2} + \cdots + \omega_{iD} = 0$. Log-contrasts form special one-dimensional "projections" of compositions. Their result is scale invariant, i.e. it does not depend on the particular units chosen for the components (mg/kg, ppm, proportions, etc.), and have many other desirable stability properties required by CoDA (Aitchison, 1986). Although log-contrasts seem to be quite complex mathematical objects, they have a clear geochemical interpretation in connection with chemical equilibrium. If we assume a reaction $aA + bB \rightleftharpoons cC + dD$, then the logarithm of the equilibrium constant is one of such log-contrasts, $\xi = \omega_A \ln A + \omega_B \ln B + \omega_C \ln C + \omega_D \ln D$, with $\omega_A = \omega_B = -1$ and $\omega_C = \omega_D = +1$, and it is directly proportional to the Gibbs Energy of the reaction $\Delta G \propto \ln K_{eq}$. Conversely, each log-contrast might be interpreted as a sort of reaction. That might not be directly meaningful for solid sample geochemistry datasets, but it still offers a way to interpret log-contrasts. Note that also the components of the above introduced alr and clr transformations can be seen as log-contrasts. Accordingly, for i-th alr variable ($i = 1, \dots, D-1$) we get

$$\ln \frac{x_i}{x_D} = \ln x_i - \ln x_D,$$

while i-th clr variable ($i = 1, \dots, D$) can be expressed as

$$\ln \frac{x_i}{g(\mathbf{x})} = \ln x_i - \frac{1}{D}(\ln x_1 + \cdots + \ln x_D).$$

The last log-ratio transformation mentioned is the (family of) ilr transformations (Egozcue et al., 2003). These are formed by $D-1$ log-contrasts which are computed using vectors of coefficients $\boldsymbol{\omega}_i = [\omega_{i1}, \omega_{i2}, \dots, \omega_{iD}]$ which are orthogonal to each other. From all possible ways of obtaining such orthogonal log-contrasts, the relevant ones for this contribution are: principal components of the clr-transformed data (Aitchison, 1982; 1986), and balances (Egozcue et al 2003; Egozcue and Pawlowsky-Glahn, 2005). Balances are simply (normalised) log-ratios of the geometric means of two groups of elements. One can obtain an ilr based on balances by choosing a binary hierarchy of association of elements (or a binary partition), such as the one represented in Figure 4. For instance the four log-ratios necessary to describe the behaviour of the subcomposition (Fe_2O_3, V, Cr, Co, Ni) in this figure are:

$$\xi_1 = \frac{1}{\sqrt{2}} \ln \frac{\mathrm{Fe2O3}}{\mathrm{V}}, \qquad \xi_2 = \frac{1}{\sqrt{2}} \ln \frac{\mathrm{Co}}{\mathrm{Ni}}, \qquad \xi_3 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{\mathrm{Cr}}{\sqrt[2]{\mathrm{Co} \cdot \mathrm{Ni}}},$$

$$\xi_4 = \frac{\sqrt{6}}{\sqrt{5}} \ln \frac{\sqrt[2]{\mathrm{Fe2O3} \cdot \mathrm{V}}}{\sqrt[3]{\mathrm{Co} \cdot \mathrm{Ni} \cdot \mathrm{Cr}}}$$

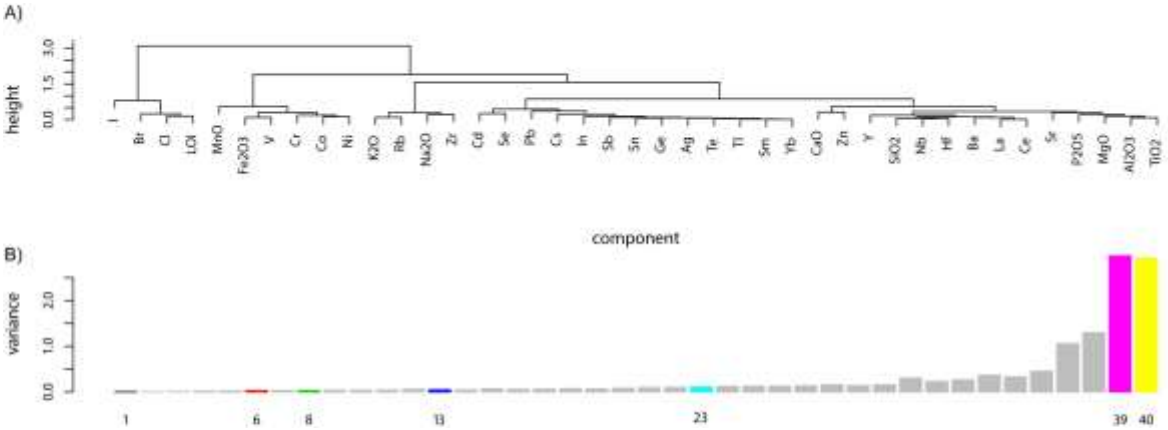

Figure 4: A) An ilr balance dendrogram and B) associated component variance obtained by choosing a binary hierarchy of association of elements (or a binary partition). The procedure is detailed in van den Boogaart and Tolosana-Delgado 2013).

In general applications of CoDA, the constants of proportionality are necessary to be able to compare the balances between them. However, in the case of mapping and given that we are going to use equal-probability intervals (Reimann, 2005), these constants are superfluous. Interested readers are referred to Pawlowsky-Glahn and Egozcue (2005) and Pawlowsky-Glahn and Buccianti (2011).

PROPOSED COMPOSITIONAL ALTERNATIVES

Univariate maps are actually produced in the hope that they might have a direct interpretation with respect to specific aspects or processes occurring in the studied region. However, we have argued that this may not be the case, because of the inherent multivariate nature of compositional data and the spurious variations induced by the closure: any process influencing a component will modify all other components (and their maps). As we cannot display or read maps displaying high-dimensional multivariate information, we are left with the question of how to choose interpretable univariate (many single element maps) or low variate (oligovariate) maps (fewer maps displaying two or more variables). All these maps will be representations of certain log-contrasts. The challenge in finding good maps is twofold: there are many more log-contrasts than there are elements in the dataset, and the most appropriate log-contrasts to use will always depend on the aim of the analysis undertaken (Pawlowsky-Glahn and Buccianti, 2011). Using single elements or log-contrasts inherently defines an approach using sub-compositions for gaining insight into geochemical processes. This may work for recognizing obvious processes, but not for processes that are under-sampled or “swamped” by dominant processes.

