Queen's University
Belfast

The single component geochemical map: Fact or fiction?

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

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

23   ABSTRACT

24 1. Single component geochemical maps are the most basic representation of spatial elemental

25   distributions and commonly used in environmental and exploration geochemistry. However,

26   the compositional nature of geochemical data imposes several limitations on how the data

27   should be presented. The problems relate to the constant sum problem (closure), and the

28   inherently multivariate relative information conveyed by compositional data. Well known is,

29   for instance, the tendency of all heavy metals to show lower values in soils with significant

30   contributions of diluting elements (e.g., the quartz dilution effect); or the contrary effect,

31   apparent enrichment in many elements due to removal of potassium during weathering.

32   The validity of classical single component maps is thus investigated, and reasonable

33   alternatives that honour the compositional character of geochemical concentrations are

34   presented. The first recommended such method relies on knowledge-driven log-ratios,

35   chosen to highlight certain geochemical relations or to filter known artefacts (e.g. dilution

36   with SiO_2 or volatiles). This is similar to the classical normalisation approach to a single

37   element. The second approach uses the (so called) log-contrasts, that employ suitable

38   statistical methods (such as classification techniques, regression analysis, principal

39   component analysis, clustering of variables, etc.) to extract potentially interesting

40   geochemical summaries. The caution from this work is that if a compositional approach is

41   not used, it becomes difficult to guarantee that any identified pattern, trend or anomaly is

42   not an artefact of the constant sum constraint. In summary the authors recommend a chain

43   of enquiry that involves searching for the appropriate statistical method that can answer the

44   required geological or geochemical question whilst maintaining the integrity of the

45   compositional nature of the data. The required log-ratio transformations should be applied

46   followed by the chosen statistical method. Interpreting the results may require a closer

47   working relationship between statisticians, data analysts and geochemists.

48   <u>Key words:</u> soil geochemistry, compositional data analysis, log-ratios, mapping.

49   INTRODUCTION

50   Over the last decade the focus of digital soil and sediment databases at a regional, national,

51   transnational and continent scale has increasingly become to provide information for a

52   range of purposes including geological and soil mapping, baseline quality documentation,

53   mineral prospecting, land and soil resource assessment, risk evaluation, environmental and

54   educational purposes, and prediction of soil provenance for forensic purposes (Smith et al.,

55   1997; Reimann et al., 1998, 2003; Morris et al., 2003; Lagacherie et al., 2007; McKinley,

56   2013). These digital soil databases are generally accompanied by geochemical atlases (e.g.

57   Caritat and Cooper, 2011a; Young and Donald, 2013; Reimann et al., 2014a,b) showing a

58   collection of distribution maps for individual geochemical elements. This has been matched

59   by a corresponding increase in published studies utilising these soil geochemical surveys

60   (e.g. Chiprés et al., 2008; Grunsky et al., 2009; Carranza, 2010; Ohta et al., 2011; Caritat and

61   Grunsky, 2013; Cheng et al., 2014; Lancianese and Dinelli, 2014; Birke et al., 2015). The

62   resolution of the ground-based sampling scheme used for the generation of these databases

63   is the best compromise between the extent of the region covered, and time and resources

64   available. Over a local to regional scale, soil sampling can be managed on a 2 km^2 grid as

65   applied by the Geological Survey of Northern Ireland (GSNI)'s Tellus project. A regional scale

66   dataset such as this is close to exhaustive sampling in terms of geological survey mapping,

67   reducing the need for interpolation. It is worth noting however that this is not exhaustive

68   sampling for mineral exploration. Generally surveys are carried out at local (1:10,000),

69   regional (1:250,000) or continental (1:1,000,000) scales. A number of authors discuss the

70   scale of geochemical mapping (e.g. Bølviken et al., 1992; Darnley et al., 1995; Reimann et al.,

71   2010).

72   The sampling strategy generally follows standard protocols (e.g. UNESCO's IGCP Global

73   Geochemical Database – Darnley et al., 1995; G-BASE standard established by BGS - Johnson

74   2005; FOREGS field handbook – Salminen et al., 1998; GEMAS field handbook – EGS, 2008;

75   North American Soil Geochemical Landscapes Project – Smith et al. 2011; National

76   Geochemical Survey of Australia field manual – Lech et al., 2007; Canadian component of

77   the North American Soil Geochemical Landscapes Project – Friske et al., 2013; China

78   Geochemical Baseline Project – Wang et al., 2015).

79   Geochemical survey data are typically represented and interpreted using single element

80   geochemical maps. The interpretability and validity of these single components have

81   repeatedly been challenged because they are prone to several artefacts: spurious negative

82   bias on correlations (Chayes, 1960), dependence of interpretation on other (potentially non-

83   reported) components (Aitchison, 1986), dependence on units (e.g. mass, molar), and

84   dependence on processes acting on some components (e.g. weathering, dilution) but

85   influencing all of them (van den Boogaart and Tolosana-Delgado, 2013). All the issues

86   mentioned are due to the fact that geochemical data constitute amounts of components

87   with relative portions of a total even if this total is unknown. The components may be

88   reported in different physical units (ppm, mg/kg or as percentages) and all the components

89   may not be reported or measured. However, each component has an amount which

90   represents its importance as part of the whole composition. The constraints of constant sum

91   or the closed nature of the relative amounts of components have implications for the

92   analysis of geochemical data. In statistics and mathematical geosciences, powerful solutions

93   to deal with these issues have been developed in a field known as Compositional Data

94   Analysis (CoDA) (Aitchison, 1986; Grunsky, 2010; Pawlowsky-Glahn and Buccianti, 2011; van

95   den Boogaart and Tolosana-Delgado, 2013; Templ et al., 2011). This paper investigates the

96   question of what the compositional nature of geochemical data means in regional

97   geochemical mapping, specifically with regard to single element (univariate) distribution

98   maps.

99   The collaborative research presented stems from the first GeoMap Workshop (held in

100   Olomouc, Czech Republic, 17-20 June 2014) that discussed the consequences of these

101   challenges and the usefulness of CoDA for regional geochemistry. Present were

102   representatives from regional geochemical surveys (Tellus Survey covering Northern Ireland

103   and Tellus Border Survey covering the border counties of the Republic of Ireland (Young and

104   Donald, 2013)), the GEMAS project covering a large part of Europe (Reimann et al. 2014a,b),

105   the soil geochemical survey of the conterminous USA (Smith et al., 2011; Drew et al., 2010)

106   and the continental scale National Geochemical Survey of Australia (Caritat and Cooper,

107   2011a,b). This paper, while acknowledging the historically important role of single

108   component maps, aims to: (1) discuss their appropriateness, (2) provide some examples to

109   highlight the problems raised above, and (3) offer some solutions to present interpretable

110   maps free of the issues arising from the compositional nature of geochemical data.

111

112   PROBLEMS RELATED TO SINGLE COMPONENT MAPS

113   Geochemical surveys generate datasets with several tens of components (between 50 and

114   up to 70 elements are commonly reported), obtained from different sample materials (soil

115   horizons, size fractions, vegetal tissues, sediments, water, etc.) and with different analytical

116   techniques (total analyses, partial or selective digests). For regional geochemistry, the key

117   applications of the data are generally either to produce and use elemental concentration

118   maps (i.e. one-component regional distribution maps) or to explore associations between

119   elements affected by geological/geochemical processes, which can also be mapped (e.g.

120   principal components). For the first of these tasks, standard practice has included producing

121   a single component map thought to represent the raw or "absolute" input data in the form

122   of dot (or point) maps, but also as interpolated maps of these raw concentrations. Reimann

123   (2005) defines the purpose of such geochemical maps as 'to display different processes in a

124   map form and to detect local deviations from the dominant process in any one sub-area'.

125   The problems discussed in this section result from the closure property of geochemical

126   compositions, i.e. the unavoidable fact that samples are *aliquots* of the geological bodies we

127   want to investigate and therefore do not really convey information about the element mass

128   distributions (Aitchison, 1986). This has implications for baseline quality mapping, mineral

129   prospecting, land and soil resource assessment or risk evaluation, though these issues are

130   beyond the scope of this contribution. Nevertheless, in the following we show that the

131   traditional meaning of closure effect (linked to closed data with unit, or any other fixed sum

132   constraint of components, and the resulting distortion of the correlation structure) should

133   be considered also in a broader sense. Namely, compositional data are primarily

134   observations that contain quantitatively expressed relative contributions of parts on a

135   whole. From this perspective, the unit sum constraint is just a representation, obtained

136   without altering the source information, conveyed by ratios between the components.

137   Therefore, even with a variable sum of geochemical concentrations, resulting, e.g., from

138   designed omitting of some components, one should be aware that the relative nature of

139   data is still present and needs to be taken into account by proper statistical processing, as

140   exemplified below by CoDA (Pawlowsky-Glahn et al., 2015).

141   **Point maps – the data "as is"**

142 It is often thought that raw one-component maps report "what is there", that they report a

143 sort of "objective ground truth". However, Reimann (2005) and others (Reimann and

144 Filzmoser, 2000; Reimann and Garrett, 2005; Reimann et al., 2005; Reimann et al., 2008)

145 highlight that to give sense to that set of spatially dispersed values, space-dependent

146 geochemical processes must be interpreted, and that these are highlighted with a proper

147 representation. Reimann (2005) discusses the advantages and disadvantages of such

148 different representations. Reimann (2005) concludes that, actually, the most important

149 issue becomes the scaling chosen to define the points (colour, size and symbol). Splitting the

150 data into groups (classes) on the basis of order statistics in exploratory data analysis (such as

151 the quartiles of boxplots, or other percentiles), Reimann suggests, may shed light on the

152 spatial structure that reflects at least a number of these processes in a map. Figure 1 shows

153 how impactful this choice of symbol/colour scaling can be with a real dataset. Obviously,

154 equidistant colour scales do not necessarily yield the most easily interpretable maps,

155 neither in raw (or "absolute") nor in logarithmic values. The proportion of the entire study

156 area actually measured by the dataset is largely exaggerated by the size of the dots. The

157 conclusion is that it may be more appropriate to use quantile-based intervals (following the

158 guidelines on scaling as discussed in Reimann, 2005) to present geochemical data to give an

159 initial assessment of the distribution of elemental concentrations. Readers are referred to

160 Reimann (2005) for further details on methods used to establish intervals in data scaling. It

161 should be pointed out that this choice of scaling does not address the problems resulting

162 from the relative character of elemental concentrations, an aspect that will be explored

163 later.

