Applied Geochemistry 75 (2016) 263–276

Contents lists available at ScienceDirect

Applied Geochemistry

journal homepage: www.elsevier.com/locate/apgeochem

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, Chemnitzerstr. 40, D-09599 Freiberg, Germany
[b] School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, BT7 1NN, UK

ARTICLE INFO

Article history:
Received 20 January 2016
Received in revised form
6 May 2016
Accepted 9 May 2016
Available online 12 May 2016

Keywords:
Centered log-ratio transformation
clr
Spurious correlation
Compositional data analysis

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).

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, 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

* Corresponding author.
E-mail addresses: r.tolosana@hzdr.de (R. Tolosana-Delgado), j.mckinley@qub.ac.uk (J. McKinley).

http://dx.doi.org/10.1016/j.apgeochem.2016.05.004
0883-2927/

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,b).

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 386 Mt (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 20 cm depth, collected on a regular grid of one sample site every 2 km^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 $\mathbf{x}_n = [x_{n1}, x_{n2}, \ldots, x_{nD}]$. Consider a sample of N of these vectors. The variance is the classical way of measuring the variability of each component,