Three approaches for selecting good ratios for various aims are discussed and illustrated with a case study example. These provide alternative approaches to presenting a single elemental component map that honour the compositional nature of the geochemical data and offer interpretable mapped outputs. The feature(s) to be mapped can be based on either (1) a geochemical understanding of the processes under consideration (knowledge-driven log-ratios based on geochemistry), (2) established mathematical projections (e.g. pwlr or clr), or (3) based on the data using statistically determined projections that best capture certain effects, patterns and elemental associations (data-driven projections). These are illustrated using the regional soil sediment geochemical survey covering Northern

Ireland, generated as part of the Tellus Survey (Young and Donald, 2013). The soil sediment dataset used comprises 6862 observations of 19 geochemical variables (Fig. 5).

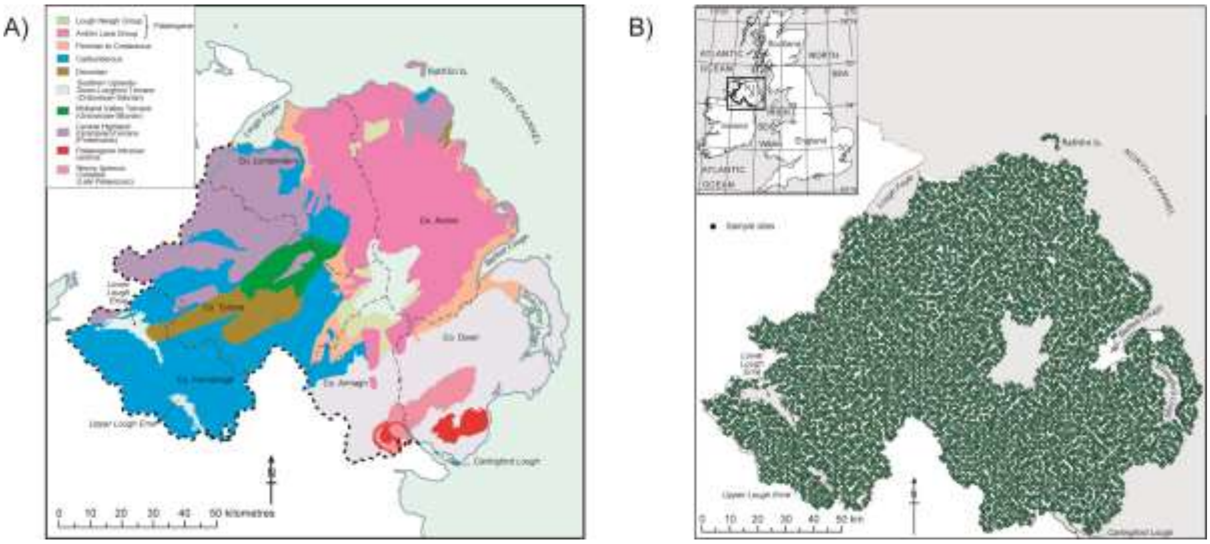

Figure 5: Case study A) Simplified geology for Northern Ireland (adapted after Mitchell, 2004) and; B) Sampling locations for the Tellus Geochemical Survey (GSNI).

1. Knowledge-driven (log)-ratios

In relation to the problem of Ni and Zr dilution by LOI (shown previously in Fig. 3), geological knowledge would suggest that it may be more useful to analyse a simple log-ratio, or a certain balance related to a reasonable equilibrium. Indeed, in this example using the (log-) ratio Ni/Zr reduces the masking effect of peat cover (LOI dilution effect) and illustrates a greater delimitation of the Palaeogene basalt-Dalradian metasediment boundary than any of the three single component maps (Fig. 3d). However, although being of possibly highest interpretational potential, ratio or log-ratio maps present one challenge: we are not used to reading them. This can be resolved by practice and it is advocated that the expected effect of less methodological artefacts promises this investment to be worth the time.

Of course, not everything relevant is always expressible in terms of (log-)ratios. In mining there is a need to predict excavation and processing costs per mass recovered, hence the ratio of value element versus mass to be mobilized is relevant, i.e. a quantity proportional to absolute concentration of the value element. In environmental studies, it is important to know how much mass of the toxic elements is present, i.e. proportional to its absolute concentration again. Thus the mass of the element per bulk mass can be a useful proxy. However, in both cases (and in many more examples), other quantities might highlight other important aspects of the problem, quantities which are often ratios or log-ratios. Mining requires the ratio of gangue elements against ore elements, because the product's quality and value depend on them. In environmental issues, well-chosen log-ratios may incorporate bioavailability effects (Barsby et al., 2012) and interactions with other components such as pH or HCO_3^-, etc. Environmental or health hazards typically depend both on absolute values and on ratios, particularly considering bioavailability fractions. Chemical reactions like those mentioned above also play an important role in mobilizing or trapping these elements. For soils and sediments that have been weathered or mixed/diluted with other material, no element shows the same percentages as in the background or source rock, but the ratios of elements unaffected by the mixing (or by pollution or weathering) are preserved. In weathering, especially of magmatic plagioclase-rich rocks, all existing measures of alteration have been defined as (log)ratios, perhaps because quantifying absolute losses of mass would be nonsensical (the removed mass was washed out of the rock-weathered product system, we do not know the starting concentrations of elements in the source rock, the degree of dilution generated from more or less mixing material with the weathered rock is unknown, etc.). Typical practice tends to be to use the ratio of mobile to immobile elements. In studies involving mineral reactions (igneous, metamorphic, diagenesis), Pearce

Element Ratios (PER) have been successfully used to model mineral paragenesis (Pearce, 1968; Madeisky and Stanley, 2010). In the molar ratio scatterplots used in PER, individual end-member minerals are represented as points and linear trends are interpretable in terms of mineral mixtures. This has been shown to be useful for instance to interpret alteration processes (e.g. Montreuil et al., 2013). Such PER can thus guide the choice of appropriate knowledge-driven log-ratios to map.