A)

B)

C)
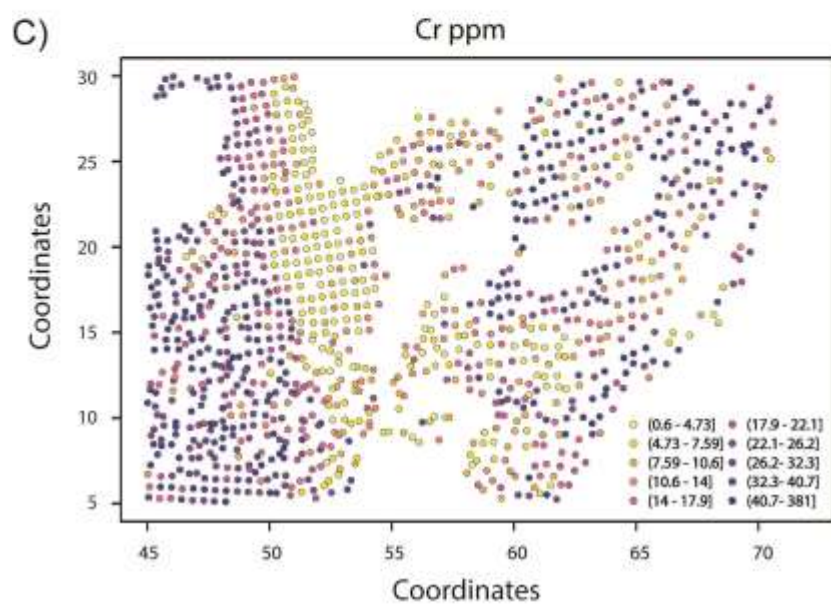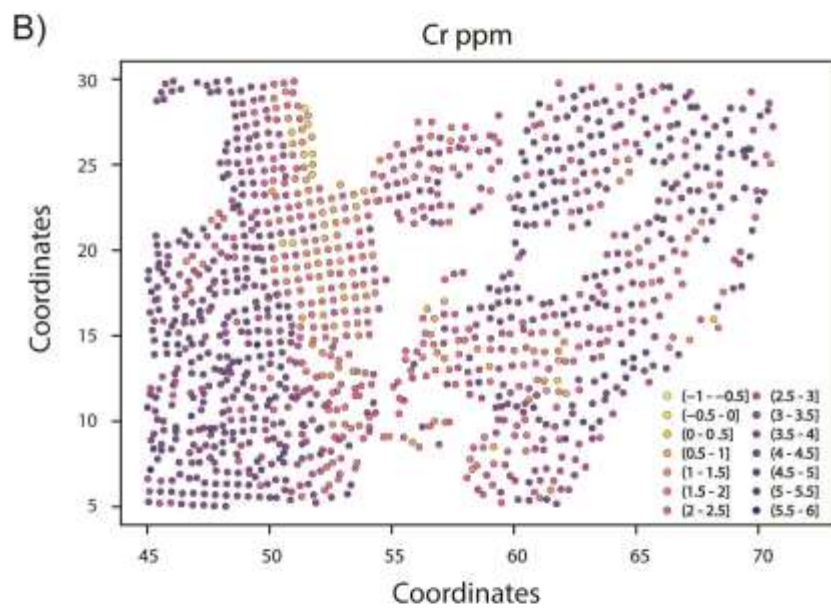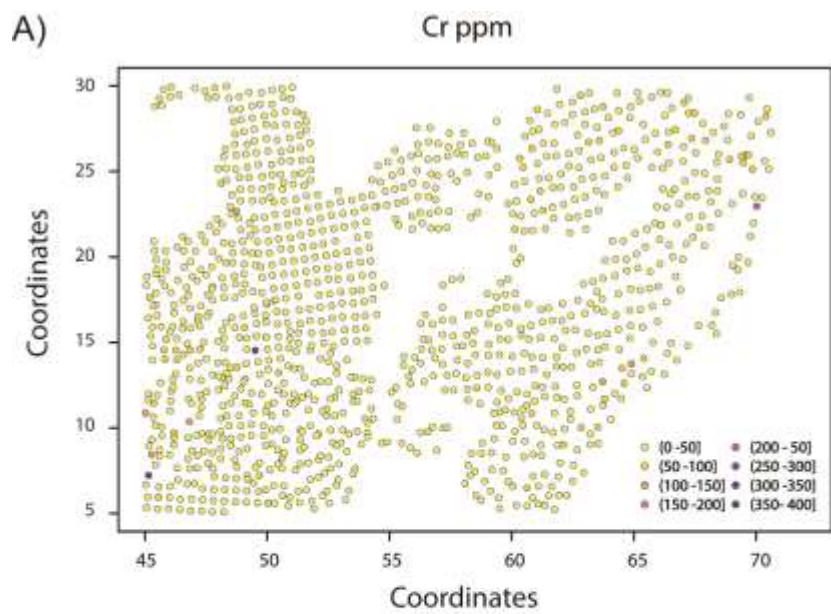
165   Figure 1: Three representations of Cr ppm concentrations from the GEMAS European soil

166   dataset (Reimann et al. 2014a,b), changing the dependence of the colour scale: (A) using

167   equidistant classes in the raw value scale, (B) using equidistant classes in the log-

168   transformed value scale, and (C) using quantile based (equal probability) classes in the raw

169   value scale. Note that the units are latitude and longitude coordinates but the orientation of

170   the image has been changed to highlight changes that are dependent on the colour scale

171   used.

172

173   **Geospatial continuity – the interpolated map**

174   It is becoming increasingly common for geochemical atlases (online and published printed

175   versions; Reimann et al., 2003, 2014a,b; Young and Donald, 2013) and research publications

176   to produce smoothed (colour) surface maps because most end-users have come to expect a

177   "nice looking" output map. Essential to the generation of an interpolated map, is the

178   demonstration of a spatially coherent pattern through the generation of semi-variograms

179   and other available geostatistical tools. The issues relating to spatial interpolation are well

180   documented elsewhere (e.g. Li and Heap, 2011) and are distinct from the question of the

181   compositional justification for single component maps, which is the focus here.

182   Nevertheless, it should be noted that when using raw 'point data' the interpolation method

183   should be chosen to ensure several aspects relevant to compositional data, in as much as

184   the interpolated values should (1) remain in the real positive space, and (2) not sum up to

185   more than 100%. Thus interpolated maps should not be obtained separately component by

186   component, as this cannot guarantee adherence to CoDA constraints (as outlined earlier).

187   Joint interpolation methods able to cope with these constraints should be preferred (for

188   example Tolosana-Delgado and van den Boogaart, 2013).

189

190   **Do compositions actually represent absolute or relative abundances?**

191   The central question of this paper is whether a single component map can support an

192   accurate interpretation on its own. As mentioned before, it is often thought that these

193   geochemical maps do actually represent absolute abundances. However, geochemical data

194   are just a particular case of compositional data, and compositional variables carry only

195   relative information of a part to a whole. This can be seen in the conventional concentration

196   units, such as mg/kg, being essentially a constant multiplied by a dimensionless unit: 1

197   mg/kg = 1 mg/1,000,000 mg, or 10^{-6} mg/mg, or simply 10^{-6}. At every sampled or

198   interpolated point, a concentration value provides information only on the relative weight

199   of one particular element to the total (Tolosana-Delgado & van den Boogaart, 2013),

200   regardless of whether we know all, some or none of the other parts of the whole. This is

201   illustrated in Figure 2 where a lower chromium (Cr) east-west zone is clearly seen in the

202   quantile scaling plot (Fig. 2a). However a corresponding increase in silica oxide (SiO_2) in the

203   same area, (as shown in Fig.2b) shows that any sensible interpretation must consider some

204   sort of possible dilution by SiO_2 that may be related to a change in geology or soil type for

205   example . This is a consequence of the constant sum (closure) effect (Chayes, 1960): if one

206   component increases (due to the addition of a pollutant or a diluent) the mass of all other

207   components remains unaffected, and their apparent proportions (concentrations) must

208   decrease. One solution to this problem is to use element ratios, rather than raw

209   concentrations, because ratios are free of any such effect.

210
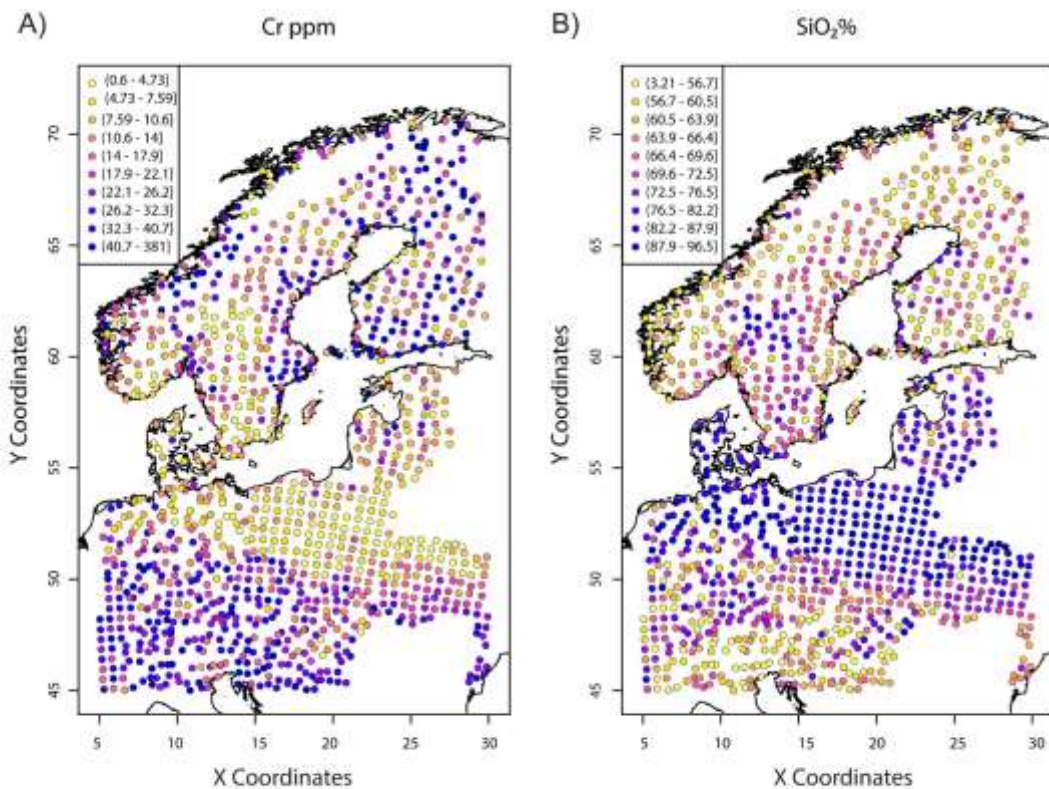
Figure 2: Illustrative maps from the GEMAS European soil dataset (Reimann et al. 2014a,b):

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.

214

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

221   soils are in residential or industrial sites. In Europe a large variety of soil guideline values for

222   agricultural and grazing land soils exists (see Appendix A in Reimann et al., 2014b).

223   Another way of looking at the relative or absolute scale is to consider a simple process

224   affecting a material. Let us imagine a fresh rock with 20% potassium oxide (K_2O) and its in-

225   situ weathered product having 10% of this oxide: what is relevant, that it lost 10% of K_2O or

226   half of it? What would be the most reasonable percentage of K_2O in the weathered material

227   if that original concentration would have been 12%? The correct answer is 6%, and not 2%.

228   These considerations are nevertheless often clearer for trace elements: if a sample has 20

229   ppm of element A and another sample, genetically derived from the first, has 10 ppm of it,

230   half of element A has been removed in the process of forming the second sample. The

231   relative scale effect applies mostly for components with lower concentrations, e.g., for trace

232   elements.

233

234   **Compositional data are always multivariate**

235   In the preceding example where the distribution of Cr was mapped (Figs. 1 and 2) it became

236   evident that any interpretation of the variation in Cr should include an awareness of the

237   dilution effect of other elements and should be done on the basis of ratios such as Cr/ SiO_2.

238   The component increase considered above and its effect on all other components may

239   appear in many other one-component geochemical maps as (high or low) outlier(s) or

240   anomalous value(s). The issue is that geochemical components form a multivariate system

241   that should be analysed as a whole, not component-wise. The following examples from the

242   regional Tellus Survey (Young and Donald, 2013) highlight this aspect. Figure 3 shows the

243   north east region of Northern Ireland where Dalradian metasediments are covered with

244   Palaeogene flood basalts. The basalt bedrock shows high nickel (Ni) concentrations (Fig. 3a)

245   with a low Ni area in the uppermost north east corner related to metasedimentary bedrock.

246   However, lower Ni values are also found in areas where the basaltic bedrock is covered by

247   peat rather than conventional soils (Fig. 3b). A similar effect is seen for zirconium (Zr),

248   though it is enriched in the metasedimentary rocks (Fig. 3c). In areas of consistent bedrock,

249   these low areas of both Ni and Zr are dilution signals formed by high volatiles (represented

250   by Loss On Ignition, LOI %) in blanket peat areas. The variation in geochemical maps can

251   therefore be influenced by dilution, enrichment or depletion in other elements, rather than

252   just by the processes specific to the element mapped. We will show later that maps based

253   on (log)-ratios can avoid this weakness.
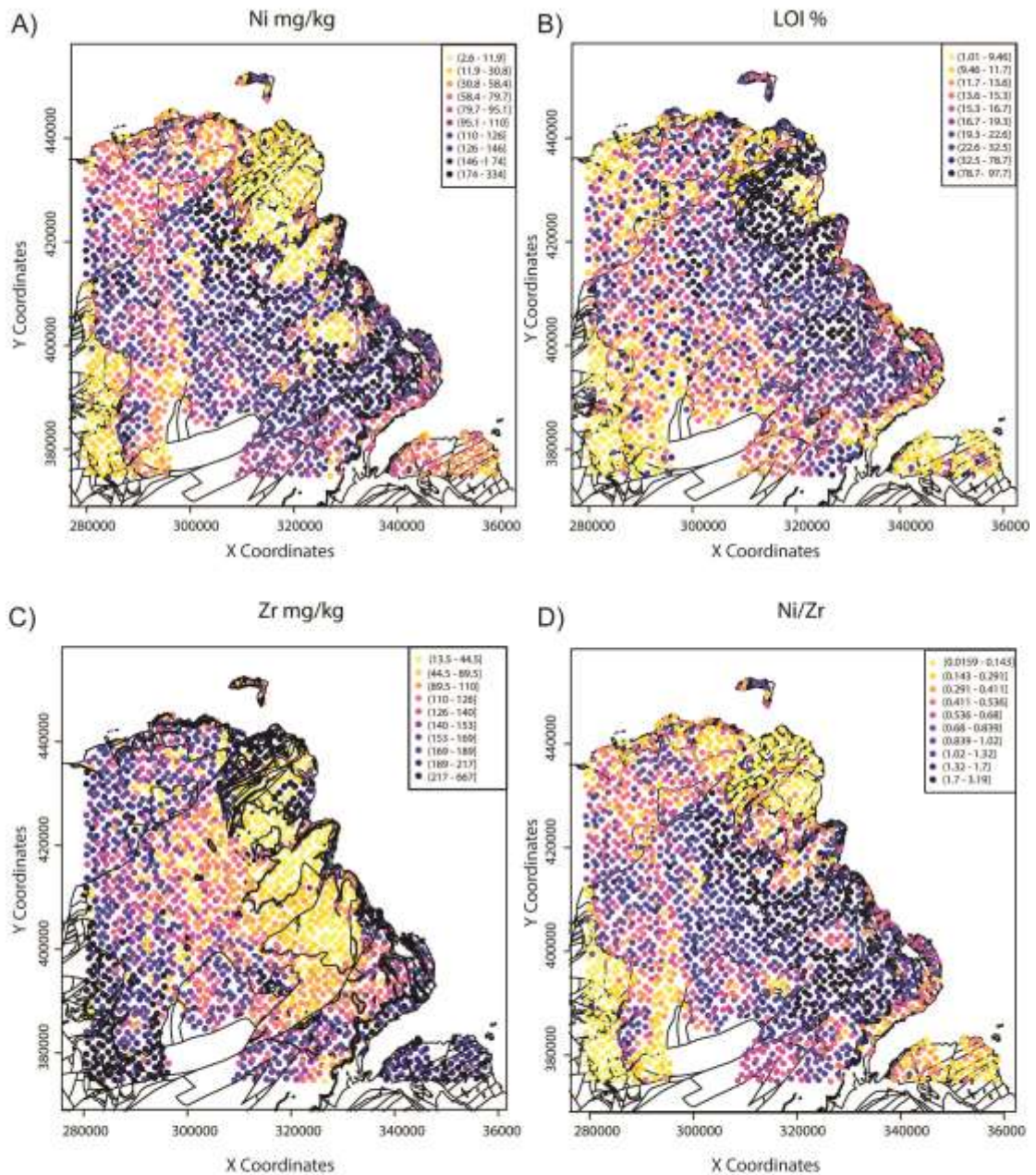
254

255   Figure 3: Four map representations using the Tellus soil survey (Young and Donald 2013)

256   showing the north east region of Northern Ireland: A) Ni mg/kg, B) Loss on Ignition (LOI %),

257   C) Zr mg/kg and D) Ni/Zr. The units are coordinates in the Irish Transverse Mercator Grid

258   (EPSG:29903).

259

261 **Generalities**

262 A common way to treat the compositional nature of geochemical data has been the

263 application of log-ratio techniques (Aitchison, 1986; Filzmoser and Hron, 2011; Pawlowsky-

264 Glahn and Buccianti, 2011; Reimann et al., 2012; van den Boogaart and Tolosana-Delgado,

265 2013), due to the fact that the ratios between the components are unaffected by constant

266 sum closure effects related to the relative nature of data (Egozcue and Pawlowsky-Glahn,

267 2011).

268 It is important to note that inherent compositional data properties apply to major, minor

269 and trace elements, so there is no reason to distinguish them methodologically, although

270 the effects of closure and other compositional properties (such as the relative scale) depend

271 on abundance and the number of analysed components. In particular, the closure effect can

272 be expected to be higher with a lower number of components and for the major

273 components, while the relative scale will apply mostly in the case of trace elements

274 (Pawlowsky-Glahn and Buccianti, 2011; Egozcue and Pawlowsky-Glahn, 2011; van den

275 Boogaart and Tolosana-Delgado, 2013). Indeed, the closure effect results from distortion of

276 the covariance structure of closed data (Aitchison, 1986), where the sum of covariances of

277 an element to the other components is equal to the negative value of its variance;

278 accordingly, for the extreme case of two-part (closed) compositions the correlation would

279 always equal -1.

280 Advancing towards the conclusions of this paper, the compositional alternative to single

281 component maps will be maps of certain well-chosen log-ratios of two or several

282   components. The next subsections explain the basics of these transformations, followed by

283   the geochemical and statistical rationale behind each proposed strategy.

284

The log-ratio transformations

286   Standard logarithmic transformations are not sufficient because it is the ratio to another

287   element that deals with the closure restraint. The use of one to one ratios does eliminate

288   the initial closure effect but ratio correlations are known to be subject to restrictions or

289   "spurious" correlations. Several families of log-ratio transformations exist in the literature.

290   Aitchison (1986) introduced the pairwise log-ratio transformation (pwlr), the additive log-

291   ratio transformation (alr) and the centred log-ratio transformation (clr), along with other

292   transformations which are not relevant for geochemical purposes. Later, Egozcue et al.

293   (2003) proposed the isometric log-ratio (ilr) transformation. None is inherently better than

294   the other, each has advantages and disadvantages. The following explanations assume that

295   the composition $\mathbf{x} = [x_1, x_2, \dots, x_D]$ involves D elements (with boldface indicating a vector

296   and the square brackets enclosing the concentrations x of elements $1, 2, \dots, D$).

297   The pairwise log-ratio transformation (pwlr) takes *all* possible pairs of elements and

298   computes their log-ratios.

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

299   The pwlr contains all the elemental information in compositional data. The treatment of

300   below detection values and zeros in compositional data is beyond the scope of this paper

301   but readers are directed to Pawlowsky-Glahn and Buccianti (2011), van den Boogaart and

302   Tolosana-Delgado (2013) and Filzmoser and Hron (2011) for further information on this

303   subject. This baseline can be used to construct more sophisticated log-ratios, as will be

304   shown later. The pwlr has the advantage that each transformed coefficient is quite simple

305   and typically interpretable on its own. Its main inconvenience is the large number of

306   resulting coefficients, and their high redundancy. For instance even after keeping only one

307   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

308   to consider $D \cdot (D-1)/2 = 780$ pairwise log-ratios. With regard to redundancy, it can be

309   seen that any row or column of that matrix can be obtained from additions and subtractions

310   from the elements of any other row or column, e.g. $\xi_{ij} = \xi_{ik} + \xi_{kj}$ for any three

311   components i, j, k. Thus, for certain tasks it will be sufficient to take one of the rows or

312   columns of the pwlr.

313   The pwlr allows us also to introduce an alternative to correlation to be used as a criterion of

314   statistical association, given the spurious character of raw Pearson correlation coefficients in

315   a compositional context. This is called the variation matrix, i.e. the variance of each of the

316   entries of the pwlr matrix: $t_{ij} = \text{Var}[\xi_{ij}]$. The variation is small for pairs of components

317   which are highly proportional, and it increases as the two components are less and less

318   dependent on each other.

319   The additive log-ratio transformation (alr) takes just one of the rows or columns of the pwlr

320   (and removes the constant zero), typically the last column:

321
$$\text{alr}(\mathbf{x}) = \left[\ln\frac{x_1}{x_D} \quad \ln\frac{x_2}{x_D} \quad \cdots \quad \ln\frac{x_{D-1}}{x_D}\right] = [\xi_{iD}],$$

322   here arranged in a row to save space. Its advantages are the low number of log-ratios

323   capturing all the information, and the simplicity of its interpretation. Its disadvantage is the

324   fact that some results might depend on the chosen common denominator x_D, and therefore

325   this must be chosen carefully. The choice usually depends on geochemical context (e.g.,

326   Thomas and Aitchison, 2005), but subjectivity cannot be completely avoided. Another

327   consideration is that the denominator should be chosen with a minimum number of, and

328   preferably no, censored values. This ensures that the log-ratio represents a realistic range of

329   values without interference of instrumentation or limits in the analytical protocols. The alr

330   transformation has as well certain geometric inconveniences (Egozcue et al., 2003; Egozcue

331   and Pawlowsky-Glahn, 2006) not relevant for this contribution.

332   Another commonly used transformation is the centred log-ratio (clr) transformation. Instead

333   of choosing a common denominator, this log-ratio transformation represents each element

334   component as a ratio to a central value:

335
$$\text{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g(\mathbf{x})} \quad \ln \frac{x_2}{g(\mathbf{x})} \quad \cdots \quad \ln \frac{x_D}{g(\mathbf{x})} \right],$$

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

337
$$g(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i} = \exp\left(\frac{1}{D} \sum_{i=1}^{D} \ln x_i \right).$$

338   The somewhat artificial role of the geometric mean becomes clear if we consider that each

339   clr variable represents nothing more than the average of all pairwise log-ratios with one of

340   the elements, i.e. the averaged sum of a column (or a row) of the pwlr matrix,

341
$$\ln \frac{x_i}{g(\mathbf{x})} = \frac{1}{D} (\xi_{i1} + \xi_{i2} + \cdots + \xi_{iD}).$$

342   Note that this averaged sum of all pairwise log-ratios contains the totality of the variability

343   of the elemental composition with respect to the set of components that was considered.