2. The pairwise log-ratio map

The second option is to rely on one of the standard log-ratio transformations, for instance the pairwise log-ratio (pwlr). The challenge in this case is the number of pairs that are possible to build: for D elements there are $D \cdot (D-1)/2$ unique maps. For example the aqua regia analyses from the GEMAS survey result in ~1600 ratio maps, and the XRF compositions add another ~400 ratio maps. It is impossible to present these within a classical atlas publication. Although this is becoming less of a problem with interactive online or electronic atlases, it is not humanly feasible to view and integrate all (or even just many) of these different perspectives. Therefore, either knowledge-driven or data-driven criteria will be required to choose which of this vast number of possibilities are most pertinent for a specific task or problem. While it is acknowledged that in both the knowledge-driven and pairwise approaches, any resultant maps may not explicitly show the impact of other variables relevant for the process being investigated, these approaches do avoid the potential for artefacts due to the constant sum constraint (closure).

3. The centred log-ratio component map

The usage of the clr transformation may be grounded in one of the following options: (1) to highlight the role of one component, (2) to see the patterns of that element relative to the average behaviour of other elements at hand in the compositional dataset (using their geometric mean), and (3) to provide the first visualization towards a deeper understanding of the underlying geochemical processes.

Mapping single clr scores seems to present a first-step simple solution for an unsupervised, blind application. However, it must be noted that clr variables have severe limitations. First, they cannot be interpreted the same way as the variables in the original concentration units, and there is no straightforward back-transformation. A further problem of the clr is its sensitivity to replacement values of censored or missing values, because the more numerous trace elements (with low concentrations and more common censorship) influence the geometric mean more than the fewer major elements (with high concentrations and rarely censored). Therefore, if we impute too low values, we shift the geometric mean to lower values. One solution could be to use a robust central value (e.g. the median of all the components, a trimmed geometric mean, etc.), although this destroys many of the nice properties of the clr and may cause problems for the interpretation of any resulting map. More importantly, as has been mentioned, the clr depends on the set of components chosen, i.e. any component cannot be analysed separately without consideration of the others. This might make interpretation of clr maps difficult, especially when the other components are driven by their own geochemical processes.

The above features are illustrated using the Tellus dataset. Figure 6a demonstrates the effect of employing the whole composition or the subcomposition of major oxides, respectively, for clr scores of these elements. In addition to a shifting of the distribution, a result of excluding the trace elements, the positive linear correlation between values of both scores is substantially disturbed for some components (e.g. in particular for Al_2O_3, MgO, SiO_2). This is also reflected in the corresponding univariate maps (as in the case of clr Al_2O_3, see Figure 6b and c). This clearly demonstrates that the selection of the subcomposition is important for the construction of clr scores. Here the use of the subcomposition of major oxides clr scores seems to be reasonable due to their higher geochemical stability in the study area. On the other hand, by employing all elements we can observe some clear regional patterns, that may indicate interesting inherent processes in the data that are not so well reflected from the major oxides subcompositional clr.

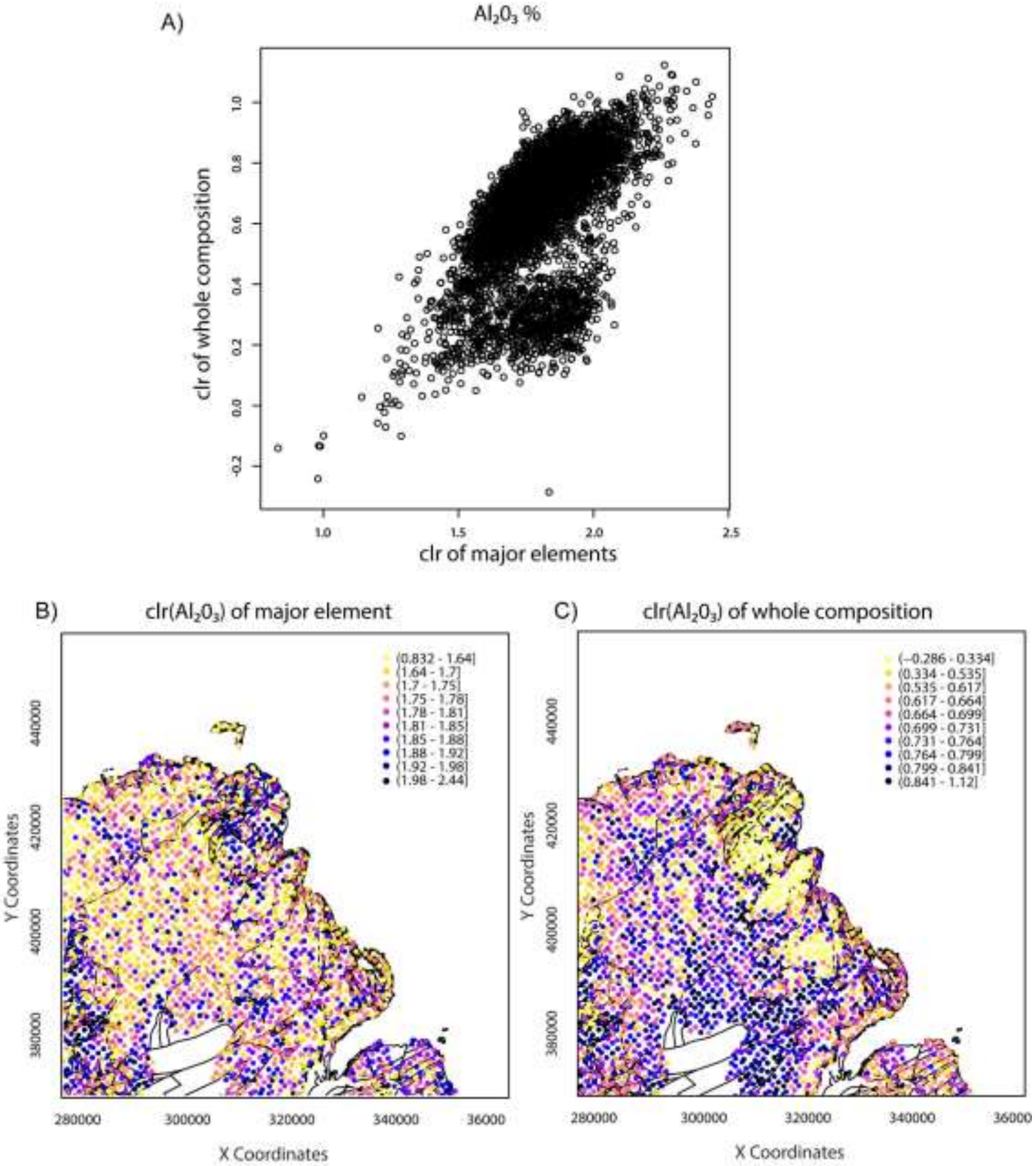

Figure 6: A) The effect of employing the clr of the whole composition and the clr of the subcomposition of the major oxides is illustrated for Al2O3. Note the substantial disturbance in the positive linear correlation. The corresponding univariate maps demonstrating the effect of employing B) the clr (Al2O3) of the subcomposition of major oxides or C) the clr (Al2O3) of the whole composition is shown. The units are coordinates in the Irish Transverse Mercator Grid (EPSG:29903).