344   Its main advantage is that each transformed score can be related to one component, which

345   might have some interpretability. However, the clr scores always depend on the geometric

346   mean in the denominator, so they cannot be simply interpreted as the original element of

347   interest independently of the other elements. Thus, just as for the individual components,

348   any single component of the clr transform will depend on the notion of the "total"

349   (meaning, the set of all components considered) and can be influenced by dilution,

350   enrichment and depletion in these components. This is the greatest disadvantage of the clr

351   transform. It should be noted that the clr has some desirable geometric characteristics not

352   relevant for this contribution (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2006). It

353   is important to note that the pwlr, alr and clr transforms do not form orthogonal bases. In

354   the case of the alr transform, the resulting ratios may not reveal the same patterns when

355   there is a change of denominator. For the clr transform, scatterplot matrices will have an

356   intrinsic distortion that may or may not reveal patterns that are meaningful.

357

358   **Log-contrasts and the isometric log-ratio transformation**

359   All transformations mentioned before can be seen as vectors of some scores that are

360   computed as *log-contrasts*, i.e. linear combination of components in log-scale:

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

362   the coefficients of which are subject to $\omega_{i1} + \omega_{i2} + \cdots + \omega_{iD} = 0$. Log-contrasts form

363   special one-dimensional "projections" of compositions. Their result is scale invariant, i.e. it

364   does not depend on the particular units chosen for the components (mg/kg, ppm,

365   proportions, etc.), and have many other desirable stability properties required by CoDA

366   (Aitchison, 1986). Although log-contrasts seem to be quite complex mathematical objects,

367   they have a clear geochemical interpretation in connection with chemical equilibrium. If we

368   assume a reaction $aA + bB \rightleftharpoons cC + dD$, then the logarithm of the equilibrium constant is

369   one of such log-contrasts, $\xi = \omega_A \ln A + \omega_B \ln B + \omega_C \ln C + \omega_D \ln D$, with $\omega_A = \omega_B = -1$

370   and $\omega_C = \omega_D = +1$, and it is directly proportional to the Gibbs Energy of the reaction $\Delta G \propto$

371    $\ln K_{eq}$. Conversely, each log-contrast might be interpreted as a sort of reaction. That might

372   not be directly meaningful for solid sample geochemistry datasets, but it still offers a way to

373   interpret log-contrasts. Note that also the components of the above introduced alr and clr

374   transformations can be seen as log-contrasts. Accordingly, for i-th alr variable ($i =$

375    $1, \dots, D-1$) we get

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

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

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

379   The last log-ratio transformation mentioned is the (family of) ilr transformations (Egozcue et

380   al., 2003). These are formed by $D-1$ log-contrasts which are computed using vectors of

381   coefficients $\boldsymbol{\omega}_i = [\omega_{i1}, \omega_{i2}, \dots, \omega_{iD}]$ which are orthogonal to each other. From all possible

382   ways of obtaining such orthogonal log-contrasts, the relevant ones for this contribution are:

383   principal components of the clr-transformed data (Aitchison, 1982; 1986), and balances

384   (Egozcue et al 2003; Egozcue and Pawlowsky-Glahn, 2005). Balances are simply (normalised)

385   log-ratios of the geometric means of two groups of elements. One can obtain an ilr based on

386   balances by choosing a binary hierarchy of association of elements (or a binary partition),

387 such as the one represented in Figure 4. For instance the four log-ratios necessary to

388 describe the behaviour of the subcomposition (Fe_2O_3, V, Cr, Co, Ni) in this figure are:

389
$$\xi_1 = \frac{1}{\sqrt{2}} \ln \frac{Fe2O3}{V}, \qquad \xi_2 = \frac{1}{\sqrt{2}} \ln \frac{Co}{Ni}, \qquad \xi_3 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{Cr}{\sqrt[2]{Co \cdot Ni}},$$

390
$$\xi_4 = \frac{\sqrt{6}}{\sqrt{5}} \ln \frac{\sqrt[2]{Fe2O3 \cdot V}}{\sqrt[3]{Co \cdot Ni \cdot Cr}}$$
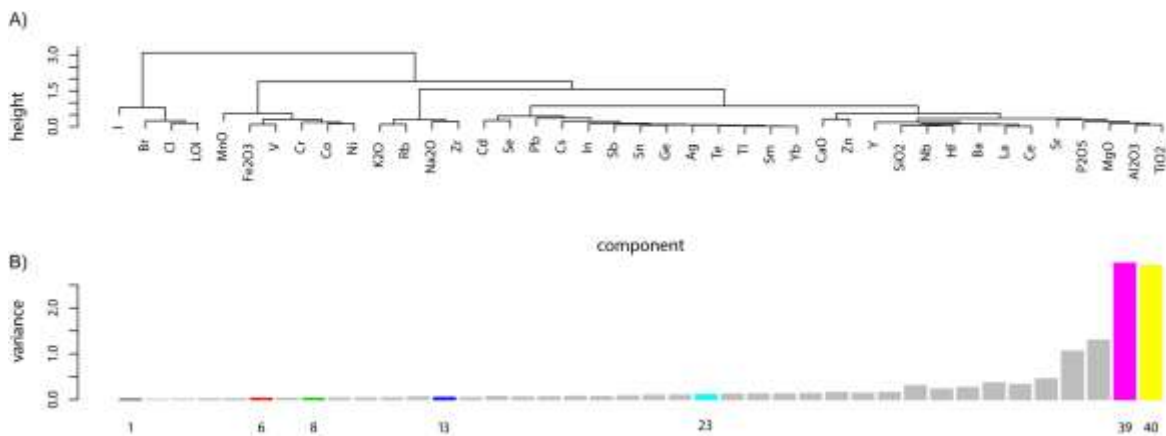
391

392 Figure 4: A) An ilr balance dendrogram and B) associated component variance obtained by

393 choosing a binary hierarchy of association of elements (or a binary partition). The procedure

394 is detailed in van den Boogaart and Tolosana-Delgado 2013).

395

396 In general applications of CoDA, the constants of proportionality are necessary to be able to

397 compare the balances between them. However, in the case of mapping and given that we

398 are going to use equal-probability intervals (Reimann, 2005), these constants are

399 superfluous. Interested readers are referred to Pawlowsky-Glahn and Egozcue (2005) and

400 Pawlowsky-Glahn and Buccianti (2011).

401

402 PROPOSED COMPOSITIONAL ALTERNATIVES

403   Univariate maps are actually produced in the hope that they might have a direct

404   interpretation with respect to specific aspects or processes occurring in the studied region.

405   However, we have argued that this may not be the case, because of the inherent

406   multivariate nature of compositional data and the spurious variations induced by the

407   closure: any process influencing a component will modify all other components (and their

408   maps). As we cannot display or read maps displaying high-dimensional multivariate

409   information, we are left with the question of how to choose interpretable univariate (many

410   single element maps) or low variate (oligovariate) maps (fewer maps displaying two or more

411   variables). All these maps will be representations of certain log-contrasts. The challenge in

412   finding good maps is twofold: there are many more log-contrasts than there are elements in

413   the dataset, and the most appropriate log-contrasts to use will always depend on the aim of

414   the analysis undertaken (Pawlowsky-Glahn and Buccianti, 2011). Using single elements or

415   log-contrasts inherently defines an approach using sub-compositions for gaining insight into

416   geochemical processes. This may work for recognizing obvious processes, but not for

417   processes that are under-sampled or "swamped" by dominant processes.

418   Three approaches for selecting good ratios for various aims are discussed and illustrated

419   with a case study example. These provide alternative approaches to presenting a single

420   elemental component map that honour the compositional nature of the geochemical data

421   and offer interpretable mapped outputs. The feature(s) to be mapped can be based on

422   either (1) a geochemical understanding of the processes under consideration (knowledge-

423   driven log-ratios based on geochemistry), (2) established mathematical projections (e.g.

424   pwlr or clr), or (3) based on the data using statistically determined projections that best

425   capture certain effects, patterns and elemental associations (data-driven projections). These

426   are illustrated using the regional soil sediment geochemical survey covering Northern

427   Ireland, generated as part of the Tellus Survey (Young and Donald, 2013). The soil sediment

428   dataset used comprises 6862 observations of 19 geochemical variables (Fig. 5).
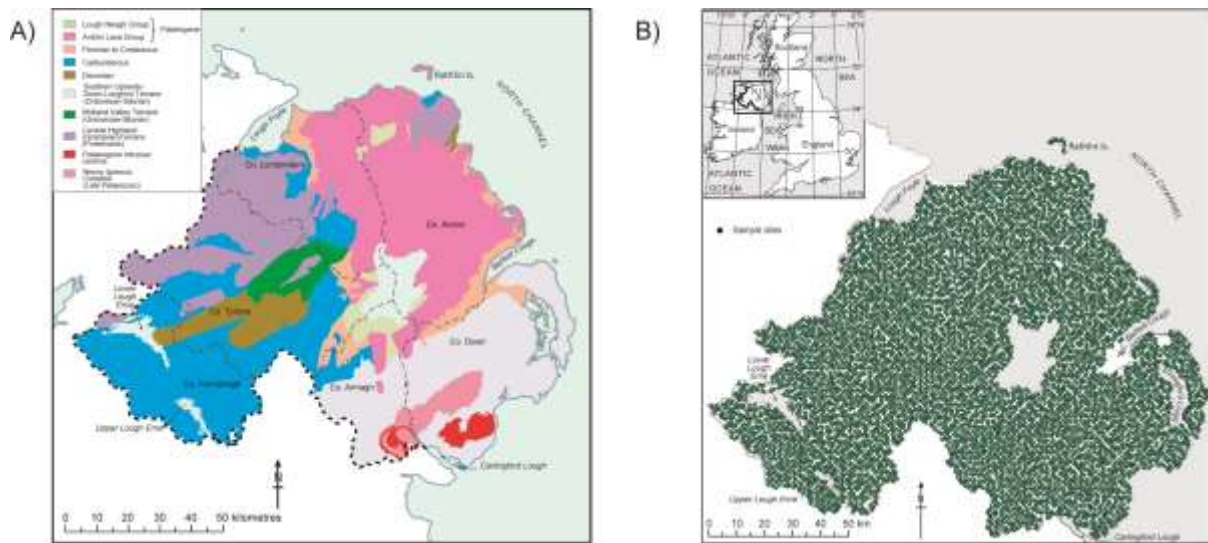
430   Figure 5: Case study A) Simplified geology for Northern Ireland (adapted after Mitchell,

431   2004) and; B) Sampling locations for the Tellus Geochemical Survey (GSNI).

432

433   **1. Knowledge-driven (log)-ratios**

434   In relation to the problem of Ni and Zr dilution by LOI (shown previously in Fig. 3), geological

435   knowledge would suggest that it may be more useful to analyse a simple log-ratio, or a

436   certain balance related to a reasonable equilibrium. Indeed, in this example using the (log-)

437   ratio Ni/Zr reduces the masking effect of peat cover (LOI dilution effect) and illustrates a

438   greater delimitation of the Palaeogene basalt-Dalradian metasediment boundary than any

439   of the three single component maps (Fig. 3d). However, although being of possibly highest

440   interpretational potential, ratio or log-ratio maps present one challenge: we are not used to

441   reading them. This can be resolved by practice and it is advocated that the expected effect

442   of less methodological artefacts promises this investment to be worth the time.