In the authors' opinion, using the geometric mean of a geologically meaningful subset of components, selected on the basis of a geological understanding, may offer a possible way to form a compromise between the two situations that we have demonstrated (Fig. 6 b and c). That is using pairwise log-ratio maps (one element to another, or any knowledge-driven log-ratio) and using the clr scores of the full composition including the problematic components (i.e. those having a high proportion of values that are censored, close to the detection limit, or rounded). The result may offer a reasonable number of single (clr score) or low) component maps. For the Tellus Survey example it would seem a sensible approach to include the major oxides Al_2O_3, CaO, Fe_2O_3, K_2O, MgO, MnO, Na_2O, P_2O_5, SiO_2 and TiO_2, with the exception of LOI and SO_3 to avoid problems with the soil water content and with the large number of missing values of SO_3. Their subcompositional clr-scores could then be mapped. For the remaining elements, comprising the trace elements, LOI and SO_3, the balance of each element against the geometric mean of the major component subcomposition could be mapped.

Another use of a subcompositional clr is to restrict attention to a knowledge-driven selection of parts, and constructing and mapping their subcompositional clr scores. If the rationale for the study is set out beforehand, then these clr maps of subcompositions should be interpretable, because the subcomposition was selected and studied for a particular reason. An example of this approach could be to use a subcomposition of potentially harmful elements that are known to adversely affect human health (for example As, Cr, Pb and Cd). In fact, this approach is reminiscent of the first method, knowledge-driven log-ratios, where the interpretability of the resulting variables is assumed since there was a knowledge-driven rationale for the approach.

4. Exploring patterns and elemental associations (data-driven projections)

Multi-element geochemical datasets are often analysed to explore patterns, links and associations between elements as part of the 'Discovery Process' (Grunsky et al., 2014). Multivariate methods available include principal component analysis (PCA), regression-related techniques (regression analysis, analysis of variance, total and partial least squares regression or canonical correlation) and grouping techniques (cluster analysis, discriminant analysis), often used to explore these patterns or links. From the point of view of mapping, most of these methods can actually be seen as a way to choose one (or a few) log-contrasts of potential interest. There is thus a range of techniques available and the issue is to choose the best one. Two families of approaches can be considered: supervised and unsupervised methods. Again, these will be exemplified using the Tellus case study.

Supervised Approaches

In the supervised approach the goal is to find which log-contrast better relates to known secondary information which is also available. For example, we may be interested in identifying the log-contrast which best distinguishes between two types of geochemical situations such as two terranes, or areas of peat coverage and peat free areas. In that case we could start from a subset of the data, where this extra information is known (i.e. a *training set*). The best log-ratio can be computed by a linear discriminant analysis derived from any log-ratio transformed composition, although due to numerical issues, ilr and alr are the best in this case (readers are referred to van den Boogaart and Tolosana-Delgado,

(2013) for further detail). Such a linear discriminant analysis has been conducted on the Tellus soil geochemical dataset on the basis of areas covered by blanket peat (Figure 7). The linear discriminant analysis suggests that many more elements and oxides (Ni, Rb, Hf, Nb, Cr, K_2O, MnO, Fe_2O_3, some with positive weight, some with negative weight; Fig.3a) are useful for blanket peat delineation than solely using LOI %, as is traditionally the case. Overall, the histograms shown in Figure 7b and 7c show that the classification between peat covered and peat free areas, produced by geostatistical Fisher discriminant analysis in this example (Tolosana-Delgado et al. 2015), is reasonably successful. On further investigation it is observed that misclassified points are found mainly at the boundaries between peat-covered areas and peat-free areas and most likely related to point-polygon mismatch between the Tellus data points and peat assigned polygons (discussed in more detail in Tolosana-Delgado et al. 2015). A full discriminant analysis approach could be applied in that results could be validated further using cross-validation, the log-contrast could be derived from a training set and the discriminant power of the analysis could be evaluated.

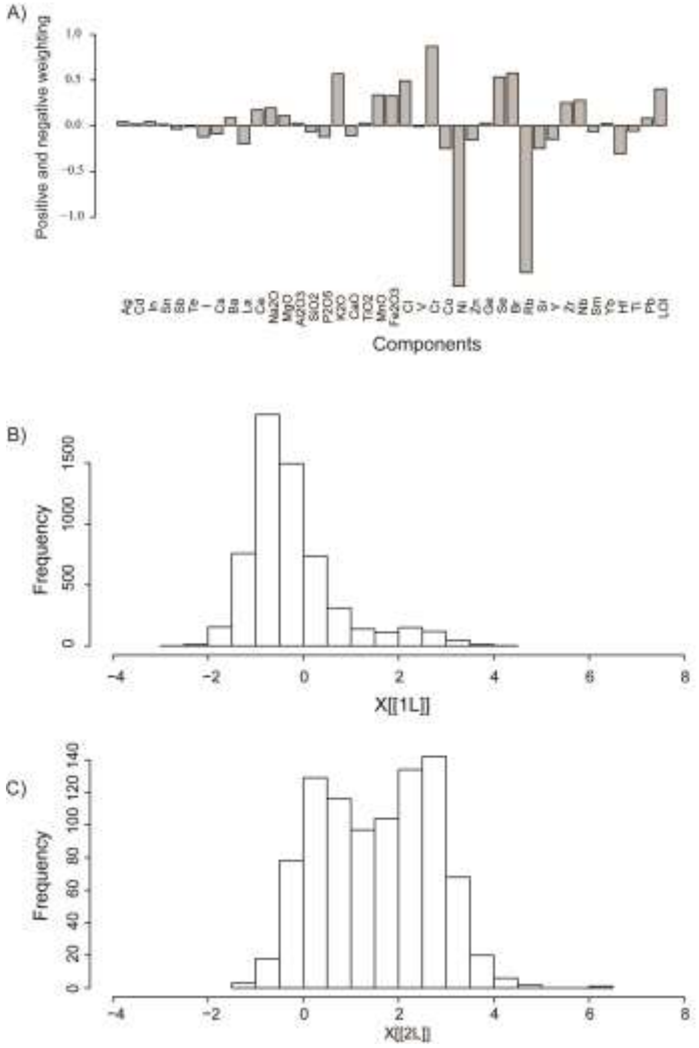

Figure 7: Illustrating a supervised method: Linear discriminant analysis using the Tellus soil geochemical dataset on the basis of blanket peat; A) elemental linear discriminant analysis, showing positive and negative weightings and histograms showing the classification between B) peat covered and C) peat free areas, produced by the geostatistical Fisher discriminant analysis. Readers are referred to Tolosana-Delgado et al. (2015) for further detail.

In a similar way, if the covariable(s) of interest is (are) a numeric quantity or a set of them (magnetics, gravimetrics, Th-channel radiometrics, total gamma-ray counts, proportion of sand, log-ratio of silt to sand, pH, etc.), compositional regression analysis could be used in order to identify the optimal log-contrast (e.g McKinley et al. in review). The composition should be expressed in log-ratios (alr or ilr) and a classical regression applied between the log-ratio scores and the chosen covariable. The explanatory power of the resultant model can be measured by its associated R-square. Thus, beyond the classical use of regression, it is possible to use this criterion to select the optimal pairwise log-ratio that maximizes the R-square of the simple regression model with the covariable of interest. This requires the exhaustive exploration of all possible pairwise log-ratios, though the selection of the “best” is automatically done by the method and not by visual inspection. Note that these techniques, although explained for one covariable or two groups, can be easily generalized towards multiple groups or covariables, providing an extra direction for each added group or covariable is known (van den Boogaart and Tolosana-Delgado, 2013).

Unsupervised approaches

Even in the absence of additional information, multivariate analysis techniques can help to identify interesting log-contrasts to map. These often show either large variability or very low variability. For example, the variation coefficient t_{ij} of two components provides a measure of their association, which can be used in cluster analysis as a measure of distance between variables. The application of cluster analysis for compositional data (compositional Q-mode clustering) is explained in detail by van den Boogaart and Tolosana-Delgado (2013). Clusters will contain elements behaving proportionally throughout the dataset. Log-ratios between elements of two different clusters should thus be similar to other log-ratios of the elements of the same clusters. Therefore, one of these log-ratios or a balance of one cluster against the other might be representative for many log-ratios, and consequently may represent a process influencing many elements in the same way. Balances of elements within the cluster will filter out these large-variability effects and focus on differences between elements behaving similar with respect to major processes. In a sense this is the approach that geochemists instinctively use when applying simple ratios. The advantage of this unsupervised CoDA approach that reflects a classical supervised approach is that these balances should thus be able to highlight processes differentiating between similar elements, i.e. second-order processes that would otherwise be overshadowed by the major processes.

In the example shown in Figure 8 using the Tellus regional soil data, a cluster dendrogram is constructed using all those components without missing values (shown in Figure 4). Although it is acknowledged that this introduces a bias, since the treatment of below detection values and zeros in compositional data is beyond the scope of this paper, this simplified approach is used here to exemplify the approach. The hierarchical cluster has been produced with a Ward cluster analysis with the variation matrix as distance between elements, the procedure is detailed in van den Boogaart and Tolosana-Delgado (2013). Six resultant ilr balances have been shown to demonstrate different features: spatial randomness (Fig. 8a; Yb vs Sm); the identification of granitic bedrock areas (Fig. 8b; Nb, Hf vs SiO_2); an artefact in the central region perhaps related to the analytical method of Ag detection (Energy Dispersive Polarised X-Ray Fluorescence, Fig. 8c, Te, Sm, Yb, Tl vs Ag); Geological bedrock units including Palaeogene basalts, a granodiorite complex and Dalradian metasediments (Fig. 8d, Rb vs K_2O); Ordovician- Devonian sandstone units and granitic rocks (Fig. 8e, Zr vs Na_2O); Palaeogene basalt detection and LOI-related peat unsupervised detection (Fig. 8f).