443   Of course, not everything relevant is always expressible in terms of (log-)ratios. In mining

444   there is a need to predict excavation and processing costs per mass recovered, hence the

445   ratio of value element versus mass to be mobilized is relevant, i.e. a quantity proportional to

446   absolute concentration of the value element. In environmental studies, it is important to

447   know how much mass of the toxic elements is present, i.e. proportional to its absolute

448   concentration again. Thus the mass of the element per bulk mass can be a useful proxy.

449   However, in both cases (and in many more examples), other quantities might highlight other

450   important aspects of the problem, quantities which are often ratios or log-ratios. Mining

451   requires the ratio of gangue elements against ore elements, because the product's quality

452   and value depend on them. In environmental issues, well-chosen log-ratios may incorporate

453   bioavailability effects (Barsby et al., 2012) and interactions with other components such as

454   pH or HCO_3^-, etc. Environmental or health hazards typically depend both on absolute values

455   and on ratios, particularly considering bioavailability fractions. Chemical reactions like those

456   mentioned above also play an important role in mobilizing or trapping these elements. For

457   soils and sediments that have been weathered or mixed/diluted with other material, no

458   element shows the same percentages as in the background or source rock, but the ratios of

459   elements unaffected by the mixing (or by pollution or weathering) are preserved. In

460   weathering, especially of magmatic plagioclase-rich rocks, all existing measures of alteration

461   have been defined as (log)ratios, perhaps because quantifying absolute losses of mass

462   would be nonsensical (the removed mass was washed out of the rock-weathered product

463   system, we do not know the starting concentrations of elements in the source rock, the

464   degree of dilution generated from more or less mixing material with the weathered rock is

465   unknown, etc.). Typical practice tends to be to use the ratio of mobile to immobile

466   elements. In studies involving mineral reactions (igneous, metamorphic, diagenesis), Pearce

467   Element Ratios (PER) have been successfully used to model mineral paragenesis (Pearce,

468   1968; Madeisky and Stanley, 2010). In the molar ratio scatterplots used in PER, individual

469   end-member minerals are represented as points and linear trends are interpretable in terms

470   of mineral mixtures. This has been shown to be useful for instance to interpret alteration

471   processes (e.g. Montreuil et al., 2013). Such PER can thus guide the choice of appropriate

472   knowledge-driven log-ratios to map.

473   **2. The pairwise log-ratio map**

474   The second option is to rely on one of the standard log-ratio transformations, for instance

475   the pairwise log-ratio (pwlr). The challenge in this case is the number of pairs that are

476   possible to build: for D elements there are $D \cdot (D - 1)/2$ unique maps. For example the

477   aqua regia analyses from the GEMAS survey result in ~1600 ratio maps, and the XRF

478   compositions add another ~400 ratio maps. It is impossible to present these within a

479   classical atlas publication. Although this is becoming less of a problem with interactive

480   online or electronic atlases, it is not humanly feasible to view and integrate all (or even just

481   many) of these different perspectives. Therefore, either knowledge-driven or data-driven

482   criteria will be required to choose which of this vast number of possibilities are most

483   pertinent for a specific task or problem. While it is acknowledged that in both the

484   knowledge-driven and pairwise approaches, any resultant maps may not explicitly show the

485   impact of other variables relevant for the process being investigated, these approaches do

486   avoid the potential for artefacts due to the constant sum constraint (closure).

487

488   **3. The centred log-ratio component map**

489   The usage of the clr transformation may be grounded in one of the following options: (1) to

490   highlight the role of one component, (2) to see the patterns of that element relative to the

491   average behaviour of other elements at hand in the compositional dataset (using their

492   geometric mean), and (3) to provide the first visualization towards a deeper understanding

493   of the underlying geochemical processes.

494   Mapping single clr scores seems to present a first-step simple solution for an unsupervised,

495   blind application. However, it must be noted that clr variables have severe limitations. First,

496   they cannot be interpreted the same way as the variables in the original concentration

497   units, and there is no straightforward back-transformation. A further problem of the clr is its

498   sensitivity to replacement values of censored or missing values, because the more

499   numerous trace elements (with low concentrations and more common censorship)

500   influence the geometric mean more than the fewer major elements (with high

501   concentrations and rarely censored). Therefore, if we impute too low values, we shift the

502   geometric mean to lower values. One solution could be to use a robust central value (e.g.

503   the median of all the components, a trimmed geometric mean, etc.), although this destroys

504   many of the nice properties of the clr and may cause problems for the interpretation of any

505   resulting map. More importantly, as has been mentioned, the clr depends on the set of

506   components chosen, i.e. any component cannot be analysed separately without

507   consideration of the others. This might make interpretation of clr maps difficult, especially

508   when the other components are driven by their own geochemical processes.

509   The above features are illustrated using the Tellus dataset. Figure 6a demonstrates the

510   effect of employing the whole composition or the subcomposition of major oxides,

511   respectively, for clr scores of these elements. In addition to a shifting of the distribution, a

512   result of excluding the trace elements, the positive linear correlation between values of

513   both scores is substantially disturbed for some components (e.g. in particular for Al_2O_3 ,

514   MgO, SiO_2). This is also reflected in the corresponding univariate maps (as in the case of clr

515   Al_2O_3, see Figure 6b and c). This clearly demonstrates that the selection of the

516   subcomposition is important for the construction of clr scores. Here the use of the

517   subcomposition of major oxides clr scores seems to be reasonable due to their higher

518   geochemical stability in the study area. On the other hand, by employing all elements we

519   can observe some clear regional patterns, that may indicate interesting inherent processes

520   in the data that are not so well reflected from the major oxides subcompositional clr.

521
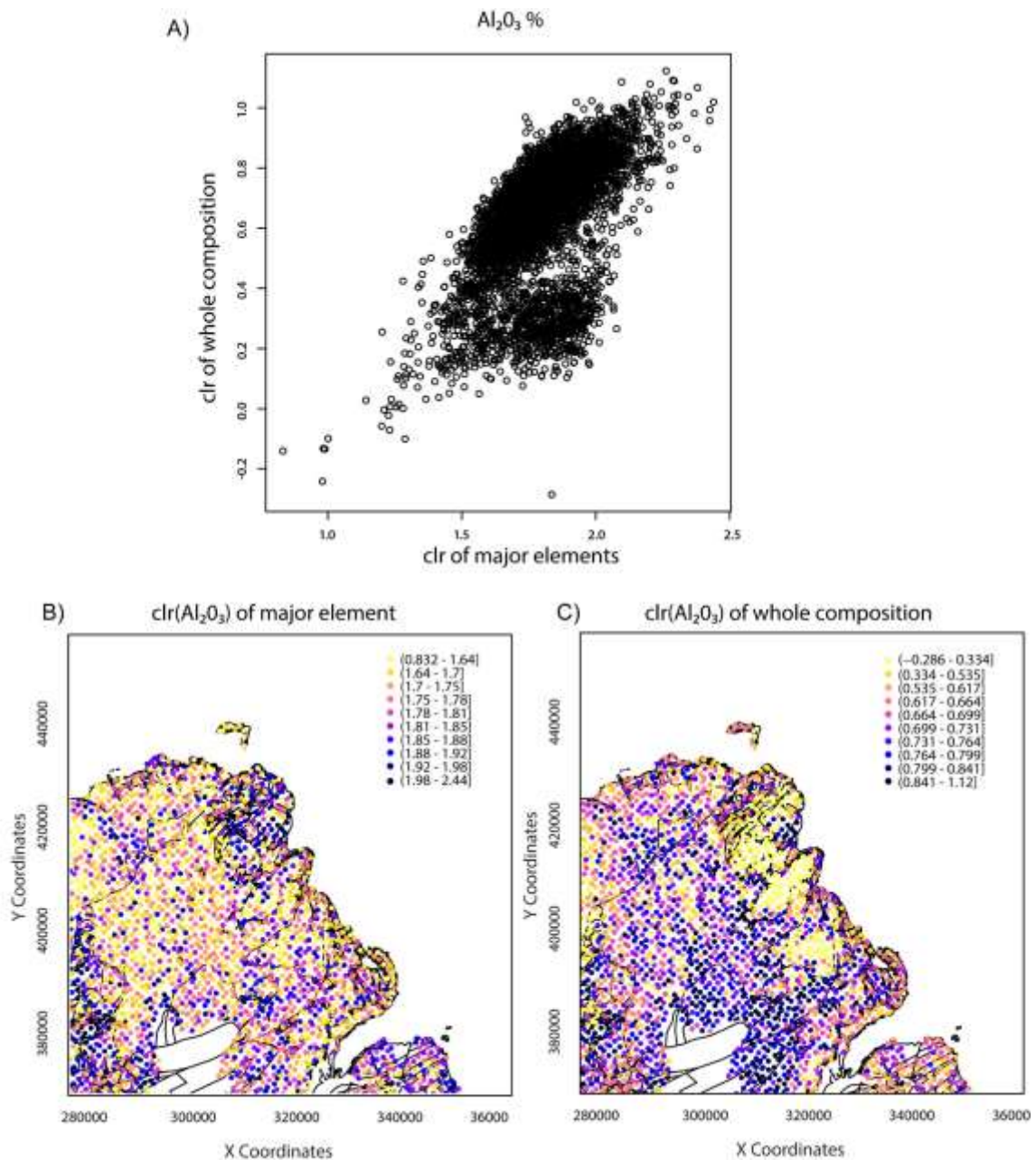
Figure 6: A) The effect of employing the clr of the whole composition and the clr of the subcomposition of the major oxides is illustrated for Al2O3. Note the substantial disturbance in the positive linear correlation. The corresponding univariate maps demonstrating the effect of employing B) the clr (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).

528

529   In the authors' opinion, using the geometric mean of a geologically meaningful subset of

530   components, selected on the basis of a geological understanding, may offer a possible way

531   to form a compromise between the two situations that we have demonstrated (Fig. 6 b and

532   c). That is using pairwise log-ratio maps (one element to another, or any knowledge-driven

533   log-ratio) and using the clr scores of the full composition including the problematic

534   components (i.e. those having a high proportion of values that are censored, close to the

535   detection limit, or rounded). The result may offer a reasonable number of single (clr score)

536   or low) component maps. For the Tellus Survey example it would seem a sensible approach

537   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,

538   with the exception of LOI and SO_3 to avoid problems with the soil water content and with

539   the large number of missing values of SO_3. Their subcompositional clr-scores could then be

540   mapped. For the remaining elements, comprising the trace elements, LOI and SO_3, the

541   balance of each element against the geometric mean of the major component

542   subcomposition could be mapped.

543   Another use of a subcompositional clr is to restrict attention to a knowledge-driven

544   selection of parts, and constructing and mapping their subcompositional clr scores. If the

545   rationale for the study is set out beforehand, then these clr maps of subcompositions should

546   be interpretable, because the subcomposition was selected and studied for a particular

547   reason. An example of this approach could be to use a subcomposition of potentially

548   harmful elements that are known to adversely affect human health (for example As, Cr, Pb

549   and Cd). In fact, this approach is reminiscent of the first method, knowledge-driven log-

550 ratios, where the interpretability of the resulting variables is assumed since there was a

551 knowledge-driven rationale for the approach.