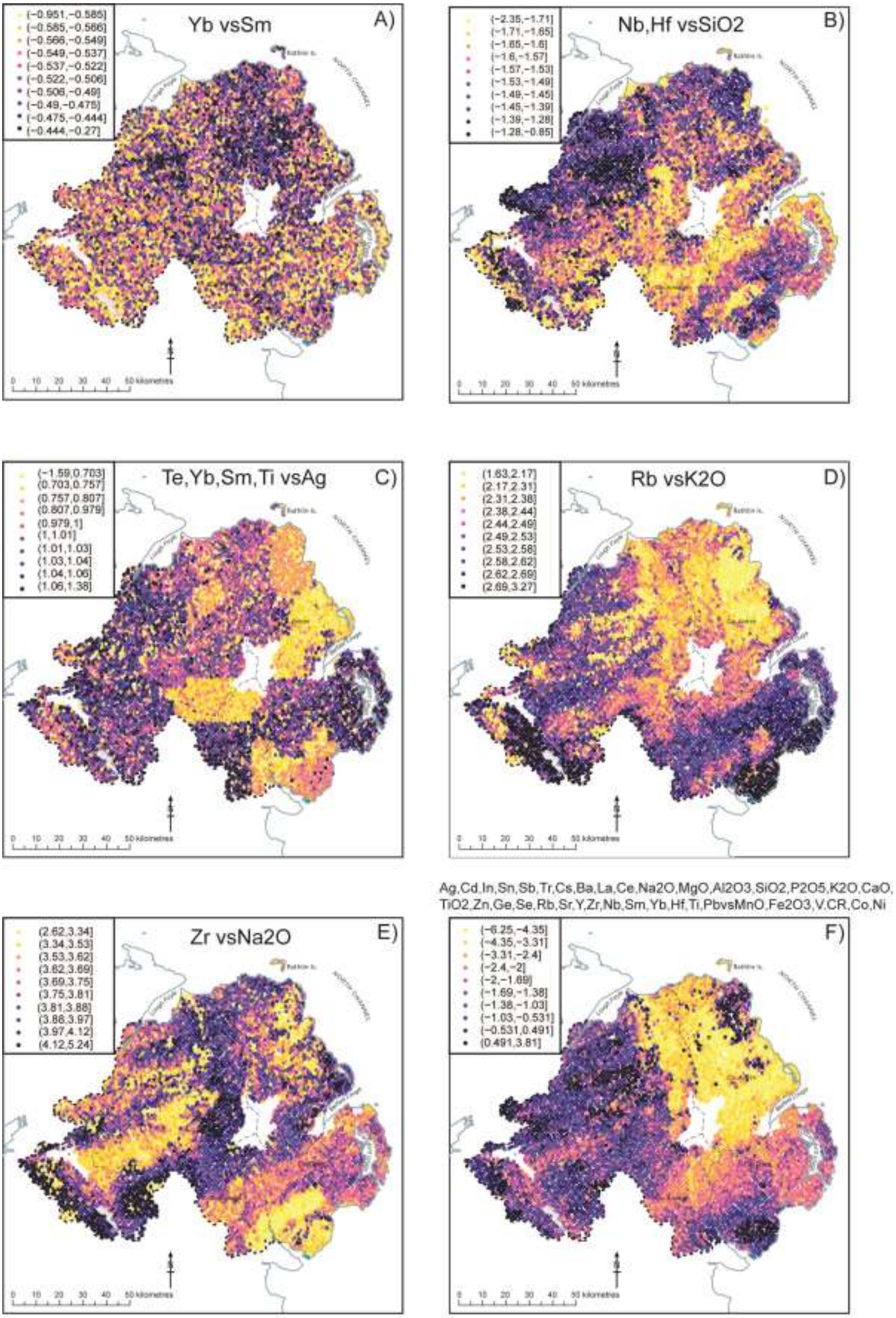
A)
Yb vsSm
(−0.951,−0.585]
(−0.585,−0.566]
(−0.566,−0.549]
(−0.549,−0.537]
(−0.537,−0.522]
(−0.522,−0.506]
(−0.506,−0.49]
(−0.49,−0.475]
(−0.475,−0.444]
(−0.444,−0.27]
B)
Nb,Hf vsSiO2
(−2.35,−1.71]
(−1.71,−1.65]
(−1.65,−1.6]
(−1.6,−1.57]
(−1.57,−1.53]
(−1.53,−1.49]
(−1.49,−1.45]
(−1.45,−1.39]
(−1.39,−1.28]
(−1.28,−0.85]
C)
Te,Yb,Sm,Ti vsAg
(−1.59,0.703]
(0.703,0.757]
(0.757,0.807]
(0.807,0.979]
(0.979,1]
(1,1.01]
(1.01,1.03]
(1.03,1.04]
(1.04,1.06]
(1.06,1.38]
D)
Rb vsK2O
(1.63,2.17]
(2.17,2.31]
(2.31,2.38]
(2.38,2.44]
(2.44,2.49]
(2.49,2.53]
(2.53,2.58]
(2.58,2.62]
(2.62,2.69]
(2.69,3.27]
E)
Zr vsNa2O
(2.62,3.34]
(3.34,3.53]
(3.53,3.62]
(3.62,3.69]
(3.69,3.75]
(3.75,3.81]
(3.81,3.88]
(3.88,3.97]
(3.97,4.12]
(4.12,5.24]
Ag,Cd,In,Sn,Sb,Tr,Cs,Ba,La,Ce,Na2O,MgO,Al2O3,SiO2,P2O5,K2O,CaO,
TiO2,Zn,Ge,Se,Rb,Sr,Y,Zr,Nb,Sm,Yb,Hf,Ti,PbvsMnO,Fe2O3,V.CR,Co,Ni
F)
(−6.25,−4.35]
(−4.35,−3.31]
(−3.31,−2.4]
(−2.4,−2]
(−2,−1.69]
(−1.69,−1.38]
(−1.38,−1.03]
(−1.03,−0.531]
(−0.531,0.491]
(0.491,3.81]
0 10 20 30 40 50 kilometres
NORTH CHANNEL
Rathlin Is.

Figure 8: The figures illustrate an unsupervised method based on the ilr balance dendrogram shown in Figure 4: Six resultant ilr balances are selected to illustrate different features: A) spatial randomness (Yb vs Sm); B) granitic bedrock areas (Nb,Hf vs SiO2); C) an artefact related to the analytical method of Ag detection (Te,Sm,Yb,Tl vs Ag); D) Geological bedrock units including Palaeogene basalts, granodiorite complex and Dalradian metasediments (Rb vs K2O); E) Ordovician- Devonian sandstone units and granitic rocks (Zr vs Na2O ; F) Palaeogene basalt detection and LOI-related peat detection.

Another popular approach traditionally used for identifying interesting log-contrasts is principal component analysis of the clr-transformed dataset and its graphical display, the compositional biplot (Aitchison and Greenacre, 2002; Caritat and Grunsky, 2013). Within a compositional approach this allows the identification of directions or log-contrasts of large or small variations based on a variance-covariance matrix. Log-contrasts of strong variation might be useful for identifying major processes strongly influencing many elements at the same time. The directions of small variation are more useful for identifying common equilibrium-like laws and anomalies by highlighting cases showing strong deviations from otherwise stable log-contrasts. These directions of lesser magnitude may also represent random processes in the data. Any of the methods mentioned before can be used in several different ways with the purpose of focussing on specific features and aspects of the dataset. In summary, there are many possible tools and approaches for selecting informative and valid univariate projections of a composition. The key point is to relate them to specific research questions to address pertinent geochemical and geological issues.

CONCLUSIONS - BEST PRACTICE AND RECOMMENDATIONS

The question that this paper has addressed is whether the single component elemental map is meaningful: can it be interpreted as fact or is it indeed no more than a fictional representation? The essence of this dilemma is that geochemical data are compositional variables and as such carry only relative information. In some cases, the classical single-component maps might carry some relevant information, but most of the time the information-rich relative data are not properly represented. On reflection, this research recommends the following:

2. An appropriate manner to represent "measured" raw or absolute geochemical values is in the form of a single component point source or ‘dot maps’ with an acknowledgement about dependence to all other variables. This may be useful for the purpose of geochemical exploration but is critical for environmental studies where stated values fall below/above guidelines. Generally, without considering or knowing the other elements of the composition, we are not able to fully interpret the component of interest. That is to say that involving more components in the interpretation can reveal the reason behind the relative increase or decrease in the component of interest.