552

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

554 Multi-element geochemical datasets are often analysed to explore patterns, links and

555 associations between elements as part of the 'Discovery Process' (Grunsky et al., 2014).

556 Multivariate methods available include principal component analysis (PCA), regression-

557 related techniques (regression analysis, analysis of variance, total and partial least squares

558 regression or canonical correlation) and grouping techniques (cluster analysis, discriminant

559 analysis), often used to explore these patterns or links. From the point of view of mapping,

560 most of these methods can actually be seen as a way to choose one (or a few) log-contrasts

561 of potential interest. There is thus a range of techniques available and the issue is to choose

562 the best one. Two families of approaches can be considered: supervised and unsupervised

563 methods. Again, these will be exemplified using the Tellus case study.

564 *Supervised Approaches*

565 In the supervised approach the goal is to find which log-contrast better relates to known

566 secondary information which is also available. For example, we may be interested in

567 identifying the log-contrast which best distinguishes between two types of geochemical

568 situations such as two terranes, or areas of peat coverage and peat free areas. In that case

569 we could start from a subset of the data, where this extra information is known (i.e. a

570 *training set*). The best log-ratio can be computed by a linear discriminant analysis derived

571 from any log-ratio transformed composition, although due to numerical issues, ilr and alr

572 are the best in this case (readers are referred to van den Boogaart and Tolosana-Delgado,

573   (2013) for further detail). Such a linear discriminant analysis has been conducted on the

574   Tellus soil geochemical dataset on the basis of areas covered by blanket peat (Figure 7). The

575   linear discriminant analysis suggests that many more elements and oxides (Ni, Rb, Hf, Nb,

576   Cr, K_2O, MnO, Fe_2O_3, some with positive weight, some with negative weight; Fig.3a) are

577   useful for blanket peat delineation than solely using LOI %, as is traditionally the case.

578   Overall, the histograms shown in Figure 7b and 7c show that the classification between peat

579   covered and peat free areas, produced by geostatistical Fisher discriminant analysis in this

580   example (Tolosana-Delgado et al. 2015), is reasonably successful. On further investigation it

581   is observed that misclassified points are found mainly at the boundaries between peat-

582   covered areas and peat-free areas and most likely related to point-polygon mismatch

583   between the Tellus data points and peat assigned polygons (discussed in more detail in

584   Tolosana-Delgado et al. 2015). A full discriminant analysis approach could be applied in that

585   results could be validated further using cross-validation, the log-contrast could be derived

586   from a training set and the discriminant power of the analysis could be evaluated.
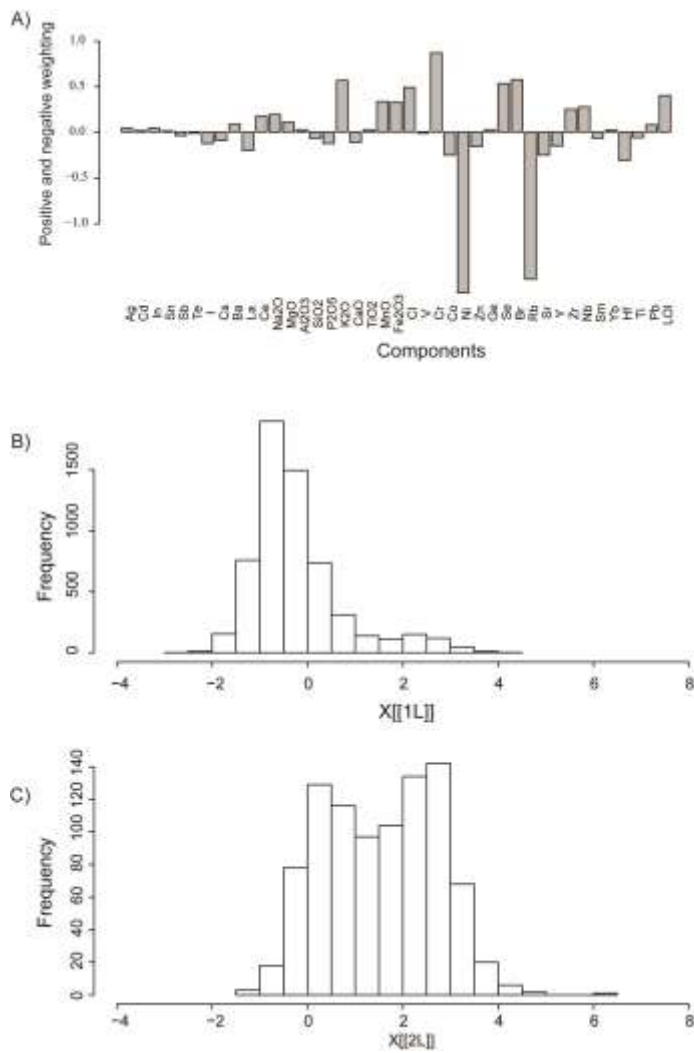
587

588 Figure 7: Illustrating a supervised method: Linear discriminant analysis using the Tellus soil

589 geochemical dataset on the basis of blanket peat; A) elemental linear discriminant analysis,

590 showing positive and negative weightings and histograms showing the classification

591 between B) peat covered and C) peat free areas, produced by the geostatistical Fisher

592 discriminant analysis. Readers are referred to Tolosana-Delgado et al. (2015) for further

593 detail.

594

595 In a similar way, if the covariable(s) of interest is (are) a numeric quantity or a set of them

596 (magnetics, gravimetrics, Th-channel radiometrics, total gamma-ray counts, proportion of

597   sand, log-ratio of silt to sand, pH, etc.), compositional regression analysis could be used in

598   order to identify the optimal log-contrast (e.g McKinley et al. in review). The composition

599   should be expressed in log-ratios (alr or ilr) and a classical regression applied between the

600   log-ratio scores and the chosen covariable. The explanatory power of the resultant model

601   can be measured by its associated R-square. Thus, beyond the classical use of regression, it

602   is possible to use this criterion to select the optimal pairwise log-ratio that maximizes the R-

603   square of the simple regression model with the covariable of interest. This requires the

604   exhaustive exploration of all possible pairwise log-ratios, though the selection of the "best"

605   is automatically done by the method and not by visual inspection. Note that these

606   techniques, although explained for one covariable or two groups, can be easily generalized

607   towards multiple groups or covariables, providing an extra direction for each added group

608   or covariable is known (van den Boogaart and Tolosana-Delgado, 2013).

609

610   *Unsupervised approaches*

611   Even in the absence of additional information, multivariate analysis techniques can help to

612   identify interesting log-contrasts to map. These often show either large variability or very

613   low variability. For example, the variation coefficient t_{ij} of two components provides a

614   measure of their association, which can be used in cluster analysis as a measure of distance

615   between variables. The application of cluster analysis for compositional data (compositional

616   Q-mode clustering) is explained in detail by van den Boogaart and Tolosana-Delgado (2013).

617   Clusters will contain elements behaving proportionally throughout the dataset. Log-ratios

618   between elements of two different clusters should thus be similar to other log-ratios of the

619   elements of the same clusters. Therefore, one of these log-ratios or a balance of one cluster

620 against the other might be representative for many log-ratios, and consequently may

621 represent a process influencing many elements in the same way. Balances of elements

622 within the cluster will filter out these large-variability effects and focus on differences

623 between elements behaving similar with respect to major processes. In a sense this is the

624 approach that geochemists instinctively use when applying simple ratios. The advantage of

625 this unsupervised CoDA approach that reflects a classical supervised approach is that these

626 balances should thus be able to highlight processes differentiating between similar

627 elements, i.e. second-order processes that would otherwise be overshadowed by the major

628 processes.

629 In the example shown in Figure 8 using the Tellus regional soil data, a cluster dendrogram is

630 constructed using all those components without missing values (shown in Figure 4).

631 Although it is acknowledged that this introduces a bias, since the treatment of below

632 detection values and zeros in compositional data is beyond the scope of this paper, this

633 simplified approach is used here to exemplify the approach. The hierarchical cluster has

634 been produced with a Ward cluster analysis with the variation matrix as distance between

635 elements, the procedure is detailed in van den Boogaart and Tolosana-Delgado (2013). Six

636 resultant ilr balances have been shown to demonstrate different features: spatial

637 randomness (Fig. 8a; Yb vs Sm); the identification of granitic bedrock areas (Fig. 8b; Nb, Hf

638 vs SiO_2); an artefact in the central region perhaps related to the analytical method of Ag

639 detection (Energy Dispersive Polarised X-Ray Fluorescence, Fig. 8c, Te, Sm, Yb, Tl vs Ag);

640 Geological bedrock units including Palaeogene basalts, a granodiorite complex and

641 Dalradian metasediments (Fig. 8d, Rb vs K_2O); Ordovician- Devonian sandstone units and

642 granitic rocks (Fig. 8e, Zr vs Na_2O); Palaeogene basalt detection and LOI-related peat

643 unsupervised detection (Fig. 8f).
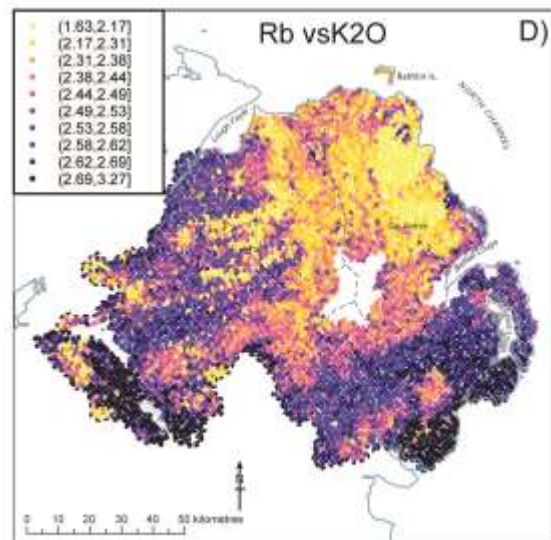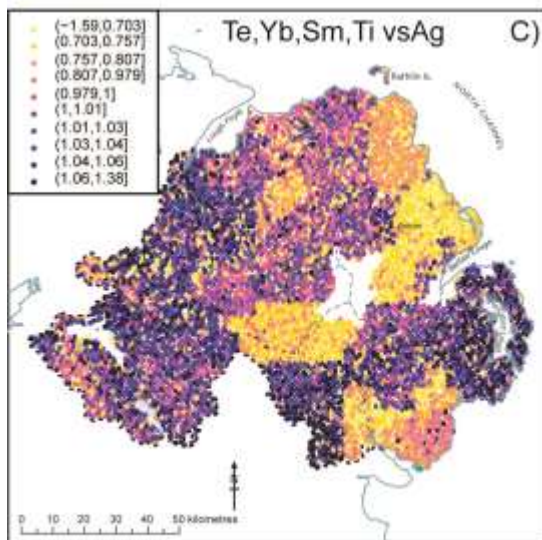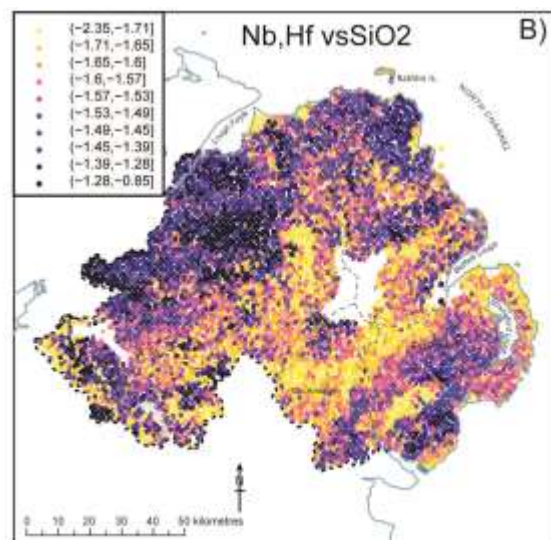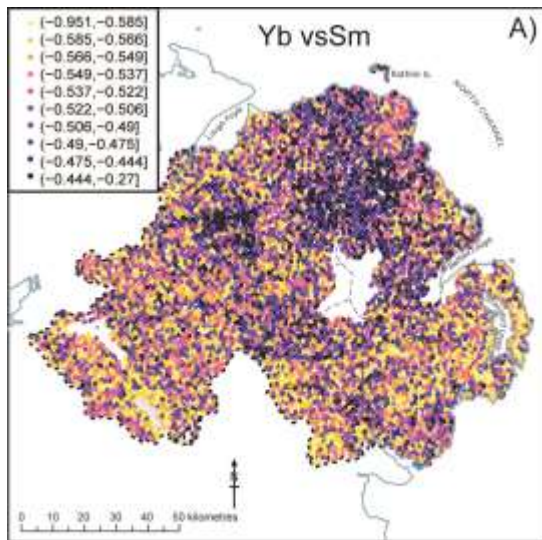
Ag,Cd,In,Sn,Sb,Tr,Cs,Ba,La,Ce,Na2O,MgO,Al2O3,SiO2,P2O5,K2O,CaO,
TiO2,Zn,Ge,Se,Rb,Sr,Y,Zr,Nb,Sm,Yb,Hf,Ti,PbvsMnO,Fe2O3,V.CR,Co,Ni