3. A complementary compositional approach is required to ensure that mapped outputs are accurate and robust to interpret the underlying geochemical processes and highlight any anomalies of interest. One suggestion is to look at log-ratio maps, which may include either knowledge-driven log-ratios, pairwise log-ratios, any reasonably aggregated log-ratios or data-driven projections, in addition to single element maps whenever possible. In this case, the authors propose the following chain of representations:

- in paper format, the maps of clr-transformed major components together with the balance of each trace component to the major components; in electronic interactive formats, this can be complemented with maps of each pwlr score;
- maps of clr- or ilr-based principal components, those with highest variability and also those with very low variability; it is worth noting that one should be careful if including all elements for principal component analysis as trace ones (resulting in highest variability due to relative scale of compositions) are often burdened by measurement errors;
- maps of the balances resulting from a compositional cluster analysis based on the variation matrix.

4. No matter whether absolute or log-ratio maps are used, most commonly it will be necessary to use quantile-based colour and symbol scales to highlight the patterns captured. In log-ratio maps, equal length intervals for colours and symbols are often useful as well. However this is close to useless in absolute maps except to highlight a few high-value outliers.
5. If a compositional approach is not used, it becomes difficult to guarantee that any identified pattern, trend or anomaly is not an artefact of the constant sum constraint. Nevertheless, it should be noted that if the elements do not sum up to the same constant, the relative scale of compositions still needs to be taken in account.
6. The evaluation of individual component maps is a first step towards a comprehensive understanding of geochemical processes that are reflected in the material that has been analysed geochemically. It is the initial step in the “Discovery Process” from which geologically meaningful log-contrasts can be derived. There are so many possible supervised and unsupervised data-driven methods, that it is necessary to: (1) clearly state the

geological/geochemical question to ask; (2) search for the appropriate statistical method that can answer that question; (3) ensure that the data available are sufficient to answer that question with the selected method; (4) apply the required log-ratio transformations and the chosen statistical method; and (5) learn to interpret the results. This will most often require statisticians, data analysts and geochemists working together.

Acknowledgments

The authors thank and acknowledge the participants of the first GeoMap Workshop (held in Olomouc, Czech Republic, 17-20 June 2014) from which this collaborative research stems. This includes 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). GSNI are gratefully 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. Patrice de Caritat publishes with permission from the Chief Executive Officer, Geoscience Australia.

References

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., Greenacre, M., 2002. Biplot of compositional data. Applied Statistics, 51 (4), 375-392.

Barsby, A., McKinley, J., Ofterdinger, U., Young, M., Cave, M., Wragg, J., 2012. Bioaccessibilty of trace elements in Northern Ireland. Science of the Total Environment, 433, 397-417. doi:10.1016/j.scitotenv.2012.05.099

Birke, M., Rauch, U., Stummeyer, J., 2015. How robust are geochemical patterns? A comparison of low and high sample density geochemical mapping in Germany, Journal of Geochemical Exploration, 154, 105-128. http://dx.doi.org/10.1016/j.gexplo.2014.12.005

Bolviken, B., Stokke, P. R., Feder, J., Jossang, T., 1992. The fractal nature of geochemical landscapes, J Journal of Geochemical Exploration, 43, 91– 109.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).

Caritat, P. de, Grunsky, E.C., 2013. Defining element associations and inferring geological processes from total element concentrations in Australian catchment outlet sediments: multivariate analysis of continental-scale geochemical data. Applied Geochemistry, 33, 104-126. DOI: 10.1016/j.apgeochem.2013.02.005

Carranza, E.J.C., 2010. Mapping of anomalies in continuous and discrete fields of stream sediment geochemical landscapes. Geochemical Exploration Environmental Analysis, 10, 171–187.

CCME (Canadian Council for Ministers of the Environment), 2014. Canadian Environmental Quality Guidelines. http://ceqg-rcqe.ccme.ca/

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

Cheng, Z., Xie X., Yao W., Feng J., Zhang Q., Fang J. 2014. Multi-element geochemical mapping in Southern China. Journal of Geochemical Exploration, 139, 183–192.

Chiprés, J.A., Salinas, J.C., Castro-Larragoitia, J., Monroy, M.G., 2008. Geochemical mapping of major and trace elements in soils from the Altiplano Potosino, Mexico: a multiscale comparison. Geochemical Exploration Environmental Analysis, 8, 279–290

Darnley A.G., Björklund A., Bølviken B., Gustavsson N., Koval V., Plant J.A., Steenfelt A., Tauchid M., Xie X.J. 1995, A global geochemical database for environmental and resource management-recommendations for international geochemical mapping—final report of IGCP Project 259. Earth Sciences, UNESCO Publishing, 19.

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, doi:10.1016/j.scitotenv.2010.08.004.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., C. Barceló-Vidal (2003) Isometric log-ratio 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, 795-828.

Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. In Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V., (Eds), Compositional Data Analysis in the Geosciences. From Theory to Practice. Geological Society, London, Special Publications 264,145-159.

Egozcue, J.J., Pawlowsky-Glahn, V., 2011. Basic concepts and procedures, in V. Pawlowsky-Glahn and A. Buccianti, (Eds.), Compositional Data Analysis. Theory and Applications. Wiley, Chichester, pp. 12-28.

EGS (EuroGeoSurveys Geochemistry Working Group), 2008. EuroGeoSurveys Geochemical mapping of agricultural and grazing land soil of Europe (GEMAS) - Field manual. NGU Report 2008.038. ISSN 0800-3416. Geological Survey of Norway.

Filzmoser, P., Hron, K., 2011. Robust statistical analysis, in V. Pawlowsky-Glahn and A. Buccianti, (Eds.), Compositional Data Analysis. Theory and Applications. John Wiley & Sons, Chichester (UK), 59-72.

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,* *doi:*10.1144/geochem2012-140

Grunsky, E.C., Drew, L.J., Sutphin, D.M., 2009. Process recognition in multi-element soil and stream-sediment geochemical data. Applied Geochemistry, 24, 1602–1616.

Grunsky, E.C. 2010.The interpretation of geochemical survey data; Geochemistry, Exploration, Environment Analysis, 10(1), 27-74.

Grunsky, E.C., Mueller, U.A., Corrigan, D.,2014. A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: Applications for predictive geological mapping. Journal of Geochemical Exploration, 141, 15-41. doi 10.1016/j.gexplo.2013.07.013.

Johnson, C.C., 2005. 2005 G-BASE field procedures manual. British Geological Survey, 65pp. (IR/05/097) (Unpublished).

Lagacherie, P., McBratney, A.B., Voltz, M., (Eds.) 2007. Digital Soil Mapping, an introductory perspective. Developments in Soil Science, 31. Amsterdam.

Lancianese, V., Dinelli, E., 2014 in press. Different spatial methods in regional geochemical mapping at high density sampling: An application on stream sediments of Romagna Apennines, Northern Italy. J. Geochemical Exploration http://dx.doi.org/10.1016/j.gexplo.2014.12.014

Lech, M.E., Caritat, P. de, McPherson, A.A., 2007. National Geochemical Survey of Australia: Field Manual. Geoscience Australia Record, 2007/08, 53 pp.

Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. Ecological Informatics, 6, 228-241.

Madeisky, H.E., Stanley, C.R., 2010. Lithogeochemical exploration of meta- somatic zones associated with volcanic-hosted massive sulfide deposits using Pearce element ratio analysis. International Geology Review, 35, 1121–1148.

McKinley, J.M., 2013, 'How useful are databases in environmental and criminal forensics?' Geological Society of London Special Publication, 384, 109-119., http://dx.doi.org/10.1144/SP384.9

Mitchell, W.I., 2004. The Geology of Northern Ireland – Our Natural Foundation. Geological Survey of Northern Ireland, Belfast.

Montreuil, J. -F., Corriveau, L., Grunsky, E.C., 2013. Compositional data analysis of hydrothermal alteration in IOCG systems, Great Bear magmatic zone, Canada: To each alteration types its own signature, Geochemistry: Exploration, Environment, Analysis vol. 13, no. 4, 2013; p. 229-247, doi:10.1144/geochem2011-101.

Morris, P.A., Pirajno, F., and Shevchenko, S., 2003. Proterozoic mineralization identified by integrated regional regolith geochemistry, geophysics and bedrock mapping in Western Australia. Geochemistry: Exploration, Environment, Analysis, Geological Society of London, 3 (1), 13-28.

Ohta, A., Imai, N., Terashima, S., Tachibana, Y., 2011. Regional geochemical mapping in eastern Japan including the nation's capital, Tokyo. Geochemistry: Exploration, Environment, Analysis, 11, 211–223.

Pawlowsky-Glahn, V., Buccianti, A., (eds.) 2011. Compositional Data Analysis. Theory and Applications. Wiley, Chichester.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and analysis of compositional data. Wiley, Chichester.

Pearce, T.H., 1968. A contribution to the theory of variation diagrams. Contributions to Mineralogy and Petrology, 19, 142–157.

Reimann, C. and Filzmoser, P., 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environmental Geology, 39(9), 1001–1014.

Reimann C., Siewers U., Tarvainen T., Bityukova L., Eriksson J., Gilucis A., Gregorauskiene V., Lukashev V.K., Matinian N.N., Pasieczna A., 2003. Agricultural Soils in Northern Europe: A Geochemical Atlas. Geologisches Jahrbuch, Sonderhefte, Reihe D, Heft SD 5, Schweizerbart'sche Verlagsbuchhandlung, Stuttgart.

Reimann, C., 2005. Geochemical mapping – technique or art? Geochemistry, Exploration, Environment, Analysis 5, 359-370.

Reimann, C., Garrett, R.G., 2005. Geochemical background – concept and reality. Science of the Total Environment 350, 12-27.

Reimann, C., Filzmoser, P., Garrett, R.G. 2005. Background and threshold: critical comparison of methods of determination. The Science of the Total Environment, 346, 1–16.

Reimann, C., Filzmoser, P., Garrett, R., Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. Wiley, Chichester.

Reimann, C., ÄyrÄs, M., Cheklushin, V., Bogatyrev, I., Boyd, R., Caritat, P. de, Dutter, R., Finne, T.E., Halleraker, J.H., Jæger, Ø., Kashulina, G., Lehto, O., Niskavaara, H., Pavlov, V., Räisänen, M.L., Strand, T., Volden, T., 1998. Environmental Geochemical Atlas of the Central Barents Region. ISBN 82-7385-176-1. NGU-GTK-CKE Special Publication, Geological Survey of Norway, Trondheim, Norway.

Reimann,C., Matschullat, J., Birke, M., Salminen, R., 2010. Antimony in the environment – lessons from geochemical mapping. Applied Geochemistry *25, 175-198.*

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), Schweizerbarth, Hannover. +DVD

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.

Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., GEMAS PROJECT TEAM, 2012. The concept of compositional data analysis in practise – Total major element concentrations in agricultural and grazing land soils of Europe. Science of the Total Environment 426, 196-210.

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

Smith R.J., Dransfield M.H., Rajagopalan S., 1997. Regional Geophysical Exploration in Arid Terrain. In Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration. Gubins A.G. (Ed.) Applications of Regional Geophysics and Geochemistry, 105, 821–828

Smith, D.B., Cannon, W.F., Woodruff, L.G., 2011. A national –scale geochemical and mineralogical survey of soils of the conterminous United States. Applied Geochemistry, 26, S250-S255.

Templ, M., Hron, K., Filzmoser, P. 2011. robCompositions: an R-package for robust statistical analysis of compositional data. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis. Theory and Applications. Wiley, Chichester, 341-355.

Thomas, C. W., Aitchison, J. 2005. Compositional data analysis of variability and process: A case study. Mathematical Geology 37, 753-772.

Tolosana-Delgado, R., van den Boogaart, K.G. 2013. Joint consistent mapping of high-dimensional geochemical surveys. Mathematical Geosciences 45, 983-1004.

Tolosana-Delgado, R., McKinley, J., van den Boogaart, K. G., 2015. Geostatistical Fisher discriminant analysis, Proceedings of IAMG 2015, Freiberg: IAMG office, 978-3-00-050337-5

van den Boogaart, K.G., Tolosana-Delgado, R., 2013. Analyzing compositional data with R. Springer, Heidelberg.

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.