644
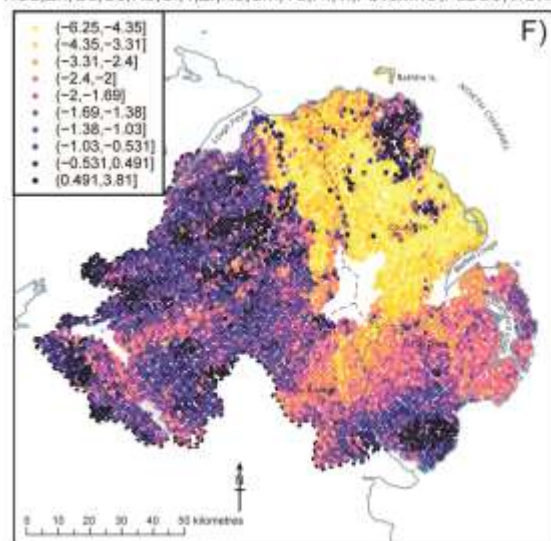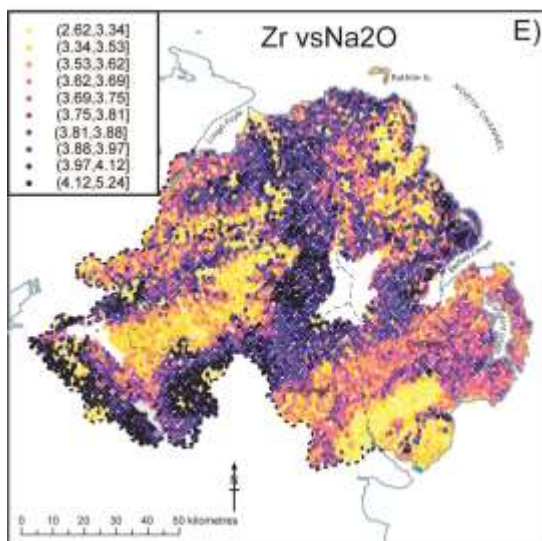
645   Figure 8: The figures illustrate an unsupervised method based on the ilr balance

646   dendrogram shown in Figure 4: Six resultant ilr balances are selected to illustrate different

647   features: A) spatial randomness (Yb vs Sm); B) granitic bedrock areas (Nb,Hf vs SiO2); C) an

648   artefact related to the analytical method of Ag detection (Te,Sm,Yb,Tl vs Ag); D) Geological

649   bedrock units including Palaeogene basalts, granodiorite complex and Dalradian

650   metasediments (Rb vs K2O); E) Ordovician- Devonian sandstone units and granitic rocks (Zr

651   vs Na2O ; F) Palaeogene basalt detection and LOI-related peat detection.

652

653   Another popular approach traditionally used for identifying interesting log-contrasts is

654   principal component analysis of the clr-transformed dataset and its graphical display, the

655   compositional biplot (Aitchison and Greenacre, 2002; Caritat and Grunsky, 2013). Within a

656   compositional approach this allows the identification of directions or log-contrasts of large

657   or small variations based on a variance-covariance matrix. Log-contrasts of strong variation

658   might be useful for identifying major processes strongly influencing many elements at the

659   same time. The directions of small variation are more useful for identifying common

660   equilibrium-like laws and anomalies by highlighting cases showing strong deviations from

661   otherwise stable log-contrasts. These directions of lesser magnitude may also represent

662   random processes in the data. Any of the methods mentioned before can be used in several

663   different ways with the purpose of focussing on specific features and aspects of the dataset.

664   In summary, there are many possible tools and approaches for selecting informative and

665   valid univariate projections of a composition. The key point is to relate them to specific

666   research questions to address pertinent geochemical and geological issues.

667   CONCLUSIONS - BEST PRACTICE AND RECOMMENDATIONS

668 The question that this paper has addressed is whether the single component elemental map

669 is meaningful: can it be interpreted as fact or is it indeed no more than a fictional

670 representation? The essence of this dilemma is that geochemical data are compositional

671 variables and as such carry only relative information. In some cases, the classical single-

672 component maps might carry some relevant information, but most of the time the

673 information-rich relative data are not properly represented. On reflection, this research

674 recommends the following:

675 2. An appropriate manner to represent "measured" raw or absolute geochemical values is in

676 the form of a single component point source or 'dot maps' with an acknowledgement about

677 dependence to all other variables. This may be useful for the purpose of geochemical

678 exploration but is critical for environmental studies where stated values fall below/above

679 guidelines. Generally, without considering or knowing the other elements of the

680 composition, we are not able to fully interpret the component of interest. That is to say that

681 involving more components in the interpretation can reveal the reason behind the relative

682 increase or decrease in the component of interest.

683 3. A complementary compositional approach is required to ensure that mapped outputs are

684 accurate and robust to interpret the underlying geochemical processes and highlight any

685 anomalies of interest. One suggestion is to look at log-ratio maps, which may include either

686 knowledge-driven log-ratios, pairwise log-ratios, any reasonably aggregated log-ratios or

687 data-driven projections, in addition to single element maps whenever possible. In this case,

688 the authors propose the following chain of representations:

689   • in paper format, the maps of clr-transformed major components together with the

690   balance of each trace component to the major components; in electronic interactive

691   formats, this can be complemented with maps of each pwlr score;

692   • maps of clr- or ilr-based principal components, those with highest variability and

693   also those with very low variability; it is worth noting that one should be careful if

694   including all elements for principal component analysis as trace ones (resulting in

695   highest variability due to relative scale of compositions) are often burdened by

696   measurement errors;

697   • maps of the balances resulting from a compositional cluster analysis based on the

698   variation matrix.

699 4. No matter whether absolute or log-ratio maps are used, most commonly it will be necessary

700   to use quantile-based colour and symbol scales to highlight the patterns captured. In log-

701   ratio maps, equal length intervals for colours and symbols are often useful as well. However

702   this is close to useless in absolute maps except to highlight a few high-value outliers.

703 5. If a compositional approach is not used, it becomes difficult to guarantee that any identified

704   pattern, trend or anomaly is not an artefact of the constant sum constraint. Nevertheless, it

705   should be noted that if the elements do not sum up to the same constant, the relative scale

706   of compositions still needs to be taken in account.

707 6. The evaluation of individual component maps is a first step towards a comprehensive

708   understanding of geochemical processes that are reflected in the material that has been

709   analysed geochemically. It is the initial step in the "Discovery Process" from which

710   geologically meaningful log-contrasts can be derived. There are so many possible supervised

711   and unsupervised data-driven methods, that it is necessary to: (1) clearly state the

712   geological/geochemical question to ask; (2) search for the appropriate statistical method

713   that can answer that question; (3) ensure that the data available are sufficient to answer

714   that question with the selected method; (4) apply the required log-ratio transformations

715   and the chosen statistical method; and (5) learn to interpret the results. This will most often

716   require statisticians, data analysts and geochemists working together.

717

729   **References**

730   Aitchison, J., 1982. The statistical analysis of compositional data (with discussion).

731   Journal of the Royal Statistical Society, Series B (Statistical Methodology), 44, 139–

732   177.

733   Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics

734   and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with

735   additional material by The Blackburn Press).

736   Aitchison, J., Greenacre, M., 2002. Biplot of compositional data. Applied Statistics, 51 (4),

737   375-392.

738   Barsby, A., McKinley, J., Ofterdinger, U., Young, M., Cave, M., Wragg, J., 2012. Bioaccessibilty

739   of trace elements in Northern Ireland. Science of the Total Environment, 433, 397-417.

740   doi:10.1016/j.scitotenv.2012.05.099

741   Birke, M., Rauch, U., Stummeyer, J., 2015. How robust are geochemical patterns? A

742   comparison of low and high sample density geochemical mapping in Germany, Journal

743   of Geochemical Exploration, 154, 105-128.

744   http://dx.doi.org/10.1016/j.gexplo.2014.12.005

745   Bolviken, B., Stokke, P. R., Feder, J., Jossang, T., 1992. The fractal nature of geochemical landscapes, J

746   Journal of Geochemical Exploration, 43, 91– 109.Caritat, P. de, Cooper, M.,

747   2011a. National Geochemical Survey of Australia: The Geochemical Atlas of

748   Australia. Geoscience Australia Record, 2011/20 (2 Volumes), 557.

749   Caritat, P. de, Cooper, M., 2011b. National Geochemical Survey of Australia: Data Quality

750   Assessment. Geoscience Australia Record, 2011/21 (2 Volumes).

751   Caritat, P. de, Grunsky, E.C., 2013. Defining element associations and inferring geological

752   processes from total element concentrations in Australian catchment outlet

753   sediments: multivariate analysis of continental-scale geochemical data. Applied

754   Geochemistry, 33, 104-126. DOI: 10.1016/j.apgeochem.2013.02.005

755   Carranza, E.J.C., 2010. Mapping of anomalies in continuous and discrete fields of stream

756   sediment geochemical landscapes. Geochemical Exploration Environmental Analysis,

757   10, 171–187.

758   CCME (Canadian Council for Ministers of the Environment), 2014. Canadian Environmental

759   Quality Guidelines. http://ceqg-rcqe.ccme.ca/

760   Chayes, F. 1960. On correlation between variables of constant sum. Journal of

761   Geophysical Research 65 (12), 4185–4193.

762   Cheng, Z., Xie X., Yao W., Feng J., Zhang Q., Fang J. 2014. Multi-element geochemical

763   mapping in Southern China. Journal of Geochemical Exploration, 139, 183–192.

764   Chiprés, J.A., Salinas, J.C., Castro-Larragoitia, J., Monroy, M.G., 2008. Geochemical mapping

765   of major and trace elements in soils from the Altiplano Potosino, Mexico: a multiscale

766   comparison. Geochemical Exploration Environmental Analysis, 8, 279–290

767   Darnley A.G., Björklund A., Bølviken B., Gustavsson N., Koval V., Plant J.A., Steenfelt A.,

768   Tauchid M., Xie X.J. 1995, A global geochemical database for environmental and

769   resource management-recommendations for international geochemical mapping—

770   final report of IGCP Project 259. Earth Sciences, UNESCO Publishing, 19.

771   Drew, L.D., Grunsky, E.C., Sutphin, D.M., Woodruff, L.G., 2010 Multivariate analysis of the

772   geochemistry and mineralogy of soils along two continental-scale transects in North

773   America, Science of the Total Environment, 409, 218-227,

774   doi:10.1016/j.scitotenv.2010.08.004.

775   Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., C. Barceló-Vidal (2003) Isometric

776   log-ratio transformations for compositional data analysis. Mathematical Geology

777   35(3), 279–300.

778   Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in

779   compositional data analysis. Mathematical Geology, 37, 795-828.

780   Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. In

781   Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V., (Eds), Compositional Data

782   Analysis in the Geosciences. From Theory to Practice. Geological Society, London,

783   Special Publications 264,145-159.

784   Egozcue, J.J., Pawlowsky-Glahn, V., 2011. Basic concepts and procedures, in V. Pawlowsky-

785   Glahn and A. Buccianti, (Eds.), Compositional Data Analysis. Theory and Applications.

786   Wiley, Chichester, pp. 12-28.

787   EGS (EuroGeoSurveys Geochemistry Working Group), 2008. EuroGeoSurveys Geochemical

788   mapping of agricultural and grazing land soil of Europe (GEMAS) - Field manual. NGU

789   Report 2008.038. ISSN 0800-3416. Geological Survey of Norway.

790   Filzmoser, P., Hron, K., 2011. Robust statistical analysis, in V. Pawlowsky-Glahn and A.

791   Buccianti, (Eds.), Compositional Data Analysis. Theory and Applications. John Wiley &

792   Sons, Chichester (UK), 59-72.

793   Friske, P.W.B., Rencz, A.N., Ford, K.L., Kettles, I.M., Garrett, R.G., Grunsky, E.C., McNeil, R.J.,

794   Klassen, R.J., 2013. Overview of the Canadian component of the North American Soil

795   Geochemical Landscapes Project with recommendations for acquiring soil

796   geochemical data for environmental and human health risk assessments

797 Geochemistry: Exploration, Environment, Analysis*, 13, 267-283,*

798 *doi:*10.1144/geochem2012-140

799 Grunsky, E.C., Drew, L.J., Sutphin, D.M., 2009. Process recognition in multi-element soil and

800 stream-sediment geochemical data. Applied Geochemistry, 24, 1602–1616.

801 Grunsky, E.C. 2010.The interpretation of geochemical survey data; Geochemistry,

802 Exploration, Environment Analysis, 10(1), 27-74.

803 Grunsky, E.C., Mueller, U.A., Corrigan, D.,2014. A study of the lake sediment geochemistry of

804 the Melville Peninsula using multivariate methods: Applications for predictive

805 geological mapping. Journal of Geochemical Exploration, 141, 15-41. doi

806 10.1016/j.gexplo.2013.07.013.

807 Johnson, C.C., 2005. 2005 G-BASE field procedures manual. British Geological Survey, 65pp.

808 (IR/05/097) (Unpublished).

809 Lagacherie, P., McBratney, A.B., Voltz, M., (Eds.) 2007. Digital Soil Mapping, an introductory

810 perspective. Developments in Soil Science, 31. Amsterdam.

811 Lancianese, V., Dinelli, E., 2014 in press. Different spatial methods in regional geochemical

812 mapping at high density sampling: An application on stream sediments of Romagna

813 Apennines, Northern Italy. J. Geochemical Exploration

814 http://dx.doi.org/10.1016/j.gexplo.2014.12.014

815 Lech, M.E., Caritat, P. de, McPherson, A.A., 2007. National Geochemical Survey of Australia:

816 Field Manual. Geoscience Australia Record, 2007/08, 53 pp.

817   Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in

818   environmental sciences: Performance and impact factors. Ecological Informatics, 6,

819   228-241.

820   Madeisky, H.E., Stanley, C.R., 2010. Lithogeochemical exploration of meta- somatic zones

821   associated with volcanic-hosted massive sulfide deposits using Pearce element ratio

822   analysis. International Geology Review, 35, 1121–1148.

823   McKinley, J.M., 2013, 'How useful are databases in environmental and criminal forensics?'

824   Geological Society of London Special Publication, 384, 109-119.,

825   http://dx.doi.org/10.1144/SP384.9

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

827   Survey of Northern Ireland, Belfast.

828   Montreuil, J. -F., Corriveau, L., Grunsky, E.C., 2013. Compositional data analysis of

829   hydrothermal alteration in IOCG systems, Great Bear magmatic zone, Canada: To each

830   alteration types its own signature, Geochemistry: Exploration, Environment, Analysis

831   vol. 13, no. 4, 2013; p. 229-247, doi:10.1144/geochem2011-101.

832   Morris, P.A., Pirajno, F., and Shevchenko, S., 2003. Proterozoic mineralization identified by

833   integrated regional regolith geochemistry, geophysics and bedrock mapping in

834   Western Australia. Geochemistry: Exploration, Environment, Analysis, Geological

835   Society of London, 3 (1), 13-28.

836   Ohta, A., Imai, N., Terashima, S., Tachibana, Y., 2011. Regional geochemical mapping in

837   eastern Japan including the nation's capital, Tokyo. Geochemistry: Exploration,

838   Environment, Analysis, 11, 211–223.

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

840   Applications. Wiley, Chichester.

841   Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and analysis of

842   compositional data. Wiley, Chichester.

843   Pearce, T.H., 1968. A contribution to the theory of variation diagrams. Contributions to

844   Mineralogy and Petrology, 19, 142–157.

845   Reimann, C. and Filzmoser, P., 2000. Normal and lognormal data distribution in

846   geochemistry: death of a myth. Consequences for the statistical treatment of

847   geochemical and environmental data. Environmental Geology, 39(9), 1001–1014.

848   Reimann C., Siewers U., Tarvainen T., Bityukova L., Eriksson J., Gilucis A., Gregorauskiene V.,

849   Lukashev V.K., Matinian N.N., Pasieczna A., 2003. Agricultural Soils in Northern

850   Europe: A Geochemical Atlas. Geologisches Jahrbuch, Sonderhefte, Reihe D, Heft SD 5,

851   Schweizerbart'sche Verlagsbuchhandlung, Stuttgart.

852   Reimann, C., 2005. Geochemical mapping – technique or art? Geochemistry, Exploration,

853   Environment, Analysis 5, 359-370.

854   Reimann, C., Garrett, R.G., 2005. Geochemical background – concept and reality. Science of

855   the Total Environment 350, 12-27.

856   Reimann, C., Filzmoser, P., Garrett, R.G. 2005. Background and threshold: critical

857   comparison of methods of determination. The Science of the Total Environment, 346,

858   1–16.

859   Reimann, C., Filzmoser, P., Garrett, R., Dutter, R., 2008. Statistical Data Analysis Explained:

860   Applied Environmental Statistics with R. Wiley, Chichester.

861   Reimann, C., ÄyrÄs, M., Cheklushin, V., Bogatyrev, I., Boyd, R., Caritat, P. de, Dutter, R.,

862   Finne, T.E., Halleraker, J.H., Jæger, Ø., Kashulina, G., Lehto, O., Niskavaara, H., Pavlov,

863   V., Räisänen, M.L., Strand, T., Volden, T., 1998. Environmental Geochemical Atlas of

864   the Central Barents Region. ISBN 82-7385-176-1. NGU-GTK-CKE Special Publication,

865   Geological Survey of Norway, Trondheim, Norway.

866   Reimann,C., Matschullat, J., Birke, M., Salminen, R., 2010. Antimony in the environment –

867   lessons from geochemical mapping. Applied Geochemistry *25, 175-198.*

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

869   of Europe's agricultural soils – Part A: Methodology and interpretation of the GEMAS

870   data set. Geologisches Jahrbuch (Reihe B 102), Schweizerbarth, Hannover. +DVD

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

872   of Europe's agricultural soils – Part B: General background information and further

873   analysis of the GEMAS data set. Geologisches Jahrbuch (Reihe B 103), Schweizerbarth,

874   Hannover.

875   Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E.,

876   Ladenberger, A., GEMAS PROJECT TEAM, 2012. The concept of compositional data

877   analysis in practise – Total major element concentrations in agricultural and grazing

878   land soils of Europe. Science of the Total Environment 426, 196-210.

879   Salminen, R., Tarvainen, T., Demetriades, A., Duris, M., Fordyce, F.M., Gregorauskiene, V., Kahelin, J.

880   Kivisilla, H., Klaver, G., Klein, H.,Larson, J.O., Lis, J., Locutura, J., Marsina, K., Mjartanova, H.,

881   Mouvet, C., O'Connor, P., Odor, L., Ottonello, G., Paukola, T., Plant, J.A., Reimann, C.,

882   Schermann, O., Siewers, U., Steenfelt, A., Van der Sluys, J., de, Vivo B., Williams, L., 1998.

883   FOREGS geochemical mapping field manual. Geological Survey of Finland, Guide 47.

884 Smith R.J., Dransfield M.H., Rajagopalan S., 1997. Regional Geophysical Exploration in Arid

885   Terrain. In Proceedings of Exploration 97: Fourth Decennial International Conference

886   on Mineral Exploration. Gubins A.G. (Ed.) Applications of Regional Geophysics and

887   Geochemistry, 105, 821–828

888 Smith, D.B., Cannon, W.F., Woodruff, L.G., 2011. A national –scale geochemical and

889   mineralogical survey of soils of the conterminous United States. Applied

890   Geochemistry, 26, S250-S255.

891 Templ, M., Hron, K., Filzmoser, P. 2011. robCompositions: an R-package for robust statistical

892   analysis of compositional data. In V. Pawlowsky-Glahn and A. Buccianti (Eds.),

893   Compositional Data Analysis. Theory and Applications. Wiley, Chichester, 341-355.

894 Thomas, C. W., Aitchison, J. 2005. Compositional data analysis of variability and process: A

895   case study. Mathematical Geology 37, 753-772.

896 Tolosana-Delgado, R., van den Boogaart, K.G. 2013. Joint consistent mapping of high-

897   dimensional geochemical surveys. Mathematical Geosciences 45, 983-1004.

898 Tolosana-Delgado, R., McKinley, J., van den Boogaart, K. G., 2015. Geostatistical Fisher

899   discriminant analysis, Proceedings of IAMG 2015, Freiberg: IAMG office, 978-3-00-

900   050337-5

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

902   Springer, Heidelberg.

903 Wang, X. and the CGB Sampling Team, 2015. China geochemical baselines: Sampling

904   methodology. Journal of Geochemical Exploration, 148, 25-39.

905   Young, M.E., Donald, A.E., 2013. A Guide to the Tellus Data. Belfast: Geological Survey of

906   Northern Ireland, Belfast.